AI or Reality?

Using Deep Learning to Classify AI-Generated Images

Rupesh Swaminathan, Sandhya Dasarathy, & Kylie Heagy

Introduction

- AI image generation software is a rapidly developing field
  - Projected growth of 10%/year
  - $20+ billion industry by 2030
- Combines aspects of both machine learning and neural networks
- AI models are becoming better trained, producing more realistic images
  - Introduces issue of how to detect these images

Problem Description

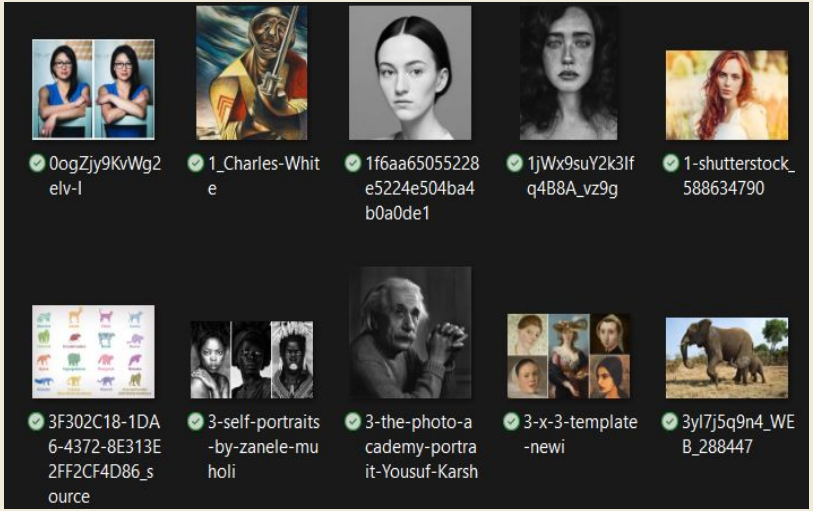
- AI image generation has benefits and drawbacks
  - Benefits: cost effective + saves time for marketing
  - Drawbacks: environmental impact, copyright issues, **misinformation**

- Our aim
  - Build a CNN model that can accurately detect whether an image is AI generated or real
  - Compare with off-the-shelf models

Dataset Description

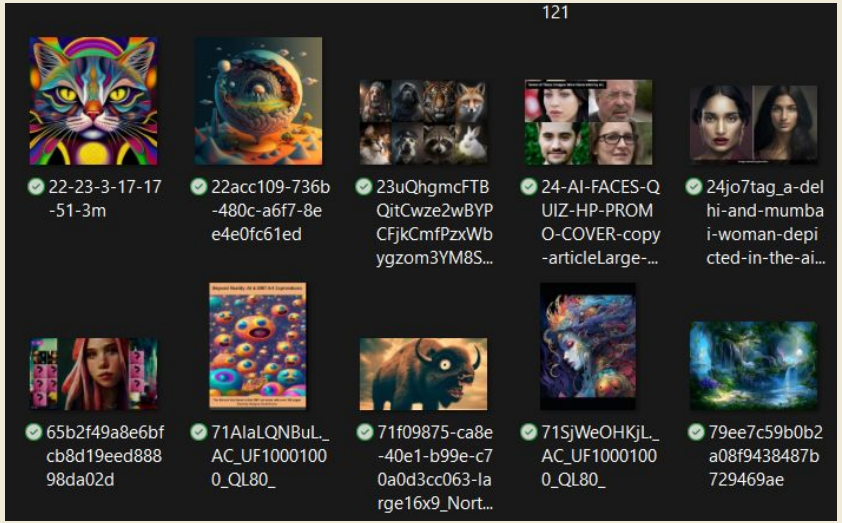
Source: Kaggle dataset — *AI-Generated Images vs Real Images* cashbowman/ai-generated-images-vs-real-images

- **Total images**: 970
  - **RealArt**: 436 photorealistic images from public-domain photo sites
  - **AiArtData**: 539 images generated by various AI models
- **File formats**: .jpg and .png
- **Image resolutions**: Ranging from **256×256** up to **1024×1024 pixel**
  - Ensures model will learn from **visual content**, not trivial cues like image size or metadata
- **Preprocessing**:
  - Resize all images to **128×128 pixel**
  - Pixel values normalized between **[0, 1]**
- **Train/Validation Split**:
  - **Stratified 80/20 split** to preserve original class balance
  - **776 images** for training, **194 images** for validation
  - No overlap between train and validation sets



Real Art Dataset

Ai Art Dataset



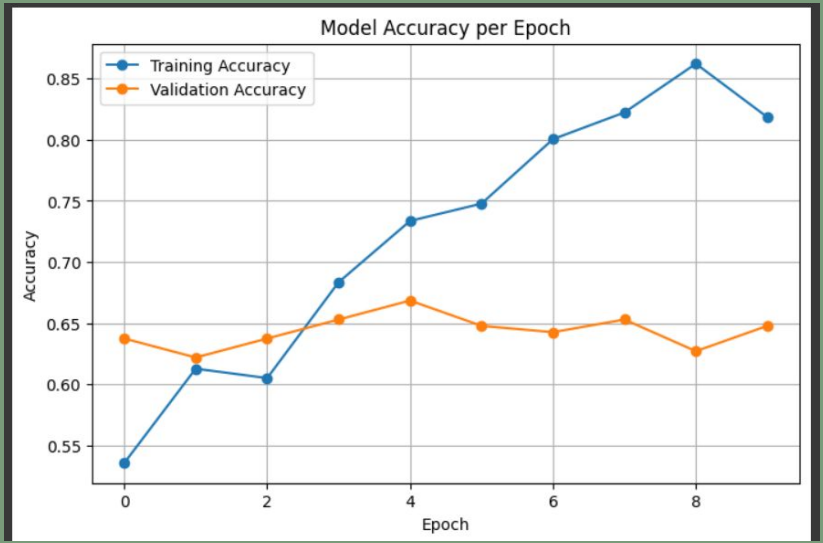
Methodology

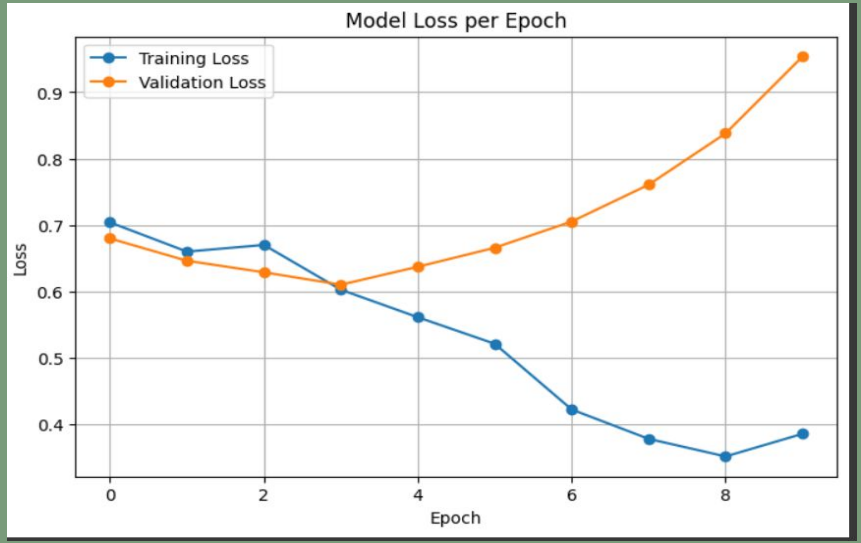
- **Model 1 (CNN with flow_from_directory)**:
  - Data loaded using directory-based generator with internal validation split
  - CNN architecture: 3 × Conv2D + MaxPooling → Flatten → Dense(128) → Dropout(0.5) → Dense(1, sigmoid)
  - Basic augmentation and shorter training (10 epochs)
- **Model 2 (CNN with custom DataFrame and stratified split)**:
  - Data loaded via DataFrame for explicit stratified splitting
  - Stronger augmentations: flip, rotation, zoom, shift (training only)
  - CNN: 3 × Conv2D + MaxPooling → Flatten → Dense(256) → Dropout(0.5) → Dense(1, sigmoid)
  - Trained with callbacks (early stopping, LR reduction) for 30 epochs
- **Transfer Learning**:
  - Experiment with pretrained models like **ResNet** and **DenseNet**
  - Goal: Improve accuracy and reduce overfitting by leveraging learned features
- **Evaluation**: Accuracy, F1-score, confusion matrix, ROC AUC, Accuracy & Loss plots per epoch

Results

Model	accuracy	loss	val_accuracy	val_loss
CNN-1	0.78	0.40	0.64	0.95
CNN-2	0.74	0.53	0.69	0.58
ResNet50	0.59	0.65	0.61	0.66
DenseNet121	0.89	0.28	0.78	0.53

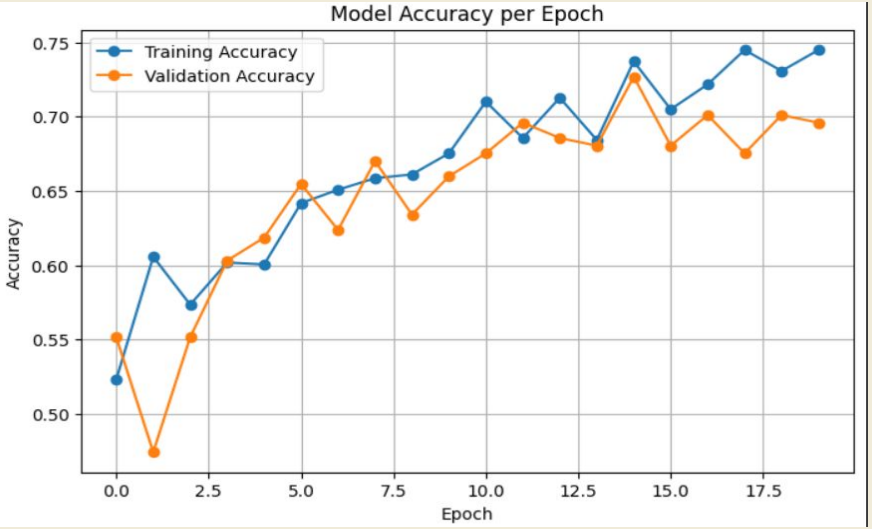
CNN: Model - 1

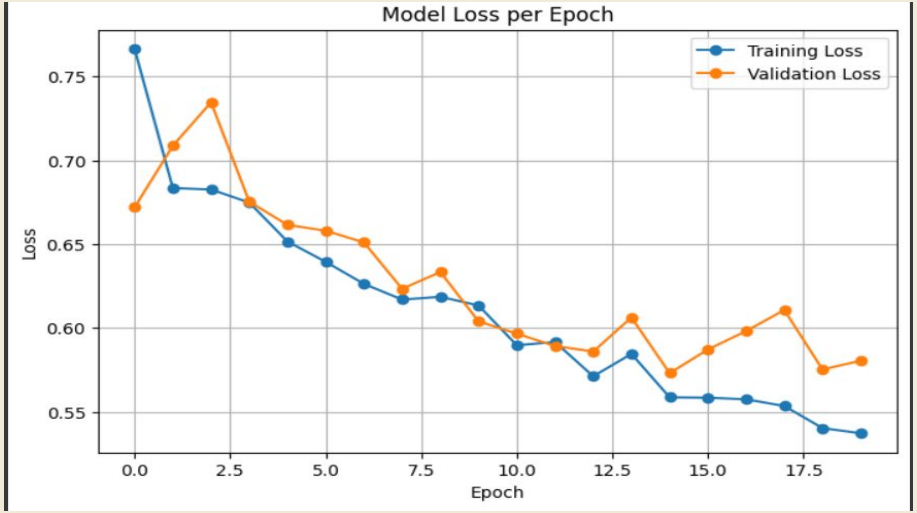




Overfitting
Training accuracy keeps rising; validation accuracy stagnates and even degrades.
Validation loss increases sharply while training loss drops — clear classic overfitting pattern.

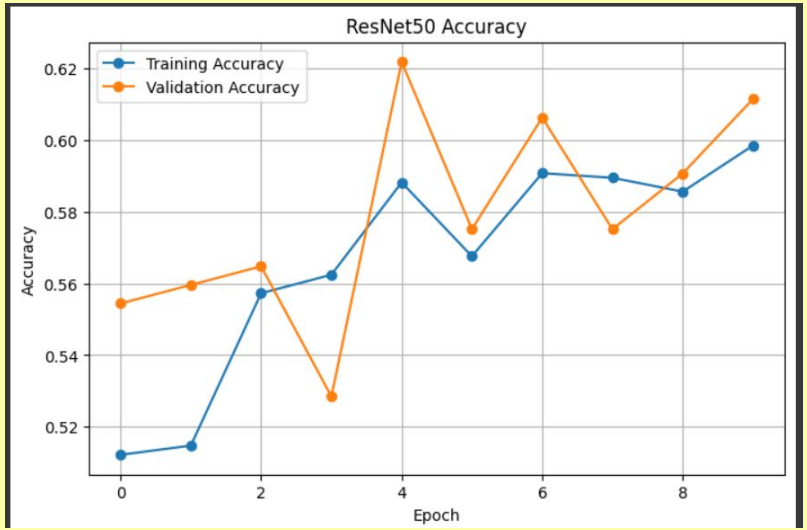
CNN Model - 2

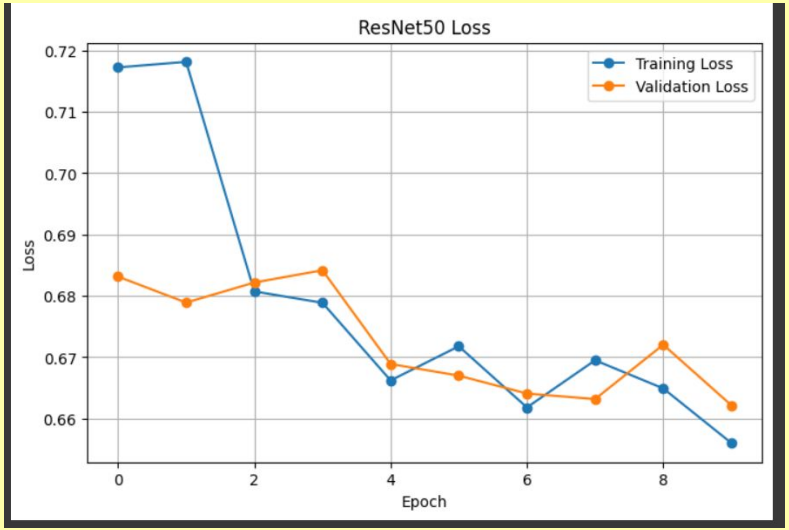




- Overfitting would show training accuracy far outpacing validation accuracy (and training loss well below validation loss).
- In these plots the two curves track closely, with only a small final gap (~0.74 vs. ~0.70 in accuracy and ~0.53 vs. ~0.58 in loss).
- Because training and validation metrics improve in parallel and the final gap is modest, the model is not overfitting.

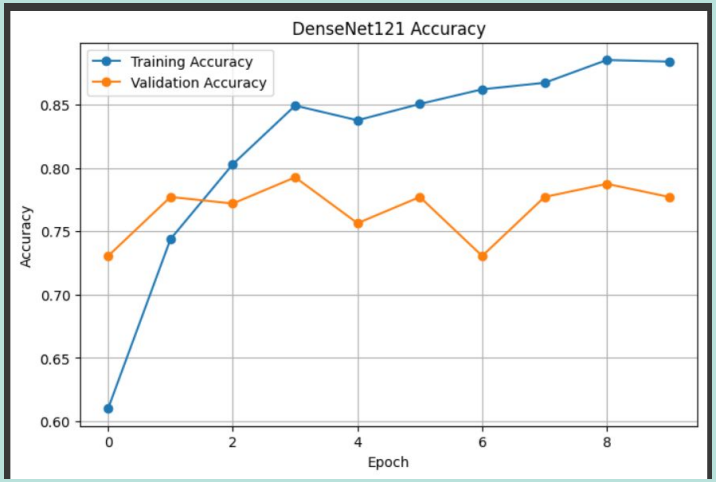
ResNet50

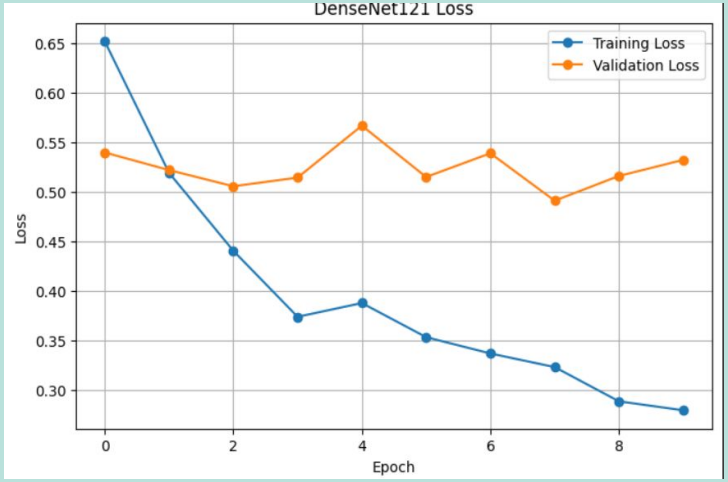




The model achieves similar performance on training and validation sets-indicating good generalization-but overall accuracy is moderate, suggesting capacity or data issues rather than overfitting.

DenseNet121





From graphs, DenseNet121 achieves strong training metrics but shows signs of overfitting, as evidenced by the widening gap between training and validation curves.

Discussion

What's Achieved:

- Our best-performing CNN model achieved ~72% validation accuracy, with an F1-score of 0.71.
- The model performed slightly better on AI-generated images compared to real photos, possibly because AI art tends to have more uniform textures and patterns.
- ROC-AUC around 0.75 indicates moderate discriminative ability.

What could be Improved:

- Add more data for better CNN model training.
- Use higher resolution (224x224) for better feature learning.
- We could probably test the model on social media platforms for practical implications.

thank you.

References

Bowman, C. (2024, February 10). *Ai generated images vs real images*. Kaggle. https://www.kaggle.com/datasets/cashbowman/ai-generated-images-vs-real-images

Jingnan, H. (2024, October 18). *AI-generated images have become a new form of propaganda this election season*. NPR. https://www.npr.org/2024/10/18/nx-s1-5153741/ai-images-hurricanes-disasters-propaganda

NBCUniversal News Group. (2024, May 29). *AI image misinformation has surged, Google Researchers find*. NBCNews.com. https://www.nbcnews.com/tech/tech-news/ai-image-misinformation-surged-google-research-finds-rcna154333

Why is AI image recognition important and how does it work?. Cloudinary. (2025, March 13). https://cloudinary.com/guides/ai/why-is-ai-image-recognition-important-and-how-does-it-work