**Automated Classification of Mass User Document Uploads using Machine Learning**

**Final Year Project Proposal**

**Session FALL 2025**

**A 4th Year Student**

BSc. (Hons.)BS in CS



Department of CS

Fast School of Computing

Fast National University, Karachi Campus

20 September 2025

**Project Registration**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Project ID (for office use) | | |  | | | | |
| Type of project | | | [\*] Traditional [ ] Industrial [ ] Continuing | | | | |
| Nature of project | | | [\*] Development [ ] Research & Development [ ] Research | | | | |
| Sustainable Development Goals(SDGs) | | | [ ] Good Health and Well-Being [ ] Quality Education  [\*] Industry, Innovation, and Infrastructure [ ] Gender Equality  [ ] Decent Work and Economic Growth [ ] Climate Action | | | | |
| Area of specialization | | | [\*] Artificial Intelligence (AI) [ ] Data Science and Analytics  [ ] Internet of Things (IoT) [ ] Blockchain  [ ] Mobile App Development [\*] Web Development [ ] Cybersecurity  [ ] Game Development [ ] Natural Language Processing (NLP)  [ ] Other | | | | |
| **Project Group Members** | | | | | | | |
| Sr.# | Reg. # | Student Name | | CGPA | Email ID | Phone # | Signature |
| (i) | Group Leader | Abdul Ahad | | 2.4 | [K224353@nu.edu.pk](mailto:K224353@nu.edu.pk) | 03352662745 | Abdul Ahad |
| (ii) | Group Member | Musaddiq Kamal | | 3.15 | [K224432@nu.edu.pk](mailto:K224432@nu.edu.pk) | 03357248062 | Musaddiq |
| (iii) | Group Member | Ashar Usmani | | 3.6 | [K224581@nu.edu.pk](mailto:K224581@nu.edu.pk) | 03282158013 | Ashar Usmani |
| **Declaration:** FYP group members have cleared all prerequisite courses For FYP-I as per their degree requirements. | | | | | | | |

**Project Abstract**

National identity authorities and similar organizations manage millions of documents daily, ranging from CNICs and passports to birth and academic certificates. Manual handling of these documents is slow, error-prone, and difficult to scale when users submit large batches at once. This project proposes the design and implementation of an **automated document classification and field extraction system** capable of processing mass user uploads.

The system follows a hybrid pipeline: uploaded documents are first preprocessed and passed through an OCR engine to extract raw text. A lightweight LLM-based extractor identifies and labels relevant fields, which are then validated with rule-based checks for formats such as CNIC numbers and dates. To ensure stability, we propose an **order-invariant embedding strategy** that converts extracted fields into a consistent numerical representation, regardless of field order or OCR variation. These embeddings are fed into a classifier (logistic regression or small neural network) that predicts the document type, while a router module decides when to trust rules, when to rely on ML, and when to escalate to human review. The system also incorporates **bulk upload support**, a feedback interface for human corrections, and an out-of-distribution detector to flag unseen document types.

This work demonstrates a practical and extensible approach for organizations like NADRA to improve document management workflows, reduce manual verification load, and prepare for scaling toward millions of documents with higher accuracy and efficiency.

# **Introduction**

Organizations such as NADRA handle millions of official documents daily, including CNICs, passports, birth certificates, and academic records. These documents are critical for identity management, verification, and service delivery. At present, much of the document processing workflow depends on manual inspection, which is slow, labor-intensive, and prone to human error. The challenge becomes more severe when users submit large batches of documents, where traditional approaches struggle to maintain accuracy and efficiency.

Document classification and information extraction are essential steps in automating this workflow. Classification allows the system to identify the type of document (e.g., CNIC vs. passport), while extraction enables retrieval of important fields such as ID numbers, names, and dates. Conventional rule-based systems can achieve high precision on well-structured documents but fail when faced with layout variations, noisy scans, or unforeseen document formats. On the other hand, machine learning models generalize better but may misclassify documents or produce low-confidence predictions when trained on limited data.

This project aims to design and implement a **hybrid document classification system** that combines the strengths of rules, machine learning, and human verification. Our approach begins with mass user upload support, where multiple documents are ingested simultaneously and passed through preprocessing and OCR. A lightweight LLM-based extractor identifies candidate fields, which are validated using deterministic rules such as CNIC format checks and date parsers. Extracted fields are then converted into an **order-invariant embedding representation**, ensuring stable classification regardless of field order or OCR noise.

# **Success Criterion**

# The project will be considered successful if the system can reliably process mass user uploads of at least 500–1000 documents in a single batch without failure, while maintaining acceptable speed and resource usage. OCR should extract text with a character error rate below 10%, and the LLM-based extractor must identify key fields with at least 80% precision and recall. The proposed order-invariant embedding method should consistently represent documents regardless of field order, enabling the classifier to achieve at least 85% accuracy with clear separation between document types. The hybrid router must demonstrate improved reliability by combining rules, machine learning, and human-in-the-loop verification, while also rejecting at least 90% of unsupported document types instead of misclassifying them. Finally, the system should support a feedback mechanism for human corrections, preserve user privacy by anonymizing sensitive values, and demonstrate scalability and reproducibility through benchmarking and documented deployment.

# **Related work**

# Several studies have addressed document classification and field extraction using OCR, rule-based methods, and machine learning combinations. The work RDU: A Region-based Approach to Form-style Document Understanding [1] proposed a layout-aware transformer model that localizes target fields via region proposals, combining text coordinate features and visual layout for improved extraction accuracy across different document types. Similarly, DeepReader [2] extracted structured information from document images by detecting visual entities such as tables and boxes and mapping those into relational schemas to support rich querying. More recently, DocParser [3] presented an end-to-end OCR-free method for extracting information from visually rich documents, avoiding the two-step OCR + extraction pipeline and achieving state-of-the-art performance in certain benchmarks.

# **Project Rationale**

The motivation for this project stems from the growing need to automate large-scale document handling in organizations such as NADRA and similar authorities that process millions of identity-related records daily. Current manual and semi-automated processes are slow, prone to human error, and difficult to scale when handling bulk user uploads. These inefficiencies lead to delays in verification, increased operational costs, and potential risks of misclassification that directly affect service quality for citizens.

From a technological perspective, the limitations of existing rule-based systems and generic OCR solutions are clear: they often fail when documents vary in layout, structure, or quality. On the other hand, machine learning models trained only on text embeddings lack robustness in real-world conditions where scanned documents contain noise, blur, or non-standard formats. A modern approach must therefore combine both textual and visual features to ensure stable and accurate document classification.

Educationally, this project provides an opportunity to gain practical expertise in cutting-edge techniques such as optical character recognition, lightweight language models, embedding generation, and convolutional neural networks (CNNs) for document image feature extraction. The integration of these methods into a scalable pipeline for bulk document uploads allows the team to address a real-world challenge while building skills in machine learning, computer vision, and system integration that are highly valuable in today’s industry.

### **4.1 Aims and Objectives**

This project aims to develop an automated document classification and field extraction system that can process bulk user uploads efficiently and accurately. The system will combine OCR, LLM-based field extraction, text embeddings, and CNN-based image features to classify documents such as CNICs, passports, and certificates. The key objectives include:

1. Implementing an OCR pipeline to extract text from scanned documents, followed by lightweight LLM models to identify and label relevant fields.
2. Designing an embedding framework that integrates both **textual embeddings** (from extracted fields) and **visual embeddings** (from CNN-based document image analysis) to capture complementary features.
3. Training and evaluating classification models such as logistic regression, deep neural networks, and CNN-augmented hybrid models to predict document type with high accuracy.
4. Developing preprocessing steps to anonymize sensitive fields (e.g., names, CNIC numbers) while retaining structural signals useful for classification.
5. Supporting bulk user uploads, automated processing, and a feedback interface where human reviewers can correct low-confidence outputs for future retraining.

The overall objective is to create a scalable, privacy-aware, and accurate system for document classification and field extraction, demonstrating both applied impact and technical rigor.

### **4.2 Scope of the Project**

The project’s scope includes the development of a hybrid document processing pipeline capable of handling large batches of uploads. Core capabilities include OCR-based text extraction, LLM-based field identification, and CNN-based visual feature extraction from scanned images to ensure classification accuracy even when textual information is incomplete or noisy.

The system will be trained and evaluated on a dataset of labeled identity documents (e.g., CNICs, passports, birth certificates). Both textual and visual features will be integrated into embeddings, enabling the classifier to leverage structural layouts as well as semantic field information. Preprocessing will address challenges such as noisy scans, variable orientations, and inconsistent field ordering.

Machine learning development will cover multiple approaches, including simple classifiers (logistic regression, SVM), deep neural networks, and hybrid models that combine text and image embeddings. Evaluation will focus on classification accuracy, extraction precision/recall, and scalability when processing hundreds of documents simultaneously.

The user interface will provide functionality for mass uploads, visual feedback of extracted fields, confidence scores, and options for human corrections. Looking ahead, the system could be extended to support additional document types, integrate predictive analytics for fraud detection, and provide APIs for integration with existing organizational workflows.

Deployment considerations include ensuring that sensitive data is anonymized or securely stored, optimizing system performance for real-time or near-real-time document processing, and maintaining compatibility with widely used OCR engines and scanning devices.

# **Proposed Methodology and Architecture**

The proposed methodology follows a three-layer framework: **document acquisition and preprocessing, feature extraction,** and **classification with confidence evaluation.**

In the first layer, scanned documents uploaded by users are preprocessed through noise removal, orientation correction, and resizing. OCR is then applied to extract raw text.

The second layer performs feature extraction. A lightweight LLM identifies and labels key fields from the OCR output, which are transformed into order-invariant embeddings. In parallel, a CNN processes the document image to capture visual layout features such as structure, logos, and formatting. The textual and visual embeddings are then combined into a unified representation.

The final layer applies supervised models such as logistic regression, DNNs, or CNN-augmented classifiers to predict the document type. A router module uses confidence thresholds to decide whether to accept the prediction, escalate it for human review, or reject unsupported document types.

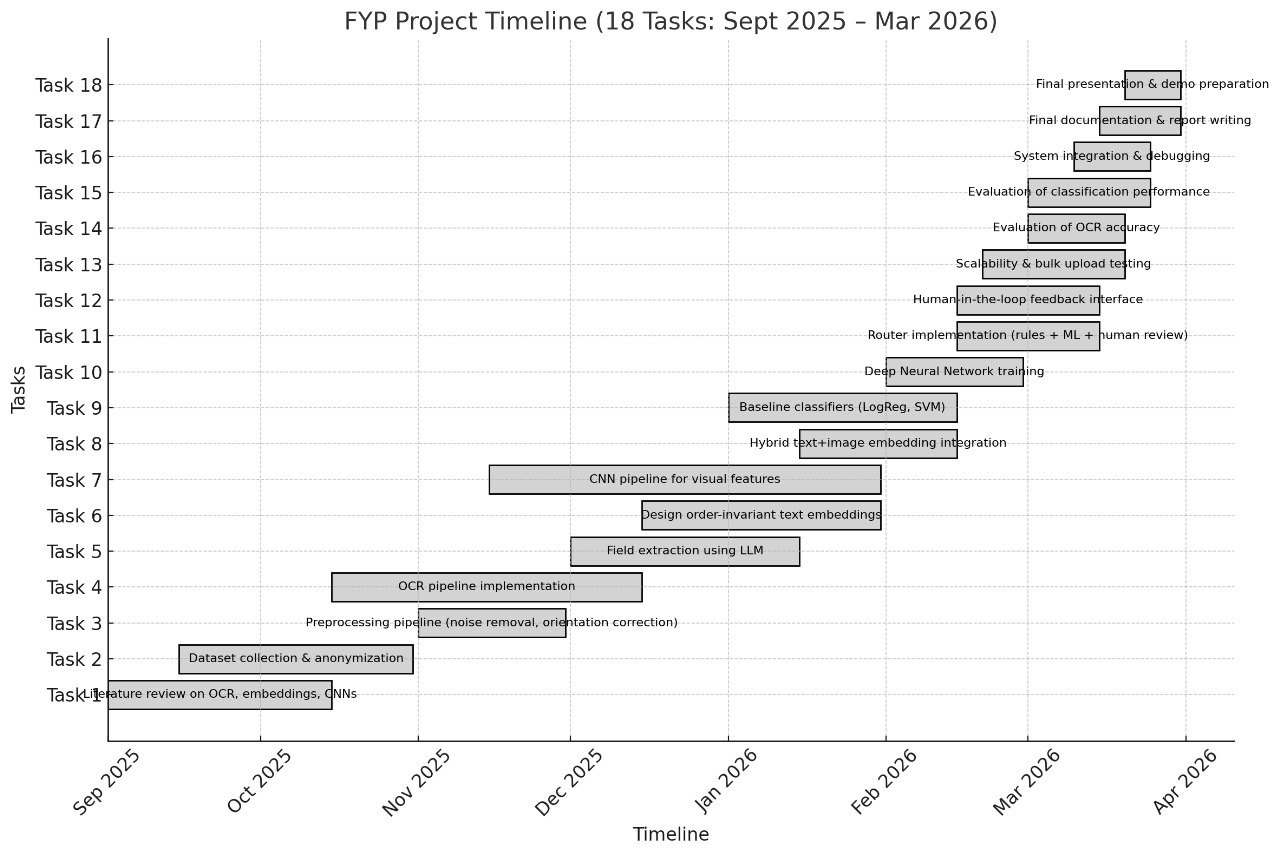
This structured methodology integrates OCR, LLM-based extraction, and CNN visual features to deliver scalable, accurate, and privacy-preserving document classification for bulk user uploads.

# **Individual Tasks**

|  |  |  |
| --- | --- | --- |
| **Team Member** | **Activity** | **Tentative Date** |
| Abdul Ahad (Group Leader) | Survey research papers on OCR, embeddings,and document classification; design preprocessing pipeline (noise removal, orientation correction); assist in evaluation of text embeddings | September 2025 – December 2025 |
| Abdul Ahad (Group Leader) | Integrate router module (rules + ML + human review); prepare documentation and final report | January 2026 - March 2026 |
| Musaddiq kamal | Collect and anonymize sample documents; prepare dataset structure; implement and test OCR pipeline | September 2025- December 2025 |
| Musaddiq Kamal | Develop LLM-based field extractor; evaluate extraction accuracy; support final presentation prep | January 2026 - March 2026 |
| Muhamamd Ashar | Implement CNN pipeline for extracting visual features; design hybrid text + image embedding model | September 2025 – December 2025 |
| Muhammad Ashar | Develop classification model combining text and image embeddings | January 2026 - March 2026 |

1. Individual Task

# **Gantt Chart**



# **Tools and Technologies**

The project will be developed using a combination of machine learning, computer vision, and natural language processing tools. Python will serve as the primary programming language due to its rich ecosystem of AI and data processing libraries. OCR will be implemented using open-source engines such as **Tesseract** or alternatives like EasyOCR to extract raw text from scanned documents. For field extraction, lightweight **LLM models** (e.g., DistilBERT or instruction-tuned small models) will be employed to identify and label relevant fields. Text embeddings will be generated using frameworks such as **SentenceTransformers**, while **Convolutional Neural Networks (CNNs)** built with **TensorFlow** or **PyTorch** will be used to capture visual layout and image features. The classification component will utilize scikit-learn and deep learning libraries to implement models ranging from logistic regression to hybrid DNN architectures. A simple **Flask or Django** backend will manage mass user uploads, while a web-based interface will be built using **HTML, CSS, and JavaScript** to support human-in-the-loop feedback. Data storage and logging will be managed through **SQLite or PostgreSQL**, with options for secure handling of anonymized records. To ensure reproducibility and scalability, the system will be containerized with **Docker** and version-controlled using **GitHub**. Together, these technologies provide a robust stack for building a scalable, accurate, and privacy-aware document classification system.

# **References**

[1] J. Yu, J. Li, and Z. Tang, “RDU: A Region-based Approach to Form-style Document Understanding,” arXiv preprint arXiv:2206.06890, Jun. 2022. [Online]. Available: [https://arxiv.org/abs/2206.06890](https://arxiv.org/abs/2206.06890?utm_source=chatgpt.com)

[2] S. Katti, C. Reisswig, S. Guder, and S. Brarda, “DeepReader: Information extraction from document images via visual entity detection and relational reasoning,” arXiv preprint arXiv:1812.04377, Dec. 2018. [Online]. Available: [https://arxiv.org/abs/1812.04377](https://arxiv.org/abs/1812.04377?utm_source=chatgpt.com)

[3] X. H. Li, J. Tang, and X. Zhang, “DocParser: An End-to-End OCR-Free Document Information Extraction Model,” arXiv preprint arXiv:2304.12484, Apr. 2023. [Online]. Available: [https://arxiv.org/abs/2304.12484](https://arxiv.org/abs/2304.12484?utm_source=chatgpt.com)