Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту
(повна назва)
Кафедра Інформатики
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

ДОСЛІДЖЕННЯ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ ПОСТАНОВКИ МЕДИЧНИХ ДІАГНОЗІВ
(тема)

Виконав:
студент 2 курсу, групи ІНФм-21-1
Кривчикова Д.І.
(прізвище, ініціали)

Спеціальності 122 Комп'ютерні науки
(код і повна назва спеціальності)

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва освітньої програми)

Керівник доц. Руденко Д.О.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ Кобилін О.А.
(підпис) (прізвище, ініціали)

2022 р.

Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту
(повна назва)

Кафедра Інформатики
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва освітньої програми)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 2022 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Кривчиковій Дар'ї Ігорівні
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методів інтелектуального аналізу даних для постановки медичних діагнозів
затверджена наказом по університету від 9 листопада 2022 року № 1469Ст
2. Термін подання студентом роботи до екзаменаційної комісії 26 листопада 2022 р.
3. Вихідні дані до роботи: науково-методична та науково-технічна література, матеріали конференцій, дані інтернет-мережі, бібліотека sklearn, бібліотека tensorflow, мова програмування Python, середовище розроблення Jupyter Notebook.

4. Перелік питань, що потрібно опрацювати в роботі
1. Проаналізувати сучасні стан засобів медичної діагностики.
2. Проаналізувати існуючі застосування штучного інтелекту в медицині.
3. Оглянути застосування методів LSSVM та CNN.
4. Оглянути сутність методів LSSVM та CNN.
5. Здійснити вибір інструментального засобу для програмної реалізації.
6. Програмно реалізувати обрані методи.
7. Протестувати методи на обраному датасеті та провести аналіз результатів.
8. Виявити перспективи подальшої роботи.

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри) модель методу, графіки з характеристиками даних, тестові зображення. реконструйовані тестові зображення.

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Консультант з дотримання діючих стандартів та норм	Доцент Творошенко І.С.		

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	09.11.2022	
2	Аналіз завдання, підбір літератури	09.11.22-10.11.22	
3	Аналіз літератури з досліджуваної проблеми	10.11.22-12.11.22	
4	Аналіз обраних методів	12.11.22-13.11.22	
5	Програмна реалізація	13.11.22-15.11.22	
6	Тестування методів	15.11.22-16.11.22	
7	Оформлення пояснювальної записки	16.11.22-19.11.22	
8	Перевірка на плагіат	23.11.2021	
9	Рецензування	24.11.2021	
10	Підготовка презентації та доповіді	25.11.2021	
11	Занесення роботи в електронний архів	26.11.2021	
12	Попередній захист кваліфікаційної роботи	05.11.2021	

Дата видачі завдання 9 листопада 2022 р.

Студент _____
(підпис)

Керівник роботи _____ _доц. Руденко Д. О._
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ/ABSTRACT

Пояснювальна записка до кваліфікаційної роботи: 64 с., 3 табл., 11 рис., 38 джерел.

МЕТОД НАЙМЕНШИХ КВАДРАТІВ ОПОРНИХ ВЕКТОРІВ, КЛАСИФІКАЦІЯ ЗОБРАЖЕНЬ, МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ, ЗГОРТКОВА НЕЙРОННА МЕРЕЖА.

Об’єктом дослідження є датасет медичних даних в форматі зображень.

Метою дослідження є практичне застосування таких методів як LS-SVM та CNN та порівняння підходів класичних методів машинного навчання та нейромереж до трактування медичних зображень.

Був проведений аналіз застосування методів інтелектуального аналізу в діагностиці в цілому і методів розпізнавання зображень в окремих випадках. Були проаналізовані методи найменших квадратів опорних векторів та згорткова нейронна мережа, їх математичний апарат та принципи роботи. Методи були апробовані на наборі даних з зображеннями тканин з раком молочної залози. З порівняння отриманих результатів можна зробити висновки, що згорткова мережа більше підходить для діагностики при використанні медичних зображень.

LSSVM, CLASSIFICATION OF IMAGES, METHODS OF INTELLECTUAL ANALYSIS, CONVOLUTIONAL NEURAL NETWORK.

The object of the research is a dataset of medical data in image format.

The purpose of the research is the practical application of such methods as LS-SVM and CNN and the comparison of classical machine learning and neural network approaches to the interpretation of medical images.

An analysis of the application of intellectual analysis methods in diagnostics in general and image recognition methods in individual cases was conducted. The methods of least squares of support vectors and convolutional neural network, their mathematical apparatus and working principles were analyzed. The methods were tested on a dataset with breast cancer tissue images. By comparing the obtained results, it can be concluded that the convolutional network is more suitable for diagnosis using medical images.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

LSSVM – Least-Squares Support Vector Machine (метод найменших квадратів опорних векторів)

QP – Quadratic Programming (квадратичне програмування)

RBF – Radial Basis Function (радиально-базисна функція)

SVM – Support Vector Machine (метод опорних векторів)

SVR – Support Vector Regression (регресія опорних векторів)

МРТ – магнітно-резонансна томографія

КТ – комп’ютерна томографія

МІС – медично-інформаційні системи

ДР – діабетична ретинопатія

ML – Machine Learning (машинне навчання)

GP – Genetical Programming (генетичне програмування)

PLS – Partical Least Squares (часткові найменші квадрати)

CNN – Convolutional Neural Network (згорткова нейронна мережа)

RNN – Recurrent Neural Network (рекурентна нейронна мережа)

FNN – Feedforward Neural Network (нейронна мережа прямого поширення)

ВСТУП

Медична діагностика задає основу правильного лікування. Але нажаль, медичний діагноз не може бути повністю об’єктивним і залежить не тільки від даних пацієнта, але й від компетенції лікаря і навіть його поточного стану. Дослідження показують, що діагноз пацієнта може змінюватися залежно від спеціаліста, або навіть від часу огляду в разі, коли діагностикою займається один й той же лікар [1].

Методи інтелектуального аналізу даних, хоч і є відносно молодою галуззю серед комп’ютерних наук, стають все більш актуальними з кожним днем. Це відбувається завдяки збільшенню об’єму інформації та вдосконаленню методів її збору та обробки. Завдяки наявності великих об’ємів історичних даних, методи інтелектуального аналізу можуть давати достовірний рівень передбачень. Тому зараз такі методи застосовуються в багатьох галузях, зокрема і в медицині. Одне з самих перспективних застосувань інтелектуального аналізу в медицині – саме діагностика різноманітних захворювань.

Актуальність дослідження полягає у постійному розвитку медичної діагностики, в тому числі і завдяки методам машинного навчання та інтелектуального аналізу. Кожного дня в цій галузі з’являються нові дослідження та розроблюються нові продукти та системи. Тому дослідження використання методів інтелектуального аналізу даних з метою діагностики є перспективним полем для наукових пошуків [2, 3].

1 АНАЛІЗ ЗАСТОСУВАННЯ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ В СФЕРІ МЕДИЧНОЇ ДІАГНОСТИКИ

1.1 Аналіз сучасних систем діагностики

Експериментальні дослідження процесів розумової діяльності людини показали, що 85% часу йде на пошук необхідної інформації, друк, побудову графіків тощо – тобто створення передумов для розумової роботи. Це стосується й роботи медичного дослідника або практикуючого лікаря. Для автоматизації робіт на кожному з етапів діагностично-лікувального процесу застосовують медичні інформаційні системи (МІС) .

Створення медичної інформаційної системи переслідує кілька цілей:

– підвищення якості діяльності медичних працівників і установ охорони здоров’я шляхом організації досконалої (відповідної рівню використовуваних технічних засобів) обробки медичної інформації, у тому числі шляхом удосконалювання процесів керування та планування;

– полегшення праці медичних працівників, ліквідація трудомістких малоефективних процесів ручної обробки й аналізу медичних даних;

– забезпечення ефективного обміну інформацією з іншими інформаційними системами.

За призначенням МІС класифікують на:

– системи, основною функцією яких є накопичення даних (автоматизовані системи обробки даних та (або) інформації, автоматизовані інформаційні та інформаційно-довідкові системи);

– діагностичні та консультаційні системи;

– системи, що забезпечують медичне обслуговування.

Найбільш загальні задачі МІС, що вирішуються у клінічних установах:

– об’єктивізація трактування результатів досліджень (по деяким даним, невірне тлумачення результатів рентгенологічного, електрокардіологічного та

лабораторних досліджень приводить у 30% випадків до помилкового діагнозу);

– автоматизація обробки інформації на етапі попередньої роботи медичного персоналу по визначенню діагнозу та виробленню тактики лікування (лікар приймає остаточне рішення з питань діагностики і лікування хворого);

– автоматизація лабораторних досліджень: біохімічних, електрофізіологічних, рентгенорадіологічних інших;

– створення баз (банків) даних: накопичення відомостей про кожного хворого для подальшого аналізу матеріалу, організації обробки цієї інформації відповідним математичним забезпеченням (у тому числі системами управління базами даних);

– створення баз знань: накопичення знань експертів в області медицини й системи охорони здоров'я, необхідних для розробки експертних систем діагностики, лікування та реабілітації, профілактичних оглядів, експертизи, планування та управління;

– упорядкування потоку інформації усередині медичної установи (задачі організаційного керування, задачі кадрові, матеріально-технічного постачання, статистичні звіти, оцінка діяльності відділень лікарень по деяких показниках тощо).

Щоб бути хорошим діагностом, лікар повинен мати дуже великий набір знань та вміти пов'язувати набір симптомів з можливою хворобою, а також робити попередні висновки про її лікування та наслідки. За статистикою, що чим раніше пацієнти починають лікування, тим більше шансів зберегти своє здоров'я.

1.2 Використання методів штучного інтелекту в сфері медицини

1.2.1 Огляд напрямків застосування методів інтелектуального аналізу даних

Медицина та охорона здоров’я є плідним підґрунтям для впровадження новітніх рішень з застосуванням методів інтелектуального аналізу [4, 5]. Сучасні розробки виводять медичну сферу на якісно новий рівень – від діджиталізації медичних записів для розпізнавання ліків до моніторингу пацієнтів після лікування.

В останні роки медичні установи почали використовувати деякі методи штучного інтелекту, щоб пришвидшити та поліпшити процес діагностики та аналізу. Однією з цілей цих введень є те, щоб лікарі отримували підтримку в проведенні діагностики та швидше досягали фази лікування. Завдяки машинному навчанню, яке використовується для аналізу діагностичних звітів або зображень, як наприклад КТ, виявлення аномалій або навіть злоякісних утворень, може бути проведено з надлюдським рівнем якості. Аналітика та гіпотези, сформовані моделями з використанням машинного навчання, переглядаються досвідченими лікарями та допомагають їм приймати правильні рішення щодо своїх пацієнтів. Таким чином, спеціалісти можуть більш ефективно виявляти пацієнтів з найсерйознішими станами та вірно визначати пріоритети, наслідком чого стане також підвищення рівня успіху у запобіганні хвороб, які ще можна попередити. Маючи більшу кількість точних даних про індивідуальні особливості пацієнта та детальні симптоми, є можливість отримати точніші рецепти та більш персоналізовану допомогу. Завдяки точним рецептам знижується вірогідність виникнення у пацієнта побічних ефектів від медикаментів, оскільки лікування спеціально розроблене для потреб пацієнта.

Діагностика та прогнозування передбачають тлумачення зображень ураженої ділянки, отриманих за допомогою рентгену, магнітно-резонансної томографії, комп’ютерної томографії, позитронно-емісійної томографії,

однофотонної емісійної комп’ютерної томографії або ультразвукового сканування. Тлумачення зображення передбачає виявлення аномалій, визначення їх розташування та меж, а також оцінку їх розмірів і тяжкості. Дефіцит експертів-людей та їхня втома, висока плата за консультації та грубі процедури оцінки обмежують ефективність тлумачення зображення. Крім того, форми, розташування та структури медичних аномалій дуже варіабельні. Це ускладнює діагностику навіть для лікарів-спеціалістів. Тому лікарі-експерти часто відчувають потребу в допоміжних інструментах, щоб допомогти в точному розумінні медичних зображень.

Це є мотивацією для створення інтелектуальних систем розпізнавань зображень у сфері медицини [6, 7].

Передбачувальна аналітика є одною з найпопулярніших напрямків застосування методів інтелектуального аналізу в сфері в охорони здоров’я. Вона дозволяє додати бажані історичні дані у модель машинного навчання та отримувати досить точні прогнози шляхом навчання на цих даних [8]. Подібні моделі можуть знаходити кореляції та асоціації симптомів, звичок, хвороб та робити значущі прогнози на їх основі. Лікарі та інші медичні спеціалісти можуть використовувати такий підхід для прогнозування різних захворювань за способом життя пацієнта, його звичками в харчування та інших сферах, та різними видами діяльності, наприклад професією чи хобі. Також передбачувальна статистика може застосовуватися лікарями при вже відомому захворюванні та спрогнозувати його перебіг та погіршення стану здоров’я пацієнта. Це дозволяє ефективніше вживати профілактичних заходів, які зменшать ризик погіршення стану пацієнтів та його тяжкість [9].

Передбачувальна аналітика також обширно застосовується та відіграє важливу роль у підвищенні ефективності логістики в цілому, що не оминає ланцюги фармацевтичних поставок. Моделі машинного навчання передбачують загальні місцеві потреби та сезонні тренди фармацевтичної логістики в певній лікарні, що запобігає виникненню дефіциту ліків у разі надзвичайних ситуацій.

Незважаючи на велику накоплену кількість медичних даних, рівень діагностики все ще не є достатнім, насамперед через проблеми обробки та інтерпретації цих даних.

Згідно з останнім дослідженням Національних академій наук, техніки та медицини [10], «щороку в США близько 5 відсотків дорослих пацієнтів отримують неправильний діагноз. Це налічує понад 12 мільйонів людей. Більше того, результати дослідження посмертного обстеження показують, що діагностичні помилки спричиняють приблизно 10 відсотків смертей пацієнтів».

Але застосування інформаційних технологій та науки про даних може надати інструменти та методи для вилучення реальної цінності з неструктурованої медичної інформації, та в підсумку посприяти підвищенню ефективності, доступності та персоналізації охорони здоров'я. За даними Коледжу керівників управління інформацією в галузі охорони здоров'я США «кількість медичних установ, які приймають рішення на основі даних, повільно, але неухильно зростає. У 2015 році тільки 15% лікарень США використовували аналіз даних і прогнозну аналітику для запобігання повторної госпіталізації. Через рік 31% установ заявили, що роблять це більше року» [11].

Найпоширеніші напрямки застосування методів штучного інтелекту та машинного навчання включають у себе:

– для певних захворювань, таких як рак, рання діагностика підвищує виживання та здоровий спосіб життя, а також зменшує витрати на лікування. Машинне навчання (зокрема, згорткові нейронні мережі) виявилося потужною допомогою в цьому контексті. Алгоритм добре підходить для задачі багатокласової класифікації (прогнозування зображення як автомобіля, велосипеда чи фургона) або бінарної класифікації (наприклад – передбачення наявності на медичному зображенні злоякісної пухлини чи ні);

– остеоартрит є одним із тих захворювань, які неможливо виявити, доки воно не завдасть шкоди, та за статистикою, воно розвинеться у 1 з 10 людей.

До сьогодні часу лікарі не могли виявити це, поки не почалося пошкодження кісток. На магнітно-резонансній томографії хряща колінного суглоба відмінності майже непомітні для досвідчених лікарів. За допомогою машинного навчання тепер можна вводити навчальний набір зображень минулих пацієнтів, у яких була діагностована хвороба, і здорових, алгоритм виявляє ці закономірності та може передбачити, хто буде сприйнятливий до захворювання через 3 роки;

– діабетична ретинопатія (ДР) є найшвидше зростаючою причиною сліпоти. Сьогодні у світі налічується 415 мільйонів людей з діабетом, і кожен потенційно ризикує отримати діагноз ДР. Ключем до запобігання сліпоти є регулярне обстеження, оскільки ДР не виявляє жодних симптомів, доки не стане надто пізно до моменту втрати зору. Під час процесу скринінгу лікарі шукають плями або крововиливи, щоб виявити наявність ДР, і оцінюють їх за шкалою, щоб показати здоров'я ока пацієнта;

– у крайніх випадках (відсутність захворювання або кінцева стадія) лікарі досить послідовні у своїх діагнозах. Але в середньому діапазоні діагноз не є достовірним. Це пов'язано з тим, що навіть людське око досвідченого лікаря не дуже ефективно вирішує цю проблему розпізнавання зображень;

– ВІЛ вражає близько 36 мільйонів людей у всьому світі. Потрібне довічне лікування антиретровірусними (не вбивають вірус, але перешкоджають його подальшому розвитку) препаратами. ВІЛ є складним, оскільки швидко мутуючий вірус призводить до стійкості до ліків. Це означає, що лікарі повинні з часом міняти препарати, щоб перехитрити вірус. Це означає, що дуже важлива послідовність введення препаратів;

– використовуючи просту ілюстрацію, якщо необхідно лікувати пацієнта різною комбінацією ліків і якщо один із препаратів викликає у пацієнта стійкість до інших ліків, потрібно буде подумати про послідовність призначених ліків і звернути пильну увагу на те, як ліки впливають на здоров'я пацієнта для тривалого лікування;

– майбутній довгостроковий догляд залежить від поточного та попереднього замовлення препаратів навіть для подібного вірусного навантаження. Дослідження показали, що люди погано вміють відстежувати історію минулих ліків і з'ясовувати вплив цих препаратів на майбутній догляд за допомогою аналітики в реальному часі;

– навчання з підкріпленням включає в себе глибоку нейронну мережу (агент), яка намагається орієнтуватися в невідомому середовищі для досягнення відчутного результату з регулярними інтервалами. Завдяки цьому постійному зворотному зв'язку модель розуміє дії, які винагороджують її позитивно чи негативно. У нашому прикладі модель продовжує відстежувати численні біомаркери з кожним прийомом препарату (і надає зворотний зв'язок у формі позитивної винагороди або негативного попередження) та використовує ці знання з часом для визначення найкращої стратегії довгострокового догляду.

1.2.2 Аналіз методів штучного інтелекту, які використовуються в сфері діагностики захворювань

Системи розшифровки зображень, які використовують методи машинного навчання, швидко розвиваються в останні роки. Методи ML включають навчання дерева рішень , кластеризацію, опорні векторні машини, *k*-середні найближчі сусіди, обмежені машини Больцмана і випадкові ліси [12–15]. Передумовою для ефективної роботи методів ML є виділення дискримінантних ознак. І ці функції, як правило, невідомі, а також є дуже складним завданням, особливо для додатків, що включають тлумачення зображення, і все ще є темою багатьох досліджень. Логічним кроком для подолання було створення інтелектуальних машин, які могли б вивчати функції, необхідні для тлумачення зображення, та витягувати його самостійно [16, 17]. Однією з таких розумних і успішних моделей є модель

згорткової нейронної мережі (CNN), яка автоматично вивчає необхідні функції та виділяє їх для тлумачення медичного зображення.

Аналіз літератури по темі підтвердив, що вже існує безліч досліджень з використанням методів штучного інтелекту в сфері діагностики. Як приклад можна навести такі з них:

Дослідження Дабовси та інших використовували нейронну мережу зворотного поширення для діагностики захворювань шкіри, щоб досягти найвищого рівня точності. Автори використовували реальні дані, зібрані з дерматологічного відділення.

Ансарі та інші використовували рекурентну нейронну мережу для діагностики вірусного гепатиту захворювання печінки та досягли 97,59%, тоді як нейронна мережа прямого зв'язку досягла 100%.

Овасіс та інші отримали 97,057 площі під кривою за допомогою залишкової нейронної мережі та довготривалої короткочасної пам'яті для діагностики захворювань шлунково-кишкового тракту.

Скаане та та інші дослідили властивість цифрового томосинтезу молочної залози на період і виявили рак у резидентів на основі скринінгу. Вони провели обстеження подвійного аналізу, залучивши жінок віком 50–69 років, і порівняли повнопольну цифрову мамографію та інструмент збору даних із повнопольною цифровою мамографією. Накопичення інструменту створення даних призвело до незначного підвищення чутливості на 76,2% і значного збільшення на 96,4%.

Тігга та інші мали на меті оцінити ризик діабету серед пацієнтів на основі їх способу життя, розпорядку дня, проблем зі здоров'ям тощо. Вони експериментували з 952 опитувальниками, зібраними за допомогою офлайн- та онлайн-анкет. Те саме було застосовано до бази даних Pima Indian Diabetes. Класифікатор випадкового лісу виявився найкращим алгоритмом.

Кетрін та інші надали огляд типів даних, які зустрічаються під час встановлення хронічного захворювання. Використовуючи різні алгоритми

машинного навчання, вони пояснили теорію екстремальних значень для кращої кількісної оцінки тяжкості та ризику хронічних захворювань.

Гонсалвес та інші мали на меті прогнозувати ішемічну хворобу серця, використовуючи історичні медичні дані за допомогою технології машинного навчання. Представлена

робота підтримувала три методи навчання під наглядом під назвою «Наївний Байєс», «Машина опорних векторів» і «Дерево рішень», щоб знайти кореляції в ішемічній хворобі серця, що допомогло б покращити швидкість прогнозування.

Шабут та інші запровадив іспит для вдосконалення розумного, універсального, уповноваженого майстра для розігрування запрограмованого відкриття туберкульозу. Вони застосували адміністрований метод штучного інтелекту для досягнення паралельного групування з вісімнадцятої нижчої хвилини затінення запиту. Їхній тест показав точність 98,4%, особливо для ідентифікації антигену туберкульозного антигену на портативному столі.

Тран та інші надав глобальні тенденції та розробки додатків штучного інтелекту, пов’язаних з інсультом та серцевими захворюваннями, щоб виявити прогалини в дослідженнях і запропонувати майбутні напрямки досліджень.

Ратход та інші запропонували автоматизовану систему пошуку захворювань шкіри на основі зображень із використанням класифікації машинного навчання.

Срінівасу та та інші запропонували ефективну модель, яка може допомогти лікарям ефективно діагностувати захворювання шкіри. Система об’єднала нейронні мережі з MobileNet V2 і довготривалою короткостроковою пам’яттю із рівнем точності 85%, що перевищує інші найсучасніші глибокі моделі нейронних мереж глибокого навчання. Ця система використовувала техніку для аналізу, обробки та генерування даних зображення, передбачених на основі різних характеристик. Як наслідок, він дав більшу точність і генерував швидше результати порівняно з традиційними методами. Тією з групою вчених було використано алгоритм AW-HARIS для виконання

автоматизованої сегментації зображень комп'ютерної томографії для виявлення аномалій у печінці людини. Помічено, що запропонований підхід перевершив у більшості випадків з точністю 78%.

1.3 Існуючі застосування класифікатора LS-SVM в медицині

1.3.1 Класифікація типів раку

Метод SVM як класифікатор є досить популярним в полі в класифікації раку відколи стали доступні дані з про експресію генів з мікрочипів.

Перше з таких досліджень провів Голуб, спробувавши лінійний SVM для класифікації двох різних типів лейкемії за допомогою даних мікрочипів експресії генів [18]. У цьому дослідженні як навчальна вибірка використовувалися дані 38 пацієнтів. Алгоритм був навчений розпізнавати різницю між двома різними формами лейкемії. Чіпи Affymetrix Hgu6800 охоплювали 7129 генних ознак і кожна з них зважувалася, відносячи кожен конкретний зразок к одному з класів. Навчену модель SVM після використовували для тестування інших незалежних даних інших 34 пацієнтів. Це дослідження продемонструвало чудові показники SVM при класифікації даних з високою розмірністю та з низьким розміром вибірки. Згодом Вапнік покращив точність зваженого методу голосування SVM, зменшивши рівень помилок з 2 помилки з 34, що становить 6% до 0% [19, 20]. Але в цьому дослідженні до розробки моделі не проводився відбір ознак.

Вищезгадані SVM – це двійкові класифікатори зразків. Рак найчастіше є неоднорідним, і тому потребує багатокласової класифікації. SVM, як и LSSVM має можливість розширення для роботи з багатокласовими, використовуючи так званий підхід «один проти одного». Для багатокласових проблем метод опорних векторів навчатиметься незалежно по кожному класу, який на момент навчання буде помічений як позитивний, а всі інші класи будуть складати негативні випадки. Навіть при багатокласовій класифікації, яка є набагато

складнішою, ніж звичайна бінарна, порівнюючи в різні найсучасніші методи на численних наборах даних експресії генів було виявлено, що метод опорних векторів є одним з найефективніших, завдяки його особливості в гарній роботі на даних, що мають високу розмірність та малий розмір вибірки.

Як і його попередник, метод найменших квадратів опорних векторів досить часто застосовується в полі діагностики раку. З існуючих досліджень можна навести декілька прикладів.

Комак та інші провели дослідження, спрямоване на діагностику захворювань печінки за допомогою нового гібридного методу машинного навчання. Шляхом гібридизації LSSVM з попередньою обробкою нечіткого зважування було отримано метод вирішення цієї проблеми діагностики шляхом класифікації захворювань печінки. Стадія попередньої обробки нечіткого зважування була розроблена вперше. Цей набір даних про захворювання печінки є дуже часто використовуваним набором даних у літературі, що стосується використання систем класифікації для діагностики захворювань печінки, і він використовувався в цьому дослідженні для порівняння ефективності класифікації запропонованого нами методу з іншими дослідженнями. Була отримана точність класифікації 94,29%, що є найвищим показником, досягнутим досі [21].

Полат та Гунес [22] провели діагностику раку молочної залози за допомогою алгоритму класифікатора опорних векторів найменших квадратів. Надійність LS-SVM перевіряється за допомогою точності класифікації, аналізу чутливості та специфічності, методу k-кратної перехресної перевірки та матриці плутанини. Отримана точність класифікації становить 98,53% і є дуже багатообіцяючою порівняно з методами класифікації, про які повідомлялося раніше. Отже, за допомогою LS-SVM отримані результати показують, що використаний метод може зробити ефективну інтерпретацію та вказати на можливість розробки нової інтелектуальної системи діагностики допомоги.

Хуанг [23] та інші запропонували метод прогнозування раку молочної залози, заснований на опорному векторі найменших квадратів, щоб точно передбачити рак молочної залози. Дослідження проводилося на основі даних на основі 469 пацієнтів. В результаті точність передбачення LS-SVM вища, ніж у звичайного методу опорних векторів, що забезпечує новий метод для діагностики раку молочної залози.

1.3.2 Застосування методу в інших сферах медицини

Виявлення біомаркерів передбачає вибір біологічно значущої або асоційованої експресії генів, ДНК або мікро-РНК з даних великих розмірів та моделювання балів на основі вибраних особливостей, що допомагають діагностувати рак, прогнозувати або відповідати на лікування. Цей процес можна розглядати як відбір ознак для класифікації (рак проти не раку, доброякісний проти злоякісного, реакція на ліки проти класів без реакції). Існує два основних методи вибору ознак: методи фільтрації та обгорткові методи. Ху побудував алгоритм SVM, заснований на принципі мінімізації структурного ризику для ідентифікації 38 маркерів, що беруть участь у розвитку мозку, на основі одноклітинних транскриптомних даних [24]. Для прогнозування раку молочної залози було застосовано вибір функцій SVM на основі профілювання метаболітів РНК в сечі.

Джаган [25] та інші провели дослідження прогнозування кількості відходів, які утворюються в лікарні, що може допомагати управлінню ними для кількох видів діяльності, таких як зберігання, транспортування та утилізація. В їх роботі використовується опорна векторна машина, опорна векторна машина найменших квадратів і генетичне програмування, щоб оцінити швидкість утворення медичних відходів. У разі прогнозування показника, тип лікарні, місткість і зайнятість ліжок використовувалися як вхідні дані SVM, LSSVM і GP. SVM базується на статистичній теорії навчання,

яка надає елегантний інструмент для моделювання нелінійних систем. Ці SVM, LSSVM і GP були використані як методи регресії. Результати показують, що продуктивність розроблених моделей перевищувала 90%.

Форд та Ленд [26] представили дослідження аналізу експресії генів на мікрочипах. Це швидкий і недорогий метод аналізу профілів експресії генів для прогнозу раку. Дані мікроматриць, що отримані в результаті онкологічних досліджень, зазвичай містять тисячі значень експресії з кількома випадками. Частковий метод найменших квадратів – це метод зменшення розмірності, який створює регресійну модель найменших квадратів у зменшеному просторі розмірів. Добре відомо, що метод опорних векторів перевершує регресійні моделі найменших квадратів. У цьому дослідженні була замінена модель PLS моделлю SVM у зменшеному розмірному просторі PLS. Щоб перевірити метод, була проведена робота з загальнодоступним набором даних із бази даних Gene Expression Omnibus, що містить рівні експресії генів, клінічні дані та час виживання пацієнтів з недрібноклітинною карциномою легенів. Використовуючи 5-кратну перехресну перевірку та аналіз робочих характеристик приймача, було показано порівняння продуктивності класифікатора між традиційною моделлю PLS і гібридом PLS/SVM. Результати показують, що заміна регресії найменших квадратів на SVM підвищує якість моделі.

Також різні модифікації методу опорних векторів можуть застосовуватися при створенні та виготовленні медикаментів. Основні проблеми, пов’язані зі створенням ліків від раку, включають побічні ефекти медикаментів, високу токсичність та резистентність пацієнта до діючих та існуючих протипухлинних препаратів. Традиційний процес створення ліків включає ітеративну процедуру пошуку сполук, які можуть активно діяти проти біологічної цілі. Цей процес займає багато часу та експоненційно зростає при виборі з великої множини сполук. Експериментальні методи, що використовуються для виявлення та створення медикаментів, є дорогими та трудомісткими. Сьогодні SVM може допомогти цьому процесу скринінгу,

використовуючи гіперплощини з максимальним полем. Подібна гіперплощина відокремлює активну сполуку від неактивної сполуки та має найбільшу можливу відстань від будь-якої міченої сполуки.

1.4 Застосування методу CNN в сфері медицини

Методи візуалізації використовуються для фіксації аномалій людського тіла. Отримані зображення необхідно розшифровувати для діагностики, прогнозу та планування лікування аномалій. Розшифровування медичного зображення зазвичай виконується кваліфікованими медичними працівниками. Однак обмежена доступність людей-експертів, втома та процедури грубої оцінки, пов’язані з ними, обмежують ефективність тлумачення зображення, яке виконують кваліфіковані медичні працівники. Згорткові нейронні мережі (CNN) є ефективними інструментами для тлумачення зображень. Вони перевершили експертів-людей у багатьох завданнях розшифровки зображень.

Модель CNN складається зі згорткових фільтрів, основною функцією яких є вивчення та виділення необхідних функцій для ефективного тлумачення медичного зображення. CNN почав набирати популярність у 2012 році завдяки AlexNet [27], моделі CNN, яка перемогла всі інші моделі з рекордною точністю у ImageNet challenge 2012. CNN використовувався корпоративними гігантами для надання інтернет-послуг [28], автоматичного тегування на зображеннях, рекомендацій щодо продуктів, персоналізації домашньої стрічки. Основні програми CNN – обробка зображень і сигналів, обробка природної мови та аналіз даних. CNN зробив великий прорив, коли GoogleNet використовував його для виявлення раку з точністю 89%, тоді як патологоанатоми змогли досягти точності лише 70%.

Підходи, засновані на CNN, поміщені в таблицю лідерів багатьох викликів розуміння зображень, таких як медичне обчислення зображень і комп’ютерне втручання (MICCAI), біомедичне завдання сегментації пухлини

мозку (BRATS), завдання мультимодальної сегментації пухлини мозку, завдання класифікації Imagenet, виклики Міжнародної конференції з розпізнавання образів (ICPR) [29] та завдання сегментації уражень ішемічного інсульту (ISLES). CNN став потужним вибором як техніка для розуміння медичного зображення. Дослідники успішно застосували CNN для багатьох програм розуміння медичних зображень, таких як виявлення пухлин і їх класифікація на доброякісні та злоякісні, виявлення уражень шкіри, виявлення зображень оптичної когерентної томографії , виявлення раку товстої кишки, рак крові, аномалії серця, молочної залози, грудної клітини, очей тощо. Крім того, моделі на основі CNN, такі як CheXNet [30], які використовуються для класифікації 14 різних захворювань грудної клітки, досягли кращих результатів порівняно з середньою продуктивністю експертів-лікарів.

CNN також домінують у сфері виявлення COVID-19 за допомогою рентгенів грудної клітки/КТ. Дослідження за участю CNN зараз є домінуючою темою на великих конференціях [31]. Крім того, існують спеціальні випуски, зарезервовані в відомих журналах для вирішення проблем за допомогою моделей глибокого навчання. Величезна кількість літератури, доступної по CNN, є свідченням їх ефективності та широкого використання. Однак різні дослідницькі спільноти розробляють ці програми одночасно, а результати розповсюдження розкидані в широкому та різноманітному діапазоні матеріалів конференцій і журналів.

1.5 Постановка задачі дослідження

Таким чином, методи інтелектуального аналізу є перспективним інструментом для використання в сфері медичної діагностики. Тому ставиться завдання практичного дослідження методів LS-SVM та CNN, та порівняння їх ефективності.

Об’єктом дослідження є датасет медичних даних в форматі зображень.

Метою дослідження є практичне застосування таких методів, як LS-SVM і CNN та порівняння підходів класичних методів машинного навчання та нейромереж до трактування медичних зображень.

Для досягнення мети необхідно вирішити такі завдання:

– проаналізувати існуючі методи інтелектуального аналізу, що застосовуються в діагностиці;

– провести аналіз методів LS-SVM та CNN;

– провести експериментальне дослідження методів на реальному медичному наборі даних, використовуючи їх комп’ютерну модель;

– порівняти отримані результати.

2 МАТЕМАТИЧНІ МЕТОДИ КЛАСИФІКАЦІЇ МЕДИЧНИХ ЗОБРАЖЕНЬ

2.1 Математичний апарат методу найменших квадратів опорних векторів

2.1.1 Ідея та розробка методу найменших квадратів опорних векторів

Наприкінці 90-х років 20-сторіччя В.Н. Вапником було запропоновано метод опорних векторів. Цей метод був створений для розв’язування задач класифікації та оцінювання нелінійних функцій. У цьому новому методі навчальна проблема переформульована та представлена таким чином, щоб отримати задачу квадратичного програмування. Рішення цієї проблеми квадратичним програмуванням є глобальним і унікальним. У методі опорних векторів можна вибрати кілька типів функцій ядра, включаючи лінійні, поліноміальні, RBF, MLP з одним прихованим шаром і сплайни, якщо виконується умова Мерсера. Крім того, у статистичній теорії навчання доступні межі похибок узагальнення, які виражаються у VC-вимірності (розмірність Вапника-Червоненкіса). Верхню межу цієї VC-розмірності можна розрахувати, розв’язавши іншу задачу квадратичного програмування. Незважаючи на приємні властивості SVM, все ще є деякі недоліки щодо вибору гіперпараметрів і того факту, що розмір матриці, яка бере участь у задачі QP, прямо пропорційний кількості точок навчання.

Модифікована версія класифікаторів SVM, класифікатори SVM найменших квадратів (LS-SVM), була запропонована в роботі Сайкенса та Вандевелля у 1999 році [32]. Дві норми були взяті з рівністю замість обмежень нерівності, щоб отримати лінійний набір рівняння замість задачі QP у подвійному просторі. У цьому сенсі формулювання LS-SVM пов’язане з мережами регуляризації, регресією Гаусових процесів та з рідж-регресією SVM для оцінки нелінійних функцій, але з включенням терміну упередження,

який має значення для алгоритмів. Первинно-дуальні формулювання SVM та обмеження в вигляді рівностей формулювання LS-SVM дозволяють робити розширення до періодичних нейронних мереж та нелінійного оптимального управління.

QP-проблема відповідності формулювання SVM, як правило, вирішується методами внутрішньої точки, послідовною мінімальною оптимізацією та ітеративно переваженими підходами найменших квадратів, в той час як LS-SVM приводять до набору лінійних рівнянь. У роботі Сайкенса та інших був розроблений ітераційний метод на основі спряженого градієнта для розв'язання відповідної системи Каруша-Куна-Такера [33]. Недоліком LS-SVM є те, що розрідженість втрачається завдяки вибору 2-норми. Однак це можна обійти на другому етапі процедурою обрізки, яка базується на видаленні навчальних точок, керуючись сортованим спектром значень опори, і не передбачає обчислення зворотних гессіанських матриць, як у класичних методах обрізки нейронних мереж. Це важливо з огляду на еквівалентність між розрідженим наближенням і SVM, як показано у роботі Гіросі [34].

Пряме розширення LS-SVM для вирішення багатокласових проблем було запропоновано Сайкенсом and та Вандевеллем у 1999 році. Для цього беруться додаткові результати для кодування мультикласів, як це часто робиться в класичній методології нейронних мереж. Цей підхід відрізняється від стандартних багатокласових підходів SVM. У запропонованій роботі багатоспіральна контрольна задача, яка, як відомо, важка для класичних нейронних мереж, була вирішена за допомогою LS-SVM з мінімальною кількістю бітів у кодуванні класу і дала чудові результати узагальнення. Цей підхід тут також поширюється на різні типи кодування вихідних даних для класів, такі як коди «один проти всіх», «один проти одного» та виправлення помилок.

2.1.2 Класичний метод опорних векторів

Припустимо, що вибірка лінійно роздільна, тобто існують такі значення параметрів *w, w_0*, при яких функціонал числа помилок приймає нульове значення:

$$[\mathrm{Q}(\mathrm{w}, w_0) = \sum_{i=1}^{\ell}[y_i(\langle w, x_i\rangle - w_0)] < 0]. \quad (2.1)$$

Але тоді розділяюча гіперплощина не єдина, оскільки існують і інші положення розділяючої гіперплощини, що реалізують те ж саме розбиття вибірки. Тому вимагається, щоб розділяюча гіперплощина максимально далеко відстояла від найближчих до неї точок обох класів.

Нехай *x*- і *x*+ – дві довільні точки класів -1 і +1 відповідно, що лежать на кордоні смуги. Тоді ширина смуги є:

$$\langle (x_+ - x_-), \frac{w}{\|w\|}\rangle = \frac{2}{\|w\|}. \quad (2.2)$$

Ширина смуги максимальна, коли норма вектору *w* мінімальна.

Отже, в разі, коли вибірка лінійно роздільна, потрібно знайти такі значення параметрів *w* і *w_0*, при яких норма вектору w мінімальна.

Побудова оптимальної розділяючої гіперплощини зводиться до мінімізації квадратичної форми при ℓ обмеженнях-нерівностях виду:

$$\langle w_i x_i\rangle - w_0 = \begin{cases} \leq -1, \text{ якщо } y_i = -1, \\ \geq -1, \text{ якщо } y_i = +1, \end{cases} \quad (2.3)$$

щодо n + 1 змінних w, w_0:

$$\begin{cases} \langle w, w_0\rangle \to \min, \\ y_i(\langle w, x_i\rangle - w_0) \geq 1, i = 1, \ldots, l. \end{cases} \quad (2.4)$$

По теоремі Куна-Такера ця задача еквівалентна двоїстій задачі пошуку сідлової точки функції Лагранжа:

$$\begin{cases} \mathrm{L}(\mathrm{w}, \mathrm{w}_0; \lambda) = \frac{1}{2}\langle \mathrm{w}, \mathrm{w}_0 \rangle - \sum_{\mathrm{i}=1}^{\ell} \lambda_\mathrm{i}(\mathrm{y}_\mathrm{i}(\langle \mathrm{w}, \mathrm{x}_\mathrm{i} \rangle - \mathrm{w}_0) - 1) \to \min\limits_{\mathrm{w}, \mathrm{w}_0} \max\limits_{\lambda}, \\ \lambda_\mathrm{i} \geq 0, \mathrm{i} = 1, \ldots \mathrm{l}, \\ \lambda_\mathrm{i} = 0, \text{ або } \langle \mathrm{w}, \mathrm{x}_\mathrm{i} \rangle - \mathrm{w}_0 = \mathrm{y}_\mathrm{i}, \ \mathrm{i} = 1, \ldots \mathrm{l}, \end{cases} \tag{2.5}$$

де λ – вектор двоїстих змінних.

Необхідною умовою сідлової точки є рівність нулю похідних Лагранжіана. З цього виходить, що шуканий вектор ваг w є лінійною комбінацією векторів навчальної вибірки, причому тільки тих, для яких $\lambda \neq 0$. Згідно умові доповнюючої нежорсткості на цих векторах x_i обмеження-нерівності звертаються в рівності: $\langle w, x_i \rangle - w_0 = y_i$, отже, ці вектори знаходяться на кордоні розділяючої смуги та звуться опорними векторами.

Отримуємо еквівалентну задачу квадратичного програмування, що містить тільки двоїсті змінні:

$$\begin{cases} -L(\lambda) = -\sum_{i=1}^{l} \lambda_i + \frac{1}{2}\sum_{i=1}^{l}\sum_{i=1}^{l} \lambda_i \lambda_j y_i y_j, \\ \lambda_i \geq 0, \quad i = 1, \ldots l, \\ \sum_{i=1}^{l} \lambda_i y_i = 0. \end{cases} \tag{2.7}$$

Тут мінімізується квадратичний функціонал, що має невід’ємну певну квадратичну форму, отже, дана задача має єдине рішення.

В результаті алгоритм класифікації може бути записаний в наступному вигляді:

$$a(x) = sign\left(\sum_{i=1}^{l} \lambda_i y_i \langle x_i, x\rangle - w_0\right). \quad (2.8)$$

Щоб узагальнити SVM на випадок лінійної нероздільності, дозволимо алгоритму допускати помилки на навчальних об'єктах, але при цьому будемо намагатися, щоб помилок було менше. Для цього вводиться набір додаткових змінних $\xi_i > 0$, що характеризують величину помилки на об'єктах x_i, $i = 1,..., \ell$. і штраф за сумарну помилку до мінімізуємого функціоналу:

$$\begin{cases} \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{l} \varepsilon_i \to \min\limits_{w, w_0, \varepsilon}, \\ y_i(\langle w, x_i\rangle - w_0) \geq 1 - \varepsilon_i, \quad i = 1, \ldots, l, \\ \varepsilon_i \geq 0, \quad i = 1, \ldots, l. \end{cases} \quad (2.9)$$

В разі $Y = \{-1, +1\}$ відступом об'єкта x_i від кордону класів називається величина: $m_i = y_i(\langle w, x_i\rangle - w_0)$. Алгоритм припускає помилку на об'єкті x_i тоді і тільки тоді, коли відступ m_i від'ємний. Якщо $m_i \in (-1, +1)$, то об'єкт x_i потрапляє всередину розділяючої смуги. Якщо $m_i > 1$, то об'єкт x_i класифікується вірно, і знаходиться на деякому видаленні від розділяючої смуги. Тоді функціонал числа помилок алгоритму a на вибірці X^ℓ:

$$Q(a, X^L) = \sum_{i=1}^{l} [m_i < 0]. \quad (2.10)$$

Крім того, додамо до функціоналу Q штрафний доданок $\tau \| wk \|^2$. Регуляризація часто застосовується для налаштування лінійних моделей класифікації і регресії. При наявності шумових або залежних ознак вона підвищує стійкість алгоритму по відношенню до складу вибірки і його узагальнюючу здатність.

Тоді функціонал якості Q приймає вид:

$$Q(a, X^L) = \sum_{i=1}^{l} (1 - m_i)_+ + \tau \|w\|^2 \to \min_{w, w_0}. \qquad (2.11)$$

Таким чином, принцип оптимальної розділяючої гіперплощини (або максимізації ширини розділяючої смуги) збігається з принципом регуляризації по Тихонову. Позитивна константа *C* (або τ) є керуючим параметром методу і дозволяє знаходити компроміс між максимізацією розділяючої смуги і мінімізацією сумарної помилки. Тоді функція Лагранжа:

$$\mathrm{L}(\mathrm{w}, \mathrm{w}_0; \lambda) = \frac{1}{2}\langle \mathrm{w}, \mathrm{w}_0 \rangle - \sum_{\mathrm{i}=1}^{\ell} \lambda_\mathrm{i}(\mathrm{y}_\mathrm{i}(\langle \mathrm{w}, \mathrm{x}_\mathrm{i} \rangle - \mathrm{w}_0) - 1) - \qquad (2.12)$$

$$- \sum_{i=1}^{\ell} \varepsilon_i(\lambda_\mathrm{i} + \eta_i - C),$$

де $\eta = (\eta_1, \ldots, \eta_\ell)$ – вектор змінних, двоїстих до змінних $\xi = (\xi_1, \ldots, \xi_\ell)$.

Таким чином, вирішення знову зводиться до квадратичного програмування щодо подвійних змінних λ_i. Єдина відмінність від лінійно роздільного випадку полягає в появі обмеження зверху $\lambda_i \leq C$. На практиці для побудови SVM вирішують саме це завдання, так як гарантувати лінійну роздільність вибірки в загальному випадку не представляється можливим.

Константу *C* зазвичай вибирають за критерієм ковзаючого контролю. Це трудомісткий спосіб, так як завдання доводиться вирішувати заново при кожному значенні *C*.

2.1.3 Ядра та випрямлячі простори

Існує ще один підхід до вирішення проблеми лінійної нероздільності – застосування ядер. Це перехід від початкового простору ознакових описів

об’єктів X до нового простору H за допомогою деякого перетворення $\psi: X \rightarrow H$. Якщо простір H має досить високу розмірність, то можна сподіватися, що в ньому вибірка виявиться лінійно нероздільні (легко показати, що якщо вибірка X^ℓ не суперечлива, то завжди знайдеться простір розмірності не більше ℓ, в якому вона буде лінійно роздільна). Простір H називають випрямляючим. Якщо припустити, що ознаковими описами об’єктів є вектори $\psi(x_i)$, а не вектори x_i, то побудова SVM проводиться точно так же, як і раніше

Функція $K: X \times X \rightarrow R$ називається ядром, якщо вона подана в вигляді $K(x, x') = \langle \psi(x), \psi(x') \rangle$ і при деякому відображенні $\psi: X \rightarrow H$, де H – простір зі скалярним добутком. Це означає, що скалярний добуток $\langle x, x'_i \rangle$ можна формально замінити ядром $K(x, x')$. Оскільки ядро в загальному випадку нелінійно, така заміна приводить до істотного розширення безлічі реалізованих алгоритмів $a: X \rightarrow Y$. Функція $K(x, x')$ є ядром тоді і тільки тоді, коли вона симетрична, $K(x, x') = K(x', x)$, і невід’ємно визначена:

$$\int X \int X \, K(x, x') g(x) g(x') dx dx' \geq 0, \qquad (2.13)$$

де для будь-якої g: $X \rightarrow R$.

Існує кілька «стандартних» ядер, які при найближчому розгляді призводять до вже відомими алгоритмами: поліноміальних розділяє поверхонь, двошаровим нейронних мереж, потенційним функціям і іншим.

Найбільш вживані ядра:

– поліноміальне;

– радіально-базисна функція;

– функція Гауса;

– сигмоїда.

Таким чином, ядра э одним з найелегантніших рішень длоя розв’язання проблеми нелінийної класифікації для багатьох алгоритмів, але до сих пір не створений універсальний ефектиний алгоритм для їх підбору.

2.1.4 Математична модель методу найменших квадратів опорних векторів

Формулювання класифікатору SVM було змінено в Сайкенсом та Вандервеллем у наступній формулі LS-SVM:

$$\min J_1(w, \xi) = \frac{1}{2} w^T \mathrm{w} + \gamma \sum_{i=1}^{N} \xi_i, \qquad (2.14)$$

з урахуванням обмежень рівності:

$$y_i[w^T \phi(x_i) + b] = 1 - \mathrm{e}, \quad i = 1, \ldots, N. \qquad (2.15)$$

Це формулювання складається з рівності замість обмежень нерівності та враховує квадратну похибку з терміном регуляризації, подібним до рідж-регресії. Рішення отримують після побудови лагранжіана:

$$L(w, b, e, \alpha) = J(w, b, e) - \sum_{i=1}^{N} \alpha_i y_i [w^T \varphi(x_i) + b] - 1 + e_i, \qquad (2.16)$$

де $\alpha_i \in IR$ – множники Лагранжа.

Вони можуть бути позитивними чи негативними у формулюванні LSSVM. З умов оптимальності отримуємо систему Каруша-Куна-Такера. Треба відмітити, що розрідженість втрачається, що зрозуміло з умови $\alpha_i = \gamma e_i$. Як і в стандартних SVM, ніколи не обчислюється ні w, ні $\phi(x_i)$. Тому усувається w та e поступається:

$$\begin{bmatrix} 0 & y^T \\ y & \Omega + \gamma^{-1} 1 \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 1_v \end{bmatrix}, \qquad (2.17)$$

де $y = [y_1, \ldots, y_N N]$, $1_v = [1, \ldots, 1]$, $e = [e_1, \ldots, e_N], \alpha = [\alpha_1, \ldots, \alpha_N]$.

Умова Мерсера застосовується всередині Ω матриці:

$$\Omega_{ij} = y_i y_j \phi(x_i)^T \phi(x_j) = y_i y_j K(x_i x_j), \tag{2.18}$$

де K – ядерна функція.

Потім класифікатор LS-SVM будується наступним чином:

$$y(x) = sign\left[\sum_{i=1}^{N} \alpha_i \, y_i K(x_i x_j) + b\right]. \tag{2.19}$$

Простий спосіб побудови двійкових класифікаторів з сингмоїдним ядром – це використання рецептури регресії, де для цільового кодування першого та другого класу використовуються цілі ± 1 [35]. Використовуючи обмеження рівності для цілей {-1, +1}, формулювання LS-SVM (2.14), (2.15) неявно відповідає формулюванню регресії з регуляризацією. Дійсно, помноживши помилку e_i на $y_i \in \{-1, +1\}$, умова помилки ED стає:

$$E_D = \frac{1}{2}\sum_{i=1}^{N} e_i^2 = \frac{1}{2}\sum_{i=1}^{N} (y_i e_i)^2 = \frac{1}{2}\sum_{i=1}^{N} (y_i - (w^T \varphi(x_i) + b))^2. \tag{2.20}$$

Враховуючи цю інтерпретацію регресії, умова упередженості у формулюванні LS-SVM дозволяє чітко пов’язати класифікатор LS-SVM з регуляризованим лінійним дискримінантним аналізом Фішера в просторі ознак. Лінійний дискримінант Фішера визначається як лінійна функція з максимальним відношенням розкидом між класами та розкидом всередині класу. Визначивши $N+$ та $N-$ як кількість навчальних даних класів +1 та −1 відповідно, лінійний дискримінант Фішера (у просторі ознак) з регуляризацією отримують шляхом мінімізації

$$\min_{w_F b_F} \frac{1}{2} w_F^T w_F + \frac{\gamma F}{2} \sum_{i=1}^{N} (t_i - w_F^T \varphi(x_i) + b_F)^2, \quad (2.21)$$

з відповідними цілями $t_i = -(N/N_-)$, якщо $y_i = -1$ та $t_i = N/N_+$, якщо $y_i = +1$.

Формулювання регресії (2.21) з цілями Фішера $\{-N/N_-, +N/N_+ +\}$ вибирає член зміщення b_F на основі середнього значення вибірки, тоді як цілі {-1, +1} відповідають асимптотично оптимальному наближенню найменших квадратів дискримінанта Байєса для двох гауссових розподілів. Отже, вирішуючи лінійну множину рівняння (2.17) у подвійному просторі для класифікатора LS-SVM, наприклад, за допомогою широкомасштабного алгоритму також отримують регуляризований дискримінант Фішера [36].

2.2 Опис моделі методу CNN

2.2.1 Архітектура згорткової мережі

Існує дві поширені архітектури нейронних мереж: згорткові нейронні мережі (CNN) і рекурентні нейронні мережі (RNN). CNN використовуються для розпізнавання візуальних шаблонів безпосередньо з піксельних зображень зі змінністю. RNN призначені для розпізнавання шаблонів у часових рядах, які складаються із символів або звукових/мовних сигналів. І CNN, і RNN є особливими типами багатошарових нейронних мереж.

Варто зазначити, що CNN є особливою формою прямої нейронної мережі (FNN), також відомої як багатошаровий персептрон (MLP), навчений із зворотним поширенням. Було доведено, що FNN здатні апроксимувати будь-яку вимірювану функцію з будь-якою бажаною точністю. Коротше кажучи, FNN є універсальними апроксиматорами. Успіх CNN у різних додатках сьогодні є відображенням універсальної здатності FNN до

наближення.Взагалі кажучи, CNN намагаються вивчити зв’язок між входом і виходом і зберігати набутий досвід у своїх вагових фільтрах.

CNN зазвичай має три рівні: згортковий рівень, рівень об’єднання або рівень пулінгу та повністю зв’язаний рівень, які можна побачити на рисунку 2.1.

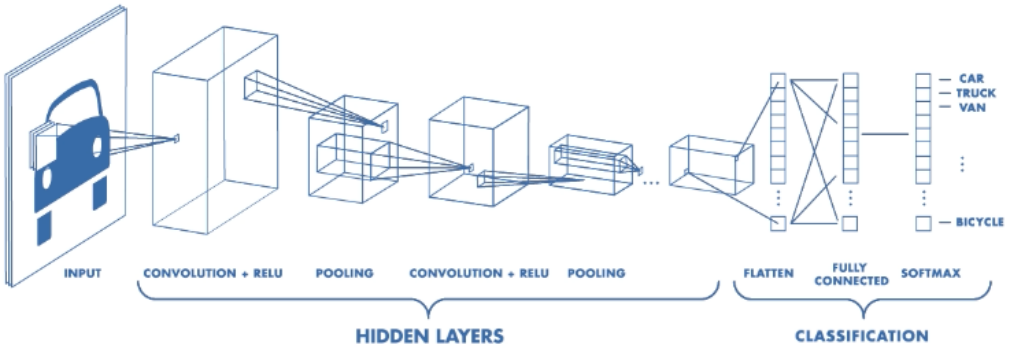

Рисунок 2.1 – Архітектура мережі CNN

Рівень згортки є основним будівельним блоком CNN. Він несе основну частину обчислювального навантаження мережі.

Цей рівень виконує скалярний добуток між двома матрицями, де одна матриця є набором параметрів, які можна дізнатися, інакше відомих як ядро, а інша матриця є обмеженою частиною сприйнятливого поля. Ядро просторово менше, ніж зображення, але більш глибоке. Це означає, що якщо зображення складається з трьох (RGB) каналів, висота і ширина ядра будуть просторово малими, але глибина поширюється на всі три канали.

Під час прямого проходу ядро ковзає по висоті та ширині зображення, створюючи представлення зображення цієї сприйнятливої області. Це створює двовимірне представлення зображення, відоме як карта активації, яка дає відповідь ядра на кожну просторову позицію зображення. Розмір ковзання ядра називається кроком. Якщо є вхід розміром $W \times W \times D$ і кількістю ядер із просторовим розміром F із кроком S і розміром заповнення P, тоді розмір вихідного об’єму можна визначити за такою формулою:

$$W_{out}\frac{W-F+2P}{S}+1. \quad (2.22)$$

Згортка використовує три важливі ідеї, які мотивували дослідників комп'ютерного зору: розріджену взаємодію, спільне використання параметрів та еквіваріантне представлення.

Тривіальні рівні нейронної мережі використовують множення матриці на матрицю параметрів, що описують взаємодію між блоком введення та виведення. Це означає, що кожен вихідний блок взаємодіє з кожним вхідним блоком. Однак згорткові нейронні мережі мають розріджену взаємодію. Це досягається шляхом зменшення розміру ядра, ніж вхід, наприклад, зображення може мати мільйони чи тисячі пікселів, але під час його обробки за допомогою ядра можливо виявити значущу інформацію, яка складається з десятків або сотень пікселів. Це означає, що потрібно зберігати менше параметрів, що не тільки зменшує вимоги до пам'яті моделі, але й покращує статистичну ефективність моделі.

Якщо обчислення однієї функції в просторовій точці (x_1, y_1) є корисним, то воно також має бути корисним у іншій просторовій точці, наприклад (x_2, y_2). Це означає, що для одного двовимірного зрізу, тобто для створення однієї карти активації, нейрони змушені використовувати той самий набір ваг. У традиційній нейронній мережі кожен елемент вагової матриці використовується один раз і ніколи не повертається, тоді як згорточна мережа має спільні параметри, тобто для отримання виходу ваги, застосовані до одного входу, такі ж, як і вага, застосована в іншому місці.

Завдяки спільному використанню параметрів, шари згорткової нейронної мережі матимуть властивість еквіваріантності до трансляції. У ньому сказано, що якщо певним чином змінили вхідні дані, вихідні дані також будуть змінені таким же чином.

Рівень об'єднання замінює вихідні дані мережі в певних місцях шляхом отримання сумарної статистики найближчих виходів. Це допомагає зменшити

просторовий розмір представлення, що зменшує необхідну кількість обчислень і ваги. Операція об’єднання обробляється для кожного фрагмента представлення окремо.

Існує кілька функцій об’єднання, наприклад середнє значення прямокутного околу, норма L2 прямокутного та зважене середнє на основі відстані від центрального пікселя. Однак найпопулярнішим процесом є максимальне об’єднання, яке повідомляє про максимальний результат із сусідства.

Якщо у нас є карта активації розміром $W \times W \times D$, об’єднуюче ядро просторового розміру F і крок S, тоді розмір вихідного об’єму можна визначити за такою формулою:

$$W_{out} \frac{W-F}{S} + 1. \quad (2.23)$$

У всіх випадках об’єднання забезпечує певну інваріантність перекладу, що означає, що об’єкт буде розпізнаним незалежно від того, де він з’являється на кадрі. Нейрони в цьому шарі мають повний зв’язок з усіма нейронами в попередньому та наступному шарі, як це видно у звичайному FCNN. Ось чому його можна обчислити як зазвичай шляхом множення матриці з наступним ефектом зміщення.

2.2.2 Функції активації

Оскільки згортка є лінійною операцією, а зображення далеко не лінійні, шари нелінійності часто розміщують безпосередньо після шару згортки, щоб додати нелінійність до карти активації.

Існує кілька типів нелінійних операцій, найпопулярніші з яких:

– сигмовидна має математичну форму $\sigma(к) = 1/(1+e^{-к})$. Він бере дійсне число та «роздавлює» його в діапазон від 0 до 1. Однак дуже небажаною

властивістю сигмовидної є те, що коли активація відбувається в будь-якому хвості, градієнт стає майже нульовим. Якщо локальний градієнт стає дуже малим, то при зворотному поширенні він фактично «вбиває» градієнт. Крім того, якщо дані, що надходять до нейрона, завжди позитивні, тоді на виході sigmoid будуть або всі позитивні, або всі негативні дані, що призводить до зигзагоподібної динаміки градієнтних оновлень ваги;

– Tanh здавлює дійсне число до діапазону [-1, 1]. Подібно до сигмовидних нейронів, активація насичується, але – на відміну від сигмовидних нейронів – її вихід зосереджений на нулі;

– ReLU стала дуже популярною за останні кілька років. Він обчислює функцію $f(к)=max\ (0,к)$. Іншими словами, активація – це просто нульовий поріг. У порівнянні з sigmoid і tanh, ReLU більш надійний і прискорює конвергенцію в шість разів. На жаль, недоліком є те, що ReLU може бути крихким під час навчання. Великий градієнт, що протікає через нього, може оновити його таким чином, що нейрон більше ніколи не оновлюватиметься. Однак можливо працювати з цим, встановивши відповідну швидкість навчання.

2.2.3 Переваги та особливості методу

CNN дійсно пропонують дуже потужний інструмент для обробки та розуміння зображень. Однак залишається кілька відкритих проблем щодо інтерпретації CNN та ширшого застосування. Деякі з них наведено нижче:

– під час навчання CNN архітектура CNN (включно з номером рівня та номером фільтра на кожному рівні тощо) повинна бути визначена заздалегідь. Враховуючи фіксовану архітектуру, ваги фільтрів оптимізуються за допомогою наскрізної системи оптимізації. Загалом, менші CNN можуть добре впоратися з простими завданнями. Однак досі немає чітких вказівок у проектуванні архітектури CNN для класу програм. Точка зору якоря-вектора

заохочує нас ретельно вивчати властивості вихідних даних. Добре розуміння розподілу вихідних даних сприяє розробці більш ефективної архітектури CNN і більш ефективного навчання;

– проте надійність CNN ставиться під сумнів останніми дослідженнями. Це цікава тема, щоб зрозуміти причини цих проблем, щоб розробити більш стійкі до помилок CNN;

– навчання CNN потребує великої кількості позначених даних. Збирати мічені дані дорого. Крім того, правила маркування можуть відрізнятися для одного набору даних від іншого навіть для тих самих програм. Важливо зменшити навантаження на маркування та дозволити навчання CNN з використанням частково та гнучко маркованих даних. Іншими словами, потрібно перейти від навчання під суворим контролем до навчання зі слабким контролем, щоб зробити CNN широко застосовними;

– ефективне навчання зворотному поширенню є важливим, оскільки сьогодні CNN стають дедалі складнішими. Було запропоновано кілька схем прискорення зворотного поширення. Іншим є введення ретельно підібраного шуму для досягнення швидшої конвергенції. В цьому полі необхідні нові розробки.

3 КОМП'ЮТЕРНЕ МОДЕЛЮВАННЯ МЕТОДІВ КЛАСИФІКАЦІЇ МЕДИЧНИХ ЗОБРАЖЕНЬ

3.1 Обгрунтування вибору програмних засобів моделювання

Python – інтерпретована об'єктно-орієнтована мова програмування високого рівня зі строгою динамічною типізацією. Структури даних високого рівня разом із динамічною семантикою та динамічним зв'язуванням роблять її привабливою для швидкої розробки програм, а також як засіб поєднування наявних компонентів. Python підтримує модулі та пакети модулів, що сприяє модульності та повторному використанню коду. Інтерпретатор Python та стандартні бібліотеки доступні як у скомпільованій, так і у вихідній формі на всіх основних платформах. В мові програмування Python підтримується кілька парадигм програмування, зокрема: об'єктно-орієнтована, процедурна, функціональна та аспектно-орієнтована. Ця мова має наступні переваги:

– наявність сторонніх модулів під різні задачі;

– широкий вибір бібліотек (NumPy для числових обчислень, Pandas для аналізу даних тощо);

– зручні структури даних;

– ідеально підходить для прототипів – забезпечує більше функціональності з меншим кодуванням;

– висока ефективність. Чистий об'єктно-орієнтований дизайн Python забезпечує розширений контроль процесів, а мова оснащена відмінними можливостями обробки тексту та інтеграції, а також власною структурою модульного тестування, що робить її ефективнішою.

Саме за ці властивості більшість спеціалістів що займаються методами інтелектуального аналізу та машинним навчанням віддають перевагу саме Python.

Для проведення практичного дослідження було обране середовище Jupiyter-notebook, інтерфейс якого можна побачити на рисунку 3.1.

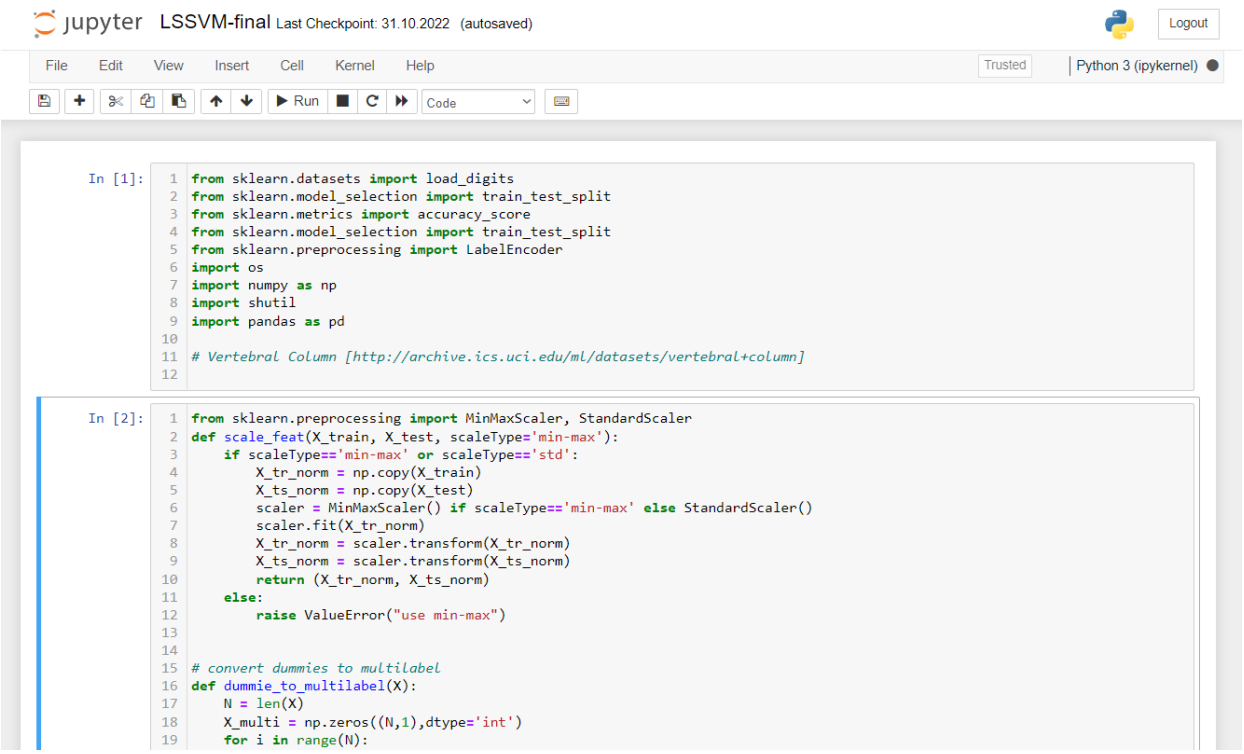

Рисунок 3.1 – Інтерфейс Jupiter Notebook

Jupyter Notebook (раніше IPython Notebook) – це інтерактивне обчислювальне середовище для створення записних документів на базі Інтернету. Jupyter Notebook створено з використанням кількох відкритих бібліотек, зокрема IPython, ZeroMQ, Tornado, jQuery, Bootstrap і MathJax. Документ Jupyter Notebook – це REPL на основі браузера, що містить упорядкований список клітинок вводу/виводу, які можуть містити код, текст (з використанням Markdown), математику, графіки та мультимедіа. Під інтерфейсом блокнот – це документ JSON, що відповідає схемі з версіями, зазвичай закінчується розширенням «.ipynb».

Основними частинами блокнотів Jupyter є: метадані, формат блокнота та список клітинок. Метадані – це словник даних визначень для налаштування та відображення блокнота. Формат блокнота – це номер версії програмного

забезпечення. Список клітинок – це різні типи клітинок для Markdown (відображення), коду (для виконання) і виведення клітинок типу коду.

В даній дослідницькій роботі були використані такі програмні пакета та модулі Python:

– pandas – бібліотека програмного забезпечення, написана на мові програмування Python для обробки та аналізу даних. Зокрема, вона пропонує структури даних і операції для роботи з числовими таблицями та часовими рядами. Pandas дозволяє імпортувати дані з різних форматів файлів, таких як значення, розділені комами, JSON, Parquet, таблиці бази даних SQL або запити, а Microsoft Excel. Pandas дозволяє виконувати різні операції обробки даних, такі як об'єднання, зміна форми, вибір, а також очищення даних і особливості обробки даних. Розробка pandas ввела в Python багато порівнянних функцій роботи з DataFrames, які були встановлені в мові програмування R. Бібліотека pandas побудована на основі іншої бібліотеки NumPy, яка орієнтована на ефективну роботу з масивами замість функцій роботи з DataFrames;

– numpy – це бібліотека для мови програмування Python, яка підтримує великі багатовимірні масиви та матриці, а також велику колекцію математичних функцій високого рівня для роботи з цими масивами. Математичні алгоритми, реалізовані на інтерпретованих мовах (наприклад, Python), часто працюють набагато повільніше тих же алгоритмів, реалізованих на компілюваних мовах (наприклад, Фортран, C, Java). Бібліотека NumPy забезпечує реалізацію вичислювальних алгоритмів (у вигляді функцій і операторів), оптимізованих для роботи з багатомірними масивами. У результаті будь-який алгоритм, який може бути виражений у вигляді послідовності операцій над масивами (матрицями) і реалізований за допомогою NumPy, працює так само швидко, як і аналогічний код, що виконується в MATLAB;

– scipy – це безкоштовна бібліотека Python із відкритим кодом, яка використовується для наукових і технічних обчислень. SciPy містить модулі

для оптимізації, лінійної алгебри, інтеграції, інтерполяції, спеціальних функцій, обробки сигналів і зображень, розв'язування звичайних рівнянь та інших поширених у науці та інженерії завдань. Базовою структурою даних, яку використовує SciPy, є багатовимірний масив, наданий модулем NumPy;

– scikit-learn (також відомий як sklearn) – це безкоштовна бібліотека машинного навчання для мови програмування Python. Він містить різні алгоритми класифікації, регресії та кластеризації, включаючи машини опорних векторів, випадкові ліси, посилення градієнта, *k*-середні та DBSCAN, і розроблений для взаємодії з числовими та науковими бібліотеками Python. Scikit-learn здебільшого написаний на Python, і широко використовує NumPy для високопродуктивної лінійної алгебри та операцій з масивами. Крім того, деякі основні алгоритми для підвищення продуктивності були написані на Cython. Машини опорних векторів реалізовані обгорткою Cython навколо LIBSVM; логістичної регресії та лінійних опорних векторних машин за допомогою подібної оболонки навколо LIBLINEAR;

– matplotlib – бібліотека мови програмування Python для візуалізації даних двомірної та трійчастої графіки. Виявлені зображення можуть бути використані в якості ілюстрацій у публікаціях. Matplotlib є гнучким, легко конфігурованим пакетом, який разом із NumPy, SciPy та IPython надає можливості, подібні MATLAB. Зараз пакет працює з кількома графічними бібліотеками, включаючи wxWindows і PyGTK. Пакет підтримує багато видів графіків і діаграм, наприклад лінійні графіки, діаграми розсіювання, столбчаті діаграми, гістограми, кругові діаграми;

– TensorFlow – це платформа з відкритим кодом для машинного навчання. Він має комплексну, гнучку екосистему інструментів, бібліотек і ресурсів спільноти, яка дозволяє дослідникам просувати найсучасніші технології машинного навчання, а розробникам – легко створювати й розгортати програми на основі ML. Спочатку TensorFlow був розроблений дослідниками та інженерами, які працюють у команді Google Brain в організації Google Machine Intelligence Research для проведення машинного

навчання та глибоких досліджень нейронних мереж. Система є достатньо загальною, щоб її також можна було застосовувати в багатьох інших областях. TensorFlow надає бібліотеку готових алгоритмів чисельних обчислень, реалізованих через графи потоків даних (data flow graphs). Вузли в таких графах реалізують математичні операції або точки входу/виводу, в той час як ребра графу представляють багатовимірні масиви даних (тензори), які перетікають між вузлами. Вузли можуть бути закріплені за обчислювальними пристроями і виконуватися асинхронно, паралельно обробляючи разом все підходящі до них тензори, що дозволяє організувати одночасну роботу вузлів в нейронної мережі за аналогією з одночасною активацією нейронів в мозку;

– Keras – це бібліотека програмного забезпечення з відкритим кодом, яка надає інтерфейс Python для штучних нейронних мереж. Keras діє як інтерфейс для бібліотеки TensorFlow. Keras містить численні реалізації часто використовуваних будівельних блоків нейронної мережі, таких як шари, цілі, функції активації, оптимізатори та безліч інструментів, які полегшують роботу з зображеннями та текстовими даними, щоб спростити кодування, необхідне для написання глибокого коду нейронної мережі. Окрім стандартних нейронних мереж, Keras підтримує згорточні та рекурентні нейронні мережі. Він підтримує інші поширені службові рівні, як-от відключення, пакетна нормалізація та об’єднання. Це також дозволяє використовувати розподілене навчання моделей глибокого навчання на кластерах графічних процесорів (GPU) і тензорних процесорів (TPU).

3.2 Опис обраного датасету

3.2.1 Характеристики датасету

Рак молочної залози є поширеним раком у жінок і однією з основних причин смерті серед жінок у всьому світі. Інвазивна протокова карцинома (IDC) є найпоширенішим типом раку молочної залози з приблизно 80% усіх

діагностованих випадків. Рання точна діагностика відіграє важливу роль у виборі правильного плану лікування та покращенні виживаності пацієнтів. В останні роки були зроблені зусилля, щоб передбачити та виявити всі типи раку за допомогою штучного інтелекту.

Відповідний набір даних є першим важливим кроком для досягнення такої мети. Гістопатологічний аналіз тканин, який проводить патологоанатом, відіграє важливу роль у діагностиці та прогнозі багатьох типів раку, наприклад молочної залози.

У цьому дослідженні було використано набір даних із 162 гістопатологічних зображень раку молочної залози, а саме набір даних гістопатологічної анотації та діагностики раку молочної залози (BreCaHAD), який дозволяє дослідникам оптимізувати та оцінити корисність запропонованих ними методів. Набір даних включає різні злоякісні випадки.

Зображення були отримані з архівних випадків хірургічної патології, які були заархівовані для навчальних цілей.

Ноттінгемська система класифікації – це міжнародна система класифікації раку молочної залози, рекомендована Всесвітньою організацією охорони здоров’я, де оцінка трьох морфологічних ознак (утворення канальців, ядерний плеоморфізм і мітотична кількість) використовується для підрахунку балів для визначення остаточного ступеня раку.

Щоб отримати ці ознаки, патологоанатом анотує або позначає гістологічні зображення, пофарбовані H&E, як мітоз, апоптоз, ядра пухлини, непухлинні ядра, канальці та нетубули. Зразки випадків зібрані з різних сценаріїв, починаючи від гістологічних структур із чіткими межами до слабодиференційованих структур із відсутністю типових ознак.

IDC є найпоширенішим підтипом усіх видів раку молочної залози. Щоб призначити ступінь агресивності зразка, патологоанатоми зазвичай зосереджуються на регіонах, які містять IDC. Як наслідок, одним із поширених етапів попередньої обробки для автоматичної оцінки агресивності є окреслення точних областей IDC всередині цілого слайда.

З оригінального набору даних було виділено 277 524 зображення розміром 50×50. Датасет складається з 198 738 IDC негативних і 78 786 IDC позитивних зразків, всього 277524 зображення. Ім’я файлу кожного зразка має такий формат uxXyYclassC.png, де:

– u – ідентифікатор пацієнта (10253idx5);

– X – координата *x*, звідки було обрізано цей патч;

– Y – *y*-координата, звідки було обрізано цей патч;

– C вказує на клас, де 0 – негативний IDC, а 1 – позитивний IDC.

Один з прикладів назви зразка –10253idx5x1351y1101class0.png.

3.2.2 Дослідницький аналіз датасету

Перед тим як перейти до безпосередньо до класифікації даних, було проведений дослідницький аналіз датасету, щоб дізнатися характеристики даних та зробити попередні висновки про них. По-перше, була перевірена кількість патчів на одного пацієнта, та розподілення цієї кількості. Перший з графіків наведений на рисунку 3.2.

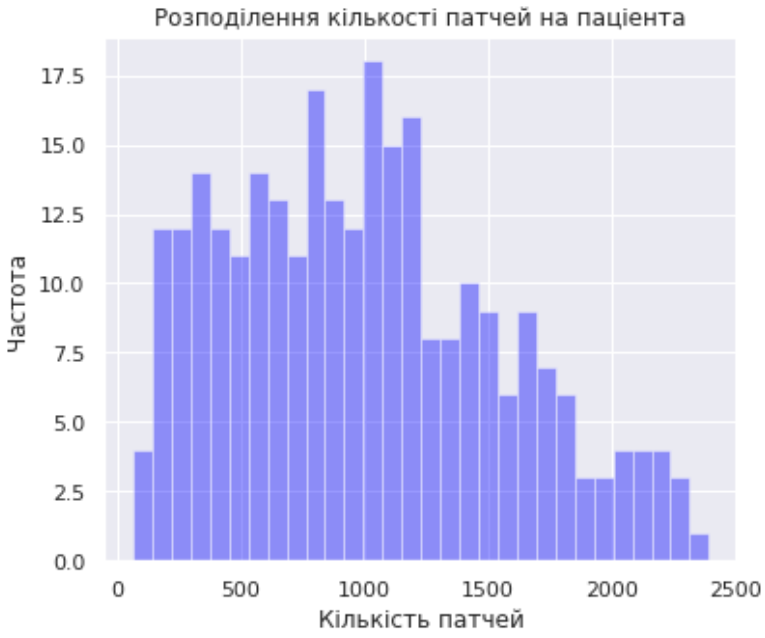

Рисунок 3.2 – Розподілення кількості патчей на пацієнта

З графіку розподілення патчей можна побачити, що кількість зображень на пацієнта досить сильно різниться. З цього можна зробити висновок, що зображення для різних пацієнтів можуть мати різну роздільну здатність.

Наступним було дослідження того, скільки відсотків зображення займають пухлини. Графік можна побачити нижче на рисунку 3.3.

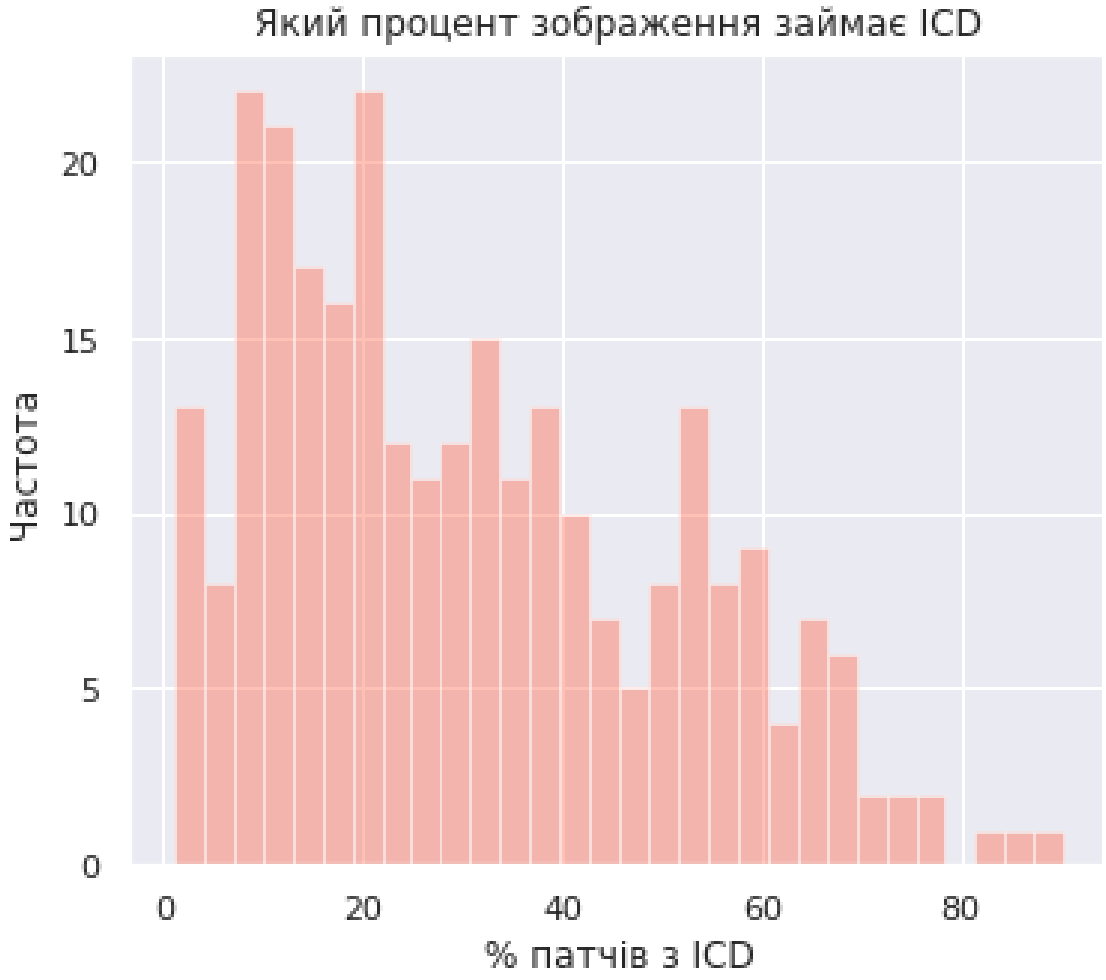

Рисунок 3.3 – Процентаж патчів з ICD

Деякі пацієнти мають понад 80% плям, які показують IDC. Отже, тканина сповнена раку або лише частина грудей була покрита шматком тканини, який зосереджений на раку IDC. Можна поставити питання, чи охоплює зріз тканини пацієнта всю область інтересу.

Також був зроблений графік розподілення класів даних, його можна побачити нижче на рисунку 3.4.

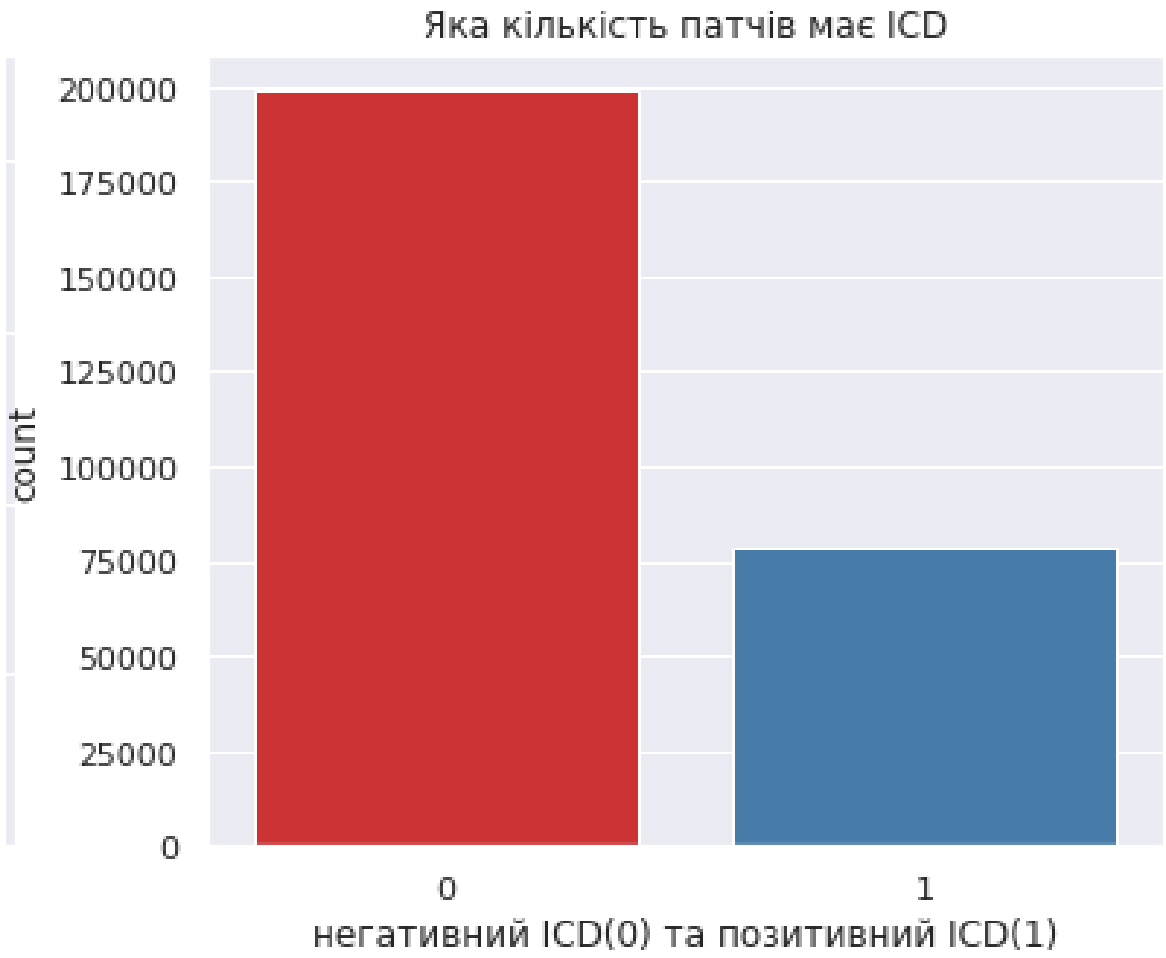

Рисунок 3.4 – Розподілення класів

Зображення показує що, класи з присутністю IDC і відсутністю IDC незбалансовані. Тому можливо перед застосуванням методів інтелектуального аналізу потрібно буде провести оверсемплінг даних заради їх балансування.

Також можна подивитися на вибірки здорових патчів та патчів, що мають рак, та проаналізувавши візуально, відмітити деякі ознаки.

Набори зображень зі здорових та уражених вибірок наведені нижче на рисунках 3.5 та 3.6.

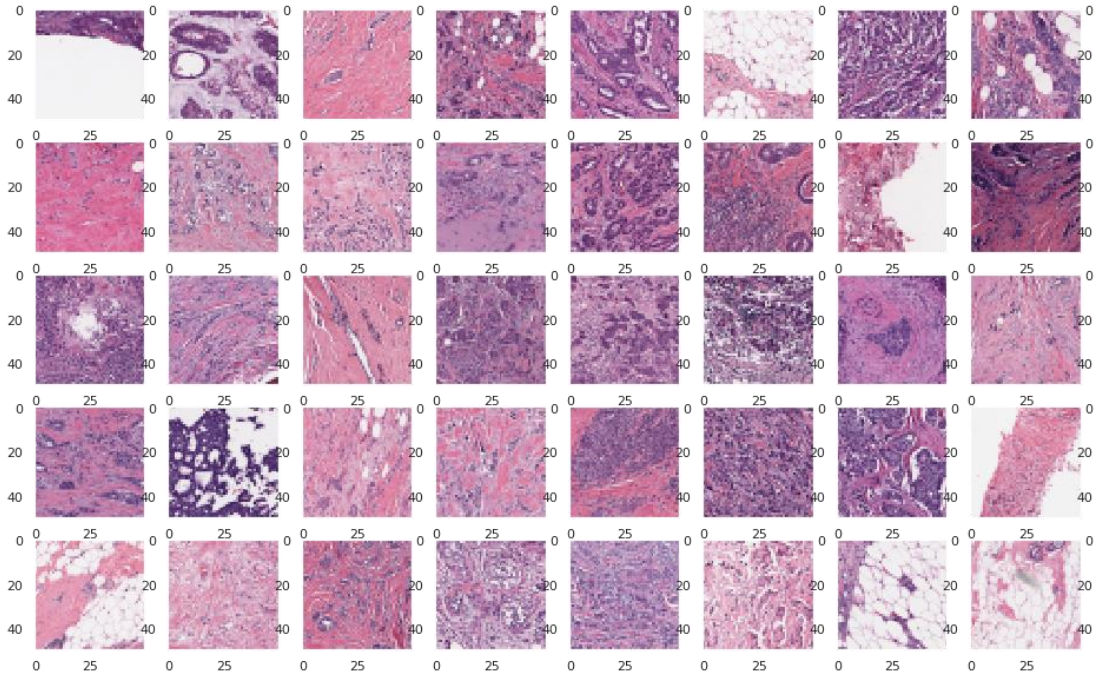

Рисунок 3.5 – Зображення уражених патчів

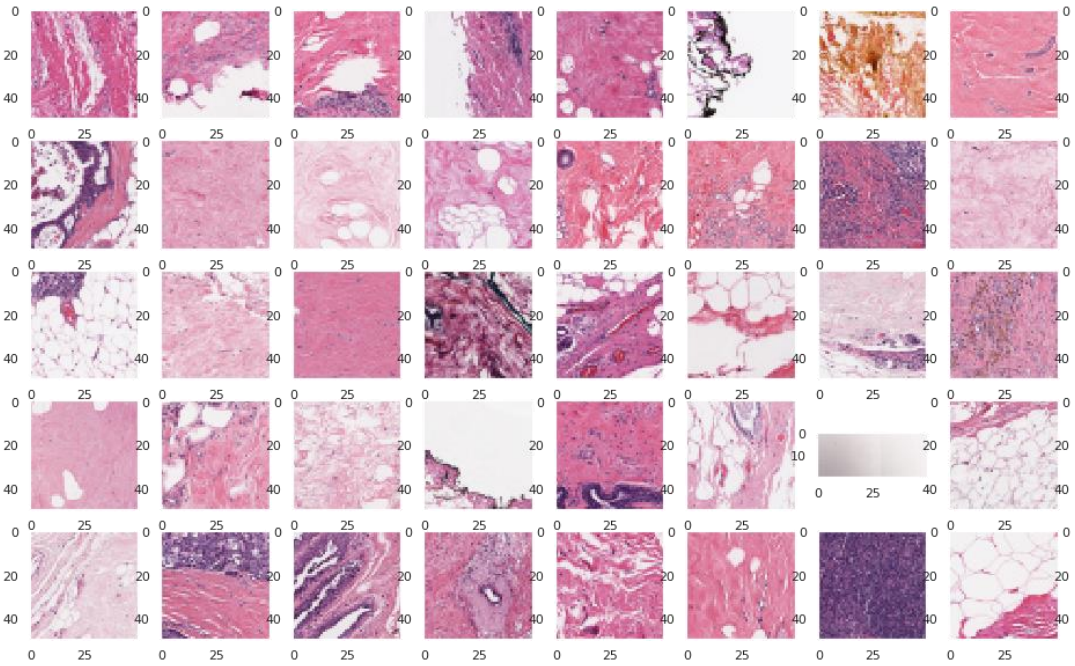

Рисунок 3.6 – Зображення здорових патчів

Зображення показують, що іноді можна знайти неповні патчі розміром менше 50×50 пікселів. По візуальним характеристикам патчі з раком виглядають більш фіолетовими і скупченими, ніж здорові. Це може бути типово для раку та характерно для клітин і тканин протоків. Як підтвердження

другої причини, можна зазначити, що деякі здорові патчі теж мають інтенсивний фіолетовий колір.

Наступним етапом дослідницького аналізу було бінарне моделювання пухлин. Для цього були витягнуті координати пухлин з назв патчів для того, щоб реконструювати повну пухлину (рис. 3.7).

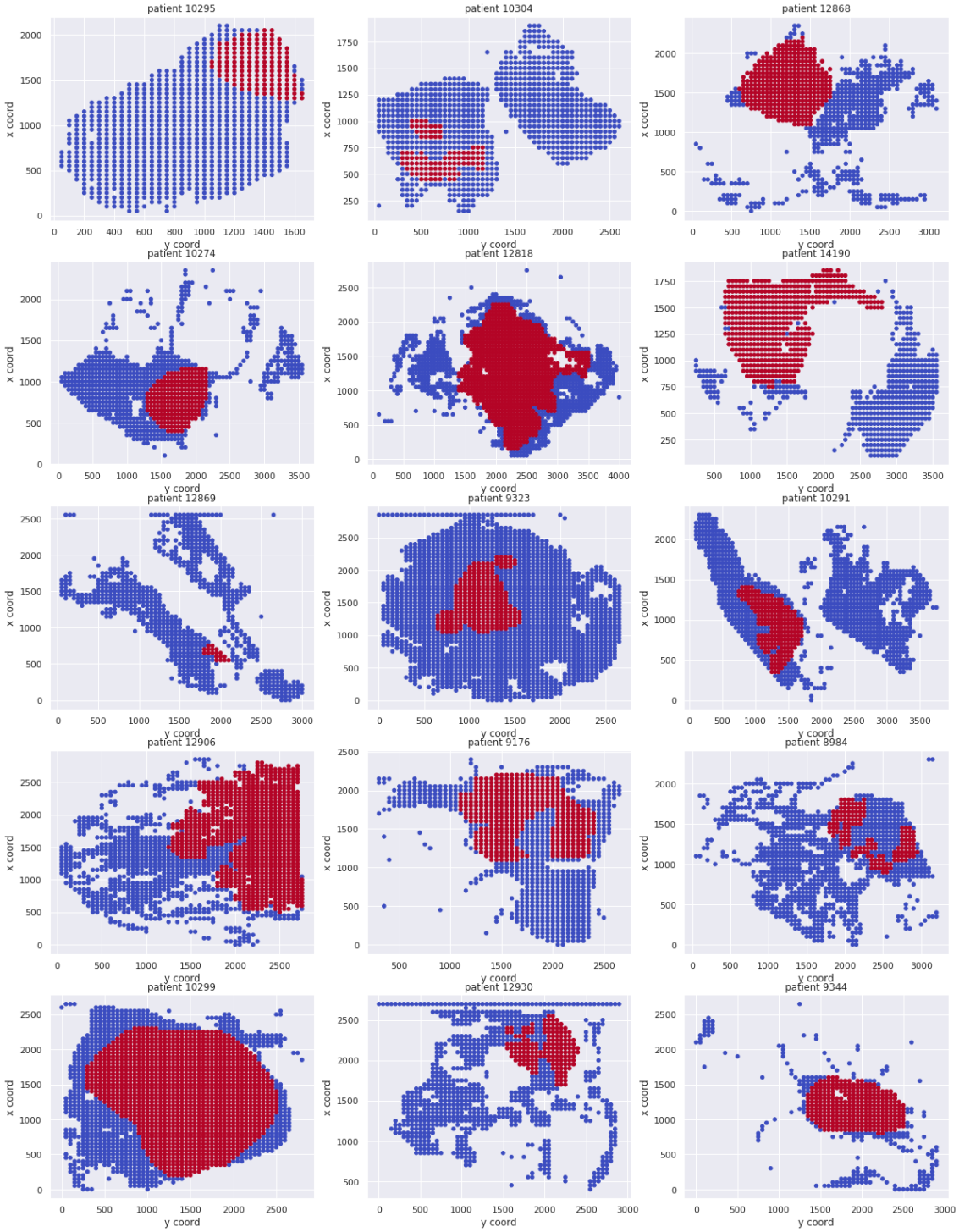

Рисунок 3.7 – Реконструйовані зображення пухлин

Поглибити аналіз можна, наклавши отримані зображення пухлин на початкові зображення. Рисунок 3.8, наведений нижче, – приклад початкового цілого зображення.

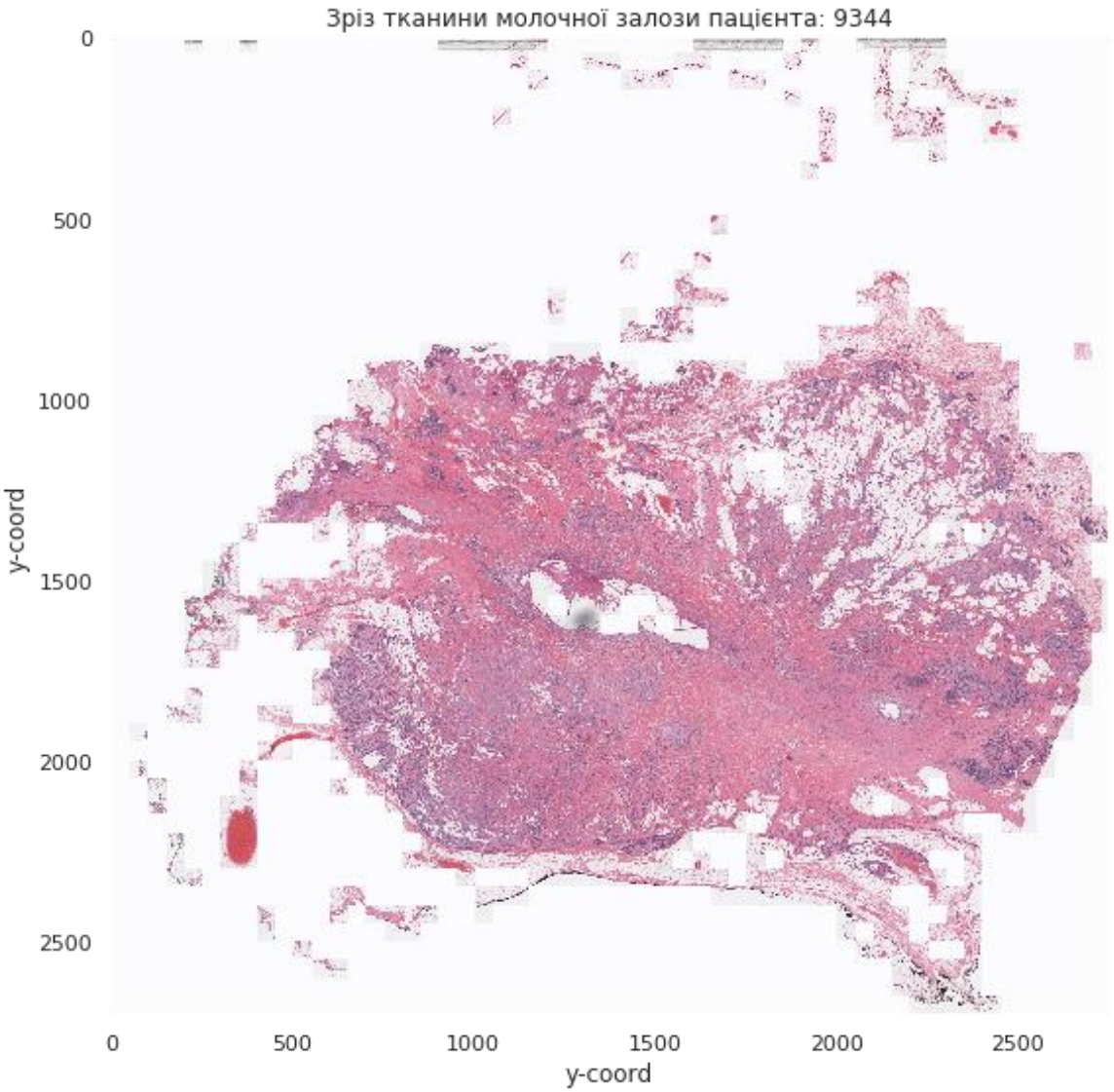

Рисунок 3.8 – Початкове зображення

Далі приведено зображення від цього ж пацієнта з виділеною реконструюванню пухлиною, що можна побачити на рисунку 3.9.

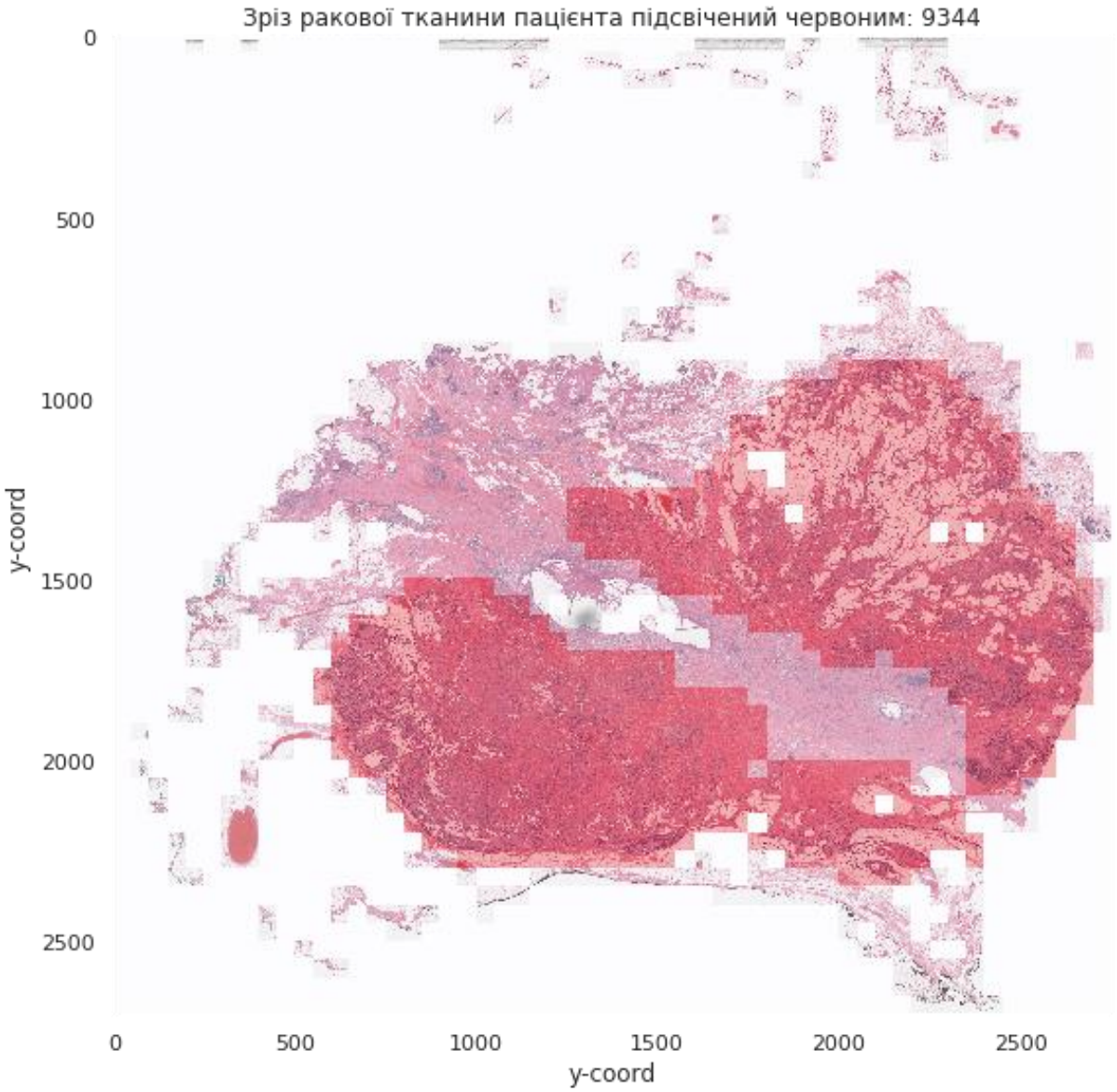

Рисунок 3.9 – Зображення з реконструюванню пухлиною

Порівнявши зображення, можна відмітити, що більш темні фіолетові частини мають більшу вірогідність бути раковою тканиною, ніж більш світлі рожеві. Але як відмічено вище, це не однозначна кореляція та більш темний колір може бути обумовлений наявністю в цьому місці протоків.

3.3 Хід проведення дослідження за допомогою комп’ютерних моделей

3.3.1 Хід проведення дослідження за допомогою методу LSSVM

Першим етапом застосування комп`ютерної моделі стала підготовка даних та препроцесинг зображень. Для цього методу був використаний не цілий датасет, а його частина. В ідеалі експеримент потрібно проводити на цілому датасеті, але для покращення часу роботи та якості класифікації, була відібрана частина зображень. Перевага надавалася пацієнтам з більшою кількістю зображень та в результаті було використано 22259 зображень. Хід проведення дослідження відбувався таким чином:

– початком роботи став збір всіх метаданних по кожному зображенню в датафрейм з такими даними, як шлях до зображення, id пацієнта, *x* та *y* координати початку патчу та позначення класу;

– наступний крок – переведення зображення в чорно-білий режим та перетворення в числовий масив numpy. Після зменшення розмірності зображень їх розмір становить 2500×1;

– для того, щоб збільшити якість роботи моделі, були видалені пошкоджені зображення. В ході видалення кількість робочих зображень зменшилася з 22259 до 22155;

– подальшою обробкою вхідних даних стало застосування методу головних компонент. На попередньому кроці вдалося зменшити загальну довжину наших масивів зображень із 7500 до 2500, перетворивши зображення на градації сірого. Тепер ціллю є ще більше зменшити кількість вимірів, застосувавши PCA до масивів зображень. Функція PCA приймає відсоток від 0 до 1, наприклад, n_components = 0,8, означитиме повернення власних векторів, на які припадає 80% відсотків відхилення від у в масивах зображень. Після проведеного аналізу можемо визначити, що на перші 150 компонентів припадає понад 80% дисперсії в даних. Функція PCA також приймає будь-яку кількість абсолютних компонентів, наприклад: маючи 2500 початкових

компонентів, щоб отримати потрібні перші 150 компонентів, потрібно встановити n_components = 150;

– при розділенні даних на тренувальну та тестову вибірки, потрібно врахувати декілька нюансів. Перше – це співвідношення ракових і не ракових зображень, які наявні в датасеті. Як було визначено раніше, дані не розділені рівномірно. В вибірці присутньо набагато більше неракових зображень, ніж зображень з раком, і це зміщення може вплинути на фінальну модель. Це також може створити проблему під час оцінювання моделі. Давайте візьмемо приклад, де у нас є набір даних, що складається зі 100 точок, 90 з них є раковими, а 10 з них не є раковими, і модель класифікувала весь набір даних як раковий, це дасть точність 90%, однак це не відображає того, що відбувається. У даному випадку було зменшено вибірку неракових даних, щоб зробити цільовий розподіл більш збалансованим. В ідеалі потрібно було б штучно збільшили вибірку даних;

– останнім етапом підготовки даних є розбиття на тренувальну та тестову вибірки. Дані були розбиті в співвідношенні 70 відсотків для тренування та 30 для тесту за допомогою sklearn train_test_split.

Після попередньої обробки даних було проведено тренування моделі. Як описано у попередніх розділах, особливість методу найменших квадратів опорних векторів в тому, що вони можуть мати різні ядра для роботи з нелінійними даними. В ході цієї роботи було протестовано три ядра: лінійне, радиально-базисне, та поліноміальне. Детальні параметри кожної моделі можна побачити в таблиці 3.1 нижче.

Таблиця 3.1 – Параметри методу

Ядро	**Гамма**	**Ступінь полінома/Сигма**
Лінійне	1	-
Поліноміальне	1	3 (ступінь полінома)
Радиально-базисне	0,1	0,5 (сигма)

Гамма-параметр визначає, наскільки далеко сягає вплив окремого навчального прикладу, при цьому низькі значення означають «далеко», а високі значення означають «близько». Гамма-параметри можна розглядати як величину, зворотну радіусу впливу зразків, обраних моделлю як опорний вектор.

Сигма, як зазвичай визначається в розподілі Гауса, є стандартним відхиленням. Вона визначає ширину розподілу Гауса для радиально-базисної функції в складі ядра методу опорних векторів.

3.3.2 Хід проведення дослідження за допомогою методу CNN

На початку роботи з даними для їх підготовки до тренування в згортковій нейронній мережі були проведені ті ж самі дії по збору даних в структуру датафрейму та виділення координат, ідентифікаційного номеру пацієнта та діагнозу. На наступному етапі був знову проведений undersampling шляхом зменшення кількості не ракових зображень. Після цього деякі зображення обох класів були горизонтально та вертикально перевернуті, для того щоб підвищити різноманітність даних. Враховуючи те, що патчі можуть бути перевернуті при початковому зборі чи обробці, перевертання при тренуванні допоможуть моделі вирізняти ракові тканини незалежно до положення патчу.

Основним етапом проведення дослідження було налаштування та тренування моделі. На вхід подаємо патчі розмірністю 50×50×3. Після цього була сформована модель. Її склад та параметри можна побачити в таблиці нижче, де Conv2D – шар згортки, MaxPooling2 – шар субдескретизації, Dropout – шар для запобігання оверфіттінгу, Flatten – шар для зменшення розмірності, та Dense – повнозв’язний шар (табл. 3.2).

Таблиця 3.2 – Шари згорткової мережі

Шар мережі	**Розмірність вихідних значень**
conv2d_2 (Conv2D)	(22, 22, 32)
conv2d_3 (Conv2D)	(19, 19, 32)
max_pooling2d_2 (MaxPooling2)	(9, 9, 32)
dropout (Dropout)	(9, 9, 32)
flatten_2 (Flatten)	(2592)
dense_4 (Dense)	(256)
dense_5 (Dense)	(2)

Базова обрана кількість епох навчання дорівнює 60, але при навчанні моделі був застосований параметр EarlyStop, який зупиняє модель, коли різниця в якості між двома епохами стане незначною.

3.4 Аналіз отриманих результатів

Після застосування методів найменших квадратів опорних векторів та згорткової нейромережі на обраному датасеті були отримані такі результати (табл. 3.3).

Таблиця 3.3 – Отримані результати

Метод	**Отримана точність**
CNN	0,8676
LSSVM (linear)	0,6692
LSSVM (poly)	0,7397
LSSVM (rbf)	0,7658

Також для згоркової мережі можна навести графік того, як змінювалася точність в ході навчання (рис. 3.10).

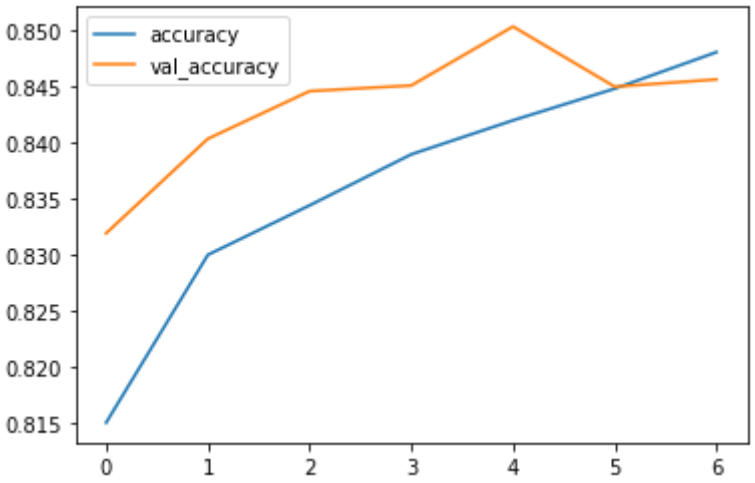

Рисунок 3.10 – Графіки точності моделі на тестовому та валідаційному датасеті

Дивлячись на наведені вище результати, можна побачити, що для цієї задачі нейронна згорткова мережа видає кращий результат, ніж метод найменших квадратів опорних векторів. Також суттєва різниця є між різними ядрами методу LSSVM – різниця в точності між методом з лінійним ядром та методом з радіально-базисним ядром майже дорівнює різниці між методом з радіально-базисним ядром та згортковою мережею.

Отримані результати підтверджують той факт, що згорткові мережі в цілому краще справляються з роботою з зображеннями завдяки їх особливій архітектурі. Але навіть при найкращому результаті з отриманих в ході експериментального дослідження, потрібно покращити якість роботи за допомогою більш глибокої підготовки даних, наприклад зменшення шуму чи оверсемплінгу, чи більш точного налаштування моделі.

Для методу найменших квадратів опорних векторів можливе покращення якості могла б дати поглиблена обробка даних та кастомізовані ядра, а також більш точний підбір параметрів ядер. Очевидно, більш перспективними ядрами є поліноміальне та радиально-базисне. Також

необхідно відзначити, що LSSVM може краще проявити себе на менших датасетах.

3.5 Подальші перспективи дослідження

В ході даного дослідження було порівняно два сильні методи для класифікації зображень – метод найменших квадратів опорних векторів та згорткову нейронну мережу. З отриманих результатів можна зробити висновок, що на даному етапі згорткові мережі справляються з класифікацією зображень трохи краще. Це відкриває широкі можливості для застосування методу в сфері медичного аналізу.

Але не всі медічні дані є зображеннями – в цьому полі є дуже багато реальних датасетів з числовими або категоріальними даними. І з такими даними метод найменших квадратів опорних векторів буде справлятися краще, тому що згорткову нейронну мережу важко адаптувати від дані, що не складаються з зображень.

Хоча в цьому полі теж проводяться деякі дослідження – науковці з японського Інституту фізико-хімічних досліджень (RIKEN) представили алгоритм з назвою Deepinsight [37], який дозволяє виділяти ознаки с числових даних, та конвертувати їх в зображення. Такі зображення можна використовувати при роботі зі згортковими мережами. Тому наступним шагом дослідження може стати порівняння роботи методу найменших квадратів опорних векторів на неграфічному датасеті та роботи згорткової мережі на тому ж датасеті, але в вигляді згенерованих зображень. Окрім кращої роботи на неграфічних даних, метод LSSVM показує кращу якість класифікації на менших датасетах.

Також треба відмітити можливість більш глибоких досліджень по цій темі з в площині модифікації базових методів під конкретні датасети та задачі, наприклад застосування кастомного ядра для методу найменших квадратів

опорних векторів або модифікації архітектури згорткової мережі. Можливо, поліпшити результати застосування обраних методів інтелектуального аналізу могла б попередня сегментація зображень.

Для обраного датасету та подібним йому також перспективним напрямком розвитку досліджень може стати застосування спеціальних претренованих нейроних мереж, наприклад таких як VGG-16, ResNet50, EfficientNet та подібних. Такі мережі мають переваги в простоті використання та якості розпізнавання завдяки попередньому тренуванню саме на різноманітних видах та категоріях зображень.

ВИСНОВКИ

У рамках кваліфікаційної роботи були досліджені методи інтелектуального аналізу даних в сфері медичної діагностики. Обрані методи інтелектуального аналізу – метод найменших квадратів опорних векторів та згорткова нейронна мережа були протестовані на датасеті з реальних медичних даних. Для проведення практичного дослідження був проведений аналіз предметної області, а саме застосування методів інтелектуального аналізу в діагностиці в цілому і методів розпізнавання зображень в окремих випадках. Потім були проаналізовані методи найменших квадратів опорних векторів та згорткова нейронна мережа, їх математичний апарат та принципи роботи. Після проведеної підготовчої роботи методи були апробовані на наборі даних з зображеннями тканин з раком молочної залози. З порівнянням отриманих результатів можна зробити висновки, що згорткова мережа більше підходить для діагностики з використанням медичних зображень.

Елементами новизни в кваліфікаційній роботі є застосування методу найменших квадратів опорних векторів для класифікації медичних зображень та порівняння цього методу зі згортковою нейронною мережею, яка традиційно вважається хорошим методом для цієї задачі.

Методи інтелектуального аналізу проявили себе перспективним інструментом для використання в сфері медичної діагностики, тому отримані результати можна використовувати в подальших дослідженнях.

Результати дослідження апробовано у вигляді тез доповідей під час XXXVII Міжнародної науково-практичної конференції «Modern ways of solving the latest problems in science» [38].

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Brown, D. E. (2008). Introduction to data mining for medical informatics. *Clinics in laboratory medicine*, *28*(1), 9-35.

2. Гороховатський, В. О., & Творошенко, І. С. (2021). Методи інтелектуального аналізу та оброблення даних: навч. посібник.

3. Руденко Д.О., Танянський О.С. (2021). Принципи передобробки даних для машинного навчання.

4. Zeleniy, O., Rudenko, D., Lyubchenko, V., & Lyashenko, V. (2022). Image Processing as an Analysis Tool in Medical Research.

5. Shafronenko, A., Bodyanskiy, Y., Pliss, I., & Popov, S. (2020, September). Evolving neo-fuzzy system for distorted data online processing. In 2020 10th International Conference on Advanced Computer Information Technologies (ACIT) (pp. 352-355). IEEE.

6. Shafronenko, A. Y., & Rudenko, D. A. (2020). ONLINE RECURRENT METHOD OF CREDIBILISTIC FUZZY CLUSTERING. BBK 91, 37.

7. A Shafronenko, Y Bodyanskiy, D Rudenko (2020) Neuro-fuzzy clustering of Distorted Data Using Cat Swarm Optimization

8. Tvoroshenko, I., & Mahomet, A. (2021). About classification of the methods in design of medical information systems.

9. Tvoroshenko I., and Gorokhovatskyi V. (2022) The Application of Hybrid Intelligence Systems for Dynamic Data Analysis, *International Journal of Engineering and Information Systems*, 6(2), pp. 40–48.

10. 7 Ways Data Science Is Reshaping Healthcare. URL: https://www.altexsoft.com/blog/datascience/7-ways-data-science-is reshaping-healthcare/ (дата звертання 20.09.2022).

11. Survey: 8 in 10 Hospital Leaders Say Predictive Analytics is Important to Future Yet Only One-Third Use It. URL: https://chimecentral.org/survey-8-10-

hospital-leaders-say-predictive-analytics-important-future-yet-one-third-use/ (дата звертання 20.09.2022)

12. Путятін, Є. П., Гороховатський, В. О., & Матат, О. О. (2006). Методи та алгоритми комп'ютерного зору: навч. посібник.

13. D. Rudenko, O.Serhiienko, O. Zeleniy, V. Lyashenko (2022) Model for Predictive Analysis of International Trade Based on the Dynamics of Stock Indices

14. Gorokhovatsky V. (2014) Structural Analysis and Intellectual Data Processing in Computer Vision. SMIT: Kharkiv, Ukraine, 316 p.

15. Daradkeh Y.I., Gorokhovatskyi V., Tvoroshenko I., and Zeghid M. (2022) Cluster representation of the structural description of images for effective classification, *Computers, Materials & Continua*, 73(3), pp. 6069–6084.

16. Гороховатський В.О., Руденко Д.О., Сірик Т.О. (2019) Дослідження системи ієрархічних ознак при блочному поданні опису у складі множини ключових точок зображення.

17. Гороховатский В.А. (2003) Распознавание изображений в условиях неполной информации. Харків: ХНУРЭ, 112 с.

18. Camlica, Z., Tizhoosh, H. R., & Khalvati, F. (2015, December). Medical image classification via SVM using LBP features from saliency-based folded data. In *2015 IEEE 14th international conference on machine learning and applications (ICMLA)* (pp. 128-132). IEEE.

19. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286(5439), 531-537.

20. Vapnik, V., & Mukherjee, S. (1999). Support vector method for multivariate density estimation. Advances in neural information processing systems, 12.

21. Comak, E., Polat, K., Güneş, S., & Arslan, A. (2007). A new medical decision making system: least square support vector machine (LSSVM) with fuzzy weighting pre-processing. *Expert Systems with Applications*, *32*(2), 409-414.

22. Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital signal processing*, *17*(4), 694-701.

23. Liu, C., Zhou, B., Li, Q., Chen, Y., Qin, G., & Hu, G. (2018). Breast Tumor Classification Diagnosis Based on LS-SVM.

24. Hu, Y., Hase, T., Li, H. P., Prabhakar, S., Kitano, H., Ng, S. K., ... & Wee, L. J. K. (2016). A machine learning approach for the identification of key markers involved in brain development from single-cell transcriptomic data. BMC genomics, 17(13), 19-29.

25. Jagan, J., Dalkiliç, Y., & Samui, P. (2016). Utilization of SVM, LSSVM and GP for predicting the medical waste generation. In *Smart cities as a solution for reducing urban waste and pollution* (pp. 224-251). IGI Global.

26. Ford, W., & Land, W. (2014). A latent space support vector machine (LSSVM) model for cancer prognosis. *Procedia Computer Science*, *36*, 470-475.

27. Jusman, Y., Ng, S. C., & Abu Osman, N. A. (2014). Intelligent screening systems for cervical cancer. *The Scientific World Journal*, *2014*.

28. Mavroforakis, M. E., Georgiou, H. V., Dimitropoulos, N., Cavouras, D., & Theodoridis, S. (2006). Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector machine classifiers. *Artificial intelligence in medicine*, *37*(2), 145-162.

29. Larochelle H, Bengio Y (2008) Classification using discriminative restricted Boltzmann machines. In: Proceedings of the 25th international conference on machine learning, pp 536–543

30. Ho T. (1995) Random decision forests. In: Proceedings of the 3rd IEEE international conference on document analysis and recognition, vol 1, pp 278–282

31. Aruna, S., & Rajagopalan, S. P. (2011). A novel SVM based CSSFFS feature selection algorithm for detecting breast cancer. International journal of computer applications, 31(8), 14-20.

32. Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. Neural processing letters, 9(3), 293-300.

33. Girosi, F. (1998). An equivalence between sparse approximation and support vector machines. Neural computation, 10(6), 1455-1480.

34. Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

35. Friedman, J. H. (1989). Regularized discriminant analysis. Journal of the American statistical association, 84(405), 165-175.

36. Baudat, G., & Anouar, F. (2000). Generalized discriminant analysis using a kernel approach. Neural computation, 12(10), 2385-2404.

37. Sharma, A., Vans, E., Shigemizu, D., Boroevich, K. A., & Tsunoda, T. (2019). DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Scientific reports*, *9*(1), 1-7.

38. Кривчикова, Д. (2022). Аналіз методів інтелектуального аналізу даних для постановки медичних діагнозів.