Development of a memory-efficient machine learning pipeline for fake image detection using statistical features

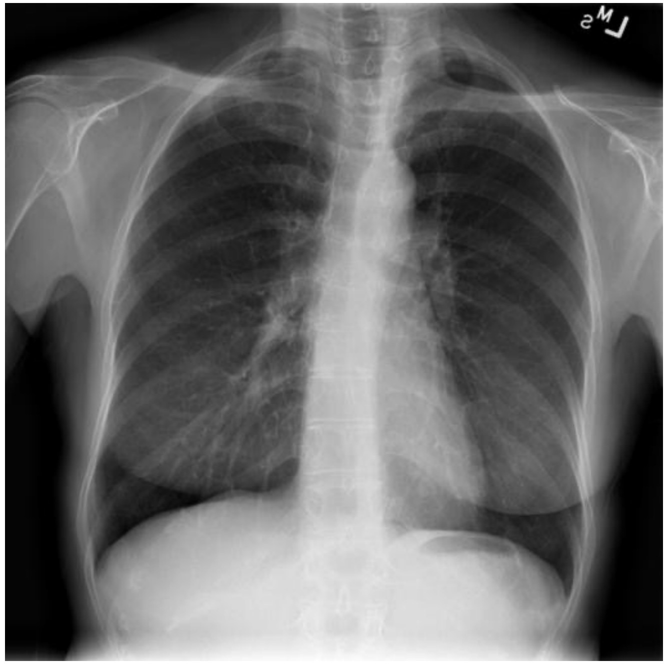
Shiwon Kim,
Digital Healthcare Lab,
Department of Digital Analytics,
Yonsei University College of Computing

Problem Definition

Synthetic Chest X-ray Image | Generated Image | Open Dataset

- syntheticData_AI_Detector
  - non_synthetic
    - Normal
    - Pneumonia
  - synthetic
    - Normal
    - Pneumonia

Type: real (non-synthetic)
Diagnosis: normal



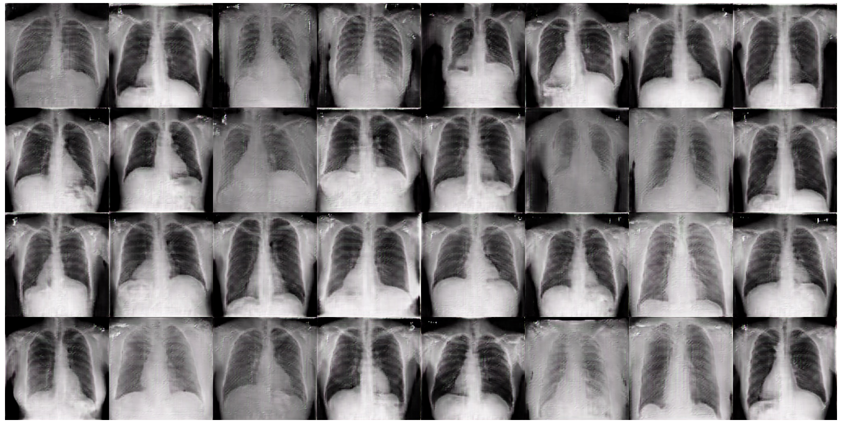
Type: synthetic
Diagnosis: normal

Problem Definition

Synthetic Chest X-ray Image | **Generated Image** | Open Dataset

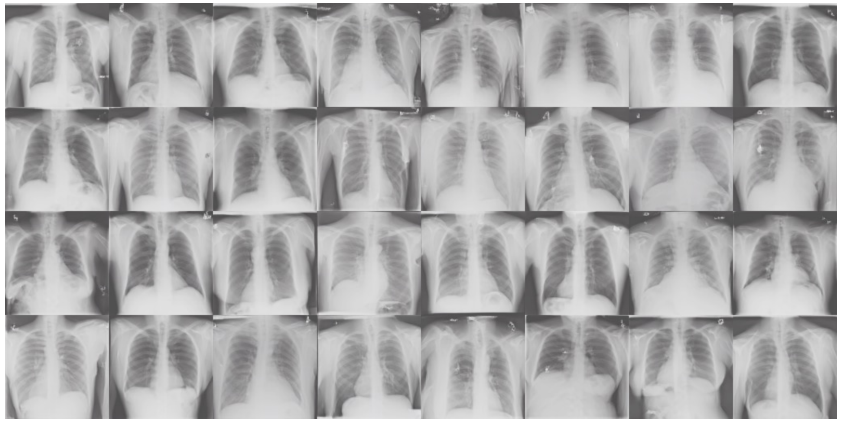
GAN

- Generative adversarial networks
- DCGAN (Deep Convolutional GAN)[1]



Diffusion

- Probabilistic generative models
- DDPM (Denoising Diffusion Probabilistic Models)[2]



Problem Definition

Synthetic Chest X-ray Image Generated Image **Open Dataset**

a. CIFAKE: Real and AI-Generated Synthetic Images[3]
b. OpenForensics: Multi-Face Forgery Detection and Segmentation In-The-Wild[4]

Information	CIFAKE	OpenForensics
# of images	60,000 real / 60,000 fake	115,325 (multi-face)*
Generation method	Stable diffusion	GAN
Pair-wise	Y	N
Multi-object	N	Y

Table 1. Information about the datasets used for fake image detection experiments.

* 16,027 real faces and 173,660 forged faces in 115,325 annotated images

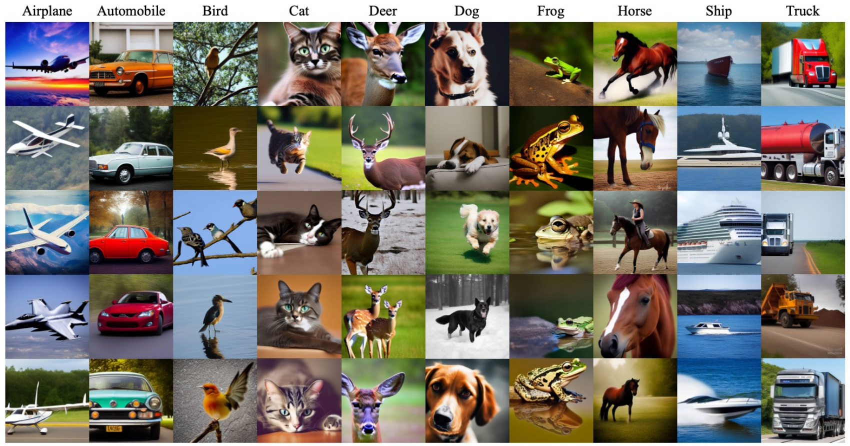
Problem Definition

Synthetic Chest X-ray Image

Generated Image

Open Dataset

a. CIFAKE: Real and AI-Generated Synthetic Images[3]
b. OpenForensics: Multi-Face Forgery Detection and Segmentation In-The-Wild[4]



CIFAKE

OpenForensics

Methods

Fake Image Detection Pipeline

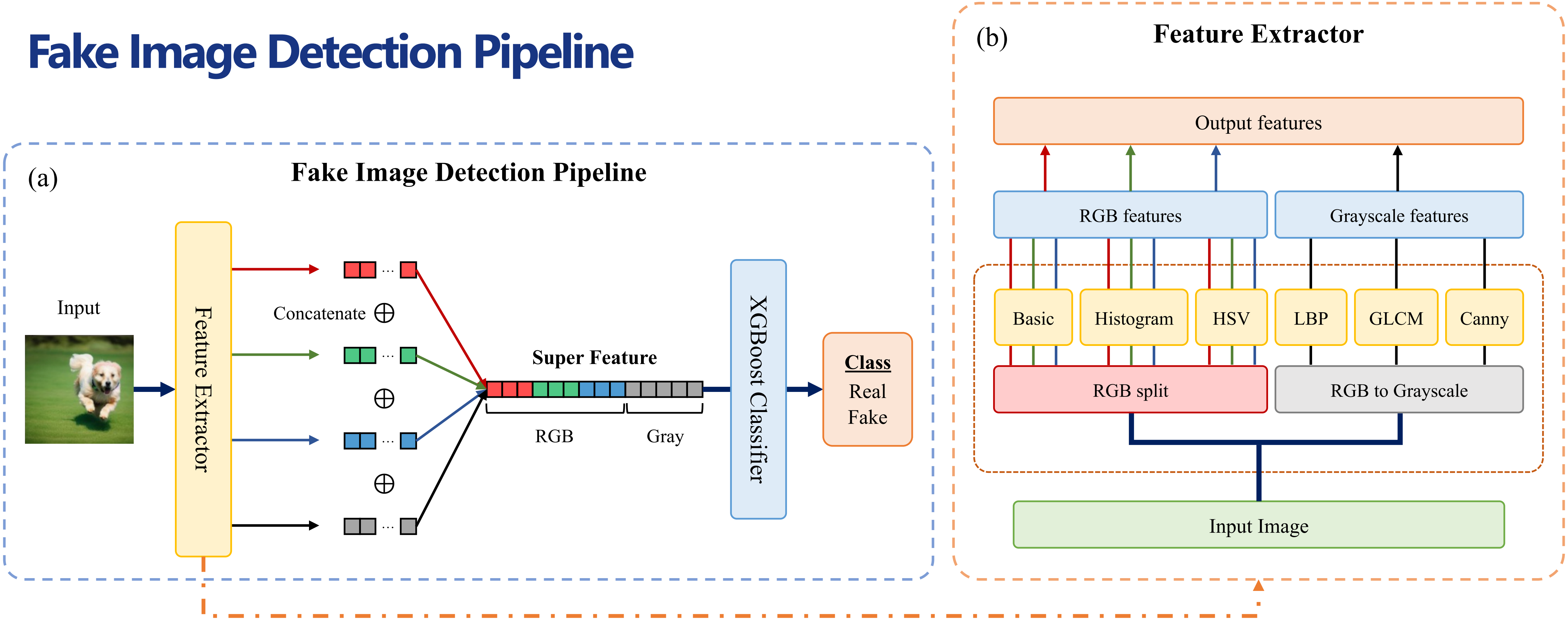


Fig 1. Overview of the fake image detection pipeline.

Methods

Fake Image Detection Pipeline

- Feature extraction
  I. Basic statistical values
    - Mean, standard deviation, median, skewness, kurtosis

  II. Image texture pattern
    - LBP (Local Binary Patterns), Haralick texture & GLCM (Gray-Level Co-Occurrence Matrix)

  III. Image pixels distribution
    - Image histogram, HSV color space

  IV. Image edge / boundary
    - Canny edge

Methods

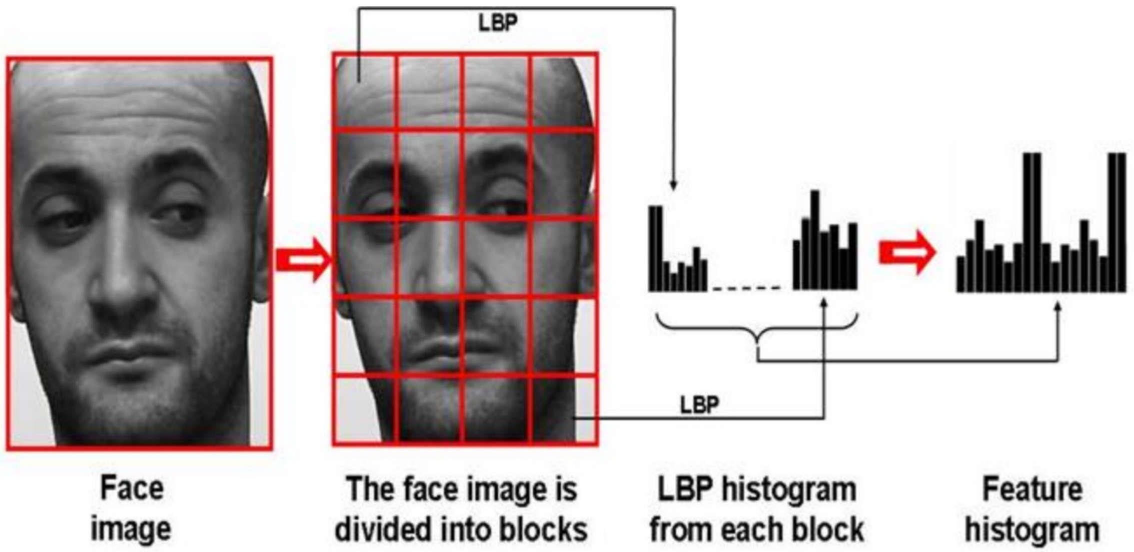
Feature Information

- Basic statistical values
  - Mean $\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$
  - Standard deviation $\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N}$
  - Median
  - Skewness $Skew = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(X_i - \bar{X})}{\sigma}\right]^3$
  - Kurtosis $Kurt = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(X_i - \bar{X})}{\sigma}\right]^4$

Methods

Feature Information

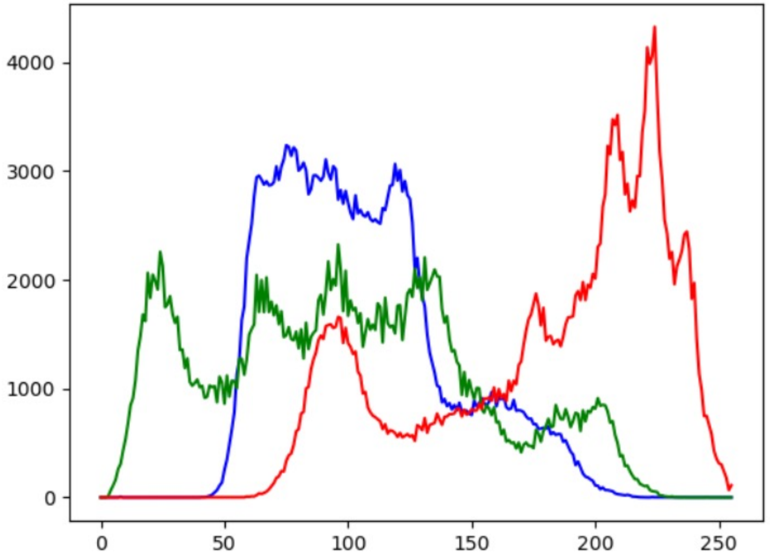
- LBP (Local Binary Pattern)[5]
  - The LBP methodology has led to significant progress in texture analysis
  - Has been very successful in **computer vision** problems such as face analysis and motion analysis



Methods

Feature Information

- Image Histogram[6]
  - Displays image characteristics with image pixel values on the x-axis and frequency on the y-axis
  - A graph showing the distribution of the number of bright and dark pixels



Methods

Feature Information

- GLCM (Gray-Level Co-Occurrence Matrix)[7,8]
  - In statistical texture analysis, texture features are computed from the **statistical distribution** of observed **combinations of intensities** at specified positions relative to each other in the image
  - The GLCM method is a way of extracting the second order statistical texture features

neighbour pixel value ---> ref pixel value:	0	1	2	3
0	0,0	0,1	0,2	0,3
1	1,0	1,1	1,2	1,3
2	2,0	2,1	2,2	2,3
3	3,0	3,1	3,2	3,3

0	1	0	2
0	2	1	1
3	1	0	0
0	0	2	3

4 x 4 image

2	1	3	0
2	1	0	0
0	1	0	1
0	1	0	0

GLCM

0.16	0.08	0.25	0
0.16	0.08	0	0
0	0.08	0	0.08
0	0.08	0	0

Nomalized GLCM

Methods

Feature Information

- GLCM (Gray-Level Co-Occurrence Matrix)[7,8]
  - GLCM-based features

$$Homogenity = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(i-j)^2}$$

$$Entropy = \sum_{i,j=0}^{N-1} -\log(P_{ij})P_{ij}$$

$$Contrast = \sum_{i,j=0}^{N-1} P_{i,j}(i-j)^2$$

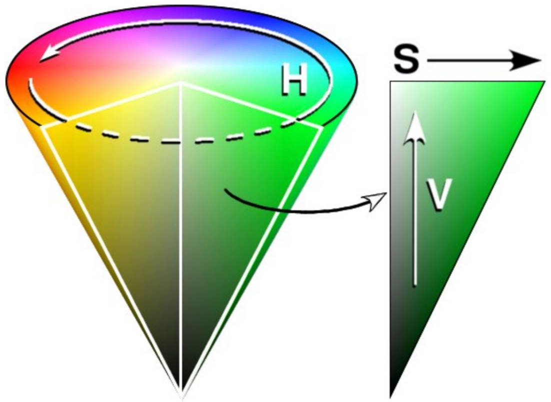
$$Correlation = \sum_{i,u=0}^{N-1} P_{i,j}\frac{(i-\mu)(j-\mu)}{\sigma^2}$$

$$Energy = \sum_{i,j=0}^{N-1} (P_{ij})^2$$

Methods

Feature Information

- HSV Color Space[9]
  - **H**ue
  - **S**aturation
  - **V**alue



- Canny Edge detection[10]
  - Gaussian filtering
  - Gradient calculation
  - Non-maximum suppression
  - Hysteresis edge tracking

Results

Fake Image Detection Performance

a. CIFAKE

- 5-folds average / (Train + valid) 80% and test 20%

Metrics	Ours	ResNet-50
Accuracy	**0.900**	0.779
Sensitivity (recall)	**0.902**	0.788
Specificity	**0.897**	0.795
AUROC	**0.965**	0.871
Training time	1269s	**1201s**
GPU memory allocation	**285MB**	1991MB

Table 2. Comparison of our method and ResNet-50 on AI-generated image detection with CIFAKE dataset.

Results

Fake Image Detection Performance

b. OpenForensics

- 5-folds average / (Train + valid) 80% and test 20%

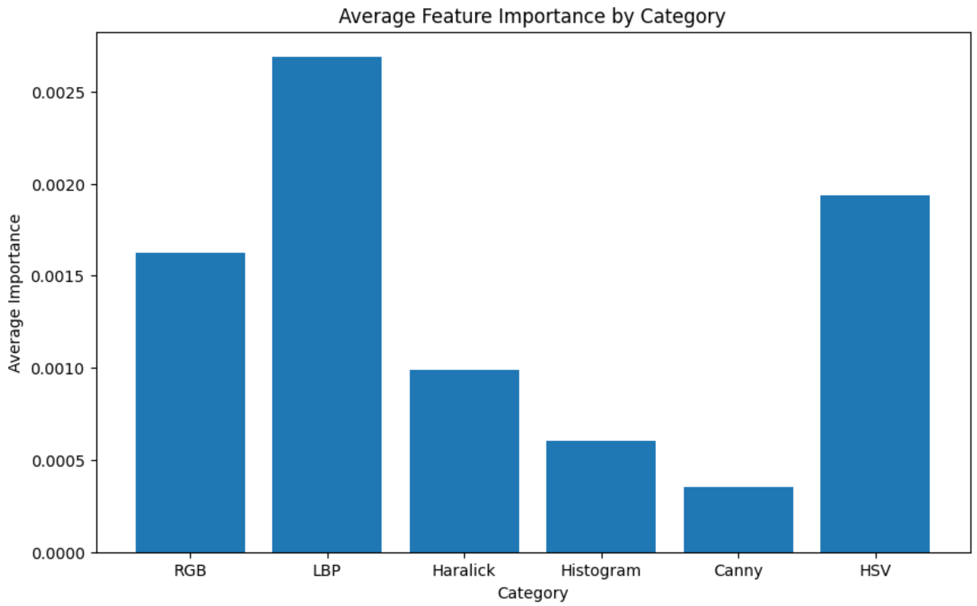
Metrics	Ours	ResNet-50
Accuracy	**0.668**	0.559
Sensitivity (recall)	**0.658**	0.534
Specificity	**0.673**	0.584
AUROC	**0.729**	0.580
Training time	**694s**	803s
GPU memory allocation	**319MB**	1971MB

Table 3. Comparison of our method and ResNet-50 on forged face detection with OpenForensics dataset.

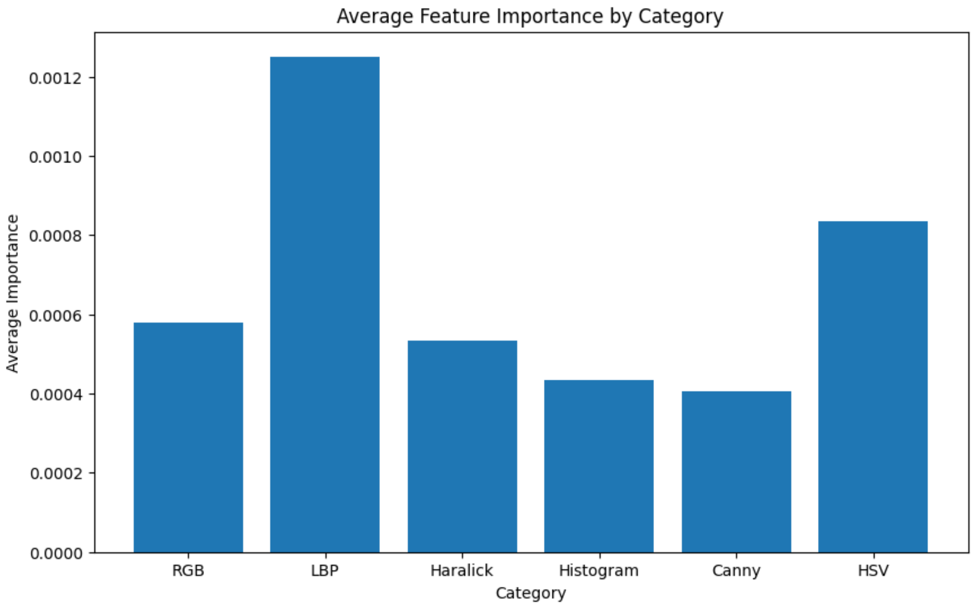
Results

Model Explainability

- Average feature importance by category
  - Basic statistics (RGB) / LBP / GLCM (Haralick) / Histogram / Canny edge / HSV



CIFAKE



OpenForensics

Results

Ablation Study

- Average performance by number of features
  - CIFAKE / 5-folds average

Metrics	Accuracy	Sensitivity	Specificity	AUROC
(1) LBP	0.696	0.676	0.718	0.765
(2) LBP+HSV	0.818	0.821	0.819	0.904
(3) LBP+HSV+RGB	0.838	0.845	0.833	0.921
(4) LBP+HSV+RGB+HAR	0.844	0.845	0.844	0.924
(5) LBP+HSV+RGB+HAR+HIS	0.898	0.903	0.900	0.967
(6) LBP+HSV+RGB+HAR+HIS+CAN	0.900	0.902	0.897	0.965

Table 4. Average performance by number of features on CIFAKE.

Results

Ablation Study

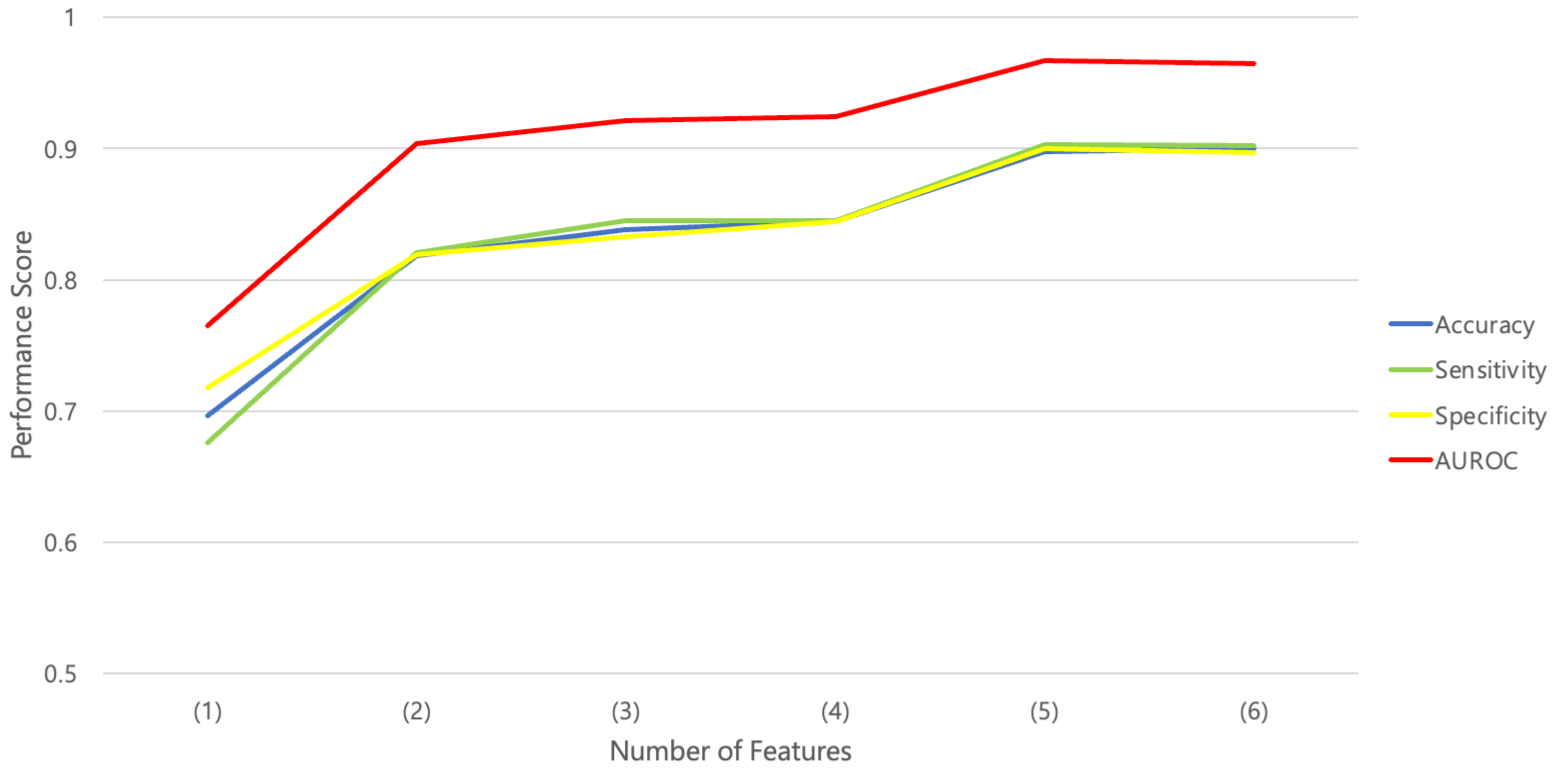


Fig 2. Visualization of average performance by number of features on CIFAKE.

Conclusion

- Developed an accurate but memory-efficient fake image detection pipeline
  - Proved the superiority of our method over ResNet-50 via 5-fold validation on various datasets
  - Quantitatively verified the efficiency of our pipeline in terms of **time and memory consumption**
- Analyzed feature importance for enhanced model explainability
  - Proved the robustness of our method in detecting fake images created using generative AI models
  - Observed **the effectiveness of each feature** in improving the performance of fake image detection

References

[1] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434 (2015).

[2]Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.

[3] Bird, J.J., Lotfi, A. (2023). CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. arXiv preprint arXiv:2303.14126.

[4] Le, Trung-Nghia et al. "OpenForensics: Large-Scale Challenging Dataset For Multi-Face Forgery Detection And Segmentation In-The-Wild." 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021): 10097-10107.

[5] Pietikäinen, Matti, et al. Computer vision using local binary patterns. Vol. 40. Springer Science & Business Media, (2011).

[6] Raju, P. Daniel Ratna, and G. Neelima. "Image segmentation by using histogram thresholding." International Journal of Computer Science Engineering and Technology 2.1 (2012): 776-779.

[7] Mohanaiah, P., P. Sathyanarayana, and L. GuruKumar. "Image texture feature extraction using GLCM approach." International journal of scientific and research publications 3.5 (2013): 1-5.

[8] 윤종일, and 김종배. "GLCM 특징정보 기반의 자동차 종류별 분류 방안." 한국정보처리학회 학술대회논문집 18.1 (2011).

[9] Sural, Shamik, Gang Qian, and Sakti Pramanik. "Segmentation and histogram generation using the HSV color space for image retrieval." Proceedings. International Conference on Image Processing. Vol. 2. IEEE, 2002.

[10] Bao, Paul, Lei Zhang, and Xiaolin Wu. "Canny edge detection enhancement by scale multiplication." IEEE transactions on pattern analysis and machine intelligence 27.9 (2005): 1485-1490.

Appendix

		Synthetic CXR	Generated Image	
	Performance Metrics		DCGAN	DDPM
ResNet-50	Accuracy	0.909	0.993	0.880
	Sensitivity (recall)	0.889	0.987	0.852
	Specificity	0.928	1.000	0.906
	AUROC	0.970	0.998	0.953
Ours	Accuracy	**0.994**	**1.000**	**0.964**
	Sensitivity (recall)	**0.998**	**1.000**	**0.972**
	Specificity	**0.992**	**1.000**	**0.953**
	AUROC	**0.995**	**1.000**	**0.970**

Table A1. Comparison of our method and ResNet-50 on Synthetic CXR dataset and DCGAN-/DDPM-generated images.

Appendix

Feature Analysis

- Average performance when only using features from a single category
  - CIFAKE / 5-folds average

Performance Metrics	Types of Features					
	RGB Features			Grayscale Features		
	RGB	Histogram	HSV	LBP	Haralick	Canny
Accuracy	0.742	0.728	0.760	0.696	0.838	0.688
Sensitivity	0.754	0.717	0.760	0.676	0.840	0.660
Specificity	0.728	0.736	0.759	0.718	0.838	0.715
AUROC	0.744	0.808	0.845	0.765	0.920	0.753

Table A2. Average performance of individual feature types.

YONSEI UNIVERSITY
COLLEGE OF MEDICINE