Milestone 2 Report: Forensic Feature Extraction & Identification

Project: TraceFinder – Scanner Identification System

1. Objective

The goal of Milestone 2 was to move beyond simple image storage and develop a mathematical representation of scanner "fingerprints." We aimed to extract texture and frequency-based features and train a machine learning model to distinguish between different scanner hardware sources.

2. Data Characterization

We utilized a balanced dataset of patches representing 5 distinct scanner models.

- **Target Classes:** Canon120-1, Canon200, EpsonV39-1, EpsonV500, and HP.
- **Input Format:** 512×512 Grayscale patches (standardized from original inputs).
- **Total Sample Size:** 1,348 forensic patches.

3. Methodology

3.1 Forensic Feature Engineering

Unlike standard image recognition, forensic identification relies on "micro-textures." We extracted three types of features:

1. **Local Binary Patterns (LBP):** We used a 10-bin histogram of uniform LBP patterns to capture surface texture variations and sensor grain.
2. **Frequency Analysis (FFT):** By applying a Fast Fourier Transform, we calculated the Mean and Standard Deviation of the frequency spectrum to detect periodic noise patterns specific to scanner motors and sensors.
3. **Noise Variance:** Calculated using a Laplacian filter to measure the statistical high-frequency "salt-and-pepper" noise.

3.2 Model Architecture

- **Preprocessing:** Data was normalized using StandardScaler to ensure features with large ranges (like FFT) didn't overshadow smaller features (like LBP).
- **Algorithm:** Multinomial **Logistic Regression** using the lbfgs solver.
- **Training Strategy:** 80/20 Train-Test split.

4. Results & Performance

The model achieved an **Overall Accuracy of 82.22%**.

4.1 Classification Metrics

Scanner Model	Precision	Recall	F1-Score
Canon120-1	0.68	0.71	0.69
Canon200	0.75	0.96	0.84
EpsonV39-1	0.85	0.75	0.80
EpsonV500	0.86	0.82	0.84
HP	0.96	0.84	0.90

4.2 Analysis

- **Strongest Performer:** The **HP** scanner was the most distinct, with a 96% Precision.
- **Observation:** The **Canon200** had the highest Recall (96%), meaning the model rarely missed a Canon200 sample, though it occasionally confused it with the Canon120-1.
- **Conclusion:** The high F1-scores across all classes (averaging ~0.82) indicate that the engineered features (LBP and FFT) are highly effective at capturing the unique digital "fingerprint" of the scanners.

5. Summary

Milestone 2 successfully demonstrates that forensic features can identify scanner hardware with over 80% accuracy without needing deep learning. The current pipeline is lightweight and robust. For Milestone 3, we will explore Deep Learning (CNNs) to attempt to push the accuracy beyond 90%.