МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Тернопільський національний економічний університет
Факультет комп'ютерних інформаційних технологій
Кафедра комп'ютерної інженерії

ЛУЦИК Юлія Андріївна

Алгоритми класифікації гістологічних і цитологічних зображень на основі методу Random Forest / Histological and Cytological Images Classification Algorithms Based on Random Forest Method

спеціальність: 123 - Комп'ютерна інженерія
магістерська програма - Комп'ютерна інженерія

Магістерська робота

Виконала студентка групи КІм-21
Ю. А. Луцик
_____

Науковий керівник:
д.т.н., професор, О. М. Березький
_____

Магістерську роботу допущено
до захисту:

"11" 01 2018 р.

Завідувач кафедри
_____ **О. М. Березький**

ТЕРНОПІЛЬ - 2018

Тернопільський національний економічний університет
Факультет комп'ютерних інформаційних технологій
Кафедра комп'ютерної інженерії
Освітній ступінь «магістр»
спеціальність: 123 - Комп'ютерна інженерія
магістерська програма - Комп'ютерна інженерія

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ О.М. Березький
"11" 11 2016 р.

ЗАВДАННЯ
НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ
Луцик Юлії Андріївни
(прізвище, ім'я, по батькові)

1. Тема магістерської роботи «Алгоритми класифікації гістологічних і цитологічних зображень на основі методу Random Forest / Histological and Cytological Images Classification Algorithms Based on Random Forest Method» керівник роботи д.т.н., проф. О. М. Березький
затверджені наказом по університету від 11 жовтня 2016 р. № 669.

2. Строк подання студентом роботи «15» січня 2018 року.

3. Вихідні дані до магістерської роботи
Об'єкт дослідження – процес обробки гістологічних та цитологічних зображень.
Предмет дослідження – метод Random Forest для класифікації біомедичних зображень.

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити):
- дослідити цитологічні та гістологічні зображення раку молочної залози;
- ознайомитись із сучасними алгоритмами розпізнавання;
- проаналізувати програмно-апаратний комплекс;
- проаналізувати різновиди методу Random forest;
- розробити програмний модуль на основі методу Random forest;
- підготувати вхідні дані;
- виконати тестування розробленого модуля.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень):
- тема, мета, завдання, методи досліджень, наукова новизна, практичне

2

значення;

– об'єкт і предмет дослідження;
– актуальність;
– аналіз цитологічних та гістологічних зображень;
– методи класифікації зображень;
– автоматична класифікація та методи створені на основі цієї класифік
– алгоритм класифікації методом Random forest;
– структура програмного модуля;
– приклад класифікації цитологічних зображень за допомогою алгори Random forest;
– приклад класифікації гістологічного зображення за допомо алгоритму Random forest;
– приклад тестової вибірки діагностування для класифікації раку;
– вікно тестування програмного модуля.

6. Консультанти розділів магістерської роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдан прийн
Антиплагіат	Мельник Г.М., доцент		
Нормо-контроль	Гураль І. В., викладач		

7. Дата видачі завдання «21» 11 2017 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів магістерської роботи	Строк виконання етапів магістерської роботи	Пр
1	Аналіз методів та алгоритмів розпізнавання гістологічних та цитологічних зображень	3.11.2016 – 1.01.2017	
2	Алгоритми класифікації зображень на основі методу Random forest	2.01.2017 – 31.05.2017	
3	Програмна реалізація та дослідження алгоритмів Random Forest	1.06.2017 – 25.01.2018	
4	Нормо-контроль, попередній захист	16.01.2018 – 2.02.2018	
5	Захист	5.02.2017	

Студент _____ Луцик Ю. А.
(підпис)

Керівник магістерської роботи _____ д.т.н., проф., О. М. Берез
(підпис)

ЗМІСТ

ВСТУП

На етапах постановки задачі машинного навчання і формування даних не завжди зрозуміло, які ознаки важливі для побудови оптимального алгоритму, тому часто в даних використовується багато лишньої (шумової) інформації.

Поява цієї шумової інформації погіршує якість роботи алгоритму і сповільнює його роботу. Тому в багатьох випадках перед розв'язком задачі класифікації, регресії і прогнозування потрібно вибирати ті ознаки, які найбільш інформативні. Відбір важливих ознак (наприклад, відбір генів, відповідних певному типу раку) може допомогти розшифрувати механізми дослідження [1].

Дерево рішень – це один з методів автоматичного аналізу великих масивів даних. Ідея, яка лежить в основі методу дерева рішень, полягає у розбитті множин можливих значень вектора ознак (незалежних змінних) на множини, які не перетинаються і підгонці простої моделі для кожної такої множини. Отже, мета цього методу полягає в тому, щоб створити модель, яка прогнозує значення цільової змінної на основі декількох змінних на вході.

Метою роботи є розробка алгоритмів класифікації гістологічних та цитологічних зображень на основі методу Random Forest

Об'єктом дослідження є розпізнавання гістологічних та цитологічних зображень. Предметом дослідження є метод Random Forest.

Досягнення поставленої мети реалізовано з використанням методів інтелектуального аналізу даних та машинного навчання.

Наукова новизна дипломної роботи полягає в тому, що було запропоновано та відлагоджено алгоритм, що досягає найкращої точності для поставленої задачі, а саме визначає клас до якого належить зображення.

Актуальність теми дослідження полягає в тому, що сучасний стан визначення захворювання клітин раком вимагає точних і якісних результатів.

1 АНАЛІЗ МЕТОДІВ ТА АЛГОРИТМІВ РОЗПІЗНАВАННЯ ГІСТОЛОГІЧНИХ ТА ЦИТОЛОГІЧНИХ ЗОБРАЖЕНЬ

1.1 Аналіз гістологічних та цитологічних зображень

Використання математичних методів є необхідним для вдосконалення наукових досліджень в біології, фармакології та медицині.

Патологічні процеси на тканинному та клітинному рівнях дослідження зазвичай проводяться за допомогою оптичної мікроскопії. Нечіткість та слабка контрастність біологічних та біомедичних зображень є основними проблемами в задачах вимірювання та розпізнавання. Для їх подолання потрібні професійні навики дослідника з коригування результатів. Автоматизація процесу обробки гістологічних та цитологічних досліджень дозволить істотно підвищити продуктивність медичного персоналу, що працює з гістологічними препаратами.

Одним з додатків комп'ютерної обробки даних є цифрова обробка зображень. Теоретичні дослідження в цій області почалися в 1960-і рр. і ґрунтувалися на методах поліпшення якості зображення і завданнях дистанційного зондування. З плином часу область застосування обробки зображень значно розширилася. Так, в якості одного з розвитку напрямів виділився аналіз біомедичних зображень.

Технічне обладнання, що застосовується для дослідження зображень клітин і клітинних систем, легко перерахувати. Це цитофотометри, які використовують інтерактивне виділення клітин, а також апарати для аналізу крові та каріотипування, що працюють з висококонтрастними зображеннями. Основною причиною повільного розвитку засобів автоматизації в гістології є висока варіабельність і слабка контрастність більшості гістологічних структур.

Однак останнім часом швидкий розвиток цифрової і аналогової техніки відкриває нові можливості перед розробниками. Наприклад, збільшення швидкодії обчислювальної техніки дозволяє використовувати складні, критичні до часу алгоритми, а завдяки появі кольорових телевізійних датчиків високої роздільної здатності можна отримувати і обробляти кольорові зображення. Засоби введення зображень в комп'ютер з кожним днем все більше вдосконалюються. Сучасні сканери дозволяють вводити повноколірні зображення з великою глибиною яскравості (16 біт на піксель). Останнім часом активно впроваджуються цифрові камери, мікроскопи, які дозволяють вводити повноколірні зображення розміром до 6 000 х 6 000 точок, в той час як аналогові відеокамери не дозволяють формувати сигнал з розгорткою більше 760 рядків.

Саме сучасні технічні можливості дозволяють значно розширити коло досліджень і відкривають нові шляхи вирішення завдань, що стосуються аналізу зображень. Використання засобів автоматизації сприяє підвищенню ефективності роботи дослідника і отримання більш якісних і точних результатів вимірювання характеристик об'єктів біомедичних зображень [2-5].

Біомедичні зображення прийнято класифікувати за способом їх отримання і галузі, до якої вони належать. Виділяють кілька видів біомедичних зображень, основними серед них є анатомічні (фотографії, рентгенівські знімки, зображення, отримані методами УЗД, ядерно-магнітного резонансу (ЯМР), моделі комп'ютерного томографа) і гістологічні (зображення оптичної та електронної мікроскопії) [6, 7].

Легкість отримання і досить високу якість аналізу анатомічних зображень обумовлюють велику кількість робіт з цієї тематики. Недостатній розвиток цієї галузі пов'язано з проблемами автоматичного виділення гістологічних об'єктів на зображеннях [2, 3, 8].

Відповідно до сучасних уявлень діагностика і подальше рішення про вибір тактики лікування будь-якого захворювання повинні бути засновані на

цілому комплексі параметрів, включаючи топографічні і візуально-прогностичні характеристики. Біомедичні зображення, будучи джерелом цих параметрів, забезпечують інформацію про анатомічний і функціональний стан організму. Технічний прогрес постійно розширює рамки можливостей отримання зображень, додаючи все нові їх типи, що, в свою чергу, призводить до розширення діагностичних і лікувальних можливостей. Слід підкреслити, що нерідко для отримання повної і об'єктивної картини захворювання використовується не один, а цілий комплекс візуально-діагностичних методів. Наприклад, при деяких захворюваннях нирок в єдиний діагностичний процес об'єднуються методи УЗД, КТ і магнітно-резонансної томографії.

У медичній практиці за характером отримання і області використання розрізняють три типи зображень:

– анатомічні;

– гістологічні (включаючи цитологічні);

– фізіологічні.

Фізіологічні зображення мають специфічні методи отримання і обробки, тому розглядатися не будуть.

В даний час існує досить багато способів отримання біомедичних зображень:

Анатомічні зображення можна отримати за допомогою рентгенографії, комп'ютерної томографії, магнітно-резонансної Томографії, УЗД, мікроскопії, термографії, ендоскопічними та оптичними методами. Для отримання гістологічних зображень використовуються термографія, електронна і оптична мікроскопія.

Всі зображення для вище перелічених способів сильно розрізняються за своїми характеристиками і вимагають спеціалізованої обробки. Зупинимося на зображеннях, отриманих методами оптичної мікроскопії, основну групу яких складають гістологічні зображення. У оптичної

мікроскопії існує кілька методів вивчення зображень: світле поле, темне поле, фазовий контраст, інтерференційні методи контрастування, флуоресцентні методи.

На гістологічних зображеннях оптичної мікроскопії визначити інформативні об'єкти непросто. Це пов'язано із слабоконтрастним зображенням клітин і клітинних структур, складністю організації тканини і її структури, наявністю в полі зору різної групи клітин і значною неоднорідністю тканини як фону (Рисунок 1.1).

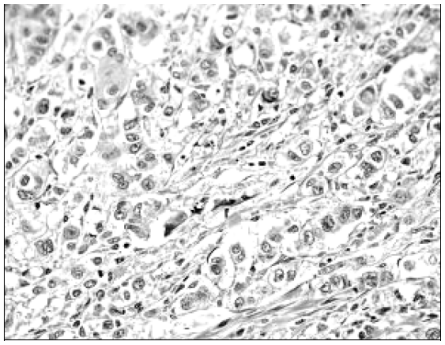


Рисунок 1.1 – Зображення оптичної мікроскопії

Зображення оптичної мікроскопії дуже сильно залежить від якості приготованого препарату і використовуваного обладнання. Тому зображення оптичної мікроскопії можуть бути як чорно-білими, так і кольоровими. Наявність великої кількості різнотипних об'єктів на гістологічних зображеннях є серйозним недоліком при аналізі цих зображень.

Як окремий клас зображень, отриманих в оптичній мікроскопії, можна виділити цитологічні зображення (Рисунок 1.2), які є окремим випадком гістологічних.

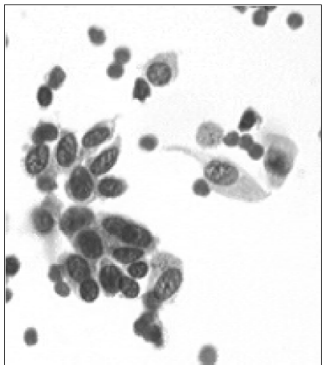
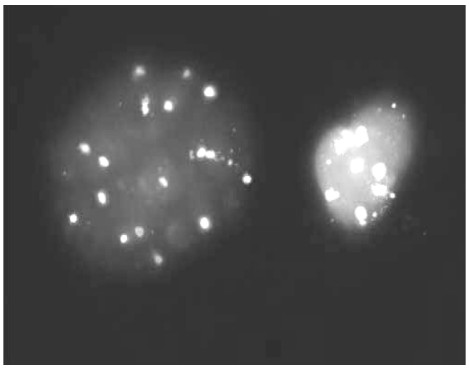
а)  б) 

Рисунок 1.2 – Цитологічні зображення: а) оптичної мікроскопії, б) флуоресцентного мікроскопа

На анатомічному зображенні оптичної мікроскопії представлені органи і їх фрагменти, найчастіше це зображення зародків, або ембріонів. На рівні тканини характеристика таких зображень нагадує гістологічну, з тією різницею, що орган або його фрагмент зазвичай геометрично локалізовані.

При обробці біомедичних зображень оптичної мікроскопії найбільш важливу роль відіграють характеристики зашумленості зображення, однорідність фону, оптичні та геометричні характеристики об'єктів (Таблиця 1.1.).

Таблиця 1.1 – Особливості зображень оптичної мікроскопії

Зображення оптичної мікроскопії	Тип зображення	Наявність великих шумів і артефактів	Нечіткі контури об’єктів	Геометричні спотворення	Наявність різнотипних об’єктів	Неоднорідність фону
Гістологічні	Напівтонове і кольорове	Присутні	Присутні	Можливі	Присутні	Присутня
Цитологічні		Можливі	Можливі	Можливі	Присутні	Присутня
Анатомічні		Присутні	Присутні	Присутні	Присутні	Присутня

Відомо, що клітини представляють собою елементарні одиниці, з яких побудовані живі організми, і що організм складається з різних тканин. Відмінності між тканинами полягають у призначенні клітин цих тканин для виконання певних функцій, необхідних організму.

Існує п'ять основних типів тканин [5, 9-11]: епітеліальна, сполучна, нервова, м'язова, системи тканин внутрішніх середовищ. Кожна з них має індивідуальну структуру і свої особливі функції. Достатньо лише вказати, що клітини кожної тканини структурно спеціалізовані для здійснення однієї або декількох функцій.

Клітини епітеліальної тканини зазвичай виконують захисні та секреторні функції. Тому вони розташовуються шарами і щільно прилягають один до одного.

Сполучна тканина має безліч функцій, але головним її завданням є підтримка цілісності інших тканин і живлення їх через власні кровоносні судини. Дуже часто зображення тканини визначається як фон з розташованими на ньому гістологічними об'єктами (клітинами, волокнами і судинами).

Нервова тканина регулює і координує фізіологічні процеси окремих тканин і систем організму, переробляє сигнали пам'яті, внутрішнього і зовнішнього середовища, володіє такими властивостями, як подразливість і провідність. Тонкі нервові волокна, що відходять від нервових клітин, тягнуться від головного і спинного мозку до всіх органів і тканин, забезпечуючи швидкий зв'язок між різними частинами організму, утворюючи білу речовину і периферичні нерви.

М'язова тканина також має складну структуру. М'язові волокна зібрані в пучки, які, в свою чергу, є складовою частиною більш складних об'єднань.

На основі структурних і геометричних характеристик можна виділити наступні типи гістологічних тканин:

– чіткої текстури (тканини з щільно прилеглими один до одного клітинами зі схожою геометричною структурою);

– складної текстури (тканини з щільно прилеглими один до одного клітинами з різною геометричною структурою);

– з окремо розміщеними клітинами, орієнтованими в просторі за формою;

– з «довільно» розміщеними клітинами (зі складно обумовленими геометричною структурою і положенням в просторі);

– з поздовжньо розміщеними волокнами, судинами або довгими клітинами, відносно орієнтованими в просторі;

– з поздовжньо розміщеними волокнами, судинами або довгими клітинами без явної орієнтації в просторі.

Зображення живих клітин досить стабільні і зручні для аналізу. Приблизно в центрі розташовано ядро, яке часто за своїм показником заломлення відрізняється від решти клітини. На пофарбованих препаратах видно, що ядро обмежено ядерною мембраною, або оболонкою. В ядрі є одне або кілька округлих темних тілець, які називаються ядерця.

Ядро розташоване в цитоплазмі, що займає зовнішню і зазвичай більшу частину клітини. Кожен компонент цитоплазми виконує свої певні функції, причому в різних клітинах ці компоненти мають різний вигляд в залежності від цих функцій. Слід зазначити, що за допомогою світлового мікроскопа в цитоплазмі виявляються лише деякі з її компонентів. Дослідження ж всіх компонентів і структури цитоплазми проводиться з використанням електронного мікроскопа. Цитоплазма містить безліч спеціалізованих ультраструктур, званих органеллами клітини, їх наявність обумовлює оптичну неоднорідність клітини.

Клітини розрізняються за своїми розмірами, існують деякі верхні і нижні межі розмірів: в більшості випадків діаметр клітини не виходить за межі 0,01 - 0,1 мм [11, 12].

На основі структурних і геометричних характеристик можна виділити наступні типи клітин:

– площинні об'єкти щодо правильної геометричної форми;

– площинні об'єкти довільної геометричної форми;

– сильно витягнуті об'єкти.

За оптичними характеристиками і методом фарбування клітини діляться на три класи:

– з пофарбованим ядром;

– з незабарвленим ядром;

– без ядра.

Однією з найважливіших характеристик клітин і ядер є їх величина [11]. На зрізах товщиною близько 7 мкм, отриманих з клітин діаметром більше 15 мкм, не обов'язково видно ядра. Коли на зріз потрапляє клітинне ядро, воно часто виглядає менше, ніж в дійсності. Крім того, розміри самої клітини на зрізі можуть здаватися менше, оскільки зображення залежить від того, яка частина клітини потрапила в площину зрізу. Тому клітини на зрізі можуть сильно відрізнятися за величиною, будучи насправді однаковими.

На багатьох зображеннях, отриманих з гістологічних зрізів, присутні різного роду перешкоди. Деякі з них класифікуються як артефакти [5, 9].

Артефактом (від лат. Art - мистецтво і factum - зроблений) називають те, що створено штучно. У гістології артефактами називають ознаки або структури, що виникають на препаратах випадково або внаслідок поганої обробки. Спотворення справжнього вигляду тканини можуть виникнути на будь-якому етапі виготовлення препарату. Оскільки артефакти іноді зустрічаються на препаратах, що вивчаються в лабораторії, корисно з самого початку навчитися впізнавати найбільш поширені з них, щоб надалі не брати до уваги [11, 12].

При аналізі будь-яке зображення ділиться на об'єкти і фон. Мета процесу сегментації - визначити області, що відповідають об'єктам, а решту частини зображення віднести до фону. Для того щоб правильно вибрати

методи сегментації, необхідно визначити об'єкти як набір ознак і характеристик.

В основу класифікації гістологічних зображень з метою автоматичного аналізу і вимірювань найзручніше покласти топологічні властивості гістологічного зображення в цілому. Це дозволить визначити порядок обробки зображення гістологічного препарату з метою виділення всіх об'єктів.

Одним з цікавих властивостей гістологічних препаратів є те, що в залежності від ступеня оптичного збільшення їх зображень одні об'єкти виділяються більш явно, а інші втрачають своє явне визначення. Кожне значення оптичного збільшення виділяє певну групу топологічних характеристик тканини і її складових. Тому найзручніше розглядати гістологічні об'єкти в залежності від ступеня збільшення гістологічних препаратів.

Загальне зображення гістологічних препаратів утворюється різними фрагментами тканини, що складаються з груп однорідних клітин або волокон (Рисунок 1.3). Ці фрагменти є об'єктами, які цікавлять дослідників, і зазвичай мають текстурний характер. Найчастіше для виділення фрагментів застосовуються методи зростання областей.

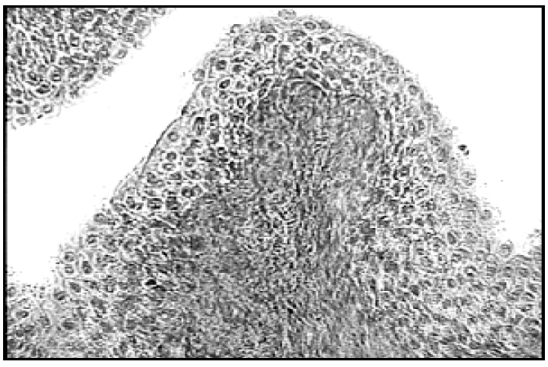


Рисунок 1.3 – Гістологічне зображення

Зображення тканини аналізується при невеликих оптичних збільшеннях, які дають можливість отримати образ тканини, що складається

з різних фрагментів. Тому, якщо розглядати обробку послідовно у вигляді дерева, цей вид зображень можна віднести до першого рівня.

На другому рівні обробки розташовуються зображення клітин, волокон і судин як досліджувані об'єкти, що утворюють фрагменти тканини. На цьому рівні починає грати свою роль співвідношення напівтонових характеристик фону і об'єктів. Найчастіше фон утворюють порожнечі, дуже дрібні клітини, волокна, будь-які частинки. Крім того, фон зазвичай включає в себе різні перешкоди і шуми, що виникають при отриманні зображення. Тому зображення фону характеризується приблизно однаковим рівнем яскравості пікселів, за винятком поодиноких викидів, утворених шумами. Яскравість пікселів зображення фону залежить від розподілу щільності тканини, якості зрізу та освітлення і наявності електронних перешкод [9].

Більш складним об'єктом є зображення клітини. Для геометричної характеристики клітини важливим параметром служить її ширина, клітини також відрізняються по геометричній формі. Крім того, при ідентифікації тканини велику роль відіграє розміщення клітини в тканині. Однак для клітин характерно безліч геометричних форм, отже, класифікація їх за формою на початковому етапі не є оптимальною. Методи, які вибираються для сегментації клітини в тканині, залежать від напівтонових характеристик зображень клітини, фону і їх взаємозв'язку. За цими характеристиками зображення клітин можна розділити на три групи:

1. Зображення окремо лежачих клітин одного типу, причому для кожного пікселя цих об'єктів значення яскравості не може збігатися зі значенням кожного пікселя, що належить фону. У цьому випадку застосовується порогова сегментація.

2. Зображення окремо лежачих клітин одного типу, при цьому яскравість пікселів фону змінюється рівномірно і не робить різких стрибків. Застосовується морфологічна сегментація.

3. Зображення клітин з яскравістю пікселів фону, здатних приймати будь-яке значення, причому на зображенні є інші об'єкти (клітини, волокна, шуми). Сегментація виконується за допомогою об'єднання областей.

Зрозуміло, використовуючи геометричні та топологічні характеристики клітин, можна продовжити класифікацію клітини з метою визначення її типу: за формою, розмірами, складом (наявності ядерець або включень).

Для зручності класифікації зображень клітин і волокон в поздовжньому розрізі вони розглядаються як площинні об'єкти, а ядра і радіальна неоднорідність всередині волокон визначаються аналогічно включень.

Наступним типом в ієрархії зображень будуть зображення клітинних ядер, основною характеристикою яких є те, що ядра представляються вкладенями в зображення клітини. Тому, з огляду на співвідношення зображень ядра і клітини, ядра можна розділити на три групи, такі ж, як і для клітин.

Подібна класифікаційна схема об'єктів і структур на гістологічних зображеннях найбільш сприятлива не тільки для автоматичного аналізу тканини і її складових, а й для вибору методів виділення об'єктів на гістологічних зображеннях.

Цитологія – це наука, що вивчає клітини. Завдання цитолога - встановити, як виглядає жива клітина і як вона виконує свої функції. Один з методів вивчення клітини - мікроскопування. Сучасний світловий мікроскоп збільшує об'єкти в 3 000 разів і дозволяє бачити найбільші органели клітини, спостерігати рух цитоплазми, поділ клітин.

Безумовно, в гістологічному препараті клітини є одними з основних елементів, що роблять тканину видимої. Зображення клітин поза тканиною прийнято вважати цитологічними. Цитологічні зображення отримують на основі всіляких мазків. Даний клас зображень нагадує гістологічні зображення клітинного рівня. Основною його відмінністю є фон, який оточує клітини. Якщо в гістологічних зображеннях фон складається переважно з

елементів тканини, то в цитологічних він не володіє явними поглинають світло властивостями. Його можна вважати оптично однорідним [13].

Головним об'єктом на цитологічному зображенні є клітина, тому класифікація об'єктів на цих зображеннях збігається зі схемою гістологічних зображень для окремо лежачих клітин.

Слід зазначити, що зображення клітин на цитологічних препаратах бувають двох типів: з окремо лежачими клітинами і з клітинами, об'єднаними в конгломерати. У другому випадку дослідження клітин ускладнюється через необхідність їх поділу, а в разі нашарування клітин одну на одну дослідження ускладнюється настільки, що в багатьох випадках не виходить задовільний результат.

Дослідження клітини мають велике значення для діагностики захворювань, так як саме в клітинах починають розвиватися патологічні зміни, що призводять до виникнення хвороби.

1.2 Аналіз методів та алгоритмів класифікації зображень

Можливості ранньої і точної діагностики, а отже, лікування, в останні роки різко зросли. В значній мірі це пов’язано з розвитком різних методів дослідження, які дають лікарю зображення нормальних та патологічних змін органів і тканин – медичні діагностичні зображення. Медичні зображення органів (medical imaging) – головне джерело інформації при встановленні діагнозу. З швидким зростанням загального рівня комп’ютеризації та технічного оновлення медико-профілактичних закладів України гостро постала проблема систематизувати набуту графічну інформацію, отриману в процесі діагностики, лікування та профілактики.

Автоматичні системи аналізу зображення були розроблені порівняно недавно. Їх робота заснована на методах, які відносять до розділів технічного зору, а тривала мініатюризація інформаційної техніки дозволяє створювати

малогабаритні матричні фотоприймачі з досить високою роздільною здатністю і спеціалізованими процесами обробки відеоінформації [13]. На даний момент не всі автоматичні системи можуть аналізувати об'єкти з нечіткими контурами, а так як більшість об'єктів є розмитими, аналізатори зображення включають можливість контролю обробки з боку користувача. Незважаючи на це, у автоматичних систем є багато переваг. По-перше, кількісний облік та класифікація ознак більш об'єктивні. По-друге, розрахунки і вимірювання виконуються набагато швидше, ніж вручну. По-третє, працюючи тривалий час, дослідник втомлюється і починає пропускати потрібні структури і допускає помилки; це виключено при використанні автоматики.

В даний час розрізняють три типи систем аналізу і обробки зображень:

– загального призначення;

– загального призначення, доповнені спеціалізованими можливостями;

– спеціалізовані системи для вирішення приватних завдань.

Для реалізації набору цих функцій використовуються наступні математичні операції:

– лінійна і нелінійна фільтрація;

– арифметико-логічні операції;

– математична морфологія;

– порогова сегментація;

– інтерактивні вимірювання;

– автоматичні вимірювання;

– статистичний аналіз.

Окремі системи аналізу зображень включають в себе Фур'є-аналіз і обробку в спектральній області. Найбільш відомі системи цього типу - VIDAS, VIDEOPLAN (Kontron electronics), Кантінент 720, Quantinent 2000 (Leica). Останнім часом найбільш відомі і популярні наступні системи:

Optimas (Macromedia Inc.), Діаморф (Діаморф, Росія, Москва), KS-500 (Kontron electronics), Халькон (MVTec). Між собою системи відрізняються наборами методів порогової сегментації, наявністю складних операцій математичної морфології і наборами виміряних параметрів.

Так як задачі розпізнавання об'єктів заключаться в класифікації зображень на основі певних критеріїв, то важливим етапом є вибір оптимального класифікатора. Серед існуючих методів класифікації можна виділити наступні:

– ймовірностний критерій якості класифікації;
– оптимальна стратегія статистичної класифікації;
– класифікатор Байєса;
– мінімаксний класифікатор;
– класифікатор Неймана-Пірсона.

Класифікатор Байєса є найбільш широко розповсюдженим з перелічених класифікаторів і застосовується при наявності повної апріорної інформації про класи, тобто коли відомі функція правдоподібності для кожного з класів, матриця штрафів, апріорні ймовірності для кожного з класів [20]. Класифікатор Байєса ґрунтується на основі принципу максимуму апостеріорної ймовірності, що базується на трьох гіпотезах:

1. Множина $X \times Y$ є імовірнісним простором з імовірнісною мірою p. Прецеденти $(x_1, y_1), \dots, (x_1, y_1)$ з'являються випадково і незалежно у відповідності з розподілом P.

2. Відомі щільності розподілу класів $p_y(x) = p\left(x\middle|K_y\right), y \in Y$, що називаються функціями правдоподібності.

3. Відомі ймовірності появи об'єктів кожного з класів $P_y = P\left(K_y\right), y \in Y$, що називаються апріорними ймовірностями.

Базуючись на даних гіпотезах, принцип максимуму апостеріорної ймовірності записується в наступному вигляді:

$$F(x) = arg\max_{y \in Y} P\left(K_y \middle| x\right) = \arg\max_{y \in Y} p_y(x) P_y \quad (1.1)$$

Доведено, що такий вибір вирішального правила є оптимальним з погляду мінімізації загального ризику. Основна проблема, полягає в тому, що на практиці гіпотези 2 і 3 майже ніколи не виконуються. Спроби оцінити ці функції розподілу по навчальній вибірці могли б привести до деякого результату, якби не погана обумовленість задачі, що приводить до вироджених рішень.

Існують системи виявлення об'єктів зображення, що базуються на "наївному" баєсовому методі. Даний метод ґрунтується на побудові емпіричної щільності розподілу ймовірностей класів по навчальній вибірці за припущення про незалежність компонентів вектора ознак [21].

Практичні результати є наступними:

1. Алгоритм побудови "наївного" баєсового класифікатора схильний до перенавчання;

2. Алгоритм побудови "наївного" баєсового класифікатора чутливий до шуму, тому що базується на емпіричних функціях щільності розподілу;

3. Швидкість роботи самого класифікатора висока, основний час може займати обчислення вектора ознак.

На основі сполучення баєсового підходу і теорії графів утворюють байєсові мережі. Суть даного підходу в тому, що будують граф, кожна вершина якого відповідає якому-небудь компоненту вектора ознак, дуги позначають причинно-наслідковий зв'язок. Побудова мережі може бути здійснена автоматично шляхом аналізу кореляції компонентів вектора ознак.

Проблемою для баєсової мережі є погана обумовленість, тому що велика розмірність вектора ознак робить граф зв'язків складним для побудови й аналізу. Також сильно зростає обчислювальна складність. Одним з варіантів розв’язання даної проблеми є скорочення розмірності вектора ознак, що приводить до погіршення узагальнюючої здатності [22].

Класифікація за статистичними признаками можлива після попередньої обробки зображення за допомогою фільтрів, що дозволяють зменшити варіацію параметрів ознак.

Наряду з статистичними методами класифікації в обробці зображень використовують евристичні методи. Це ряд підходів, що можна розділити на наступні групи.

Евристичні методи є історично найбільш ранніми, вони досить прості в реалізації і працюють з високою швидкістю, однак жорстко запрограмовані правила позбавляють систему гнучкості і стійкості. Як правило, евристичні системи орієнтовані на відносно вузький клас задач.

Метод порівняння з шаблоном є більш універсальним підходом, однак даний метод вимагає наявності дуже точного шаблона об'єкта зображення. Шаблон може бути складною структурою і допускати різні деформації і перетворення, таким чином, сприяючи інваріантості системи до просторових спотворень об'єкта зображення і змін освітленості. Системи, засновані на порівняння з шаблоном, найчастіше використовуються для розв’язання задач відстеження об'єктів у відео з ініціалізацією на першому кадрі – до початку роботи системи існує загальна модель шаблона, при ініціалізації вона уточнюється і коректується під час роботи системи. При добре заданому шаблоні досягається висока точність і дуже низький рівень збоїв. Інваріантість до просторових спотворень і зміни освітлення залежить від складності шаблона.

Методи з навчанням по прецедентах є найбільш загальним підходом. Задача розпізнавання об'єктів зображення зводиться до задачі класифікації і для неї застосовується добре розроблений математичний апарат побудови моделі (навчання) по прецедентах. Модель будується автоматично по заздалегідь зібраному наборі прецедентів – зображень, для яких відомо, чи є вони зображеннями об'єкта чи ні. Спостереженням, у даному випадку, є деякий "вектор ознак", отриманий з вихідного зображення за допомогою перетворення, що відображає зображення в просторі дійсних векторів.

Гіпотеза, що підлягає перевірці – приналежність зображення до класу зображень шуканого об'єкта. Таким чином, система розпадається на два модулі – модуль перетворення зображення у вектор ознак і модуль класифікації. Задачею модуля перетворення є найбільш повне й інформативне представлення зображення у виді числового вектора. Задачею модуля класифікації є перевірка гіпотези приналежності зображення класу зображень об'єкта на підставі спостереження, яким є вектор ознак. Серед евристичних найбільш поширені методи:

- опорних векторів (SVM, supporting vectors method);
- Sparse network of Winnows (SNoW);
- посилення слабких класифікаторів (classifier boosting).

Метод SVM полягає в побудові оптимального лінійного класифікатора. Класичний алгоритм полягає в побудові лінійної поверхні (гіперплощини), яка рівновіддалена від опуклих оболонок класів, опукла оболонка будується по прецедентах. Стверджується, що така гіперплощина буде оптимальна, з погляду загального ризику, щодо будь-яких інших можливих гіперплощин. Якщо така гіперплощина не існує (класи лінійно не роздільні), то для здійснення нелінійної класифікації застосовується ядрове перетворення, що проектує вихідний простір у простір ще більшої, можливо нескінченної, розмірності [23].

Метод опорних векторів був успішно застосований для задачі розпізнавання об'єктів зображення. Метод характеризується наступним:

– висока стійкість до перенавчання;

– чутливість до шуму може регулюватися за рахунок зменшення точності;

– в системах розпізнавання об'єктів метод дає прискорення в декілька разів у порівнянні з нейронними мережами.

SNoW – особливий вид нейронної мережі. Вектор ознак бінарний. Мережа складається з двох (по числу можливих класів) лінійних нейронів,

зв'язаних з компонентами вектора ознак. Класифікація проходить за принципом “переможець забирає все”. Метод SNoW вважається досить ефективним методом для розв’язання задач виявлення об'єктів зображення. Відповідно до деяких досліджень SNoW перевершує по своїм параметрам метод опорних векторів.

Classifier boosting – це підхід до розв’язання задачі класифікації, шляхом комбінування примітивних класифікаторів в один більш сильний класифікатор. Основна ідея методу полягає в ітеративній мінімізації опуклого функціонала помилки класифікації, шляхом додавання в набір класифікаторів чергового слабкого класифікатора. Для систем розпізнавання об'єктів зображення був застосований каскадний підхід, який полягає в побудові каскаду із комплексу слабких класифікаторів, що працює за принципом послідовних наближень. Характеристики каскадного методу перевершують всі інші системи [24].

По показникам роботи в реальних системах розпізнавання об'єктів зображення найбільш вдалими виявилися алгоритми boosting (посилення слабких класифікаторів) і SNoW. Обидва підходи забезпечують високу швидкість, високий рівень розпізнавання і низький рівень похибок другого роду. Метод опорних векторів досить сильно уступає по показниках перерахованим вище підходам, тому що має дуже низький відсоток вірних виявлень.

Випадковий ліс (random forest) – алгоритм машинного навчання, запропонований Лео Брейманом і Адель Катлер, що полягає у використанні ансамблю дереврішень [14]. Алгоритм поєднує в собі дві основні ідеї: метод баггінга (bagging – від bootstrap aggregation, тобто дерева навчаються по випадковим підвибіркам) і метод випадкових підпросторів. Алгоритм застосовується для задач класифікації, регресії і кластеризації.

На сьогодні найбільш поширений і використовуваний метод дешифрування це візуальне дешифрування знімка. Однак цей метод є трудомістким і досить тривалим, тому актуальним є дослідження способів

автоматичного дешифрування (автоматичної класифікації). Автоматичною класифікацією називають процес розбиття пікселів неперервного растрового зображення на категорії на основі їх спектральних значень, в результаті чого кожному пікселю присвоюється нове значення [15]. На даний час існують два підходи у реалізації автоматичної класифікації: керована класифікація (класифікація „з учителем") та некерована (класифікація „без учителя", кластеризація). На основі цих підходів створено багато методів, основні з яких показані на рис 1. При керованій класифікації відбувається аналіз пікселів у межах кожного еталонного полігона і створення спектральних сигнатур для кожного типу покриття. За порівнянням спектральних значень пікселів зі створеними сигнатурами виконується класифікація зображення (Рисунок 1.4).

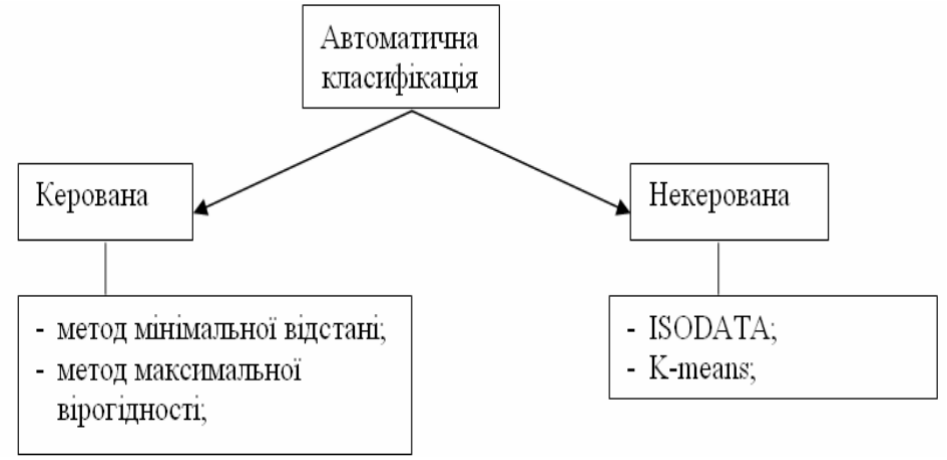


Рисунок 1.4 – Основні методи автоматичної класифікації

Класифікація за методом мінімальної відстані полягає в розрахунку евклідової відстані значень відбиття піксела до середнього спектрального значення кожного сигнатури. Піксел призначається до класу, відстань до котрого є найменшою. Класифікація за методом максимальної вірогідності вважається однією з оптимальних, оскільки базується на ймовірністних принципах. Дисперсія значень відбиття в еталонному полігоні описується функцією імовірності щільності, яка базується на статистиці Байєса [16].

Алгоритми некерованої класифікації (їх часто називають алгоритмами кластеризації) застосовують за відсутності апріорної інформації про об'єкт зйомки. Кластерний аналіз дозволяє виділяти контури з неконтрастною по спектральній яскравості структурою, наприклад рослинність, відкриті грунти, воду, хмари та інші об'єкти. З використанням алгоритмів кластеризації виконується автоматичне розділення зображення на групи пікселів, подібних за спектральним характеристикам (кластери). Ці алгоритми потребують мінімум початкової інформації (число класів, кількість ітерацій) [17].

Кластеризація зображення за алгоритмом ISODATA грунтується на різниці між середніми значеннями кластерів (мінімальній спектральній відстані між центрами класів) [18]. Метод K-means є подібним до методу ISODATA. Головна відмінність алгоритмів ISODATA і K-means полягає в тому, що на стадії ініціалізації алгоритму ISODATA відбувається розподіл пікселів, тоді як для алгоритму K-means відбувається розподіл значень математичних очікувань. В таблиці 1.2 подані основні характеристики методів автоматичної класифікації.

Таблиця 1.2 – Основні характеристики методів автоматичної класифікації

Методи	Потреба у «навчанні»	Швидкість	Переваги	Недоліки
Метод мінімальної відстані	+	Швидкий	Після класифікації немає некласифікованих пікселів	Не враховується дисперсія між сигнатурами еталонних полігонів
Метод максимальної вірогідності	+	Повільний	Вважається найбільш точним, так як працює на ймовірнісних принципах	Сигнатури з великим значенням коваріації сильно підкреслюються і тому потребують нормалізації

ISODATA	-	Швидкий	Не потребує попереднього навчання та менш залежний від людського чинника	Невідповідність створених кластерів потрібним класам, тому вимагає подальшого об’єднання або розбиття кластерів користувачем
K-means	-	Швидкий	Досить добре працює з частково «навченими» кластерами, тобто для частини точок відомо, до якого класу вони належать	Потрібно точно знати необхідну кількість кластерів, тому спочатку використовують інші методи кластеризації, де отримують кількість кластерів і початкове розбиття

Алгоритми керованої класифікації доцільно застосовувати тоді, коли є додаткова інформація про об’єкти на знімку. Якщо така інформація відсутня, то слід користуватися алгоритмами некерованої класифікації. Для підвищення точності дешифрування часто використовують спільно і керовану і некеровану класифікації.

1.3 Програмні засоби обробки біомедичних зображень

Для диференційної діагностики тиреоїдних захворювань розроблена система «Контур», що функціонує на базі персонального комп'ютера, світлового мікроскопа, кольорової телекамери і захоплювача кадрів. Поряд з високими технічними характеристиками система має потужну програмну підтримку.

Система «Контур» є додатком Microsoft Windows, побудованим на основі інтерфейсу складових документів. Система підтримує завантаження цитологічних растрових зображень основних графічних форматів. Єдина графічна оболонка дозволяє одночасно проводити обробку зображень, сегментацію і морфометричний аналіз.

Система надає можливість морфометричної оцінки наступних об'єктів:

– клітин лімфоїдного ряду тканини щитовидної залози (одиноких об'єктів);

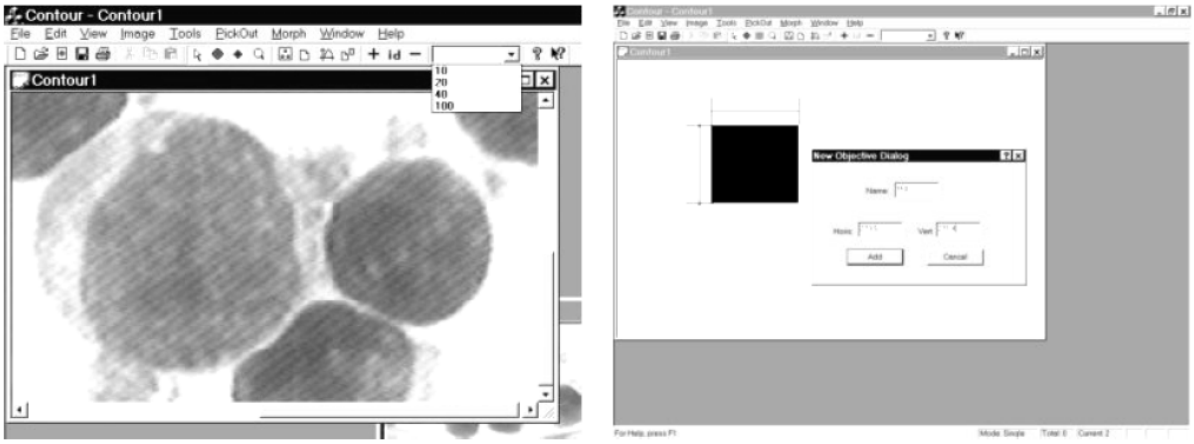
– ядер тиреоцитів (пар об'єктів);

– агрегатів тиреоцитів (складових об'єктів).

Робота програми може відбуватися в ручному, напівавтоматичному або автоматизованому режимах. Основним режимом роботи є автоматизований.

При роботі в ручному режимі користувач має можливість одночасно проводити операції над растровим і векторним вмістом документа, що є безперечною перевагою розробленої системи. При маніпуляції векторними об'єктами відбувається автоматичне (в залежності від поточної позиції курсора) перемикання між різними режимами роботи векторного редактора. Це дозволяє користувачеві зменшити потенційні втрати і в цілому підвищує зручність роботи з системою.

Розроблена система «Контур» є пакетом програм для морфометрії біологічних об'єктів. Система має єдину графічну оболонку, яка дозволяє одночасно проводити обробку зображень, сегментацію і морфометричний аналіз (рис. 1.5, а).



а) б)

Рисунок 1.5 – Знімки екранів роботи програми «Контур»: а) загальний вид системи; б) процес калібрування

Звичайна послідовність операцій полягає в застосуванні автоматизованих алгоритмів морфометрії, і тільки в разі потреби здійснюється перехід в інші режими.

Основні функціональні можливості системи:

1) Вибір досліджуваних об'єктів – система надає можливість морфометричної оцінки наступних об'єктів: ядер тиреоцитів (поодинокі об'єкти, режим Single); клітин лімфоїдного ряду (пари об'єктів, режим Pair); агрегатів тиреоцитів (складові об'єкти, режим Multiple).

2) Завантаження растрових зображень – розроблене програмне забезпечення підтримує завантаження цитологічних растрових зображень основних растрових форматів. В якості таких форматів було обрано такі: BMP (бітові карти Windows), JPG (формат Joint Photograph Group), TIFF (формат корпорації Adobe).

3) Вибір об'єктива мікроскопа – для морфометричної оцінки біологічних об'єктів необхідно знати їх реальні фізичні розміри в мікрометрах. Дана інформація відсутня в растровому описі, де зберігаються тільки пікселі зображення, але для кожного такого знімка є інформація про об'єктив мікроскопа, за допомогою якого був здійснений даний знімок. Для правильного зіставлення фізичних розмірів зображення з растрової

інформацією користувач повинен вибрати об'єктив за допомогою керуючого елемента панелі інструментів «Change Objective». У випадаючому списку (рис. 1.3, а) можна відзначити необхідний об'єктив серед вже відкаліброваних. У разі якщо такого об'єктива в списку не виявилося, в програмі передбачена можливість калібрування довільного об'єктива.

4) Калібрування об'єктива – у системі реалізований режим калібрування об'єктива мікроскопа. Це дозволяє користувачеві легко налаштувати програму для роботи з різними мікроскопами і методиками постановки діагнозу. Режим калібрування об'єктива активізується щоразу, коли в сесії завантажено якесь растрове зображення.

Зазвичай калібрування нового об'єктива вимагає завантаження знімка, на якому присутній стандартний об'єкт із заздалегідь відомими в мікрометрах розмірами по горизонталі і вертикалі. Такий об'єкт називається калібрувальною міткою.

Для виділення біологічних об'єктів в системі виконуються операції сегментації і контурування. На виході цих операцій створюється векторний опис біологічних об'єктів. Підготовлені таким чином дані з одного зображення передаються для подальшої статистичної обробки. У поточну сесію завантажується нове зображення, до якого також застосовуються зазначені операції. Векторний опис може бути отриманий в одному з трьох наявних режимів [25].

Система «Контур» має в своєму розпорядженні повноцінний вбудований векторний редактор. Редактор забезпечує набір стандартних операцій над векторними об'єктами, таких як створення і видалення об'єкта, додавання і видалення точок для обраного контуру, поділ об'єкта на кілька складових частин. При роботі в ручному режимі користувач отримує можливість одночасно проводити операції над растровими і векторними вмістами документа, що є безперечною перевагою розробленої системи.

Перехід в напівавтоматичний режим відбувається при натисканні правої кнопки миші на область зображення, на якій візуально спостережувані

об'єкти не виділені при ручному або автоматизованому режимі. При цьому з'являється спливаюче меню, яке пропонує пошук необхідних об'єктів поблизу точки. Якщо об'єкти є парою ядро і клітина, то буде виділена відповідна пара контурів.

Автоматизований режим виділення об'єктів цитологічних зображень ґрунтується на алгоритмах аналізу та сегментації кольорових зображень. Його метою є зведення до мінімуму частки ручної праці при проведенні діагностики.

Результатом роботи програми є векторне подання контурів об'єктів. Воно в точності відповідає контурному представленні, що отримується при ручному введенні. Після запуску автоматичного виділення об'єктів дослідник візуально оцінює якість виділення контурів і при необхідності може скорегувати форму або видалити помилково виділений об'єкт.

Програмою передбачені наступні можливості роботи з експертною системою для постановки діагнозу:

1. Вибір нового методу постановки діагнозу – активується за допомогою основного меню програми «Morph | New »або кнопкою« New Data »на панелі інструментів.

2. Перегляд поточних даних методу – активується за допомогою основного меню програми «Morph | Show »або кнопкою« Show Data »на панелі інструментів.

3. Додавання даних з поточного знімка до аналогічних даних попередніх знімків. Перегляд поточних даних методу – активується за допомогою основного меню програми «Morph | Add »або кнопкою« Add Data »на панелі інструментів.

У системі є можливість використовувати три методи постановки діагнозу відповідно до трьох типів об'єктів: по ядрах тиреоцитів (поодинокі об'єкти, режим Single); по клітинах лімфоїдного ряду (пари об'єктів, режим Pair); по агрегатам тиреоцитів (складові об'єкти, режим Multiple).

Експертній системі для постановки діагнозу передається векторне опис об'єктів (ядер, клітин, агрегатів), заданий набором точок {x, y} і представляє собою замкнутий багатокутник.

Автоматизована система «Біоскан» розроблена для обробки і аналізу зображень та орієнтована на застосування в гістології та цитології. Вона є універсальним комплексом для наукових досліджень, в основі яких лежить вимір об'єктів на зображенні [26]. Система може вирішувати більшість завдань аналізу зображень, але використовується найчастіше в оптичній мікроскопії:

- для дослідження судин, клітин і клітинних популяцій;
- дослідження впливу протезів на тканину;
- вимірювання характеристик волокон;
- вивчення деформації протезів;
- клітинного підрахунку;
- визначення характеристик органів (переважно у ембріонів);
- визначення характеристик різних областей тканини, наприклад області запалення.

Спеціалізація системи виражена у великому наборі функцій і понять медичної морфометрії і можливості їх комбінації. При вирішенні завдань оптичної мікроскопії дуже зручним є використання мультифазних зображень, що дозволяють проводити класифікацію та виділяти гістологічну структуру клітин і тканини безпосередньо під час процесу аналізу, при виконанні індивідуальних вимірювань і визначенні характеристики поля.

– Надалі система «Біоскан» знайшла застосування в наступних областях медицини:

- діагностика в клінічній онкології;
- експертиза в патологічної анатомії;
- експериментальна онкологія;
- аналіз клітинного росту;

– контроль за впливом лікарських препаратів в фармакології;

– вимірювання геометричних і оптичних характеристик гістологічних об'єктів в специфічних наукових дослідженнях.

Повний набір класичних функцій аналізу і обробки зображень робить можливим застосування цієї системи і в багатьох інших областях, пов'язаних з обробкою зображень.

«Біоскан» являє собою високо інтегровану гнучку аналітичну систему, основний принцип побудови якої – отримання гранично широких можливостей для обробки та аналізу зображень оптичної мікроскопії за рахунок розвинутого оптимізованого програмного забезпечення при мінімальній складності апаратної частини.

Особливістю системи є одномоніторне виконання. За допомогою установки в комп'ютері спеціального пристрою введення зображення на програмному рівні вирішена задача отримання зображення безпосередньо з телекамери. Цей пристрій має прямий доступ до пам'яті і дозволяє передавати зображення різного програмно-регульованого формату, що необхідно для пошуку об'єкта дослідження в реальному масштабі часу, регулювання освітленості, наведення на різкість. Вбудований інтерпретатор мови високого рівня дозволяє виконувати програму аналізу зображень в автоматичному режимі, складати користувачем додаткові алгоритми виділення і аналізу об'єктів для спеціалізованих завдань. На програмному рівні вирішена задача інтерактивних (напівавтоматичних) вимірювань і редагування зображень за допомогою маніпулятора «миша». Передбачена можливість підключення пристроїв введення графічної інформації для більш точної роботи з фотографіями, рентгенограмами і малюнками. Вбудовані в систему функції обробки і аналізу зображень реалізують більшість відомих в даний час підходів до отримання, корекції, перетворенню, вимірюванню, реконструкції та зберігання зображень, в тому числі методи математичної морфології для бінарних напівтонових і кольорових зображень [27].

Система «Біоскан» має високорозвинений екранний інтерфейс, що функціонує в середовищі Microsoft Windows. Управління програмним пакетом може реалізовуватися через пункти меню і через команди, описані в невеликих підпрограма вбудованої мови інтерпретатора (Рисунок 1.6). Всі функції доступні і з підпрограм інтерпретатора, і з меню. Окремі функції можуть викликатися через панель інструментів, яка також реалізована в цій системі. При роботі з меню і панеллю інструментів використовуються динамічні вікна, в яких описуються і задаються основні параметри і характеристики викликаних функцій. Підпрограми інтерпретатора редагуються і запускаються на виконання.

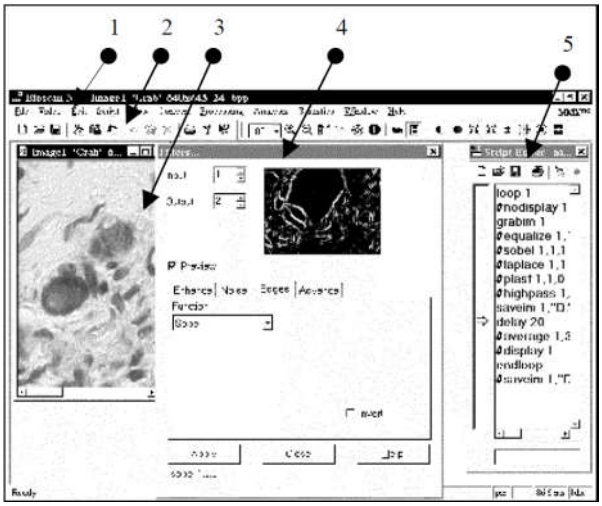


Рисунок 1. 6 – Робоча панель програми «Біоскан»: 1) меню, 2) панель інструментів, 3) вікно із зображенням, 4) вікно виклику функцій, 5) редактор підпрограм

Програмне забезпечення системи «Біоскан» має повний набір функцій по обробці зображень, вимірам об'єктів на них, а також з елементарної обробці об'єктів, що дозволяє гнучко вирішувати більшість завдань оптичної мікроскопії і задовольняє вимогам, що пред'являються до сучасних систем.

1.4 Аналіз завдання та постановка задач дипломної роботи

Дерево рішень – це один з методів автоматичного аналізу великих масивів даних. Ідея, яка лежить в основі методу дерева рішень, полягає у розбитті множин можливих значень вектора ознак (незалежних змінних) на множини, які не перетинаються і підгонці простої моделі для кожної такої множини. Отже, мета цього методу полягає в тому, щоб створити модель, яка прогнозує значення цільової змінної на основі декількох змінних на вході [19].

Області використання метода дерева рішень наступні:

– опис даних: дерева рішень дозволяють зберігати інформацію про вибірку даних у компактній і зручній для обробки формі, що містить у собі точні описи об’єктів;

– класифікація: дерева рішень ефективно вирішують задачі класифікації, тобто співвіднесення об’єктів до одного з раніше описаних класів;

– регресія: якщо змінна має непевні значення, то дерева рішень дозволяють визначити залежність цієї цільової змінної від незалежних (вхідних) змінних.

– Застосування класифікації гістологічних та цитологічних зображень на основі методу Random Forest дозволить забезпечити точність встановлення діагнозів.

Для вирішення цього завдання необхідно:

– дослідити предметну область;

– ознайомитись із сучасними алгоритмами розпізнавання;

– проаналізувати програмно-апаратний комплекс;

– розробити алгоритм на основі методу Random forest;

– спроектувати працездатний класифікатор;

– підготувати вхідні дані;

– виконати тестування розробленого модуля.

2 АЛГОРИТМИ КЛАСИФІКАЦІЇ ЗОБРАЖЕНЬ НА ОСНОВІ МЕТОДУ RANDOM FOREST

2.1 Основні ідеї методу Random forest

Випадковий ліс (random forest) – алгоритм машинного навчання, запропонований Лео Брейманом і Адель Катлер, що полягає у використанні ансамблю дереврішень [18]. Алгоритм поєднує в собі дві основні ідеї: метод баггінга (bagging – від bootstrap aggregation, тобто дерева навчаються по випадковим підвибіркам) і метод випадкових підпросторів. Алгоритм застосовується для задач класифікації, регресії і кластеризації.

У задачах регресії їх відповіді усереднюються, а в задачах класифікації приймається рішення голосуванням за більшістю [19]. Нехай дана навчальна вибірка D розміру n'. Генерується m нових вибірок D_i розміру n', вибором з D випадково з поверненням. Деякі спостереження можуть потрапити до вибірки кілька разів, деякі можуть не потрапити взагалі. Якщо $n' = n$ і n велике, то частка різних спостережень в D_i буде $(1 - 1 / e) \approx 63{,}2\%$. Далі вивчається m класифікаторів на кожній вибірці D_i. При класифікації нової точки, ці класифікатори голосують і відносять точку до класу, за який проголосувала більшість. У методі випадкових підпросторів (random subspace method, RSM) класифікатори навчаються на різних підмножинах признакового опису, які також виділяються випадковим чином.

Розглянемо алгоритм побудови випадкового лісу. Нехай навчальна вибірка складається з N прикладів, розмірність простору ознак дорівнює M, і заданий параметр m (в задачах класифікації зазвичай $m \approx \sqrt{M}$).

Всі дерева комітету будуються незалежно один від одного за такою процедурою:

1. Згенеруємо випадкову підвибірку з повторенням розміром n з навчальної вибірки. (Таким чином, деякі приклади потраплять в неї кілька разів, а приблизно $N/3$ прикладів не ввійдуть в неї взагалі);

2. Побудуємо вирішальне дерево, що класифікує приклади такої підвибірки, причому в ході створення чергового вузла дерева будемо вибирати ознаку, на основі якої проводиться розбиття, не з усіх M ознак, а лише з m випадково обраних. Вибір найкращого з цих m ознак може здійснюватися різними способами. В оригінальному коді Бреймана використовується критерій Джині. У деяких реалізаціях алгоритму замість нього використовується критерій приросту інформації;

3. Дерево будується до повного вичерпання підвибірки.

Одна з переваг випадкових лісів полягає в тому, що для оцінки ймовірності помилкової класифікації немає необхідності використовувати крос-перевірку (cross-validation) або тестову вибірку. Оцінка ймовірності помилкової класифікації випадкового лісу здійснюється методом "Out-Of-Bag" (OOB), який полягає у наступному. Так як, кожна бутстреп вибірка не містить приблизно 37% спостережень вихідної навчальної вибірки (оскільки вибірка з поверненням, то деякі спостереження в неї не потрапляють, а деякі – декілька раз). Класифікуємо деякий вектор $x \in D$. Для класифікації використовуються тільки ті дерева випадкового лісу, які будувалися по бутстреп вибіркам, що не містять x, і як звичайно використовується метод голосування. Частота помилково класифікованих векторів навчальної вибірки при такому способі класифікації і представляє собою оцінку ймовірності помилкової класифікації випадкового лісу методом OOB. Практика застосування оцінки OOB показала, що у випадку, якщо кількість дерев велика, то ця оцінка має високу точність.

На сьогоднішній день існує дуже багато алгоритмів, що реалізують дерева рішень: CART, C4.5, NewId, ITrule, CHAID, CN2 і т.д [28]. Але найбільшого поширення і популярність отримали наступні два:

– CART (Classification and Regression Tree) - це алгоритм побудови бінарного дерева рішень - дихотомічної класифікаційної моделі. Кожен вузол дерева при розбитті має тільки двох нащадків. Як видно з назви алгоритму, вирішує завдання класифікації і регресії;

– C4.5 - алгоритм побудови дерева рішень, кількість нащадків у вузла не обмежена. Не вміє працювати з безперервним цільовим полем, тому вирішує тільки завдання класифікації.

Більшість з відомих алгоритмів є "жадібними алгоритмами". Якщо один раз був обраний атрибут, і по ньому було вироблено розбиття на підмножини, то алгоритм не може повернутися назад і вибрати інший атрибут, який дав би краще розбиття. І тому на етапі побудови можна сказати дасть обраний атрибут, в кінцевому підсумку, оптимальне розбиття.

При побудові дерев рішень особлива увага приділяється наступним питанням: вибору критерію атрибута, за яким піде розбиття, зупинки навчання і відсікання гілок. Розглянемо всі ці питання по порядку.

Для побудови дерева на кожному внутрішньому вузлі необхідно знайти таку умову (перевірку), яка б розбивала множину, асоційовану з цим вузлом на підмножини. В якості такої перевірки повинен бути обраний один з атрибутів. Загальне правило для вибору атрибута можна сформулювати наступним чином: обраний атрибут повинен розбити множину так, щоб одержувані в результаті підмножини складалися з об'єктів, що належать до одного класу, або були максимально наближені до цього, тобто кількість об'єктів з інших класів в кожному з цих множин було якомога менше [29].

Були розроблені різні критерії, але ми розглянемо тільки два з них:

Алгоритм C4.5, вдосконалена версія алгоритму ID3 (Iterative Dichotomizer), використовує теоретико-інформаційний підхід. Для вибору найбільш підходящого атрибута, пропонується наступний критерій:

$$Gain\ (X) = Info\ (T) - Info\ x(T) Gain\ (X) = Info\ (T) - Info\ x(T), \quad (2.1)$$

де $Info\ (T)\ Info\ (T)$ - ентропія безлічі TT, а

$$Info\ x(T) = \Sigma i\ =\ 1n|Ti||T| * Info\ (Ti) Info\ x(T) = \Sigma i\ = 1n|Ti||T|\ *\ Info\ (Ti), \quad (2.2)$$

Множини $T1, T2, \ldots Tn$ отримані при розбитті вихідної множини T з перевірки X. Вибирається атрибут, що дає максимальне значення за критерієм (2.1).

Вперше це було запропоновано Р. Куінленом в розробленому ним алгоритмі ID3. Крім вищезгаданого алгоритму C4.5, є ще цілий клас алгоритмів, які використовують цей критерій вибору атрибута.

Алгоритм CART використовує так званий індекс Gini (в честь італійського економіста Corrado Gini), який оцінює "відстань" між розподілами класів (статистичний критерій).

$$Gini\ (T)\ =\ 1 - \sum_{j=i}^{n} p_j^2, \quad (2.3)$$

де T - поточний вузол, а p_j- ймовірність класу j в вузлі T.

CART був запропонований Л.Брейманом (L.Breiman) та ін.

На додаток до основного методу побудови дерев рішень були запропоновані наступні правила:

1) Використання статистичних методів для оцінки доцільності подальшого розбиття, так звана "рання зупинка" (prepruning). В кінцевому результаті "рання зупинка" процесу побудови доцільна в плані економії часу навчання, але тут доречно зробити одне важливе застереження: цей підхід будує менш точні класифікаційні моделі і тому рання зупинка вкрай небажана. Визнані авторитети в цій галузі Л.Брейман і Р. Куінлен радять буквально наступне: "Замість зупинки використовуйте відсікання".

2) Обмежити глибину дерева. Зупинити подальшу побудову, якщо розбиття веде до дерева з глибиною, яка перевищує задане значення.

3) Розбиття має бути нетривіальним, тобто отримані в результаті вузли повинні містити не менше заданої кількості прикладів.

Дуже часто алгоритми побудови дерев рішень дають складні дерева, які "переповнені даними", мають багато вузлів і гілок. Такі "гіллясті" дерева дуже важко зрозуміти. До того ж гіллясте дерево, що має багато вузлів, розбиває навчальну множину на все більшу кількість підмножин, що складаються з дедалі меншої кількості об'єктів.

Цінність правила, справедливого скажімо для 2-3 об'єктів, є вкрай низькою, і в цілях аналізу даних таке правило практично непридатне. Набагато краще мати дерево, що складається з малої кількості вузлів, яким би відповідало велика кількість об'єктів з навчальної вибірки. І тут виникає питання: а чи не побудувати всі можливі варіанти дерев, відповідні обучающему безлічі, і з них вибрати дерево з найменшою глибиною? На жаль, це завдання є NP-повною, це було показано Л. Хайфілем (L. Hyafill) і Р. Ривестом (R. Rivest), і, як відомо, цей клас задач не має ефективних методів вирішення.

Для вирішення вищеописаної проблеми часто застосовується так зване відсікання гілок (pruning). Нехай під точністю (розпізнавання) дерева рішень розуміється відношення правильно класифікованих об'єктів при навчанні до загальної кількості об'єктів з навчальної множини, а під помилкою - кількість неправильно класифікованих. Припустимо, що нам відомий спосіб оцінки помилки дерева, гілок і листя. Тоді, можливо використовувати наступне просте правило:

– побудувати дерево;

– відсікти або замінити піддерево, ті гілки, які не призведуть до зростання помилки.

На відміну від процесу побудови, відсікання гілок відбувається знизу вгору, рухаючись з листя дерева, відзначаючи вузли як листя, або замінюючи їх піддерево. Хоча відсікання не є панацеєю, але в більшості практичних

завдань дає хороші результати, що дозволяє говорити про правомірність використання подібної методики.

Дерева рішень є прекрасним інструментом в системах підтримки прийняття рішень, інтелектуального аналізу даних (data mining) [30].

До складу багатьох пакетів, призначених для інтелектуального аналізу даних, вже включені методи побудови дерев рішень. В областях, де висока ціна помилки, вони послужать відмінною підмогою аналітика або керівника.

Дерева рішень успішно застосовуються для вирішення практичних завдань в наступних областях:

– Банківська справа. Оцінка кредитоспроможності клієнтів банку при видачі кредитів;

– Промисловість. Контроль за якістю продукції (виявлення дефектів), випробування без руйнувань (наприклад перевірка якості зварювання) і т.д.;

– Медицина. Діагностика різних захворювань;

– Молекулярна біологія. Аналіз будови амінокислот.

Це далеко не повний список областей де можна використовувати дерева рішень. Ще не досліджені багато потенційних областей застосування.

Випадкові ліси мають цілу низку переваг, в порівнянні з деревами рішень, що і зумовило їх широке застосування, а саме:

1. Випадкові ліса забезпечують істотне підвищення точності, так як дерева в ансамблі слабокорельовані внаслідок подвійної ін'єкції випадковості в індуктивний алгоритм – за допомогою баггінга і використання методу випадкових підпросторів при розщепленні кожної вершини;

2. Методично і алгоритмічно складне завдання обрізання повного дерева рішень знімається , оскільки дерева у випадковому лісі не обрізаються (це також призводить до високої обчислювальної ефективності);

3. Відсутня проблема перенавчання (overfitting) (навіть при кількості ознак, що перевищує кількість спостережень навчальної вибірки і великій

кількості дерев). Тим самим знімається складна проблема відбору ознак, необхідна для інших ансамблевих класифікаторів;

4. Простота застосування: єдиними параметрами алгоритму є кількість дерев в ансамблі і кількість ознак, випадково відібраних для розщеплення в кожній вершині дерева;

5. Легкість організації паралельних обчислень.

Недоліками випадкових лісів у порівнянні з деревами рішень є відсутність візуального представлення процесу прийняття рішень і складність інтерпретації їх рішень.

Метод випадкових лісів швидко отримав визнання як в статистичному співтоваристві, так і в середовищі дослідників, і в даний час є одним з найбільш відомих методів класифікації і непараметричної регресії. Також він має вбудовану оцінку якостей та показує високу ефективність при роботі з великою кількістю ознак (наприклад, в хімії, біології і т.д.). Цей метод, став основою для розробки різних його модифікацій, які є важливими для додатків. Крім того, він дав імпульс для розробки методів, в основі яких лежить близька схема, яка не використовує дерева рішень в якості базового класифікатора.

Розглянувши сутність та переваги методу випадкових лісів, можна зробити висновок про перспективність його застосування в задачах класифікації даних, зокрема зображень [31].

З моменту винаходу оригінального Random Forest було розроблено декілька його модифікацій. Метод Online Random Forest опирається на ідеї онлайн бегінга і повністю рандомізованих дерев. Таким чином, виходить отримати достатньо швидкий онлайн алгоритм, який дозволяє використовувати його для задачі відстежування об’єктів. Метод Streaming Random Forest користується оцінкою Хефдинга при побудові дерев рішень та обтинанні гілок як стратегії забування. Сучасний метод Mondrian Forests вводить так звані Мондріановські дерева, які включають залежність від часу в кожному розбитті.

2.2 Алгоритм On-line Random Forests

Випадкові ліси (RFs) часто використовуються в багатьох додатках комп'ютерного зору та машинного навчання. Їх популярність полягає у високій обчислювальній ефективності під час підготовки та оцінки при досягненні впроваджених результатів. Проте, в більшості випадків RF використовується в автономному режимі. Це обмежує їх придатність для багатьох практичних проблем, наприклад, при навчанні дані надходять послідовно або вихідний розподіл безперервно змінюється.

У цьому пункті ми розглянемо різновид алгоритму випадкового лісу. Ми об'єднуємо ідеї онлайн-бегінгу, надзвичайних випадкових лісів і пропонуємо зростаючі on-line дерева рішень. Крім того, ми додамо тимчасову схему зважування для адаптивного відкидання деяких дерев на основі out-of-bag-error в задані інтервали часу і, отже, створення нових дерев. Експерименти по набору даних загального машинного навчання показують, що наш алгоритм сходиться до виконання в автономному режимі RF. Ми демонструємо зручність on-line RF на завданні інтерактивної сегментації в режимі реального часу.

Кожне дерево в лісі, побудоване і випробуване незалежно від інших дерев, отже, загальні процедури підготовки та тестування можуть бути виконані паралельно. Під час навчання, кожне дерево отримує новий набір бутстраповського навчання та породжені підвибірки із заміною оригінального навчального набору. Ми маємо на увазі ті зразки, які не включені в хід підготовки дерева, як і Out-Of-Bag (OOB) зразки цього дерева. Ці зразки можуть бути використані для обчислення Out-Of-Bag-Error (OOBE) дерева, а також ансамбль, який має низьку оцінку похибки [32].

Випробування на кожному вузлі дерев рішень вибираються спочатку для створення набору випадкових тестів, а потім вибирається кращий серед

них відповідно до деякого вимірювання якості (наприклад, посилення інформації або індекс Джині). Дерева, як правило, будують до їх повного розміру без відтинання.

Позначимо t дерево ансамблю як $F\ (X, \theta t)\colon X\ \rightarrow\ Y$, де θt випадковий вектор захоплюючи різні стохастичні елементи дерева. Для позначення стислості ми, як правило, представляємо дерева, як ft (x) $=$ P (x, θt). Ми також позначаємо весь ліс як $F\ =\ \{f1,\cdots,f\mathrm{T}\}$, де T являє собою кількість дерев у лісі. Ми можемо описати розрахункову ймовірність для передбачення класу k для розщеплення:

$$p(k|x) = \frac{1}{T}\sum_{t=1}^{T} p_t(k|x), \tag{2.4}$$

де $p_t\ (k \mid x)$ розрахункова щільність класу міток листя дерева t, де x спадає.

Остаточна мультикласова функція рішення лісу визначається [33]:

$$C(x) = \arg\max_{k\in y} p(k|x). \tag{2.5}$$

Брейман визначив межу класифікації позначеної проміжком (x, y), як

$$m_l(x, y) = p(y|x) - \max_{\substack{k\in y \\ k\neq y}} p(k|x). \tag{2.6}$$

Очевидно, що для правильної класифікації повинно виконуватись нерівність ml (x, y)> 0. Тому, помилка узагальнення подається формулою:

$$GE = E_{(X,Y)}(m_l(x, y) < 0), \tag{2.7}$$

де очікування вимірюється на весь розподіл (x, y). Брейман також показав, що ця помилка має верхню межу у формі [33]:

$$GE \leq \bar{p}\frac{1-s^2}{s^2}, \quad (2.8)$$

де $\bar{p}$ є середньою кореляцією між парами дерев у лісі, а s - сила ансамблю (тобто очікувана величина межі над усім розподілом). Іншими словами, для низької похибки узагальнення окремі дерева повинні бути високо незалежними і мати високу точність.

Ми поєднали он-лайн беггінг, випадковий вибір об'єктів та новий метод побудови дерев, що дозволяє створювати дерева рішень в режимі он-лайн. Експерименти з даними машинного навчання показують, що метод наближається до невідповідного аналога. Крім того, ми продемонстрували зручність використання нашого методу для завдання візуального відстеження та інтерактивної сегментації зображення. Цей метод досить стабільний, працює в режимі реального часу і його легко реалізувати.

2.3 Алгоритм Streaming Random Forests

Багато сучасних додатків мають справу з потоками даних, концептуально нескінченними послідовностями записів даних, часто перебуваючи при високих швидкостях потоку. Стандартні методи отримання інформації, як правило, припускають, що записи можуть бути доступні кілька разів, і тому, природно, не поширюються на потокові дані. Алгоритми для аналізу потоків повинні отримати всю необхідну інформацію із записів тільки одного, або, можливо, кількох, шляхів над даними. Алгоритм Streaming RF будує кілька дерев прийняття рішень, а також класифікує немарковані записи на основі безлічі голосів дерев. Ми оцінюємо точність класифікації алгоритму Streaming RF на декількох наборах даних, а також показуємо, що точність прирівнюється зі стандартним алгоритмом Random Forest.

Багато додатків, таких як моніторинг інтернет-трафіку, телекомунікаційний білінг, спостереження за астероїдами поблизу землі, а також замкнутого телебачення та відстеження продажів створюють величезну кількість даних для аналізу. Звичайно, ці дані не завжди зберігаються. Такі дані є концептуально нескінченною, в режимі реального часу та упорядкованою послідовністю записів, і тому вони моделюються як потік. Для того, щоб витягувати дані з потоку, існуючі машинні алгоритми пошуку даних необхідно адаптувати, щоб відобразити властивості потоків [34-36].

Видобуток потокових даних привернув велику увагу, з дослідженням, наприклад, виявлення змін потоків даних, ведення статистики потоків даних, класифікації потоків даних та потоку даних кластеризації. У пункті ми розглядаємо проблему класифікації даних потоку. Класифікація, як правило, вважається необхідною для виконання трьох етапів, кожен з яких пов'язаний з даними. На першому етапі модель побудована з позначеними навчальними даними; на другому етапі модель перевіряється за допомогою раніше небачених мічених даних (дані тесту); і на третьому етапі, моделювання розгортаються немарковані дані. У класифікації потоку є лише один потік даних, тому проблема повинна бути сформульована іншим способом. Ми припускаємо налаштування, в якому містяться деякі записи у потоці, які використовуються для побудови або тестування моделі, а метою є прогнозування класу незамічених записів. У цьому налаштуванні існує кілька різних сценаріїв, залежно від того, як приклади тренувань (помічені) наводяться через потік. Класифікатор може бути побудований стандартним (автономним) способом, але повинен бути побудований досить швидко, щоб не відставати від швидкості прибуття помічених записів. Подібним чином, класифікація повинна бути достатньо швидкою, щоб йти в ногу із швидкістю прильоту немаркованих записів. Деякі підмножини або суфікс позначених записів можуть бути використані як дані тесту [37-39].

Стандартний алгоритм Random Forest може бути застосований для потокової передачі даних. Труднощі щоб зробити це полягають в тому, що алгоритм Random Forest робить кілька проходів над навчальними даними, двома різними способами. По-перше, шлях до даних відкритий щоб створити навчальні дані для кожного дерева, яке буде побудоване. По-друге, для кожного дерева, шлях проводиться наскрізь (деяких стовпців даних) для кожного внутрішнього вузла дерева, щоб генерувати відліки, які використовуються, щоб вирішити, який атрибут і точку поділу вибрати для нього.

Потоковий алгоритм не може дозволити собі кілька шляхів до даних. Для створення кожного дерева, використовується окремий пакет або блок помічених записів. В результаті алгоритм Streaming RF вимагає значно більше даних, ніж стандартний алгоритм для побудови набору дерев. Було б можливо використовувати окрему партію даних, щоб прийняти рішення по кожному внутрішньому вузлі, але для цього буде потрібно ще більшу кількість даних, а також це приведе до збільшення часу, перш ніж надійні класифікації зможуть бути зроблені для немаркованих записів. Замість цього, ми адаптуємо ідеї з потокових алгоритмів дерева рішень [40, 41].

Алгоритм Streaming Random Forest будує велику кількість дерев, так само, як стандартний алгоритм Random Forest. За цей час, алгоритм займає необхідну кількість дерев в якості параметра, але розширення, в яких число дерев походить від точності класифікації, або до набору дерев були додані нові члени, а старі видалені щоб реагувати на прості зміни даних.

При надходженні нових маркованих записів, вони прямують вниз поточного дерева, на основі його значень атрибутів і нерівностей внутрішніх вузлів, поки він не досягне граничного вузла.

У граничному вузлі, значення атрибута запису сприяє підрахункам класу, які використовуються для обчислення Gini індексів. Для того, щоб підтримувати інформацію про розподіл значень атрибутів і їх відношення до класу міток, атрибути дискретизовані в інтервалах фіксованої довжини. Межі

між цими інтервалами є можливими розщепленими точками для кожного атрибута. Процедура прийняття рішення, коли і як змінити граничний вузол у інший вид вузла трохи складніше.

Параметр, N_{min}, використовується, щоб вирішити, як часто, щоб перевірити, чи слід розглядати граничний вузол для трансформації. Всякий раз, коли вузол накопичує N_{min} мічених записів, індекс Gini і Хефдинг тести застосовуються. Якщо тест Хефдинга виконується, то граничний вузол записує досить записів, щоб визначити найкращий атрибут і розподіл значення. Тест індекса Gini потім використовується, щоб вибрати кращий атрибут і точку поділу, а межа вузла перетворюється у внутрішній вузол з нерівністю на основі цього атрибута і точці поділу [42]. Дві похідних цього вузла стають новими граничними вузлами.

Якщо число записів, які досягли граничного вузла перевищує поріг називається вікном вузла, а вузол ще не був розділений, алгоритм перевіряє, щоб побачити, якщо цей вузол повинен бути замість цього перетворюється в лист. Якщо вузол накопичив записи, які в основному з одного класу, то вона перетворюється в лист. В іншому випадку, вузол перетворюється на внутрішній вузол, використовуючи найкращий атрибут і розділяє точку далі.

2.4 Алгоритм Mondrian Forests

Ансамблі випадкових дерев рішень, як правило, називають випадковими лісами, які широко використовуються для задач класифікації та регресії машинного навчання та статистики. Найбільш популярні варіанти випадкового лісу (наприклад, випадковий ліс Бреймана і надзвичайно випадкові дерева) працюють на вибірці навчальних даних. Онлайн методи в даний час користуються великим попитом. Існуючі онлайн випадкові ліси вимагають більшої кількості навчальних даних, ніж їх аналоги для

досягнення порівняльної прогностичної продуктивності. Використовуючи процеси Мондріана (Roy і Teh, 2009) ми будуємо ансамблі випадкових дерев рішень, які називаються Mondrian Forests [43].

Використовуючи властивості процесів Мондріана, ми представляємо ефективний онлайн алгоритм, який узгоджується з його серійною копією на кожній ітерації. Не тільки онлайн Mondrian ліс швидший і точніший, ніж останні пропозиції по онлайн методу випадкових лісів, але вони майже збігаються з точність методів випадкового лісу навчених на тому ж наборі даних.

Для наших цілей, дерево прийняття рішень по R^D буде ієрархічним, двійковим розбиттям R^D і правило для прогнозування мітки контрольних точок даних підготовки даних. На рисунку 2.1(а) показана структура дерева та розділ дерева рішень, тоді як на рисунку 2.1(б) показано дерево Mondrian.

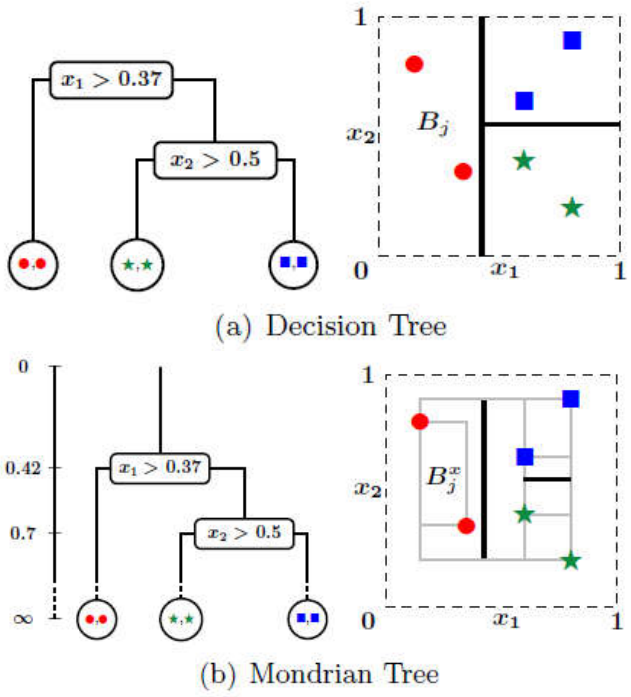


Рисунок 2.1 – Порівняння Decision Tree та Mondrian Tree

Зверніть увагу, що дерево Mondrian вбудоване в вертикальну вісь часу, причому кожен вузол пов'язаний з часом розщеплення, і розбиття

відбувається лише в межах даних тренувань у кожному блоці (позначається сірими прямокутниками).

Ми розглянули Mondrian Forest, новий клас випадкових лісів, який може бути вивчений поступово ефективним чином. MF перевершує існуючі онлайн випадкові ліси з точки зору часу навчання, а також кількість навчальних примірників, необхідних для досягнення певної точності тесту. Примітно, що MF досягає конкурентної точності випробувань до партії випадкових лісів навчених на одній і тій же фракції даних. MF не може обробляти безліч непотрібних функцій. Один із способів використання міток для керівництва шпагат здійснюється за допомогою недавно запропонованого послідовного алгоритму Монте-Карло для дерев рішень [44].

У реалістичній настройці потокових даних, де не можна зберігати приклади навчання для декількох шляхів, MF вимагатиме значно менше прикладів, ніж ORF-Saffari для досягнення такої ж точності.

Порівнюючи точність перевірки з часом навчання на usps, satimages та letter datasets, ми спостерігаємо, що MF, щонайменше на порядок швидше, ніж перепідготовлені пакетні версії та ORF-Saffari. Для ORF-Saffari, ми намічаємо точність перевірки в кінці кожного додаткового шляху; отже, він містить додаткові маркери в порівнянні з верхнім рядком, результати яких відбуваються після одного шляху. Повторна підготовка пакету RF з використанням 100 міні-партій несправедлива до MF; в налаштуваннях потокових даних, де модель оновлюється, коли з'являється новий екземпляр навчання, MF буде значно швидшим, ніж перепідготовлені версії пакетів. На рисунку 2.2 зображено порівняльні графіки різновидів RF.

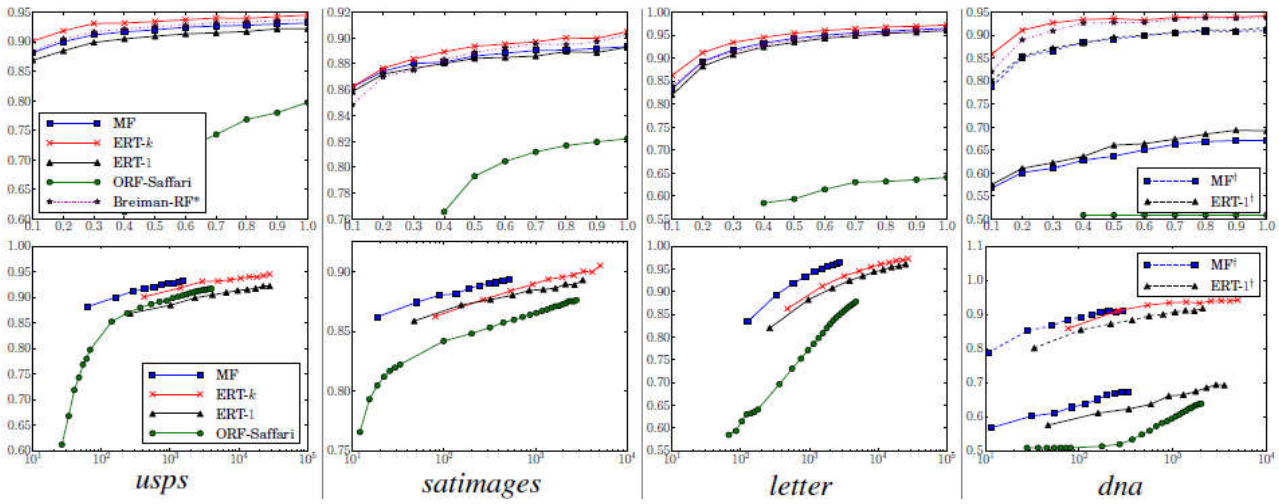


Рисунок 2.2 – Порівняння різних методів Random Forest

Результати на різних наборах даних: вісь y - точність перевірки в обох рядках. Вісь x - це частка даних тренувань для верхнього ряду та часу навчання (у секундах) для нижнього ряду. Ми використовували попередньо вибрану тестову та навчальну вибірку. Для набору даних usps: D = 256; K = 10; N_{train} = 7291; N_{test} = 2007; для satimages набору даних: D = 36; K = 6; N_{train} = 3104; N_{test} = 2000; набір даних для letter datasets: D = 16; K = 26; N_{train} = 15000; N_{test} = 5000; для даних dna: D = 180; K = 3; N_{train} = 1400; N_{test} = 1186.

3 ПРОГРАМНА РЕАЛІЗАЦІЯ ТА ДОСЛІДЖЕННЯ АЛГОРИТМУ РОЗПІЗНАВАННЯ ЗОБРАЖЕНЬ

3.1 Узагальнена структура програмного модуля

Розпізнавання образів (pattern recognition) — це розділ теорії штучного інтелекту (artificial intelligence), що вивчає методи класифікації об'єктів. За традицією об'єкт, що піддається класифікації, називається образом (pattern). Образом може бути цифрова фотографія (розпізнавання зображень), буква або цифра (розпізнавання символів), запис мови (розпізнавання мови) тощо.

В межах теорії штучного інтелекту розпізнавання образів включається в більш широку наукову дисципліну — теорію машинного навчання (machine learning), метою якої є розробка методів побудови алгоритмів, що здатні навчатися [45].

Існує два підходи до навчання: індуктивне і дедуктивне. Індуктивне навчання, або навчання за прецедентами, засноване на виявленні загальних властивостей об'єктів на підставі неповної інформації, отриманих емпіричним шляхом. Дедуктивне навчання передбачає формалізацію знань експертів у вигляді баз знань (експертних систем тощо). В нашому курсі нас буде цікавити лише індуктивне навчання, тому будемо вважати машинне навчання і навчання за прецедентами синонімами.

Слід зауважити, що, як кожна математична дисципліна, розпізнавання образів має власний математичний апарат, який включає математичну статистику, методи оптимізації, дискретну математику, алгебру і геометрію.

Розпізнавання образів має широке застосування і використовується при створенні усіх комп'ютерних систем, на які покладаються інтелектуальні функції, тобто функції, пов'язані із прийняттям рішень замість людини: медична діагностика, криміналістична експертиза, пошук інформації та інтелектуальний аналіз даних тощо [46].

Прецедент — це об’єкт, приналежність якого до заданого класу визначена заздалегідь. Прецедентом може бути, наприклад, набір ознак пацієнта із відомим діагнозом, з яким слід порівнювати набір ознак людини, діагноз якої ще невідомий.

Кожний образ являє собою набір чисел, що описують його властивості і називаються ознаками (feature). Упорядкований набір ознак об’єкта називається вектором ознак (feature vector). Вектор ознак — це точка в просторі ознак (feature space).

Класифікатор, або вирішальне правило (decision rule) — це функція, яка ставить у відповідність вектору ознак образу клас, до якого він належить.

Задачу розпізнавання образів можна розділити на ряд підзадач:

1. Генерування ознак (feature generation) — вимірювання або обчислення числових ознак, що характеризують об’єкт.

2. Вибір ознак (feature selection) — визначення найбільш інформативних ознак для класифікації (в цей набір можуть входити не лише первинні ознаки, але й функції від них).

3. Побудова класифікатора (classifier construction) — конструювання вирішального правила, на підставі якого здійснюється класифікація.

4. Оцінка якості класифікації (classifier estimation) — обчислення показників правильності класифікації (точність, чутливість, специфічність, помилки першого та другого роду).

Структурна схема – схема, яка визначає основні функціональні частини виробу, їх взаємозв'язки та призначення. Структурна схема призначена для відображення загальної структури пристрою, тобто його основних блоків, вузлів, частин та головних зв'язків між ними. Із структурної схеми повинно бути зрозуміло, навіщо потрібний даний пристрій і як він працює в основних режимах роботи, як взаємодіють його частини. Побудуємо загальну структурну схему для програмного модуля (Рисунок 3.1).

Рисунок 3.1 – Загальна структура програмного модуля

У цій схемі відображається послідовність дій для класифікації зображення. Щоб працювати з програмним модулем нам потрібні вхідні зображення, тобто в даному випадку цитологічні та гістологічні зображення. Для класифікації зображення потрібно пройти такі етапи:

а) попередня обробка;

б) фільтрація;

в) сегментація.

Попередня обробка зображення є важливою сферою цифрової обробки зображень, тому що величезна кількість різноманітних дослідженнь та промислових процессів виконується із застосуванням різного роду зображень. Під час формування зображень неминучим є його викривлення. Попередня обробка зображення проводиться із урахуванням априорної інформації про природу викривлень, та причину їх появлення. Найбільш розповсюдженним видом викривлення зображення є шум – це змінення яскравості або кольору деяких точок зображення під дією різних факторів.

Приклад попередньої обробки цитологічного зображення наведено на рисунку 3.2.

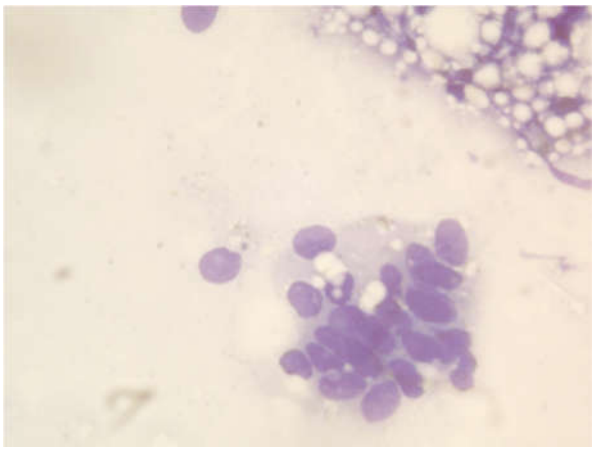


Рисунок 3.2 – Попередня обробка цитологічного зображення

Наступним етапом є фільтрація. При фільтрації яскравість (сигнал) кожної точки вихідного зображення, спотвореного завадою, замінюється деяким іншим значенням яскравості, яке в меншій мірі було спотворене завадою. Фільтрація зображень здійснюється в просторовій і частотній областях. При просторовій фільтрації зображень перетворення виконується безпосередньо над значеннями відліків зображення. Результатом фільтрації є оцінка корисного сигналу зображення. Наявність шумів на зображені 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.

З комп'ютерної точки зору, сегментація — це процес розділення цифрового зображення на декілька сегментів. Мета сегментації полягає у спрощенні і/або зміні представлення зображення для полегшення його аналізу. Сегментацію зображень зазвичай використовують для виділення об'єктів та меж (лінії, криві, і т. д.) на зображеннях. Точніше, сегментація зображень — це процес присвоєння таких міток кожному пікселю зображення, що пікселі з однаковими мітками мають спільні візуальні характеристики.

Після обробки зображень проводимо виділення кількісних ознак ядер клітин. Виділяємо такі ознаки, як площа, окружність, довжина головної осі та

діаметр Фарета. Отримавши кількісні ознаки класифікуємо зображення за допомогою класифікатора.

Для проектування програмного забезпечення використаємо інструмент Visual Paradigm for UML. Необхідність побудови діаграм полягає в тому, що це дає можливість детально зрозуміти саму програму, з чого вона складатиметься, що вона міститиме, які поля будуть заповнюватися, які класи будуть використовуватися і які зв'язки між класами будуть задіяні.

У середовищі VP існує 13 типів діаграм у межах пакету UML Diagrams, користуючись якими розробник має можливість створювати вичерпний опис розроблюваної ПС.

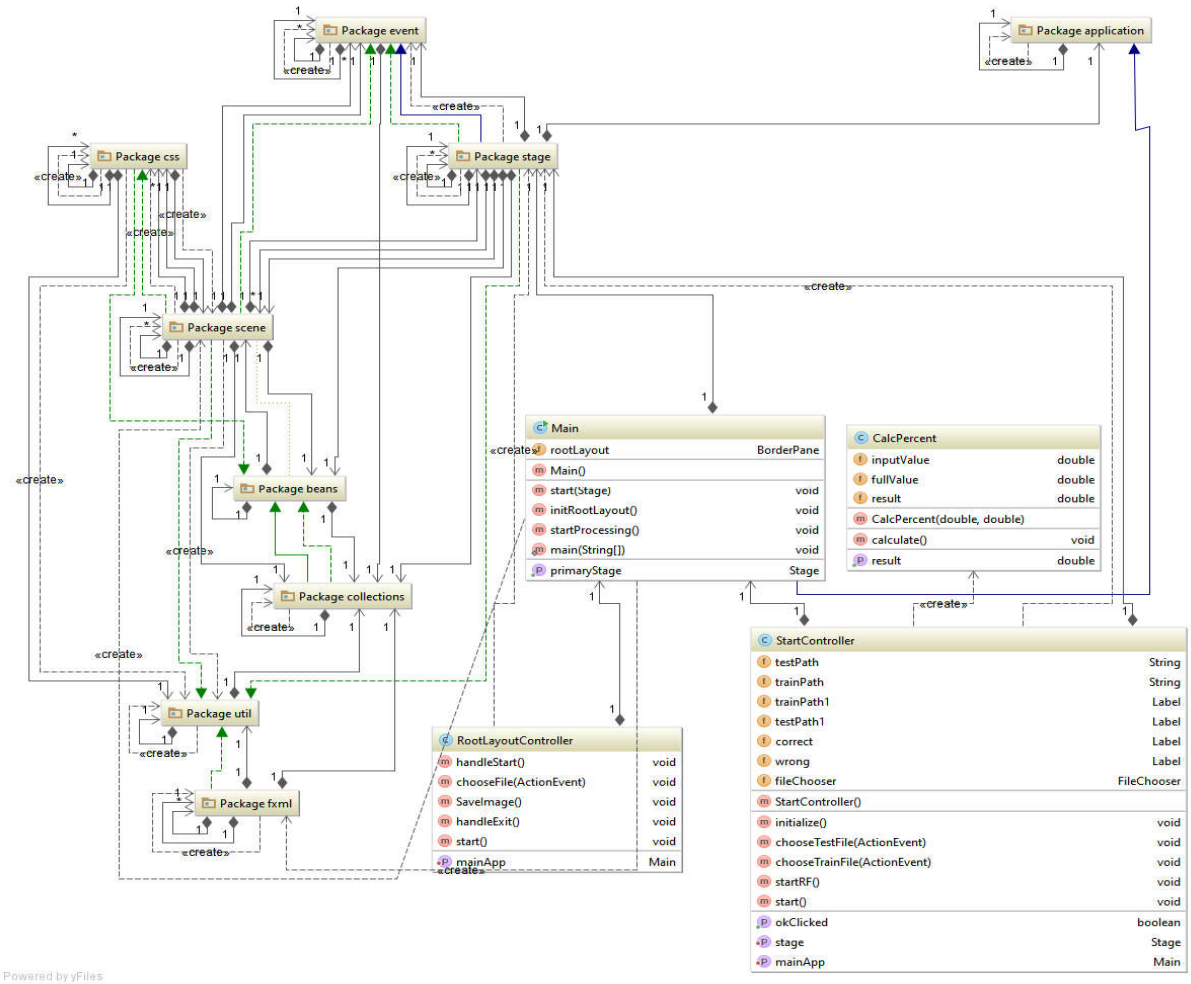
Діаграма класів — статичне представлення структури моделі. Відображає статичні (декларативні) елементи, такі як: класи, типи даних, їх зміст та відношення. Діаграма класів, також, може містити позначення для пакетів та може містити позначення для вкладених пакетів.

Клас (class) у мові UML використовується для визначення множини об'єктів, які мають однакову структуру, поведінку та відношення із об'єктами із інших класів. Графічно клас зображується у вигляді прямокутника, який додатково може бути розділений горизонтальними лініями на розділи або секції. В цих розділах вказуються ім'я класа, атрибути (змінні) та операції (методи).

Базовими типами відношень або зв'язками між класами в мові UML є: відношення залежності, відношення асоціації, відношення узагальнення, відношення агрегації та відношення композиції. Кожен з них має власне графічне позначення на діаграмі класів. Відношення асоціації є основним типом залежностей між акторами та прецедентами, що вказує на те, що деякий актор відіграє певну роль у взаємодії з системою. Відношення включення дає можливість відобразити те, що деякі функції одного прецеденту А обов'язково мають бути задіяні при використанні іншого прецеденту В. На діаграмі включення позначається пунктирною прямою лінією зі стрілкою на кінці, яка направлена від того прецеденту, що включає

в себе деякий інший і має назву «include». Відношення розширення дає можливість відобразити те, що при опрацюванні системою деякого прецеденту X є можливим (але не обов'язковим) використання функціональності іншого прецеденту Y. На діаграмі включення позначається пунктирною прямою лінією зі стрілкою на кінці, яка направлена від того прецеденту, що є розширенням та має назву «extend». Відношення узагальнення дозволяє показати певну ієрархію акторів та/або самих прецедентів, тобто позначити, який з них є супертипом, а який — підтипом певного класу об'єктів. На діаграмі це позначається суцільною прямою лінією з трикутною стрілкою на кінці, яка направлена до прецеденту, що є супертипом.

Побудована діаграма класів для розроблюваного програмного модуля представлена в додатку А.



Клас «Main» є головним класом та точкою входу в систему. Поля та методи класу «Main» наведено на рисунку 3.3.

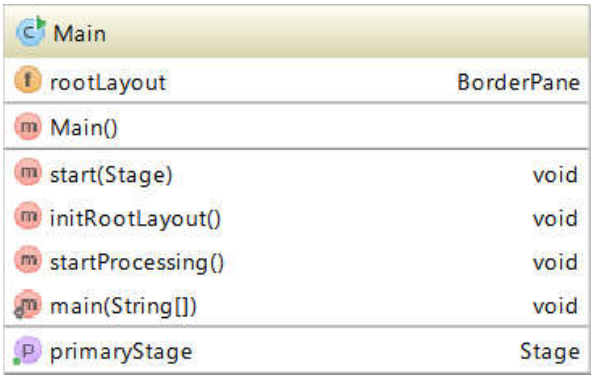
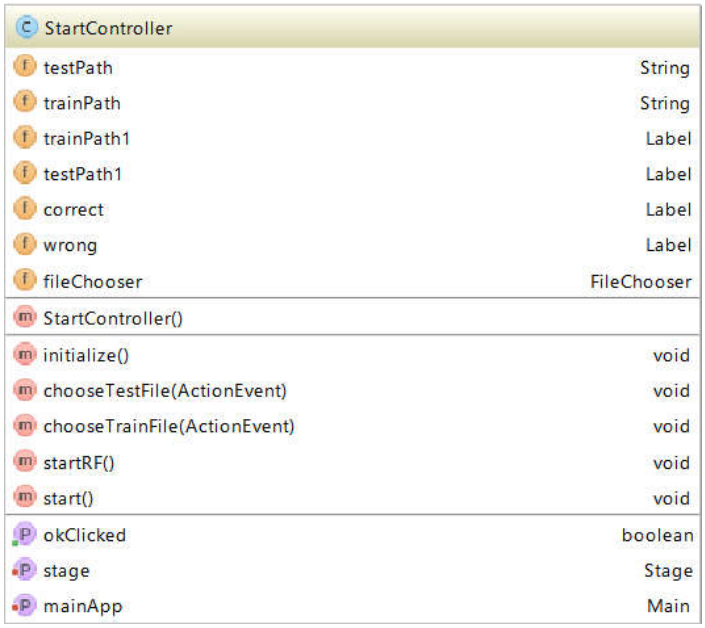


Рисунок 3.3 – Клас «Main»

Поле «rootLayout» з типом даних «BorderPane» відповідає за побудови форми головного вікна графічного інтерфейсу. «Main()» - конструктор класу, який викликається при створенні об’єкта головного класу .Метод «initRootLayout()» відповідає за завантаження складових елементів вікна графічного інтерфейсу. Метод «startProcessing()» відповідає за заповнення елементів форми відповідним контентом.

Стурктуру класу «StartController» наведено на рисунку 3.4.



Риснуок 3.4 – Стурктура класу «StartController»

Даний клас вміщує поля, методи та властивості для обробки подій, здійснених, користувачем та для виводу результатів проведення досліджень на екран.

Поле «testPath» та «trainPath» типу «String» зберігають шлях до текстових файлів для навчальної та тестової вибірки. Поля «testPath1» та «trainPath1» типу «Label» призначені для виводу візуальної інформації про шлях до файлу на головне вікно програми. Поле «correct» відповідає за вивід інформації про відсоток та кількість правильно класифікованих елементів з тестової вибірки. Поле «wrong» відповідає за вивід інформації про відсоток та кількість неправильно класифікованих елементів з тестової вибірки.

Метод «StartController()» є конструктором даного класу та викликається при створенні об’єкту класу «StartController» у класі «Main». Метод «initialize()» відповідає за ініціалізацію базових елементів класу, описаних вище.

У методів «chooseTestFile» та «chooseTrainFile» вхідним параметром є елемент «ActionEvent», що зберігає вибраний користувачем шлях до файлу відповідно тестової та навчальної вибірки. У методі «startRF()» відбувається створення об’єкта класу для здійснення класифікації вхідних даних методом «Random Forest». Вхідними параметрами є шлях до текстових файлів.

Структура класу «CalcPercent» наведено на рисунку 3.5.

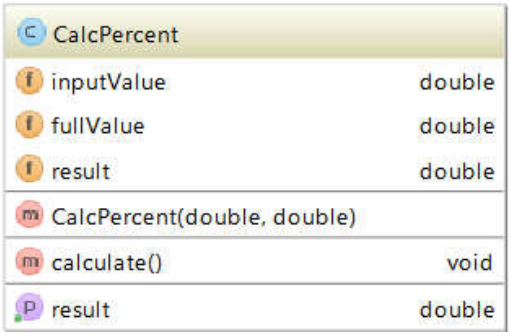


Рисунок 3.5 - Структура класу «CalcPercent»

Поля даного класу призначені для підрахунку відсотку правильно класифікованих даних. Типом даних полів «inputValue», «fullValue», «result» є «double», тип даних з плаваючою крапкою. Вхідними параметрами для конструктора «CalcPercent(double, double)» є два числа – кількість правильно класифікованих елементів та загальна кілкьість елемннтів тестової вибірки. Метод «calculate()» здійснює обчислення відсотку правильно класифікованих записів.

Пакет «util» вміщує допоміжні класи, потрібні для роботи методу «Radom Forest». Наприклад , класи, необхідні для вводу-виводу інформації, розбору текстового файлу та елементи тощо.

У пакеті «fxml» знаходяться файли розмітки fxml, що відповідають за вивід таких елементів форми , як кнопки, текстові поля та ін. Приклад fxml коду для виводу елементів форми наведено нижче:

```
<Label text="Навчальна вибірка" GridPane.columnIndex="0" GridPane.rowIndex="1" />
<Button mnemonicParsing="false" onAction="#chooseTrainFile" text="Вибрати файл" GridPane.columnIndex="1" GridPane.rowIndex="1" />
<Label text="Тестова вибірка" GridPane.columnIndex="0" GridPane.rowIndex="2" />
<Button mnemonicParsing="false" onAction="#chooseTestFile" text="Вибрати файл" GridPane.columnIndex="1" GridPane.rowIndex="2" />
```

Пакет «css» відповідає за зберігання каскадних таблиць стилів, що визначають розміщення , колір елементів головного вікна програми.

3.2 Інтерфейс програмного модуля

Для реалізації даного програмного модуля використано середовище програмування IntelliJ IDEA . Мовою програмування, яка буде застосована в цьому середовищі є Java.

IntelliJ IDEA — комерційне інтегроване середовище розробки для різних мов програмування (Java, Python, Scala, PHP та ін.) від компанії JetBrains [47]. Системні вимоги для роботи з цим програмним середовищем наведені у таблиці 3.1.

Таблиця 3.1 – Системні вимоги для програмного середовища

	Windows	OS X	Linux
Версія ОС	Microsoft Windows 10/8/7/Vista/2003/XP (включаючи 64-bit)	Mac OS X 10.5 або вище, аж до MacOS 10.12 (Sierra)	GNOME або KDE стільниця
Оперативна пам’ять	1 GB ОЗП мінімум, 2 GB ОЗП рекомендується		
Простір на диску	300 MB на твердому диску + принаймі 1 GB для кешування		
Версія JDK	JDK 1.8 починаючи з 2016.1		
Роздільна здатність	1024*768 мінімальна роздільна здатність		

Java — об'єктно-орієнтована мова програмування, випущена компанією SunMicrosystems у 1995 році як основний компонент платформи Java. Синтаксис мови багато в чому походить від C та C++. У офіційній реалізації, Java програми компілюються у байткод, який при виконанні інтерпретується віртуальною машиною для конкретної платформи. Всі дані і дії групуються в класи об'єктів. Виключенням з повної об'єктності є примітивні типи (int, float тощо). Через це, Java не вважається повністю об'єктно-орієнтовною мовою.

Програми на Java утворені з визначень класів та інтерфейсів. Класи

містять змінні та сталі, які утримують дані, методи, які виконують дії, та конструктори, які створюють екземпляри класів — об'єкти. Дані можуть мати простий тип (наприклад байт, ціле число, символ) або бути посиланням на об'єкт. Мова Java є статично типізованою [48].

В даній дипломній роботі розроблено програмний модуль для класифікації цитологічних та гістологічних зображень методом Random Forest. При роботі з програмою використовується простий і зручний інтерфейс.

Інтерфейс — сукупність засобів і методів взаємодії між елементами системи. Залежно від контексту, поняття застосовне як до окремого елемента (інтерфейс елемента), так і до зв'язувань елементів (інтерфейс спряження елементів).

Цей термін використається практично у всіх галузях науки й техніки. Його значення ставиться до будь-якого сполучення взаємодіючих сутностей. Під інтерфейсом розуміють не тільки роботу пристрою, але й правила (протокол) взаємодії цих пристроїв.

Інтерфейс користувача — сукупність засобів для обробки та відображення інформації, максимально пристосованих для зручності користувача; різновид інтерфейсів, в якому одна сторона представлена людиною (користувачем), інша — машиною/пристроєм. Представляє собою сукупністю засобів і методів, за допомогою яких користувач взаємодіє з різними, найчастіше складними, машинами, пристроями і апаратурою.

При розробці інтерфейсу необхідно враховувати наступні принципами:

1. Стандартизація. Рекомендується використовувати стандартні, перевірені багатьма програмістами і користувачами інтерфейсні рішення.

2. Зручність і простота роботи. Інтерфейс повинен бути інтуїтивно зрозумілим. Бажано, щоб усі дії легко запам’ятовувалися і не вимагали стомлюючих процедур.

3. Зовнішній дизайн. Важливо, щоб інтерфейс не стомлював зір. Він повинен бути розрахований на тривалу роботу користувача з додатком

протягом дня.

До сукупності засобів, за допомогою яких користувач взаємодіє з різними програмами і пристроями, відносяться:

1. Інтерфейс командного рядка: інструкції комп'ютеру даються шляхом введення з клавіатури текстових рядків (команд).

2. Графічний інтерфейс користувача: програмні функції представляються графічними елементами екрану.

3. Діалоговий інтерфейс: наприклад, пошук.

4. Природно-мовний інтерфейс: користувач «розмовляє» з програмою на рідній йому мові.

5. Тактильний інтерфейс: кермо, джойстик і так далі.

6. Нейрокомп’ютерний інтерфейс (англ. brain-computerinterface): відповідає за обмін між нейронами і електронним пристроєм за допомогою спеціальних імплантованих електродів

Для написання даної програми було вирішено обрати діалоговий інтерфейс, адже при використанні програми буде вестись діалог між користувачем та комп’ютером за допомогою графічних елементів: меню, вікон, тощо.

Для забезпечення діалогу між програмою та користувачем було вирішено використовувати діалогове вікно, розроблене за SDI (SingleDocumentInterface) типом інтерфейсу.

Діалогове вікно — це особливий тип вікна, яке задає запитання і дозволяє вибрати варіанти виконання дії, або ж інформує користувача. Діалогові вікна зазвичай відображаються тоді, коли програмі для подальшої роботи потрібна відповідь користувача.

Одновіконний (SDI) інтерфейс – це тип інтерфейсу, у якому надається можливість роботи тільки з одним документом в одному вікні. Під документом у цьому випадку потрібно розуміти форму, призначену для роботи з даними, а не з конкретним документом. Цей тип інтерфейсу підходить для програм, створених для роботи з документом одного типу з

невеликою кількістю полів.

Інтерфейс програмного модуля реалізовано на основі інтерфейсу типу SDI, програма складається з одного діалогового вікна в якому містяться всі необхідні елементи керування.

При запуску програми відкривається вікно, яке зображене на рисунку 3.6.

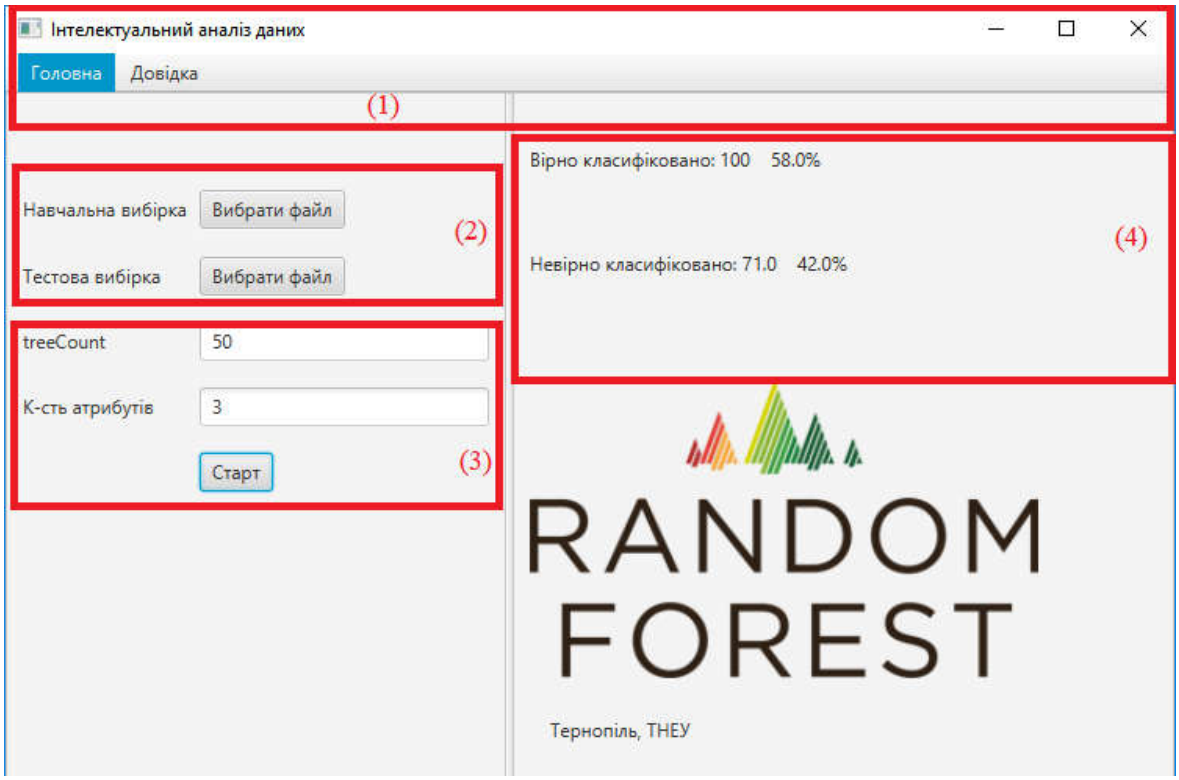


Рисунок 3.6 – Графічний інтерфейс головного вікна програми

Після завантаження програми буде відкрито вікно, що містить область меню і робочу область програми. У верхній частині вікна знаходиться головне меню (1), яке визначає всі можливі функції системи. Нижче, займаючи переважну за величиною частину екрану, знаходиться робоча область, де розташовуються такі блоки: блок вибору файлу навчальної та тестової вибірки (2), блок вибору параметрів класифікації(3) та блок виводу результатів (4).

Область меню складається з меню вікна, заголовка вікна, головного меню і кнопок управління вікном (рисунок 3.7).

Рисунок 3.7 – Меню програми

Меню «Головна» відкриває головне вікно програми, а меню «Довідка» містить інформацію про програму та інструкцію як нею користуватись.

У блоці вибору файлу навчальної та тестової вибірки потрібно вказати шлях до файлу, який містить кількісні дані гістологічних або цитологічних зображень. Після вибору цих файлів з даними, у блоці вибору параметрів вказуємо кількість дерев та атрибутів. Кількість вводиться з клавіатури вручну. У блоці виводу результатів відображається кількість прокласифікованих зображень та точність класифікації (рисунок 3.6 (4)).

3.3 Тестування програмного модуля

Тестування — один з розділів діагностики. Тестування застосовується в техніці, медицині, психіатрії, освіті для визначення придатності об'єкта тестування для виконання тих чи інших функцій. У завдання тестування не входить визначення причин невідповідності заданим вимогам (специфікаціям). Якість тестування і достовірність його результатів значною мірою залежить від методів тестування та складу тестів.

Процес тестування включає:

– подачу тестового набору;

– визначення реакції об'єкта тестування на тестовий набір;

– оцінку реакції і висновки.

Тестовий набір складається з окремих тестів і розробляється таким чином, щоб забезпечити повне або значне покриття множини ймовірних впливів на об'єкт тестування. Цим, також, визначається складність розробки як окремих тестів, так і тестових наборів.

У технічній діагностиці застосовуються формалізовані методи розробки мінімальних, необхідних і достатніх тестів перевірки працездатності (відповідності специфікаціям).

Для коректної роботи програми запускаємо IntelliJ IDEA з правами адміністратора. Під час компіляції програми не виявлено ніяких помилок, це означає, що програма працює вірно і можна її виконати. Результат виконання програми показаний на рисунку 3.6.

Для проведення класифікації за допомогою математичних методів необхідно мати формальний опис об'єкта, яким можна оперувати, використовуючи математичний апарат класифікації. Таким описом найчастіше виступає база даних. Кожний запис бази даних несе інформацію про деяку властивість об'єкта. Набір вхідних даних (вибірку даних) розбивають на дві множини: навчальна (training set) і тестова (test set) [28]. У навчальну вибірку входять об'єкти, для яких відомі значення як незалежних, так і залежних змінних. На підставі навчальної вибірки будується модель визначення значення залежної змінної. Її часто називають функцією класифікації. Для одержання максимально точної функції до навчальної вибірки пред'являються такі основні вимоги:

– кількість об'єктів, які входять у вибірку, повинне бути досить великим. Чим більше об'єктів, тим побудована на їхній основі функція класифікації буде точніше;

– у вибірку повинні входити об'єкти, які представляють усі можливі класи;

– для кожного класу вибірка повинна мати достатню кількість об'єктів.

Тестова (test set) множина також містить вхідні й вихідні значення параметрів. Тут вихідні значення використовуються для перевірки працездатності моделі [29].

Для подальшої роботи з програмою нам потрібно створити файли навчальної та тестової вибірки для гістологічних та цитологічних зображень відповідно.

Створюємо по два файли (навчальна та тестова вибірки) для кожного виду зображень, які містять біля двохсот можливих варіантів даних. У файлах знаходиться інформація про ядра клітин. Це така інформація, як площа, окружність, довжина головної осі та діаметр Фарета, а також клас до якого відноситься ця клітина. Приклад вибірок наведено на рисунку 3.8 та 3.9.

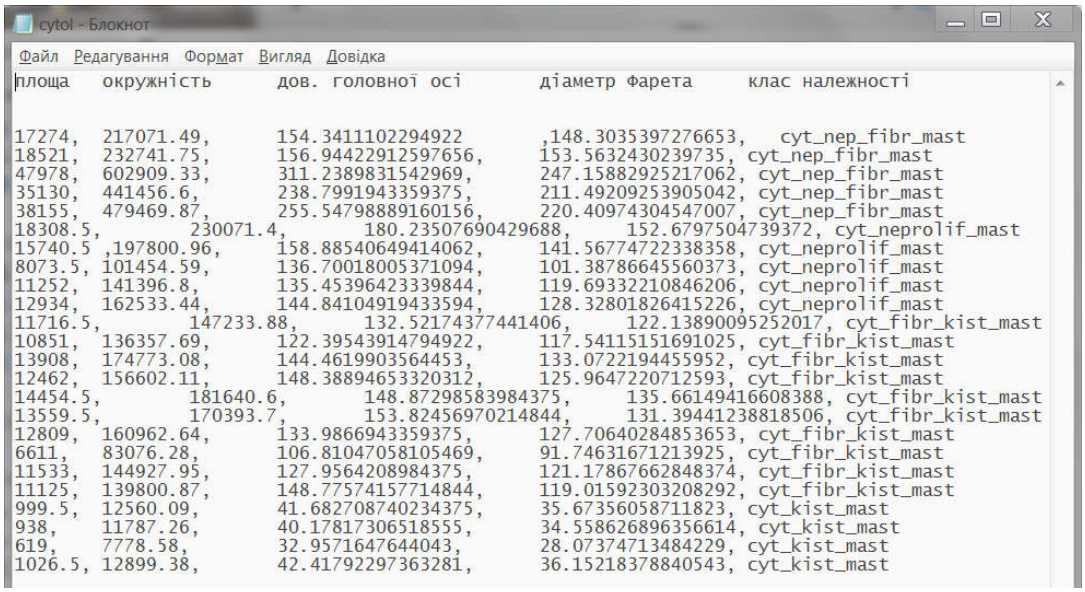
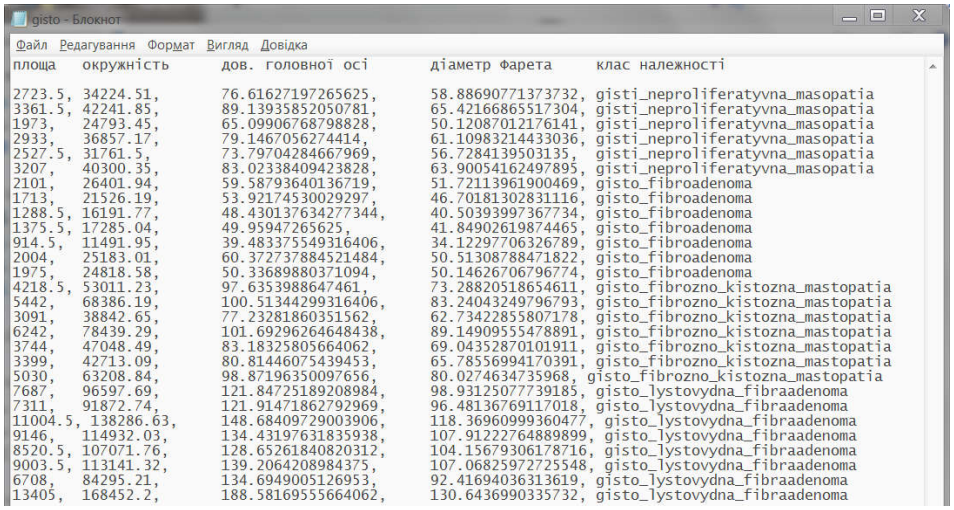


Рисунок 3.8 – Вибірка для цитологічних зображень

У цитологічних зображеннях виділено такі класи: непроліферативна мастопатія, непроліферативна фібро-мастопатія, фіброзно-кістозна мастопатія та кістозна мастопатія.


gisto - Блокнот

Файл Редагування Формат Вигляд Довідка

площа	окружність	дов. головної осі	діаметр Фарета	клас належності
2723.5,	34224.51,	76.61627197265625,	58.88690771373732,	gisti_neproliferatyvna_masopatia
3361.5,	42241.85,	89.13935852050781,	65.42166865517304,	gisti_neproliferatyvna_masopatia
1973,	24793.45,	65.09906768798828,	50.12087012176141,	gisti_neproliferatyvna_masopatia
2933,	36857.17,	79.1467056274414,	61.10983214433036,	gisti_neproliferatyvna_masopatia
2527.5,	31761.5,	73.79704284667969,	56.7284139503135,	gisti_neproliferatyvna_masopatia
3207,	40300.35,	83.02338409423828,	63.90054162497895,	gisti_neproliferatyvna_masopatia
2101,	26401.94,	59.58793640136719,	51.72113961900469,	gisto_fibroadenoma
1713,	21526.19,	53.92174530029297,	46.70181302831116,	gisto_fibroadenoma
1288.5,	16191.77,	48.430137634277344,	40.50393997367734,	gisto_fibroadenoma
1375.5,	17285.04,	49.95947265625,	41.84902619874465,	gisto_fibroadenoma
914.5,	11491.95,	39.483375549316406,	34.12297706326789,	gisto_fibroadenoma
2004,	25183.01,	60.372737884521484,	50.51308788471822,	gisto_fibroadenoma
1975,	24818.58,	50.33689880371094,	50.14626706796774,	gisto_fibroadenoma
4218.5,	53011.23,	97.6353988647461,	73.28820518654611,	gisto_fibrozno_kistozna_mastopatia
5442,	68386.19,	100.51344299316406,	83.24043249796793,	gisto_fibrozno_kistozna_mastopatia
3091,	38842.65,	77.23281860351562,	62.73422855807178,	gisto_fibrozno_kistozna_mastopatia
6242,	78439.29,	101.69296264648438,	89.14909555478891,	gisto_fibrozno_kistozna_mastopatia
3744,	47048.49,	83.18325805664062,	69.04352870101911,	gisto_fibrozno_kistozna_mastopatia
3399,	42713.09,	80.81446075439453,	65.78556994170391,	gisto_fibrozno_kistozna_mastopatia
5030,	63208.84,	98.87196350097656,	80.0274634735968,	gisto_fibrozno_kistozna_mastopatia
7687,	96597.69,	121.84725189208984,	98.93125077739185,	gisto_lystovydna_fibraadenoma
7311,	91872.74,	121.91471862792969,	96.48136769117018,	gisto_lystovydna_fibraadenoma
11004.5,	138286.63,	148.68409729003906,	118.36960999360477,	gisto_lystovydna_fibraadenoma
9146,	114932.03,	134.43197631835938,	107.91222764889899,	gisto_lystovydna_fibraadenoma
8520.5,	107071.76,	128.65261840820312,	104.15679306178716,	gisto_lystovydna_fibraadenoma
9003.5,	113141.32,	139.2064208984375,	107.06825972725548,	gisto_lystovydna_fibraadenoma
6708,	84295.21,	134.6949005126953,	92.41694036313619,	gisto_lystovydna_fibraadenoma
13405,	168452.2,	188.58169555664062,	130.6436990335732,	gisto_lystovydna_fibraadenoma

Рисунок 3.9 – Вибірка для гістологічних зображень

У гістологічних зображеннях виділено такі класи: непроліферативна мастопатія, фіброаденома, фіброзно-кістозна мастопатія, листовидна фіброаденома.

Після вибору шляху до файлів вибірки вказується кількість дерев та атрибутів і нажимаємо кнопку «Старт» для початку класифікації. Для класифікації цитологічних зображень вибираємо файли cyto та cyto1, а для гістологічних – gist та gist1 відповідно. Результат класифікації цитологічних зображень відображений на рисунку 3.6(4). В даному випадку відображається кількість прокласифікованих зображень та точність класифікації у відсотках. Існують помилки 1-го та та 2-го роду. Помилка 1-го роду – це відношення правильно прокласифікованих зображень до величини вибірки, а помилка 2-го роду – це відношення кількості неправильно прокласифікованих зображень до величини вибірки. Зробивши декілька класифікацій ми можемо побувати тривимірний графік залежності (X-помилка,Y-величина тестової вибірки, Z-кількість дерев). Графіки залежності для помилок 1-го та 2-го роду зображені на рисунках 3.11 та 3.12 відповідно.

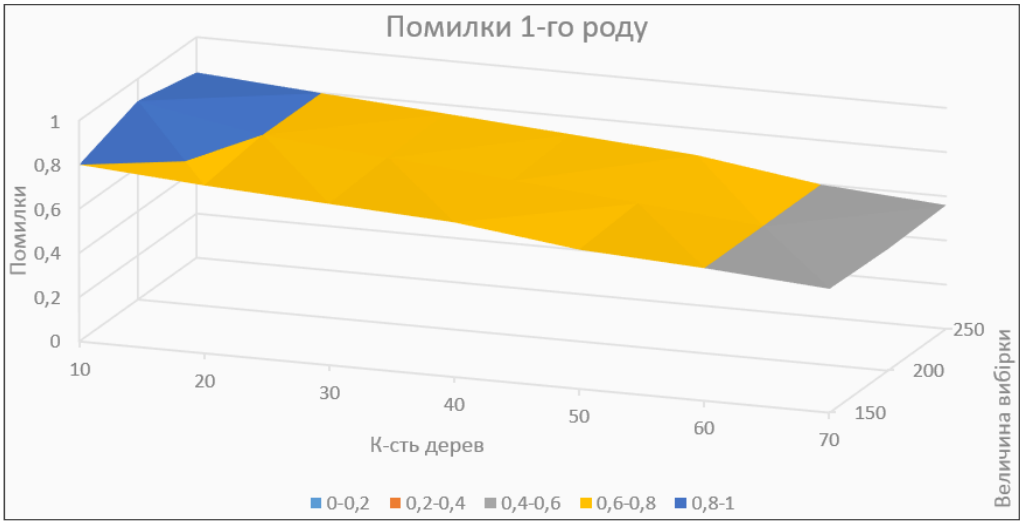


Рисунок 3.11 – Помилки 1-го роду

Зробивши аналіз рисунку 3.11 можна зробити висновок, що при збільшенні кількості дерев зменшується помилка 1-го роду. Від величини вибірки помилка змінюється в незначному розмірі.

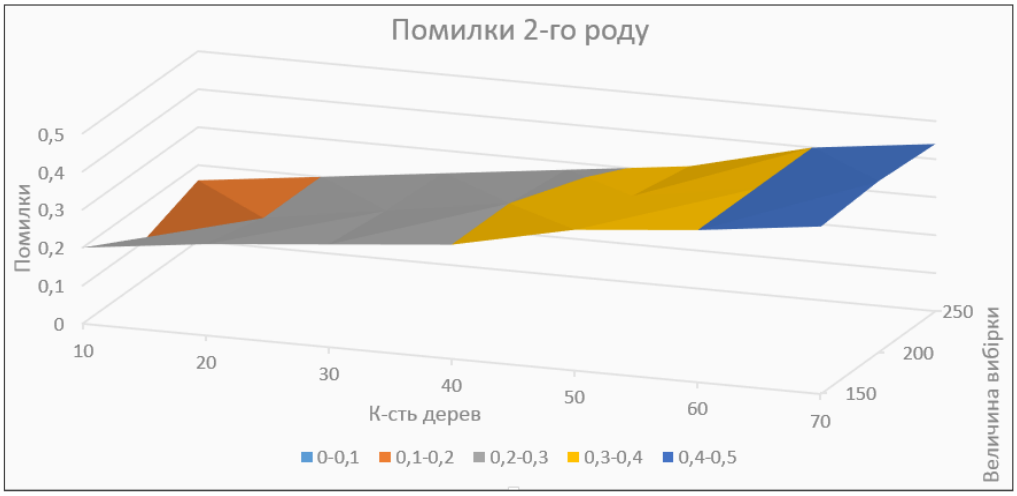


Рисунок 3.12 – Помилки 2-го роду

Проаналізувавши графік на рисунку 3.12 можна сказати, що при збільшенні кількості дерев збільшується помилка 2-го роду. Величина вибірки аналогічно як у помилки 1-го роду.

Розробляємо другу версію програми, яка буде виводити кінцевий діагноз. Для цього нам також потрібно по два файли вибірок. Головне вікно програми зображено на рисунку 3.13.

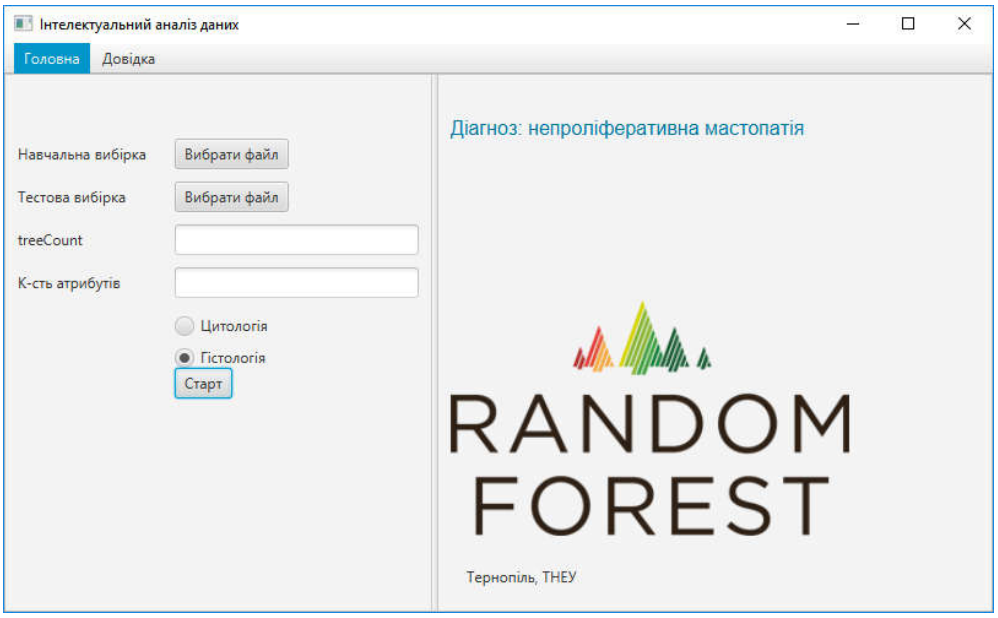


Рисунок 3.13 – Друга версія програми для виводу кінцевого діагнозу

У цій версії програми вказується шлях до вибірки для цитологічних або гістологічних зображень відповідно, кількість дерев та кількість атрибутів та зазначається для яких зображень проводиться класифікація. У результаті класифікації виводиться клас до якого належить зображення.

ВИСНОВКИ

У даній дипломній роботі було розроблено програмний модуль для класифікації гістологічних та цитологічних зображень на основі методу Random Forest.

В результаті виконання дипломної роботи:

– проаналізовано гістологічні та цитологічні зображення, розглянуто алгоритми класифікації та програмні засоби для розпізнання біомедичних зображень, проведено аналіз завдання на дипломну роботу та постановку задачі;

– охарактеризовано різновиди алгоритмів методу Random Forest, за допомогою порівнянь результатів класифікації цими алгоритмами, показано що точність класифікації прирівнюється до стандартного алгоритму Random Forest;

– створено програмний модуль на основі методу Random Forest, проведено класифікації за допомогою тестових та навчальної вибірок гістологічних та цитологічних зображень, за результатами тестування виявлено помилки 1-го та 2-го роду.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Guyon, I. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research./ I. Guyon, A. Elisseeff , 2003. – 340 с.

2. Данс Д. Рентгенодиагностические системы получения изображения // Физика визуализации изображений в медицине / Д. Данс / Под ред. С. Уэбба. Т. 1. – М.: Мир, 1991. – С. 40-106.

3. Лич М. Получение ЯМР изображений с пространственной локализацией // Физика визуализации изображений в медицине/ М. Лич / Под ред. С. Уэбба. Т. 2. – М.: Мир, 1991. – С. 105-223.

4. Суиндел Б. Рентгеновская трансмиссионная компьютерная томография // Физика визуализации изображений в медицине/ Б. Суиндел, С. Уэбб / Под ред. С. Уэбба. Т. 2. – М.: Мир, 1991. – С. 138-171.

5. Тринжаус Г. От клеток к органам./ Г. Тринжаус – М.:Мир,1972. – 320с.

6. Недзьведь А.М. Компьютерная обработка изображений сосудов и волокон биологических препаратов и изменение их геометрических характеристик / А.М. Недзьведь, Ю.Г. Ильич, Г.М. Карапетян и др. // Тез. докл. III науч. конф. по распознаванию и анализу изображений. Кн. 2. – Мн., 1995. – С. 110-113.

7. Frame A.J. Structural analysis of retinal vessels / A.J. Frame, P.E. Undrill, J.A. Olson et al. // Proc. of Sixth international conference on image processing and its applications. – 1997. – V. 2. – P. 824-827.

8. Subsol G. Non rigid registration for building 3D anatomical atlases/ G. Subsol, J.Ph. Thirion, N. Ayache // Proc. of 12-th IAPR international conference on pattern recognition. V. 1. – Jerusalem, Israel, 1994. – P. 576-557.

9. Афанасьев Ю.И., Гистология/ Ю.И. Афанасьев, Н.А. Юрина: Учеб. – М.: Медицина, 1998. – 795 с.

10. Харрис Г. Ядро и цитоплазма./ Г. Харрис – М.: Мир, 1973. – 340 с.

11. Хем А. Гистология./ А. Хем, Д. Кормак: В 5 т. – М.: Мир, 1982.

12. Автандилов Г.Г. Медицинская морфометрия./ Г.Г. Автандилов – М.: Медицина, 1990. – 384 с.

13. Иваницкий Г.Р. Математическая биофизика клетки./ Г.Р. Иваницкий, В.И. Кринский, Е.Е. Сельников – М.: Наука, 1978. – 308 с.

14. Breiman L. Random forests. / L. Breiman // Machine Learning. – Vol. 45, N 1. –2001. – P. 5–32.

15. Солодянкина С.В. Возможности использования данных дистанционного зондирования в крупномасштабном ландшафтном планировании./ С.В. Солодянкина // Материалы Международной научной конференции Ландшафтное планирование для России: итоги и перспективы. Иркутск. 2006. – 20 с.

16. Кохан С.С. Сучасні підходи до класифікації космічних знімків./ С.С. Кохан //Матеріали Міжнародної науково-методичної конференції "Географічні інформаційні системи в аграрних університетах (GISAU)". Херсон. 2007.

17. Зубков И.А. Применение алгоритмов неконтролируемой классификации при обработке данных./ И.А. Зубков, В.О. Скрипачев/ ДЗЗ. ФГУП «Научный центр космических информационных систем и технологий наблюдения». Москва 2007.

18. Image Processor v.3.0 Руководство пользователя. Программа обработки данных дистанционного зондирования Земли ScanEx Image Processor v.3.0. Москва, 2008 г.

19. Чистяков С.П. Случайные леса: обзор / С.П. Чистяков // Труды Карельского научного центра РАН. – № 1. – 2013. – С. 117–136.

20. Вапник В. Н. Теория распознавания образов./ В. Н. Вапник, А. Я. Червоненкис— М.: Наука, 1974.

21. Вапник В. Н. Восстановление зависимостей по эмпирическим данным./ В. Н. Вапник — М.: Наука, 1979.

22. Дуда Р. Распознавание образов и анализ сцен./ Р. Дуда, П. Харт — М.: Мир, 1976.

23. Nello Cristianini, John Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. — Cambridge University Press, 2000.

24. Вьюгин В. Математические основы машинного обучения и прогнозирования. / В. Вьюгин — МЦМНО, 2014. — 304 с.

25. Недзьведь А.М. Обработка медицинских изображений гистологических объектов./ А.М. Недзьведь, С.В. Абламейко // Цифровая обработка изображений. Вып. 4. – Мн.: Ин-техн. кибернетики НАН Беларуси, 2000. – С. 152-164.

26. Недзьведь А.М. Представление цветных изображений для математической морфологии./ А.М. Недзьведь, С.В. Абламейко // Компьютерный анализ данных и моделирование: Сб. науч. статей V Междунар. конф. / Под ред. С.А. Айвазяна, Ю.С. Харина. – Мн.: Изд-во БГУ, 1998. – Ч. 4: К-Я. – С. 86-95.

27. Недзьведь А.М. Связность пикселей на цветном изображении при использовании операций математической морфологии./ А.М. Недзьведь, С.В. Абламейко // Информационные системы и технологии (IST’2002): Мат. междунар. конф.– Мн.: Изд-во БГУ, 2002. – С. 41-52.

28. Чубукова И.А. Data mining: учебное пособие./ И.А. Чубукова – М.: Интернет- университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. – 382 с.

29. Ситник В.Ф. Інтелектуальний аналіз даних./ В.Ф. Ситник К.: КНЕУ, 2007.

30. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP./ А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод – СПб.: БХВ-Петербург, 2007. – 384 с.

31. Паклин Н.Б. Бизнес-аналитика: от данных к знаниям: Учеб. пособие. 2-е изд., перераб. и доп./ Н.Б. Паклин, В.И. Орешков – СПб.: Питер, 2010. – 704 с.

32. Breiman L. Out-of-bag estimates. Technical report./ L. Breiman, 1996. – 655c.

33. Breiman L. Random forests. Machine Learning./ L. Breiman, 45, P. 5– 32, October 2001.

34. Kifer D. Detecting change in data streams. In Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)./ D. Kifer, S. Ben-David, J. Gehrke., P. 180–191. Toronto, Canada, 2004.

35. Krishnamurthy B. Sketchbased change detection: methods, evaluation, and applications. In Proceedings of the 3rd ACM SIGCOMM Internet Measurement Conference (IMC)./ B. Krishnamurthy, S. Sen, Y. Zhang, Y. Chen, P. 234–247. Miami Beach, FL, 2003.

36. Bulut A. A unified framework for monitoring data streams in real time. In Proceedings of the 21st International Conference on Data Engineering (ICDE)./ A. Bulut, A. Singh., P.44–55. Tokyo, Japan, 2005.

37. Zhu Y. Statistical monitoring of thousands of data streams in real time. In Proceedings of the 28th International Conference on Very Large Data Bases (VLDB)./ Y. Zhu, D. Shasha., P.358–369. Hong Kong, China, 2002.

38. Aggarwal C. Sa framework for clustering evolving data streams. In Proceedings of 29th International Conference on Very Large Data Bases(VLDB)./ C. Aggarwal, J. Han, J. Wang, and P. Yu., P. 81–92. Berlin, Germany, 2003.

39. Gaber M. Costefficient mining techniques for data streams. In Proceedings of the 1st Australasian Workshop on Data Mining and Web Intelligence (DMWI)./ M. Gaber, S. Krishnaswamy, A. Zaslavsky, P. 81–92. Dunedin, New Zealand, 2003.

40. Domingos P. Mining high-speed data streams. In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)./ P. Domingos, G. Hulten., P. 71–80. Boston, MA, August 2000.

41. Hulten G. Mining timechanging data streams. In Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)./ G. Hulten, L. Spencer, P. Domingos., P. 97–106. San Francisco, CA, August 2001.

42. Random forests. Technical Report [Електронний ресурс]/ L. Breiman, 1999. Available at www.stat.berkeley.edu.

43. Caruana R. An empirical comparison of supervised learning algorithms. In Proc. Int. Conf. Mach. Learn. (ICML)./ R. Caruana, A. Niculescu-Mizil., 2006.

44. Lakshminarayanan B. Top-down particle filtering for Bayesian decision trees. In Proc. Int. Conf. Mach. Learn. (ICML)./ B. Lakshminarayanan, D. M. Roy, Y. W. Teh., 2013.

45. Местецкий Л.М. Математические методы распознавания образов. (Курс лекций) [Електронний ресурс]. ВмиК МГУ: Москва, 2004. – Режим доступу: www.ccas.ru/frc/papers/mestetskii04course.pdf.

46. Дуда Р. Распознавание образов и анализ сцен./ Р. Дуда, П. Харт – М.: Мир, 1976.

47. Brains J. IntelliJ IDEA Editions Comparison./ Jet Brains, 2003.

48. Кей С. Java. Библиотека профессионала, том 1. Основы./ С. Кей , Корнелл Г. – 9-е издание. «Вільямс», 2013.