

**НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І
ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ**

Факультет Інформаційних технологій

ПОГОДЖЕНО

Декан факультету (Директор ННІ)

інформаційних технологій

(назва факультету (ННІ))

(підпис)

Ігор Болбот

(ім'я ПРІЗВИЩЕ)

“ ” _____ 2025 р.

ДОПУСКАЄТЬСЯ ДО ЗАХИСТУ

Завідувач кафедри

комп'ютерних наук

(назва кафедри)

(підпис)

Белла Голуб

(ім'я ПРІЗВИЩЕ)

“ ” _____ 2025 р.

МАГІСТЕРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

на тему Система обробки та зберігання даних закордонних паспортів з елементами машинного навчання

Спеціальність 122 - Комп'ютерні науки

(код і найменування)

Освітня програма Інформаційно управляючі системи та технології

(назва)

Орієнтація освітньої програми освітньо-професійна

(освітньо-професійна або освітньо-наукова)

Гарант освітньої програми

к.т.н., доцент

(науковий ступінь та вчене звання)

(підпис)

Белла Голуб

(Ім'я Прізвище)

Керівник магістерської кваліфікаційної роботи

к.т.н., доцент

(науковий ступінь та вчене звання)

(підпис)

Іван Пархоменко

(Ім'я Прізвище)

Виконав

(підпис)

Данііл Прокопенко

(Ім'я Прізвище зобувача)

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ
І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

Факультет (ННІ) Інформаційних технологій

ЗАТВЕРДЖУЮ

Завідувач кафедри комп'ютерних наук
к.т.н., доцент Белла Голуб
(науковий ступінь, вчене звання) (підпис) (ім'я ПРІЗВИЩЕ)
" 01 " листопада 2024 року

ЗАВДАННЯ

ДО ВИКОНАННЯ МАГІСТЕРСЬКОЇ КВАЛІФІКАЦІЙНОЇ РОБОТИ ЗДОБУВАЧУ

Прокопенко Данііл Олександрович

(прізвище, ім'я, по батькові)

Спеціальність 122 «Комп'ютерні науки»

(код і назва)

Освітня програма Інформаційні управляючі системи та технології

(назва)

Орієнтація освітньої програми освітньо-професійна

Тема магістерської кваліфікаційної роботи: Система обробки та зберігання даних закордонних паспортів з елементами машинного навчання

затверджена наказом від " 01 " листопада 2024р. №1964 «С»

Термін подання завершеної роботи на кафедру 01.12.2025

(число, місяць, рік)

Вихідні дані до магістерської кваліфікаційної роботи: зображення закордонних паспортів (скановані копії та фотографії різної якості), готові моделі для оптичного розпізнавання тексту, мова програмування C#, бібліотеки та фреймворки ML.NET.

Перелік питань, що підлягають дослідженню:

- Аналіз предметної області та характеристика паспортних документів.
- Дослідження та використання моделей машинного навчання для попередньої обробки паспортних зображень та підвищення точності OCR
- Оцінка результатів роботи системи та аналіз її практичної придатності.

Перелік графічного матеріалу (за потреби)

Дата видачі завдання " 01 " листопада 2024 р.

Керівник магістерської кваліфікаційної роботи

Іван Пархоменко

(підпис)

(Ім'я Прізвище)

Завдання прийняв до виконання

Данііл Прокопенко

(підпис)

(Ім'я Прізвище)

Зміст

ВСТУП	5
РОЗДІЛ 1. СИСТЕМНИЙ АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ	10
1.1 Опис предметної області та характеристика паспортних документів	10
1.2 Аналіз проблем і недоліків існуючих підходів до обробки паспортних даних	12
1.3 Огляд сучасних технологій автоматизованої обробки документів (OCR та машинне навчання)	15
1.3.1 Основні технології OCR та їх характеристики	15
1.3.2 Машинне навчання в обробці документів	18
1.4 Постановка завдання магістерського дослідження	19
РОЗДІЛ 2 МОДЕЛЮВАННЯ СИСТЕМИ	22
2.1 Методологія моделювання системи	22
2.2 Об'єктне та функціональне моделювання	24
2.2.1 Діаграма прецедентів	24
2.2.2 Діаграма послідовності	30
2.2.3 Діаграма діяльності	35
2.2.4 Діаграма класів	38
2.3 Моделювання структури даних	41
2.4 Визначення вимог до системи	43
РОЗДІЛ 3. РОЗРОБКА СИСТЕМИ	46
3.1 Вибір інструментальних засобів та середовища розробки	46
3.1.1 Середовище розробки та мова програмування	46
3.1.2 Обґрунтування вибору бібліотеки Tesseract OCR	47
3.1.3 Обґрунтування вибору ML.NET та його роль у системі	48
3.1.4 Вибір системи керування базами даних SQL Server та інструменту SSMS	50
3.2 Архітектура спроектованої системи	51
3.3 Реалізація модуля обробки зображень паспортів	53
3.3.1 Використання Tesseract OCR	53
3.3.2 Використання Tesseract у поєднанні з методами МН	59
3.4 Інтерфейс користувача	61
РОЗДІЛ 4. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ	67
4.1 Вимоги до апаратного та програмного забезпечення	67
4.2 Хід виконання дослідження	70
4.3 Аналіз результатів роботи	72
ВИСНОВКИ	78
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ	80

СПИСОК УМОВНИХ ПОЗНАЧЕНЬ

OCR – Optical Character Recognition (оптичне розпізнавання символів);

MRZ – Machine Readable Zone (машинно-зчитувана зона);

SQL – Structured Query Language (мова структурованих запитів);

UML – Unified Modeling Language (уніфікована мова моделювання);

SSMS – SQL Server Management Studio (середовище управління SQL Server);

БД – База Даних;

ML – Machine Learning (машинне навчання).

ВСТУП

Актуальність теми дослідження. Сучасні туристичні компанії використовують цифрові технології для автоматизації процесів обслуговування клієнтів. Одним із основних документів, необхідних для здійснення подорожей, є закордонний паспорт, дані якого необхідно швидко та безпомилково вводити у системи бронювання авіаквитків, готелів та інших послуг.

У більшості туристичних агентств дані з паспортів вводяться вручну — менеджери переносять інформацію зі сканів або фото у систему бронювання. Такий підхід займає багато часу і часто призводить до неточностей: навіть одна помилка в номері або прізвищі може спричинити відмову у посадці на рейс чи потребу повторного бронювання.

У цій роботі розглядається можливість використання технологій оптичного розпізнавання символів (OCR) разом із алгоритмами машинного навчання для автоматизації зчитування паспортних даних.

Отже, створення системи обробки та зберігання даних закордонних паспортів з елементами машинного навчання є актуальним завданням для туристичних фірм, оскільки воно повинно забезпечити швидкість, точність і надійність роботи з документами, що напряду впливає на якість послуг і конкурентоспроможність компанії.

Об'єктом дослідження процеси автоматичної обробки паспортних даних в інформаційних системах туристичних компаній. Об'єктом виступають цифрові зображення закордонних паспортів, що містять персональні дані клієнтів, а також технологічні етапи їх розпізнавання, верифікації та подальшого збереження у корпоративних базах даних. Особливу увагу приділено аспектам підвищення точності та швидкості обробки цих даних, що має безпосередній вплив на якість сервісного обслуговування та ефективність бізнес-процесів у сфері туризму.

Предметом дослідження є система автоматизації обробки даних закордонних паспортів із використанням елементів машинного навчання.

Предмет охоплює методи, моделі та алгоритми, які забезпечують інтелектуальну обробку зображень паспортів, включно з виділенням зон MRZ, підвищенням якості зображення, оптичним розпізнаванням символів (OCR) та автоматичною корекцією помилок за допомогою глибоких нейронних мереж. Основна увага приділяється побудові комплексної системи, здатної працювати в реальному часі та інтегруватися в існуючі інформаційні системи туристичних компаній.

Метою дослідження даної роботи є оцінка впливу методів машинного навчання на ефективність розпізнавання паспортних даних за допомогою технології OCR та розробити програмну систему, що забезпечує автоматизацію процесу обробки паспортної інформації. Досягнення поставленої мети передбачає аналіз сучасних технологій оптичного розпізнавання символів, вивчення можливостей глибоких нейронних мереж для підвищення точності та надійності обробки даних, а також практичну реалізацію системи, здатної зменшити вплив людського фактора, скоротити час введення інформації та забезпечити єдину стандартизовану структуру паспортних даних.

Для досягнення мети дослідження у магістерській роботі поставлено наступні **завдання**:

1. Провести аналіз процесів обробки та зберігання даних закордонних паспортів у туристичних компаніях.
2. Сформулювати функціональні та нефункціональні вимоги до системи автоматизації обробки паспортних даних.
3. Дослідити сучасні технології оптичного розпізнавання символів (OCR) та алгоритми машинного навчання, які можуть бути застосовані для автоматизації обробки паспортних даних.
4. Побудувати модель предметної області та архітектуру системи обробки та зберігання даних закордонних паспортів.
5. Розробити програмну реалізацію системи та інтегрувати методи OCR і машинного навчання для підвищення точності і швидкості обробки даних.
6. Провести тестування системи, оцінити її ефективність та сформулювати висновки щодо доцільності використання запропонованого рішення.

Для досягнення мети магістерської роботи та виконання поставлених завдань використовуються комплексні **методи дослідження**, що поєднують аналіз процесів, сучасні інформаційні технології та алгоритми машинного навчання.

На першому етапі проводиться системний аналіз об'єкта дослідження — процесів обробки та зберігання даних закордонних паспортів у туристичних компаніях. Це дозволяє виявити проблемні місця, оцінити ефективність існуючих підходів та сформулювати вимоги до автоматизованої системи.

Для автоматизації зчитування паспортних даних застосовується технологія оптичного розпізнавання символів (OCR), що забезпечує високоточне отримання інформації з документів, мінімізуючи людські помилки при введенні даних у системи бронювання.

Для підвищення точності і надійності обробки даних використовується машинне навчання, зокрема алгоритми для перевірки коректності зчитаних даних, виявлення аномалій та автоматичного виправлення помилок. Застосування таких методів дозволяє не лише прискорити процес обробки паспортів, але й забезпечити більшу надійність і безпеку збереження інформації.

Таким чином, комплексний підхід, що поєднує системний аналіз, OCR, алгоритми машинного навчання та організацію баз даних, забезпечує реалізацію поставленої мети дослідження та підвищує ефективність процесів обробки паспортних даних у туристичній компанії.

Наукова новизна магістерської роботи полягає у використанні методів машинного навчання для підвищення ефективності технології OCR під час автоматизованого зчитування паспортних даних.

Розроблено та реалізовано підхід до попередньої обробки зображень закордонних паспортів, який дозволяє покращити якість розпізнавання MRZ-зони, особливо у випадках низької якості зображення чи спотворень. Удосконалено процес перевірки та корекції результатів розпізнавання шляхом використання моделі машинного навчання для автоматичного виявлення помилкових символів і їх виправлення за допомогою статистичних

закономірностей у структурі MRZ-коду. Розроблена архітектура системи забезпечує інтеграцію OCR та ML-модулів із реляційною базою даних, що підвищує швидкість доступу до інформації та зменшує ймовірність втрати або дублювання даних.

Отримані результати становлять практичну та наукову цінність, оскільки доводять доцільність поєднання технологій оптичного розпізнавання символів із методами машинного навчання для оптимізації процесів обробки паспортних документів у сфері туристичних послуг.

Магістерська робота викладена на 82 сторінках машинописного тексту, містить 18 рисунків, 3 таблиць та список використаних джерел із 32 найменувань. Робота складається зі вступу, чотирьох розділів, висновків, списку використаних джерел і додатків.

У вступі обґрунтовано актуальність теми, визначено мету, об'єкт, предмет і завдання дослідження, наведено наукову новизну та практичне значення отриманих результатів.

У першому розділі проведено системний аналіз предметної області, подано опис паспортних документів, виконано огляд існуючих підходів до автоматизації обробки паспортних даних, розглянуто сучасні технології OCR та методи машинного навчання, а також сформульовано постановку завдання магістерського дослідження.

У другому розділі здійснено моделювання системи: обґрунтовано методологію моделювання, побудовано об'єктні та функціональні діаграми (прецедентів, послідовності, діяльності), спроектовано структуру бази даних і визначено вимоги до системи.

У третьому розділі описано процес розробки системи. Наведено вибір інструментальних засобів і середовища розробки, розглянуто використання технологій Tesseract OCR, ML.NET і SQL Server. Подано архітектуру системи, описано реалізацію модулів для обробки зображень паспортів та користувацького інтерфейсу.

У четвертому розділі наведено результати дослідження: визначено вимоги до апаратного та програмного забезпечення, описано хід виконання дослідження, проведено оцінку ефективності роботи системи та порівняльний аналіз результатів.

У висновках узагальнено результати дослідження, сформульовано основні наукові та практичні висновки, а також наведено рекомендації щодо подальшого розвитку системи.

РОЗДІЛ 1. СИСТЕМНИЙ АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Опис предметної області та характеристика паспортних документів

Закордонний паспорт — основний документ, що посвідчує особу громадянина під час перетину державного кордону. Його значення полягає не лише у підтвердженні особи власника, а й у наданні доступу до широкого спектра сервісів: бронювання авіаквитків, поселення у готелях, отримання віз, оформлення страхових полісів тощо. Для туристичних компаній паспорт є основним джерелом ідентифікаційних даних клієнта, від точності яких залежить коректність усіх наступних операцій, оскільки навіть невелика помилка у цих даних може спричинити проблеми з бронюванням або оформленням подорожі. [1].

Закордонний паспорт має стандартизовану структуру, яка визначається міжнародними нормами. Зокрема, відповідно до вимог Міжнародної організації цивільної авіації (ICAO), основні дані про власника документа зосереджені на біографічній сторінці паспорта. На ній розміщуються такі реквізити: прізвище та ім'я власника, дата і місце народження, стать, громадянство, дата видачі та закінчення дії документа, номер паспорта, а також фотографія та підпис власника [2].

Крім візуальної інформації, сучасні паспорти містять так звану машинозчитувальну зону (Machine-Readable Zone, MRZ), що дозволяє автоматизувати процес зчитування даних. Формат MRZ регламентується стандартом ICAO Document 9303, і зазвичай складається з двох рядків по 44 символи у паспортах-книжечках. У цій зоні міститься зашифрована у стандартизованому вигляді інформація: тип документа, код держави, прізвище та ім'я, номер паспорта, дата народження, стать, термін дії документа, а також контрольні цифри для перевірки коректності даних [3].

Також важливим є те, що MRZ дозволяє виключити суб'єктивний фактор людини під час введення інформації. Наприклад, при ручному введенні можуть виникати помилки у транслітерації прізвищ, заміні цифр або пропуску символів. Завдяки використанню MRZ система отримує дані у єдиному уніфікованому форматі, що істотно знижує ймовірність помилок і прискорює роботу з документами.

Сучасні паспорти дедалі частіше мають електронний носій (чіп), який дає можливість зберігати біометричні дані власника — цифрове фото, відбитки пальців, іноді навіть дані райдужної оболонки ока. Такі документи отримали назву біометричних паспортів (е-паспортів) [4]. Вони підвищують рівень безпеки, адже дані з чіпа складніше підробити, а також спрощують проходження контролю у багатьох країнах.

З точки зору автоматизованої обробки, паспорт має низку особливостей, що впливають на точність і швидкість розпізнавання. По-перше, якість фотографії чи сканованого документа відіграє вирішальну роль: тіні, спотворення, низька роздільна здатність або відблиски можуть суттєво ускладнити роботу алгоритмів OCR. По-друге, декоративні елементи паспорта, захисні сітки та голограми, які виконують функцію захисту від підробки, також можуть заважати розпізнаванню тексту. По-третє, різні країни використовують відмінні формати написання імен та шрифтів, що вимагає універсальності від системи автоматизованого зчитування [5].

Отже, предметна область, пов'язана з обробкою даних закордонних паспортів, характеризується поєднанням двох ключових складових: наявності стандартизованих реквізитів, які можна обробляти автоматизовано (MRZ), та низки технічних і організаційних викликів, пов'язаних із якістю зображень, захисними елементами документів і різноманітністю форматів у різних країнах. Усе це робить актуальним застосування сучасних технологій оптичного розпізнавання символів (OCR) та алгоритмів машинного навчання, здатних адаптуватися до складних умов введення і забезпечувати високу точність та надійність роботи системи.

1.2 Аналіз проблем і недоліків існуючих підходів до обробки паспортних даних

У діяльності туристичних компаній обробка паспортних даних клієнтів є ключовим процесом, оскільки саме на основі цих даних здійснюється бронювання авіаквитків, готелів, оформлення страхових полісів, візова підтримка та інші супутні послуги. Від точності і швидкості обробки паспортної інформації залежить не тільки якість обслуговування клієнта, а й репутація компанії, а також її фінансова ефективність.

Незважаючи на існування різноманітних систем та програмних рішень для обробки документів, на практиці туристичні фірми часто мають труднощі, що ускладнює роботу з паспортними даними. Однією з основних є ручне введення інформації. Багато агентств досі використовують класичні методи обробки: менеджери отримують скан-копії або фотографії паспортів, після чого вручну вносять дані до внутрішніх систем бронювання або CRM. Це уповільнює роботу та підвищує ризик помилок. Наприклад, неправильне введення імені латиницею, невірний номер паспорта або помилки у даті народження можуть призвести до анулювання бронювання, відмови у реєстрації на рейс або додаткових фінансових витрат компанії. У середньому, обробка паспортних даних вручну займає від 5 до 10 хвилин на одного клієнта, що при великому потоці туристів значно знижує швидкість обслуговування.

Ще одна проблема полягає в обмеженні ефективності наявних систем OCR. Хоча на ринку існують популярні рішення, такі як ABBYY FineReader, Tesseract або хмарні сервіси Google Vision OCR і Microsoft Azure Form Recognizer, вони часто недостатньо адаптовані для специфіки туристичної галузі. Паспортні документи можуть мати різну якість сканування, відображати захисні елементи (водяні знаки, голограми, мікротекст), містити записи різними мовами, що суттєво знижує точність розпізнавання. Як показує практика, навіть сучасні OCR-системи допускають похибки, які потребують додаткової ручної перевірки. У туристичній сфері це означає, що менеджери повинні витратити додаткові 2–

3 хвилини на перевірку кожного документа, що при великій кількості клієнтів перетворюється на значні втрати часу та ресурсів [6].

Наступним важливим аспектом є безпека і конфіденційність даних. Паспортні дані належать до категорії критично важливої персональної інформації, що підлягає захисту відповідно до законодавства. На жаль, у багатьох туристичних компаніях копії паспортів зберігаються у загальнодоступних папках або локальних комп'ютерах без належного шифрування та контролю доступу. Це створює ризик витоку або несанкціонованого доступу до персональних даних клієнтів. Такі дії можуть призвести до юридичної відповідальності та втрати довіри. Насправді навіть одна витік інформації про клієнтів може коштувати туристичній фірмі десятки тисяч гривень штрафів та значно знизити рейтинг серед партнерів і клієнтів.

Крім того, проблемою є відсутність інтелектуальної обробки даних. Сучасні системи OCR здебільшого виконують лише базове перетворення зображення документа на текст, без можливості перевірки правильності інформації, корекції помилок та адаптації алгоритмів до специфічних типів паспортів. Для туристичної компанії важливо автоматично перевіряти правильність транслітерації, термін дії паспорта відносно дати подорожі, наявність вільних сторінок для віз, відповідність введених даних стандартам авіакомпаній та готелів. Відсутність таких функцій призводить до додаткових витрат часу на перевірку, повторне введення даних та виникнення помилок у бронюваннях.

Таким чином, аналіз існуючих підходів до обробки паспортних даних у туристичних компаніях дозволяє виділити ключові проблеми, що знижують ефективність роботи з клієнтською інформацією. Основними недоліками є висока залежність від ручного введення даних, обмежена точність OCR при обробці документів різної якості та зі складними захисними елементами, низький рівень інформаційної безпеки та ризики витоку персональних даних, а також обмежене застосування інтелектуальних методів для автоматичної перевірки і корекції даних. Наслідком цих недоліків є не тільки додаткові часові витрати

працівників, а й підвищений ризик фінансових втрат і погіршення рівня обслуговування клієнтів, що прямо впливає на репутацію та конкурентоспроможність туристичної фірми.

Враховуючи зазначені проблеми, стає очевидною потреба у створенні спеціалізованої системи обробки паспортних даних саме для туристичних компаній. Така система повинна забезпечувати комплексну автоматизацію процесів введення та перевірки даних, мінімізувати вплив людського фактора та знизити ймовірність помилок. Важливим аспектом є високий рівень інформаційної безпеки, що гарантує захист персональних даних клієнтів від несанкціонованого доступу або витоку. Інтеграція елементів машинного навчання дає змогу автоматично контролювати коректність введених даних, адаптувати алгоритми до різних типів паспортів та підвищувати точність розпізнавання з часом. Реалізація такої системи дасть змогу туристичній компанії значно скоротити час обробки паспортних даних, підвищити якість обслуговування клієнтів та зменшити фінансові та репутаційні ризики, що виникають у разі помилок або затримок у роботі.

1.3 Огляд сучасних технологій автоматизованої обробки документів (OCR та машинне навчання)

1.3.1 Основні технології OCR та їх характеристики

Оптичне розпізнавання символів (OCR, Optical Character Recognition) є ключовим інструментом для автоматизованої обробки документів. Його застосування охоплює широкий спектр завдань – від оцифрування архівних матеріалів до інтелектуальної обробки сучасних документів у банківській, юридичній, транспортній та туристичній сферах. Для завдань, пов'язаних із розпізнаванням паспортних документів, технології OCR мають особливе значення, адже вони дозволяють перетворювати візуальне зображення паспорта у структурований текстовий формат, придатний для подальшої перевірки, аналізу чи збереження у базі даних.

Сьогодні існує кілька провідних OCR-систем, що активно застосовуються на практиці. Їх характеристики наведено нижче.

1. ABBYY FineReader.

Це одне з найпоширеніших комерційних рішень для OCR. FineReader забезпечує надзвичайно високу точність розпізнавання навіть у випадках, коли текст має складне форматування, містить таблиці чи багатоколонкову структуру. Система підтримує понад 190 мов, включаючи українську та російську. FineReader також надає можливості створення редагованих PDF, пошуку тексту в зображеннях та порівняння документів. Основним недоліком є висока вартість ліцензії та закритість програмного коду, що обмежує можливості кастомізації та інтеграції у спеціалізовані системи.

2. Google Cloud Vision API.

Це хмарний сервіс від Google, який надає широкий набір інструментів для аналізу зображень, серед яких є функція OCR. Vision API демонструє високу точність завдяки використанню глибокого навчання, зокрема convolutional neural networks (CNN). Основні переваги – простота інтеграції через REST API, масштабованість та можливість обробки великих обсягів даних. Недоліки

полягають у залежності від інтернет-з'єднання та вартості використання, яка збільшується зі зростанням кількості запитів. Крім того, відсутність повного контролю над процесом розпізнавання може бути критичною для систем із підвищеними вимогами до конфіденційності даних.

3. Amazon Textract.

Це рішення від Amazon, яке поєднує функціонал OCR із можливістю інтелектуального вилучення структурованих даних із документів. Textract здатен розпізнавати не лише текст, а й таблиці, поля форм, що робить його зручним для обробки фінансових або юридичних документів. Як і Google Vision API, Textract інтегрується через хмарний сервіс і орієнтований на масштабовані проекти. Основними недоліками є залежність від зовнішньої інфраструктури AWS та потреба у додаткових витратах.

4. EasyOCR.

Відкрите рішення, створене на основі бібліотек глибинного навчання (PyTorch). Підтримує понад 80 мов і дозволяє швидко інтегрувати OCR у Python-проекти. EasyOCR відрізняється простотою використання та відкритістю, однак поступається точністю комерційним рішенням і може вимагати значних обчислювальних ресурсів при роботі з великими наборами документів.

5. Tesseract OCR.

Одна з найпоширеніших відкритих систем OCR з відкритим кодом. Підтримує понад 100 мов, включаючи українську, російську та англійську. Основними перевагами Tesseract є безкоштовність, активна спільнота розробників та наявність великої кількості бібліотек для інтеграції у різні мови програмування (C#, Java, Python). Якість розпізнавання сильно залежить від попередньої обробки зображення – для досягнення високої точності часто застосовуються алгоритми бінаризації, нормалізації яскравості, видалення шумів та виправлення нахилу тексту. Попри те, що Tesseract поступається АBBYY FineReader у точності розпізнавання складних форматованих документів, для задач розпізнавання паспортних даних він є оптимальним рішенням з огляду на баланс «якість–вартість–можливості інтеграції».

Для кращої наочності характеристики розглянутих технологій наведено у таблиці 1.1.

Таблиця 1.1

Порівняльна характеристика сучасних OCR-технологій

Технологія	Тип ліцензії	Підтримка мов	Точність розпізнавання	Особливості	Недоліки
ABBYY FineReader	Комерційна	>190	Дуже висока	Обробка складних PDF, збереження форматування	Висока вартість, закритий код
Google Vision API	Комерційна (хмара)	>50	Висока	Маштабованість, ML-алгоритми Google	Залежність від інтернету, вартість
Amazon Textract	Комерційна (хмара)	>50	Висока	Розпізнавання форм і таблиць	Вартість, залежність від AWS
EasyOCR	Open Source	~80	Середня	Простота інтеграції в Python, ML-бібліотеки	Нижча точність, ресурсоємність
Tesseract OCR	Open Source	>100	Висока (з попередньою обробкою)	Широка підтримка мов, інтеграція у C#/Java/Python	Залежність від якості зображення

Виходячи з таблиці можна умовно поділити сучасні OCR-системи на дві групи: комерційні хмарні сервіси (ABBYY, Google Vision, Amazon Textract), що забезпечують максимально високу точність, але потребують додаткових витрат та доступу до інтернету, та відкриті рішення (EasyOCR, Tesseract), що відзначаються безкоштовністю, простотою інтеграції та активною підтримкою спільноти. Вибір конкретної технології визначається завданнями дослідження, обсягами даних та вимогами до інтеграції в інформаційну систему.

1.3.2 Машинне навчання в обробці документів

Сучасні системи обробки документів поступово переходять від традиційних алгоритмічних підходів до використання методів машинного навчання (ML). На відміну від класичних OCR-систем, які здебільшого ґрунтуються на шаблонному розпізнаванні символів та правил, моделі машинного навчання забезпечують гнучкість, здатність до самонавчання та високу точність при роботі з різноманітними джерелами даних.

Одним із ключових напрямів застосування ML у сфері обробки документів є розпізнавання тексту на основі нейронних мереж, зокрема згорткових (Convolutional Neural Networks, CNN) та рекурентних (Recurrent Neural Networks, RNN, LSTM). CNN ефективно працюють із вхідними зображеннями, дозволяючи автоматично виявляти локальні ознаки символів (контури, лінії, сегменти), тоді як RNN застосовуються для послідовної обробки тексту, забезпечуючи врахування контексту символів та слів. У комплексі ці архітектури формують сучасні OCR-моделі, такі як CRNN (Convolutional Recurrent Neural Network), які демонструють значно вищі показники точності порівняно з класичними підходами.

Другим завданням, яке вирішується за допомогою ML, є класифікація та структуризація документів. Алгоритми машинного навчання дозволяють автоматично розподіляти документи за категоріями (паспорти, договори, рахунки-фактури), витягати ключові атрибути (ім'я, дата народження, номер документа) та створювати семантичні зв'язки між ними. Це досягається шляхом використання методів обробки природної мови (Natural Language Processing, NLP), які застосовують векторні подання слів (Word2Vec, FastText, BERT) та трансформерні архітектури.

Особливе місце в обробці документів посідають методи навчання з підкріпленням та глибинного навчання. Вони дозволяють створювати системи, які не лише розпізнають текст, а й адаптуються до конкретного середовища використання. Наприклад, при роботі з паспортними документами модель може

вдосконалювати свої результати шляхом активного навчання (active learning), отримуючи зворотний зв'язок від користувача та підвищуючи точність у процесі експлуатації.

Крім того, машинне навчання дає можливість реалізувати модулі автоматичної перевірки автентичності документів, де застосовуються моделі для аналізу захисних елементів, водяних знаків, а також виявлення підробок за статистичними відхиленнями у структурі документа. Це особливо важливо в системах перевірки закордонних паспортів, де від точності алгоритму залежить достовірність результатів і безпека всього процесу.

Таким чином, застосування машинного навчання в обробці документів дозволяє не лише підвищити точність розпізнавання тексту, а й розширити функціонал систем — від автоматизованого введення даних до комплексного аналізу структури документа. Це робить ML одним із ключових напрямів подальшого розвитку технологій інтелектуальної обробки даних.

1.4 Постановка завдання магістерського дослідження

Основною метою цієї магістерської роботи є розробка комплексної системи обробки та зберігання паспортних даних для туристичної компанії, яка поєднує автоматизацію процесів, застосування технологій оптичного розпізнавання символів (OCR) та інтелектуальних алгоритмів машинного навчання для перевірки та корекції даних. Така система повинна забезпечити високий рівень точності, швидкості та безпеки обробки паспортних документів, що дозволить підвищити ефективність обслуговування клієнтів, оптимізувати внутрішні процеси та мінімізувати ризики фінансових втрат і репутаційних проблем.

Для туристичного бізнесу обробка паспортних даних має ключове значення. Кожна помилка у введенні даних – неправильна транслітерація імені, помилковий номер паспорта, невірна дата народження або термін дії документа – може призвести до відмови у посадці на рейс, скасування бронювання готелю або проблем із візовим оформленням. Такі випадки не лише створюють

незручності для клієнтів, але й несуть прямі економічні збитки для туристичної компанії та шкодять її репутації на ринку. У зв'язку з цим, автоматизація та інтелектуальна обробка паспортних даних стають ключовим фактором для підвищення якості сервісу і конкурентоспроможності компанії.

В межах цього дослідження передбачається вирішення таких завдань:

1. Автоматизація зчитування паспортних даних. Система повинна забезпечувати швидке і точне перетворення зображень паспортів у цифровий формат. Кожне поле паспорта – прізвище, ім'я, номер документа, дата народження, термін дії документа та інші реквізити – повинно бути виділено й оброблено автоматично. Це дозволить зменшити залежність від ручного введення даних і значно знизити ймовірність помилок, пов'язаних із людським фактором.
2. Впровадження методів машинного навчання для перевірки та корекції даних. Завдяки використанню інтелектуальних алгоритмів, система зможе перевіряти правильність транслітерації імен, форматування дат, а також відповідність даних міжнародним стандартам ІСАО. Машинне навчання дозволяє системі адаптуватися до різних форматів паспортів, враховувати особливості старих та нових документів і підвищувати точність розпізнавання з часом.
3. Забезпечення безпеки та конфіденційності даних. Оскільки паспортні дані є особистою інформацією клієнтів, система повинна гарантувати надійне зберігання даних, контроль доступу та захист від несанкціонованого доступу. Передбачено використання шифрування, системи прав користувачів та інших технологій кібербезпеки для мінімізації ризиків витоку інформації.
4. Оптимізація процесів обслуговування клієнтів. Автоматизована обробка паспортних даних дозволить значно скоротити час обслуговування одного клієнта, підвищити швидкість бронювання та уникнути затримок у наданні туристичних послуг. Це

безпосередньо підвищує рівень сервісу і покращує взаємодію з клієнтами, що важливо для збереження їхньої лояльності.

5. Створення бази даних для подальшого розвитку системи. Всі дані, що обробляються системою, будуть структуровані та збережені у базі даних, що дозволяє не лише забезпечити надійне зберігання, а й використовувати їх для навчання моделей машинного навчання, аналізу ефективності обробки документів та розвитку системи у майбутньому.

Розробка системи також передбачає створення архітектури, яка дозволяє інтегрувати модулі OCR та машинного навчання, реалізувати інтерфейс користувача для зручного введення і перевірки даних, а також налаштувати базу даних для ефективного зберігання і обробки великого обсягу паспортних документів.

Таким чином, постановка завдання магістерського дослідження полягає у створенні повноцінного програмного рішення, що дозволяє туристичній компанії:

- автоматизувати обробку паспортних даних;
- підвищити точність і надійність введення інформації;
- знизити час обслуговування клієнтів;
- забезпечити високий рівень безпеки персональної інформації;
- закласти основу для подальшого удосконалення та масштабування системи.

Розроблене рішення має не лише теоретичне значення, а й практичну користь для туристичних компаній: вона сприятиме оптимізації внутрішніх процесів туристичної компанії, зменшенню фінансових та репутаційних ризиків, а також дозволить вивчити ефективність інтеграції OCR та методів машинного навчання для автоматизованої обробки документів у реальних умовах діяльності туристичного бізнесу.

РОЗДІЛ 2 МОДЕЛЮВАННЯ СИСТЕМИ

2.1 Методологія моделювання системи

Проектування інформаційних систем неможливе без попереднього етапу моделювання, адже саме воно забезпечує перехід від абстрактних вимог замовника до чітко структурованих технічних рішень. Моделювання дозволяє виявити сутність предметної області, описати основні бізнес-процеси, визначити взаємозв'язки між ними та перевести неформалізовані вимоги у форму, зручну для подальшого аналізу й реалізації. Таким чином, моделювання виступає ключовим інструментом зниження ризиків при розробці системи, адже дозволяє уникнути непорозумінь між розробниками та замовником ще до початку етапу програмної реалізації [6].

Для моделювання систем у сучасній практиці розробки програмного забезпечення найширше застосовується Unified Modeling Language (UML) — уніфікована мова моделювання, яка є міжнародним стандартом (OMG UML Specification). UML забезпечує універсальні засоби для опису як статичних характеристик системи (її структура, класи, зв'язки між ними), так і динамічних аспектів (послідовність виконання процесів, взаємодія користувачів із системою тощо). Основна перевага UML полягає в тому, що вона є зрозумілою як для технічних спеціалістів, так і для представників бізнесу, завдяки використанню наочних діаграм [7].

У рамках цієї роботи UML застосовується для відображення двох ключових підходів до моделювання:

- Функціонального підходу, що дозволяє описати, які саме функції повинна виконувати система з точки зору користувачів. Для цього використовуються:
 - діаграми прецедентів, які показують, які дії може здійснювати користувач (зчитування даних паспорта чи їх перевірка в базі);

- діаграми послідовності, що демонструють логіку виконання певного сценарію роботи (взаємодія користувача, OCR-модуля та бази даних під час зчитування документа);
- діаграми активності, які дозволяють візуалізувати бізнес-процеси в цілому, виділяючи можливі паралельні або альтернативні гілки виконання.
- Об'єктно-орієнтованого підходу, що зосереджений на структурі системи та її внутрішній організації. Основним інструментом тут виступає діаграма класів, яка дозволяє описати об'єкти предметної області (паспорт, користувач, запис у базі даних, запит системи) та взаємозв'язки між ними. Використання цього підходу особливо важливе для підготовки до реалізації, оскільки він безпосередньо відображає майбутню архітектуру програмного продукту [8].

Застосування обох підходів у комплексі є виправданим у випадку розробки системи для туристичної компанії. По-перше, туристичний бізнес характеризується чітко визначеними бізнес-процесами (бронювання квитків, обробка паспортних даних, підтвердження поїздки), які доцільно описувати функціонально. По-друге, система працює з великими обсягами даних і має складну внутрішню логіку, що вимагає об'єктно-орієнтованого представлення.

Таким чином, методологія моделювання у даній магістерській роботі спирається на інтеграцію функціонального та об'єктно-орієнтованого підходів із використанням мови UML, що забезпечує комплексний опис системи як з точки зору бізнес-процесів, так і з точки зору її внутрішньої структури. Такий підхід дозволяє сформулювати повне уявлення про вимоги до майбутнього програмного продукту, створити теоретичне підґрунтя для його реалізації та підготувати набір моделей, зрозумілих як технічним спеціалістам, так і представникам туристичної компанії, що виступають кінцевими користувачами. У наступних підрозділах цієї роботи будуть детально розглянуті діаграми прецедентів, активності та послідовності, а також побудовано логічну модель даних і ER-діаграму, які відобразять ключові аспекти структури та поведінки розроблюваної системи.

2.2 Об'єктне та функціональне моделювання

2.2.1 Діаграма прецедентів

Діаграма прецедентів (англ. Use Case Diagram) є одним із ключових інструментів об'єктно-орієнтованого моделювання в UML (Unified Modeling Language). Вона дозволяє наочно відобразити функціональні можливості системи та взаємодію користувачів (акторів) із системою. Основна мета діаграми прецедентів — чітко показати, що саме може робити система з точки зору її користувачів, не заглиблюючись у внутрішню реалізацію, тобто не розглядати алгоритми або внутрішні структури даних, а зосередитись на зовнішній поведінці системи.

У контексті розроблюваної системи обробки та зберігання даних закордонних паспортів із елементами машинного навчання, діаграма прецедентів відображає основні сценарії роботи користувача з системою та ілюструє взаємозв'язок між функціональними блоками. Такий підхід дозволяє не лише розробникам, а й замовникам або кінцевим користувачам отримати цілісне уявлення про можливості системи та її логіку роботи.

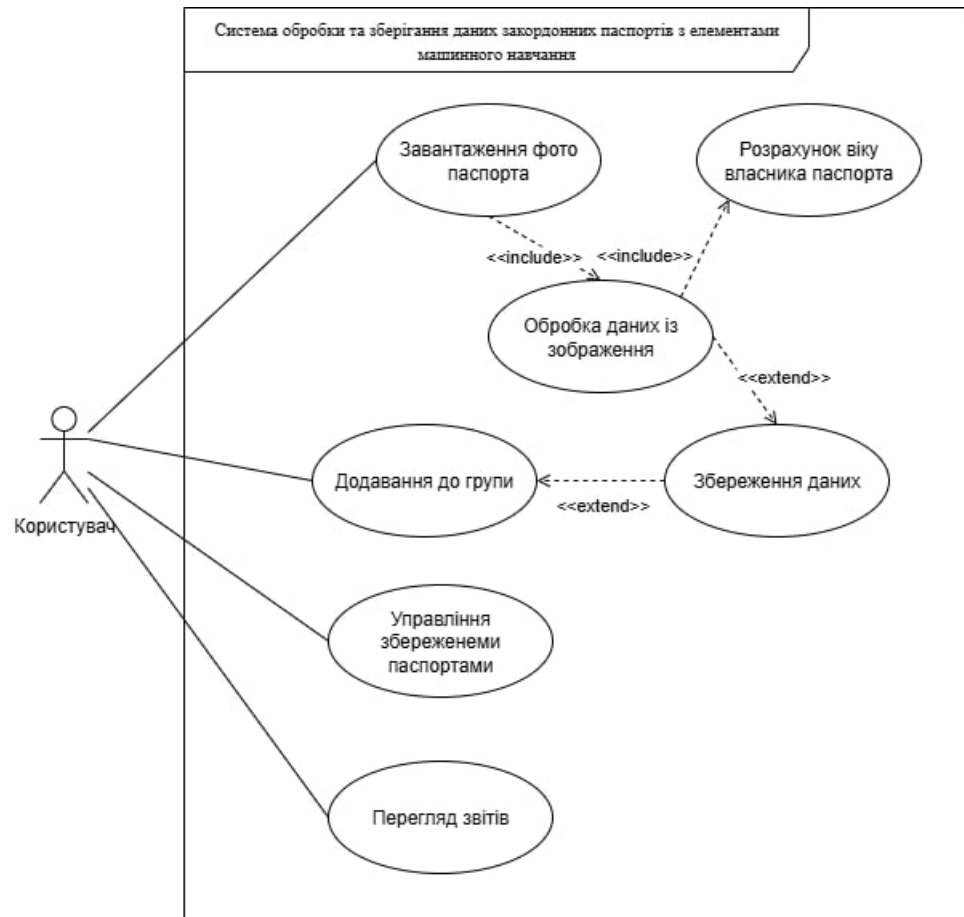


Рис. 1 Діаграма прецедентів

У представлений діаграмі прецедентів головним актором є Користувач. Він виступає зовнішнім учасником системи, який ініціює виконання основних сценаріїв. Актор взаємодіє із системою для досягнення власних цілей, зокрема — для завантаження паспортних даних, їх обробки, збереження та отримання необхідної аналітичної інформації.

Основні функціональні обов'язки актора включають:

1. Ініціювання завантаження фото паспорта для подальшої обробки.
2. Контроль збереження даних у системі.
3. Додавання оброблених записів до групи, якщо необхідно об'єднати декілька осіб.
4. Керування збереженими паспортами, що включає перегляд, редагування або видалення записів.
5. Отримання звітів про обробку даних.

Таким чином, Користувач виступає центральною ланкою взаємодії між реальним користувачем та програмною системою. Усі функціональні можливості діаграми спрямовані на задоволення потреб цього актора, що підкреслює користувацько-орієнтований підхід у проєктуванні системи.

На діаграмі виділено такі ключові функціональні прецеденти:

1. Завантаження фото паспорта — базова операція, яка дозволяє користувачу передавати зображення паспорта до системи для подальшого опрацювання. Цей прецедент є обов'язковим для виконання інших функцій системи, оскільки без наявності зображення неможливо провести автоматизовану обробку даних.
2. Обробка даних із зображення — центральний функціональний блок системи. Він включає застосування алгоритмів оптичного розпізнавання тексту (OCR) для виділення інформації з паспорта, перевірку правильності введених даних та можливе попереднє очищення та нормалізацію інформації. Обробка даних є ключовою для автоматизації введення паспортних даних та подальшого використання цієї інформації у системі.
3. Розрахунок віку власника паспорта — додатковий аналітичний прецедент, який демонструє можливості системи виконувати базові розрахунки на основі зчитаних даних. Цей функціонал може бути використаний, для перевірки відповідності вікових обмежень під час бронювання турів або отримання статистичних даних про клієнтів.
4. Збереження даних — функціонал, що забезпечує надійне збереження обробленої інформації у базі даних системи. Це дозволяє зберігати історію оброблених паспортів, забезпечувати цілісність та доступність інформації для подальшого використання, при формуванні груп туристів або створенні звітів.
5. Додавання до групи — сценарій, який дозволяє об'єднати кількох користувачів у групу, що подорожує разом. Це полегшує управління

бронюваннями, дозволяє швидко формувати списки учасників групових турів та оптимізувати обробку даних для групових запитів.

6. Управління збереженими паспортами — надає можливість редагувати, видаляти або переглядати інформацію про збережені паспорти. Цей функціонал забезпечує гнучке керування даними, що важливо для підтримки актуальності та достовірності інформації у системі.
7. Перегляд звітів — дозволяє користувачу отримувати аналітичну інформацію про оброблені дані, формувати звіти про виконані операції та стан обробки паспортних даних. Це допомагає оцінювати ефективність роботи системи та забезпечує можливість контролю за робочими процесами.

Нижче наведено докладний опис трьох основних прецедентів, які є центральними для роботи системи та безпосередньо пов'язані з ключовими функціями, описаними вище:

Прецедент «Завантаження фото паспорта»

1.1 Передумови:

- Користувач має цифрову версію фото паспорта, яку потрібно завантажити.

1.2 Тригер:

- Користувач ініціює процес завантаження фото паспорта через інтерфейс користувача.

1.3 Головний потік:

1. Користувач відкриває інтерфейс для завантаження фото.
2. Користувач вибирає файл фото паспорта для завантаження.
3. Система перевіряє тип файлу(A1) та його розмір(A2).
4. Система завантажує фото та підтверджує успішне завантаження.

1.4 Під-потіки: Відсутні

1.5 Альтернативні потоки:

- A1: Якщо користувач вибирає не підтримуваний формат файлу, система виводить повідомлення про помилку: "Непідтримуваний формат файлу", користувач повинен обрати інший файл або завершити прецедент.
- A2: Якщо файл перевищує максимальний розмір, система виводить повідомлення: "Файл занадто великий, будь ласка, виберіть менший файл", користувач повинен обрати інший файл або завершити прецедент.

Прецедент «Перевірка валідності паспорта»

1.1 Передумови:

- Користувач завантажив фото паспорта, і система готова до перевірки його валідності.

1.2 Тригер:

- Користувач натиснув кнопку "Отримати дані".

1.3 Головний потік:

1. Після завантаження фото паспорта система ініціює перевірку автентичності паспорта.
2. Система перевіряє паспорт на валідність, перевіряючи наступні параметри:
 - Порівняння з офіційними стандартами для перевірки справжності паспорта.
 - Аналіз водяних знаків, QR-кодів, шаблонів та інших елементів безпеки.
3. Якщо паспорт справжній, система підтверджує його валідність і переходить до етапу розпізнавання тексту паспорта.(A1)

1.4 Під-потоки: Відсутні

1.5 Альтернативні потоки:

- A1: Якщо система не може підтвердити його валідність то вона виводить повідомлення "Не вдалося перевірити валідність паспорта. Будь ласка, завантажте нове фото", користувач повинен завантажити нове фото або завершити прецедент.

Прецедент «Розпізнавання тексту паспорта»

1.1 Передумови:

- Фото паспорта перевірене на валідність.

1.2 Тригер:

- Після успішної перевірки валідності паспорта ініціюється процес розпізнавання тексту.

1.3 Головний потік:

1. Система ініціює процес розпізнавання тексту з завантаженого фото паспорта.
2. Алгоритм обробляє зображення, виділяючи текстові елементи (серія та номер паспорта, ім'я, дата народження тощо).
3. Система перевіряє точність розпізнаного тексту.
4. Якщо текст розпізнано коректно, система виводить отримані дані.(A1)

1.4 Під-потоки:

Відсутні

1.5 Альтернативні потоки:

- A1: Якщо система не змогла розпізнати текст то виводить повідомлення “Не вдалось розпізнати текстові елементи. Будь ласка, завантажте нове фото”, користувач повинен завантажити нове фото або завершити прецедент.

На діаграмі також використано спеціальні відносини «include» та «extend», які демонструють залежності між прецедентами:

- Відношення «include» означає обов'язкове виконання підпроцесу під час виконання основного прецеденту. Наприклад, операція «Обробка даних із зображення» включає завантаження фото паспорта та розрахунок віку користувача, що підкреслює їхню обов'язковість у процесі обробки.

- Відношення «extend» вказує на необов'язкові або додаткові функціональні можливості, які можуть бути виконані за потреби. Так, прецеденти «Додавання до групи» та «Збереження даних» можуть активуватися лише в певних умовах, що підкреслює гнучкість та масштабованість системи.

Завдяки такому представленню стає очевидною структура функціоналу системи, послідовність виконання основних сценаріїв та взаємозв'язки між різними прецедентами. Діаграма прецедентів слугує основою для подальшого детального моделювання, включаючи розробку діаграм послідовності, діаграм діяльності та діаграм класів, що забезпечує комплексний підхід до проектування та підтримки цілісності проектної документації.

2.2.2 Діаграма послідовності

Діаграма послідовності (Sequence Diagram) є важливим засобом моделювання динамічної поведінки системи, що дозволяє наочно показати порядок взаємодії між об'єктами під час виконання конкретного процесу або прецеденту використання. Вона демонструє, які повідомлення передаються між об'єктами, у якій послідовності відбуваються виклики методів і як дані переміщуються між компонентами системи. Такий підхід дозволяє чітко простежити логіку виконання операцій, виявити потенційні проблеми на етапі проектування та забезпечити узгодженість роботи різних функціональних блоків системи.

Перед тим як розглядати конкретні діаграми послідовності, важливо визначити та описати основні об'єкти, які беруть участь у процесі обробки паспортних даних. Кожен із цих об'єктів виконує свою специфічну роль у системі та взаємодіє з іншими компонентами для забезпечення коректного виконання функціональних сценаріїв. Нижче наведено детальний опис ключових об'єктів системи:

1. Користувач

Користувач є зовнішнім актором системи, який взаємодіє з нею для обробки паспортних даних. Основною його метою є завантаження файлу паспорта, контроль за процесом обробки даних та отримання результатів через інтерфейс користувача.

2. Форма

Форма є інтерфейсним об'єктом, що забезпечує взаємодію користувача із системою. Через форму користувач вводить або завантажує паспортні дані, а також отримує результати їх обробки. Форма контролює правильну передачу даних між користувачем і внутрішніми сервісами системи та відображає інформацію у зручному для користувача вигляді.

3. Клас Passport

Цей клас представляє паспортні дані, які зберігаються та обробляються системою. Об'єкти класу Passport містять інформацію, отриману від користувача та оброблену сервісами системи, включаючи серію та номер паспорта, ім'я власника, дату народження та інші дані. Passport є ключовою сутністю для перевірки валідності документа та передачі даних іншим компонентам системи.

4. Клас VerificationService

VerificationService відповідає за перевірку валідності паспортних даних. Для цього можуть застосовуватися алгоритми машинного навчання або інші методи валідації, що дозволяють виявляти помилки та невідповідності в паспортних даних. Цей сервіс забезпечує надійність і точність перевірки документів.

5. OCRService

OCRService відповідає за оптичне розпізнавання тексту з фото паспорта, завантаженого користувачем. Сервіс виділяє текстові елементи, забезпечує автоматичне введення даних у систему та передає їх у зручному форматі для подальшої обробки.

6. StorageService

StorageService керує збереженням оброблених паспортних даних у базі даних. Він забезпечує надійність зберігання, доступ до даних та їх оновлення, що дозволяє формувати історію обробки паспортів та контролювати цілісність інформації.

7. База даних

База даних слугує централізованим сховищем для всіх паспортних даних, отриманих та оброблених системою. Вона забезпечує доступ до даних для інших сервісів, формування звітів та підтримку історії обробки паспортів.

Для ілюстрації послідовності взаємодії об'єктів у системі були побудовані три діаграми послідовності, що відображають ключові етапи обробки паспортних даних.

На першому етапі користувач обирає файл із фото паспорта. Система перевіряє цей файл на відповідність допустимому формату та розміру. У разі перевищення допустимого розміру або вибору непідтриманого формату користувач отримує повідомлення про помилку та має можливість обрати інший файл або завершити прецедент. Якщо файл відповідає вимогам, користувачу стає доступною кнопка «Отримати дані», що дозволяє продовжити процес обробки. Цей етап взаємодії відображено на Рис. 2 Діаграмі послідовності “Завантаження фото паспорта”.

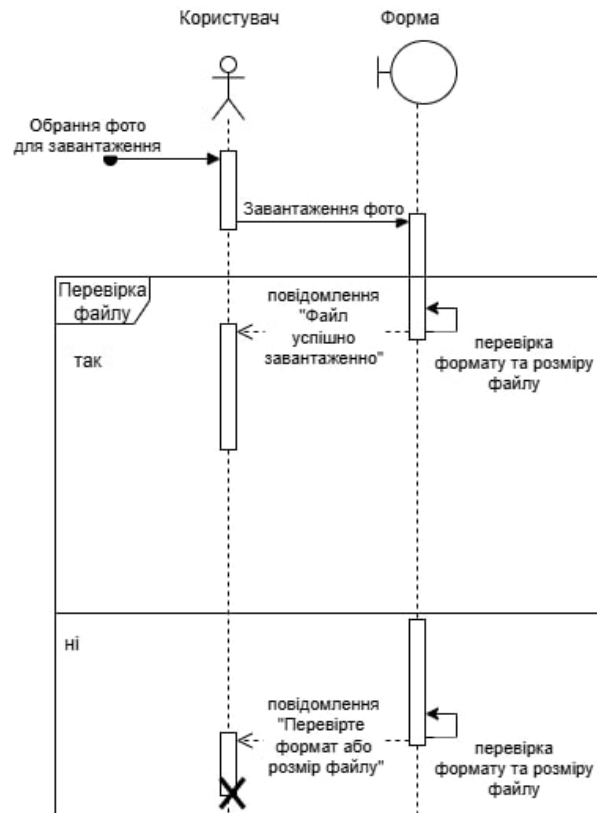


Рис. 2 Діаграмі послідовності “Завантаження фото паспорта”

На другому етапі фото паспорта передається до класу `VerificationService`, де відбувається перевірка валідності документа: перевірка наявності необхідних марок, водяних знаків та інших елементів безпеки, що підтверджують справжність паспорта. Після успішної валідації фото передається у сервіс `OCRService`, який здійснює оптичне розпізнавання тексту паспорта та виводить розпізнані дані на відповідну форму. Цей етап роботи системи відображено на Рис. 3 Діаграмі послідовності “Перевірка валідності та Розпізнавання тексту паспорта”.

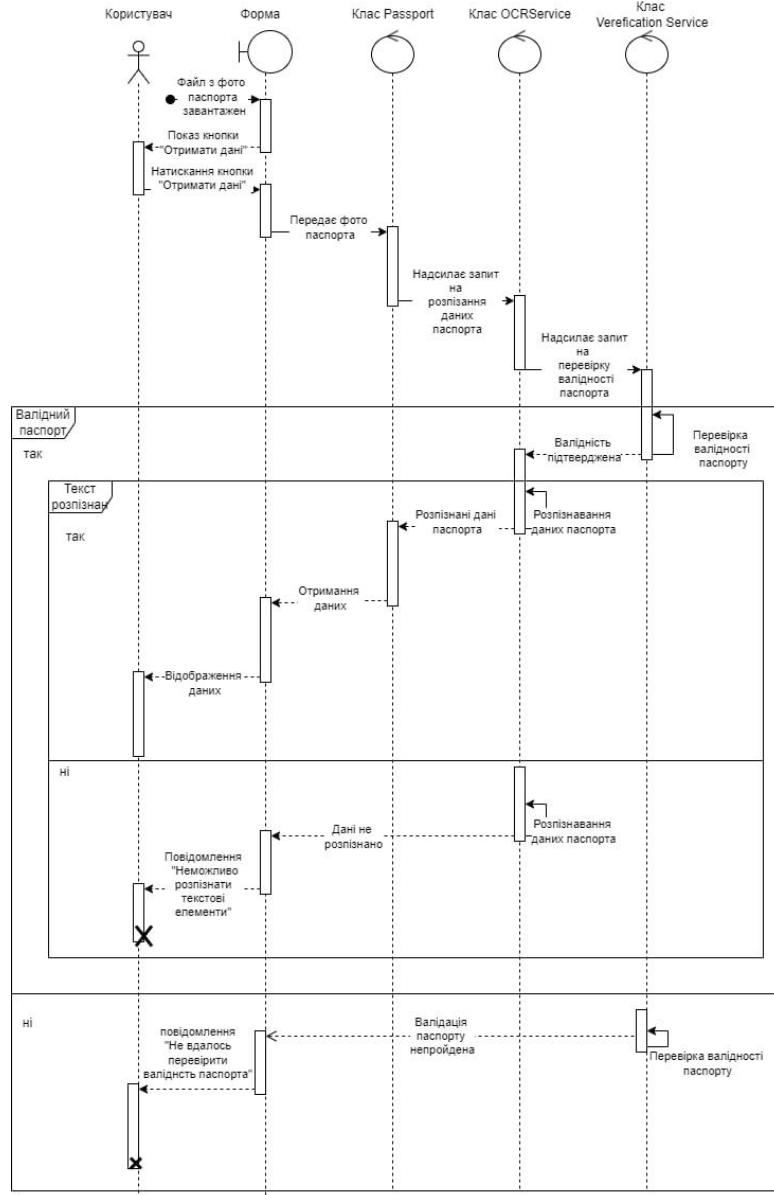


Рис. 3 Діаграмі послідовності “Перевірка валідності та Розпізнавання тексту паспорта”

На завершальному етапі користувачу пропонується зберегти отримані дані у базі даних через StorageService. Користувач може підтвердити збереження або відмовитися від нього. У разі підтвердження, дані зберігаються у базі даних для подальшого використання. Цей процес відображено на Рис. 4 Діаграмі послідовності “Збереження даних паспорта у БД”.

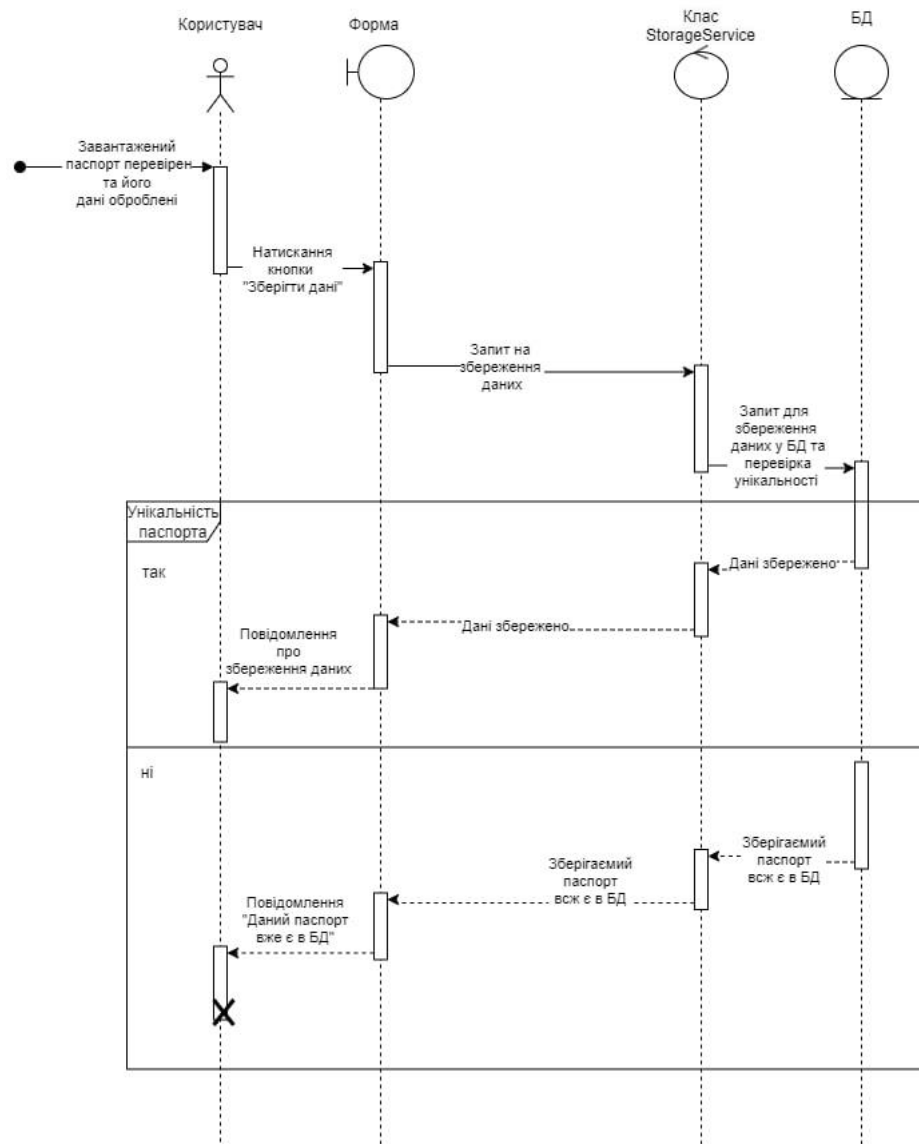


Рис. 4 Діаграмі послідовності “Збереження даних паспорта у БД”

Таким чином, усі три діаграми послідовності демонструють логічну та послідовну взаємодію користувача з системою та внутрішніми сервісами на різних етапах обробки паспортних даних, що дозволяє наочно простежити динаміку роботи системи від завантаження фото паспорта до збереження результатів обробки.

2.2.3 Діаграма діяльності

Діаграма діяльності (Activity Diagram) є одним із основних засобів функціонального моделювання системи в нотації UML. Вона використовується для відображення послідовності дій або бізнес-процесів, що виконуються у

системі, і дозволяє наочно представити логіку виконання певного процесу від початку до завершення. На відміну від діаграми послідовності, яка фокусується на взаємодії між об'єктами, діаграма діяльності зосереджується на загальному алгоритмі та потоках управління, показуючи, як виконуються дії та як переходи між ними визначають результат процесу.

Використання діаграми діяльності дозволяє формалізувати бізнес-процеси, виявити потенційні вузькі місця та підвищити зрозумілість моделі як для розробників, так і для майбутніх користувачів системи. Такий підхід забезпечує прозорість опису функціоналу системи та створює основу для подальшого проектування й реалізації.

Для системи обробки паспортних даних були побудовані три ключові діаграми діяльності, що відображають різні етапи роботи користувача та системи.

Перша діаграма діяльності представлена на рис. 5 вона ілюструє прецедент Завантаження фото паспорта. Вона складається з двох пулів: Користувач та Система. На діаграмі відображено, що користувач завантажує файл із фото паспорта, після чого система перевіряє його на відповідність допустимому формату та розмір. У залежності від результату перевірки система або приймає файл для подальшої обробки, або відхиляє його, повідомляючи користувача про помилку.

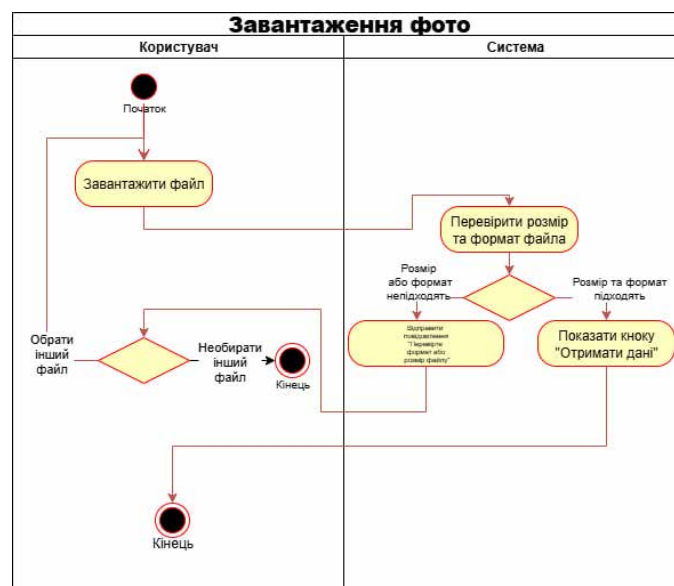


Рис. 5 Діаграма діяльності "Завантаження фото"

Друга діаграма діяльності представлена на рис. 6 вона демонструє процес обробки паспорта, який може розпочатися лише після успішного завантаження фото. Процес починається після того, як користувач натискає кнопку «Отримати дані». Система перевіряє зображення паспорта на валідність та можливість розпізнавання тексту. У разі виявлення помилок система виводить повідомлення про неможливість обробки; якщо ж перевірка пройдена успішно, система виводить на форму користувача відповідні дані, отримані з фото паспорта.

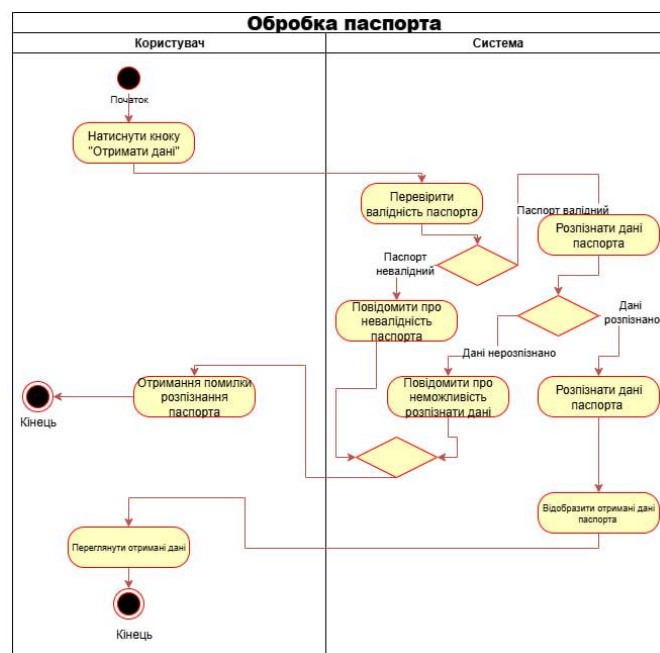


Рис. 6 Діаграма діяльності "Обробка паспорта"

Третя діаграма діяльності представлена на рис.7 вона відображає процес збереження даних паспорта у базі даних. Вона складається з трьох пулів: Користувач, Система та База даних. На діаграмі показано, що користувач ініціює збереження даних натисканням відповідної кнопки. Система перевіряє наявність аналогічного запису у базі даних; якщо запису немає, дані зберігаються, в іншому випадку користувач отримує повідомлення про існування такого паспорта.

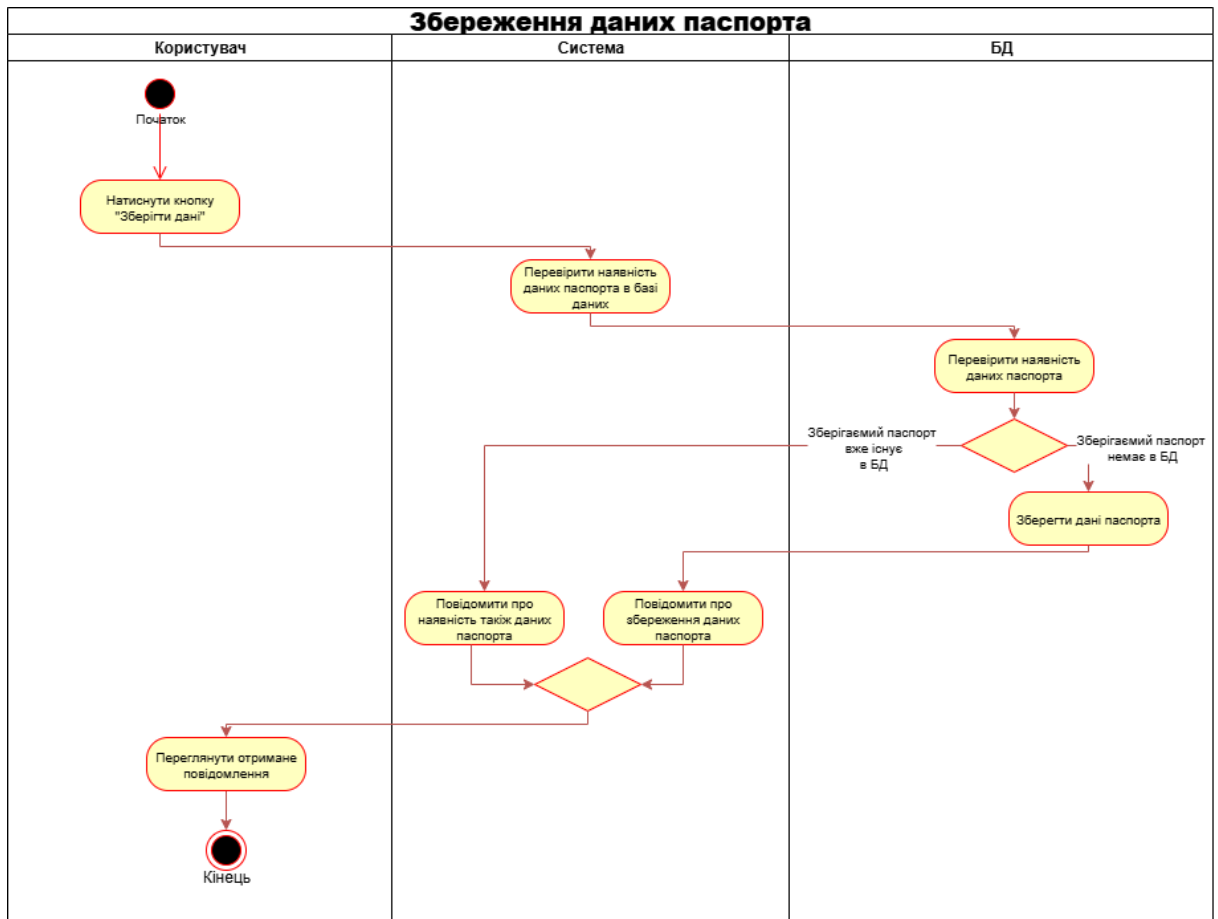


Рис. 7 Діаграма діяльності "Збереження даних паспорта"

Таким чином, усі три діаграми діяльності відображають логіку виконання ключових функціональних сценаріїв системи: завантаження фото паспорта, перевірку та обробку даних, а також їх збереження у базі даних. Вони дозволяють наочно простежити потоки управління та послідовність виконання дій, що сприяє точному розумінню функціонування системи та подальшій реалізації її алгоритмів.

2.2.4 Діаграма класів

Для відображення структури розробленої системи автоматизованої обробки паспортних документів у туристичних компаніях була побудована UML-діаграма класів. Діаграма демонструє основні сутності системи, їх атрибути, методи та взаємозв'язки, діаграма класів представлена на рис. 8.

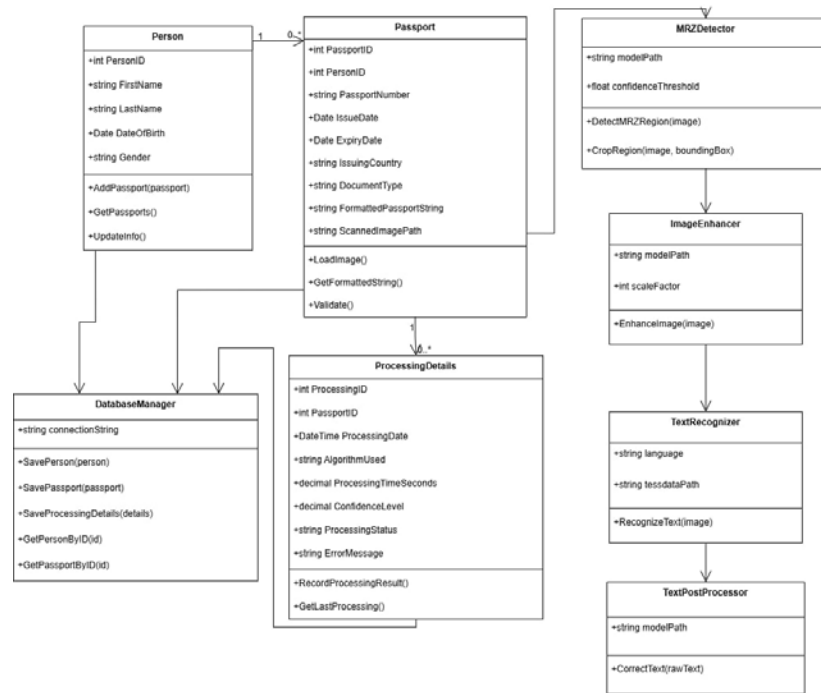


Рис. 8 Діаграма класів

На діаграмі представлені наступні класи:

1. **Person** – клас, що представляє клієнта туристичної компанії. Кожна особа може мати кілька паспортів, що дозволяє відстежувати історію документів. Основні атрибути включають `PersonID`, `FirstName`, `LastName`, `DateOfBirth`, `Gender`. Метод `AddPassport()` дозволяє додати новий паспорт до конкретної особи, `GetPassports()` повертає список усіх паспортів, пов'язаних із користувачем, а `UpdateInfo()` забезпечує редагування особистих даних.
2. **Passport** – представляє конкретний паспортний документ особи. В атрибутах зберігається унікальний `PassportID`, `PersonID`, `PassportNumber`, `IssueDate`, `ExpiryDate`, `IssuingCountry`, `DocumentType`, `FormattedPassportString` та шлях до сканованого зображення `ScannedImagePath`. Методи `LoadImage()` та `Validate()` забезпечують завантаження зображення та перевірку його формату, а `GetFormattedString()` повертає стандартний MRZ-рядок для подальшого використання.
3. **ProcessingDetails** – клас для збереження детальної інформації про обробку кожного паспорта алгоритмами OCR та машинного навчання. Включає атрибути `ProcessingID`, `PassportID`, `ProcessingDate`, `AlgorithmUsed`,

- ProcessingTimeSeconds, ConfidenceLevel, ProcessingStatus, ErrorMessage. Методи RecordProcessingResult() та GetLastProcessing() забезпечують збереження та отримання результатів обробки. Цей клас дозволяє оцінювати ефективність алгоритмів і відстежувати точність розпізнавання MRZ.
4. DatabaseManager – клас для організації взаємодії з базою даних. Забезпечує збереження та отримання інформації про осіб, паспорти та результати обробки. Основні методи: SavePerson(), SavePassport(), SaveProcessingDetails(), GetPersonByID(), GetPassportByID(). Клас гарантує централізоване управління даними та інтеграцію з іншими модулями системи.
 5. MRZDetector – відповідає за автоматичне виявлення MRZ-зони на зображенні паспорта. Використовує модель YOLOv8 для точного визначення області, навіть при складних умовах зйомки. Атрибути: modelPath, confidenceThreshold. Основні методи: DetectMRZRegion() для знаходження MRZ та CropRegion() для обрізання зображення по визначеній зоні.
 6. ImageEnhancer – клас для підвищення якості MRZ-зони перед розпізнаванням. Застосовується модель Real-ESRGAN для збільшення роздільної здатності, зменшення шумів і відновлення контурів символів. Атрибути: modelPath, scaleFactor. Основний метод: EnhanceImage(image), який повертає покращене зображення для подальшого розпізнавання.
 7. TextRecognizer – реалізує розпізнавання тексту з MRZ-зони за допомогою Tesseract OCR. Атрибути: language, tessdataPath. Метод RecognizeText(image) перетворює зображення на текстовий формат, який далі обробляється алгоритмами постобробки.
 8. TextPostProcessor – клас для виправлення типових помилок OCR та покращення якості результатів. Використовує character-level classifier для автоматичного виправлення неправильних символів та структурних

неточностей. Атрибут: `modelPath`. Метод: `CorrectText(rawText)` аналізує текст і повертає коректований MRZ-рядок.

2.3 Моделювання структури даних

Для забезпечення ефективного зберігання та обробки паспортних даних у системі була розроблена логічна структура даних, що формує основу майбутньої бази даних `PassportDB`. Основною метою моделювання структури даних є формалізація організації інформації у системі, забезпечення цілісності та взаємозв'язку даних, а також підготовка основи для реалізації функціоналу обробки та збереження паспортних даних, діаграма БД системи представлена на рис.9.

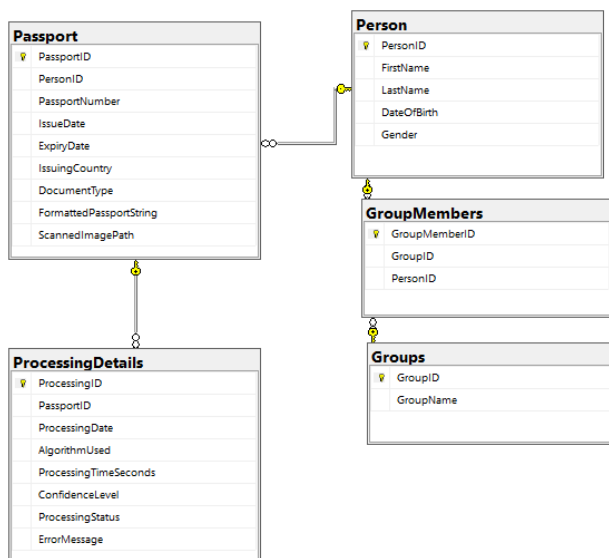


Рис. 9 База даних системи

Структура даних системи включає декілька основних сутностей, які відображають ключові об'єкти предметної області. До них належать:

1. `Person` — містить основну інформацію про користувачів системи:
 - `PersonID` — унікальний ідентифікатор користувача (PK);
 - `FirstName` — ім'я користувача;
 - `LastName` — прізвище користувача;
 - `DateOfBirth` — дата народження;
 - `Gender` — стать користувача.

2. Passport — зберігає паспортні дані користувачів та пов'язана з таблицею Person через зовнішній ключ:
 - PassportID — унікальний ідентифікатор паспорта (PK);
 - PersonID — зовнішній ключ на таблицю Person (FK);
 - PassportNumber — серія та номер паспорта;
 - IssueDate, ExpiryDate — дати видачі та закінчення дії паспорта;
 - IssuingCountry — країна видачі паспорта;
 - DocumentType — тип документа (закордонний паспорт, ID-карта тощо);
 - FormattedPassportString — відформатовані дані паспорта;
 - ScannedImagePath — шлях до збереженого фото паспорта.
3. ProcessingDetails — містить інформацію про обробку паспортних даних алгоритмами системи:
 - ProcessingID — унікальний ідентифікатор обробки (PK);
 - PassportID — зовнішній ключ на таблицю Passport (FK);
 - ProcessingDate — дата та час виконання обробки;
 - AlgorithmUsed — назва алгоритму або моделі машинного навчання;
 - ProcessingTimeSeconds — час обробки у секундах;
 - ConfidenceLevel — рівень впевненості алгоритму (%);
 - ProcessingStatus — статус обробки (успішно / помилка / частково);
 - ErrorMessage — повідомлення про помилку, якщо вона виникла.
4. Groups — таблиця для організації груп користувачів, наприклад сімей або колективів:
 - GroupID — унікальний ідентифікатор групи (PK);
 - GroupName — назва групи.
5. GroupMembers — таблиця зв'язків між особами та групами:
 - GroupMemberID — унікальний ідентифікатор запису (PK);
 - GroupID — зовнішній ключ на таблицю Groups (FK);
 - PersonID — зовнішній ключ на таблицю Person (FK).

Такий підхід до моделювання структури даних дозволяє впорядкувати інформацію, забезпечити логічні зв'язки між сутностями та уникнути дублювання. Це створює основу для надійного зберігання та ефективної обробки паспортних даних, що у свою чергу підвищує продуктивність і точність роботи всієї системи.

2.4 Визначення вимог до системи

Визначення вимог до системи є одним із ключових етапів процесу моделювання та проектування програмного забезпечення. Саме від чіткості та повноти сформульованих вимог залежить ефективність розробки, надійність системи та її відповідність потребам кінцевих користувачів. У рамках цієї магістерської роботи, яка присвячена створенню системи обробки та зберігання даних закордонних паспортів для туристичної компанії, вимоги формуються на основі аналізу предметної області, існуючих проблем ручного введення паспортних даних, а також сучасних методів автоматичного розпізнавання тексту (OCR) із використанням елементів машинного навчання.

Формування вимог ґрунтується на принципах, що забезпечують охоплення потреб туристичної компанії: скорочення часу обробки документів, підвищення точності введення даних, забезпечення безпеки та надійності роботи системи, а також створення зручного та інтуїтивного інтерфейсу для користувачів.

Система поділяється на функціональні та нефункціональні аспекти, які разом формують повне технічне завдання для розробки.

Функціональні вимоги визначають конкретні дії та можливості, які система повинна забезпечувати для користувача та для внутрішніх процесів. Основними функціональними вимогами проектованої системи є:

1. Автоматичне зчитування даних із паспортів

Система повинна забезпечувати швидке та точне перенесення паспортних реквізитів у базу даних. OCR-модуль із використанням елементів машинного навчання дозволяє розпізнавати текст навіть на документах із шумами, розмиттями чи захисними елементами.

2. **Перевірка коректності та повноти даних**
Система має виявляти помилки або пропуски у паспортних реквізитах і повідомляти користувача для виправлення. Це забезпечує зниження ризику неправильного оформлення бронювань і підвищує точність роботи компанії.
3. **Збереження та управління даними**
Дані клієнтів повинні зберігатися у структурованій базі даних із можливістю пошуку, редагування та оновлення записів. Забезпечення організованого доступу до інформації є необхідною умовою для швидкого та безпечного обслуговування.
4. **Інтерфейс користувача для працівників туристичної компанії**
Інтерфейс повинен бути зручним та інтуїтивно зрозумілим, що дозволяє менеджерам перевіряти, редагувати та контролювати дані без спеціальної технічної підготовки.
5. **Формування звітів та аналітичної статистики**
Система повинна надавати можливість формувати звіти про обсяг оброблених документів та ефективність роботи працівників, що сприяє аналізу процесів та прийняттю управлінських рішень.

Нефункціональні вимоги визначають якісні характеристики системи та умови її ефективної роботи, забезпечуючи безпеку, надійність та продуктивність:

1. **Точність та надійність розпізнавання даних.** Система повинна досягати точності розпізнавання не менше 95%, що забезпечує мінімізацію помилок при автоматичному введенні інформації.
2. **Швидкість обробки одного документа.** Час зчитування та обробки паспорта не повинен перевищувати 10 секунд, що дозволяє обслуговувати велику кількість клієнтів у короткий проміжок часу.
3. **Безпека та конфіденційність.** Персональні дані клієнтів мають бути захищені шляхом шифрування, а доступ до системи контролюватися за допомогою механізмів автентифікації та авторизації.

4. Масштабованість та гнучкість. Система повинна легко адаптуватися до зростання кількості користувачів і обсягів оброблюваних документів, а також до змін у форматах паспортів та вимог законодавства.
5. Зручність та доступність інтерфейсу. Графічний інтерфейс повинен бути інтуїтивно зрозумілим, щоб користувачі без спеціальної технічної підготовки могли швидко виконувати свої завдання.

Визначення та формалізація функціональних і нефункціональних вимог забезпечує чітку основу для подальшої розробки системи, гарантує її відповідність потребам туристичної компанії та створює передумови для високої точності, швидкості та безпеки обробки паспортних даних клієнтів.

РОЗДІЛ 3. РОЗРОБКА СИСТЕМИ

3.1 Вибір інструментальних засобів та середовища розробки

3.1.1 Середовище розробки та мова програмування

У процесі створення програмного забезпечення надзвичайно важливим є правильний вибір середовища розробки та мови програмування, оскільки від цього залежить продуктивність, зручність подальшої підтримки та можливості інтеграції системи з іншими технологіями.

Для розробки системи було обрано інтегроване середовище Microsoft Visual Studio. Воно є одним із найпотужніших інструментів для створення програмного забезпечення під платформу .NET Framework та .NET Core, забезпечує широкі можливості для роботи з різними мовами програмування, зокрема C#, VB.NET, C++ тощо. Visual Studio містить у собі зручні засоби налагодження коду, профілювання, підключення бібліотек через NuGet, а також інтегровану підтримку роботи з системами керування базами даних (SQL Server, MySQL, PostgreSQL тощо). Це значно спрощує розробку комплексних застосунків, до яких належить і система обробки паспортних документів.

Мовою програмування для реалізації проєкту обрано C# (C Sharp). Дана мова є однією з найбільш поширених у сучасній розробці програмного забезпечення та підтримується компанією Microsoft. C# поєднує у собі зручність високорівневих мов із продуктивністю низькорівневих підходів, забезпечує реалізацію об'єктно-орієнтованих принципів (інкапсуляція, спадкування, поліморфізм, абстракція) та має розвинену інфраструктуру стандартних бібліотек. Важливою перевагою C# є можливість роботи з різними технологіями одночасно: графічними інтерфейсами, мережевими підключеннями, машинним навчанням та обробкою зображень.

Для створення інтерфейсу користувача застосовано технологію Windows Forms (WinForms), яка є класичним рішенням для побудови настільних додатків у середовищі Windows. WinForms надає великий набір готових елементів

керування (таблиці, кнопки, панелі, текстові поля тощо), що дозволяє швидко та зручно розробити інтуїтивно зрозумілий інтерфейс. Крім того, WinForms легко поєднується з іншими бібліотеками, зокрема з Tesseract OCR, а також із компонентами машинного навчання (ML.NET), що є важливою складовою даної роботи.

Таким чином, поєднання Visual Studio, C# та WinForms є обґрунтованим вибором для розробки системи обробки паспортних документів. Це рішення забезпечує:

- швидку розробку та налагодження програмного забезпечення;
- простоту інтеграції з базами даних та зовнішніми бібліотеками;
- зручний графічний інтерфейс для кінцевих користувачів;
- можливість масштабування та подальшого розвитку системи, у тому числі із застосуванням елементів машинного навчання.

3.1.2 Обґрунтування вибору бібліотеки Tesseract OCR

У процесі створення системи обробки паспортних документів одним із ключових завдань є якісне розпізнавання текстової інформації з фотографій паспортів. Для вирішення цього завдання було обрано бібліотеку Tesseract OCR, яка є одним із найбільш відомих і перевірених у практиці інструментів оптичного розпізнавання символів з відкритим вихідним кодом.

Сучасна версія Tesseract 5.2.0 працює на основі глибинних нейронних мереж типу LSTM (Long Short-Term Memory), що забезпечує значно вищу точність порівняно з класичними статистичними методами. На відміну від багатьох комерційних систем, Tesseract є абсолютно безкоштовним та підтримує більше ніж 100 мов, включаючи українську та англійську, що є необхідним для обробки паспортних документів міжнародного зразка.

Вибір саме Tesseract зумовлений поєднанням кількох ключових переваг. По-перше, це його відкритий вихідний код і активна підтримка спільноти, що дозволяє адаптувати алгоритми під специфіку задачі без додаткових фінансових

витрат. По-друге, на відміну від простіших бібліотек OCR, Tesseract забезпечує високу точність розпізнавання навіть за умов низької якості зображень, що є типовим для фотографій документів, зроблених користувачами. По-третє, на відміну від комерційних сервісів (таких як ABBYY FineReader SDK чи Google Cloud Vision), Tesseract не потребує підключення до зовнішніх API, що робить систему автономною та безпечною з точки зору захисту персональних даних.

Крім того, існують офіційні обгортки для інтеграції Tesseract у середовище C#, зокрема через NuGet-пакети, що дозволяє без проблем використовувати його у Windows Forms застосунках. У межах даної роботи Tesseract застосовується для обробки та розпізнавання машинозчитуваної зони (MRZ) паспортів і витягнення ключових даних, таких як прізвище, ім'я, дата народження, номер паспорта та термін його дії. Саме поєднання високої точності, відкритості, автономності та сумісності із середовищем розробки визначило вибір Tesseract як основного інструменту OCR у цій системі.

3.1.3 Обґрунтування вибору ML.NET та його роль у системі

У спроектованій системі розпізнавання паспортних документів особливе значення має попередня обробка зображень перед виконанням OCR за допомогою Tesseract. Для реалізації цього функціоналу було обрано бібліотеку ML.NET, яка є платформою для інтеграції вже навчених моделей машинного навчання у додатки на базі .NET. Вибір ML.NET обґрунтований кількома ключовими факторами. По-перше, платформа має високу сумісність із мовою C# та середовищем WinForms, що дозволяє безпосередньо інтегрувати моделі машинного навчання у вже існуючу систему без потреби у сторонніх сервісах або складних зовнішніх бібліотеках. По-друге, ML.NET забезпечує зручний pipeline для обробки даних, виконання прогнозів та отримання результатів від нейронних мереж, що значно спрощує інтеграцію сучасних методів інтелектуальної обробки зображень. По-третє, платформа дозволяє виконувати моделі локально, що гарантує високу продуктивність, автономність системи та безпеку оброблюваних даних користувачів. Нарешті, ML.NET є гнучким і

масштабованим інструментом, що дозволяє у майбутньому підключати нові моделі або алгоритми для підвищення точності та функціональності системи.

У системі ML.NET використовується як runtime для вже навчених нейронних мереж, які були підготовлені поза межами C#-проекту. Основна модель – YOLO (You Only Look Once) – виконує детекцію паспорта та MRZ-зони на зображенні. Модель була навчена на великому наборі анотованих зображень паспортів різних країн: для кожного фото вручну вказано координати паспорта та MRZ-зони. Навчання включало стандартні методи аугментації, такі як повороти, зміни масштабу, варіації освітлення та додавання шуму, що забезпечує стійкість мережі до реальних умов зйомки, перекосів та часткової зашумленості. Після тренування мережа експортується у формат ONNX, а ML.NET у C#-проекті завантажує модель і виконує інференс – отримує координати MRZ-зони для подальшої обробки.

Отримані координати MRZ використовуються для вирізання та попередньої обробки зображення: вирівнювання нахилу (deskew), покращення контрасту, шумопригнічення та масштабування до оптимального розміру для Tesseract OCR. Це дозволяє стандартизувати зону MRZ і підвищити точність розпізнавання тексту.

Додатково застосовується постобробка тексту MRZ за допомогою нейронної мережі, яка тренувалася на парах «неправильний текст OCR → правильний MRZ». Модель здатна виправляти типові помилки OCR (наприклад, 0 ↔ O, 1 ↔ I, некоректне відображення символу <). ML.NET інтегрує цю модель у C#-проект, обробляє результати Tesseract та повертає виправлений текст MRZ.

Таким чином, ML.NET у системі виконує роль інструменту для інтеграції та виконання нейронних мереж, забезпечуючи два основні етапи машинного навчання: детекцію паспорта та MRZ на зображенні, а також корекцію помилок OCR. Підхід дозволяє поєднувати класичний OCR із сучасними методами машинного навчання, підвищуючи точність і надійність розпізнавання паспортних документів навіть у складних умовах (частково перекошені, затемнені або зашумлені зображення).

3.1.4 Вибір системи керування базами даних SQL Server та інструменту SSMS

Для зберігання та обробки інформації про паспортні документи в спроектованій системі було обрано Microsoft SQL Server як основну систему управління базами даних (СУБД) та SQL Server Management Studio (SSMS) як інтегроване середовище адміністрування і розробки баз даних. Такий вибір обґрунтований кількома ключовими факторами, що забезпечують надійність, зручність та ефективність роботи з даними у контексті реалізації системи.

По-перше, SQL Server є потужною реляційною СУБД, яка підтримує комплексне управління структурованими даними, надійні транзакції та багаторівневі механізми безпеки. Це дозволяє забезпечити цілісність і захист інформації користувачів, що є критично важливим при роботі з паспортними даними. База даних на SQL Server дозволяє реалізувати складні запити, фільтрування та сортування інформації, що значно полегшує обробку великих обсягів даних і підвищує продуктивність системи.

По-друге, SQL Server Management Studio (SSMS) забезпечує зручний інтерфейс для створення та адміністрування баз даних, розробки таблиць, запитів, процедур і тригерів, а також для контролю за безпекою та правами доступу користувачів. SSMS дозволяє швидко перевіряти коректність запитів, виконувати налагодження та аналізувати структуру бази даних, що значно спрощує процес розробки та подальшого супроводу системи.

Використання SQL Server та SSMS у спроектованій системі дозволяє реалізувати централізоване зберігання даних про паспорти та результати розпізнавання, ефективно організувати взаємодію між модулем обробки зображень і базою даних, а також забезпечує можливість масштабування системи у майбутньому. Крім того, завдяки підтримці стандарту SQL і потужним інструментам адміністрування, система отримує високий рівень гнучкості при розробці запитів, формуванні звітів та аналітичній обробці даних.

Таким чином, поєднання SQL Server та SSMS у спроектованій системі забезпечує надійну, масштабовану та ефективну платформу для зберігання та управління даними, що є невід'ємною складовою автоматизованої системи розпізнавання паспортних документів. Використання цих інструментів дозволяє забезпечити високий рівень безпеки, зручність адміністрування та гнучкість при розширенні функціоналу системи.

3.2 Архітектура спроектованої системи

Архітектура спроектованої системи базується на принципах трирівневої (layered) архітектури, що є класичним підходом до побудови сучасних програмних комплексів. Основна ідея полягає у поділі системи на окремі рівні, кожен з яких виконує чітко визначені функції і взаємодіє з іншими рівнями через стандартизовані інтерфейси. Така організація забезпечує не лише структурованість коду та зрозумілість логіки роботи, але й підвищує надійність, гнучкість і масштабованість програмного забезпечення.

На першому рівні знаходиться інтерфейс користувача, який реалізовано засобами Windows Forms у середовищі програмування C#. Цей рівень виконує роль «фронтенду» системи, забезпечуючи зручну взаємодію з кінцевим користувачем. Саме тут відбувається завантаження зображень паспортів, запуск процесів обробки, відображення результатів розпізнавання та надання можливості внесення змін у випадку виявлення неточностей. Користувацький інтерфейс також забезпечує доступ до додаткових функцій, таких як перегляд історії оброблених документів, формування вибірки даних або їх експорт. Важливою характеристикою цього рівня є інтуїтивність та простота, оскільки саме через нього здійснюється взаємодія людини з програмою.

Другий рівень становить бізнес-логіка, що фактично є «ядром» системи. Він реалізований у вигляді набору класів, які інкапсулюють алгоритми обробки даних і забезпечують виконання основних завдань. На цьому рівні здійснюється попередня обробка зображень паспортів: масштабування, корекція орієнтації, видалення спотворень та виділення машиннозчитуваної зони (MRZ). Для цього

використовується ML.NET, що дозволяє інтегрувати у систему моделі машинного навчання. Далі дані передаються до бібліотеки Tesseract OCR, яка відповідає за оптичне розпізнавання символів. Отримані результати піддаються аналізу та перевірці: обчислюється рівень точності розпізнавання, здійснюється виявлення можливих помилок, а у разі потреби користувачеві пропонується внести корективи вручну. Таким чином, бізнес-логіка виступає центральною ланкою між інтерфейсом користувача та рівнем роботи з даними, забезпечуючи цілісність процесу та правильність його виконання.

Третім рівнем є рівень доступу до даних, який реалізовано у вигляді окремого класу для взаємодії з системою керування базами даних Microsoft SQL Server. Цей рівень відповідає за усі операції, пов'язані зі збереженням та отриманням інформації: створення нових записів у базі, оновлення існуючих, пошук та вибірку даних, а також видалення застарілої інформації. Завдяки чіткій ізоляції цього модуля від бізнес-логіки система набуває більшої стійкості до змін: у випадку заміни СУБД або модифікації структури бази даних достатньо внести зміни лише у модуль доступу до даних, не зачіпаючи інші частини програми. Така інкапсуляція підвищує надійність і робить систему більш гнучкою.

Важливо підкреслити, що взаємодія між рівнями здійснюється у строго визначеному порядку. Інтерфейс користувача звертається виключно до бізнес-логіки, уникаючи безпосереднього доступу до бази даних. У свою чергу, бізнес-логіка виконує усі необхідні обчислення та підготовку даних, після чого передає їх модулю доступу до БД для збереження. Така схема не лише спрощує підтримку системи, але й робить її більш захищеною, адже зменшує ризик виникнення помилок, пов'язаних із некоректним зверненням до бази.

Запропонована архітектура має низку переваг. Вона дозволяє розмежувати відповідальність між різними частинами програмного комплексу, що спрощує супровід і тестування. Завдяки чіткій структурі легко інтегрувати нові алгоритми обробки зображень чи розпізнавання тексту без необхідності змінювати інтерфейс або рівень роботи з даними. Крім того, така архітектура створює умови

для масштабування: у разі необхідності можна розширити функціонал системи, додати нові модулі чи інтегрувати її з зовнішніми сервісами, не змінюючи при цьому основної логіки роботи.

Отже, архітектура спроектованої системи, побудована за принципами трирівневої організації, забезпечує збалансоване поєднання зручності користування, коректності виконання основних функцій та надійності збереження даних. Це створює надійну основу для подальшої експлуатації та удосконалення програмного комплексу, орієнтованого на автоматизовану обробку паспортних документів.

3.3 Реалізація модуля обробки зображень паспортів

3.3.1 Використання Tesseract OCR

В рамках реалізації системи для автоматизованої обробки паспортних даних було розроблено окремий програмний модуль OcrService, який відповідає за розпізнавання текстової інформації з паспортних документів за допомогою бібліотеки Tesseract OCR. Основна функція модуля полягає у трансформації графічного зображення паспорта у структуровані дані, придатні для подальшого зберігання та обробки в інформаційній системі. Використання Tesseract забезпечує високу точність розпізнавання MRZ-зони (Machine Readable Zone) навіть при наявності шумів чи незначних дефектів у зображенні.

Перед тим, як здійснити розпізнавання тексту, модуль OcrService завантажує зображення паспорта та приводить його до формату, сумісного з бібліотекою Tesseract OCR. Tesseract працює з типом Pix, тому будь-яке графічне зображення (JPEG, PNG, BMP) конвертується у внутрішнє представлення, яке дозволяє алгоритму ефективно аналізувати кожен піксель.

Для підвищення точності розпізнавання застосовується низка попередніх операцій обробки зображення:

1. Масштабування зображення – якщо фото паспорта має високий розмір, його зменшують до оптимального розміру для обробки, щоб Tesseract

міг швидко працювати без втрати інформації про символи. Невелике збільшення розміру дрібних деталей допомагає покращити точність розпізнавання MRZ-рядків.

2. Оптимізація кольорової палітри – оскільки паспорти мають фон різного кольору та освітлення, модуль переводить зображення у формат, який виділяє текстові області, зменшує вплив тіней, відблисків та інших шумів. Це може включати конвертацію у відтінки сірого або застосування контрастування.
3. Підготовка піксельних даних – Tesseract аналізує текст на основі структурованих піксельних масивів. Модуль конвертує зображення у відповідний формат Pix, нормалізує яскравість та контрастність, а також прибирає зайві шуми, що можуть спотворити OCR.
4. Вибір режиму сегментації сторінки (PageSegMode) – у коді застосовується PageSegMode.SingleBlock, що оптимально для MRZ-зони паспорта, оскільки вона являє собою компактний блок тексту. Такий режим дозволяє Tesseract ігнорувати зайві елементи на зображенні, наприклад фон, графічні елементи або логотипи, і фокусуватись на рядках машиннозчитуваного тексту.
5. Мова та тренувальні дані – модуль використовує спеціальну навчальну модель OCRB, яка заточена під розпізнавання MRZ у паспортних документах. Вона містить спеціальні шрифти та символи, що використовуються у машиннозчитуваних зонах. Використання правильного мовного пакета є критично важливим, оскільки стандартний англійський або універсальний набір символів може допускати помилки при розпізнаванні '<', цифр або літер O/I/L.
6. Обробка винятків та помилок – у процесі завантаження та підготовки зображення передбачена обробка можливих помилок, наприклад, пошкоджений файл, нестандартний формат або некоректне кодування пікселів. Це дозволяє уникнути аварійного завершення програми і забезпечує стабільну роботу модулю.

В цілому, підготовка зображення для Tesseract забезпечує основу для подальшого точного розпізнавання MRZ-зони. Завдяки цим крокам значно зменшується кількість помилок, пов'язаних із якістю фотографії паспорта, освітленням або кутом нахилу документа, що особливо важливо при автоматизованій обробці великих обсягів паспортних даних.

Після підготовки зображення модуль OcrService виконує пошук і виділення MRZ-зони (Machine Readable Zone) — спеціальної області паспорта, де міститься машинно-зчитуваний текст. Цей крок є критичним, адже точність OCR значною мірою залежить від того, наскільки чисто і правильно виділена зона для аналізу.

У коді ця операція реалізована методом ExtractMrzFromImage зображеному на рис.10.

```
public string ExtractMrzFromImage(string imagePath)
{
    try
    {
        using (var engine = new TesseractEngine(_tessDataPath, _language, EngineMode.Default))
        {
            using (var bmp = new Bitmap(imagePath))
            {
                using (var pix = PixConverter.ToPix(bmp))
                {
                    using (var page = engine.Process(pix, PageSegMode.SingleBlock))
                    {
                        string text = page.GetText();
                        return GetMrzFromText(text) ?? string.Empty;
                    }
                }
            }
        }
    }
    catch
    {
        return string.Empty;
    }
}
```

Рис. 10 Код методу ExtractMrzFromImage

Модуль обробки зображень паспортів у системі реалізований із застосуванням бібліотеки Tesseract OCR, яка дозволяє автоматично розпізнавати текст із фотографій документів. Основним завданням даного модуля є виділення MRZ-зони (Machine Readable Zone) паспорта та перетворення її в структуровані дані, придатні для подальшого збереження у базі даних та обробки. Для досягнення високої точності розпізнавання використовується спеціалізований мовний пакет ocrb, який адаптований під формат MRZ, що гарантує коректне визначення символів навіть у випадках низької якості зображення або при наявності шуму.

Процес роботи модуля починається із завантаження зображення паспорта у формат Bitmap. Оскільки бібліотека Tesseract працює з власним внутрішнім форматом Pix, зображення конвертується у відповідний формат за допомогою PixConverter.ToPix. Цей етап забезпечує правильне представлення пікселів, що є критично важливим для точного розпізнавання MRZ.

Розпізнавання тексту здійснюється із застосуванням режиму PageSegMode.SingleBlock, який дозволяє обробляти всю MRZ-зону як єдиний текстовий блок. Такий підхід допомагає уникнути помилок, пов'язаних із наявністю додаткових елементів на фотографії, фонового зображення, емблем чи водяних знаків, що можуть заважати розпізнаванню.

Після отримання тексту зображення проводиться детальна обробка рядків для виділення саме MRZ. Для цього застосовується нормалізація символів: усі літери переводяться у верхній регістр, пробіли та специфічні символи замінюються на стандартний знак <, а неприпустимі символи видаляються регулярними виразами. Також виконується перевірка довжини рядків та кількості роздільників <, щоб виділити дві рядкові структури довжиною 44 символи кожна, що відповідає міжнародному стандарту паспортів. У разі виникнення помилок розпізнавання або неповної MRZ застосовується резервний алгоритм, який дозволяє вибрати найбільш ймовірні рядки для подальшого парсингу.

Наступним етапом є розбір MRZ на окремі поля. У коді передбачено визначення прізвища та імені, дати народження та закінчення дії паспорта, номера паспорта, національності та статі власника. При цьому додатково виконується обчислення дати видачі паспорта відповідно до віку власника: для дітей до 16 років — термін видачі зменшується до 5 років, для дорослих — встановлюється стандартний термін 10 років. Окрім того, формується спеціальний рядок FormattedPassportString, який агрегує всі ключові дані у стандартизованому форматі для подальшої автоматизованої обробки.

Важливо зазначити, що весь процес побудований таким чином, щоб модуль залишався стійким до помилок OCR: навіть якщо частина символів

розпізнана некоректно, нормалізація та логіка вибору рядків MRZ дозволяють отримати валідні дані у більшості випадків. Такий підхід забезпечує надійне і ефективне функціонування системи при різних умовах фотографування паспортів, зменшує потребу у ручній корекції та значно підвищує точність автоматичного розпізнавання.

Після отримання MRZ-тексту з зображення паспорта необхідним етапом є нормалізація даних, оскільки OCR-алгоритми, навіть високоточні, часто генерують некоректні або нестандартні символи через низьку якість зображення, відблиски, шум або специфічні шрифти. Для цього в системі реалізовано метод `NormalizeMrzLine`, який забезпечує стандартизацію рядків MRZ перед подальшою обробкою. Основними операціями даного методу є переведення всіх символів у верхній регістр, заміна нетипових символів, таких як «, », ¡, | на стандартний символ '<', а також видалення пробілів і всіх символів, що не відповідають міжнародному стандарту MRZ (A–Z, 0–9, '<').

Окремо обробляються поля, що містять прізвище та ім'я власника документа, за допомогою методу `CleanNameField`. Цей метод замінює символ '<' на пробіли, видаляє зайві пробіли, а також приводить рядок до нормалізованого формату, що дозволяє уникнути помилок при збереженні даних у базі. Такий підхід є критично важливим, оскільки навіть незначні помилки в імені або прізвищі можуть призвести до некоректної роботи системи або проблем при подальшому використанні даних у автоматизованих процесах.

Після нормалізації виконується парсинг структурованих даних, під час якого інформація з MRZ розбивається на окремі поля. Прізвище та ім'я виділяються з першого рядка MRZ за допомогою роздільника '<<'. Дата народження витягується з другого рядка у форматі `ууММdd` і конвертується у більш зручний формат `dd.ММ.уууу`. Номер паспорта та національність очищаються від символів '<' та інших некоректних символів, що забезпечує відповідність стандартам і полегшує подальшу обробку. Стать власника визначається з одного символу ('M' або 'F'), а термін дії паспорта обчислюється на основі дати закінчення дії документа, отриманої з MRZ.

Особливої уваги заслуговує автоматичне обчислення дати видачі паспорта. Для неповнолітніх осіб (менше 16 років) від кінця терміну дії віднімається 5 років, тоді як для дорослих користувачів застосовується стандартний термін у 10 років. Така логіка дозволяє системі коректно визначати дати навіть у разі неповного або помилкового введення даних користувачем, що значно зменшує необхідність ручної корекції та підвищує точність обробки.

Критично важливою складовою модуля є контроль та обробка помилок, що забезпечує надійність і стабільність роботи системи. Модуль здійснює перевірку довжини рядків MRZ, контролює наявність обов'язкових символів, перевіряє правильність формату дат і забезпечує коректність алфавітних полів, таких як прізвище, ім'я та національність. Завдяки цьому ймовірність помилок при автоматичному розпізнаванні значно зменшується, а оператор або користувач витрачає менше часу на ручні виправлення.

Для ілюстрації роботи механізму нормалізації MRZ наведено невеликий фрагмент коду зображений на рис.11 .

```
private string NormalizeMrzLine(string line)
{
    if (string.IsNullOrEmpty(line)) return line;
    string s = line.ToUpperInvariant();
    s = s.Replace('«', '<').Replace('»', '<').Replace(';', '<').Replace(':', '<');
    s = s.Replace(" ", "");
    s = Regex.Replace(s, @"[^A-Z0-9<]", "");
    return s;
}
```

Рис. 11 Метод нормалізації MRZ

Даний метод демонструє ключові етапи підготовки тексту для подальшого парсингу та обробки структурованих даних. Завдяки такому підходу модуль забезпечує високу точність розпізнавання MRZ, стабільність роботи системи і мінімізує вплив помилок OCR на результати автоматичної обробки паспортних документів.

Таким чином, розроблений модуль OcrService забезпечує високий рівень автоматизації обробки паспортних даних, підвищує точність розпізнавання та формує стандартизовані результати, готові для подальшого використання системою.

3.3.2 Використання Tesseract у поєднанні з методами МН

У рамках дослідження було реалізовано комбінований підхід до автоматичного розпізнавання MRZ на паспортних фотографіях, який поєднує класичний OCR Tesseract із сучасними методами машинного навчання для забезпечення високої точності та надійності обробки. Основна мета полягала у створенні системи, здатної самостійно визначати наявність паспорта та MRZ-зони на зображенні, покращувати якість цієї зони для підвищення точності OCR і виправляти можливі помилки після розпізнавання тексту.

На першому етапі обробки застосовано модель YOLOv8 (You Only Look Once) для виявлення паспорта та MRZ на зображеннях. Модель YOLOv8 є сучасною глибокою згортковою нейронною мережею (CNN, Convolutional Neural Network), спеціалізованою на задачах об'єктного детектування. Архітектура YOLOv8 включає оптимізовані згорткові шари для виділення ознак, шари підвибірки (Pooling Layers) для зменшення розмірності та шари передбачення, які формують bounding boxes та ймовірності класів об'єктів.

Для навчання моделі використано корпус паспортних зображень, вручну анотованих координатами паспорта та MRZ-зони. Під час навчання застосовувалися стандартні методи аугментації даних — повороти, масштабування, зміни освітлення та додавання шуму, що забезпечило здатність моделі стабільно виявляти MRZ навіть при перекосах або зашумленості. Після навчання модель YOLOv8 здатна точно передбачати координати MRZ на нових зображеннях без ручного кодування правил. Для інтеграції моделі в .NET-проект вона була експортована у формат ONNX, а ML.NET використано для завантаження моделі, обробки прогнозів та отримання координат MRZ, які в подальшому використовуються для обрізки та попередньої обробки зображення.

Після виявлення та вирізання MRZ-зони застосовується додатковий етап покращення зображення за допомогою моделі Real-ESRGAN (Enhanced Super-Resolution Generative Adversarial Network). Ця нейронна мережа виконує підвищення роздільної здатності (super-resolution), усуває шум, покращує

контраст і чіткість символів навіть на низькоякісних або стиснених фотографіях. Real-ESRGAN базується на генеративних змагальних мережах (GAN), де дві нейронні мережі — генератор і дискримінатор — змагаються між собою: генератор створює покращені зображення, а дискримінатор оцінює їх якість, наближаючи результат до реалістичного. Модель Real-ESRGAN використовується в системі у вже навченому вигляді (претренована модель на великому наборі зображень з шумами, розмиттям і низькою роздільністю), тому не потребує додаткового навчання. Вона інтегрована як етап між детекцією YOLOv8 і OCR, забезпечуючи значне покращення видимості символів MRZ перед передачею до Tesseract. Це підвищує точність розпізнавання, особливо для фотографій з низькою якістю, тінями або невеликими нахилами документа.

На наступному етапі застосовується Tesseract OCR для розпізнавання тексту MRZ. Незважаючи на попереднє покращення, можливі залишкові помилки, пов'язані зі специфікою шрифту OCR-B або з незначними артефактами. Для їх корекції використовується додатковий етап машинного навчання, реалізований у вигляді character-level classifier модель. Ця нейронна мережа прогнозує правильні символи на основі контексту вхідного тексту, виправляючи типові помилки Tesseract, такі як неправильне розпізнавання символів 0 ↔ O, 1 ↔ I або < ↔ пробіл. Модель постобробки інтегрована в C#-проект за допомогою ML.NET, що дозволяє автоматично обробляти результати OCR та формувати фінальний, коректний MRZ-текст.

Таким чином, у запропонованій системі машинне навчання застосовується на трьох рівнях:

1. Детекція паспорта та MRZ-зони — модель YOLOv8;
2. Покращення якості зображення MRZ — модель Real-ESRGAN;
3. Постобробка результатів OCR — character-level модель.

Поєднання цих етапів дозволило досягти високої точності навіть на фотографіях з низькою якістю, перекосами чи шумами. Інтеграція ML-моделей у .NET-проект через ML.NET забезпечила універсальність і продуктивність

рішення. У результаті система демонструє підвищену стійкість та точність у порівнянні з класичним використанням OCR без етапів машинного навчання.

3.4 Інтерфейс користувача

У розробленій системі реалізовано графічний інтерфейс користувача (GUI), створений на базі технології Windows Forms (WinForms). Основна мета інтерфейсу полягає у забезпеченні зручного доступу користувача до функціоналу програми, зокрема — завантаження паспортних фотографій, виконання обробки та збереження результатів у базі даних.

На рисунку 12 подано головне вікно системи у вкладці «Обробка».

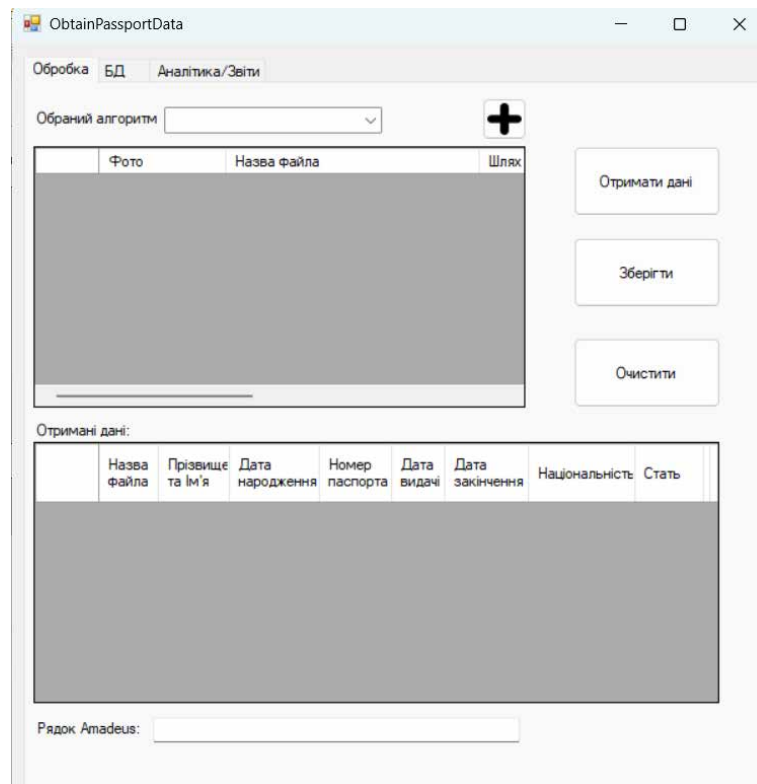


Рис. 12 Вкладка «Обработка»

Інтерфейс складається з кількох основних функціональних блоків:

- Панель завантаження даних (верхня частина вікна). Користувач може додати одне або декілька зображень паспортів для обробки, використовуючи кнопку з іконкою «+», вибір зображень показано на рис. 13.

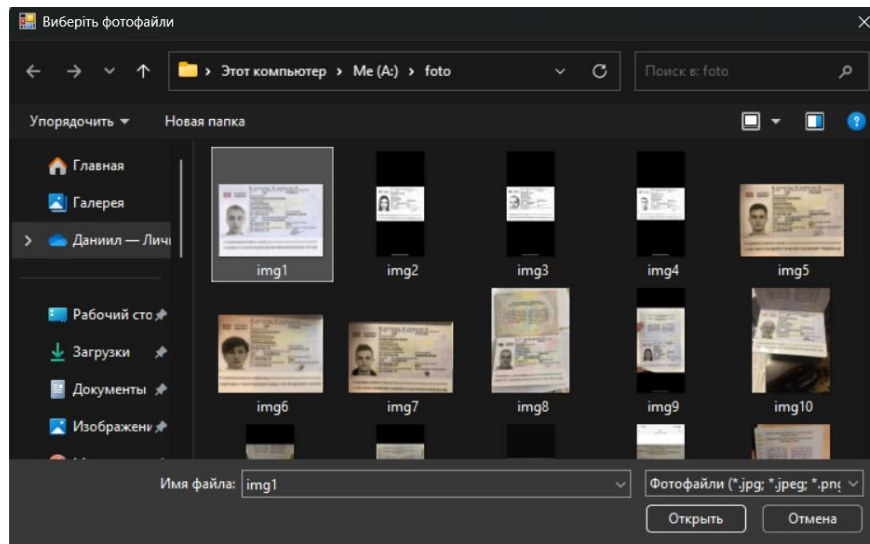


Рис. 13 Вибір файлів для обробки

- Після завантаження у таблиці відображаються такі параметри: фотографія документа, назва файлу та шлях до нього. Це забезпечує візуальний контроль користувача за вхідними даним. Також після завантаження обирається потрібний алгоритм для обробки.

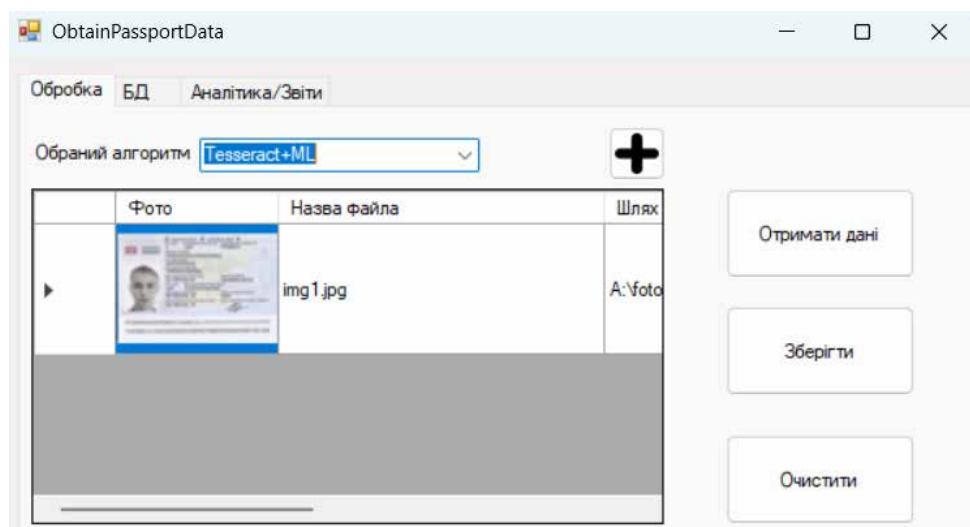


Рис. 14 Завантажене зображення

- Кнопки керування процесом обробки (права частина вікна):
 - «Отримати дані» — запуск основного алгоритму обробки, що включає виявлення MRZ-зони, попередню обробку зображення, OCR-розпізнавання та застосування ML-моделей для виправлення помилок.

- «Зберегти» — запис отриманих даних у базу даних SQL Server.
- «Очистити» — видалення проміжних результатів і очищення робочої області для завантаження нових зображень.
- Таблиця результатів (нижня частина вікна). У ній відображаються результати обробки паспорта:
 - назва файлу,
 - прізвище та ім'я,
 - дата народження,
 - номер паспорта,
 - дата видачі,
 - дата закінчення дії документа,
 - національність,
 - стать.

Таким чином, користувач одразу бачить повністю структуровану інформацію, яку можна зберегти чи використати для подальшої обробки, результати обробки показані на рис. 15.

Отримані дані:

	Назва файла	Прізвище та ім'я	Дата народження	Номер паспорта	Дата видачі	Дата закінчення	Національність	Стать
▶	img1...	ПРОКО...	03.06.2003	FV698614	08.0...	08.07.2029	UKR	М

Рядок Amadeus: `srdocshk1-p-UKR-FV698614-UKR-03jun03-M-08jul29-PROKOPENKO/DANIIL`

Рис. 15 Результати обробки

- Поле «Рядок Amadeus» (внизу вікна). Це додатковий вихідний блок, який дозволяє сформувати рядок даних у спеціальному форматі (для подальшого експорту чи інтеграції з іншими системами).

Розроблений інтерфейс орієнтований на простоту використання. Він не перевантажений елементами, має інтуїтивно зрозумілу структуру та дає змогу виконати всі ключові операції (завантаження, обробка, збереження) у кілька кліків.

Таким чином, інтерфейс користувача є важливим елементом системи, оскільки забезпечує не лише взаємодію з користувачем, але й контроль за правильністю роботи алгоритмів, візуалізацію отриманих результатів та можливість інтеграції з зовнішніми системами.

Також система дає можливість завантажувати та обробляти кілька зображень одночасно, це представлено на рис.

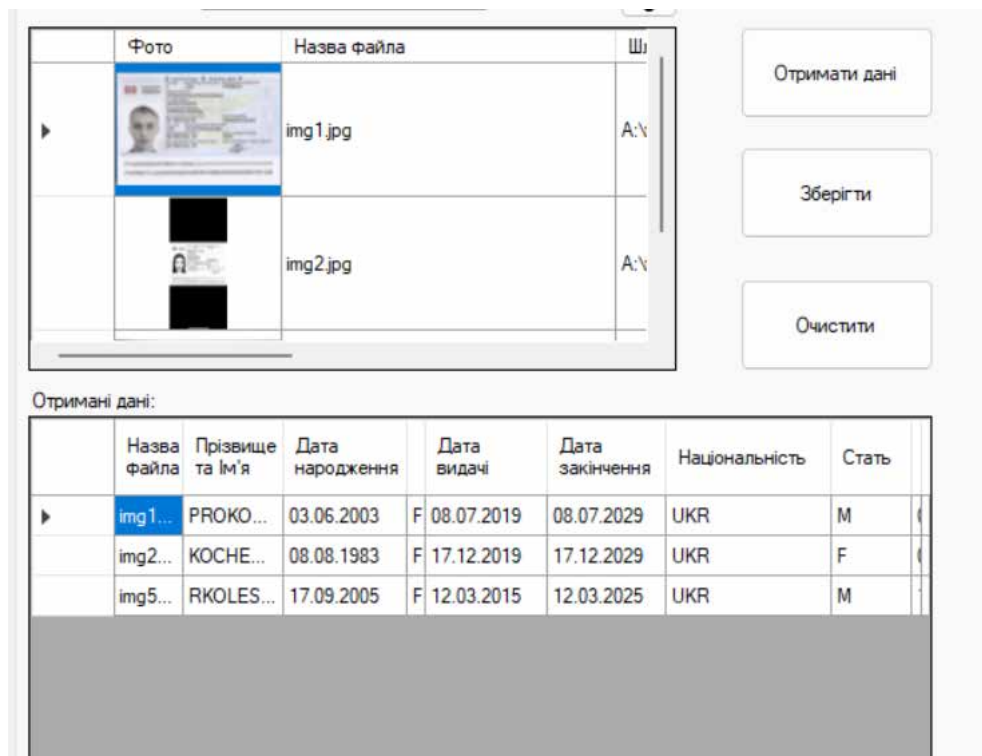


Рис. 16 Обробка декількох зображень

У вкладці «БД» реалізовано інтерфейс для роботи з базою даних, у якій зберігається інформація про розпізнані паспорти. Основна мета цієї вкладки —

надання користувачеві інструментів для перегляду, редагування та організації вже збережених записів, вкладка «БД» представлена на рис. 17.

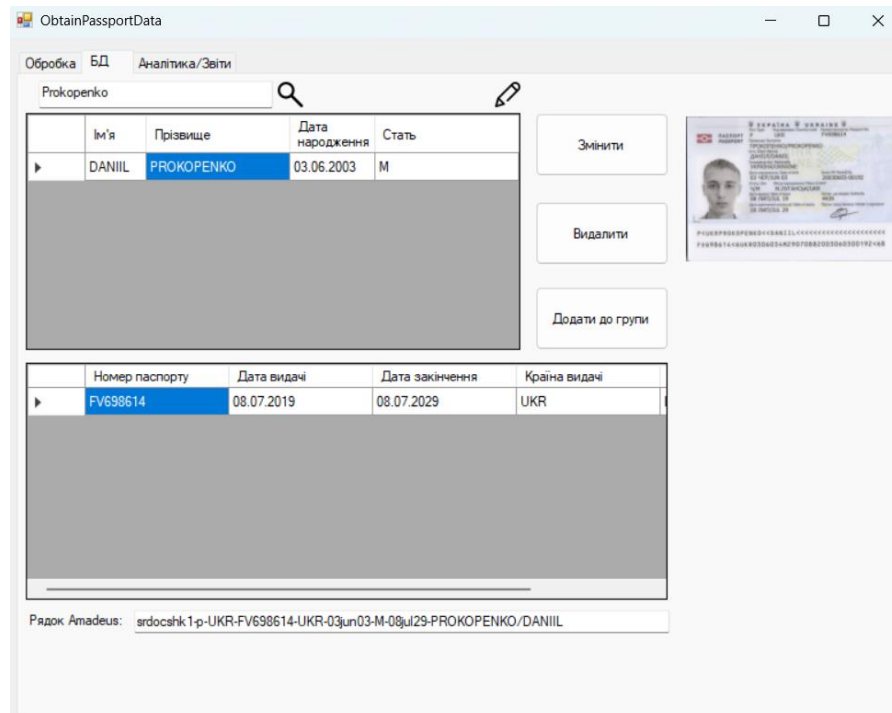


Рис. 17 Вкладка «БД»

Інтерфейс складається з таких функціональних блоків:

- Список користувачів та основні дані (верхня таблиця). У ній відображаються:
 - ім'я,
 - прізвище,
 - дата народження,
 - стать.

Користувач може швидко знайти потрібний запис, використовуючи поле пошуку або навігацію по таблиці.

- Деталізація паспорта (нижня таблиця). Після вибору конкретного користувача відображаються його паспортні дані:
 - номер паспорта,

- дата видачі,
- дата закінчення,
- країна видачі.

Це забезпечує можливість працювати з кількома паспортами одного користувача.

- Блок керування записами (права частина вікна):
 - «Змінити» — редагування даних користувача або паспорта.
 - «Видалити» — видалення обраного запису з бази даних.
 - «Додати до групи» — додавання користувача до певної групи для зручної організації (для формування аналітичних вибірок чи звітів).
- Модуль візуалізації. Справа передбачене відображення сканованого зображення паспорта користувача. Це дає змогу швидко співставити отримані OCR-дані з оригінальним документом і переконатися в їхній правильності.
- Рядок Amadeus (нижня частина вікна). Як і у вкладці «Обробка», дані автоматично конвертуються у спеціалізований формат (для подальшого використання у туристичних чи авіаційних системах).

Таким чином, вкладка «БД» виконує роль центру керування даними користувачів і паспортів. Вона поєднує перегляд, редагування та організацію інформації, а також забезпечує інтеграцію з іншими модулями системи.

РОЗДІЛ 4. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

4.1 Вимоги до апаратного та програмного забезпечення

Для забезпечення коректного функціонування системи автоматизованої обробки паспортних даних необхідно дотримуватися певних вимог до апаратного та програмного забезпечення. Виконання цих вимог гарантує стабільність роботи, оптимальну продуктивність і можливість подальшого розвитку та масштабування програмного комплексу.

Система передбачає обробку зображень високої роздільної здатності, виконання алгоритмів машинного навчання для попередньої обробки зображень, а також роботу з базою даних. Тому до апаратного забезпечення висуваються специфічні вимоги, які забезпечують швидкість виконання обчислень, коректність обробки зображень та надійне збереження інформації.

Мінімальні та рекомендовані параметри апаратного забезпечення представлені наступним чином:

1. Процесор

Для виконання алгоритмів розпізнавання тексту та попередньої обробки зображень необхідний процесор із достатньою обчислювальною потужністю. Мінімально допустимий процесор має частоту 2,0 ГГц, що дозволяє запускати систему та виконувати базові операції. Для оптимальної роботи та зменшення часу обробки зображень рекомендовано багатоядерний процесор із частотою від 2,5 ГГц, що забезпечує паралельне виконання алгоритмів і більш швидко інтеграцію моделей машинного навчання.

2. Оперативна пам'ять

Обсяг оперативної пам'яті визначає здатність системи обробляти великі масиви даних одночасно. Мінімально достатній обсяг складає 4 ГБ, що забезпечує базову роботу програми та завантаження невеликої кількості зображень. Рекомендований обсяг — 8 ГБ і більше, що дозволяє одночасно обробляти декілька документів, виконувати алгоритми

машинного навчання та підтримувати стабільність роботи системи при високому навантаженні.

3. Накопичувач (HDD/SSD)

Для зберігання програмного забезпечення, бібліотек та бази даних потрібен вільний простір не менше 2 ГБ. Використання SSD значно підвищує швидкість завантаження даних і обробки зображень, що особливо важливо при роботі з великими обсягами документів та високою роздільною здатністю сканів.

4. Відеоадаптер

Для базових операцій достатньо інтегрованої графічної карти, однак для прискореної обробки зображень і підвищення продуктивності алгоритмів машинного навчання рекомендовано використовувати відеоадаптер із апаратним прискоренням обробки графіки.

5. Монітор

Для коректного відображення інтерфейсу користувача та зручного перегляду оброблених зображень рекомендується використовувати монітор із роздільною здатністю не менше 1280×720 пікселів. Для підвищення комфортності роботи та точності візуальної перевірки даних краще використовувати монітор Full HD або з більшою роздільною здатністю.

Програмне забезпечення системи складається з операційної системи, середовища розробки, бібліотек для реалізації алгоритмів машинного навчання та OCR, а також системи керування базами даних.

1. Операційна система

Система розроблена для роботи під керуванням Windows 10 або новіших версій. Вибір ОС зумовлений сумісністю з .NET Framework, WinForms та іншими бібліотеками, необхідними для обробки зображень і виконання алгоритмів ML.

2. Середовище розробки

Microsoft Visual Studio обрано як основне середовище розробки через його

інтеграцію з мовою програмування C#, підтримку WinForms та зручні засоби налагодження коду.

3. Бібліотеки та фреймворки

- .NET Framework / .NET 6+ забезпечує платформу для реалізації бізнес-логіки системи та підтримки об'єктно-орієнтованого програмування.
- ML.NET використовується для попередньої обробки зображень, визначення орієнтації документа, виділення MRZ-зони та підвищення точності OCR.
- Tesseract OCR служить для оптичного розпізнавання символів із зображень паспортів, забезпечуючи базову функціональність системи.

4. Система керування базами даних

Microsoft SQL Server використовується для зберігання та організації даних, а SQL Server Management Studio (SSMS) забезпечує адміністрування бази та контроль її структури. Така організація дозволяє ізолювати роботу з базою даних від бізнес-логіки та інтерфейсу користувача, підвищуючи надійність і безпеку системи.

Дотримання зазначених вимог дозволяє забезпечити стабільну роботу системи, виконання алгоритмів машинного навчання та оптичного розпізнавання тексту, а також створює надійну основу для подальшого розвитку та інтеграції додаткового функціоналу.

4.2 Хід виконання дослідження

Під час виконання дослідження основна увага була зосереджена на розробці та тестуванні модулю автоматизованої обробки паспортних документів для потреб туристичного агентства. Для досягнення мети було проведено послідовну роботу, що включала аналіз предметної області, розробку алгоритмів обробки зображень та оцінку результатів їх застосування у практичних умовах.

Першим етапом стало детальне вивчення структури паспортних документів, особливо машиннозчитуваної зони (MRZ), яка містить критично важливу інформацію про власника паспорта. Було визначено, що якість зображень, що надходять для обробки, може значно варіюватися, тому для дослідження використовувалися два типи фото: високої якості та низької якості. Це дозволило оцінити стабільність роботи системи у реальних умовах, коли паспорти можуть бути сфотографовані у різних умовах освітлення та під різним кутом.

Другим етапом стало впровадження алгоритмів розпізнавання тексту за допомогою Tesseract OCR, а також комбінованого підходу Tesseract OCR + машинне навчання (ML). Алгоритми були інтегровані у єдиний модуль обробки зображень, який здійснює наступні кроки: завантаження зображення паспорта, перетворення його у формат, придатний для OCR, розпізнавання MRZ, нормалізацію тексту та витяг ключових структурованих даних (прізвище, ім'я, дата народження, номер паспорта, національність, стать та термін дії документа). Машинне навчання використовувалося для підвищення точності розпізнавання у випадках низькоякісних зображень, що дозволяло компенсувати можливі помилки базового OCR.

На третьому етапі виконувалося систематичне тестування та збір даних для аналізу результатів роботи системи. Було оцінено точність розпізнавання MRZ, середній час обробки зображень та стійкість алгоритмів для різних типів фото. Дані були зведені у таблиці, що дозволило наочно порівняти ефективність

стандартного Tesseract OCR та Tesseract OCR + ML, а також визначити переваги та обмеження кожного підходу.

Особливу увагу під час дослідження приділено практичній придатності системи у туристичному агентстві, де час обробки даних та мінімізація ручного введення є критично важливими. Було проведено експерименти, які показали, що автоматичне формування рядка паспортних даних значно скорочує час підготовки інформації та знижує ймовірність помилок оператора. Результати підтвердили доцільність використання комбінованого підходу, коли стандартний OCR доповнюється алгоритмами машинного навчання.

Таким чином, хід виконання дослідження охопив три основні напрямки: аналіз предметної області, розробку та впровадження алгоритмів обробки MRZ, а також оцінку ефективності системи у реальних умовах. Зібрані експериментальні дані дозволяють не лише підтвердити працездатність розробленого модуля, а й зробити обґрунтовані висновки щодо оптимізації алгоритмів для практичного використання у сфері туризму.

4.3 Аналіз результатів роботи

Для оцінки ефективності розробленого модуля обробки паспортних документів було проведено тестування з використанням двох підходів: Tesseract OCR та Tesseract OCR + ML. Критеріями оцінки обиралися точність розпізнавання MRZ, середній час обробки зображень, а також стійкість алгоритмів при обробці паспортів різної якості.

Таблиця 4.1

Порівняння Tesseract OCR та Tesseract OCR + ML

Критерій	Tesseract OCR	Tesseract OCR + ML
Точність розпізнавання(фото з високою якістю)	100%	100%
Точність розпізнавання (фото низькою якістю)	32.6%	79.1%
Середній час обробки зображення	1.26 с.	1.47 с.
Відсканований паспорт з гарною якістю (сер. час)	1.18 с.	1.37 с.
Фото паспорта з меншою якістю (сер. час)	1.33 с.	1.56 с.

З аналізу таблиці 4.1 можна зробити кілька ключових висновків:

1. Обидва алгоритми забезпечують максимальну точність при обробці зображень високої якості, що підтверджує стабільність базової роботи Tesseract OCR.
2. Для фото низької якості додавання машинного навчання дозволяє значно підвищити точність розпізнавання — з 32,6% до 79,1%. Це свідчить про ефективність ML-моделі у покращенні обробки складних або нечітких зображень.

3. Середній час обробки зображень для обох алгоритмів практично однаковий, що вказує на те, що використання ML не значно збільшує витрати часу на обробку.
4. При роботі з відсканованими паспортами високої якості час обробки навіть трохи скорочується при використанні ML, що демонструє оптимізацію алгоритмів та їхню адаптивність до різних типів вхідних даних.
5. Для низькоякісних фото невелике збільшення часу обробки компенсується значним покращенням точності, що робить комбінацію Tesseract + ML більш надійним рішенням для реальних сценаріїв.

На основі отриманих експериментальних даних, проведеного аналізу результатів обробки різних типів паспортних зображень та оцінки ключових характеристик роботи алгоритмів, можна детально узагальнити переваги та недоліки кожного підходу, виокремивши їхні сильні та слабкі сторони у різних умовах використання.

Таблиця 4.2

Переваги та недоліки алгоритмів

Алгоритм	Основні переваги	Основні обмеження/недоліки
Tesseract OCR	Швидка обробка простих зображень, стабільна робота з високоякісними сканами, безкоштовний	Низька точність на низькоякісних зображеннях, слабка стійкість до шумів і викривлень
Tesseract OCR + ML	Висока точність навіть на низькоякісних зображеннях, покращена обробка складних MRZ, адаптивність до різних типів документів	Незначне збільшення часу обробки, потребує попереднього навчання моделі, додаткові ресурси для ML

Аналіз таблиць 4.1 та 4.2 дозволяє зробити наступні висновки:

- Використання ML-моделі значно підвищує точність обробки паспортів у складних умовах (нечіткі зображення, низька роздільна здатність), що

робить систему придатною для реального використання у туристичних агентствах.

- Tesseract OCR залишається ефективним для простих і високоякісних сканів, але не забезпечує стабільності при поганому освітленні або нестандартному куті фотографування.
- З точки зору практичного застосування, комбінований підхід Tesseract + ML забезпечує оптимальний баланс між точністю та швидкістю.

Також Для наочного порівняння ефективності алгоритмів Tesseract OCR та Tesseract OCR + ML було побудовано діаграму за категоріями точності розпізнавання. Діаграма дозволяє оцінити розподіл результатів за трьома рівнями точності: низька (0–50 %), середня (51–70 %) та висока (71–100 %). Такий підхід дозволяє наочно демонструвати переваги комбінованого використання машинного навчання для підвищення надійності та стабільності розпізнавання MRZ на різних типах зображень, діаграма представлена на рис. 18.

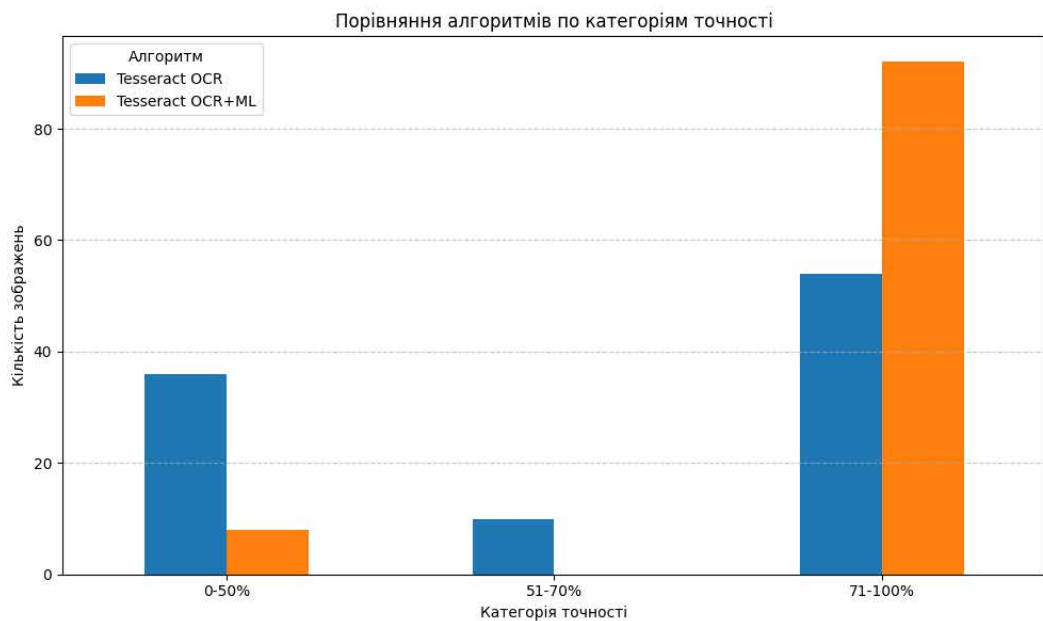


Рис. 18 Гістограма “Порівняння алгоритмів по категоріям точності”

На основі гістограми, що відображає розподіл результатів роботи алгоритмів Tesseract OCR та Tesseract OCR + ML за категоріями точності, можна зробити наступні висновки:

1. У категорії низької точності (0–50%) алгоритм Tesseract OCR продемонстрував значно гірші результати — велика кількість зображень була розпізнана з мінімальною точністю. У той час як при використанні Tesseract OCR + ML кількість таких випадків зменшилась майже у п'ять разів. Це свідчить про суттєве зниження ризику отримання некоректних результатів.
2. У середній категорії (51–70%) Tesseract OCR мав помітну кількість результатів, тоді як для Tesseract OCR + ML таких випадків практично не спостерігається. Це підтверджує, що інтеграція методів машинного навчання дозволяє уникати "середніх" результатів і підвищує стабільність роботи.
3. У високій категорії (71–100%) Tesseract OCR + ML значно перевищує Tesseract OCR за кількістю успішно розпізнаних зображень. Якщо базовий алгоритм досяг лише часткового успіху, то у поєднанні з ML кількість високоточних результатів зросла майже вдвічі.

Отже, використання додаткових методів машинного навчання у поєднанні з Tesseract OCR дозволяє значно підвищити точність розпізнавання та стабільність системи, зменшити кількість критично низьких результатів і забезпечити більшу кількість випадків із максимальною якістю. Це підтверджує доцільність інтеграції ML-моделей у систему обробки паспортних даних.

Також для практичного застосування розробленої системи було проведено експеримент, спрямований на оцінку ефективності автоматичного формування готового рядка паспортних даних у стандартному форматі для подальшого використання в системі бронювання авіаквитків. Розглянуто два сценарії: автоматичне формування через систему та ручне введення даних оператором.

При автоматичному формуванні рядка система після сканування паспорта та обробки MRZ генерує готовий рядок у форматі:

srdocshk1-p-[країна видачі паспорта]-[номер паспорта]-[національність]-[дата народження]-[стать]-[кінцевий термін дії паспорта]-[Прізвище/Ім'я]

де:

- [країна видачі паспорта] — трилітерний код країни, що видала паспорт (наприклад, UKR)
- [національність] — трилітерний код країни видачі паспорта (наприклад, UKR для України),
- [номер паспорта] — унікальний ідентифікатор документа, очищений від некоректних символів,
- [дата народження] — у форматі ddMMMyу, наприклад 03jun03,
- [стать] — символ М для чоловіка або F для жінки,
- [кінцевий термін дії паспорта] — дата закінчення терміну дії у форматі ddMMMyу,
- [Прізвище/Ім'я] — повне ім'я власника документа, де прізвище та ім'я розділяються символом /.

Таким чином, для будь-якого паспорта система автоматично формує повністю готовий рядок, який можна одразу вставляти у систему бронювання без додаткових змін або перевірок. Відповідний процес зайняв у середньому 12 секунд, що включає завантаження зображення, розпізнавання MRZ, нормалізацію тексту та форматування рядка.

Для порівняння, ручне введення тих самих даних оператором у стандартному форматі зайняло б приблизно 45 секунд, що у понад 3,5 рази довше. Крім економії часу, автоматичний підхід також забезпечує стандартизовану структуру даних, мінімізує ймовірність помилок через людський фактор і дозволяє підвищити загальну продуктивність процесу обробки паспортних документів у системі.

Цей приклад наочно демонструє, що автоматизація процесу генерації рядка `srdocshk1` не лише скорочує час обробки, а й підвищує точність та надійність введення даних у корпоративні системи бронювання та обліку.

Отже, проведений аналіз результатів роботи розробленого модуля обробки паспортних документів дозволяє зробити наступні підсумкові висновки.

Використання базового алгоритму Tesseract OCR забезпечує високу точність для відсканованих документів високої якості та є швидким рішенням для стандартних сценаріїв. Проте при роботі з низькоякісними зображеннями його ефективність значно знижується, що обмежує застосування у реальних умовах з різною якістю вхідних даних.

Комбінований підхід Tesseract OCR + ML показав значне покращення точності розпізнавання для складних та нечітких MRZ, при цьому збережено прийнятний середній час обробки зображень. Додаткові алгоритми машинного навчання дозволяють зменшити кількість критично низьких результатів і забезпечити більшу стабільність роботи системи.

Практичні експерименти підтвердили, що автоматичне формування рядка паспортних даних у стандартному форматі значно скорочує час обробки: від 12 секунд при автоматичному отриманні до 45 секунд при ручному введенні. Це не лише підвищує продуктивність операторів, а й мінімізує ймовірність помилок через людський фактор, забезпечуючи стандартизований і готовий до використання результат.

Таким чином, результати дослідження свідчать про доцільність використання комбінованого підходу з інтеграцією машинного навчання для підвищення точності, стабільності та надійності системи обробки паспортних документів. Це підтверджує практичну ефективність розробленого модуля та його готовність до впровадження у реальні робочі процеси.

ВИСНОВКИ

У межах виконання магістерської роботи створено систему автоматизованої обробки паспортних документів, орієнтовану на потреби туристичних агентств. Основною метою дослідження було розроблення модуля розпізнавання MRZ-зони, який автоматично формує структуровані дані паспорта та дає змогу інтегрувати їх у корпоративні системи бронювання. Для реалізації цієї мети виконано низку ключових етапів, зокрема розроблено модуль обробки зображень на базі Tesseract OCR і вдосконалено точність розпізнавання завдяки використанню алгоритмів машинного навчання.

Під час експериментальної частини проведено тестування алгоритмів на двох типах вхідних зображень — високої та низької якості. Результати показали, що стандартний Tesseract OCR демонструє найкращу точність на якісних сканах паспортів, однак значно гірше працює з фотографіями низької якості, де точність знижувалася до 32,6 %. Використання додаткових ML-алгоритмів дало змогу підвищити цей показник до 79,1 %, при цьому середній час обробки зображень майже не змінився. Отримані результати підтвердили ефективність комбінованого підходу, який поєднує базовий OCR із машинним навчанням і забезпечує стабільну якість обробки паспортних даних незалежно від умов зйомки.

Практичне впровадження системи показало помітну економію часу під час обробки документів: автоматичне формування готового рядка для введення в Amadeus триває близько 12 секунд, тоді як ручне введення займає приблизно 45 секунд. Крім скорочення часу, розробка забезпечує єдину структуру даних, зменшує ризик помилок, пов'язаних із людським фактором, і підвищує загальну ефективність роботи туристичного агентства.

Отже, розроблена система повністю відповідає поставленим завданням: вона забезпечує швидке, надійне й точне розпізнавання MRZ-паспортів та автоматично формує готові рядки для корпоративних систем бронювання. Створені алгоритми можуть застосовуватися у реальних умовах туристичної

діяльності, а також стати основою для подальшого вдосконалення — розширення підтримки паспортів інших держав, додавання нових типів ідентифікаційних документів і підвищення стійкості роботи за складних умов зйомки.

Підсумовуючи, результати дослідження підтвердили доцільність поєднання методів OCR і машинного навчання для автоматизації обробки паспортних даних. Такий підхід забезпечує кращу точність, швидкість і надійність роботи системи, що має практичну цінність для туристичної галузі.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Machine-readable passport. Wikipedia. URL: https://en.wikipedia.org/wiki/Machine-readable_passport (дата звернення: 18.09.2025).
2. ICAO Doc 9303. Machine Readable Travel Documents. International Civil Aviation Organization, 2021. 488 p.
3. Liu Y., Chen H., Xu J., Zhang L. MRZ code extraction from visa and passport documents using convolutional neural networks. arXiv preprint, 2020. URL: <https://arxiv.org/abs/2009.05489> (дата звернення: 18.09.2025).
4. Biometric passport. Wikipedia. URL: https://en.wikipedia.org/wiki/Biometric_passport (дата звернення: 18.09.2025).
5. Rosebrock A. OCR Passports with OpenCV and Tesseract. PyImageSearch, 2021. URL: <https://pyimagesearch.com/2021/12/01/ocr-passports-with-opencv-and-tesseract/> (дата звернення: 18.09.2025).
6. Sommerville I. Software Engineering. – 10th ed. – London : Pearson, 2015. 810 p.
7. Booch G., Rumbaugh J., Jacobson I. The Unified Modeling Language User Guide. – 2nd ed. – Boston : Addison-Wesley, 2005. 482 p.
8. Fowler M. UML Distilled: A Brief Guide to the Standard Object Modeling Language. – 3rd ed. – Boston : Addison-Wesley, 2018. 208 p.
9. Smith R. An overview of the Tesseract OCR engine. Proceedings of the International Conference on Document Analysis and Recognition (ICDAR). 2007. P. 629–633.
10. Tesseract OCR. GitHub repository. Google, 2022. URL: <https://github.com/tesseract-ocr/tesseract> (дата звернення: 20.09.2025).
11. Jaderberg M., Simonyan K., Vedaldi A., Zisserman A. Reading text in the wild with convolutional neural networks. International Journal of Computer Vision. 2016. Vol. 116, № 1. P. 1–20.

12. Goodfellow I., Bengio Y., Courville A. Deep Learning. – Cambridge : MIT Press, 2016. 775 p.
13. Krizhevsky A., Sutskever I., Hinton G. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. 2012. Vol. 25. P. 1097–1105.
14. LeCun Y., Bengio Y., Hinton G. Deep learning. Nature. 2015. Vol. 521. P. 436–444.
15. Gonzalez R. C., Woods R. E. Digital Image Processing. – 4th ed. – London : Pearson, 2018. 1168 p.
16. Bradski G., Kaehler A. Learning OpenCV 3: Computer Vision in C++ with the OpenCV Library. – Sebastopol : O'Reilly Media, 2016. 1024 p.
17. Shapiro L., Stockman G. Computer Vision. – New Jersey : Prentice Hall, 2001. 580 p.
18. Jurafsky D., Martin J. Speech and Language Processing. – 3rd ed. draft. – Stanford : Stanford University, 2021. 950 p.
19. Bishop C. Pattern Recognition and Machine Learning. – Berlin : Springer, 2006. 738 p.
20. Russell S., Norvig P. Artificial Intelligence: A Modern Approach. – 4th ed. – London : Pearson, 2021. 1136 p.
21. Koul A., Ganju S., Kasam M. Practical Deep Learning for Cloud, Mobile, and Edge. – Sebastopol : O'Reilly Media, 2019. 584 p.
22. Patel A., Singh R., Verma P. Object character recognition with neural networks. Journal of Computer Applications. 2020. Vol. 43, № 2. P. 55–62.
23. He W., Zhang X., Yin F. Efficient scene text detection with deep learning. Pattern Recognition Letters. 2018. Vol. 107. P. 36–44.
24. Almazán J., Gordo A., Fornés A., Zisserman A. Scene text recognition and retrieval. International Journal of Computer Vision. 2015. Vol. 113, № 2. P. 123–142.

25. Krishnan P., Mathew M., Jawahar C. Data augmentation strategies for text recognition. arXiv preprint, 2021. URL: <https://arxiv.org/abs/2104.08400> (дата звернення: 22.09.2025).
26. Passport OCR using machine learning. Medium, 2022. URL: <https://towardsdatascience.com/passport-ocr-using-machine-learning-123456> (дата звернення: 22.09.2025).
27. OpenCV documentation. OpenCV.org. URL: <https://docs.opencv.org/> (дата звернення: 21.09.2025).
28. Microsoft Docs. C# Guide. Microsoft, 2023. URL: <https://learn.microsoft.com/dotnet/csharp/> (дата звернення: 19.09.2025).
29. Visual Studio IDE documentation. Microsoft, 2023. URL: <https://learn.microsoft.com/visualstudio/> (дата звернення: 19.09.2025).
30. .NET Framework Guide. Microsoft, 2023. URL: <https://learn.microsoft.com/dotnet/framework/> (дата звернення: 19.09.2025).
31. Abdelnour N., Garcia C., Méric J. Improving OCR accuracy for identity documents using deep learning and image preprocessing techniques. IEEE Access. 2022. Vol. 10. P. 78543–78555. DOI: 10.1109/ACCESS.2022.3198765.
32. Nguyen H., Pham T., Le D. Machine learning–based MRZ recognition and error correction in e-passports. Journal of Imaging. 2023. Vol. 9, № 2. P. 44–57. URL: <https://www.mdpi.com/journal/jimaging> (дата звернення: 22.09.2025).