Distinguishing AI-Generated Images from Real Photography

Akin Akinlabi, Cat Weiss, Deva Empranthiri, Gia Nguyen
W207 Summer 2024
Github Repo: https://github.com/akinlaba/final_project_207/tree/main

90%

of online content could be synthetically generated using AI by 2026

Psychology Today: How AI-Generated Content Can Undermine Your Thinking Skills

Objectives

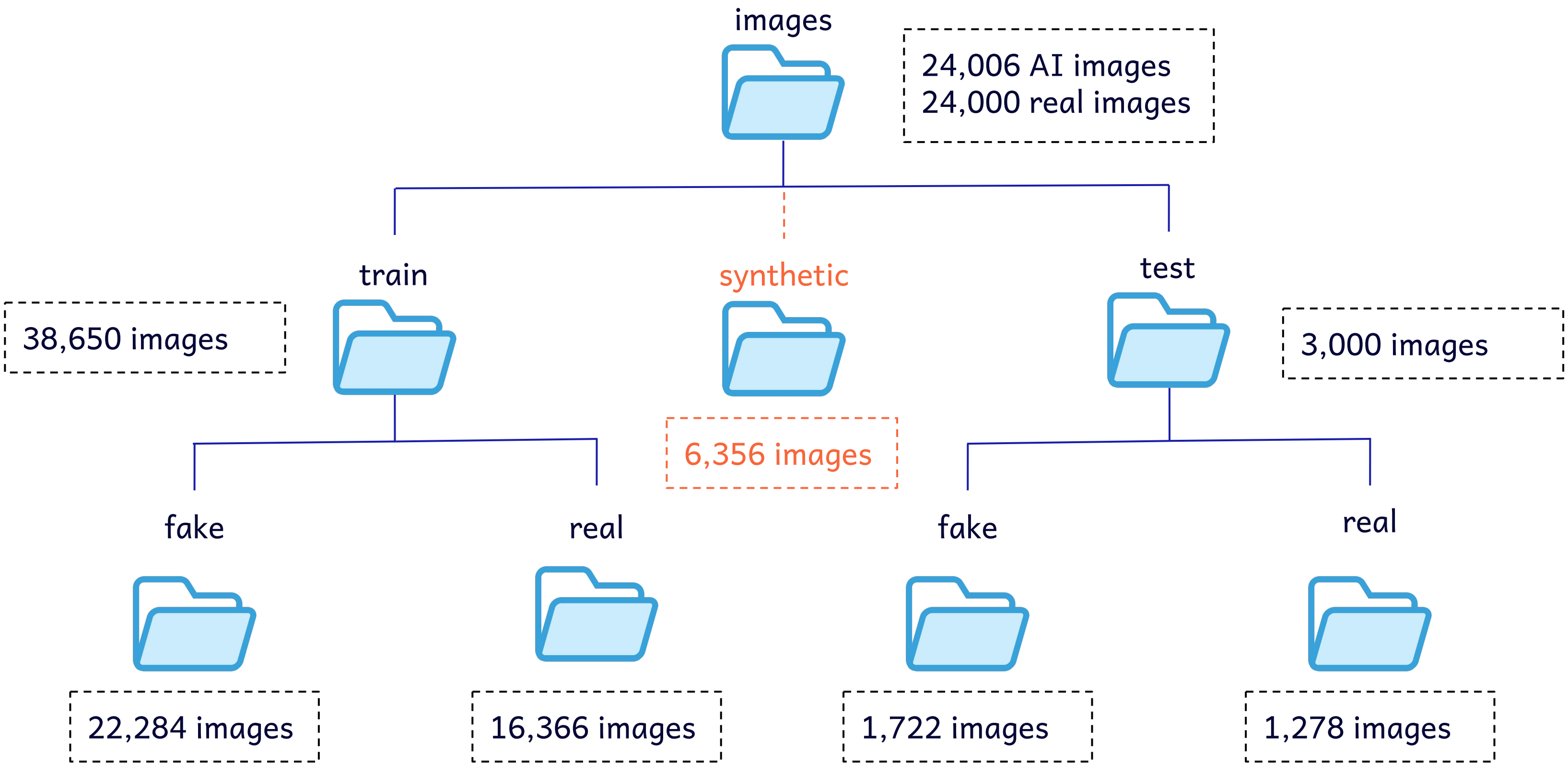
Why?

- Understand the capabilities & limitations of AI in visual content creation
- Address ethical concerns related to authenticity and manipulation in media
- Help maintain trust and integrity in digital imagery

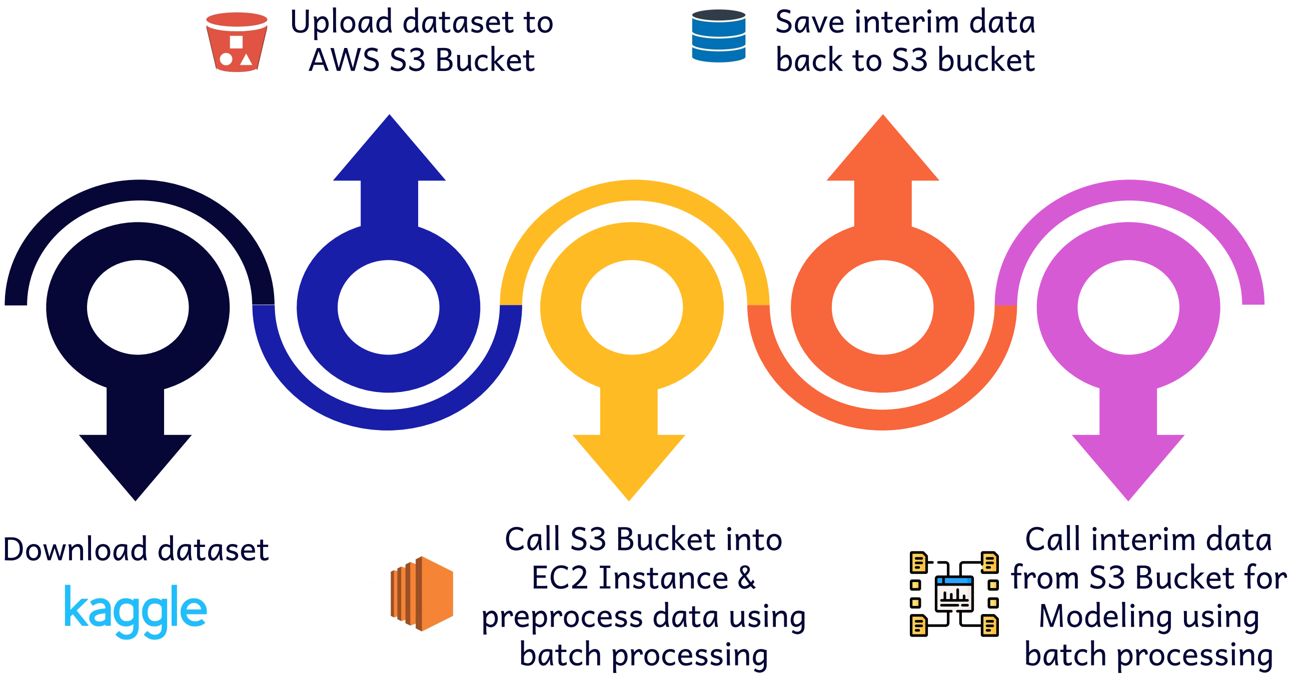
Use Cases

- Enhance media & journalism integrity by verifying the authenticity of images
- Improve AI/ML models for better image classification
- Assist in cybersecurity efforts to detect & prevent the spread of deepfakes & manipulated images

Dataset Overview & Annotation



Data Engineering Architecture



Exploratory Data Analysis

First Real Image

First Fake Image

Exploratory Data Analysis

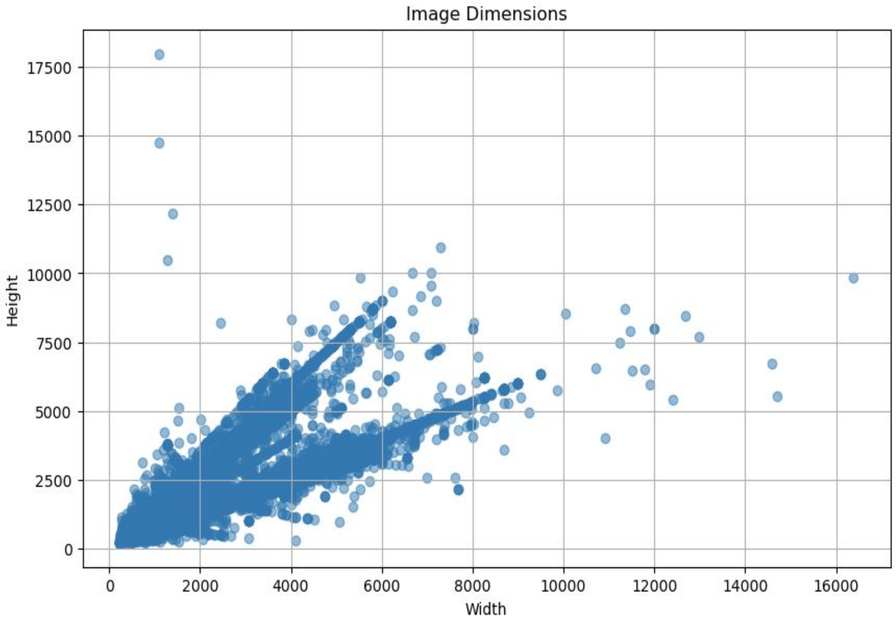


Figure 1: Scatter plot on image dimensions in pixels

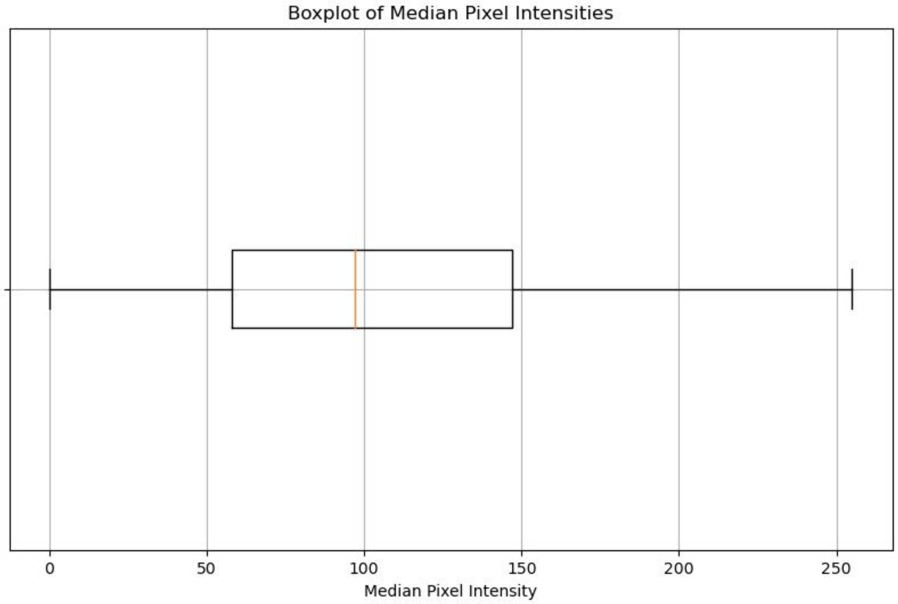


Figure 2: Box plot on median pixel intensities

Data Preparation

Data Cleaning

- Removes corrupted, or incomplete images
- Ensures that the dataset is high quality

Data Augmentation

- Creates new training samples by applying random transformations to existing images
- Techniques used: rotate, flip, adjust brightness, and adjust contrast

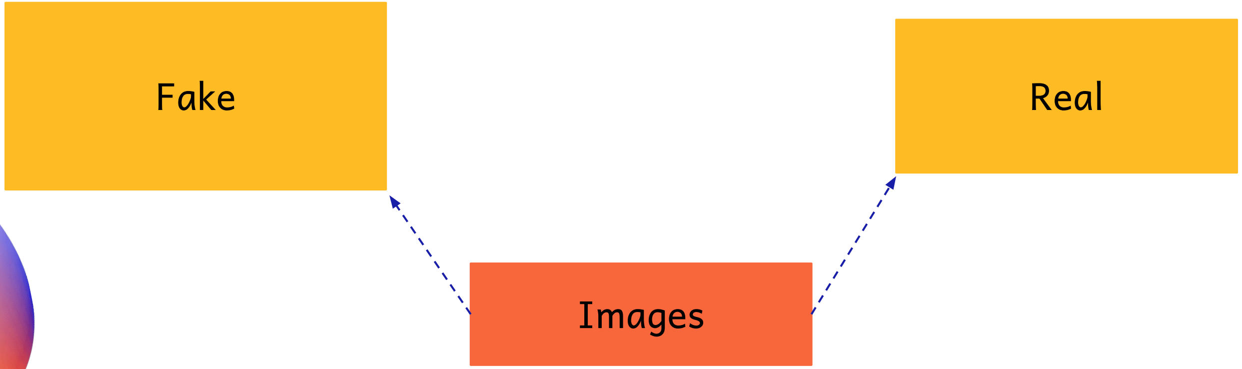
Normalization

- Adjusts the pixel values of images to a common scale
- Helps with faster convergence during training and reduces the risk of gradient vanishing

Resizing

- Changes the dimensions of all images to a consistent size
- Ensures that all images have the same input dimensions required by CNN

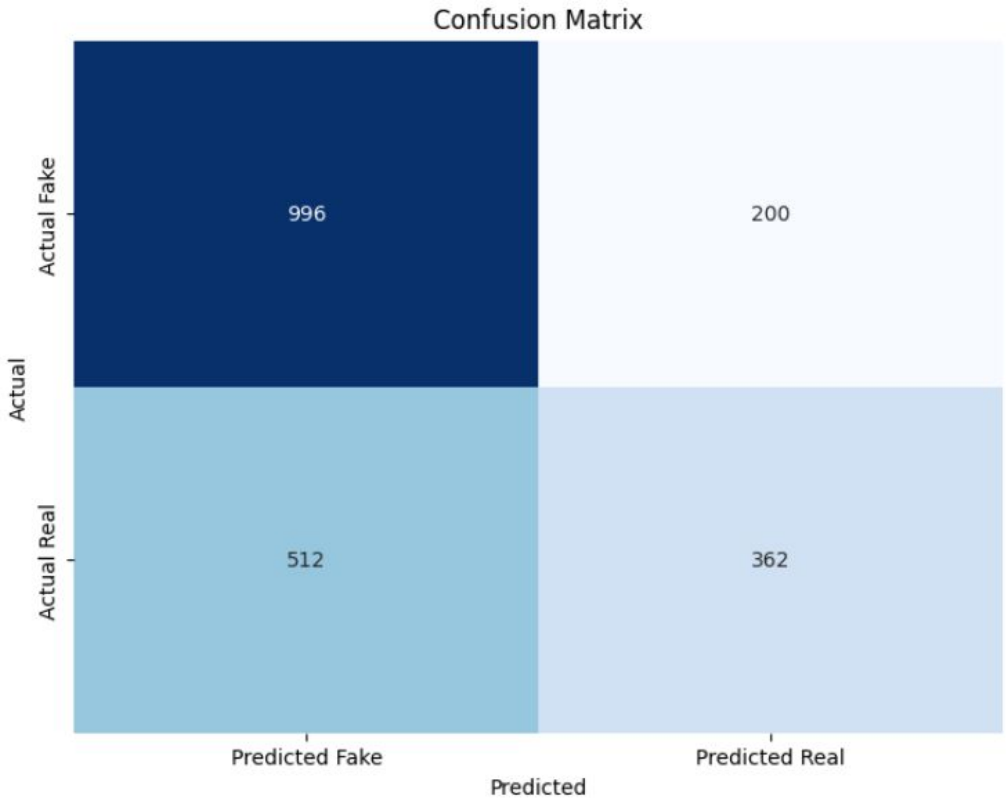
Baseline Model: Dummy Classifier



Model Performance

Training accuracy: 57.6%
Validation accuracy: 57.8%

Model 2: Simple CNN



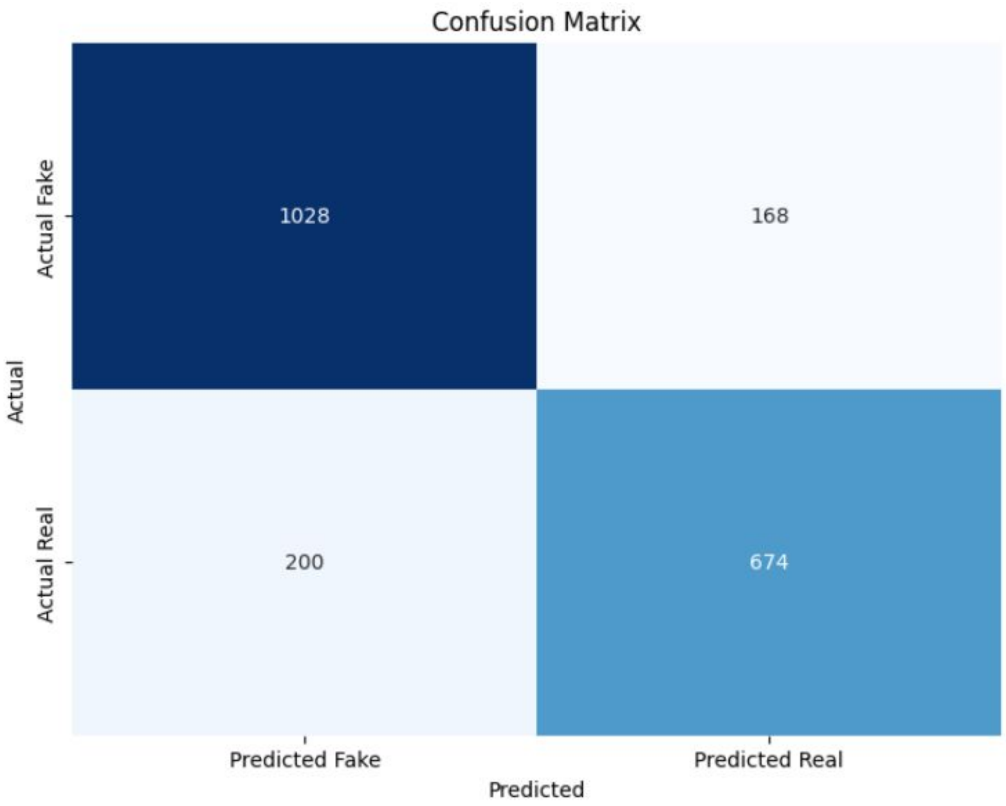
Model Parameters

```
filters: 64
pool_size: 2
dense_layer_units_1: 32
dense_layer_units_2: 24
learning rate: 0.001
```

Model Performance

Training accuracy: 73.5%
Validation accuracy: 65.6%

Model 3: CNN w/ Hyperparameter Tuning



Best Model Parameters

```
filters_1: 112
kernel_size_1: 3
filters_2: 160
kernel_size_2: 3
pool_size: 3
dense_layer_units: 256
learning rate: 0.0001
```

Best Model Performance

Training accuracy: 84.8%
Validation accuracy: 82.2%
Test accuracy: 79.8%

Results

	Baseline Model	Simple CNN	CNN with Hyperparameter Tuning
Total parameters	N/A	80,293,777	227,700,389
Training accuracy	57.6%	73.5%	**84.8%**
Validation accuracy	57.8%	65.6%	**82.2%**
Test accuracy	N/A	N/A	**79.8%**

Model Predictions Successes & Failures

Predicted Real

Predicted Fake

Actual Real

Actual Fake

Responsible Machine Learning

Source: NeurIPS

Stakeholders impacted by our work

Stakeholders include media organizations, technology companies, policymakers, cybersecurity professionals, and the general public, all of whom have a vested interest in authenticity & reliability of digital imagery.

Potential negative societal impacts

This project could potentially be exploited to aid in refining and improving the methods used to create convincing deepfakes or misleading visual content, thereby exacerbating issues related to misinformation and digital reception.

Mitigation strategies

Mitigation strategies include developing robust detection tools, implementing strict ethical guidelines for AI usage, promoting public awareness about deepfakes and advocating for policies that regulate the creation and dissemination of AI-generated content.

Limitations

Computation Power

Processing time is limited due to insufficient computational power and no access to GPUs, preventing it from processing the entire dataset

Limited Data Variety

The dataset lack sufficient diversity in terms of different styles and subjects, which can hinder the model's ability to generalize to other images

Model Performance

Model does not perform well when predicting artwork and paintings due to similarities in textual complexities

Manual Data Annotation

This introduces bias which can skew the model's learning process as manual annotation is subjective and inconsistent

Conclusions

Balanced Performance

The final model maintained a good balance between precision and recall for real images. The balance suggest that the model is reliable in identifying real images and effective in minimizing false positives.

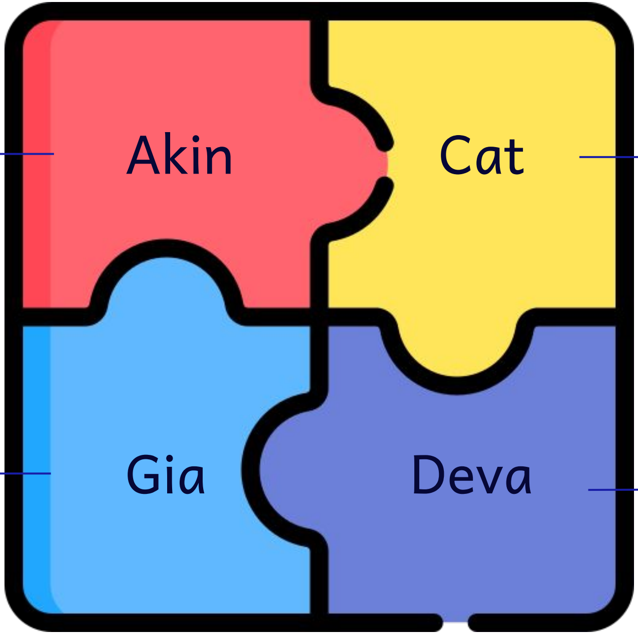
Detection of Fake Images

The final model shows a robust capability in accurately identifying fake images – crucial for use cases that depend on detection of AI-generated content and images.

Future Work

- ➔ Expand computational resources to train and test on the entire dataset
- ➔ Introduce more data varieties for better generalization
- ➔ Experiment with more hyperparameters and models
- ➔ Enhance model’s sensitivity and specificity to further reduce misclassification rates

Contributions



1. Upload files into AWS
2. Preprocessing, EDA, and modeling
3. Setting up batch processing for modeling

1. Set up AWS server
2. Set up S3 bucket
3. Preprocessing & modeling

1. Wrote initial codebase for preprocessing & modeling
2. Data annotation
3. Made slide deck

1. Model experimentation
2. Batch Preprocessing and EDA
3. Model Evaluation and Scoring

References

1. https://www.kaggle.com/datasets/tristanzhang32/ai-generated-images-vs-real-images/data?select=train
2. https://www.youtube.com/watch?v=b1K0CNYxtC0
3. https://www.tensorflow.org/tutorials/load_data/numpy
4. https://aws.amazon.com/blogs/machine-learning/performing-batch-inference-with-tensorflow-serving-in-amazon-sagemaker/
5. https://www.tensorflow.org/guide/data
6. https://techreport.com/statistics/software-web/ai-image-generator-market-statistics/#:~:text=Statistics%20tell%20us%20that%20in,marketers%20preferring%20it%20in%202023
7. https://www.psychologytoday.com/intl/blog/the-art-of-critical-thinking/202311/how-ai-generated-content-can-undermine-your-thinking
8. https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist