**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. The classifier predicts the document type, while a router decides whether to accept the result, rely on a rule-based decision, or escalate to human review.

The system is designed with scalability and extensibility in mind. Bulk upload handling allows organizations to process thousands of documents efficiently. A feedback loop enables human operators to correct extraction errors, and these corrections can later be used to improve the model through retraining. Furthermore, the system supports privacy-preserving features by removing or hashing sensitive values while retaining structural signals useful for classification.

# **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 need to modernize transformer health monitoring practices within the electrical utility industry. Current manual assessment methods create significant risks due to their inherent limitations in detecting subtle degradation patterns and predicting equipment failures with sufficient advance warning. These limitations translate directly into increased maintenance costs, reduced system reliability, and potential safety hazards for both utility personnel and consumers.

The financial implications of transformer failures extend far beyond equipment replacement costs, it also includes lost revenue from service interruptions, regulatory penalties, customer compensation claims, and emergency response expenses.

Educational objectives include gaining practical experience with cutting-edge artificial intelligence technologies while addressing real-world engineering challenges that have direct societal impact. This project provides opportunities to develop expertise in computer vision, deep learning, and industrial system integration that are highly valued in present engineering practice.

## **Aims and Objectives**

This project aims to develop an automated system for computing transformer health indexes using deep learning technology to analyze thermal images. The system will provide accurate and timely health assessments to inform maintenance decisions. Key objectives include:

1. Implementing CNN architectures for thermal image analysis to detect temperature anomalies, hotspots, and thermal patterns indicating transformer degradation.
2. Recognizing thermal signatures of issues like overheating, insulation breakdown, cooling system failures, and load-related temperature distributions.
3. Developing image processing pipelines to enhance thermal image quality, normalize temperature scales, and extract meaningful features despite varying environmental conditions.

The system will enable advanced thermal-based condition monitoring for power system equipment, improving maintenance decision-making. System deployment objectives encompass creating user-friendly interfaces that present thermal analysis results in intuitive formats accessible to maintenance personnel with varying technical backgrounds. The implementation must include real-time image processing capabilities, automated alert generation for critical thermal conditions, and reporting functionalities that support decision-making processes effectively.

## **Scope of the Project**

The project's scope includes the creation of a comprehensive health index automation system that uses thermal imaging technology and is especially made for distribution transformer applications. Thermal image processing capabilities using computer vision techniques to detect component temperature anomalies, overheating patterns, and thermal signatures that indicate different fault states impacting transformer performance are examples of core capability.

The scope of data processing includes the collection and examination of thermal images taken from distribution transformers in a range of operating scenarios. In order to manage fluctuating ambient temperatures, seasonal effects, lighting conditions, and camera positioning irregularities frequently seen in field deployment settings, the system will incorporate sophisticated image preprocessing techniques.

Machine learning model development includes implementing and comparing multiple convolutional neural network architectures. This encompasses specialized CNN designs for temperature distribution analysis, thermal anomaly detection algorithms, and classification systems for different failure modes visible through thermal imaging.

User interface development scope includes creating web-based dashboards for thermal image analysis results, mobile applications for field personnel thermal data capture, and reporting tools that translate thermal patterns into maintenance recommendations. The interface design will prioritize usability while providing comprehensive thermal analysis functionality for different user roles.

Looking ahead, opportunities exist to integrate additional sensor technologies, broaden the application to various types of electrical equipment, and incorporate predictive analytics derived from thermal trend monitoring. Nevertheless, the present project remains intentionally focused on thermal imaging-based condition assessment, ensuring the thorough development and refinement of essential computer vision functionalities.

Key factors for deployment include ensuring compatibility with existing thermal imaging hardware, implementing robust data security measures suitable for critical infrastructure, and optimizing system performance to handle real-time processing of high-resolution thermal data in demanding industrial settings.

# **Proposed Methodology and Architecture**

The methodology employs a structured framework for evaluating transformer health through computer vision analysis of thermal images. The system's architecture comprises three core layers: thermal image acquisition and preprocessing, feature extraction via deep learning models, and health index calculation with confidence evaluation.

The thermal image acquisition layer integrates with thermal imaging systems to capture detailed temperature profiles across transformer surfaces. The preprocessing stage applies image refinement algorithms, and region-of-interest detection methods tailored to the transformer's geometry and thermal properties.

The feature extraction layer leverages a customized convolutional neural network (CNN) designed to recognize thermal patterns, identifying anomalies such as hotspots, temperature gradients, and unusual thermal signatures associated with various failure modes.

The health index computation component takes the output from the thermal analysis model to generate a comprehensive health score that reflects the transformer's overall condition.

# **Individual Tasks**

|  |  |  |
| --- | --- | --- |
| **Team Member** | **Activity** | **Tentative Date** |
| Muhammad Umar (Group Leader) | Project coordination and planning, thermal image dataset collection and preprocessing, CNN architecture design and implementation | September 2025 - February 2026 |
| Muhammad Umar (Group Leader) | Literature review on thermal imaging for transformer monitoring, deep learning model training and optimization, system testing and validation | October 2025 - April 2026 |
| Umar Orakzai | Thermal image preprocessing pipeline development, computer vision algorithm implementation, feature extraction optimization for thermal patterns | October 2025 - February 2026 |
| Umar Orakzai | Model performance evaluation and validation, thermal anomaly detection algorithm development, user interface backend integration | December 2025 - March 2026 |
| Shaheer Uddin | Web application development for thermal image analysis interface, database design for thermal image storage, mobile application development | November 2025 - April 2026 |
| Shaheer Uddin | User interface design for thermal pattern visualization, reporting module implementation, system deployment and testing | December 2025 - April 2026 |

1. Individual Task

# **Gantt Chart**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | August | September | October | November | December | January | February | March | April | May | June | July |
| Task 1 |  | Project coordination and planning | | | | | |  |  |  |  |  |
| Task 2 |  | Thermal image dataset collection & preprocessing | | | | | |  |  |  |  |  |
| Task 3 |  | CNN architecture design & implementation | | | | | |  |  |  |  |  |
| Task 4 |  |  | Literature review on thermal imaging | | | | | | |  |  |  |
| Task 5 |  |  | Deep learning model training & optimization | | | | | | |  |  |  |
| Task 6 |  |  | System testing and validation | | | | | | |  |  |  |
| Task 7 |  |  | Thermal image preprocessing pipeline | | | | |  |  |  |  |  |
| Task 8 |  |  | Computer vision algorithm implementation | | | | |  |  |  |  |  |
| Task 9 |  |  | Feature extraction optimization | | | | |  |  |  |  |  |
| Task 10 |  |  |  |  | Model performance evaluation & validation | | | |  |  |  |  |
| Task 11 |  |  |  |  | Thermal anomaly detection algorithm development | | | |  |  |  |  |
| Task 12 |  |  |  |  | UI backend integration | | | |  |  |  |  |
| Task 13 |  |  |  | Web application development | | | | | |  |  |  |
| Task 14 |  |  |  | Database design for thermal image storage | | | | | |  |  |  |
| Task 15 |  |  |  | Mobile application development | | | | | |  |  |  |
| Task 16 |  |  |  |  | UI design for thermal pattern visualization | | | | |  |  |  |
| Task 17 |  |  |  |  | Reporting module implementation | | | | |  |  |  |
| Task 18 |  |  |  |  | System deployment and testing | | | | |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |

# **Tools and Technologies**

Python serves as the core programming language due to its rich ecosystem for machine learning and the availability of powerful libraries. Deep learning models are developed using TensorFlow and PyTorch, both of which offer GPU acceleration to support the high computational demands of training.

For data handling and processing, Pandas is used for manipulating structured datasets, while NumPy facilitates numerical operations. OpenCV is employed for computer vision tasks, particularly in the processing of thermal images. Machine learning workflows are further supported by Scikit-learn, which offers essential utilities for model evaluation and performance metrics. Visualization and pattern interpretation of thermal data are achieved using libraries such as Matplotlib and Seaborn.

The web application architecture combines Django for backend development and React.js for building dynamic, user-friendly frontend interfaces. This combination supports the creation of scalable and maintainable dashboards for thermal image analysis.

To manage machine learning lifecycle operations, MLflow is used for tracking experiments and managing different model versions. For cloud-based deployment, the infrastructure is hosted on scalable platforms like Amazon Web Services (AWS) or Microsoft Azure, ensuring reliable and efficient computational support.

# **References**

[1] A. S. Mogos, X. Liang, and C. Y. Chung, "Enhancing Transformer Health Index Prediction Using Dissolved Gas Analysis Data Through Integration of LightGBM and Robust-EM Algorithms," *IEEE Access*, early access, 2024.

[2] Y. Li, "The State of the Art in transformer fault diagnosis with artificial intelligence and Dissolved Gas Analysis: A Review of the Literature," *arXiv*, Apr. 2023.