УДК 004.8:091
DOI https://doi.org/10.32782/2521-6643-2025-1-69.8

Івасечко А. В., викладач кафедри інформаційно-обчислювальних систем і управління
Західноукраїнського національного університету
ORCID: 0009 0002-8623-7828

Ліп’яніна-Гончаренко Х. В., доктор технічних наук, доцент, доцент кафедри інформаційно-обчислювальних систем і управління
Західноукраїнського національного університету
ORCID: 0000-0002-2441-6292

АРХІТЕКТУРА НАПІВАВТОМАТИЗОВАНОЇ СИСТЕМИ АНОТАЦІЇ БАГАТОМОВНИХ АРХІВНИХ РУКОПИСНИХ ТЕКСТІВ

У статті розглянуто актуальну науково-прикладну проблему, пов’язану з автоматизованим розпізнаванням історичних рукописних документів, що зберігаються в архівних установах України. Особливу складність становить те, що значна частина цих документів – зокрема періоду XIV–XIX століть – має гетерогенну мовну структуру, включаючи тексти українською, польською, латинською, російською та османсько-турецькою мовами. Окрім багатомовності, історичні тексти відзначаються використанням кількох алфавітів у межах одного документа, що значно ускладнює застосування класичних методів оптичного розпізнавання символів (OCR), які зазвичай орієнтовані на сучасні одномовні друковані документи.

Для подолання вказаних викликів у роботі запропоновано архітектуру інтелектуальної системи, здатної до напівавтоматизованого маркування символів у складних текстових фрагментах. Передбачено механізм колективної участі користувачів у процесі лейблінгу та подальшої валідації міток, що дозволяє зменшити ймовірність людських помилок і підвищити якість анотованих даних. Такий підхід сприяє формуванню високоякісних корпусів даних, необхідних для навчання моделей глибинного навчання, адаптованих до розпізнавання рукописних текстів зі змішаною мовною структурою.

Ключовими функціональними особливостями запропонованої системи є підтримка мультимовності, мультиалфавітності, а також можливість масштабування процесів розмітки даних відповідно до специфіки історичних джерел. Інструмент орієнтований на створення гнучкого середовища для інтерактивної взаємодії між дослідниками гуманітарної сфери та фахівцями з обробки даних, що, у свою чергу, дозволяє розширити можливості міждисциплінарних досліджень.

Запропоноване рішення має як теоретичне, так і практичне значення. З одного боку, воно сприяє розробці нових підходів до автоматизованої обробки історичних джерел у контексті цифрової гуманітаристики. З іншого – забезпечує інструментарій для збереження, вивчення та популяризації національної історико-культурної спадщини шляхом її якісної цифрової трансформації.

Ключові слова: *розпізнавання рукописного тексту, створення датасету, вебдодаток.*

Ivasechko A. V., Lipianina-Honcharenko Kh. V. Architecture of a semi-automated annotation system for multilingual archival handwritten texts

The developed system architecture enables the creation of datasets for further processing using machine learning and deep learning techniques. This approach plays a key role in addressing the challenges associated with the automated recognition of historical handwritten documents, particularly in complex multilingual and multi-script environments. A significant portion of Ukraine’s archival heritage, especially documents dating from the 14th to the 19th centuries, contains texts written in various languages–including Ukrainian, Polish, Russian, and Ottoman Turkish–using different scripts such as Cyrillic, Latin, and Arabic.

Traditional Optical Character Recognition (OCR) systems are typically designed for printed texts and are limited in their ability to handle the variability and noise present in historical manuscripts. Furthermore, they often lack support for mixed-language documents and rare historical scripts, making them unsuitable for large-scale archival digitization projects. In contrast, the proposed system architecture not only allows for the semi-automated labeling of individual characters but also incorporates user interaction, validation mechanisms, and multilingual capabilities. These features significantly improve the quality of the labeled data and ensure its suitability for downstream machine learning tasks.

The resulting datasets can be used to train modern recognition models, including convolutional neural networks (CNNs) and transformer-based architectures, which have demonstrated high effectiveness in visual and sequence processing tasks. By generating high-quality, annotated training samples, the system contributes to the development of robust handwriting recognition solutions that can adapt to historical variation in script style, ink degradation, and complex page layouts.

Moreover, the architecture supports iterative model refinement through human-in-the-loop strategies, where user feedback is incorporated to improve recognition accuracy over time. This is particularly important in the digital humanities domain, where expert validation and domain-specific knowledge play a critical role in ensuring the reliability of computational tools. Ultimately, the proposed system facilitates the preservation, accessibility, and computational analysis of cultural heritage documents, thereby supporting historians, linguists, and archivists in their research efforts.
Key words: *handwritten text recognition, dataset creation, web application.*

Постановка проблеми. Цифровізація та обробка рукописної інформації є однією з актуальних задач сучасної комп'ютерної лінгвістики, штучного інтелекту та інформаційних технологій загалом. Значна частина історичних документів, що зберігаються в архівах України, залишаються неіндексованими та малодоступними для пошуку й автоматичного аналізу через відсутність якісних засобів розпізнавання рукописного тексту, особливо в багатомовному та мультиалфавітному середовищі. Частина таких документів – це рукописні тексти XIV–XIX століть, створені польською, російською, українською та турецькою мовами, із використанням як кирилиці, так і латиниці. Окрім того, виявлено численні приклади українських слів, записаних латиною, що ускладнює задачу автоматичної класифікації та розпізнавання.

Відповідно, постає завдання якісного розпізнавання зображень текстів, їх символів та інтеграції в інтелектуальні системи. З огляду на складність рукописних шрифтів, варіативність мов та нестандартну орфографію, наявні датасети виявляються малоефективними. Це зумовлює потребу у створенні спеціалізованого датасету для задач комп'ютерного зору, навчання моделей глибокого навчання, семантичної обробки й класифікації тексту.

Аналіз останніх досліджень та публікацій. У контексті зростаючого інтересу до цифрової гуманітаристики, обробки рукописної спадщини та розвитку систем штучного інтелекту для аналізу історичних документів, дослідження методів багатомовної анотації й розпізнавання рукописного тексту набуває особливої актуальності. Багатомовні архівні матеріали, які містять тексти, написані різними мовами й алфавітами, становлять складне, але цінне джерело для побудови навчальних корпусів та розробки інтелектуальних інструментів розпізнавання (HTR/OCR). Сучасні підходи, що поєднують методи глибокого навчання (CNN, LSTM, GAN) з механізмами валідації, дозволяють досягти високої точності навіть у складних умовах, таких як варіативність орфографії, транслітерація або відсутність стандартизованої розмітки. Огляд актуальних досліджень демонструє як значний прогрес у точності розпізнавання, так і розвиток методологій, спрямованих на підвищення якості анотацій, зокрема через колективну валідацію та донавчання моделей на локальних даних.

Одне з ключових досліджень [1] описує підхід до постійного навчання моделей для історичних рукописів, що дозволяє оновлювати моделі з новими даними без повного перенавчання. В рамках обробки застосовується трифазова система: навчальний, валідаційний та тестовий набори. Це дозволяє досягати стабільності при поступовому розширенні корпусу.

У роботі [2] здійснено тестування платформи Transkribus на латинських текстах. Автори зазначають, що середній показник точності досяг 94,7% при використанні моделей, натренованих на понад 20 000 символів. Застосування Transkribus дозволяє автоматизувати транскрипцію, зменшуючи навантаження на бібліотеки та архіви.

У дослідженні [3] застосовано CNN до рукописів мовою Ґеез із середньою точністю 89% на тестовому наборі. Були оброблені 12 000 зображень, поділені на тренувальний, валідаційний та тестовий набори (80/10/10).

У огляді датасетів [4] підкреслюється, що лише 18% з 36 відомих історичних наборів містять розмітку на рівні символів. Дослідники пропонують стандартизацію валідації: по 10 прикладів на клас для тренування, тестування і виключення.

У дослідженні [5] проведено аналіз HTR на середньовічних латинських та французьких текстах. Модель демонструє точність розпізнавання 82,5% при використанні лише 200 вручну анотованих сторінок.

У роботі [6] розглядається мультиалфавітне розпізнавання із застосуванням hybrid CNN-LSTM моделей. Найкращий результат – 93,2% точності – досягнутий на корпусі з 5 мовами (арабська, хінді, бенгальська, англійська, урду).

Дослідження [7] показали, що середній CER (Character Error Rate) зменшується на 18% при донавчанні моделей Transkribus на локальних багатомовних архівах.

В рамках проєкту DigitalMaktaba[8] було оброблено 100 документів арабською, перською та урду, створено векторне представлення текстів, що дозволяє автоматичне групування за темами з точністю 91,3%.

В дослідженні[9] запропоновано новий підхід на основі Visual Language Models та GAN. Середній CER у тесті на корпусі середньовічних текстів становив 11,2% – покращення на 7,8% порівняно з базовою моделлю CNN.

Огляд[10] акцентує увагу на системах автоматичної анотації, що використовують CNN та ResNet. Найвищу точність (95,4%) було зафіксовано при розпізнаванні англійських рукописів, що оцифровані з архівів Університету Османа (Індія).

На основі аналізу десяти (табл. 1) досліджень можна зробити висновок, що успішне розпізнавання багатомовних рукописів залежить не лише від архітектури моделі, але й від методів збору, структурування та валідації даних. Найвищу точність було досягнуто при використанні великих анотованих корпусів, у поєднанні з CNN або їхніми гібридними формами. Ефективними виявилися і стратегії донавчання на локальних корпусах (Capurro et al.) та використання нових підходів як Visual Language Models (Aguilar, 2024). У таблиці 1 нижче зведено основні кількісні результати.

Таблиця 1

Порівняння досліджень в розпізнаванні багатомовних рукописів

№	Автори	Рік	Методологія	Мова / корпус	Точність / CER (%)	Кількість прикладів
1	Schomaker [1]	2020	Lifelong learning, HTR	Історичні документи	–	–
2	Milioni [2]	2020	Transkribus + ML	Латинська	94.7	>20,000 символів
3	Gurmu [3]	2021	CNN	Ґеез	89.0	12,000 зображень
4	Nikolaidou et al. [4]	2022	Dataset analysis	36 корпусів	–	18% мають розмітку символів
5	Aguilar & Jolivet [5]	2023	HTR	Латинська, французька	82.5	200 сторінок
6	Sinwar et al. [6]	2021	CNN-LSTM	5 мов	93.2	–
7	Capurro et al. [7]	2023	Transkribus + донавчання	Багатомовні архіви	CER ↓ на 18%	–
8	Bergamaschi et al. [8]	2022	Векторизація + тематична класиф.	Арабська, перська, урду	91.3	100 документів
9	Aguilar [9]	2024	Visual LMs + GAN	Середньовічні тексти	CER 11.2 (↓7.8%)	–
10	Deepa et al. [10]	2024	CNN + ResNet	Англійська (архіви Індії)	95.4	–

Таблиця 1 дозволяє легко оцінити вплив методологічних рішень на точність розпізнавання. Такі дослідження демонструють важливість використання адаптивних моделей та гнучких підходів до валідації в умовах багатомовного середовища. Найбільш ефективними виявляються системи з можливістю локального донавчання, підтримкою розмітки на рівні символів та гібридними архітектурами, що поєднують CNN з трансформерами або рекурентними мережами.

Мета статті – системно описати архітектуру напівавтоматизованої платформи для сегментації, анотації й класифікації символів із багатомовних архівних рукописних текстів, що поєднує попередню обробку зображень, колективний лейблінг та формування мультиалфавітного датасету.

Задачі дослідження:

1. Проектування модуля попередньої обробки – визначити послідовність конвертації PDF → растрове зображення (600 dpi), бінаризації за Отсу, виділення рядків, слів і символів із фіксацією позиційних метаданих.
2. Розробка клієнт-серверного інтерфейсу анотації – описати архітектуру Flask-додатку з базою SQLite, механізмом розподілу блоків символів між 3–5 анотаторами, консенсусною валідацією та логуванням дій.
3. Формування структури мультиалфавітного датасету – спроектувати схему таблиць для збереження лейблів із полем language, автоматизувати експорт розмітки в CSV із класифікацією за мовами й алфавітами.

Виклад основного матеріалу. Систему планується реалізувати у вигляді клієнт-серверного веб-додатку, що повинно забезпечити ефективну взаємодію користувачів з інтерфейсом анотації рукописних символів та зробити її доступною.

На рисунку 1 представлено розроблену архітектуру напівавтоматизованої системи анотації багатомовних архівних рукописних текстів.

Серверна логіка реалізована за допомогою мікрофреймворку Flask, що дозволяє обробляти HTTP-запити, управляти сесіями, обробляти форму реєстрації та авторизації користувачів, а також динамічно надавати дані з бази даних для відображення на сторінці анотації. Серверна частина взаємодіє з базою даних SQLite (блок 5), у якій зберігається інформація про:

– зареєстрованих користувачів (таблиця users (блок 6));
– доступні до анотації символи (таблиця symbols (блок 7));
– результати лейблінгу (таблиця labels (блок 8)).

Кожен запис у таблиці labels містить посилання на користувача (user_id), зображення символу (symbol_id), відповідну анотацію (label), слово(word) та мову(language). Для кожної групи користувачів формується індивідуальний набір ще неанотованих символів, що гарантує уникнення дублювання результатів.

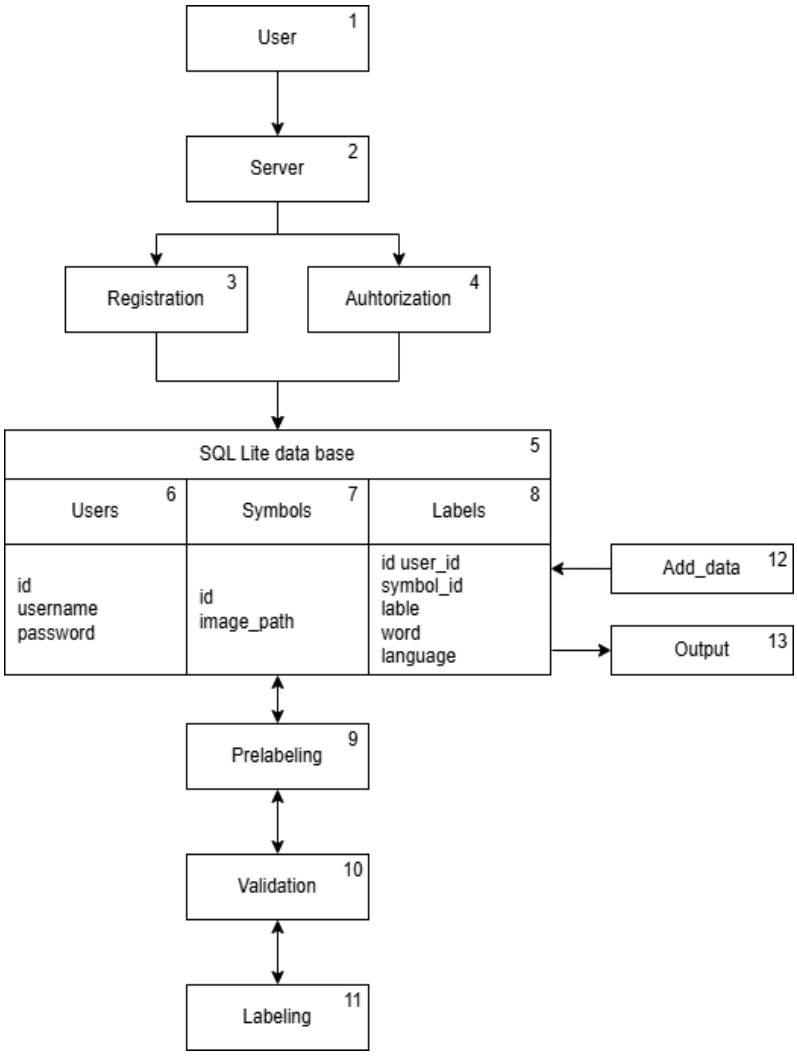

Рис. 1. Архітектура запропонованої системи

Інтерфейс користувача реалізований на основі HTML-шаблонів із використанням шаблонізатора Jinja2, що забезпечує генерацію динамічного веб-контенту та інтеграцію з серверною логікою.

В процесі роботи системи можна виділити такі ключові етапи:

1. Завантаження даних. Скрипт Add_data(блок 11) відповідає за імпорт в базу даних попередньо оброблених зображень символів у вигляді PNG-зображень розміром 64x64 пікселі сегментованих з архівних документів в форматі PDF взятих з центрального. Він здійснює ітерацію по файлах у директорії static/ symbols, перевіряє відсутність дублювання у базі даних, та додає нові записи до таблиці symbols. Кожен запис містить відносний шлях до зображення, що дозволяє коректно інтегрувати файл у систему відображення у браузері. Тепер система готова до роботи.

2. Авторизація. Після встановлення з'єднання з сервером (блок 2), користувач – User (блок1) автоматично перенаправляється на сторінку автентифікації, де має змогу або здійснити реєстрацію нового облікового запису – Registration (блок 3), або виконати вхід до системи за допомогою наявних облікових даних – Authorization (блок 4). В обох випадках відбувається звернення до таблиці users (блок 6) шляхом виконання відповідних SQL-запитів, що дозволяє здійснювати перевірку автентичності та обробку облікової інформації.

Після успішної авторизації користувача система перенаправляє його на сторінку інтерфейсу розмітки символів (блок 9), яка забезпечує інтерактивну взаємодію з даними. Зокрема, реалізовано можливість перегляду зображення символу, введення відповідної текстової мітки та надсилання результатів розмітки на сервер. Подані анотації фіксуються в таблиці labels (блок 8) за допомогою SQL-запитів, із прив'язкою до ідентифікатора користувача та символу.

Для візуалізації зображення символу використовується HTML-компонент <img>, шлях до якого формується динамічно через функцію Flask url_for('static', filename=…), що забезпечує сумісність із системою розподілу статичних ресурсів. Після кожного завершеного етапу розмітки сторінка автоматично

оновлюється та завантажує наступне зображення для анотації. У разі відсутності нерозмічених символів, інтерфейс інформує користувача про завершення доступного набору розмітки відповідним повідомленням.

3. Лейблінг. Процес лейблінгу організовано із урахуванням принципів колективного розмічування з багаторівневою перевіркою якості. Кожен документ, що завантажується до системи, може містити значну кількість символів – наприклад, до 10 000 одиниць. Для забезпечення ефективного розподілу навантаження та паралельної обробки, усі символи поділяються на дискретні блоки фіксованого розміру, наприклад, по 2000 символів у кожному.

Кожен блок призначається групі з трьох-п'яти користувачів, які незалежно здійснюють анотацію кожного символу в межах виділеного фрагмента. Такий підхід забезпечує можливість багатоверсійної оцінки кожного елементу та дозволяє реалізувати механізм консенсусної перевірки результатів.

З огляду на те, що джерельні документи, які підлягають обробці, можуть бути написані різними мовами – зокрема українською, польською, російською, а також містити фрагменти транслітерованого українського тексту латиницею – виникає потреба у фіксації мовної приналежності кожного зразка символу. Для вирішення цієї задачі передбачається додавання до інтерфейсу анотації відповідного елемента взаємодії з користувачем – зокрема поля вибору мови або відповідної позначки.

Після завершення первинної розмітки система формує агреговану таблицю анотацій, в якій для кожного символу здійснюється порівняння відповідей користувачів. У випадку, коли усі або більшість респондентів надали ідентичну мітку, результат автоматично вважається достовірним, цей процес називається валідацією (блок 10). У разі розбіжностей передбачається застосування механізму голосування, тобто присвоюється та мітка яку назвала більшість користувачів. Контекстна перевірка дозволяє зменшити кількість помилок, пов'язаних із хибним прочитанням окремих літер, зокрема у випадках, коли символи мають схожу графічну форму (наприклад, «і» та «ї», або «н» та «п»). Після валідації відбувається остаточне маркування, або ж Labeling (блок 11) правильний варіант записується в базу даних labels (блок 8).

Запропонована багатокористувацька стратегія розмітки дозволяє суттєво підвищити надійність результатів анотації, знизити ризики суб'єктивного сприйняття з боку окремих користувачів, та створити передумови для формування якісних навчальних вибірок для систем автоматичного розпізнавання рукописного тексту.

4. Експорт даних. Після завершення процесу розмітки даних, окремий скрипт Output (блок 13) формує CSV-файл із усіма анотаціями. Він здійснює об'єднання таблиць symbols (блок 7) і labels (блок 8) за допомогою операції JOIN та експортує дані у форматі: label,image_path,word,language

Цей файл може бути використаний для подальшого навчання моделей розпізнавання символів на основі глибокого навчання.

Висновки. У межах проведеного дослідження розроблено архітектуру інтелектуальної системи для напівавтоматизованої анотації рукописних символів історичних документів у багатомовному та мультал-фавітному середовищі. Створена архітектура повинна забезпечити функціонування сервісу, що включає реєстрацію користувачів, розподіл завдань розмітки, інтерактивне введення міток, фіксацію результатів у централізованій базі даних та динамічне відображення зображень символів, також наповнення бази даних символами для розмітки та експорт її результатів в форматі csv для подальшого використання в якості датасету для навчання нейронних мереж.

Список використаних джерел:

1. Schomaker, L. Lifelong learning for text retrieval and recognition in historical handwritten document collections. Handwritten historical document analysis, recognition and retrieval–state of the art and future trends. World Scientific, London, 2020. 221-248.

2. Milioni, N. Automatic transcription of historical documents: Transkribus as a tool for libraries, archives and scholars. DiVA Portal. 2020.

3. Gurmu, M. G. Offline handwritten text recognition of historical Ge'ez manuscripts using deep learning techniques. ResearchGate PDF. 2021.

4. Nikolaidou, K., Seuret, M., Mokayed, H., & Liwicki, M. (2022). A survey of historical document image datasets. International Journal on Document Analysis and Recognition (IJDAR), 25(4), 305-338.

5. Aguilar, S. T., & Jolivet, V. Handwritten text recognition for documentary medieval manuscripts. HAL. 2023.

6. Sinwar, D., Dhaka, V. S., Pradhan, N., & Pandey, S. Offline script recognition from handwritten and printed multilingual documents: a survey. International Journal on Document Analysis and Recognition (IJDAR). 2021. 24(1), 97-121.

7. Capurro, C., Provatorova, V., & Kanoulas, E. Experimenting with training a neural network in transkribus to recognise text in a multilingual and multi-authored manuscript collection. Heritage, 2023. 6(12), 7482-7494.

8. Bergamaschi, S., De Nardis, S., Martoglia, R., Ruozzi, F., Sala, L., Vanzini, M., & Vigliermo, R. A. Novel perspectives for the management of multilingual and multialphabetic heritages through automatic knowledge extraction: The digitalmaktaba approach. Sensors, 2022. 22(11), 3995.

9. Aguilar, S. T. Handwritten Text Recognition for Historical Documents using Visual Language Models and GANs. HAL. 2024

10. Deepa, A., Srija, B., Jabeen, M., Kankar, K. R., Lakshmi, B. J., & Negi, A. A Review on Automated Annotation System for Document Text Images. In 2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU) (pp. 1-6). IEEE. 2024.

References:

1. Schomaker, L. (2020). Lifelong learning for text retrieval and recognition in historical handwritten document collections. Handwritten historical document analysis, recognition and retrieval–state of the art and future trends. World Scientific, London, 221-248.

2. Milioni, N. (2020). Automatic transcription of historical documents: Transkribus as a tool for libraries, archives and scholars. DiVA Portal.

3. Gurmu, M. G. (2021). Offline handwritten text recognition of historical Ge'ez manuscripts using deep learning techniques. ResearchGate PDF.

4. Nikolaidou, K., Seuret, M., Mokayed, H., & Liwicki, M. (2022). A survey of historical document image datasets. International Journal on Document Analysis and Recognition (IJDAR), 25(4), 305-338.

5. Aguilar, S. T., & Jolivet, V. (2023). Handwritten text recognition for documentary medieval manuscripts. HAL.

6. Sinwar, D., Dhaka, V. S., Pradhan, N., & Pandey, S. (2021). Offline script recognition from handwritten and printed multilingual documents: a survey. International Journal on Document Analysis and Recognition (IJDAR), 24(1), 97-121.

7. Capurro, C., Provatorova, V., & Kanoulas, E. (2023). Experimenting with training a neural network in transkribus to recognise text in a multilingual and multi-authored manuscript collection. Heritage, 6(12), 7482-7494.

8. Bergamaschi, S., De Nardis, S., Martoglia, R., Ruozzi, F., Sala, L., Vanzini, M., & Vigliermo, R. A. (2022). Novel perspectives for the management of multilingual and multialphabetic heritages through automatic knowledge extraction: The digitalmaktaba approach. Sensors, 22(11), 3995.

9. Aguilar, S. T. (2024). Handwritten Text Recognition for Historical Documents using Visual Language Models and GANs. HAL.

10. Deepa, A., Srija, B., Jabeen, M., Kankar, K. R., Lakshmi, B. J., & Negi, A. (2024, March). A Review on Automated Annotation System for Document Text Images. In 2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU) (pp. 1-6). IEEE.