Machine Learning for Cryptography

Report 1: Distinguishing Encrypted from Raw Data

Name: Mateusz Tabaszewski

October 24, 2025

1 Introduction

The experiments described in this report investigates if it is possible to use machine learning models (Logistic Regression and Random Forest Classifier) to differentiate between unencrypted and encrypted/random data by analyzing chosen statistical features. To conduct this experiment, 20 of the biggest Python files from the official PyTorch GitHub repository were downloaded and chopped into 32 KB chunks of raw data. Then, three AES algorithm variations were applied (ECB, CBC, CTR) along with random compressed data (gzip) to create a training dataset. Ratios of examples from different datasets were adjusted in such a way as to create a balanced dataset with equal number of positive (encrypted/random) and negative (unencrypted) examples. Additionally, an experiment was conducted comparing the training results of all types of data against one another (recognizing random data vs. AES in ECB mode; AES in ECB mode vs. AES in CBC mode, etc.). Both classification tasks for both models were implemented based on the following features computed for the binary data:

- Mean Entropy
- Chi-Squared P-value
- Compressibility
- Serial Correlation
- FFT Flatness
- ECB Indicator

It should be added that for the first experiments, the trained models were also evaluated on a new, unseen dataset comprised of Tensorflow's Python files and random data to evaluate the validity of the findings.

2 Results and Analysis

The calculated features on all data types being fed into the models as input can be seen in Table 1. It should be added that the results are the mean values of the features calculated for each training example of the given data type, for complete analysis for concatenated data, please see the corresponding Jupyter notebook and its outputs.

Table 1: Statistical Features: Raw vs. Encrypted Data

Metric	**Raw Data**	**AES ECB**	**AES CBC**	**AES CTR**	**Random**
Mean Entropy (bits)	0.962	1.000	1.000	1.000	1.000
Mean Entropy (bits/byte)	4.449	7.983	7.994	7.994	7.994
Chi-Squared P-value	0.000	0.008	0.503	0.511	0.511
Compressibility	0.172	0.842	1.000	1.000	1.000
Serial Correlation	0.628	−0.003	0.000	0.000	0.000
FFT Flatness	0.306	0.560	0.560	0.560	0.560
ECB Indicator	0.182	0.182	0.000	0.000	0.000

From the presented results, the following observations can be drawn:

- **Entropy:** Raw data has a much higher value for mean entropy, showing the redundancy and structure in the data as opposed to encrypted/random data. However, this pattern is harder to notice in entropy calculated on the level of bits.
- **Chi-Squared Test:** Both Raw Data and ECB have much lower P-value for Chi-Squared Test while the remaining methods are much closer to uniform distribution.
- **Compressibility:** Raw data is much more compressible, showing (yet again) redundancy and structure of the data, as opposed to encrypted/random data. However, it should be noted that ECB mode of AES also displays a noticeably higher level of compressibility than other tested algorithms.
- **Serial Correlation:** Raw data shows strong correlation between neighboring bytes, while encrypted data displays near-zero correlation, indicating randomness.
- **FFT Flatness:** Raw data exhibits a slightly lower value of the FFT Flatness when compared to the random and encrypted datasets.
- **ECB Indicator:** The ECB Indicator value is higher for both raw and AES ECB data when compared to the remaining methods.

The distribution of the entropy values for all data can be seen in Figure 1. The distribution is somewhat bimodal, most likely related to the raw vs. encrypted/random data.

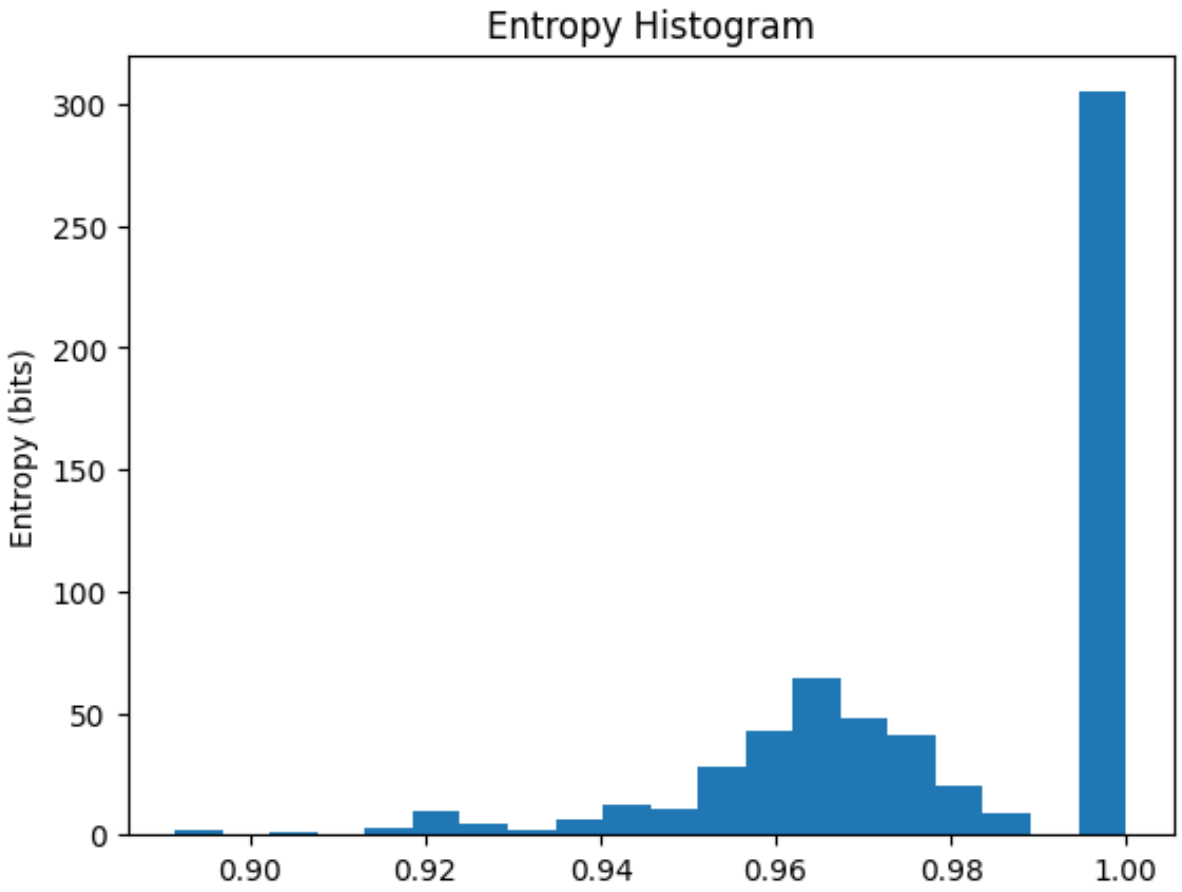

Figure 1: Distribution of entropy values for all considered data.

The performance of the models was summarized in Table 2. All tested models achieve perfect classifications for both train and test data. This suggests that indeed, the raw data can be easily distinguished from encrypted and random data, and that the provided statistical features provide a very clear separation between these two classes.

Table 2: Binary Classification: Encrypted vs. Raw Data

Model	**Accuracy**	**ROC AUC**	**TPR @ 1% FPR**
Logistic Regression (Train)	1.000	1.000	1.000
Logistic Regression (Test)	1.000	1.000	1.000
Random Forest (Train)	1.000	1.000	1.000
Random Forest (Test)	1.000	1.000	1.000

The pairwise comparison heatmap can be seen in Figure 2. The plots show which datasets can be differentiated between one another by utilizing the same features as in the previous experiment.

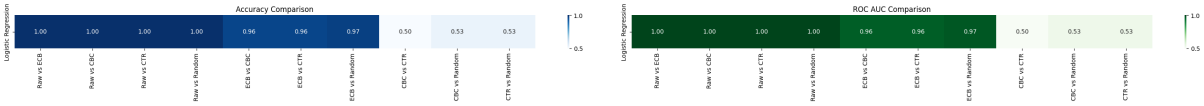

Figure 2: Comparison of pairwise classification of dataset types.

The comparison showcases that raw data can be easily differentiated from any other dataset by a Logistic Regression algorithm, and so can the AES-ECB algorithm, which is likely due to it being the "least random", i.e. due to the fact that each block is independent, some structure may still be found. However, differentiating other types of datasets by Logistic Regression results in performance that can be considered almost random.

3 Conclusions

The experiment proved that encrypted and raw data can be easily distinguished by machine learning methods when combined with static feature calculation. The tested models produce reliable results that can be tested on a dataset from a completely different repository for comparable results.