Research Project Report on the Topic:
Development of a Method for Assessing and Identifying Artificially Generated Images
Разработка метода оценки и выявления искусственно сгенерированных изображений

#БПАД231, 2nd year of study

Dubenskiy Konstantin Mikhailovich

Project Supervisor:

Lukyanchenko Peter Pavlovich

Senior Lecturer Faculty of
Computer Science, HSE University

Moscow 2025

Subject area

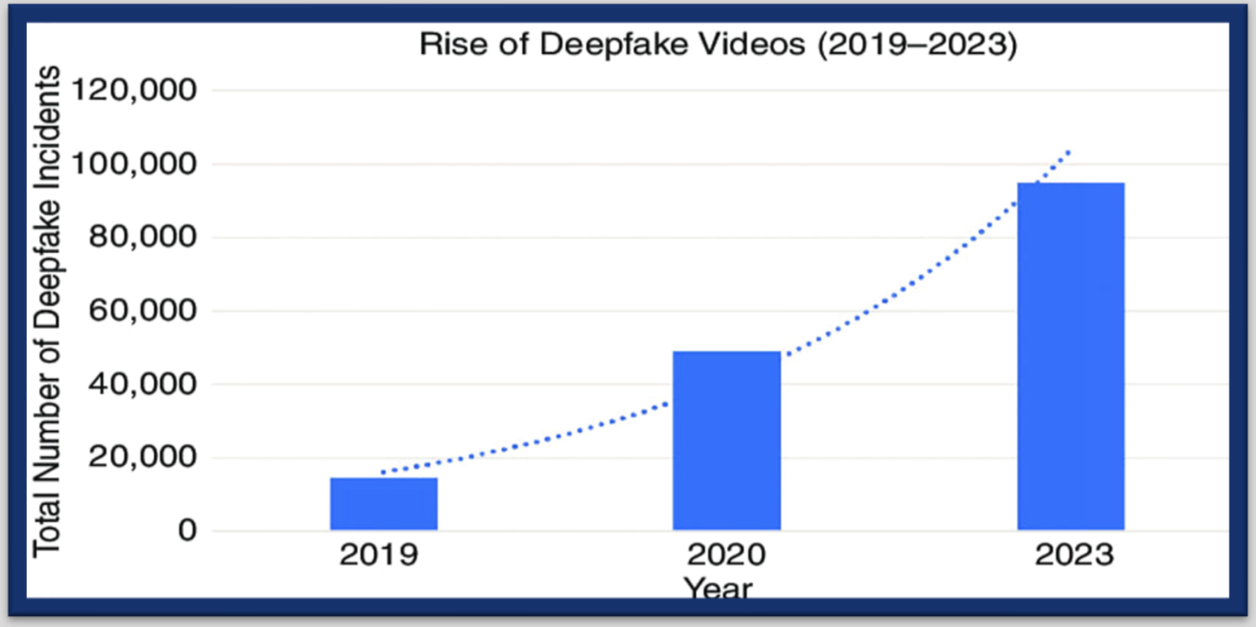
Modern generative models (GANS) allow you to create photorealistic images. This creates a need for automated methods of recognizing fake images for cybersecurity.

Real image by phone

Fake image

The relevance of the work

The amount of media content created is growing exponentially. Since 2019, the number of deepfakes has increased by about 5 times. Recognizing such images manually is very difficult, and requires an automated and accurate way to assess the likelihood of artificial origin of the image.



Goal

To investigate and evaluate existing methods for detecting artificially generated images based on interpreted features.

Tasks

- Implement 5 feature-based image analysis methods.
- Automate the selection of parameters for each method.
- Combine the methods into a single model.
- Conduct computational experiments and evaluate accuracy.

Functional requirements

- Dataset of real and fake images
- Processing without meta tags
- Output of probability from 0 to 1
- The ability to evaluate each method separately and in combination

The entropy method

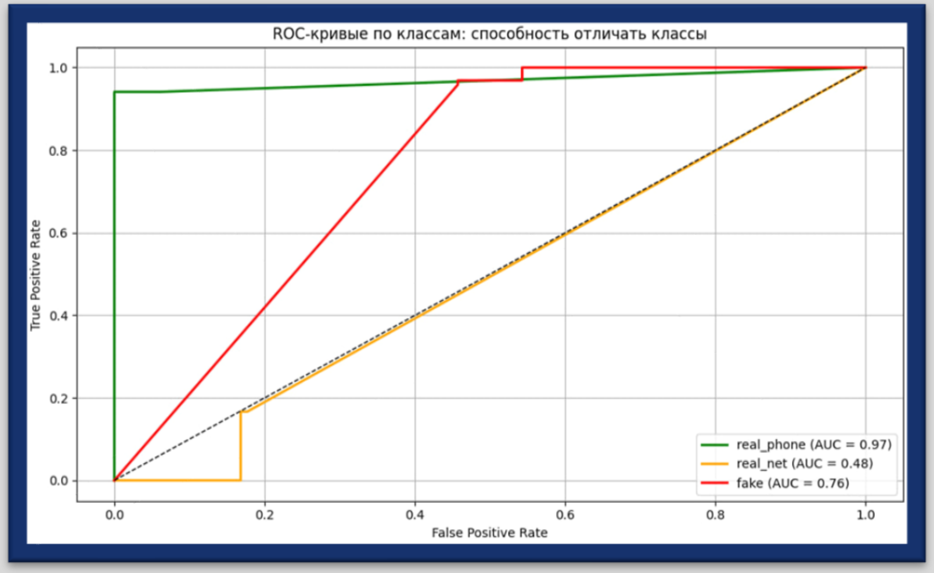
The method analyzes local entropy — how much information (diversity) differs in channels R, G, and B. In real images (for example, the sky, walls), channels have similar entropy, so large "consistent" zones are formed. The generated images have more randomness and small areas of overlap.

Signs used

- The entropy in the channel R, G, B
- he difference between them is in the local window
- Average size of clusters of matches

Parameters

- radius is the radius of the window for local entropy (more = smoother result)
- tolerance — how strongly the channels should match (RGB) (less = stricter)
- min_size, max_size — which clusters are considered the norm.



Result

AUC ≈ 0.76 (After setting up the parameters)
The method is quite accurate

The Fourier transform method

This method analyzes the image in the frequency domain. Generators often leave regular textures and artificial noises, especially in GAN.
The Fourier transform helps to distinguish them: fake images have a higher proportion of high-frequency components.

Signs used

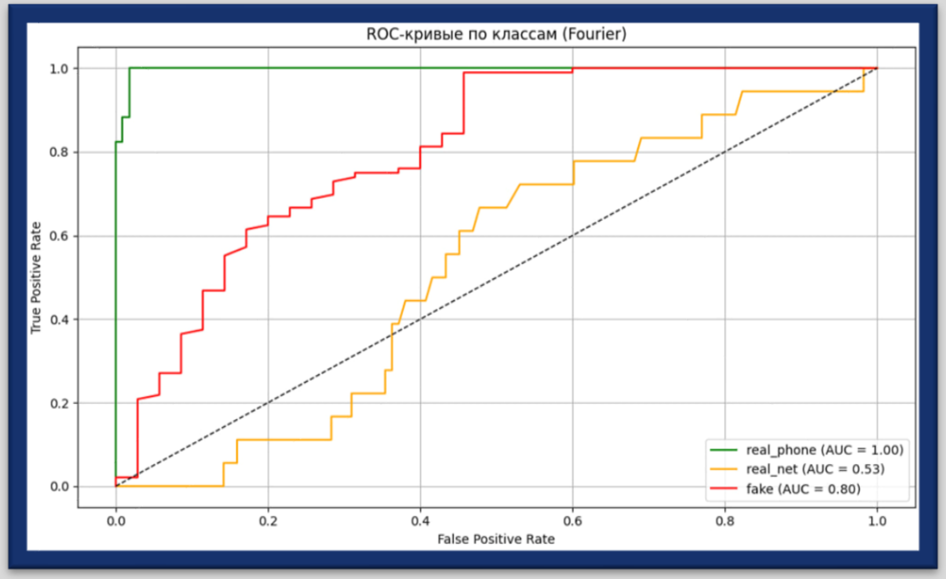
- The spectrum module (log scale)
- The center of mass of the spectrum
- The share of energy in the central zone vs in the periphery

Parameters

- The radius of the central area (which is considered "low" frequencies)
- The scale of spectrum normalization
- Threshold for the "center / total energy" ratio

Result

AUC ≈ 0.80
The method proved to be the most accurate among the single ones.
Confidently distinguishes between regular noises and sharp frequency patterns.



The method of color distributions

Converting the image to the YCrCb color space. In real images, the color channels (especially Cr, Cb) are smoothly distributed.
Fake has abnormal peaks, offsets, and unnatural color ratios.

Signs used

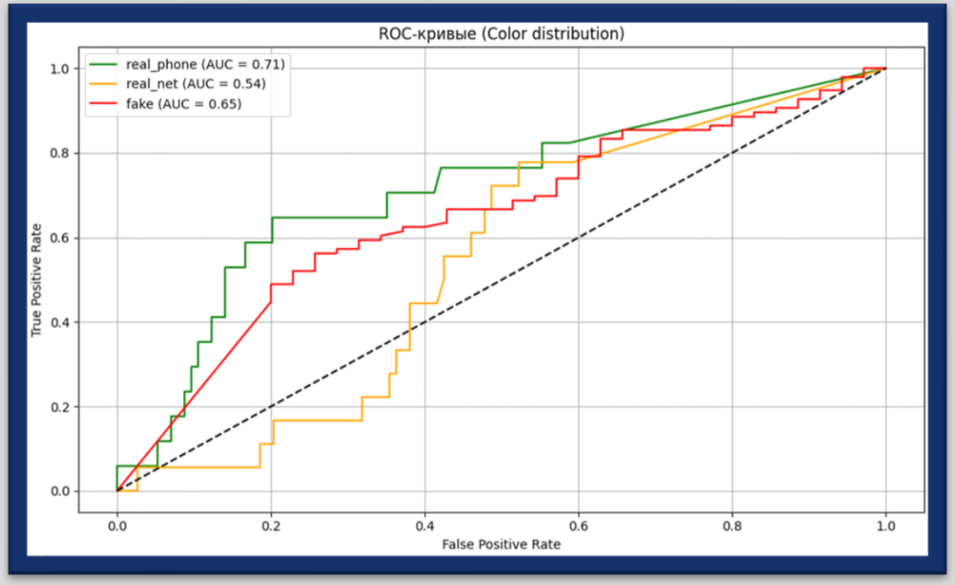
- Average values and standard deviations of Cr, Cb
- The shape of the channel histograms
- The difference between channels

Parameters

- Weighting factor for each channel
- The threshold of deviation from the norm
- Normalization method (logarithmization)

Result

AUC ≈ 0.65
The method works best in combination with others. By itself, it is sensitive, but it can be wrong in the case of artistic filters.



Metadata analysis method

Camera images almost always contain EXIF: camera, date, geo, software.
Generated images often do not contain EXIF or contain suspicious fields (Software = “Stable Diffusion”)

Signs used

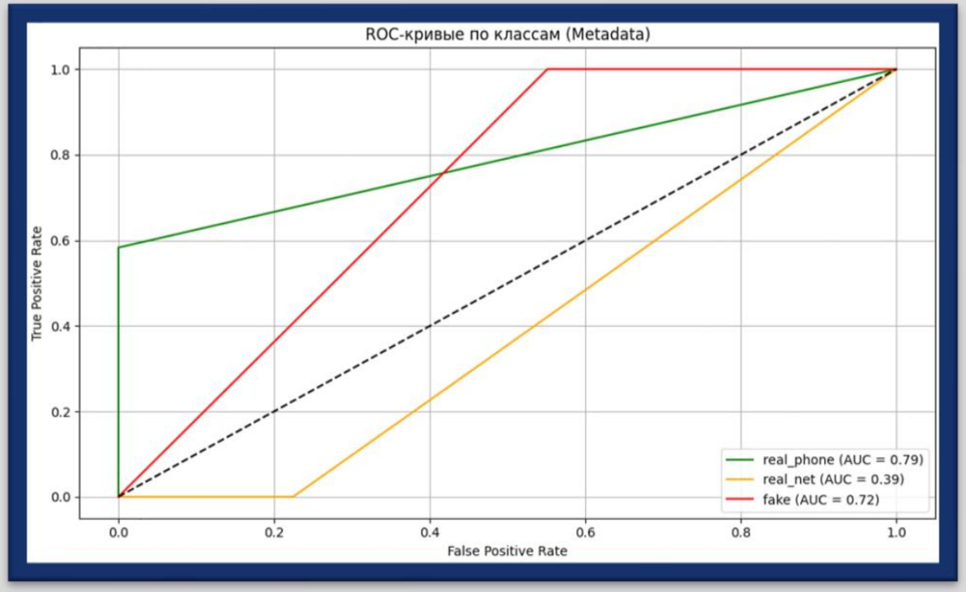
- EXIF availability
- The presence of the "camera" field
- The name of the programs
- DPI, ICC, dimensions
- JPEG quantization

Parameters

- Weight of each feature (w_exif, w_camera etc.)
- Which DPI is considered suspicious
- What is considered an "unnatural" size?

Result

AUC ≈ 0.53–0.85
It depends very much on the images. The original real photos are accurate.
It may behave erratically on jpeg photos from messengers.



A method for detecting generation artifacts

The method finds abnormal borders and textures in the image. Generators often create unrealistic abrupt transitions (around the eyes, background, and clothing).

Signs used

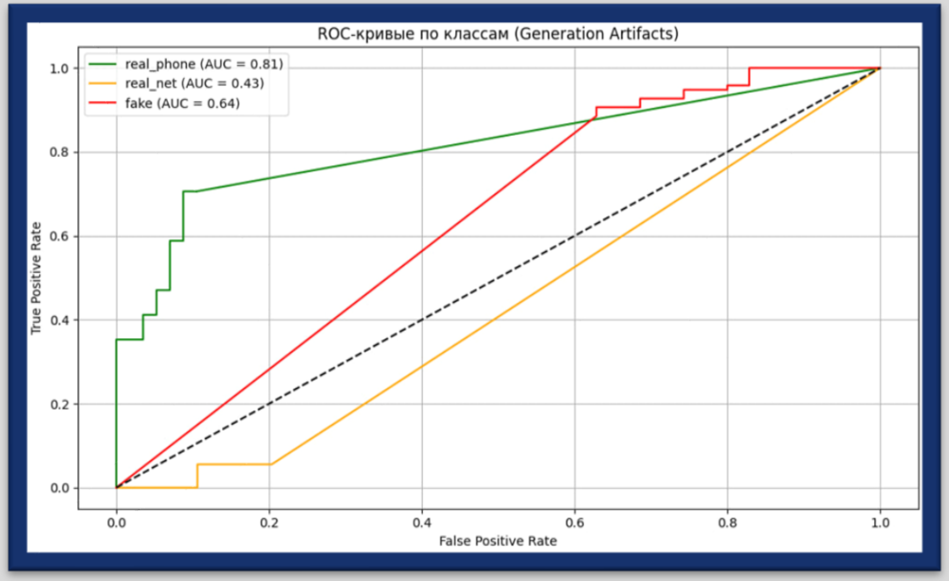
- Gradients according to Laplace and Sobel
- Average gradient strength
- Density of "artifact" zones

Parameters

- Gradient strength threshold
- Window size (if we count local artifacts)
- Normalization scale

Result

AUC ≈ 0.64
The method provides additional features, especially for GAN. Best results in combination.



Architecture

The application is divided into modules:

1. Methods (5 separate files)
2. Parameter selection scripts
3. Scripts for applying the method to the dataset and displaying the results
4. Analysis and visualization

Experiment planning

The dataset is divided into 3 classes:

1. real_phone — photos taken on the phone (contain real EXIF data)
2. real_net - images taken from the Internet (EXIF is often missing or damaged)
3. fake images generated by AI

Basic metrics

- AUC (Area Under Curve) - How well does the method distinguish between classes
- ROC curve - visual interpretation of the model's operation
- F1-score - the balance between precision and recall
- Accuracy - the proportion of correct predictions

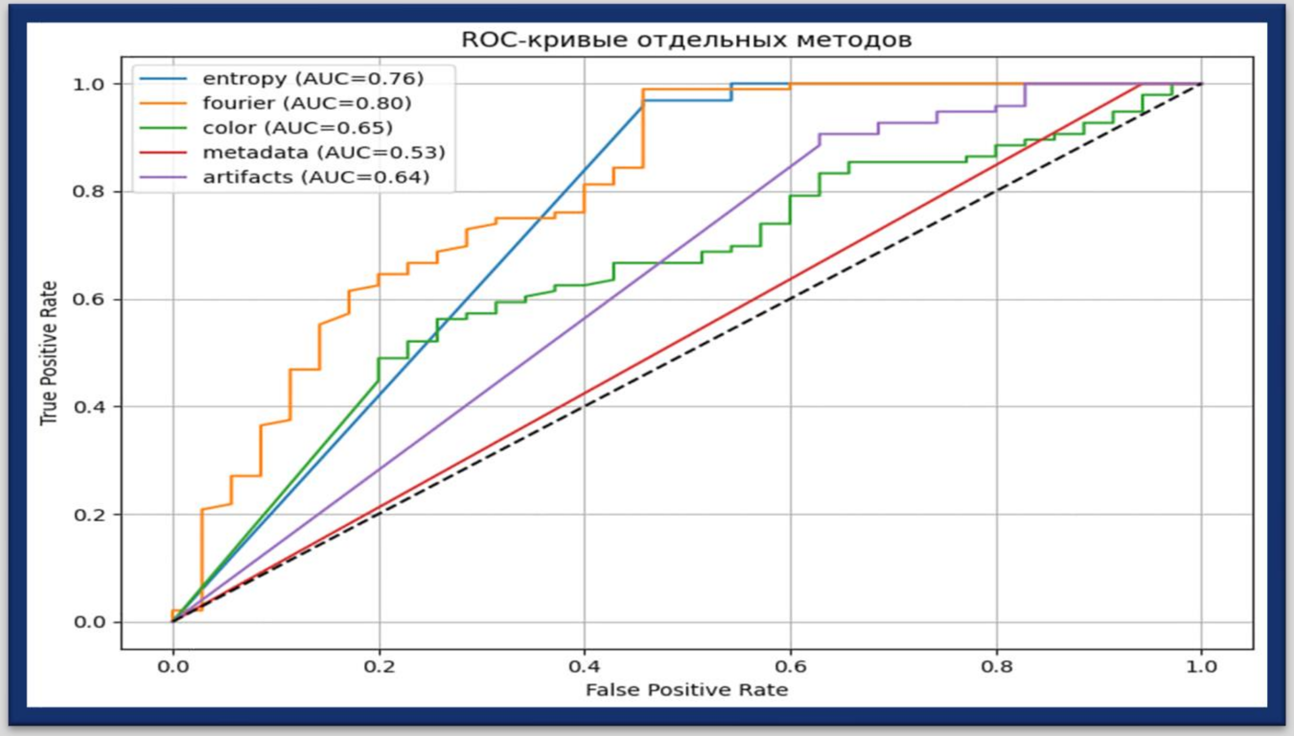
real_phone

fake

real_net

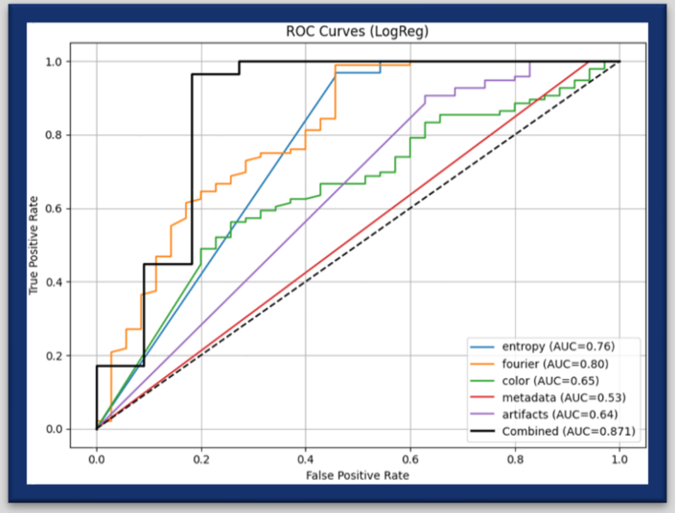
Experimental results

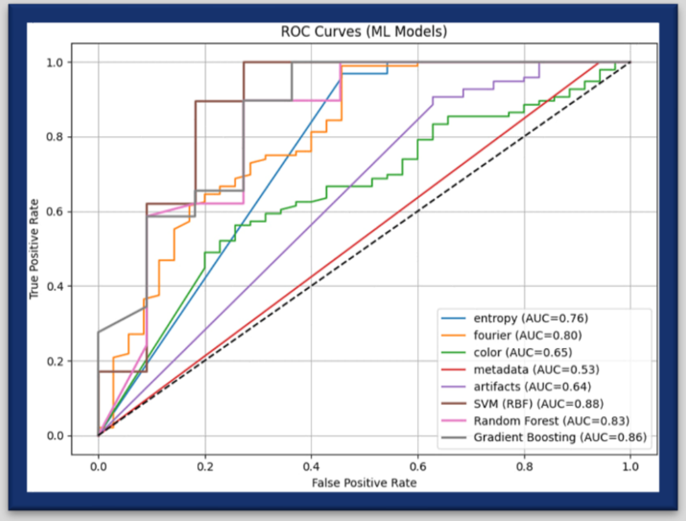
ROC curves, optimal parameters, and AUC values were calculated for each method.

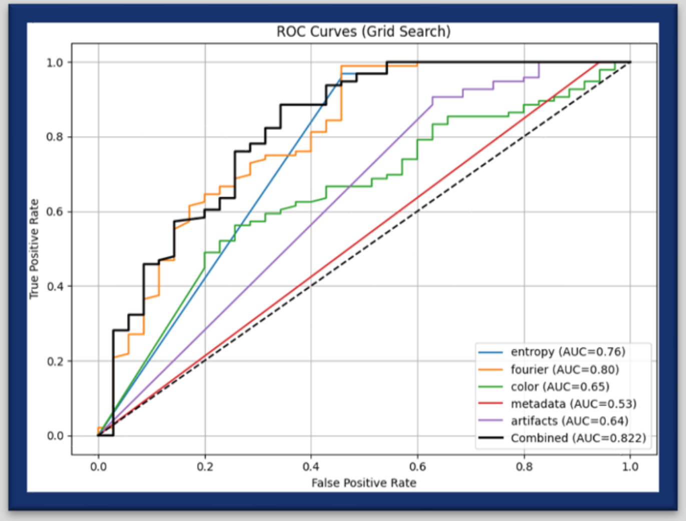


Combined models

The methods of logistic regression, weight sorting, and machine learning (SVM, Boosting) were used.







The main conclusion

The best single method: Fourier (AUC = 0.80)
Combined SVM (RBF) model gave AUC = 0.88

Directions for further work

- Video analysis (by frames)
- Adding new features (EVERYTHING, symmetries)
- Model training on new generators (Midjourney, DALE 3)
- GUI interface development

List of sources

- Zhang et al., 2019 – "Detecting AI-Generated Images via Entropy Analysis"
- URL: https://blog.frohrer.com/detecting-ai-generated-images-using-entropy-analysis/
- [2] Durall et al., 2020 – "Unmasking DeepFakes with simple features"
- [3] McCloskey Albright, 2018 – "Detecting GAN-generated images through color inconsistencies"
- [4] Verdoliva, 2020 – "Media Forensics and DeepFakes: An Overview"
- [5] Wang et al., 2020 – "CNN-generated images are surprisingly easy to spot. . . for now"