TRACEFINDER – FORENSIC SCANNER IDENTIFICATION SYSTEM

Abstract

Digital document forensics plays a vital role in identifying the authenticity and origin of scanned documents. This project, TraceFinder, focuses on identifying the source scanner of a scanned document using machine learning and deep learning techniques. The system analyzes scanner-specific artifacts through handcrafted features and convolutional neural networks. Explainability techniques and a user-friendly deployment interface are incorporated to enhance transparency and usability.

---

1. Introduction

With the increasing use of digital documents, verifying their authenticity has become a significant challenge. Scanners introduce unique noise patterns and texture artifacts during document digitization. Identifying these patterns helps determine the scanner source. TraceFinder aims to automate scanner identification using image processing, machine learning, deep learning, and explainability techniques.

---

2. Problem Statement

To design and implement an automated system capable of identifying the scanner brand or model used to digitize a document by analyzing scanner-specific artifacts in scanned images.

---

3. Objectives

- Collect and organize scanned document datasets
- Extract scanner-specific features
- Train classical ML and CNN models
- Apply explainability techniques
- Deploy the system with a simple user interface

---

4. Dataset Description

The dataset consists of scanned document images collected from multiple scanner sources and categories:

- Official document scans

- Wikipedia document scans
- Tampered images
- Flatfield images
- Original documents

Each image is labeled with scanner information and metadata.

---

5. System Architecture

Image Upload

↓

Preprocessing

↓

Feature Extraction / CNN

↓

Prediction

↓

Confidence Score

↓

Explainability

↓

Result Logging & Download

---

6. Methodology

Milestone 1 – Dataset Collection & Labeling

- Image collection from multiple scanners
- Automatic labeling using Python scripts
- Metadata extraction and analysis

Milestone 2 – Preprocessing

- Image resizing and normalization

- Noise and quality analysis

Milestone 3 – Feature Engineering & Classical ML

- LBP, FFT, and Noise Residual feature extraction
- Classification using SVM, Random Forest, and Logistic Regression

Milestone 4 – Deep Learning & Explainability

- CNN model trained on raw images
- Image augmentation applied
- Grad-CAM based explainability implemented

Milestone 5 – Deployment

- Streamlit-based web interface
- Image upload and prediction
- Confidence score display
- Prediction logging and download

---

7. Results & Model Comparison

Model	Description	Performance
Classical ML	Handcrafted features	Moderate
CNN	Automatic feature learning	Higher
Explainability	Grad-CAM	Visual insights

---

8. Explainability Analysis

Grad-CAM visualizations highlight regions contributing to scanner identification. Due to limited training data, placeholder visualizations were generated to validate the explainability pipeline.

---

9. Deployment & User Interface

A Streamlit-based interface allows users to upload scanned images and receive scanner predictions with confidence scores. Prediction logs can be downloaded for forensic analysis.

---

10. Conclusion

TraceFinder successfully demonstrates an end-to-end forensic scanner identification system. The integration of machine learning, deep learning, explainability, and deployment ensures scalability and practical applicability.

---

11. Future Work

- Expand dataset with more scanners
- Improve CNN performance
- Apply advanced explainability methods
- Deploy as a cloud-based service

---

12. References

- Digital Image Forensics literature
- TensorFlow & Scikit-learn documentation
- Grad-CAM research papers