Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Інформаційних радіотехнологій та технічного захисту інформації
(повна назва)

Кафедра радіотехнологій інформаційно-комунікаційних систем
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

ГЮІК.ХХХХХХ.009 ПЗ
Дослідження інтелектуальної системи пошуку знань з баз даних
(тема)

Виконав: студент 2 курсу, групи ІКТм-18-1
Серенко Іван Максимович
(прізвище, ініціали)
спеціальності 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма
інформаційно-комунікаційні технології
(повна назва освітньої програми)

Керівник д.т.н., професор Кузьомін О. Я.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ Цопа О. І.
(підпис) (прізвище, ініціали)

2019 р

Не містить відомостей заборонених для відкритого публікування.

Керівник д.т.н., професор Кузьомін О. Я.

Студент Серенко І.М.

Харківський національний університет радіоелектроніки

Факультет Інформаційних радіотехнологій та технічного захисту інформації
Кафедра___ радіотехнологій інформаційно-комунікаційних систем ___

Рівень вищої освіти_____другий (магістерський)_____

Спеціальність _____122 Комп'ютерні науки_____
(код і повна назва)

Тип програми _____освітньо-професійна_____
(освітньо-професійна або освітньо-наукова)

Освітня програма ____інформаційно-комунікаційні технології_____
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 20 ___ р.

ЗАВДАННЯ

НА АТЕСТАЦІЙНУ РОБОТУ

студентові_____Серенко Івану Максимовичу_____
(прізвище, ім'я, по батькові)

1. Тема роботи __Дослідження інтелектуальної системи пошуку знань з баз даних
затверджена наказом по університету від___21_ __11__ __2019 р. № ___1730_____
2. Термін подання студентом роботи до екзаменаційної комісії ___24 грудня 2019 р.
3. Вихідні дані до роботи __Науково-технічні публікації та інтернет джерела з тематики атестаційної роботи_____

4. Перелік питань, що потрібно опрацювати в роботі _ вступ, аналіз вимог до вхідних даних, постановка задачі атестаційної роботи, реалізація програмного модуля, висновки_____

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри)
Слайди презентації

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
Основна частина	Проф.кафр.інформатики Кузьомін О.Я.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Аналіз літератури та Інтернет-джерел	10.10.19	виконано
2	Постановка задачі	25.10.19	виконано
3	Дослідження сучасного стану вирішення	08.11.19	виконано
4	Дослідження моделей та методів побудови систем	20.11.19	виконано
5	Написання пояснювальної записки	03.12.19	виконано
6	Підготовка презентації	11.12.19	виконано
7	Перевірка на плагіат	11.12.19	виконано
8	Нормоконтроль	16.12.19	виконано
9	Захист	24.12.19	виконано

Дата видачі завдання __01__ __10__ __2019__ р.

Студент _____
(підпис)

Керівник роботи _____ _____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка до магістерської атестаційної роботи містить:

79с., 4 розділи, 34 рис., 5 табл., 64 джерел.

ЗГОРТКОВА НЕЙРОННА МЕРЕЖА, БАЗА ЗНАНЬ, PYTHON, ГЛУБОКЕ НАВЧАННЯ, ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ.

У цьому дослідженні було навчиено глибоку згорткову нейронну мережу на зображеннях слайдів, отриманих з атласу генома раку, щоб точно і автоматично класифікувати їх на LUAD, LUSC або нормальну легеневу тканину. Продуктивність нашого методу порівнянна з такою у патологів, з середньою площею під кривою (ППК) 0,97. Крім того, ми навчили мережу передбачати десять найбільш часто мутуючих генів в LUAD. Ми виявили, що шість з них - STK11, EGFR, FAT1, SETBP1, KRAS і TP53 - можуть бути передбачені по зображеннях патології, з ППК від 0,733 до 0,856, як виміряно на утримуваної популяції. Ці результати показують, що моделі глибокого навчання можуть допомогти патології у виявленні підтипу раку або генних мутацій. Цей підхід може бути застосований до будь-якого типу раку.

ABSTRACT

The explanatory note to the master's certification work contains:

79 pages, 4 sections, 34 figures, 5 tables, 64 sources.

CONVOLUTIONAL NEURAL NETWORK, KNOWLEDGE BASE, PYTHON, DEEP LEARNING, INTELLECTUAL DATA ANALYSIS

In this study, we trained a deep convolutional neural network on whole-slide images obtained from The Cancer Genome Atlas to accurately and automatically classify them into LUAD, LUSC or normal lung tissue. The performance of our method is comparable to that of pathologists, with an average area under the curve (AUC) of 0.97. Furthermore, we trained the network to predict the ten most commonly mutated genes in LUAD. We found that six of them—STK11, EGFR, FAT1, SETBP1, KRAS and TP53—can be predicted from pathology images, with AUCs from 0.733 to 0.856 as measured on a held-out population. These findings suggest that deep-learning models can assist pathologists in the detection of cancer subtype or gene mutations. Our approach can be applied to any cancer type

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

ІАД – інтелектуальний аналіз даних;

БД – база даних;

БЗ – база знань;

ШІ – штучний інтелект;

ML – machine learning;

DM – data mining;

CNN – convolutional neural network;

LUAD – lung adenocarcinoma;

LUSC – lung squamous cell carcinoma.

ВСТУП

У сучасній медицині ефективність роботи персоналу і надання медичної допомоги ЛПЗ залежить від лікувальних призначень та рекомендацій лікаря, який приймає рішення. Тому для забезпечення якості медичних послуг особлива увага приділяється на проблему раціональної постановки клінічного діагнозу захворювання. Ця проблема ускладняється відсутністю у лікарів достатнього досвіду, швидкого розвитку медицини і брак часових ресурсів на підвищення кваліфікації та досвіду роботи персоналу, наслідком чого стають використання дублюючих досліджень і даремно проведені дорогі і непотрібні лікування.

Об’єктом дослідження даної магістерської дисертації є медична система що повинна покращити результати лікарської діагностики за допомогою алгоритмів постановки діагнозу. Предметом дослідження є комплекс теоретичних, методологічних та практичних проблем побудови інтелектуальних систем підтримки прийняття рішень в задачах медичної діагностики.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ І ПОСТАНОВКА ЗАВДАННЯ

1.1 Сучасні проблеми

З бурхливим ростом розвитку електронної техніки ми отримали можливість збирати і зберігати безпрецедентну кількість інформація разом з тим виникла проблема в способах обробки такої кількості інформації. Аналіз такого обсягу сирих даних вже не під силу людині, хоча необхідність проведення такого аналізу очевидна. Знання, отримані з таких "сирих" даних можуть бути корисні в прийнятті рішень. Для здійснення такого аналізу за допомогою обчислювальних потужностей комп'ютера виникла необхідність в наділення обчислювальних систем здібностями до інтелектуального аналізу. Дослідження в цій області об'єднують під загальною назвою Data Mining (DM) [1].

Data Mining - це процес виявлення в "сирих" даних раніше невідомих нетривіальних практично корисних і доступних інтерпретації знань, необхідних для прийняття рішень в різних сферах людської діяльності.

Результатом такого аналізу повинна бути нетрівіальноа і раніше неізвестноя інформація. Корисність полягає в тому, що ці знання можуть приносити певну вигоду при їх застосуванні. Знання повинні мати наступні властивості:

- повинні передбачати значення одних ознакою на основі інших і описувати нові зв'язки між властивостями;
- бути застосовні і на нових даних з деякою мірою вірогідності;
- бути в зрозумілому для звичайного користувача вигляді.

Використовувані в DM алгоритми, вимагають великої кількості обчислень. Раніше цей факт стримував дослідження в даній області. Однак останнім часом стрімке зростання обчислювальних потужностей дозволив домогтися значних успіхів в інтелектуальній обробці даних. Хоча і зараз виникають ідеї для реалізації яких потужності сучасних обчислювальних систем недостатньо.

Завдання, які вирішуються методами DM, діляться на описові та передбачувальні:

- описові - дають наочне опис наявних прихованих закономірностей;
- передбачувальні - роблять прогнози для тих випадків, для яких даних ще немає.

Для вирішення перерахованих вище завдань було розроблено безліч методів і алгоритмів. Розвиток DM лежить на стику таких областей:

- штучний інтелект;
- машинне зір;
- математична статистика;
- математичне програмування;
- візуалізація;
- теорія баз даних;
- теорія інформації;
- та інші.

Тому очевидно, що методи і алгоритми в більшій мірі запозичені (з деякими доповненнями та змінами) з цих дисциплін.

1.2 Сфери застосування алгоритмів DM

1.2.1 Роздрібна торгівля

З розвитком електронних способів оплати товару виникла можливість зв'язування даних про покупки з конкретною людиною через його банківський рахунок. Що в свою чергу дозволило складати профіль покупця і поліпшити рекламу в цілому для всіх покупців, а також передбачати намір конкретного клієнта про покупку певного товару і запропонувати йому знижку заздалегідь. Крім того, це дозволило оптимізувати стратегію запасання товару на складах в кожній конкретній торговій точці.

Наступним рівнем є аналіз профілів клієнтів і поділ їх на групи за такими властивостями як платоспроможність, залученість та інше.

1.2.2 Банківська справа

В першу чергу інтелектуальний аналіз даних дозволяє виявляти і запобігати спробам шахрайських дій аналізуючи транзакції користувачів і порівнюючи їх з прецедентами.

Сегментація клієнтів на групи допомагає банку вести свою маркетингову політику більш оптимально і цілеспрямовано пропонуючи певним клієнтам найбільш результативні послуги, а також мінімізувати ризики, пов'язані з кредитними позиками.

1.2.3 Страхування

Також, як і з банківською справою за багато років накопичуватися велика кількість даних про клієнтів і прецедентах шахрайства. Аналіз проведених страхових виплат дозволяє переглянути стратегію видача страхових полісів і мінімізувати ризики в подальшому.

1.2.4 Телекомунікації

Інтелектуальний аналіз даних дозволяє поліпшити якість послуг, що надаються за рахунок оптимізації завантаження на телекомунікаційний лінії зв'язку, а також допомагають компаніям більш енергійна просувати свої програми маркетингу і ціноутворення щоб більш стрімко нарощувати базу клієнтів.

1.2.5 Прикладна хімія

Методи DM як же широко застосовуються в прикладної хімії. Дві головні завдання, що стоять перед дослідниками в цій галузі це: знаходження нових властивостей вже відомих хімічних сполук, а також синтез нових елементів і сполук на основі аналізу існуючих. Так як варіантів хімічних сполук безліч, а корисними виявляються одиниці і простий перебір всіх можливих значень може зайняти нескінченну кількість часу, необхідний інтелектуальний підхід до аналізу і синтезу нових хімічних сполук.

1.2.6 Медицина сфера

Відомо безліч інтелектуальних систем для постановки діагнозів. Більшість з них побудовано на основі правил написання сполучень симптомів різних

хвороб. За допомогою яких систем з'являється можливість не тільки діагностувати захворювання з високою точністю, але і призначити лікування з урахуванням протипоказань.

Це найбільш важливо і складно напрямок оскільки не дивлячись на велику кількість даних про історії захворювань і лікування накопичене за весь час "свідомої медицини", велика частина даних представлена в паперовому, а найчастіше письмовому вигляді, описаних хоча і формальним, але все ж людською мовою.

На даний момент всесвітня організація охорони здоров'я (ВООЗ) [2] поширює офіційну статистику, в якій йдеться про три основних причин смертності в світі. Згідно даній статистиці на першому місці стоять хвороби серцево-судинної системи, потім рак і інші хвороби, в тому числі ДТП.

На рисунку 1.1 представлена статистика смертності в 2017 році за версією Центру по контролю і профілактиці захворювань (Center of Disease Control and prevention - CDC) [3].

Фахівці з медичної школи університету Джонса Хопкінса прийшли до висновку, що якщо визнати лікарські помилки, то вони займуть третє місце серед основних причин смерті після захворювання серцево - судинної системи та раку. Також вчені вважають, що несвоєчасні, недостатньо ретельно проведені або помилкові діагностики, так само як і неправильні або надмірні лікування, ймовірно, є причинами більш 250000 смертей в країні за 2017 рік.

Саме тому інтелектуальний аналіз з метою структурування дані в медичній сфері є пріоритетним на сьогоднішній день і був вибраний як основний напрямок для даної дослідницької роботи.

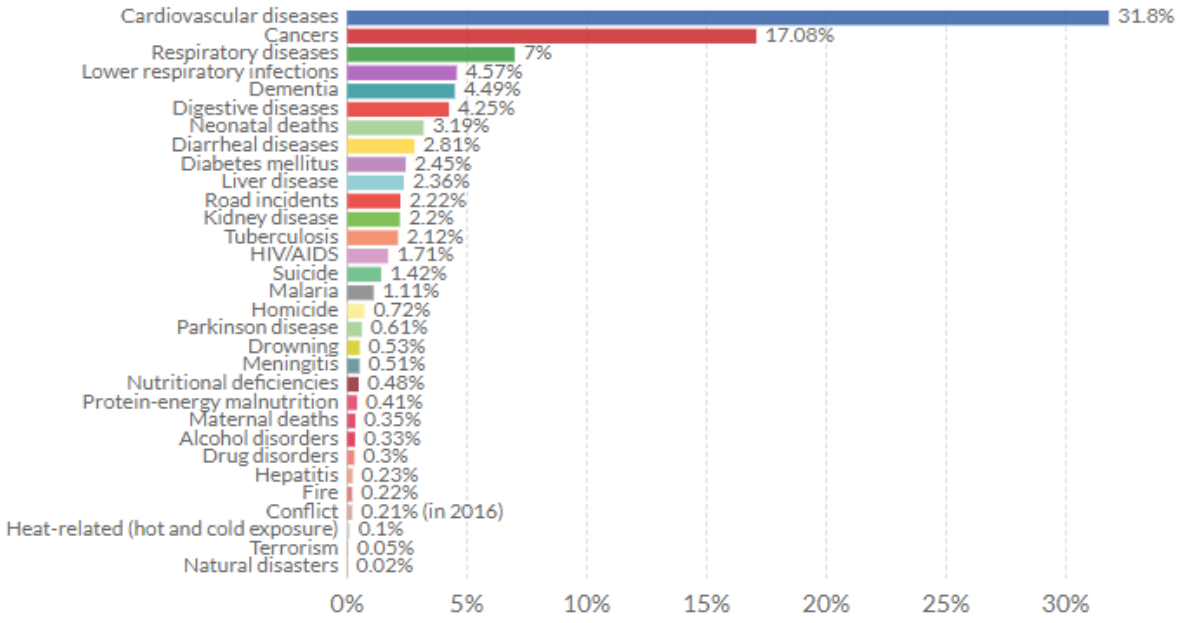

Рисунок 1.1 – Причини смертності. Статистика CDC

1.3 Існуючі рішення

Застосування методів штучного інтелекту в медицині зазвичай пов'язують з виникненням експертних систем в середині восьмидесятих років минулого століття як результату 30-річного академічного періоду досліджень в області штучного інтелекту. Обґрунтоване міркування з цього приводу висловив Е. Фейгенбаум на 5-й Об'єднаної конференції по штучному інтелекту в 1977 році. Суть їх полягала в тому, що фахівці досягають високих результатів, накопичуючи знання і досвід; якщо ж інтелектуальні програми будуть влаштовані так, що зможуть діяти подібним чином, то і вони зможуть досягти високих результатів.

Як було сказано вище, існує велика кількість експертних систем, призначених структурувати і зручно представити знання в медичній сфері для допомоги медичному персоналу в прийнятті рішень установки діагнозу і призначення лікування. Як приклад деякі з них представлені нижче.

Система **CASNET** [7] призначена для діагностики і вибору стратегії лікування Глаукоми. Система **DXplain** [8] - приклад інтелектуальної системи підтримки клінічних рішень, використовується для асистування в процесі

діагностики і містить у своїй базі знань симптоми, лабораторні дані і процедури, що зв'язують їх зі списком діагнозів. Вона забезпечує підтримку і обґрунтування диференціальних діагнозів і наступних досліджень [9]. В її основі даних міститься 4500 клінічних маніфестацій, які зв’язані асоціативними зв'язками більш ніж з 2000 різних нозологій.

У систему вбудований модуль автоматичного машинного навчання, який дозволяє патологу створювати нові правила без участі інженера по знаннях. В даний час створено 2300 таких правил. На побудову кожного нового правила потрібно близько хвилини. Щодня система коментує 100 звітів в області газового складу артеріальної крові, тесту толерантності глюкози та ін. Система **Puff** [12] призначена для інтерпретації результатів функціонального пульмонологічного тесту. Вона використовує прецедентну інформацію; в її базі прецедентів містяться десятки тисяч випадків. Є комерційна версія системи, кілька сотень копій якої впроваджено в ряді країн.

Застосування методів штучного інтелекту в медицині зазвичай пов'язують з виникненням експертних систем в середині вісімдесятих років минулого століття як результату 30-річного академічного періоду досліджень в області штучного інтелекту. Основне міркування з цього приводу висловив Е. Фейгенбаум на 5-й Об'єднаної конференції по штучному інтелекту в 1977 році. Сутність їх полягала в тому, що фахівці досягають високих результатів, накопичуючи знання і досвід; якщо ж інтелектуальні програми будуть влаштовані так, що зможуть діяти подібним чином, то і вони зможуть досягти високих результатів.

Однією з перших систем, які використовували знання для вирішення завдань була експертна система **DENDRAL** [5], розроблена в Стенфордському університеті і призначена для породження формул хімічних сполук на основі спектрального аналізу. В даний час **DENDRAL** поставляється покупцям разом зі спектрометром. Першою власне медичної експертної системою стала система **MYCIN** [6], призначена для діагностики та лікування інфекційних захворювань крові.

Система **Germwatcher** [10] була розроблена в допомогу лікарняним епідеміологу. Містить великий обсяг даних за різними мікробіологічними культурам. Включає базу знань, засновану на правилах, яка використовується для генерації гіпотез про можливих інфекціях.

Система **PEIRS** [11] інтерпретує і коментує звіти за хімічними патологій.

Серед вітчизняних розробок відзначимо систему для синдромной діагностики невідкладних станів у дітей **ДІН** [13], створену в Московському НДІ педіатрії та дитячої хірургії. Ця система містить інформацію про 42 синдромах, які представляють собою список діагностичних пропозицій - гіпотез. Так як вибір лікування багато в чому визначається прогнозом можливих ускладнень, в системі описані взаємозв'язки синдромів, які визначаються причинно-наслідковими, тимчасовими і асоціативними відносинами.

Програмний комплекс **АЙБОЛИТ** [14] перед-призначений для діагностики, класифікації та корекції терапії гострих розладів кровообігу у дітей. Він створений в Центрі серцево-судинної хірургії імені О.М. Бакулєва і активно застосовується при оперативних втручаннях і виборі післяопераційного лікування в умовах реанімаційного відділення. Система включає математичну модель кровообігу, «реагує» на інформацію, що надходить з датчиків поточну інформацію. Вона дозволяє не тільки проводити діагностику і оцінку стану хворого, але і допомагати при виборі і подальшої корекції лікувальних заходів.

Система **HELP** [15] - повна госпітальна інформаційна система, заснована на технологіях штучного інтелекту. Вона підтримує не тільки стандартні функції госпітальних інформаційних систем, але функції підтримки прийняття рішень. Ці функції інкорпоровані в рутинні додатки госпітальної системи. Вони підтримують клінічний процес тривожними сигналами і нагадуваннями, інтерпретацією даних, виробленням пропозицій з управління процесом лікування та клінічних протоколів. Ці функції можуть активуватися зі звичайних додатків або включатися самостійно після введення клінічних даних в комп'ютерну історію хвороби.

Відзначимо ще систему **SETH** [16], область застосування якої аналіз токсичності лікарських засобів. Система заснована на моделюванні експертних міркувань, для кожного токсикологічного класу враховують клінічні симптоми і застосовуються дози. Система виконує моніторинг лікувального процесу, спрямований на контроль взаємодії взаємовиключення ліків.

Diagnos.ru [17] - є єдиною інтерактивною телемедичною системою діагностики, а також найбільшою в світі за кількістю груп діагностованих хвороб і категорій пацієнтів. На даний час діагностується більше 240 захворювань і більше 600 нозологічних одиниць. Нозологічна одиниця - певна хвороба, яку виділяють як самостійну, як правило, на основі встановлених причин, механізмів розвитку і характерних клінічних проявів.

Система діагностики складається з двох частин (програм). Перша здійснює аналіз медичних даних, використовуючи ряд сучасних математичних методів. Друга - та, що розташовується безпосередньо на сайті Діагноз.ru - виконує збір інформації про пацієнта і видає діагноз на основі готових відповідностей, створених аналізатором.

Технічно система діагностики являє собою штучний інтелект на базі нечіткої логіки. Відповіді на питання, які дає користувач, надходять у чіткому вигляді. На основі визначених сполучень відповідей формуються нечіткі припущення з синдромам і захворювань, в яких система може зміцнитися або розчаруватися. Залежно від тих чи інших припущень користувачеві задаються нові питання. Наприкінці видається Можливий діагноз.

Інший принцип роботи системи - комплексність умов для постановки діагнозу. Наприклад, наявність тільки однієї болі в животі ніколи не дасть навіть мінімальної ймовірності якого-небудь захворювання. Зате сукупність навіть незначно виражених ознак може значно підвищити ймовірність відповідного діагнозу. Система діагностики видає відсоткову ймовірність захворювань, а не просто припущення щодо їх наявності-відсутності.

Завдяки тому, що діагностика охоплює всі системи органів людини, з'явилася можливість діагностувати системні захворювання зі складними

поліорганних ураженнями. Що система діагностики дає пацієнтам і лікарям? Для пацієнта це список можливих хвороб із зазначенням ймовірності, рекомендації по відвідуванню фахівців і коментарі до поточного стану. Для лікарів - можливість подивитися анамнез і відразу без додаткових питань мати висновок. На рисунку 1.2 представлений веб інтерфейс системи.

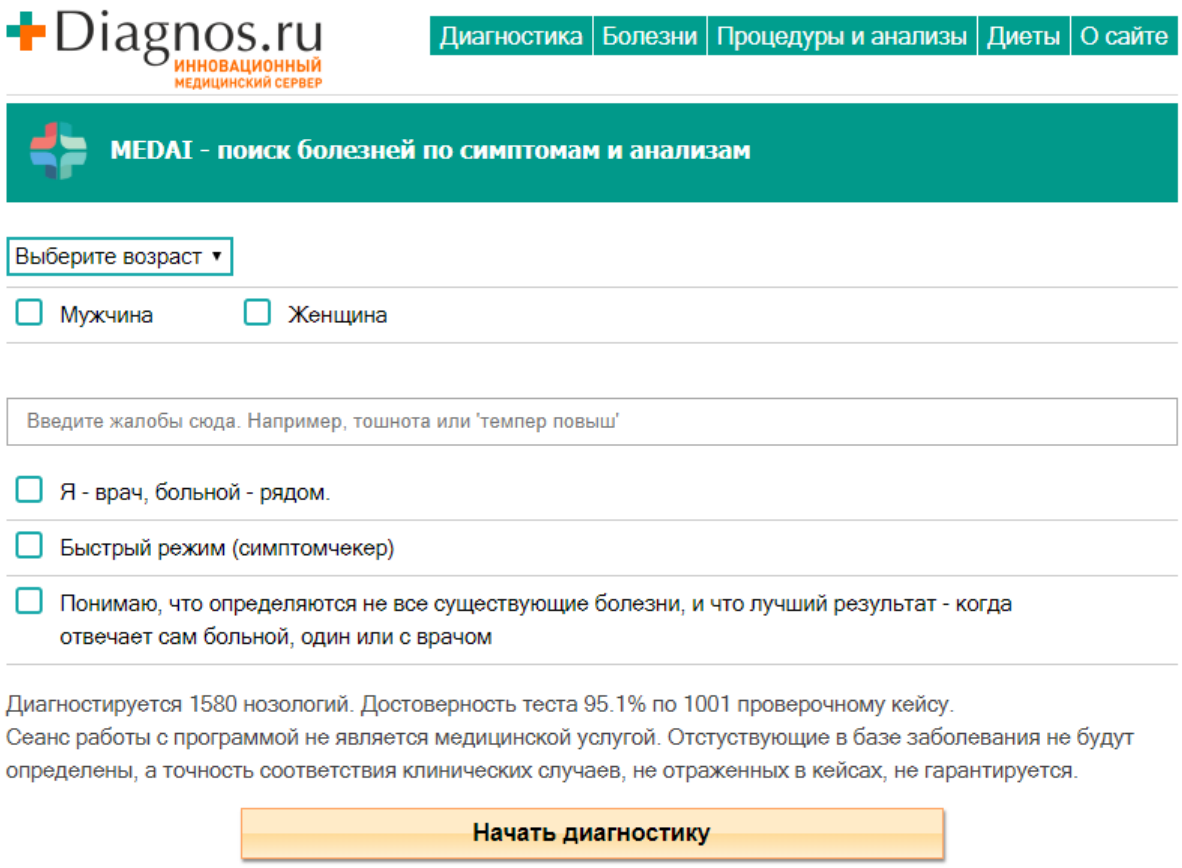

Рисунок 1.2 – інтерфейс Diagnos.ru

Homeopath-expert.com [18] - Тут можна з високою точністю знайти гомеопатичний препарат, ґрунтуючись на симптомах. Щоб дана експертна система дала правильний результат, потрібно ввести симптоми в поле для швидкого пошуку АБО скористатися опитувальником, відзначаючи відповідні симптоми галочкою. Необхідно вибрати не менше чотирьох симптомів. На рисунку 1.3 представлений веб інтерфейс системи.

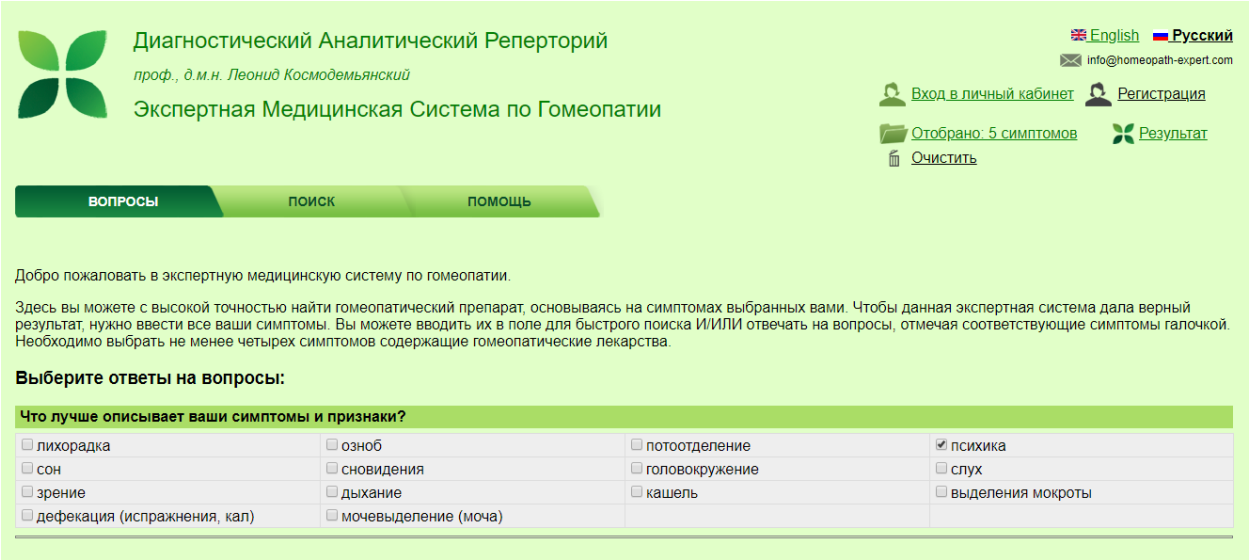

Рисунок 1.3 – інтерфейс Homeopath-expert.com

Такі системи досить непогано визначає захворювання по явними ознакам, однак через те що в її основі лежить модель дерево рішень, підтримка та розширення такої системи дуже трудомісткий і малоефективноий процесс. Так як система враховує тільки явні симптоми вона може бути корисна лише для постановка попереднього діагнозу.

1.4 Огляд існуючих методів аналізу даних

Кілька основних методів, які використовуються для інтелектуального аналізу даних, описують тип аналізу і операцію по відновленню даних. На жаль, різні компанії і рішення не завжди вживають одні й ті ж терміни, що може погіршити плутанина і уявну складність.

Розглянемо деякі ключові методи і приклади того, як використовувати ті або інші інструменти для інтелектуального аналізу даних.

1.4.1 Асоціація

Асоціація (або відношення) [4], ймовірно, найбільш відомий, знайомий і простий метод інтелектуального аналізу даних. Для виявлення моделей робиться просте зіставлення двох або більше елементів, часто одного і того ж типу.

Наприклад, відстежуючи звички покупки, можна помітити, що разом з полуницею зазвичай купують вершки.

Створити інструменти інтелектуального аналізу даних на базі асоціацій або відносин неважко. Наприклад, в InfoSphere Warehouse є майстер, Який видає конфігурації інформаційних потоків для створення асоціацій, Досліджуючи джерело вхідної інформації, підстава прийняття рішень і вихідну інформацію. На рисунку 1.4 наведено Відповідний приклад для зразка бази даних.

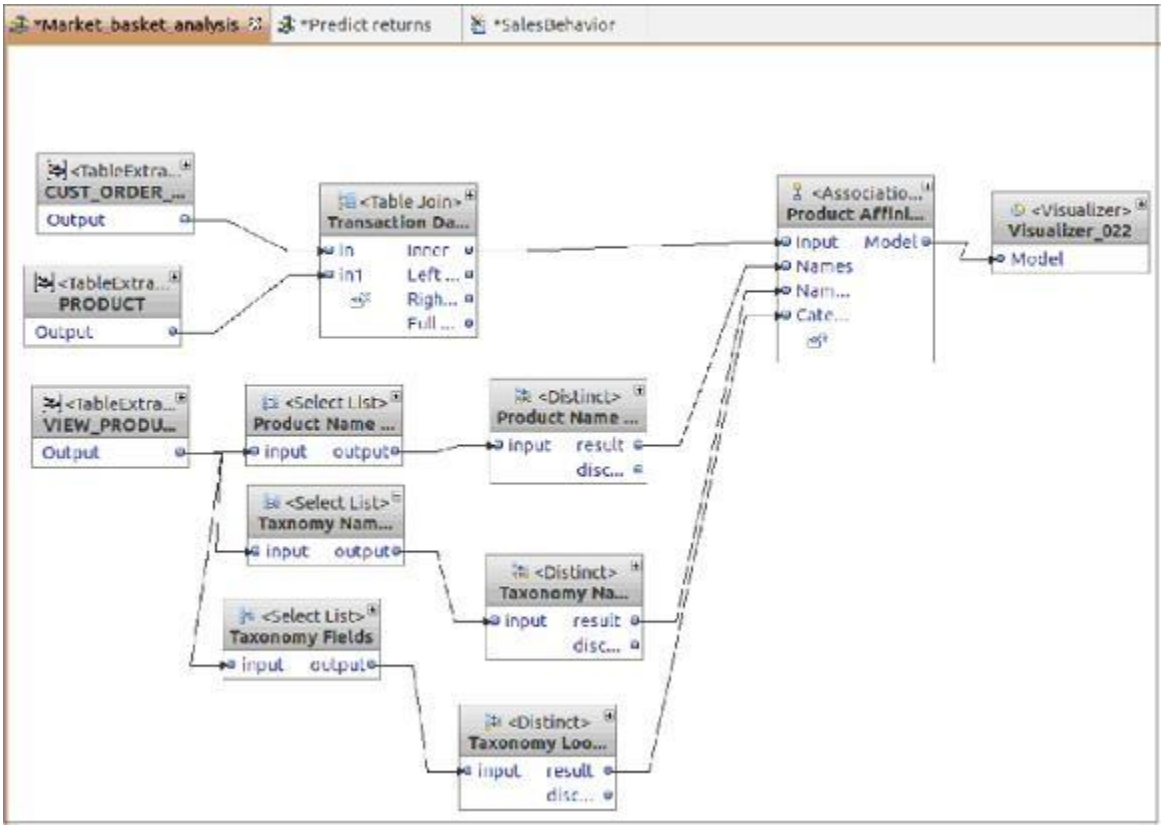

Рисунок 1.4 - Інформаційний потік, який використовується при підході асоціації

1.4.2 Класифікація

Класифікацію [4] можна використовувати для отримання уявлення про тип покупців, товарів або об'єктів, Описуючи кілька атрибутів для ідентифікації певного класу. Наприклад, автомобілі легко класифікувати за типом (седан, позашляховик, кабріолет), визначивши різні атрибути (кількість місць, форма

кузова, провідні колеса). Вивчаючи новий автомобіль, можна віднести його до певного класу, Порівнюючи атрибути з відомим визначенням. Те ж принципи можна застосувати і до покупців, наприклад, Класифікуючи їх за віком і соціальної групи.

Крім того, класифікацію можна використовувати в якості вхідних даних для других методів. Наприклад, для визначення класифікації можна застосовувати дерева прийняття рішень. Кластеризація дозволяє використовувати загальні атрибути різних класифікацій з метою виявлення кластерів.

1.4.3 Кластеризація

Досліджуючи [4] один або більше атрибутів або класів, можна згрупувати окремі елементи даних разом, отримуючи структуроване висновок. На простому рівні при кластеризації використовується один або кілька атрибутів в якості основи для визначення кластера подібних результатів. Кластеризація корисна при визначенні різної інформації, тому що вона корелюється з іншими прикладами, так що можна побачити, де подібності та діапазони узгоджуються між собою.

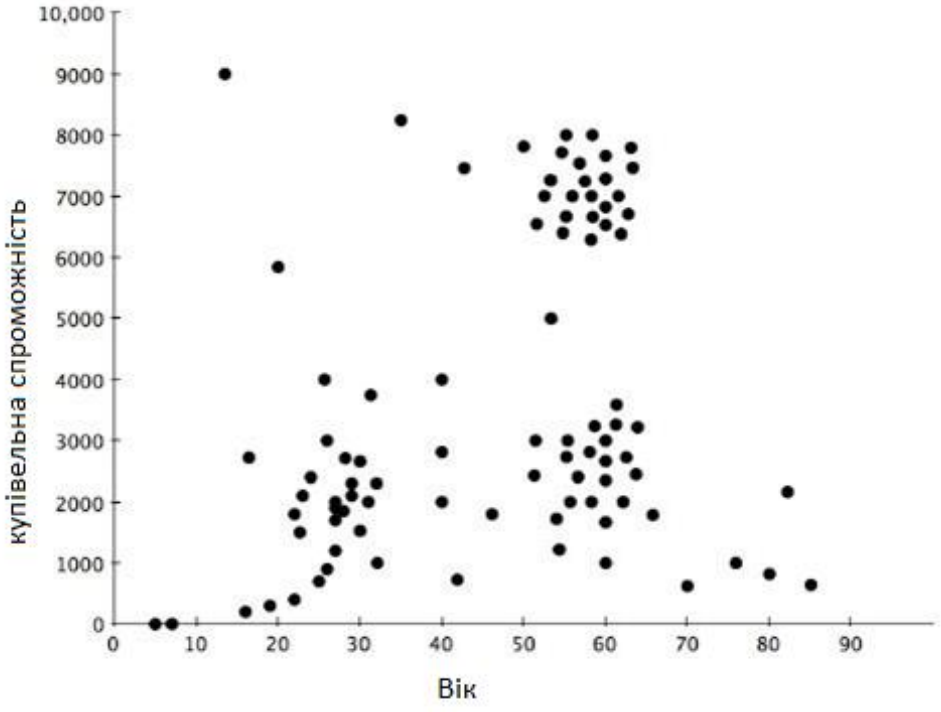

Рисунок 1.5 - Кластеризація

Метод кластеризації працює в обидві сторони. Можна припустити, що в певній точці є кластер, а потім використовувати свої критерії ідентифікації, щоб перевірити це. Графік, зображений на рисунку 1.5, демонструє наочний приклад. Тут вік покупця порівнюється з вартістю покупки. Розумно очікувати, що люди у віці від двадцяти до тридцяти років (до вступу в шлюб і появи дітей), а також в 50-60 років (коли діти покинули будинок) мають більш високий наявний дохід.

У цьому прикладі видно два кластери, один в районі $ 2000 / 20-30 років і інший в районі $ 7000-8000 / 50-65 років. В даному випадку ми висунули гіпотезу і перевірили її на простому графіку, Який можна побудувати за допомогою будь-якого відповідного ПО для побудови графіків. Для більш складних комбінацій потрібен повний аналітичний пакет, особливо якщо потрібно автоматично основі рішення на інформації в найближчій сусіда.

Така побудова кластерів є багаторазовим собою Спрощений приклад так званого образу найближчого сусіда. ОКРЕМИХ покупців можна розрізняти по їх буквальною близькості один до одного на графіку. Досить імовірно, що покупці з одного і того ж кластера поділяють і інші загальні атрибути, і це припущення можна використовувати для пошуку, класифікації та інших видів аналізу членів набору даних.

Метод кластеризації можна застосувати і в зворотний бік: З огляду на певні вхідні атрибути, виявляючи різні артефакти. Наприклад, недавнє дослідження чотиризначних PIN-кодів виявили кластери чисел в діапазонах 1-12 і 1-31 для першої і другої пар. Зобразивши ці пари на графіку, можна побачити кластери, пов'язані з датами (дні народження, ювілеї).

Слід чітко зрозуміти різницю між класифікацією і кластеризацией документів. Класифікація це віднесення кожного обекту в певний клас із заздалегідь відомими параметрами, отриманими на етапі навчання. Число класів строго обмежена. Кластеризація - розбиття безлічі обэктыв на кластери - підмножини, параметри яких заздалегідь невідомі.

1.4.4 Прогнозування

Прогнозування [4] - це широка тема, яка простягається від передбачення відмов компонентів обладнання до виявлення шахрайства і навіть прогнозування прибутку компанії. У поєднанні з іншими методами інтелектуального аналізу даних прогнозування передбачає аналіз тенденцій, класифікацію, зіставлення з моделлю і відносини. Аналізуючи минулі події або екземпляри, можна передбачати майбутнє.

Наприклад, Використовуючи дані по авторизації кредитних карт, можна об'єднати аналіз дерева рішень минулих транзакцій людина з класифікацією і зіставлення з історичними моделями з метою виявлення шахрайському транзакцій. Якщо покупка авіаквитків в США збігається з транзакціями в США, то цілком ймовірно, що ці транзакції справжні.

1.4.5 Послідовні моделі

Послідовні моделі [4], які часто використовуються для аналізу довгострокових даних, - корисний метод виявлення тенденцій, або регулярних повторень подібних подій. Наприклад, за даними про покупців можна визначити, що в різні пори року вони купують певні набори продуктів. За цією інформацією додаток прогнозування купівельної корзини, грунтуючись на частоті і історії покупок, може автоматично припустити, що в корзину будуть додані ті чи інші продукти.

Стосовно до медичної системи, послідовна модель на прогнозування можуть бути використана для пероедбачення епідемій, аналізуючи дані пацієнтів з прив'язкою до певної місцевості.

1.4.6 Дерева рішень

Дерево рішень [4], пов'язане з більшістю інших методів (головним чином, класифікації та прогнозування), можна використовувати або в рамках критеріїв відбору, або для підтримки вибору певних даних в рамках загальної структури. Дерево рішень починають з простого питання, який має дві відповіді (іноді більше). Кожна відповідь призводить до наступного питання, допомагаючи класифікувати та ідентифікувати дані або робити прогнози.

Дерева рішень часто використовуються з системами класифікації інформації про властивості і з системами прогнозування, де різні прогнози можуть грунтуватися на минулому історичному досвіді, який допомагає побудувати структуру дерева рішень і отримати результат.

Недоліком даного методу є той факт що побудова дерева рішень має здійснюватися експертом або групою експертів з досить високої кваліфікації.

На рисунку 1.6 наведено приклад класифікації несправних станів.

На практиці дуже рідко використовується тільки один з цих методів. Класифікація та кластеризація - подібні методи. Використовуючи кластеризацию для визначення найближчих сусідів, можна додатково уточнити класифікацію. Дерева рішень часто використовуються для побудови і виявлення класифікацій, які можна простежувати на історичних періодах для визначення послідовностей і моделей.

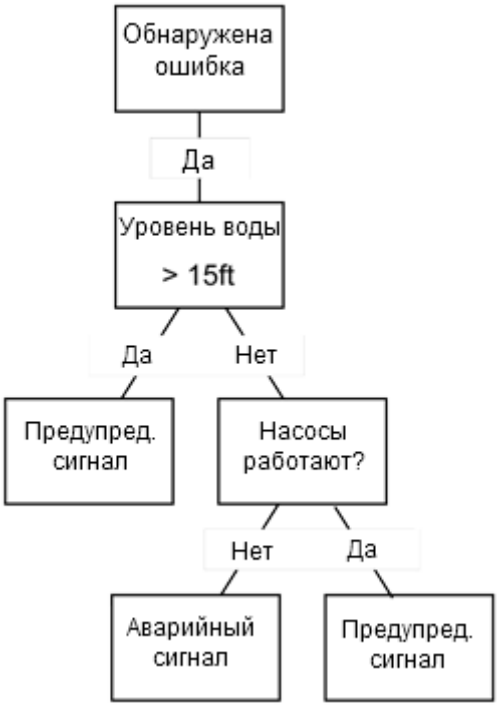

Рисунок 1.6 - Дерево рішень

1.4.7 Обробка з запам'ятовуванням

При всіх основних методах часто має сенс записувати і згодом вивчати отриману інформацію. Для деяких методів це абсолютно очевидно. Наприклад,

при побудові послідовних моделей і навчанні з метою прогнозування аналізуються історичні дані з різних джерел і примірників інформації.

В інших випадках цей процес може бути більш яскраво вираженим. Дерева рішень рідко будуються один раз і ніколи не забуваються. При виявленні нової інформації, подій і точок даних може знадобитися побудова додаткових гілок або навіть зовсім нових дерев.

Деякі з цих процесів можна автоматизувати. Наприклад, побудова прогностичної моделі для виявлення шахрайства з кредитними картами зводиться до визначення ймовірностей, які можна використовувати для поточної транзакції, з подальшим відновленням цієї моделі при додаванні нових (підтверджених) транзакцій. Потім ця інформація реєструється, так що наступного разу рішення можна буде прийняти швидше.

1.5 Підсумок та постановка задачі

Провівши аналіз існуючих методів і моделей побудови інтелектуальних систем аналізу великих даних була складена зведена таблиця (таблиця 1.1) їх порівняння за чотирма критеріями.

- складність реалізації - відображає рівень кваліфікації необхідний для розробки моделі;
- складність підтримки відображає рівень кваліфікації необхідні для підтримки і розширення моделі;
- ефективність - коефіцієнт залежно користь від трудовитрат;
- універсальність - коефіцієнт показує ступінь застосовності моделі до різних областей знань.

Таблиця 1.1 - Критеріальна оцінка моделей

	Складність реалізації (менше краще)	Складність підтримки (менше краще)	Эфеекативність (більше краще)	Універсальність (більше краще)
Ассоціація	0.5	0.8	0.2	0.8
Кластеризація	0.6	0.1	0.9	1
Классифікасія	0.4	0.2	0.7	0.9
Послідовна модель	0.7	0.7	0.3	0.2
Дерево рішень	0.9	0.9	0.5	0.3
Прогнозування	1	0.3	0.8	0.1

Виходячи з розглянутого вище, робимо висновок що найкращим методом інтелектуального аналізу даних є класифікація, особливо в зв'язку з бурхливим розвитком універсальних класифікаторів, тобто нейронних мереж. Тому в подальших розділах ми розглянемо типи нейронних мереж способи їх навчання і існуючі архітектурні рішення, а також їх придатність в медичній сфері. В результаті спробуємо розробити модель яка допоможе вирішити проблему з високим відсотком помилкових діагнозів, тобто буде системою підтримки прийняття рішень. Попередня структура система представлених на рисунку 1.7.

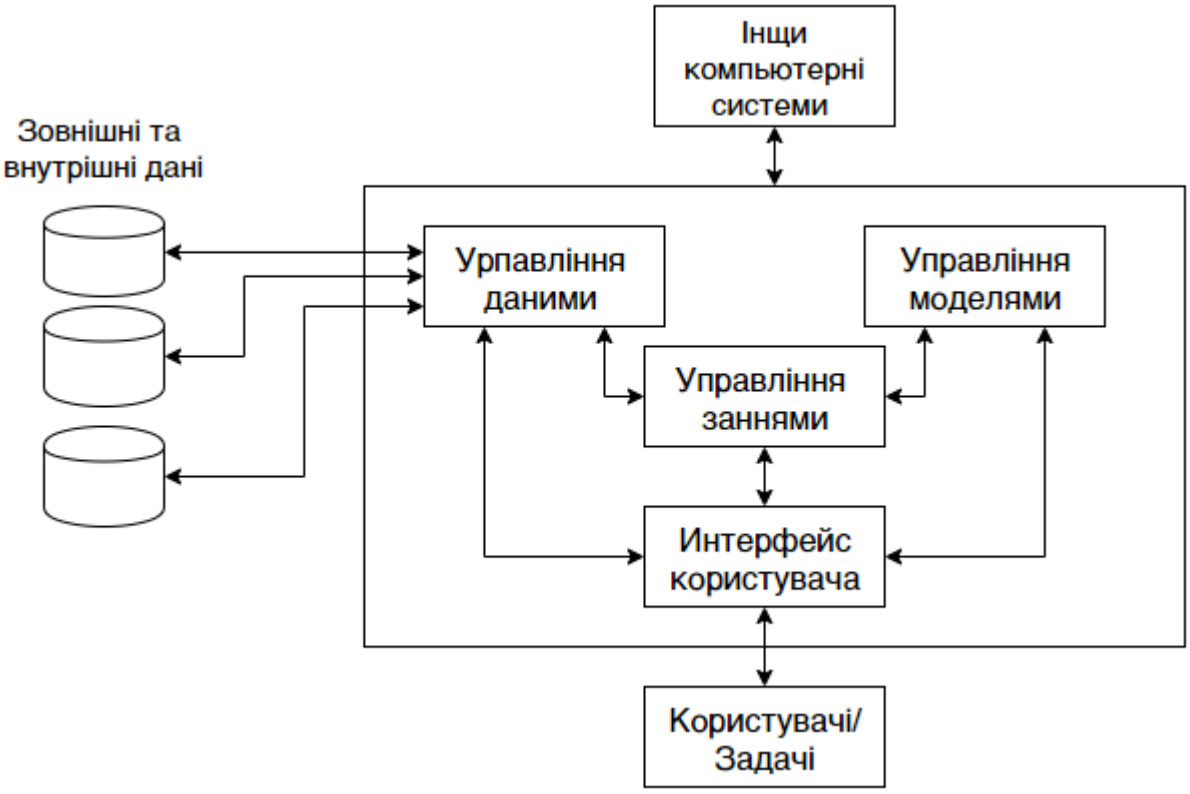

Рисунок 1.7 – Структурна схема системы що проектуется

2 ТЕХНОЛОГІЇ РОЗРОБКИ СИСТЕМИ

2.1 Нейронні мережі

Комп'ютерний зір складається з різних проблем, таких як класифікація зображень, локалізація, сегментація і об'єкт виявлення. Серед них класифікація зображень може розглядатися як фундаментальна проблема і служить основою для інших проблем з комп'ютерним зором. До 90-х років тільки традиційні підходи машинного навчання використовувалися, щоб класифікувати зображення. Але точність і масштаб завдання класифікації були обмежені декількома проблемами, такими як ручний процес вилучення ознак і т. Д. В останні роки глибока нейронна мережа (DNN), також званий глибоким навчанням [19], успішно примененяется до великих масивів даних з використанням алгоритму зворотного поширення [20]. Серед DNN, згорткві нейронна мережа демонструє відмінні показники в задачах комп'ютерного зору, особливо в класифікації зображень.

Згорткова нейронна мережа (CNN або ConvNet) є особливий тип багатошарової нейронної мережі, натхненний механізм оптичної системи живих істот. Хьюбел і Візель [21] виявили, що клітини зорової кори тварин виявляють світло в маленькому сприйнятливому поле. Мотивований цією роботою, в 1980 році Куніхіко Фукусіма представив неокогнітрон [22] яка є багатошаровою нейронною мережею, здатної розпізнавати візуальні патерни, ієрархічно, через навчання. Ця мережа розглядається як теоретичне база для CNN. У 1990 LeCun et al. представив практичну модель CNN [23] і розробив LeNet-5 [24]. Тренування алгоритмом зворотного поширення [25] допоміг LeNet-5 розпізнавати візуальні шаблони в наборі пікселів безпосередньо, без використання будь-якого окремого механізму виділеніе ознак. Також менше зв'язків і параметрів CNN, ніж у звичайній нейронної мережі прямого зв'язку з аналогічний розмір мережі, полегшує навчання моделі. Але в той же час, не дивлячись на всі переваги, продуктивність CNN в складних завданнях, таких як класифікація зображень з

високою роздільною здатністю, було обмежено відсутністю великої кількості даних для навчання, відсутність кращий методів оптимізації і недостатня обчислювальних потужностей.

В даний час у нас є великі набори даних з розміченій зображеннями з високою роздільною здатністю в тисячу категорій, як ImageNet [26], LabelMe [27] і т. З появою потужних графічних процесорів і кращих методів оптимізації, CNN забезпечує видатну продуктивність в задачах класифікації зображень. У 2012 році нейросеть глибокої згортки, звана AlexNet [28], розроблена Крижевський і співавт показав відмінні результати в ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [29]. Успіх AlexNet став натхненням інших моделей CNN, таких як ZFNet [30], VGGNet [31], GoogleNet [32], ResNet [33], DenseNet [34], CapsNet [35], SENet [36] та інщі. В наступні роки.

У цьому розділі ми спробували дати огляд досягнень CNN в області класифікації зображень. Загальний вигляд архітектур CNN представлений в розділі II. Розділ III описує архітектуру і деталі навчання різних моделей CNN. У розділі IV представлено порівняння між різними моделями CNN. Висновки в розділі V.

2.2 Згорткова нейронна мережа

Типовий CNN складається з одного або декількох блоків шарів згортки і підвибірки, після цього один або кілька повністю підключених шарів і вихідний шар, як показано на рисунку 2.1.

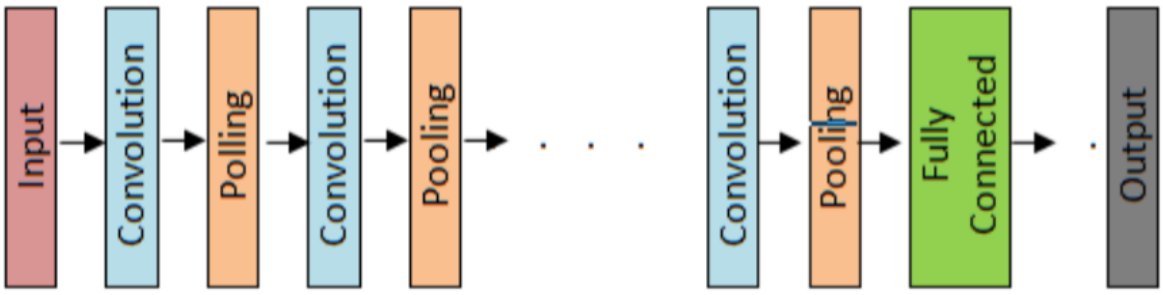

Рисунок 2.1 – Типова структура CNN

2.2.1 Згортковый шар

Згортковый шар є центральною частиною CNN. Зображення зазвичай носять стаціонарний характер. Це означає формування однієї частини зображення таке ж, як і будь-який інший частина. Таким чином, функція, вивчена в одному регіоні, може відповідати картина в іншому регіоні. У великому зображенні ми беремо маленький сегмент і проносимо його через всі точки на великому зображенні (Вхід). Проходячи в будь-який момент, ми звертаємо їх в одну точку (Вихід). Кожен невеликий сигмент зображення, який обходить більше зображення називається фільтром або ядром (Kernel). Фільтри налаштовуються на основі методу зворотного поширення ощибки. На рисунку 2.2 показана типова згорткові операція.

2.2.2 Субдіскретізація або пул

Об'єднання (Pooling) просто означає стиснення зображення. Воно приймає невелику область згорнутого виходу в якості вхідних даних і виробляє подвиборку для виробництва одног виходу точки. Існують різне методи обєднання, такі як:

- максимальна об'єднання (Max pooling);
- середнє об'єднання (mean pooling);
- середній пул (avg pooling);
- інші.

При максимальному пулінгу (pooling) береться максимальне значення в регіоні за вихідну, як показано на рисунку 2.3. Об'єднання зменшує кількість параметрів, які потрібно обчислити, і робить мережу інваріантної до перекладів за формою, розміром і масштабом.

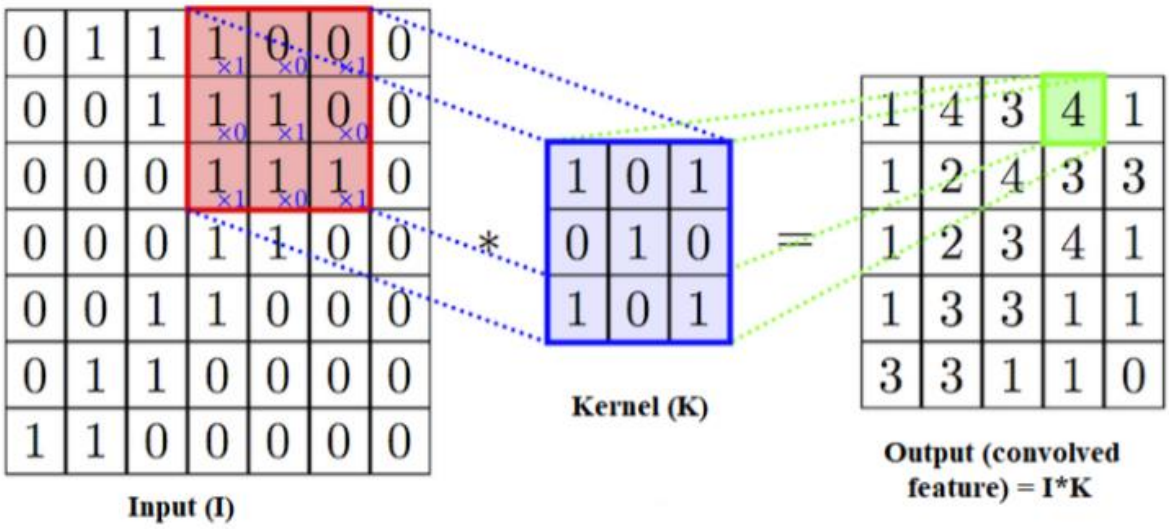

Рисунок 2.2 – Згортковый шар

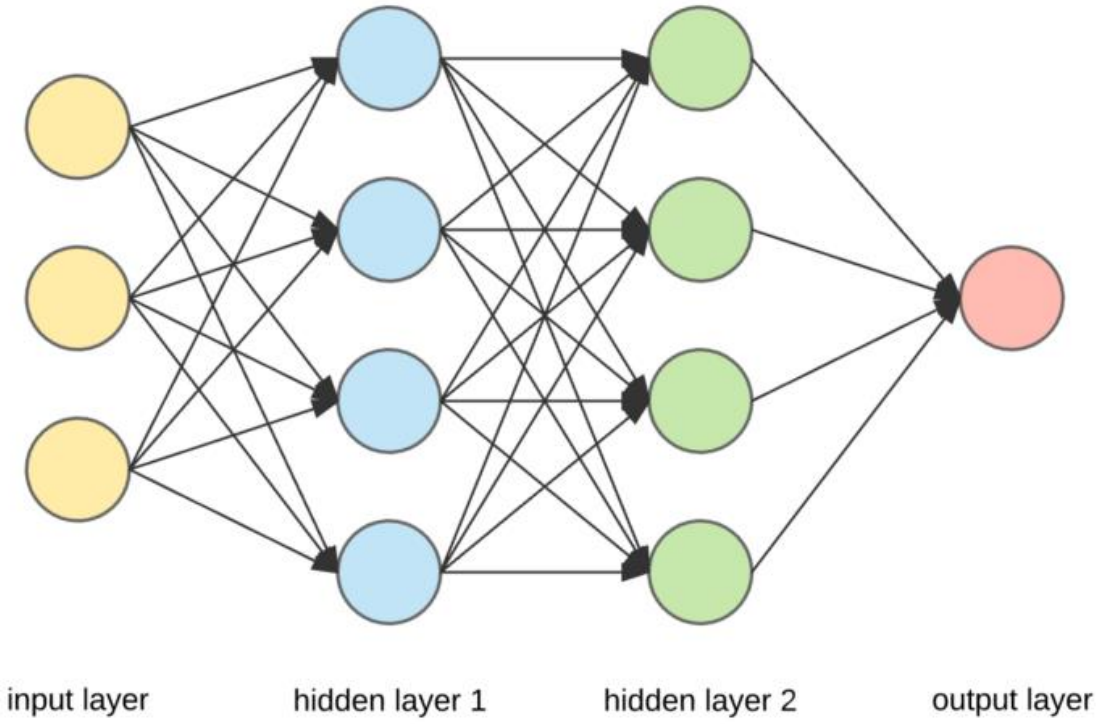

Рисунок 2.3 – Повнозв’язный шар шар

2.2.3 Повнозв’язный шар (шар FC)

Остання секція CNN - це в основному один або кілька повнозв'язних шарів (рис. 4). Цей шар приймає дані від всіх нейронів в попередньому шарі і виконує роботу з індивідуальним нейрон в поточному шарі для генерації виведення.

2.3 Різні моделі CNN для класифікації зображення

2.3.1 LeNet-5:

У 1998 році LeCun et al. представив CNN для класифікації рукописних цифр. Їх модель CNN, названа LeNet-5 [24] як показано на рисунку 2.4, має 7 зважених (учнів) шарів. Серед них три (C1, C3, C5) згортальних шари, два (S2, S4) середній пул шарів, один (F6) повністю пов'язаний шар і один вихідний шар. Для активації нелінійності перед операцією об'єднання використовувалася сігмоідной функція.

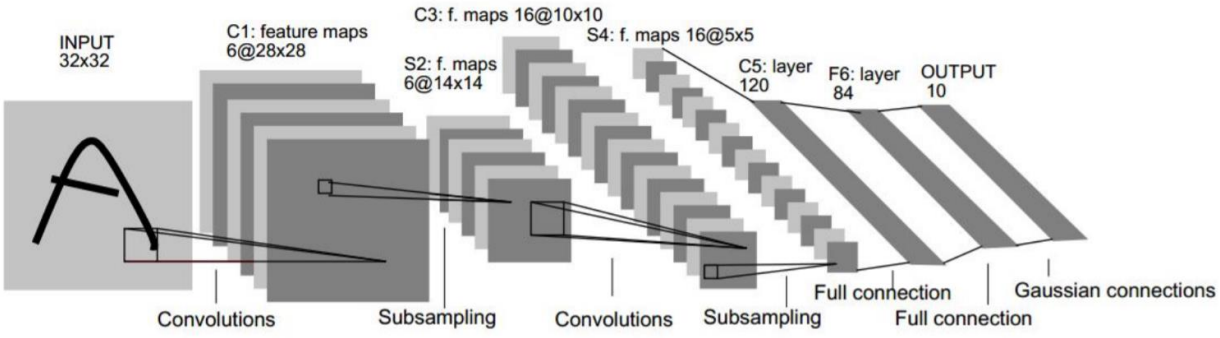

Рисунок 2.4 – Архітектура LeNet-5

У таблиці 2.1 приведені різні шари, розмір фільтрів використовуються в кожному шарі згортки, розмір вихідної карти об'єктів і загальна кількість параметрів, необхідних для шару LeNet-5.

Таблиця 2.1 – Архытектура LeNet-5

Шар	Розмір фільтра	Кількість фіильтрів	Розмір вихідного шару	Кількість параметрів
Convolution(C1)	5 × 5/1	6	28 × 28 × 6	156
Sub-sampling(S2) 12	2 × 2/2		14 × 14 × 6	12
Convolution(C3)	5 × 5/1	16	10 × 10 × 16	1516
Sub-sampling(S4)	2 × 2/2		5 × 5 × 16	32
Convolution(C5)	5 × 5	120	1 × 1 × 120	48120
Fully Connected(F6)	2 × 2		14 × 14 × 6	10164
OUTPUT				84

1) Використовуваний набір даних: для навчання і тестування LeNet-5, використовувалася база рукописних цифр MNIST [38]. База даних містить 60 тис. тренувальних і 10 тис. тестових даних. Розмір вхідного зображення цієї моделі в основному становить 32 × 32 пікселя, що більше, ніж найбільший символ (20 × 20 пікселів) в базі даних, оскільки центральна частина чутливого поля багата ознаками. Розмір вхідних зображень нормується і центрується в поле 28 × 28.

2) Деталі навчання: Автори навчили кілька версій LeNet-5 з використанням стохастичного градієнтного спуску (SGD) з 20 ітераціями для всіх систему адаптації за сеанс зі зниженою швидкістю глобального навчання і імпульсом 0,02. У 1990-х роках LeNet-5 був досить хороший. LeNet-5 і LeNet-5 (з перекручуванням) досягли частоти точності 0,95% і 0,8% відповідно, в наборі даних MNIST. Але в міру того, як обсяг даних, дозвіл зображення і кількість класів проблеми класифікації згодом збільшувалися, нам потрібна була більш глибока згорткова мережа і потужна машина з графічним процесором для навчання моделі.

2.3.2 AlexNet-2012:

У 2012 році Алексом Крижевський і співавторами була розроблена архітектура глибокої CNN, званої AlexNet [28], щоб класифікувати дані ImageNet [26]. Архітектура AlexNet така ж, як LeNet-5, але набагато більше. Вона складається з 8 учнів шарів. Серед них 5 згортальних шарів і 3 полносвязанних шару. Використання випрямленою лінійної функції (ReLU) як активационной функції після згорткового шару і повнозв'язних шарів допомогли навчити їх модель швидше, ніж аналогічні мережі з функцією активації - гіпербаліческій тангенц. У мережі використовувалася локальна нормалізація реакції (LRN), звана «brightness normalization», після першого і другого згорткового шару який допомагає узагальненню. Вони використовували max-pool шар після кожного рівня LRN і п'ятого згорткового шару. На рисунку 2.5 показані архітектура AlexNet. У таблиці 2.2 представлена специфікація AlexNet.

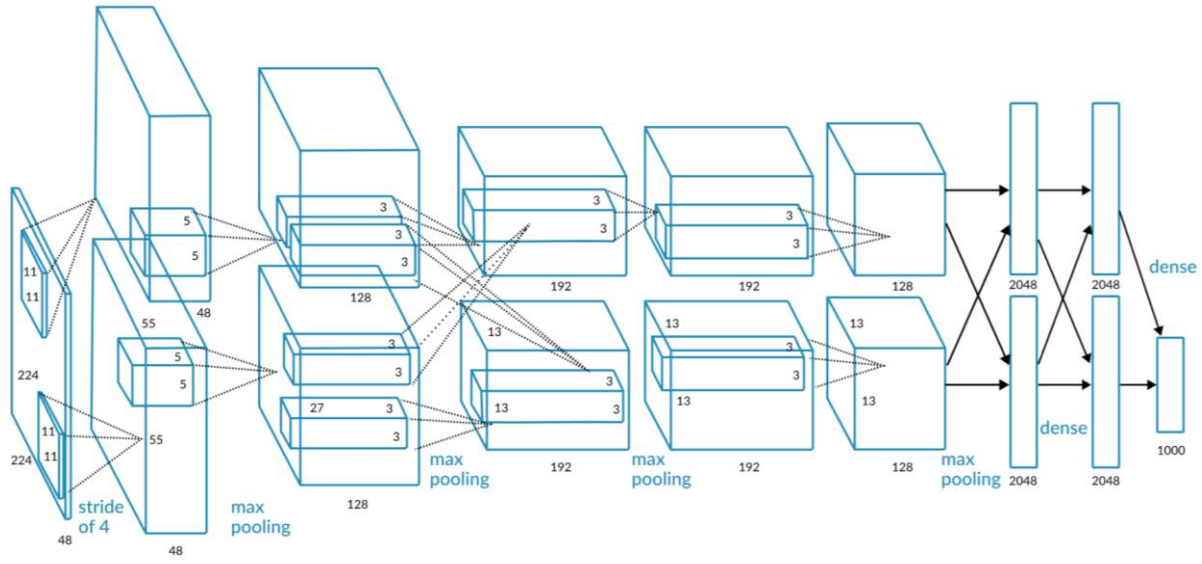

Рисунок 2.5 – Архітектура AlexNet

Таблица 2.2 - Архітектура AlexNet

Шар	Розмір фільтра	Кількість фіильтрів	Розмір вихідного шару	Кількість параметрів
Conv-1	11 × 11/4	96	55 × 55 × 96	34848
pool-1	3 × 3/2		27 × 27 × 96	
Conv-2	5 × 5/1	256	27×27×256	614400
pool-2	3 × 3/2		13×13×256	
Conv-3	3 × 3/1	384	13×13×384	981504
Conv-4	3 × 3/1	384	13×13×384	1327104
Conv-5	3 × 3/1	256	13×13×256	884736
pool3	3 × 3/2		6 × 6 × 256	
FC6			1 × 1 × 4096	37748736
FC7			1 × 1 × 4096	16777216
FC8			1 × 1 × 1000	4096000

1) Використовуваний набір даних: команда розробки AlexNet використовували для класифікація 1.2 мільйона зображень у високому дозволом і 1000й класів з ILSVRC - 2010 і ILSVRC - 2012.

2) Деталі навчання: з зображення зі змінною роздільною здатністю ImageNet, AlexNet використовував знижену дискретизацию і центрованої 256 × 256 піксельний зображення. Щоб зменшити перенавчання, вони використовували дані часу виконання, а також метод регуляризації, званий випадання (dropout). У доповненні даних, вони витягли перекладений і горизонтально відображені 10 випадкових плям розміром 224 × 224 зображення, а також використовується аналіз основних компонентів (principal component analysis - PCA) для зміщення каналів RGB зображень. Автори навчили AlexNet з використанням стохастичного градієнтного спуску (SGD) з розміром партії 128,

зниженням ваги 0,0005 і імпульсом 0,9. Зниження ваги працює як регуляризатора і зменшує помилку навчання. AlexNet пройшов навчання на двох графічних процесорах NVIDIA GTX-580 3 ГБ використання крос-графічного розпаралелювання протягом п'яти-шести днів.

Автори помітили, що видалення будь-якого середнього шару погіршує продуктивність мережі. Отже, результат залежить від глибини мережі.

2.3.3 GoogLeNet

Архітектура GoogLeNet [32], відрізняється від звичайного CNN. В ній збільшилася кількість одиниць в кожному шарі з використанням паралельних фільтрів, які називаються inception - модуль розміром 1 × 1, 3 × 3 і 5 × 5 в коженому шарі згортки. Також було збільшено кількість шарів до 22. На рисунку 2.6 показані 22 шару GoogLeNet. При розробці цієї моделі вони вважали обчислювальний бюджет фіксованим. Так що модель може бути використана в мобільні і вбудовані системи. Вони використовували серію вагові фільтри Габора різного розміру на початку архітектура для обробки декількох масштабів. Зробити архітектуру обчислювально ефективні, вони використовували початковий модуль зі зменшенням розмірності замість наївної версії початкового модуля. Рисунку 2.7 і рисунок 2.8 показують обидва початкових модуля. Незважаючи на 22 шару, кількість параметри, які використовуються в GoogLeNet, в 12 разів менше, ніж у AlexNet, але його точність значно краще. Вся згортка, Шари скорочення і проекції використовують нелінійність ReLU. Они використовували середній пул шар замість повністю підключений шари. Крім деяких початкових модулів, вони використовували допоміжні класифікатори, які в основному менше CNN, щоб боротьба зі зникненням градієнта і проблема переоснащення.

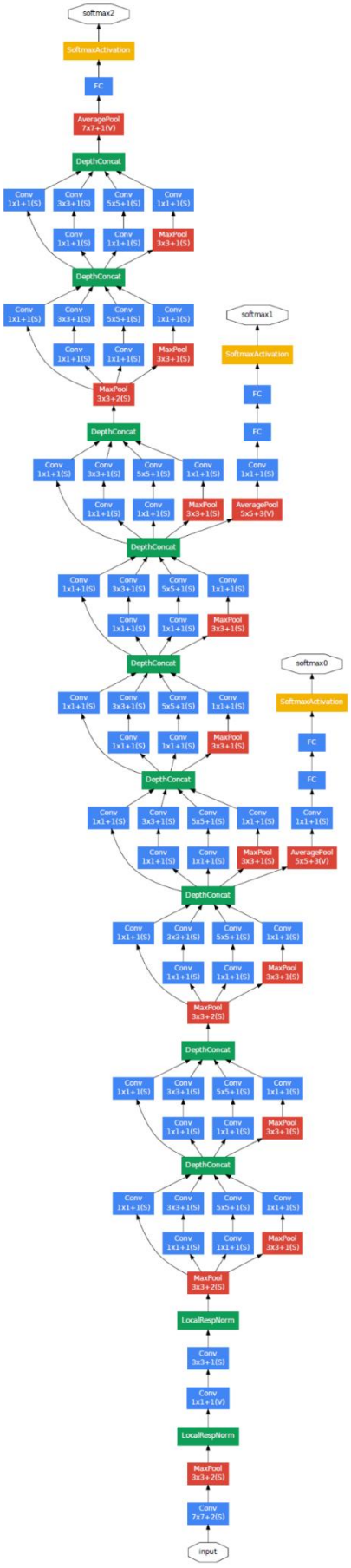

Рисунок 2.6 – Архітектура GooleLeNet

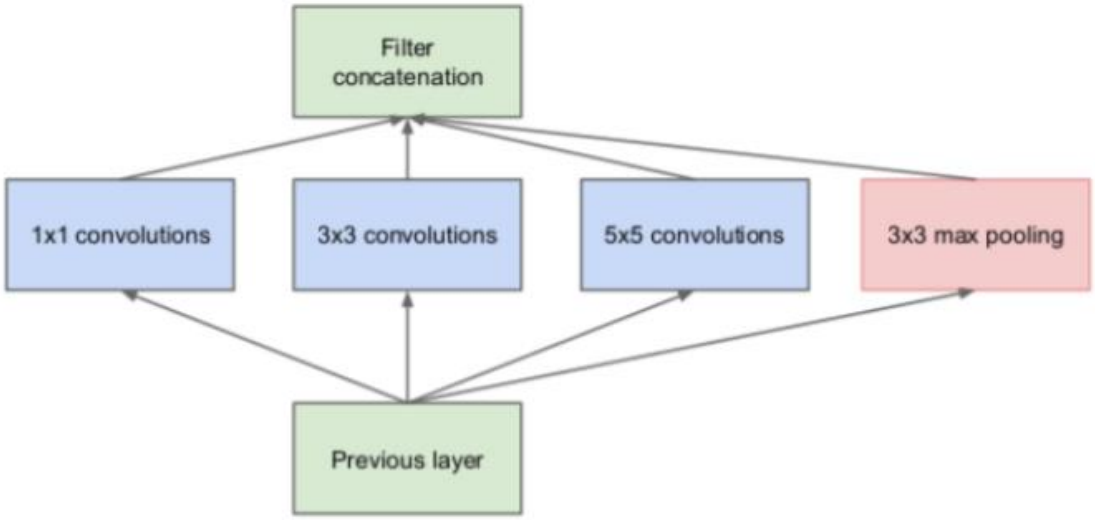

Рисунок 2.7 - Схема блоку GoogleNet

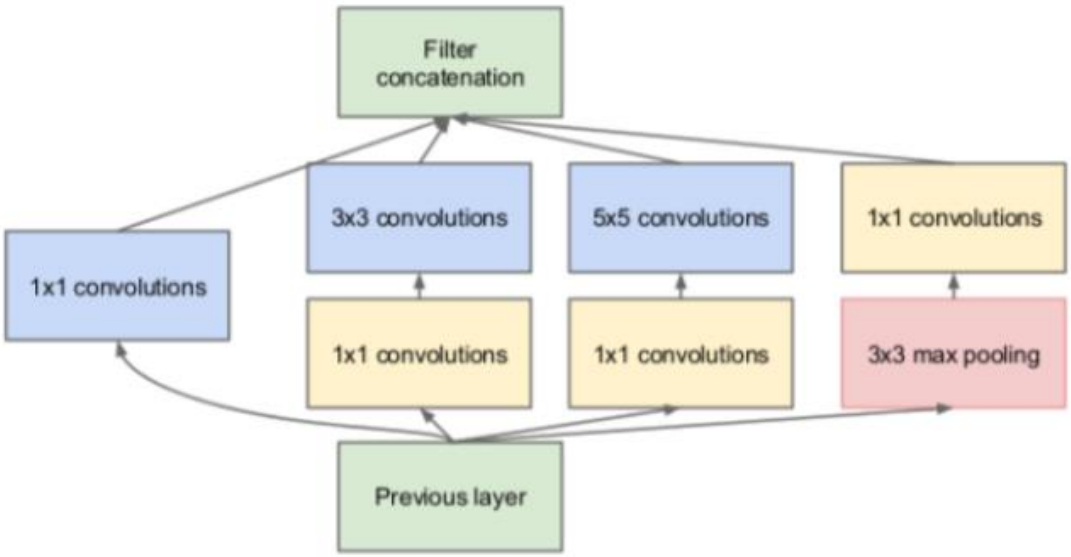

Рисунок 2.8 – Схема блоку GoogleNet

1) Подробиці навчання: GoogLeNet, реалізація на базі процесора, була навчена з використанням розподіленої машини DistBelief [32] система навчання з використанням помірної кількості моделі і даних розпаралелювання. Вони використовували асинхронний SGD з імпульсом 0,9 і постійний графік навчання. використовуючи різні

Вибірка і випадкове упорядкування вхідних зображень, вони навчили 7 ансамблів GoogLeNet з такою ж инициализацией. На відміну від AlexNet вони використовували змінене зображення 4 шкал з коротше розмірність 256, 288, 320 і 352 відповідно. загальна кількість культур на зображення становить 4 (шкали) × 3 (зліва, справа) і центральна площа / масштаб) × 6 (4 кута і центру 224 × 224 кадрування і площа змінюються до 224 × 224) × 2 (дзеркальне відображення з усіх шести культур) = 144.

2.3.4 SENet

У 2017 році була сконструёована «Squeeze-and-Excitation мережа» (SENet) [36] що стала переможцем ILSVRC-2017. Вона знизила число помилок в топ-5 до 2,25%. Її основний внесок - «Стиснення і збудження» (SE), як показано на рисунку 2.9. Тут Ftr: X → U є згорткова операція. Функція стиснення (Fsq) виконує середній пул на окремому каналі карти об'єктів U і створити дескриптор размерного каналу 1 × 1 × C. функція збудження (Fex) являє собою самозатворний механізм, що складається з з трьох шарів - два повністю пов'язаних шару і шар нелінійності ReLU між ними. Приймає стислий висновок і отримати ваги модуляції на канал. Застосовуючи збуджений висновок на карті об'єктів U, U масштабується (Fscale) генерувати остаточний висновок (Xe) блоку SE.

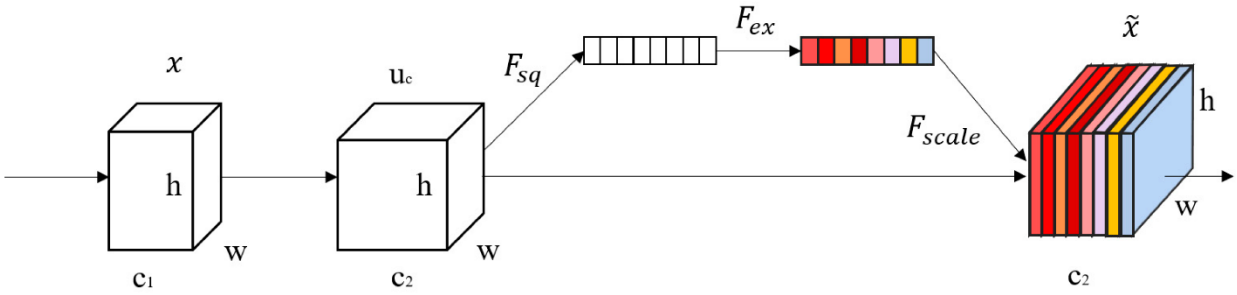

Рисунок 2.9 - Блок стиснення і збудження

Цей блок SE можна скласти докупи, щоб зробити SENet, який узагальнювати різні дані дуже добре. Автори розробили різні SENets, включаючи ці блоки в кілька складних Моделі CNN, такі як VGGNet [31],

GoogLeNet [32], ResNext (Варіант ResNet), Inception-ResNet, MobileNet, ShuffleNet.

1) Деталі навчання: автори навчені і протестовані їх модельні варіанти на ImageNet, CIFAR-10 і CIFAR-100.

Вони навчили оригінальні моделі CNN і ці моделі з блоками SE, і порівняйте компроміс між швидкістю і точністю. Они показали, що їх моделі перевершують оригінальні моделі з трохи збільшуючи час навчання / тестування.

2.4 Порівняння та вибір архітектури

У таблиці 2.3 показано порівняльні показники різніих CNN (від AlexNet до DenseNet) в наборі даних ImageNet, а саме процент помылок в наборі даних перевірки (валідація) Топ-1 і топ-5 та на топ-5 тестовому наборі даних.

У цьому розділі ми обговорили досягнення CNN в задачах класифікації зображень. Ми показали тут що хоча AlexNet, ZFNet і VGGNet слідували архітектурі традиційної моделі CNN, такої як LeNet-5 їх мережі більше і глибше. Ми переконалися що комбінація inception модулів із звичайною моделю CNN, як в GoogLeNet і ResNet отримали кращу точність, ніж повторна використання тих же будівельних блоків знов і знов. DenseNet зосередився на повторному використанні функції для посилення особливість поширення. Незважаючи на те, що CapsNet досягла найсучасніших досягнень в MNIST [38], вона ще не досягла достатньої продуктивності CNN для набору даних зображень з високою роздільною здатністю такі як ImageNet. Результат SENet для набору даних ImageNet дає нам надію, що вона може виявитися корисною для інших завдань що вимагають сильних дискримінаційних ознак.

Тож для реалізації моделі за основу було взято архітектура сімейства GoogLeNet, а саме inception v3 (рис. 2.6).

Таблиця 2.3 - Порівняльна продуктивність різних конфігурацій CNN.

Назва та рік	Тип мережі	Кількість шарів	топ-1 (валідація)	Топ-5 (валідація)	Топ-5 (тест)
AlexNet 2012	1 CNN	8	40.7%	18.2%	-
	5 CNN	-	38.1%	16.4%	16.4%
	1 CNN	-	39.0%	16.6%	-
	7 CNN	-	36.7%	15.4%	15.3%
ZFNet 2013	1 CNN	8	38.4 %	16.5%	-
	5 CNN - (a)	-	36.7 %	15.3%	15.3%
	1 CNN with layers 3, 4, 5: 512, 1024, 512 maps-(b)	-	37.5 %	16.0%	16.1%
	6 CNN, combination of (a)&(b)	-	36.0 %	14.7%	14.8%
VGGNet 2014	ensemble of 7 ConvNets (3-D, 2-C & 2-E)	-	24.7%	7.5%	7.3%
	ConvNet-D(multi-crop& dense)	16	24.4 %	7.2%	-
	ConvNet-E(Multi-crop& dense)	19	24.4 %	7.1%	-
	ConvNet-E(Multi-crop& dense)	19	24.4 %	7.1%	7.0%
	Ensemble of multi-scale ConvNets D&E (multi-crop & dense)	-	23.7%	6.8%	6.8%
GoogLe-Net 2014	1 CNN with 1 crop	22	-	-	10.07 %
	1 CNN with 10 crops	-	-	-	9.15%
	1 CNN with 144 crops	-	-	-	7.89%
	7 CNN with 1 crop	-	-	-	8.09%
	1 CNN with 10 crops	-	-	-	7.62%
	1 CNN with 144 crops	-	-	-	6.67%
ResNet 2015	ResNet-18	18	27.88%	-	-
	plain layer	18	27.94%	-	-
	Plain layer	34	28.54%	10.02	-
	ResNet-34 (zero-padding shortcuts), 10 crop testing -(a)	34	25.03%	7.76%	-

Продовження таблиці 2.3

Назва та рік	Тип мережі	Кількість шарів	топ-1 (валідація)	Топ-5 (валідація)	Топ-5 (тест)
ResNet 2015	ResNet-34 (projection shortcuts to increase dimension, others are identity shortcuts), 10 crop testing-(b)	34	24.52%	7.46%	-
	ResNet-34 (all shortcuts are projection), 10 crop testing-(c)	34	24.52%	7.46%	-
	ResNet-50 (with bottleneck layer), 10 crop testing	50	22.85%	6.71%	-
	ResNet-101 (with bottleneck layer), 10 crop testing	101	21.75%	6.05%	-
	ResNet-152 (with bottleneck layer), 10 crop testing	152	21.43%	5.71%	-
	ResNet-34 (b)	34	21.84%	5.71%	-
	ResNet-34 (c)	34	21.53%	5.60%	-
	ResNet-50	50	20.74%	5.25%	-
	ResNet-101	101	19.87%	4.60%	-
	ResNet-152	152	19.38%	4.49%	-
	Ensemble of 6 models	-	-	-	3.57%
DenseNet 2016	DensNet-121 +	121	23.61%	6.66%	-
	DenseNet-169 +	169	22.80%	5.92%	-
	DenseNet-201 +	201	22.58%	5.54%	-
	DenseNet-264 +	264	20.80%	5.29%	-
SENet 2017	SE-ResNet-50	50	23.29%	6.62%	-
	SE-ResNext-50	50	21.10%	5.49%	-
	SENet-154 (crop size 320 × 320/299 × 229)	-	17.28%	3.79%	-
	SENet-154(crop size 320 × 320)	-	16.88%	3.58%	-

3 ПРОЦЕС РОЗРОБКИ ТА ТЕСТУВАННЯ

3.1 Опис проблеми

За даними Американського онкологічного товариства і Центру статистики раку [n-1], щорічно захворюють понад 150000 пацієнтів з раком легенів (очікується 154 050 на 2018 рік), в той час як щороку діагностується ще 200000 нових випадків (очікується 234 030 для 2018). Це один з найпоширеніших видів раку в світі через не тільки куріння, але і впливу токсичних хімічних речовин, таких як радон, азбест і миш'як. LUAD і LUSC є двома найбільш поширеними типами недрібноклітинного раку легенів [39], і кожен з них пов'язаний з окремими посібниками по лікуванню. За відсутності певних гістологічних ознак це важлива відмінність може бути складним і трудомістким, і вимагає наявності підтверджуючих иммуногистохимических плям. Класифікація типу раку легені є ключовим діагностичним процесом, оскільки доступні варіанти лікування, включаючи звичайну хіміотерапію і останнім часом таргетная терапія відрізняється для LUAD і LUSC [63]. Крім того, діагноз LUAD спонукає до пошуку молекулярних біомаркерів і сенсибилизирующих мутацій і, таким чином, матиме великий вплив на варіанти лікування [41]. Наприклад, мутації рецептора епідермального фактора росту (EGFR), присутні приблизно в 20% LUAD, і перебудови анапластичної рецепторной лімфоми тирозинкінази (ALK), присутні в <5% LUAD [42], в даний час мають таргетной терапію, схвалену Food and Управління по лікарських засобів (FDA) [43]. Мутації в інших генах, таких як KRAS і пухлинний білок P53 (TP53), зустрічаються дуже часто (близько 25% і 50% відповідно), але до сих пір виявилися особливо складними мішенями для ліків [42]. Біопсія легень зазвичай використовується для діагностики типу та стадії раку легенів. Віртуальна мікроскопія забарвлених зображень тканин зазвичай виходить при збільшенні від 20 × до 40 ×, створюючи дуже великі двовимірні зображення (від 10000 до > 100000 пікселів в кожному вимірі), які часто складно провести візуальним оглядом вичерпним чином. Крім того, точна

інтерпретація може бути утруднена, і відмінність між LUAD і LUSC не завжди ясно, особливо в погано диференційованих пухлинах; в цьому випадку для точної класифікації рекомендуються додаткові дослідження [44]. Щоб допомогти експертам, недавно був вивчений автоматичний аналіз повних зображень раку легенів для прогнозування результатів виживання і класифікації [45]. В останньому випадку Yu et al. [45] об'єднали традиційні методи визначення порогу і обробки зображень з методами машинного навчання, такими як класифікатори випадкових лісів, машини опорних векторів (SVM) або наївні байєсовські класифікатори, досягнувши ППК ~ 0,85 при розрізненні нормального і пухлинного слайди, і ~ 0,75 на відміну від слайдів LUAD і LUSC. Зовсім недавно для класифікації пухлин молочної залози, сечового міхура і легенів було використано глибоке навчання, досягнувши ППК 0,83 при класифікації типів пухлин легенів на предметних стеклах з Атласу генома раку (TCGA) [46]. Аналіз значень ДНК в плазмі також виявився хорошим предиктором наявності недрібноклітинного раку з ППК ~ 0,94 [47] при розрізненні LUAD від LUSC, тоді як використання імунохімічних маркерів дає ППК ~ 0.941 [48]. Ось, ми демонструємо, як область може ще більше отримати вигоду з глибокого навчання, представляючи стратегію, засновану на згортальних нейронних мережах (CNN), яка не тільки перевершує методи в раніше опубліковані роботи, але також досягає точності, які можна порівняти з патологами. Найбільш важливим є те, що наші моделі зберігають свої експлуатаційні якості при тестуванні на незалежних наборах даних як заморожених, так і залитих в малин тканин з парафіном (FFPE), а також на зображеннях, отриманих з біопсій. Розвиток нових, недорогих і більш потужних технологій (зокрема, графічних процесорів (GPU)) уможливило навчання більших і складніших нейронних мереж [49]. Це призвело до розробки декількох глибоких CNN, які здатні виконувати складні завдання візуального розпізнавання. Такі алгоритми вже успішно використовувалися для сегментації [50] або класифікації медичних зображень [51] і, більш конкретно, для додатків з цілими слайдами зображень, таких як виявлення ядер [52], сегментація

ниркової тканини [49] і локалізація клубочків [53], діагностика раку молочної залози [54], аналіз пухлин товстої кишки [55], класифікація гліоми при пухлинах головного мозку [56] ідентифікація епітеліальної тканини при раку передміхурової залози [57] і діагностика остеосаркоми [58].

CNN також були вивчені з точки зору класифікації патернів легкого при комп'ютерної томографії (КТ), досягнувши f-показника ~ 85,5%. Для вивчення автоматичної класифікації повних слайдів зображень раку легкого була використана архітектуру inception v3 і повні слайди зображень гематоксилин-еозіновой (H&E) тканини легені з TCGA, отриманих хірургічним видаленням з подальшої підготовки замороженого зрізу. Як було сказано в предведущей чолі - в 2014 році Google переміг в конкурсі ImageNet по великомасштабного візуальному розпізнавання, розробивши архітектуру GoogleNet, яка підвищила стійкість до перекладу і здатності до нелінійного навчання завдяки використанню елементів мікроархітектури, званих inception. Кожен блок inception включає в себе кілька модулів нелінійної згортки з різним дозволом. Inception архітектура особливо корисна для обробки даних в декількох дозволах, що робить цю архітектуру підходящої для задач патології. Ця мережа вже була успішно адаптована до інших конкретних типів класифікацій, таким як рак шкіри і виявлення діабетичної ретинопатії.

3.2 Постановка завдання для розроблюваного модуля

Предпологаются результат - система глибокого навчання для автоматичного аналізу зображень гістопатології. Мета проектування полягає в тому, щоб розробити модуль глибокого навчання для автоматичного аналізу слайдів пухлини з використанням загальнодоступних зображень повних слайдів, доступних в TCGA, і згодом провести тестування. Характеристики набору даних TCGA і наша загальна обчислювальна стратегія узагальнені на Рисунок 3.1-3.4. Ми використовували 1634 повних слайдів зображення з бази даних Genomic Data Commons 1176 пухлинних тканин і 459 нормальних тканин (рис. 3.1). Тисяча

шістсот тридцять чотири повних слайд-зображення були розділені на три групи: навчання, валідаційні (перевірочні) та тестові (рис. 3.2). Важливо, що це гарантує, що наша модель ніколи не буде навчатися і тестується на фрагментах, отриманих з одного і того ж зразка пухлини.

Оскільки розміри зображень всього слайда занадто великі, щоб використовувати їх в якості прямого введення в нейронну мережу (рис. 3.3), мережа замість цього була навчена, перевірена і протестована з використанням фрагментів розміром 512 × 512 пікселів, отриманих з неперекриваючіхся «патчів» цілих слайдів зображень. Це призвело до того, що на слайд доводилося від десятків до тисяч фрагментів, в залежності від вихідного розміру (рис. 3.4).

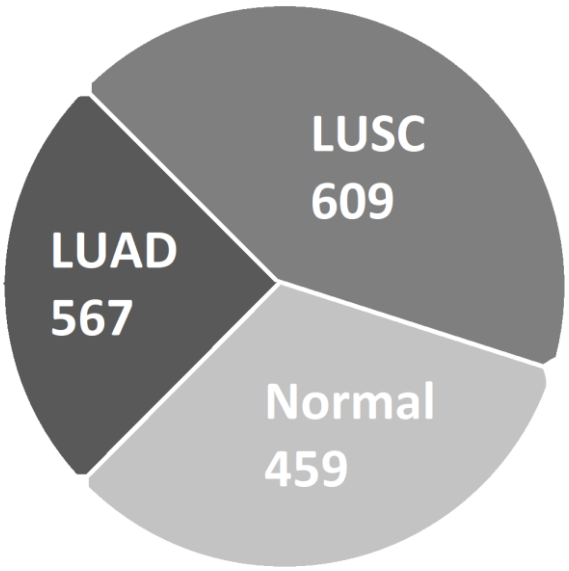

Рисунок 3.1 – Диаграма розподілу даних

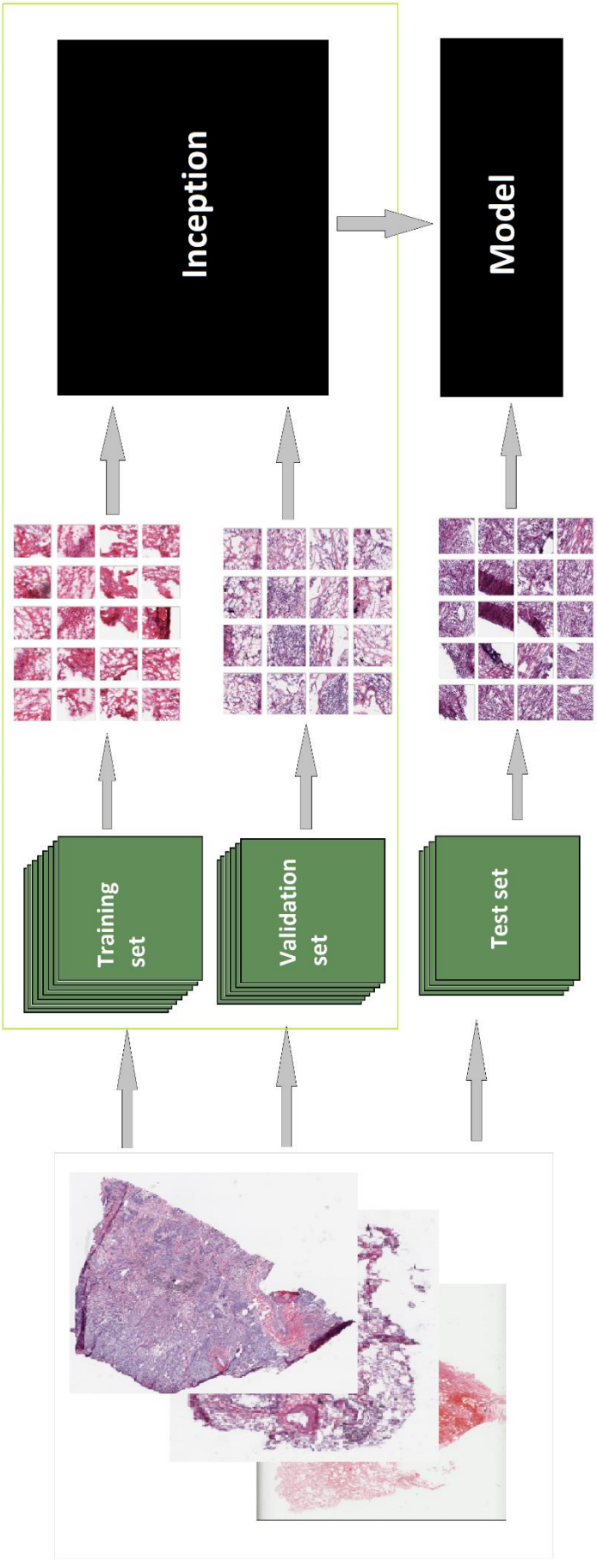

Рисунок 3.2 – Алгоритм системи

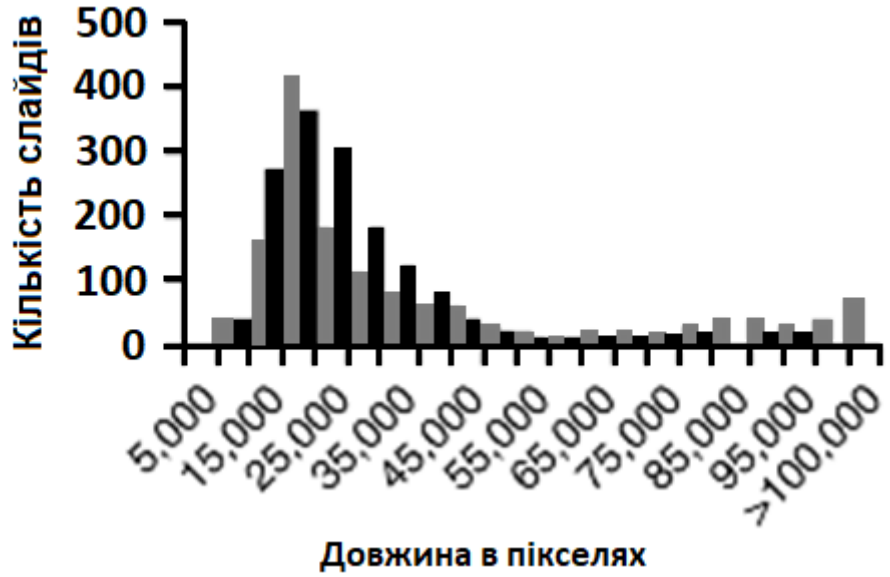

Рисунок 3.3 – Діаграма розподілу розмірностей слайдів

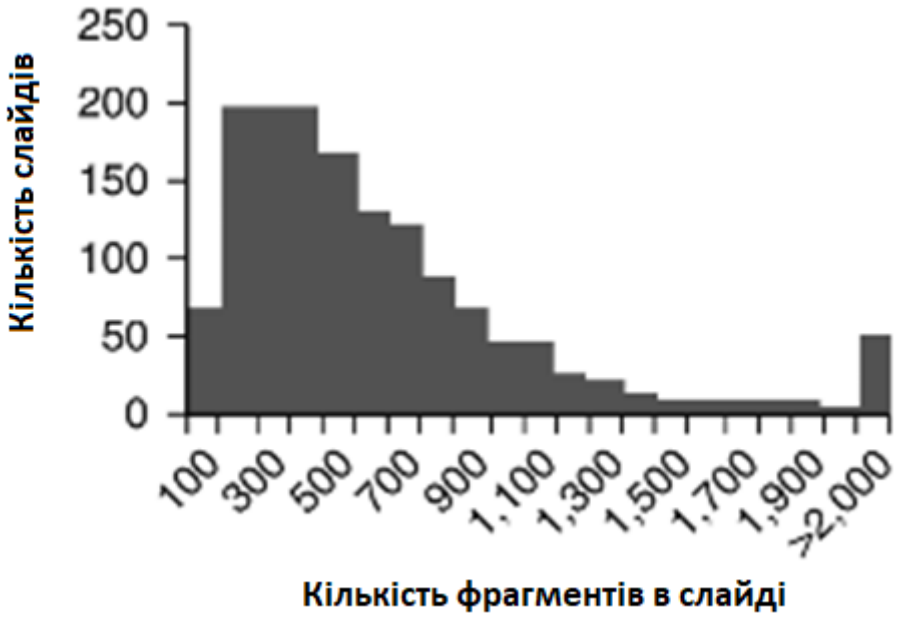

Рисунок 3.4 – Діаграма розподілу фрагментів слайдів

3.3 Результати тестування

Основні дії на вищій стратегії, викладені на Рисунок 3.1-3.4, ми представляємо два основні результати. По-перше, ми розбираємо моделі

класифікацій, які класифікують повні слайд-зображення на здорових легенях, LUAD або LUSC з високою точністю близько 0,97 і співставляють з результатами професійних патологів. Ефективність наших класифікаційних моделей була перевірена на декількох незалежних наборах даних: біопсії та образи хірургічної рецензії, підготовлені у вигляді заморожених зрізів або зрізів тканей FFPE. По-друге, починаючи з областей LUAD, як предсказує модель класифікації LUAD порівнюючи з LUSC та зі звичайною класифікацією, ми використовуємо той самий алгоритм, що зображено на рисуноках 3.1-3.4 для взуття нових моделей, щоб передбачити мутаційний статус часто мутованих генів. при наявності анокарциномів легких із використанням зображень всього слайда в якості єдиного входу. Весь робочий процес нашого вибільного аналіза суммірується на додатковій (рис. 3.1 – 3.4) моделі глубокого взуття генерують точну діагностику гістопатологічних зображень легких. Використовує обчислювальний конвеєр, показаний на рис. 3.1 – 3.4, ми сначала вчили початок v3 розташувати пухлина за порівнянням з нормальною. Для того, щоб оцінити точність набору тестів, результати класифікації за фрагмент агрегувались на основі слайди або шляхом усереднення ймовірностей, отриманих на кожному фрагменті, або шляхом підрахунку відсотків фрагментів, позитивно класифікованих, таким чином, генеруючи класифікацію на слайд. Перший підхід дав ППК 0,990, а другий - 0,993 (рис. 3.5 і таблиця 3.1) для класифікації за порівнянням з пухлиною, перетворенням ППК ~ 0,85, досягнутого підходом на основі визнання, з ~ 0,94, досягнутого аналізу ДНК плазми [47] і сопоставимим або найкращим, що містить молекулярного профілю.

Затем ми перевірили ефективність нашого підходу до більш складної задачі відмінностей LUAD та LUSC. Для того, щоб зробити це, ми спершу перевірили, можливо, вони перетворили нейтронні мережі, перетворившись на опублікований підхід, заснований на функції, навіть при використанні простого переносного взуття. Для цього знання останніх рівнів inception v3 - більш раннє взуття в наборі даних ImageNet для ідентифікації 1000 різних класів - були ініціалізаторами випадкової системи, і забуті взуття для наших задач

класифікацій. Після зведених статистичних даних для кожної людини слайди (рис. 3.6) цей процес прийшов до ППК 0,847 (таблиця 3.1); то є, прискорення ППК ~ 0,1 за порівнянням з найкращими результатами, отриманими з використаними характеристиками зображення у форматі з випадковим класифікатором. Виробництво може бути додатково вдосконалене путем повного взуття, що створюється, що приводить до ППК 0,950, коли агрегація виконує справжнє використання серед потенційних можливостей для кожного елементу мозаїчного зображення (рис. 3.7). Ети значень ППК покращують ще на 0,002, коли фрагменти, більш ранні класифіковані як «нормальні» за першим класифікатором, не включаються в процес агрегації (таблиця 3.1). Ми додатково оцінили моделі продуктивності глубокого взуття, вбрання та тестування переходять до прямої трехсторонній класифікації на три типи зображень (звичайні, LUAD, LUSC). Такий підхід дозволив отримати найвищу продукцію при вдосконаленні всіх ППК за крайньою мірою до 0,968 (таблиця 3.1). Помімо роботи з фрагментми з 20-кратним збільшенням, ми дослідили вплив збільшення і поля зрізання фрагментів на виробництві наших моделей. Належно до особливостей низького розрізнення (гнезда клеток, круглі узори) також можна бути полезними для класифікуючих типів легких, ми використали слайди, показавши більше поля полегшення, для взуття моделей створили мозаїчні рисунок розміром 512 × 512 пікселів. при 5-кратному збільшенні. Двоєчні та трехсторонні мережі, взуті на таких слайдах, привели до аналогічних результатів (таблиця 3.1). На додатковому рисунок 11, 12 та в додатковій таблиці 2 обобщених та встановлених характеристик різних підводів, розглянутих у цьому дослідженні та в попередній роботі.

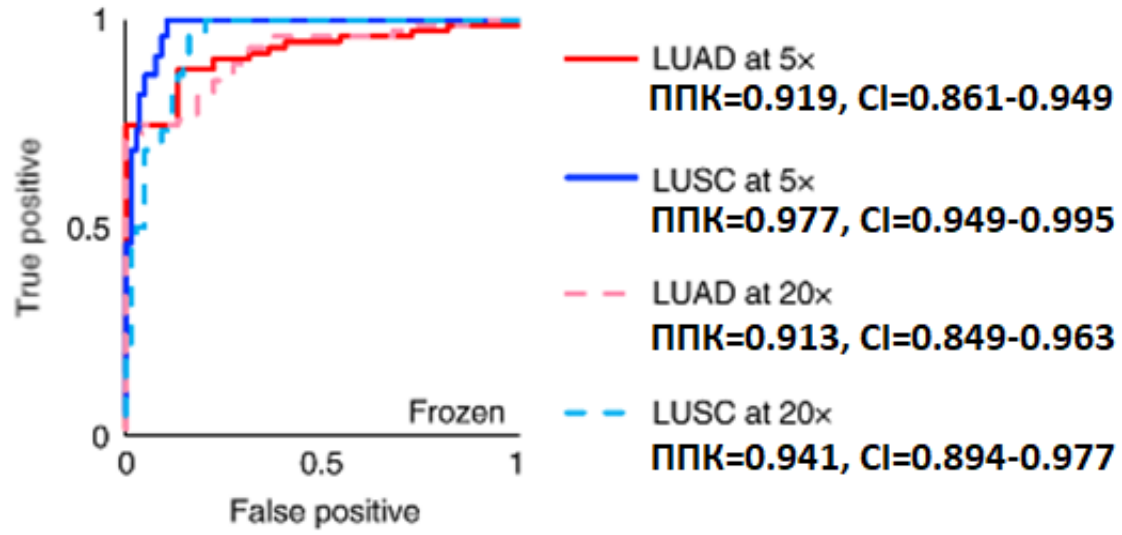

Рисунок 3.5 – Результати класіфікації зображень заморожених тканин

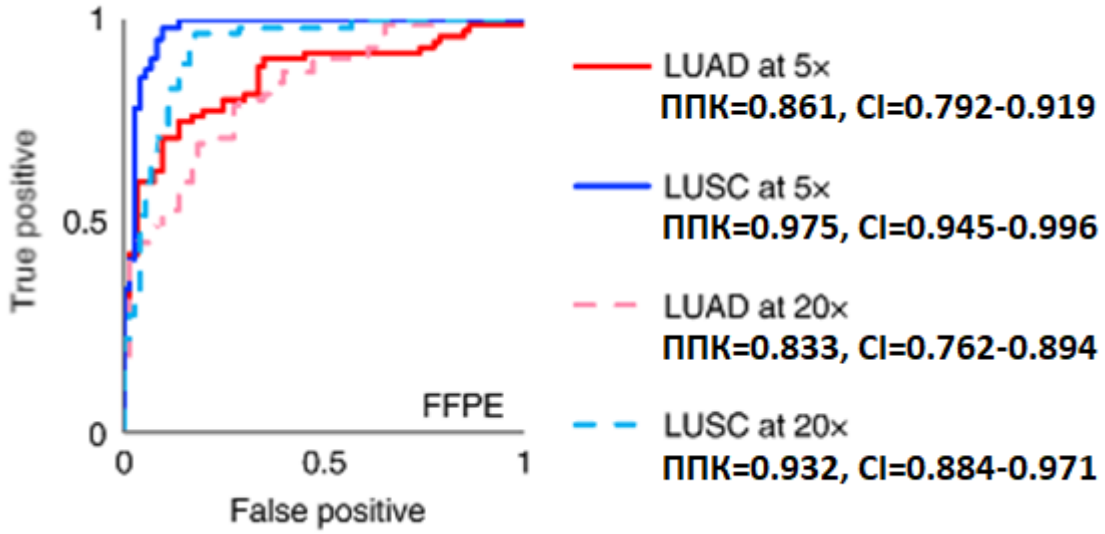

Рисунок 3.6 – Результати класіфікації зображень тканин оброблених методом FFPF

Що стосується тимчасових витрат, патологоанатома може знадобитися від одного до декількох хвилин, щоб проаналізувати предметне скло в залежності від складності розпізнавання кожного випадку. Крім того, при відсутності певних гістологічних ознак потрібні підтверджують імуногістохімічні плями, які можуть затримувати діагностику на термін до 24 год. Час обробки слайда нашим алгоритмом залежить від його розміру; в даний час для розрахунку ймовірностей

класифікації на 500 фрагментів потрібно близько 34с (середнє число фрагментів на слайді <500) в одному графічному процесорі GTX 1060. З огляду на можливість використання декількох графічних процесорів для паралельної обробки фрагментів, класифікація з використанням нашої моделі може бути виконана за кілька секунд. Час сканування кожного слайда з використанням сканера Aperio (Leica) в даний час складає 2-2,5 хвилини для слайда зі швидкістю 20×, але зі схваленням FDA 2017 року нового надшвидкого цифрового сканера патології від Philipps [59] цей крок, ймовірно, буде більше не буде вузьким місцем в найближчому майбутньому.

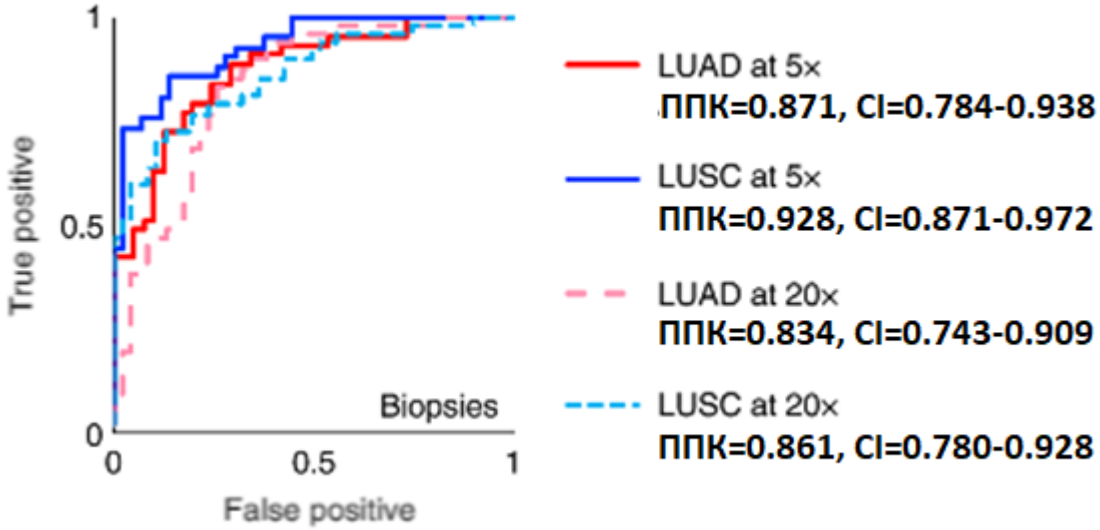

Рисунок 3.7 – Результати класіфікації зображень тканин після біопсії

Тестування на незалежних когортах демонструє узагальненість моделі нейронної мережі. Потім модель була оцінена на незалежних наборах даних зображень повних слайдів раку легкого, отриманих з заморожених зрізів (98 слайдів) і зрізів FFPE (140 слайдів), а також біопсії легенів (102 слайда), отриманих в Langone Medical при Нью-Йоркському університеті (NYU). Центр (рис. 5 - 8). В цьому випадку діагноз патологів, заснований на морфології і доповнений імуногістохімічними плямами (TTF-1 і p40 для LUAD і LUSC відповідно) при необхідності, використовувався в якості золотого стандарту

(тобто використовувався в якості основного фактора для оцінки ефективності з нашого підходити). Кожне зображення TCGA складається майже виключно з клітин LUAD, клітин LUSC або нормальної тканини легені. В результаті кілька зображень в двох нових наборах даних містять ознаки, які мережа не була навчена розпізнавати, що робить задачу класифікації складнішою. Ми спостерігали, що особливості, включаючи згусток крові, кровоносні судини, запалення, некротичні області і області коллапсірующая легкого, іноді позначають як LUAD, бронхіальний хрящ іноді позначають як LUSC, а фіброзні рубці можна неправильно класифікувати як нормальні або LUAD. Як показано на додатковій рис. 14, зображення TCGA мають значно більш високий вміст пухлини в порівнянні з незалежними наборами даних, і зміст пухлини корелює зі здатністю алгоритму узагальнювати ці нові невидимі зразки. Щоб зменшити зміщення, що породжується деякими з цих специфічних особливостей, які знаходяться за межами областей пухлини, і перевірити тільки здатність нашої мережі диссаціювати області LUAD, LUSC і нормальних тканин, ППК на рис. 5 - 13 були розраховані для областей з високим рівнем зміст пухлини, які були вручну відібрані патологом. З огляду на, що на деяких старих слайдах також були виявлені нові типи артефактів (тьмяне забарвлення, нерівномірне фарбування, бульбашки повітря під кришкою слайда, що призводять до можливих спотворень), результати, отримані на цих незалежних когортах, дуже обнадіюють. При збільшенні в 20 разів більша кількість фрагментів повністю покривається деякими з цих «невідомих» ознак, тоді як при збільшенні в 5 разів поле зору більше і містить елементи, відомі по класифікованих (пухлинні або нормальні клітини) безлічі інших фрагментів, що дозволяє більш точна класифікація на фрагменті. Це, в свою чергу, призводить до більш точної класифікації для кожного слайда. Взяті разом, ці спостереження можуть пояснити, чому ППК класифікатора на фрагментах з 5-кратним збільшенням в основному вище, ніж на фрагмента з 20-кратним збільшенням. Цікаво, що хоча слайди із секцій FFPE і біопсії були збережені з використанням методу, відмінного від тих, що були в базі даних TCGA, продуктивність залишається

задовільною (Рис. 6). Що стосується біопсій, ми помітили, що низька продуктивність була пов'язана не тільки з регіонами, де фіброз, запалення або кров також були присутні, але також в дуже погано диференційованих пухлинах. Зрізи, отримані з біопсій, зазвичай набагато менше, що зменшує кількість фрагментів на слайді, але продуктивність нашої моделі залишається незмінною для 102 протестованих зразків (ППК ~ 0,834-0,861 при 20-кратному збільшенні і 0,871-0,928 при 5-кратному збільшенні; Рис. 8), і точність класифікації не корелює з розміром вибірки (Рис. 4) (r2 = 9,5 * 10 ^ -5). Примітно, що при поділі набору даних ми помітили, що наша модель здатна також класифікувати ці складні випадки: при 20-кратному збільшенні ППК для LUAD і LUSC для цих складних випадків становили 0,809 (довірчий інтервал (ДІ), 0,639-0,940) і 0,822 (ДІ, 0,658-0,951), відповідно, що лише трохи нижче, ніж слайди, які вважаються очевидними для патологів (для LUAD, ППК 0,869 (ДІ, 0,753-0,961) і для LUSC, 0,883 (ДІ, 0,777-0,962)).

Нарешті, ми перевірили, чи можна замінити ручну процес вибору пухлини автоматичним комп'ютерним вибором. Наприклад, щоб перевірити ефективність моделі відбору пухлин на біопсіях, ми навчили модель розпізнавати область пухлини на заморожених зразках і зразках FFPE, потім застосували цю модель до біопсії і, нарешті, застосували навчений TCGA тристоронній класифікатор до області пухлини, обраної за допомогою моделі автоматичного вибору пухлини. ППК на фрагменту моделі автоматичного відбору пухлин (використовуючи вибір пухлини патолога як еталон) становив 0,886 (ДІ, 0,880-0,891) для біопсій, 0,797 (ДІ, 0,795-0,800) для заморожених зразків і 0,852 (ДІ, 0,808-0,895) для зразків FFPE. Як показано на додатковому рисунку 14 (крайній правий стовпчик кожного графіка), ми спостерігали, що автоматичний вибір приводив до продуктивності, порівнянної з ручним вибором (трохи краще ППК в замороженому режимі, без відмінностей в FFPE і трохи гірше в біопсії, див. Також додатковий рис. 15). Прогнозування мутаційного статусу гена по зображеннях з повних слайдів. Потім ми сфокусувалися на слайдах LUAD і перевірили, чи можна навчити CNN передбачати мутації генів, використовуючи

зображення в якості єдиного входу. Для цього з TCGA були завантажені дані про генні мутації для відповідних зразків пацієнтів. Щоб переконатися, що навчальний і тестовий набори містять досить зображень з мутують генами, ми відібрали тільки ті, які були мутував принаймні в 10% доступних пухлин. На кожному слайді LUAD для цього завдання використовувалися тільки фрагменти, що класифіковані як LUAD нашої класифікаційної моделлю, щоб уникнути зміщення мережі для вивчення специфічних для LUAD і специфічних для LUSC мутацій і замість цього зосередитися на розрізненні мутацій, які слід виключно на фрагменти LUAD. Inception v3 був змінений, щоб дозволити багатоканальна класифікація: навчання та перевірка були проведені на ~ 212 000 фрагментів з ~ 320 слайдів, а тестування - на ~ 44 000 фрагментів з 62 слайдів. Значення ППК для серин/треонін-протеїнкінази 11 (STK11), EGFR, нетипового кадгерінов 1 FAT (FAT1), зв'язуючого білка SET 1 (SETBP1), KRAS і TP53 були між 0,733 і 0,856 (таблиця 3.1). Очікується, що наявність більшої кількості даних для навчання істотно поліпшить показники. Як згадувалося раніше, EGFR вже має цільову терапію. STK11, також відомий як печінкова киназа 1 (LKB1), є пухлинним супрессором, інактивованих в 15-30% недрібноклітинного раку легені [60], і також є потенційною терапевтичної мішенню: повідомлялося, що фен-формин, мітохондріальний інгібітор, збільшує виживаність у мишей [37]. Також було показано, що мутації STK11 в поєднанні з мутаціями KRAS приводили до більш агресивним пухлин [61]. FAT1 є Ортолог гена жиру дрозофіли, який бере участь у багатьох типах раку, і його інактивація, як передбачається, збільшує зростання ракових клітин [62]. Вважається, що мутація гена-супресора пухлини TP53 призводить до стійкості до хіміотерапії, що призводить до зниження виживаності при дрібноклітинному раку легені.

Мутація в SETBP1, подібно KEAP1 і STK11, була ідентифікована як одна з сигнатур LUAD [64]. Нарешті, для кожного гена ми порівняли класифікацію, досягнуту нашою моделлю глибокого навчання, з частотою алелі (рис. 3c). Серед генних мутацій, передбачених з високим ППК, в чотирьох з них ймовірності класифікації (згідно з нашою моделі) були пов'язані з частотою алелі: FAT1,

KRAS, SETBP1 і STK11, демонструючи, що ці ймовірності можуть відображати відсоток клітин ефективно залежить від мутації. Подивившись, наприклад, на прогнози, виконані для всього слайд-зображення з рис. 4a, наш процес успішно ідентифікував TP53 (частота алелей 0,33) і STK11 (частота алелей 0,25) як два гена, які найбільш ймовірно мутували (рис. 4a). Теплова карта показує, що майже всі фрагменти LUAD, за прогнозами, демонструють ознаки, подібні TP53-мутанту (рис. 4b), і дві основні області з ознаками, подібними STK11-мутантів (рис. 4c). Цікаво, що коли класифікація застосовується до всіх мозаїчним елементам, це показує, що навіть мозаїчні елементи, класифіковані як LUSC, представляють мутації TP53 (рис. 4d), тоді як мутант STK11 обмежений фрагментами LUAD (рис. 4e). Ці результати реалістичні, враховуючи, що, як згадувалося раніше, мутація STK11 є ознакою LUAD [64], в той час як мутація TP53 частіше зустрічається при всіх ракових захворюваннях людини.

Подальша робота над інструментами візуалізації моделі з глибоким ухилом [64] допоможе визначити та охарактеризувати функції, які використовуються нейронною мережею.

Хоча наш поточний аналіз не визначає конкретні особливості, які використовуються мережею для виявлення мутацій, наші результати наводять на думку. що такі кореляції генотип-фенотип обнаружіми. Визначення статусу мутації за гістологічною зображенню і обхід додаткового тестування важливі, зокрема, при раку легені, так як ці мутації часто несуть як прогностичну, так і прогностичну інформацію. Зовсім недавно Chiang et al. [66] експериментально продемонстрували зв'язок між визначальною мутацією і унікальною морфологією підтипу раку молочної залози. Було показано, що деякі мутації з високими ППК, виділені в нашому дослідженні (наприклад, мутації в STK11, TP53 і EGFR), впливають на полярність клітин і форму клітин 49-51 - дві особливості, які зазвичай не оцінюються під час патологічного діагнозу. Ми відзначаємо, що наша модель не була здатна виявляти мутації ALK, хоча такі пухлини були пов'язані зі специфічними гістологічними ознаками, такими як суцільна структура з клітинами персня з кільцем або слизисто-муцинозная

форма 52,53. Хоча поширеність мутацій ALK дуже низька (за повідомленнями, в межах 1,8-6,4% [67]), їх наявність зазвичай визначається за допомогою імуногістохімії, так як пухлини з цією мутацією можуть реагувати на інгібітори ALK6,7.

Таблиця 3.1 - ППК досягнута мережею, навченої мутацям (з 95% ДИ)

Мутація	На фрагмент (тайл) ППК	На слайд ППК	
		средня прогнозована ймовірність	відсоток позитивно класифікованих фрагментів
STK11	0.845 (0.838–0.852)	0.856 (0.709–0.964)	0.842 (0.683-0.967)
EGFR	0.754 (0.746–0.761)	0.826 (0.628–0.979)	0.782 (0.516-0.979)
SETBP1	0.785 (0.776–0.794)	0.775 (0.595–0.931)	0.752 (0.550–0.927)
TP53	0.674 (0.666–0.681)	0.760 (0.626–0.872)	0.754 (0.627–0.870)
FAT1	0.739 (0.732–0.746)	0.750 (0.512–0.940)	0.750 (0.491–0.946)
KRAS	0.814 (0.807–0.829)	0.733 (0.580–0.857)	0.716 (0.552–0.854)
KEAP1	0.684 (0.670–0.694)	0.675 (0.466–0.865)	0.659 (0.440–0.856)
LRP1B	0.640 (0.633–0.647)	0.656 (0.513–0.797)	0.657 (0.512–0.799)
FAT4	0.768 (0.760–0.775)	0.642 (0.470–0.799)	0.640 (0.440–0.856)
NF1	0.714 (0.704–0.723)	0.640 (0.419–0.845)	0.632 (0.405–0.845)

Щоб підтвердити, що наші моделі можуть бути застосовані до незалежних когортам, ми перевірили прогнозування мутанта EGFR, використовуючи 63 зображення повних слайдів зразків резекції легені з відомим мутаційним статусом EGFR: 29 мутантів EGFR і 34 зразки EGFR дикого типу. Цей незалежний набір даних має деякі важливі відмінності від набору даних TCGA, які можуть негативно вплинути на оцінку моделі на основі TCGA: (i) зразки були заморожені, а замість цього були збережені з використанням FFPE, і (ii) тільки 22 зразків були послідовність отримання даних для підтвердження мутаційного статусу EGFR з високою специфічністю і чутливістю; інші зразки (тобто 65%

тестового набору) були проаналізовані за допомогою імуногістохімічних (IHC) пятен55, методики, відомої своєю високою специфічністю, але низькою чутливістю 56,57 і яка ідентифікує тільки дві найбільш поширені мутації55 (p.L858R і p.E746_A750del) . З іншого боку, дані з набору даних TCGA, використаного для навчання, були ідентифіковані за допомогою інструментів секвенування наступного покоління (NGS) Illumina HiSeq 2000 або Genome Analyzer II. Тому наша модель TCGA була навчена розпізнавати не тільки p.L858R і p.E746_A75-del, але і багато інших мутанти і делеции EGFR, такі як, наприклад, p.G719A, p.L861Q або p.E709_T710delinsD. Незважаючи на ці застереження, ми вважали, що все ще буде важливо продемонструвати, що наші моделі, засновані на TCGA, можуть, принаймні, працювати значно краще, ніж випадкові, в незалежній когорти NYU. Дійсно, результати показали ППК 0,687 (ДІ 0,554-0,811), причому більш високий ППК (0,750; ДІ 0,500-0,966) в зразках, підтверджених секвенуванням, ніж в тестованих IHC (ППК, 0,659; ДІ 0,485-). 0,826). Хоча заснований на секвенування ППК, рівний 0,75, нижче, ніж той, який був оцінений в тестовому наборі TCGA (0,83), ми вважаємо, що більша частина цієї різниці може бути пов'язана з різницею в підготовці зразка (заморожений в порівнянні з FFPE). Ми помітили, що розбіжність (~ 0,08) схоже на різницю, що спостерігається в ППК LUAD з набору даних TCGA (0,97) і набору даних FFPE (0,83). У задачі класифікації ця проблема була вирішена шляхом зменшення збільшення до 5×. Однак це марно для завдання передбачення мутації, тому що, по-видимому, в 20 разів необхідно захопити ознаки предсказательной зображення (модель передбачення мутації TCGA EGFR в 5 разів має випадкове спектакль). Проте, ми вважаємо, що 0,75 ППК, які ми отримали для перевіреного секвенування підмножини EGFR-мутантів, демонструє, що модель може узагальнювати незалежні набори даних.

4 ОПИС РОЗРОБЛЕНОЇ СИСТЕМИ

4.1 Структура та інтерфейс

Ни рисунке 4.1 изображена схема вариантов использования разработанного модуля.

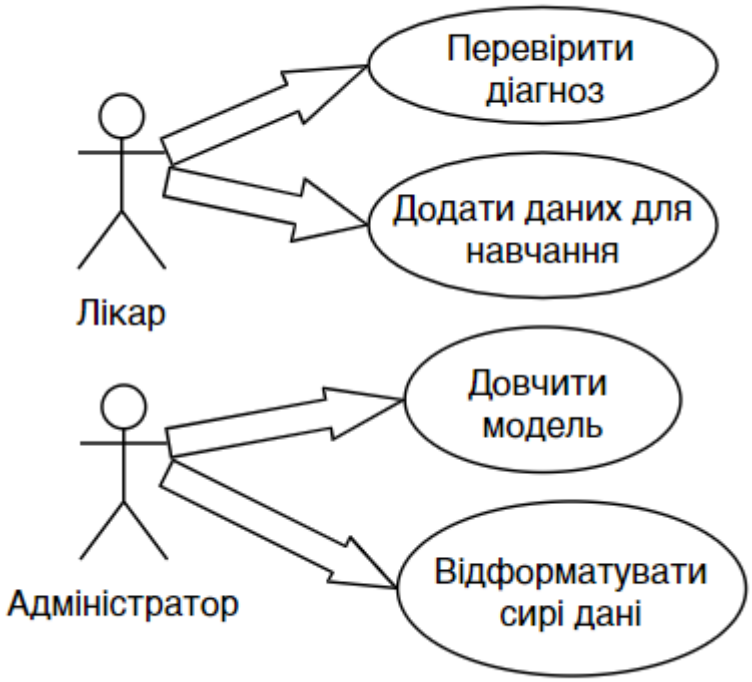

Рисунок 4.1 – Диаграма Use-Case

Структурна схема системи в цілому представлена на рисунку 4.2. У поточній роботі було реалізовано мудуль розпізнавання зображень тканин легенів як один з багатьох модулів системи, що може бути вбудованим у будь яке середовище від web-додатку до професійного програмного продукту що вікористовується безпосередньо у лікарнях при аналізі тканин у лабораторіях.

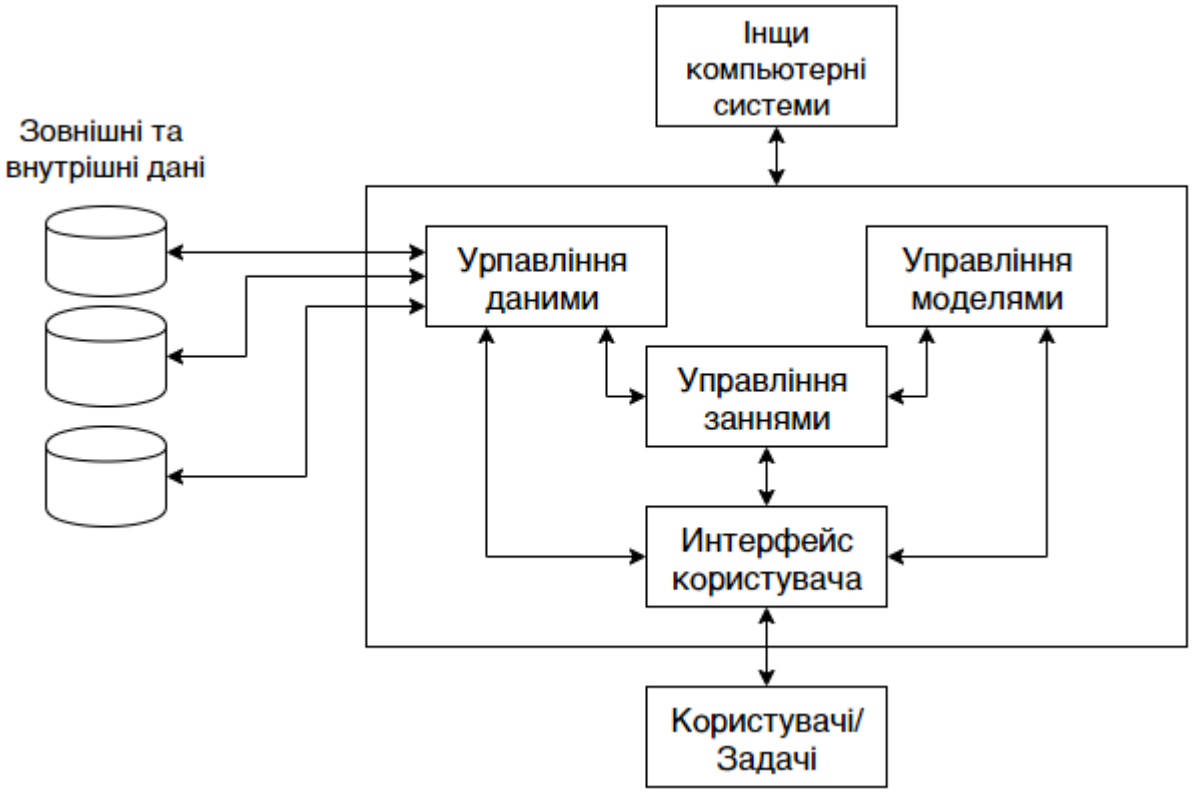

Рисунок 4.2 – Структурна схема системы

4.2 Вхідні та вихідні параметри

Як було сказано у розділі 3, для обучення та тестування системы були використовані слайди тканини легень вісокої розділової спроможності (TCGA) що були порізані на фрагменти 512 на 512 пікселів. На рисунках 4.3 – 4.5 представленій візуальне представлення цих данних. Кожен слайд має власній уникальний номер (назва директорії рис. 4.3) для синхронізації з документов опису слайдів (рис. 4.5). В кожній директорії знаходяться близько 500 фрагментів первинного слайду (рис. 4.4).

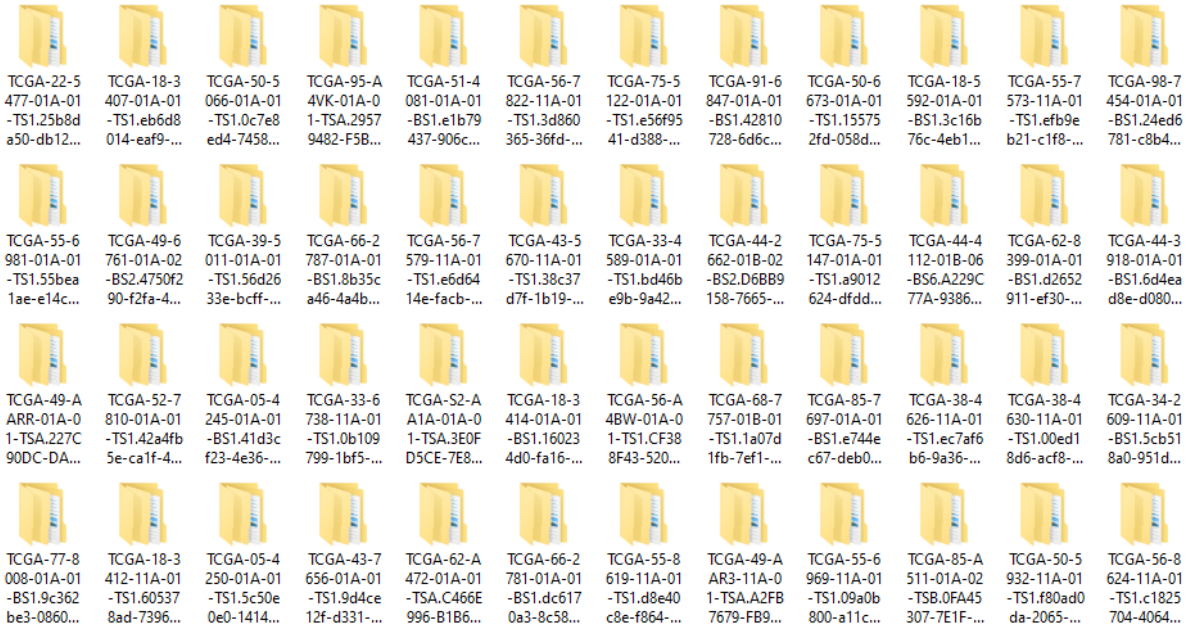

Рисунок 4.3 – Директорія з підготовленими тестовими данними

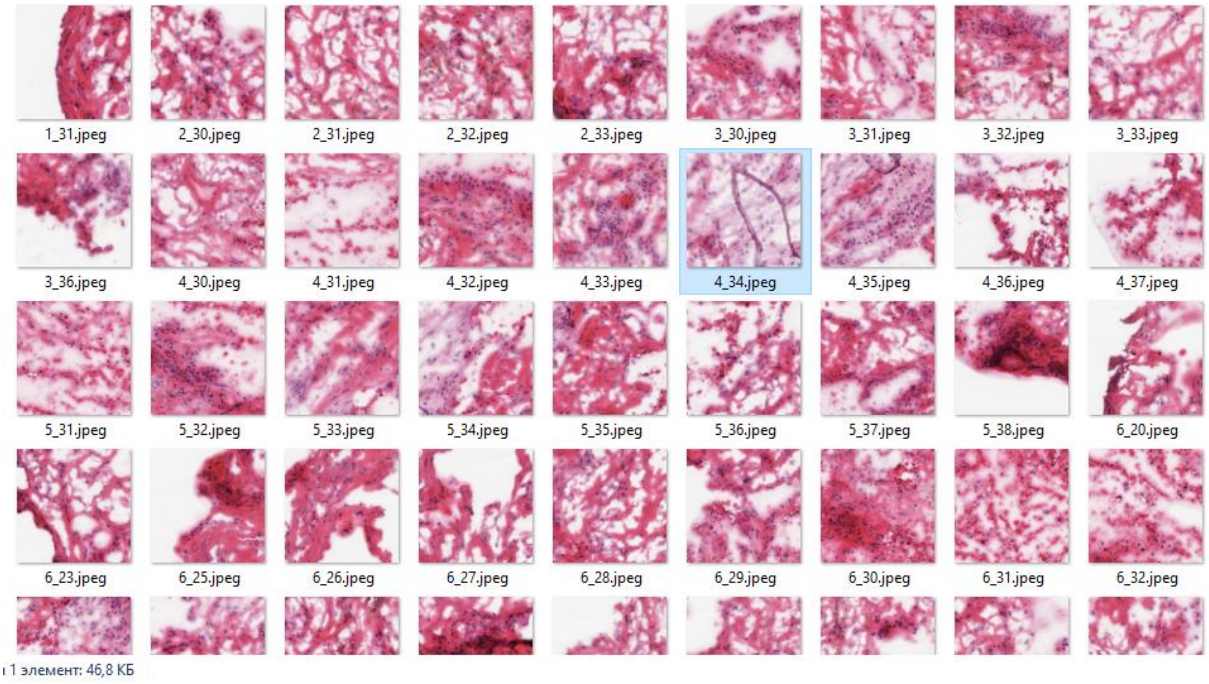

Рисунок 4.4 – Фрагменты слайду тестовых зображень

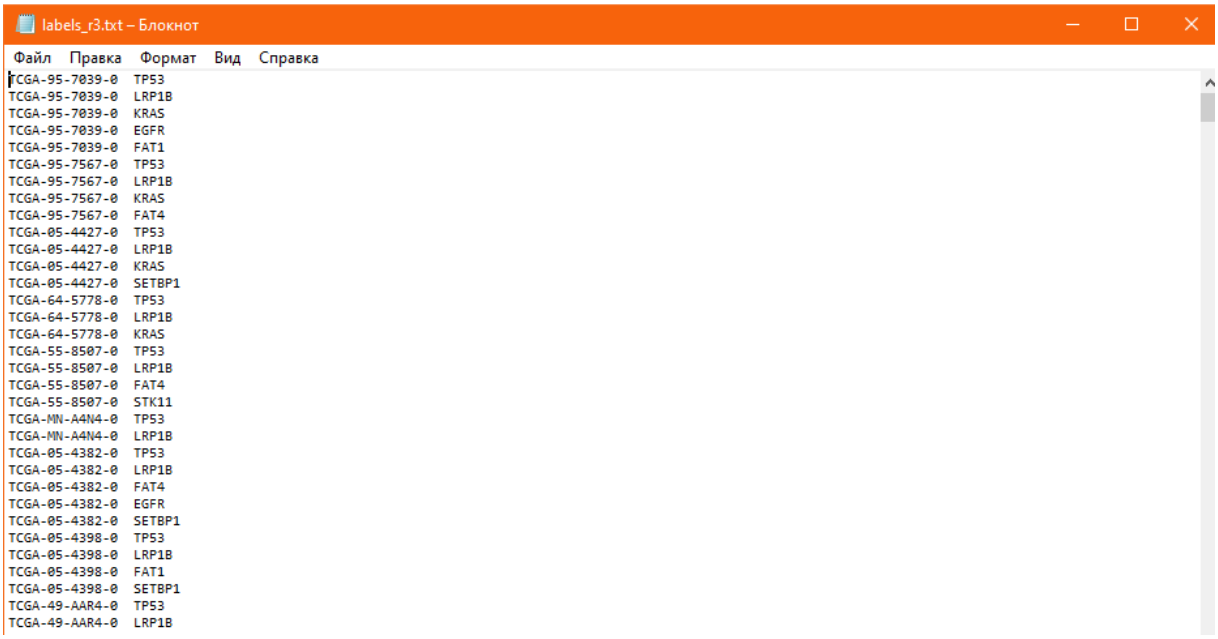

Рисунок 4.5 – Документ мапінгу (mapping) тестових слайдів та типів раку зображеніх на них

На рисунку 4.6 представлено скрипт для підготовки медициньських зображень формату “.svs” до формату придатного до роботи з нейроною мережею.

Рисунок 4.6 – Скрипт для підготовки данных

4.3 Використання модуля

Алгоритм використання (тестування) модуля:

1) Підготувати слайди біопсії тканин легень в форматі “.svs”;

2) Запустити скрипт форматування (preprocessing) данних «data_set.sh» (рис. 4.6) ;

3) Запустити скрипт тестування «test.sh» (рис. 4.7).

Результатом процессу тестування є файл out_All_Stats.txt. (рисунок 4.10)

test.sh

```
export CHECKPOINT_PATH='./code/training/tmp_9a_10m'
export OUTPUT_DIR='./r2_test'
export DATA_DIR='./r1_TFRecord_train'
export LABEL_FILE='./code/example_TCGA_lung/labelref_r1.txt'

# Best checkpoints
declare -i count=108000
declare -i NbClasses=3

# create temporary directory for checkpoints
mkdir  -p $OUTPUT_DIR/tmp_checkpoints
export CUR_CHECKPOINT=$OUTPUT_DIR/tmp_checkpoints

export TEST_OUTPUT=$OUTPUT_DIR/test_$count'k'
mkdir -p $TEST_OUTPUT

ln -s $CHECKPOINT_PATH/*-$count.* $CUR_CHECKPOINT/.
touch $CUR_CHECKPOINT/checkpoint
echo 'model_checkpoint_path: "'$CUR_CHECKPOINT'/model.ckpt-'$count'"' > $CUR_CHECKPOINT/checkpoint
echo 'all_model_checkpoint_paths: "'$CUR_CHECKPOINT'/model.ckpt-'$count'"' >> $CUR_CHECKPOINT/checkpoint

# Test
python ./code/testing/xClasses/nc_imagenet_eval.py --checkpoint_dir=$CUR_CHECKPOINT
--eval_dir=$OUTPUT_DIR --data_dir=$DATA_DIR  --batch_size 300  --run_once --ImageSet_basename='train-'
--ClassNumber $NbClasses --TVmode='test' --mode='0_softmax'

mv $OUTPUT_DIR/out* $TEST_OUTPUT/.
```

Рисунок 4.7 - Скрипт для запуску тестування моделі “test.sh”

Алгоритм роботи з модулем для його дообучення на додатковіх даних:

1) Підготувати слайди біопсії тканин легень в форматі “.svs”;

2) Запустити скрипт форматування (preprocessing) данних «data_set.sh» (рис. 4.6) ;

3) Запустити скрипт тестування «train.sh» (рис. 4.8).

4) Провести валідації роботи системи запустивши скрипт тестування «valid.sh» (рис. 4.9). Якщо процен помилки на валідаційних даних задовільний провести тестування системи, інакше повернутися до шагу 3.

train.sh

```
#!/bin/bash
#SBATCH --gres=gpu:1
#SBATCH --job-name=TFR_Vset
#SBATCH --cpus-per-task=1
#SBATCH --output=rq_TFR_%A_%a.out
#SBATCH --error=rq_TFR_%A_%a.err
#SBATCH --mem=20G

module load numpy/intel/1.13.1
module load cuda/8.0.44
module load tensorflow/python2.7/1.0.1
module load bazel/gnu/0.4.3

mkdir r1_results

bazel build inception/imagenet_train

bazel-bin/inception/imagenet_train --num_gpus=4 --batch_size=400
--train_dir='r1_results' --data_dir='r1_TFRecord_train' --ClassNumber=3
--mode='0_softmax' --NbrOfImages=923893 --save_step_for_chekcpoint=2300
--max_steps=230001
```

Рисунок 4.8 - Скрипт для запуску тестування моделі “train.sh”

valid.sh

```
export CHECKPOINT_PATH='./DeepPATH_code/training/tmp_9a_10m'
export OUTPUT_DIR='./r2_valid'
export DATA_DIR='./r1_TFRecord_valid'
export LABEL_FILE='./DeepPATH_code/example_TCGA_lung/labelref_r3.txt'

# Best checkpoints
declare -i count=108000
declare -i NbClasses=3

# create temporary directory for checkpoints
mkdir  -p $OUTPUT_DIR/tmp_checkpoints
export CUR_CHECKPOINT=$OUTPUT_DIR/tmp_checkpoints

export TEST_OUTPUT=$OUTPUT_DIR/test_$count'k'
mkdir -p $TEST_OUTPUT

ln -s $CHECKPOINT_PATH/*-$count.* $CUR_CHECKPOINT/.
touch $CUR_CHECKPOINT/checkpoint
echo 'model_checkpoint_path: "'$CUR_CHECKPOINT'/model.ckpt-'$count'"' > $CUR_CHECKPOINT/checkpoint
echo 'all_model_checkpoint_paths: "'$CUR_CHECKPOINT'/model.ckpt-'$count'"' >> $CUR_CHECKPOINT/checkpoint

# Test
python ./code/testing/xClasses/nc_imagenet_eval.py --checkpoint_dir=$CUR_CHECKPOINT
--eval_dir=$OUTPUT_DIR --data_dir=$DATA_DIR  --batch_size 300  --run_once --ImageSet_basename='valid-'
--ClassNumber $NbClasses --TVmode='valid' --mode='0_softmax'

mv $OUTPUT_DIR/out* $TEST_OUTPUT/.
```

Рисунок 4.9 - Скрипт для запуску тестування моделі “valid.sh”

Файл Правка Формат Вид Справка

```
TCGA-95-7039-0=EGFR:0.000,FAT1:0.000,FAT4:0.002,KEAP1:0.000,KRAS:0.000,LRP1B:0.000,NF1:0.000,SETBP1:0.001,STK11:0.000,TP53:0.997;
TCGA-05-4427-0=EGFR:0.000,FAT1:0.013,FAT4:0.007,KEAP1:0.005,KRAS:0.000,LRP1B:0.810,NF1:0.010,SETBP1:0.002,STK11:0.000,TP53:0.000;
TCGA-64-5778-0=EGFR:0.000,FAT1:0.000,FAT4:0.002,KEAP1:0.000,KRAS:0.983,LRP1B:0.000,NF1:0.007,SETBP1:0.006,STK11:0.001,TP53:0.002;
TCGA-55-8507-0=EGFR:0.977,FAT1:0.020,FAT4:0.004,KEAP1:0.010,KRAS:0.001,LRP1B:0.040,NF1:0.000,SETBP1:0.001,STK11:0.000,TP53:0.001;
TCGA-MN-A4N4-0=EGFR:0.000,FAT1:0.999,FAT4:0.002,KEAP1:0.000,KRAS:0.000,LRP1B:0.000,NF1:0.000,SETBP1:0.001,STK11:0.000,TP53:0.000;
TCGA-05-4382-0=EGFR:0.000,FAT1:0.000,FAT4:0.002,KEAP1:0.000,KRAS:0.000,LRP1B:0.000,NF1:0.000,SETBP1:0.001,STK11:0.000,TP53:0.997;
TCGA-95-7567-0=EGFR:0.000,FAT1:0.013,FAT4:0.007,KEAP1:0.005,KRAS:0.000,LRP1B:0.810,NF1:0.010,SETBP1:0.002,STK11:0.000,TP53:0.000;
TCGA-67-3771-0=EGFR:0.977,FAT1:0.020,FAT4:0.004,KEAP1:0.010,KRAS:0.001,LRP1B:0.040,NF1:0.000,SETBP1:0.001,STK11:0.000,TP53:0.001;
TCGA-95-7547-0=EGFR:0.000,FAT1:0.013,FAT4:0.007,KEAP1:0.005,KRAS:0.000,LRP1B:0.810,NF1:0.010,SETBP1:0.002,STK11:0.000,TP53:0.000;
TCGA-49-AAR4-0=EGFR:0.000,FAT1:0.000,FAT4:0.002,KEAP1:0.000,KRAS:0.000,LRP1B:0.000,NF1:0.000,SETBP1:0.001,STK11:0.000,TP53:0.997;
TCGA-55-A4DG-0=EGFR:0.000,FAT1:0.013,FAT4:0.007,KEAP1:0.005,KRAS:0.000,LRP1B:0.810,NF1:0.010,SETBP1:0.002,STK11:0.000,TP53:0.000;
TCGA-05-4410-0=EGFR:0.977,FAT1:0.020,FAT4:0.004,KEAP1:0.010,KRAS:0.001,LRP1B:0.040,NF1:0.000,SETBP1:0.001,STK11:0.000,TP53:0.001;
TCGA-78-8662-0=EGFR:0.000,FAT1:0.999,FAT4:0.002,KEAP1:0.000,KRAS:0.000,LRP1B:0.000,NF1:0.000,SETBP1:0.001,STK11:0.000,TP53:0.000;
TCGA-NJ-A4YF-0=EGFR:0.000,FAT1:0.013,FAT4:0.007,KEAP1:0.005,KRAS:0.000,LRP1B:0.810,NF1:0.010,SETBP1:0.002,STK11:0.000,TP53:0.000;
TCGA-44-7662-0=EGFR:0.000,FAT1:0.013,FAT4:0.007,KEAP1:0.005,KRAS:0.000,LRP1B:0.810,NF1:0.010,SETBP1:0.002,STK11:0.000,TP53:0.001;
TCGA-49-AARE-0=EGFR:0.000,FAT1:0.013,FAT4:0.007,KEAP1:0.005,KRAS:0.000,LRP1B:0.810,NF1:0.010,SETBP1:0.002,STK11:0.000,TP53:0.000;
TCGA-MN-A4N4-0=EGFR:0.000,FAT1:0.999,FAT4:0.002,KEAP1:0.000,KRAS:0.000,LRP1B:0.000,NF1:0.000,SETBP1:0.001,STK11:0.000,TP53:0.000;
TCGA-05-4382-0=EGFR:0.000,FAT1:0.000,FAT4:0.002,KEAP1:0.000,KRAS:0.000,LRP1B:0.000,NF1:0.000,SETBP1:0.001,STK11:0.000,TP53:0.997;
TCGA-95-7567-0=EGFR:0.000,FAT1:0.013,FAT4:0.007,KEAP1:0.005,KRAS:0.000,LRP1B:0.810,NF1:0.010,SETBP1:0.002,STK11:0.000,TP53:0.000;
TCGA-67-3771-0=EGFR:0.977,FAT1:0.020,FAT4:0.004,KEAP1:0.010,KRAS:0.001,LRP1B:0.040,NF1:0.000,SETBP1:0.001,STK11:0.000,TP53:0.001;
TCGA-95-7547-0=EGFR:0.000,FAT1:0.013,FAT4:0.007,KEAP1:0.005,KRAS:0.000,LRP1B:0.810,NF1:0.010,SETBP1:0.002,STK11:0.000,TP53:0.000;
TCGA-49-AAR4-0=EGFR:0.000,FAT1:0.000,FAT4:0.002,KEAP1:0.000,KRAS:0.000,LRP1B:0.000,NF1:0.000,SETBP1:0.001,STK11:0.000,TP53:0.997;
TCGA-55-A4DG-0=EGFR:0.000,FAT1:0.013,FAT4:0.007,KEAP1:0.005,KRAS:0.000,LRP1B:0.810,NF1:0.010,SETBP1:0.002,STK11:0.000,TP53:0.000;
TCGA-95-7547-0=EGFR:0.000,FAT1:0.013,FAT4:0.007,KEAP1:0.005,KRAS:0.000,LRP1B:0.810,NF1:0.010,SETBP1:0.002,STK11:0.000,TP53:0.000;
TCGA-49-AAR4-0=EGFR:0.000,FAT1:0.000,FAT4:0.002,KEAP1:0.000,KRAS:0.000,LRP1B:0.000,NF1:0.000,SETBP1:0.001,STK11:0.000,TP53:0.997;
TCGA-55-A4DG-0=EGFR:0.000,FAT1:0.013,FAT4:0.007,KEAP1:0.005,KRAS:0.000,LRP1B:0.810,NF1:0.010,SETBP1:0.002,STK11:0.000,TP53:0.000;
TCGA-05-4410-0=EGFR:0.977,FAT1:0.020,FAT4:0.004,KEAP1:0.010,KRAS:0.001,LRP1B:0.040,NF1:0.000,SETBP1:0.001,STK11:0.000,TP53:0.001;
TCGA-78-8662-0=EGFR:0.000,FAT1:0.999,FAT4:0.002,KEAP1:0.000,KRAS:0.000,LRP1B:0.000,NF1:0.000,SETBP1:0.001,STK11:0.000,TP53:0.000;
TCGA-NJ-A4YF-0=EGFR:0.000,FAT1:0.013,FAT4:0.007,KEAP1:0.005,KRAS:0.000,LRP1B:0.810,NF1:0.010,SETBP1:0.002,STK11:0.000,TP53:0.000;
TCGA-44-7662-0=EGFR:0.000,FAT1:0.013,FAT4:0.007,KEAP1:0.005,KRAS:0.000,LRP1B:0.810,NF1:0.010,SETBP1:0.002,STK11:0.000,TP53:0.001;
TCGA-49-AARE-0=EGFR:0.000,FAT1:0.013,FAT4:0.007,KEAP1:0.005,KRAS:0.000,LRP1B:0.810,NF1:0.010,SETBP1:0.002,STK11:0.000,TP53:0.000;
TCGA-MN-A4N4-0=EGFR:0.000,FAT1:0.999,FAT4:0.002,KEAP1:0.000,KRAS:0.000,LRP1B:0.000,NF1:0.000,SETBP1:0.001,STK11:0.000,TP53:0.000;
TCGA-05-4382-0=EGFR:0.000,FAT1:0.000,FAT4:0.002,KEAP1:0.000,KRAS:0.000,LRP1B:0.000,NF1:0.000,SETBP1:0.001,STK11:0.000,TP53:0.997;
TCGA-95-7567-0=EGFR:0.000,FAT1:0.013,FAT4:0.007,KEAP1:0.005,KRAS:0.000,LRP1B:0.810,NF1:0.010,SETBP1:0.002,STK11:0.000,TP53:0.000;
```

Рисунок 4.10 – Файл “out_All_Stats.txt”. Результати тестування

Логи процесу обучення зображено на рисунку 4.11.

```
true_label: [0.0, 0.0, 1.0] Percent_Selected: 0.000000  0.000000    1.000000    Average_Probability: 0.028655   0.388514    0.582832
true_label: [1.0, 0.0, 0.0] Percent_Selected: 0.301299  0.002597    0.696104    Average_Probability: 0.306132   0.245624    0.448244
true_label: [1.0, 0.0, 0.0] Percent_Selected: 0.343254  0.000000    0.656746    Average_Probability: 0.345403   0.228446    0.426151
true_label: [1.0, 0.0, 0.0] Percent_Selected: 0.842593  0.000000    0.157407    Average_Probability: 0.487952   0.195572    0.316476
true_label: [0.0, 1.0, 0.0] Percent_Selected: 0.698225  0.000000    0.301775    Average_Probability: 0.447135   0.169560    0.383304
true_label: [1.0, 0.0, 0.0] Percent_Selected: 0.079545  0.000000    0.920455    Average_Probability: 0.219084   0.290405    0.490511
true_label: [1.0, 0.0, 0.0] Percent_Selected: 0.771930  0.000000    0.228070    Average_Probability: 0.481788   0.189255    0.328957
true_label: [1.0, 0.0, 0.0] Percent_Selected: 0.538462  0.000000    0.461538    Average_Probability: 0.384511   0.213162    0.402327
true_label: [1.0, 0.0, 0.0] Percent_Selected: 0.644518  0.000000    0.355482    Average_Probability: 0.436756   0.187425    0.375819
true_label: [1.0, 0.0, 0.0] Percent_Selected: 0.393750  0.000000    0.606250    Average_Probability: 0.356554   0.227159    0.416286
true_label: [0.0, 1.0, 0.0] Percent_Selected: 0.110599  0.000000    0.889401    Average_Probability: 0.162894   0.306063    0.531043
true_label: [0.0, 0.0, 1.0] Percent_Selected: 0.050820  0.000000    0.949180    Average_Probability: 0.170285   0.307039    0.522676
true_label: [1.0, 0.0, 0.0] Percent_Selected: 0.360000  0.000000    0.640000    Average_Probability: 0.305405   0.260995    0.433600
true_label: [0.0, 0.0, 1.0] Percent_Selected: 0.003067  0.000000    0.996933    Average_Probability: 0.081380   0.342194    0.576426
true_label: [0.0, 1.0, 0.0] Percent_Selected: 0.101010  0.000000    0.898990    Average_Probability: 0.186007   0.287384    0.526609
true_label: [1.0, 0.0, 0.0] Percent_Selected: 0.041667  0.000000    0.958333    Average_Probability: 0.239346   0.264745    0.495910
true_label: [0.0, 0.0, 1.0] Percent_Selected: 0.024444  0.000000    0.975556    Average_Probability: 0.133933   0.326005    0.540061
true_label: [0.0, 0.0, 1.0] Percent_Selected: 0.761905  0.000000    0.238095    Average_Probability: 0.466616   0.178199    0.355185
true_label: [0.0, 0.0, 1.0] Percent_Selected: 0.000000  0.000000    1.000000    Average_Probability: 0.064845   0.361048    0.574107
true_label: [1.0, 0.0, 0.0] Percent_Selected: 0.797101  0.000000    0.202899    Average_Probability: 0.469473   0.195007    0.335520
true_label: [0.0, 0.0, 1.0] Percent_Selected: 0.010246  0.000000    0.989754    Average_Probability: 0.059018   0.363582    0.577399
true_label: [1.0, 0.0, 0.0] Percent_Selected: 0.625000  0.000000    0.375000    Average_Probability: 0.434040   0.194183    0.371777
```

Рисунок 4.11 – Логи процессу обучення нейронної мережі

ВИСНОВКИ

Наше дослідження демонструє, що згорткові нейронні мережі, такі як Google Inception, можуть бути використані для діагностики раку легенів по гістопатологічним слайдам: він майже однозначно класифікує нормальних і пухлиних тканин (~ 0,99 ППК) і розрізняти типи раку легенів з високою точність (0,97 ППК), досягнувши чутливості і специфічності, порівнянних з спеціалістом патологом.

Висока точність нашої моделі була досягнута, незважаючи на наявність різних артефактів на зображеннях TCGA, які були пов'язані з процедурами підготовки і збереження зразків. Однак зображення TCGA, використовувані для навчання глибокої нейронної мережі, можуть не повною мірою відображати різноманітність і неоднорідність тканин, які зазвичай перевіряють патології, що може включати додаткові ознаки, такі як некроз, кровоносні судини і запалення. Для перенавчання мережі потрібно більше слайдів, що містять такі функції, щоб ще більше підвищити її продуктивність. Незважаючи на це і той факт, що процес був навчений на заморожених зображеннях, тести показують дуже багатообіцяючі результати за класифікацією пухлин також з зрізів FFPE.

В цілому, це дослідження демонструє, що глибоко вивчені згорткові нейронні мережі можуть бути дуже корисним інструментом для допомоги патології в їх класифікації повних зображень легень. Ця інформація може мати вирішальне значення при застосуванні відповідної та спеціалізованої цільової терапії для пацієнтів з раком легені, збільшуючи тим самим обсяг і ефективність точної медицини, яка націлена на розробку мультиплексного підходу з індивідуально підібраною терапією.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Ian H. Witten, Eibe Frank and Mark A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. — 3rd Edition. — Morgan Kaufmann, 2011. — P. 664. — ISBN 9780123748560.

2. https://www.who.int/ru (Дата звернення - 05.09.2019)

3. https://www.cdc.gov/ (Дата звернення - 08.09.2019)

4. https://www.ibm.com/developerworks/ru/library/ba-data-mining-techniques/index.html (Дата звернення - 08.09.2019)

5. Lindsay, Robert K., Bruce G. Buchanan, E. A. Feigenbaum, and Joshua Lederberg. DENDRAL: A Case Study of the First Expert System for Scientific Hypothesis Formation. Artificial Intelligence 61, 2 (1993): 209-261.

6. E. H. Shortliffe. Computer-Based Medical Consultations: MYCIN. Elsevier/North Holland, New York NY, 1976.

7. http://www.casnet.com (Дата звернення - 12.10.2019)

8. http://dxplain.org/dxp/dxp.pl (Дата звернення - 13.10.2019)

9. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain – an evolving diagnostic decision-support system. JAMA. 1987; 258: 67-74.

10. Doherty J, Noirot LA, Mayfield J, Ramiah S, Huang C, Dunagan WC, Bailey TC. Implementing GermWatcher, an enterprise infection control application. AMIA Annu Symp Proc.2006:209-13.

11. Glenn Edwards, Paul Compton, Ron Malor, Ashwin Srinivasan, Leslie Lazarus. Peirs: A pathologistmaintained expert system for the interpretation of chemical pathology reports. Pathology. 1993, Vol. 25, No.1, Pages 27-34

12. Aikins JS, Kunz JC, Shortliffe EH, Fallat RJ. PUFF: an expert system for interpretation of pulmonary function data. Comput Biomed Res. 1983 Jun;16(3):199-208.

13. Кобринский Б.А. Автоматизированные диагностические и информационно-аналитические системы в педиатрии//Русский медицинский журнал. – 1999. т. 7. - №4. с. 35-42.

14. Бураковский В.И., Бокерия Л.А., Газизова Д.Ш., Лищук В.А. и др. Компьютерная технология интенсивного лечения: контроль, анализ, диагностика, лечение, обучение. – М.: НЦ ССХ РАМН, 1995.

15. Gardner RM, Pryor TA, Warner HR. The HELP hospital information system: update 1998. Int J Med Inf. 1999 Jun; 54(3):169-82.

16. Darmoni SJ, Massari P, Droy JM, Mahe N, Blanc T, Moirot E, Leroy J. SETH: an expert system for the management on acute drug poisoning in adults. Comput Methods Programs Biomed. 1994 Jun;43(3-4):171-6

17. http://www.diagnos.ru (Дата звернення - 15.11.2019)

18. http://www.homeopath-expert.com (Дата звернення - 18.11.2019)

19. Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 5 2015.

20. R. Hecht-Nielsen, “Theory of the backpropagation neural network,” in International 1989 Joint Conference on Neural Networks, 1989, pp. 593– 605 vol.1.

21. D. H. Hubel and T. N. Wiesel, “Receptive fields and functional architecture of monkey striate cortex,” Journal of Physiology (London), vol. 195, pp. 215–243, 1968.

22. K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” Biological Cybernetics, vol. 36, no. 4, pp. 193–202, Apr 1980. (https://doi.org/10.1007/BF00344251) (Дата звернення - 07.10.2019)

23. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” Neural Computation, vol. 1, no. 4, pp. 541–551, Dec 1989.

24. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

25. Y. L. Cun, “A theoretical framework for back-propagation,” 1988.

26. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in CVPR09, 2009.

27. B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, “Labelme: A database and web-based tool for image annotation,” International Journal of Computer Vision, vol. 77, no. 1, pp. 157–173, May 2008. (https://doi.org/10.1007/s11263-007-0090-8) (Дата звернення - 10.10.2019)

28. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. (http://papers.nips.cc/paper/ 4824-imagenet-classification-with-deep-convolutional-neural-networks. pdf) (Дата звернення - 11.11.2019)

29. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “Imagenet large scale visual recognition challenge,” Int. J. Comput. Vision, vol. 115, no. 3, pp. 211–252, Dec. 2015. (http://dx.doi.org/10.1007/s11263-015-0816-y) (Дата звернення - 11.11.2019)

30. M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in Computer Vision – ECCV 2014, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 818–833.

31. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” CoRR, vol. abs/1409.1556, 2014. (http://arxiv.org/abs/1409.1556) (Дата звернення - 16.11.2019)

32. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

33. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

34. G. Huang, Z. Liu, and K. Q. Weinberger, “Densely connected convolutional networks,” CoRR, vol. abs/1608.06993, 2016. (http://arxiv.org/abs/1608.06993) (Дата звернення - 18.11.2019)

35. S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” CoRR, vol. abs/1710.09829, 2017. (http://arxiv.org/abs/1710.09829) (Дата звернення - 17.11.2019)

36. J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” CoRR, vol. abs/1709.01507, 2017. (http://arxiv.org/abs/ 1709.01507) (Дата звернення - 10.10.2019)

37. Y. LeCun and C. Cortes, “MNIST handwritten digit database,” 2010. (http://yann.lecun.com/exdb/mnist/)(Дата звернення - 14.11.2019)

38. Travis, W. D. et al. International Association for the Study of Lung Cancer/ American Thoracic Society/European Respiratory Society international multidisciplinary classification of lung adenocarcinoma. J. Thorac. Oncol. 6, 244–285 (2011).

39. Hanna, N. et al. Systemic therapy for stage IV non–small-cell lung cancer: American Society of Clinical Oncology clinical practice guideline update. J. Clin. Oncol. 35, 3484–3515 (2017).

40. Chan, B. A. & Hughes, B. G. Targeted therapy for non–small cell lung cancer: current standards and the promise of the future. Transl. Lung Cancer Res. 4, 36–54 (2015).

41. Terra, S. B. et al. Molecular characterization of pulmonary sarcomatoid carcinoma: analysis of 33 cases. Mod. Pathol. 29, 824–831 (2016).

42. Blumenthal, G. M. et al. Oncology drug approvals: evaluating endpoints and evidence in an era of breakthrough therapies. Oncologist 22, 762–767 (2017).

43. Thunnissen, E., van der Oord, K. & den Bakker, M. Prognostic and predictive biomarkers in lung cancer. A review. Virchows Arch. 464, 347–358 (2014).

44. Yu, K.-H. et al. Predicting non–small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat. Commun. 7, 12474 (2016).

45. Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O. & Hajirasouliha, I. Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. EBioMedicine 27, 317–328 (2018).

46. Sozzi, G. et al. Quantification of free circulating DNA as a diagnostic marker in lung cancer. J. Clin. Oncol. 21, 3902–3908 (2003).

47. Terry, J. et al. Optimal immunohistochemical markers for distinguishing lung adenocarcinomas from squamous cell carcinomas in small tumor samples. Am. J. Surg. Pathol. 34, 1805–1811 (2010).

48. Schmidhuber, J. Deep learning in neural networks: an overview. Neural Netw. 61, 85–117 (2015).

49. Qaiser, T., Tsang, Y.-W., Epstein, D. & RajpootEma, N. Tumor segmentation in whole slide images using persistent homology and deep convolutional features. In Medical Image Understanding and Analysis: 21st Annual Conference on Medical Image Understanding and Analysis. (Eds. Valdes Hernandez, M. & González-Castro, V.) 320–329 (Springer International Publishing, New York, 2018).

50. Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19, 221–248 (2017).

51. Xing, F., Xie, Y. & Yang, L. An automatic learning-based framework for robust nucleus segmentation. IEEE Trans. Med. Imaging 35, 550–566 (2016).

52. Simon, O., Yacoub, R., Jain, S., Tomaszewski, J. E. & Sarder, P. Multi-radial LBP features as a tool for rapid glomerular detection and assessment in whole slide histopathology images. Sci. Rep. 8, 2032 (2018).

53. Cheng, J.-Z. et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Sci. Rep. 6, 24454 (2016).

54. Sirinukunwattana, K. et al. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Trans. Med. Imaging 35, 1196–1206 (2016).

55. Ertosun, M. G. & Rubin., D. L. Automated grading of gliomas using deep learning in digital pathology images: a modular approach with ensemble of convolutional neural networks. In AMIA Annual Symposium Proceedings. 1899–1908 (American Medical Informatics Association, Bethesda, MD, USA).

56. Bulten, W., Kaa, C.A.H.-d., Laak, J.d. & Litjens, G.J. Automated segmentation of epithelial tissue in prostatectomy slides using deep learning. In Medical Imaging 2018: Digital Pathology. Vol. 10581 (Eds. Tomaszewski, J. E. & Gurcan, M. N.) 105810S (International Society for Optics and Photonics, Bellingham, WA, USA, 2018).

57. Mishra, R., Daescu, O., Leavey, P., Rakheja, D. & Sengupta, A. Histopathological Diagnosis for Viable and Non-viable Tumor Prediction for Osteosarcoma Using Convolutional Neural Network. In International Symposium on Bioinformatics Research and Applications Vol. 10330 (Eds. Cai, Z., D. Ovidiu, & Li, M.) 12–23 (Springer International Publishing, New York, 2018).

58. Grossman, R. L. et al. Toward a shared vision for cancer genomic data. N. Engl. J. Med. 375, 1109–1112 (2016).

59. Abels, E. & Pantanowitz, L. Current state of the regulatory trajectory for whole slide imaging devices in the USA. J. Pathol. Inform. 8, 23 (2017).

60. Sanchez-Cespedes, M. et al. Inactivation of LKB1/STK11 is a common event in adenocarcinomas of the lung. Cancer Res. 62, 3659–3662 (2002).

61. Kandoth, C. et al. Mutational landscape and significance across 12 major cancer types. Nature 502, 333–339 (2013).

62. Zeiler, M.D. & Fergus, R. Visualizing and understanding convolutional networks. In European Conference on Computer Vision. 818–833 (Springer International Publishing, New York, 2015).

63. Chiang, S. et al. IDH2 mutations define a unique subtype of breast cancer with altered nuclear polarity. Cancer Res. 76, 7118–7129 (2016).

64. Dearden, S., Stevens, J., Wu, Y.-L. & Blowers, D. Mutation incidence and coincidence in non small-cell lung cancer: meta-analyses by ethnicity and histology (mutMap). Ann. Oncol 24, 2371–2376 (2013).