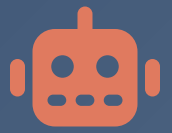


AI-Generated Face Detection

Using Classical Machine Learning

Μηχανική Μάθηση

ΔΠΜΣ «Τεχνητή Νοημοσύνη» 2025/26

Αριστοτέλης Γκάρο mtn2503 · Κωνσταντίνος Δημόπουλος mtn2507 · Απόστολος Σταύρου mtn2526

Introduction

The Problem: Deepfakes & Synthetic Media

The Rise of GANs

StyleGAN (NVIDIA) generates photorealistic faces by separating high-level attributes from stochastic variation. These faces are virtually indistinguishable from real photographs.

Risks & Threats

- **Disinformation:** Bot networks using AI faces to feign legitimacy
- **Fraud:** Synthetic identities for KYC bypass
- **Harassment:** Non-consensual imagery creation

Current Solutions: Deep Learning

State-of-the-art detection uses CNNs (XceptionNet, EfficientNet) achieving >99% accuracy.

Critical Limitations:

- **High Computational Cost:** Requires high-end GPUs
- **"Black Box" Nature:** No insight into why images are flagged

Our Approach

Use **Classical Machine Learning** for a transparent, CPU-efficient alternative.

REAL

FAKE

REAL

FAKE

Goals

Project Objectives

1

Identify Artifacts

Determine which specific visual domains (Texture, Frequency, Color, or Shape) contain the most persistent StyleGAN artifacts that can be exploited for detection.

2

Resource Efficiency

Demonstrate that a lightweight classical ML classifier can achieve reasonable accuracy without requiring GPU acceleration, making it possibly suitable for edge devices and low-resource environments.

3

Interpretability

Provide clear visual explanations and feature importance analysis for classification decisions, understanding exactly why an image is flagged as fake rather than treating the model as a black box.

Data

Dataset: 14k Real and Fake Faces

Total Images

14,000

Balanced Dataset

Real Faces

7,000

FFHQ Dataset

AI-Generated

7,000

StyleGAN

Data Split

Set	Count	Purpose
Training	10,000	Model fitting & CV
Validation	2,000	Parameter tuning
Test	2,000	Final evaluation

Preprocessing

- **Resize:** 128×128 pixels (cv2.INTER_AREA)
- **Grayscale:** cv2.COLOR_BGR2GRAY (except color features)
- **Standardization:** StandardScaler for feature normalization
- **Format:** .npz files for efficient pipeline processing

Data Source: "140k Real and Fake Faces" - Flickr-Faces-HQ (real) vs StyleGAN (fake), maintaining 50/50 split across all subsets.

Methodology

Feature Extraction Pipeline

H HOG
Histogram of Oriented Gradients

Purpose: Captures shape and edge information

Why: StyleGAN shows structural inconsistencies in peripheral areas (ears, hair, background)

Features: 1,764

L LBP
Local Binary Patterns

Purpose: Texture descriptor for surface analysis

Why: GAN skin is often oversmoothed; real skin has complex texture from pores, wrinkles

Features: 59

C Color
Statistical Color Analysis

Purpose: Global color distribution and contrast

Why: GANs struggle with color constancy, producing unnatural saturation or color casts

Features: 6

G Gabor
Gabor Filters

Purpose: Frequency domain analysis

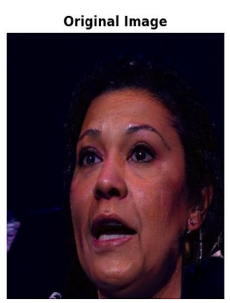
Why: Detects checkerboard artifacts from upsampling layers in GAN generators

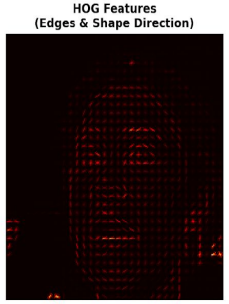
Features: 8

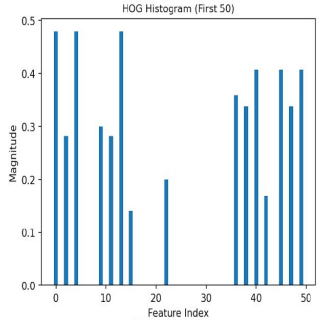
Total Feature Vector: 1,837 dimensions

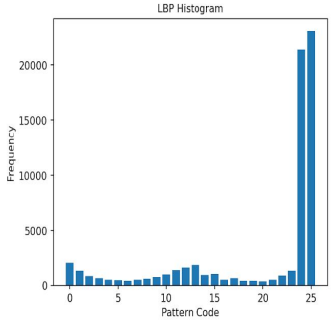
Concatenated features capture shape, texture, color, and frequency information

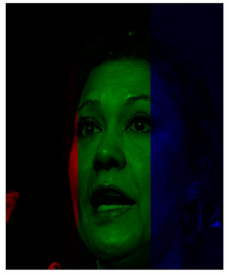
Feature Visualization

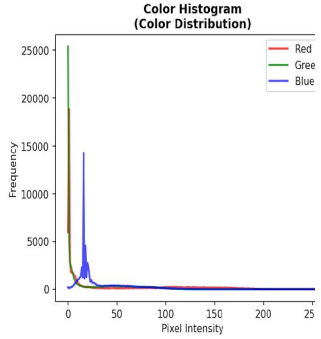


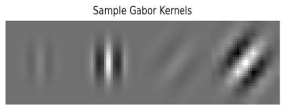


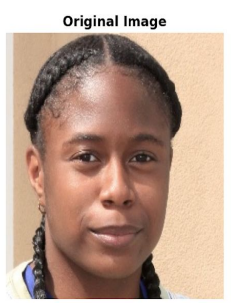


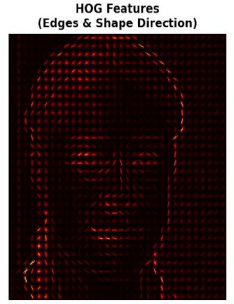


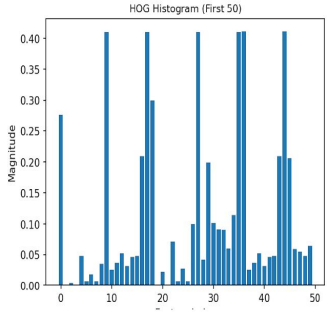


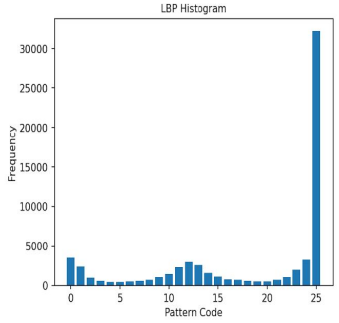


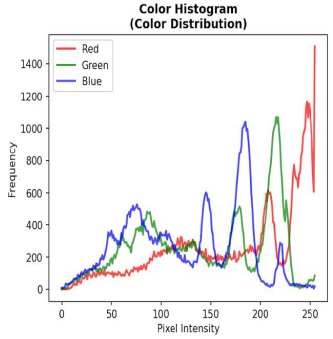


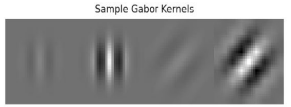












FAKE

REAL

Methodology

Classification Algorithms

Support Vector Machine

Approach: Finds optimal hyperplane maximizing margin between classes using RBF kernel

Strengths: Excellent with high-dimensional data, robust decision boundary

Config: RBF kernel, C=10, γ='scale'

Random Forest

Approach: Ensemble of uncorrelated decision trees using bagging

Strengths: Feature importance scores, robust to high dimensionality

Config: 350 trees, max_depth=50

Gradient Boosting

Approach: Sequential ensemble focusing on residuals of previous trees

Strengths: Catches hard-to-classify examples, minimizes prediction errors

Config: 200 estimators, lr=0.2, max_depth=5

k-Nearest Neighbors

Approach: Instance-based learning using majority vote of k closest neighbors

Strengths: Captures local patterns, effective for irregular boundaries

Config: k=1, Minkowski distance, uniform weights

Results

Classifier Performance Comparison

Classifier	Accuracy	F1-Score	ROC-AUC
SVM	**85.15%**	**0.8481**	**0.9010**
Random Forest	74.90%	0.7546	0.7912
Gradient Boosting	77.90%	0.7962	0.8063
kNN	60.90%	0.4155	0.6090

Winner: SVM

SVM achieved superior performance across all metrics with **85.15% accuracy** and **0.9010 ROC-AUC**, demonstrating its effectiveness for high-dimensional feature spaces and non-linear decision boundaries.

Analysis

Key Findings: Feature Analysis & PCA

Feature Importance Discovery

Unlike deep learning models, our analysis revealed that **HOG (shape/edges)** was the dominant discriminator.

Insight: StyleGAN's primary weakness lies in its inability to generate consistent edges, particularly in peripheral areas like ears, hair boundaries, and background transitions.

PCA Dimensionality Reduction

Applied PCA with **95% variance threshold** to reduce feature dimensionality:

Before PCA	After PCA
1,837 features	480 features

Impact: 74% reduction with minimal accuracy loss

Training Time Improvement

85%

47s → 7s (SVM)

Gabor Filter Cost

90.4%

of preprocessing time

Accuracy Loss

0.65%

when removing Gabor

Recommendation: Exclude Gabor filters from production pipelines. They account for 90.4% of preprocessing time (0.234s vs 0.024s for all others) but contribute minimally to accuracy.

Analysis

Model Extrapolation: Can we predict “fakeness” in other datasets?

False Positives

Fake images misclassified as real (1000/1000)

Misclassification: High-quality SDXL outputs with exceptional realism and natural-looking texture variation that successfully bypass detection mechanisms

False Negatives

Real images misclassified as fake (157/1000)

“Misclassification": Model possibly learned that professionally edited real photos may look smooth and synthetic

True Positives

Real images classified as real (843/1000)

Classification: Model successfully identified 843 real images by detecting smoothness patterns.

True Negatives

Fake images classified as fake (0/1000)

“Misclassification": The model flagged every single fake image as real, demonstrating complete inability to recognize other than StyleGan photos

Comparison

Computational Efficiency vs. Deep Learning

Our Approach: Classical ML

Accuracy	85.15%
Training Time On standard CPU (no GPU required)	< 8 min
Hardware No specialized hardware needed	CPU Only
Interpretability Clear feature importance analysis	✓ High

Deep Learning: CNNs

Accuracy XceptionNet, EfficientNet	~99%
Training Time Requires GPU acceleration	Hours
Hardware High-end NVIDIA GPUs	GPU Required
Interpretability Black box nature	✗ Low

Summary

Conclusions & Future Work

Key Takeaways

1. SVM + HOG/LBP Fusion: Provides a robust, interpretable, CPU-efficient alternative for AI-generated face detection with 85.15% accuracy.

2. Edge Detection Dominance: HOG features (shape/edges) were the strongest discriminator, revealing StyleGAN's weakness in generating consistent edges.

3. Computational Efficiency: Under 8 minutes training on CPU vs. hours on GPU for deep learning, making it viable for resource-constrained environments.

4. Interpretability: Insight on why images are flagged, enabling targeted improvements and trust in the system.

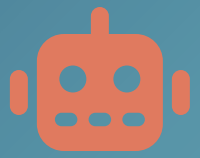
Future Work

- **Ensemble Methods:** Combine multiple classifiers for improved accuracy
- **Additional Features:** Explore wavelet transforms and frequency analysis
- **Architecture Transfer:** Test on other GAN architectures (ProGAN, BigGAN)
- **Hybrid Approaches:** Combine classical features with lightweight CNNs

Practical Impact

This work demonstrates that classical ML techniques remain relevant in the deep learning era for specific applications requiring:

- ✓ Transparency and explainability
- ✓ Deployment on edge devices
- ✓ Rapid prototyping and iteration
- ✓ Low infrastructure costs



Thank You

Questions & Discussion