Implementation and Evaluation of AI-Generated Image Detection Systems

DuckAI

Benjamin Ullrich, Olajide Yusuf, Neha Darawan, Chinmayee Mayekar
Dept. of Electrical and Computer Engineering, Dr. Hao Wang (AAI/CPE/EE 595-B)

Introduction

With generative AI becoming increasingly widespread, models such as ChatGPT, DeepSeek, Sora, and Imagen have the potential to produce unprecedented amounts of misinformation. To protect individuals from scams, fraud, copyright violations, defamation, and other harms, effective methods for detecting AI-generated content are essential. This project narrows its focus to AI-generated images and will survey and implement existing detection techniques, followed by proposing improvements or combined approaches to achieve stronger detection performance.

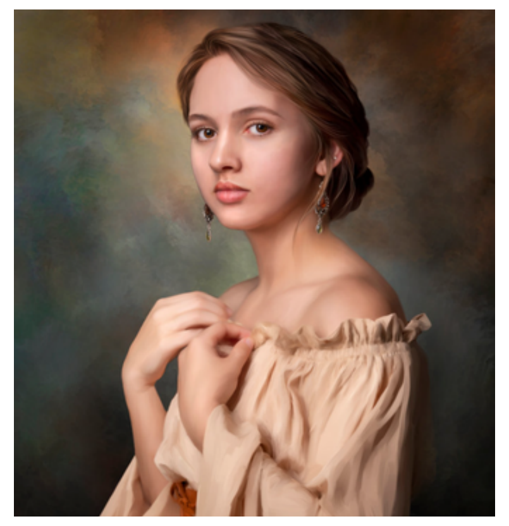
Methodology

- Collected and preprocessed AI-generated and real images.
- Fine-tuned ResNet152V2 for deep-learning classification.
- Trained Logistic Regression using VGG16 feature extraction.
- Applied Random Forest (HOG + color histograms) and KNN (flattened pixels).
- Evaluated all models using accuracy, reports, and confusion matrices.

Data

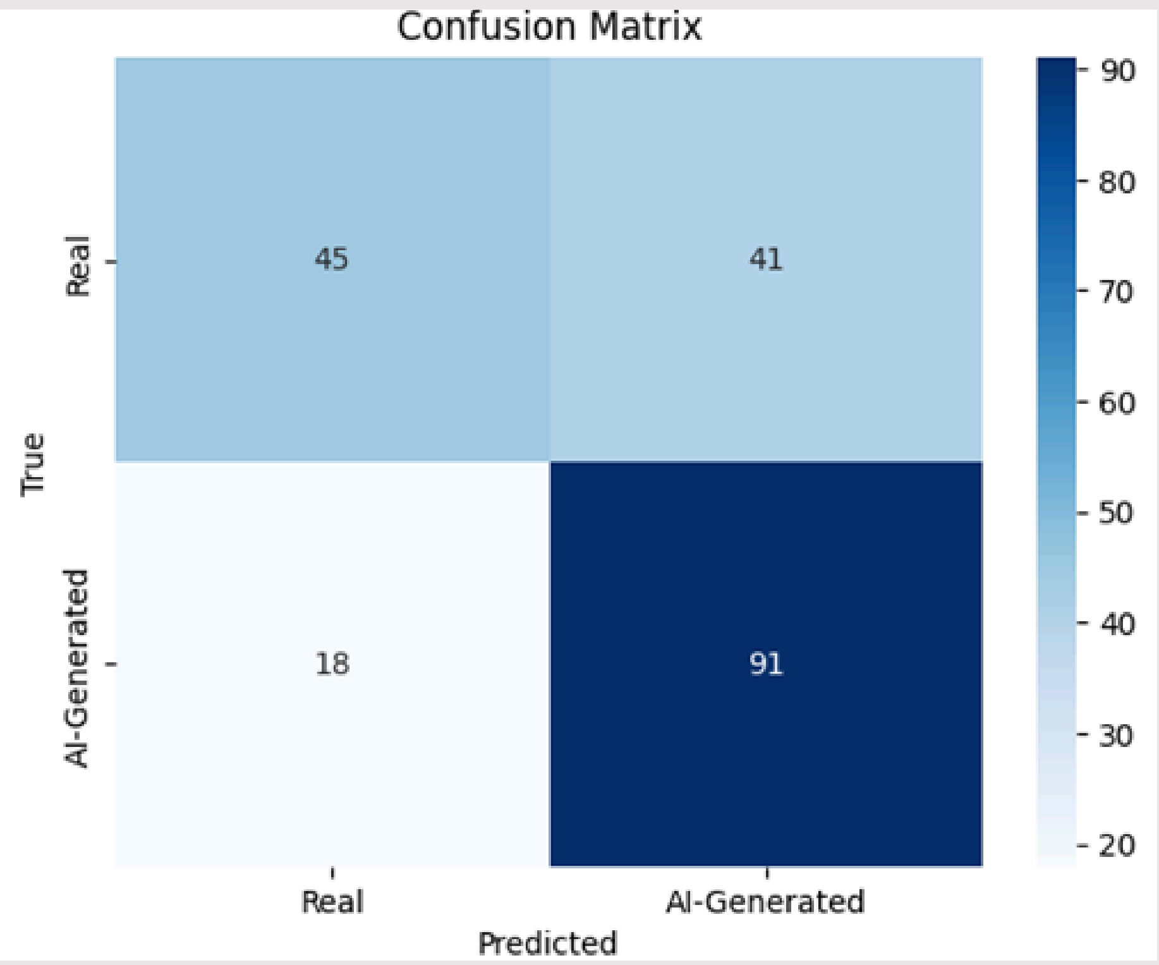
Initial Dataset:
To establish which model types are effective for this type of task, we used a relatively small web-scraped dataset of 1,000 images (obtained from Kaggle)

AI
←
Real
→

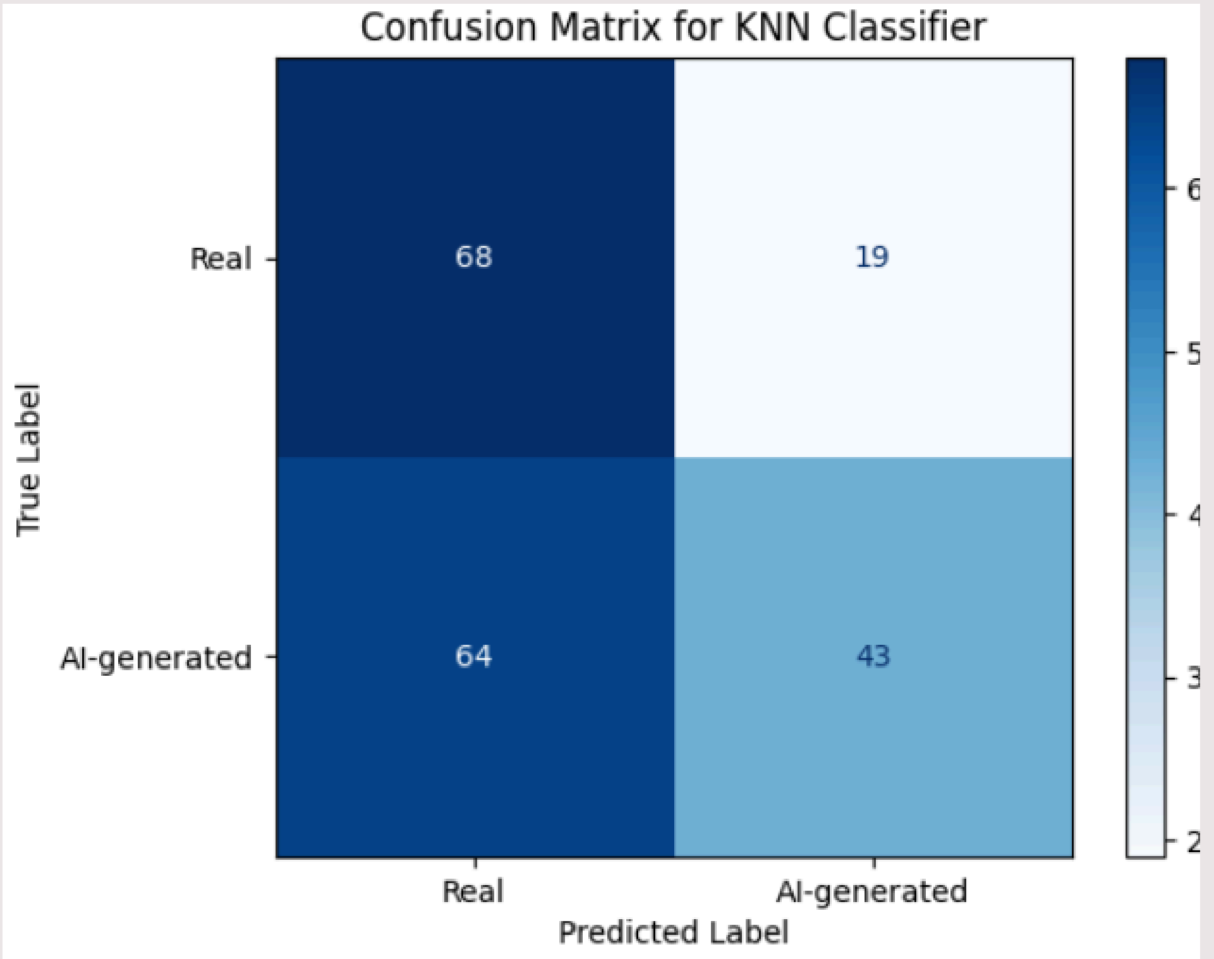


Improved Dataset:
Once different approaches were tested, the full-scale model was trained and tested on a significantly larger, more recent dataset (approx. 80k images) with human-generated content from Shutterstock and AI-generated content from DeepMedia.

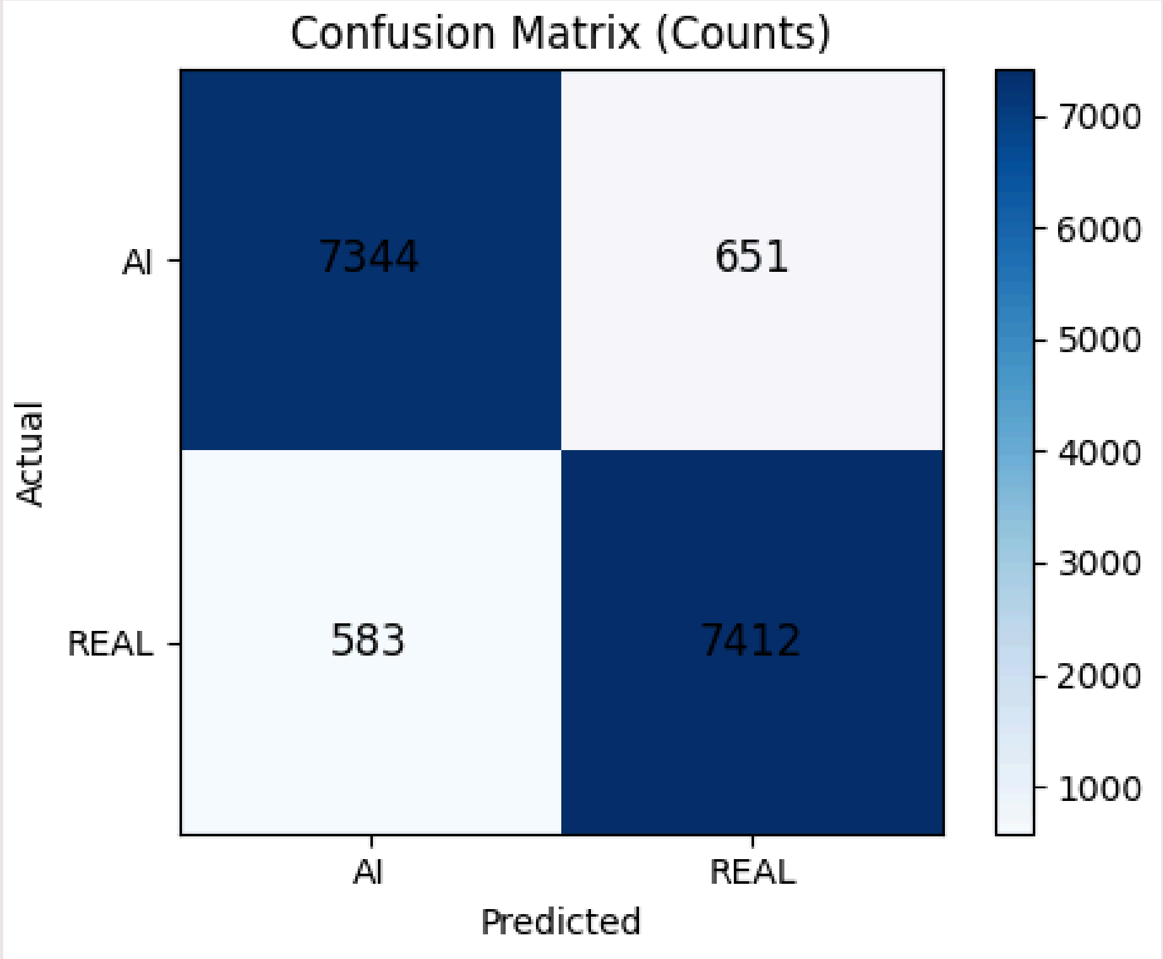
AI
←
Real
→



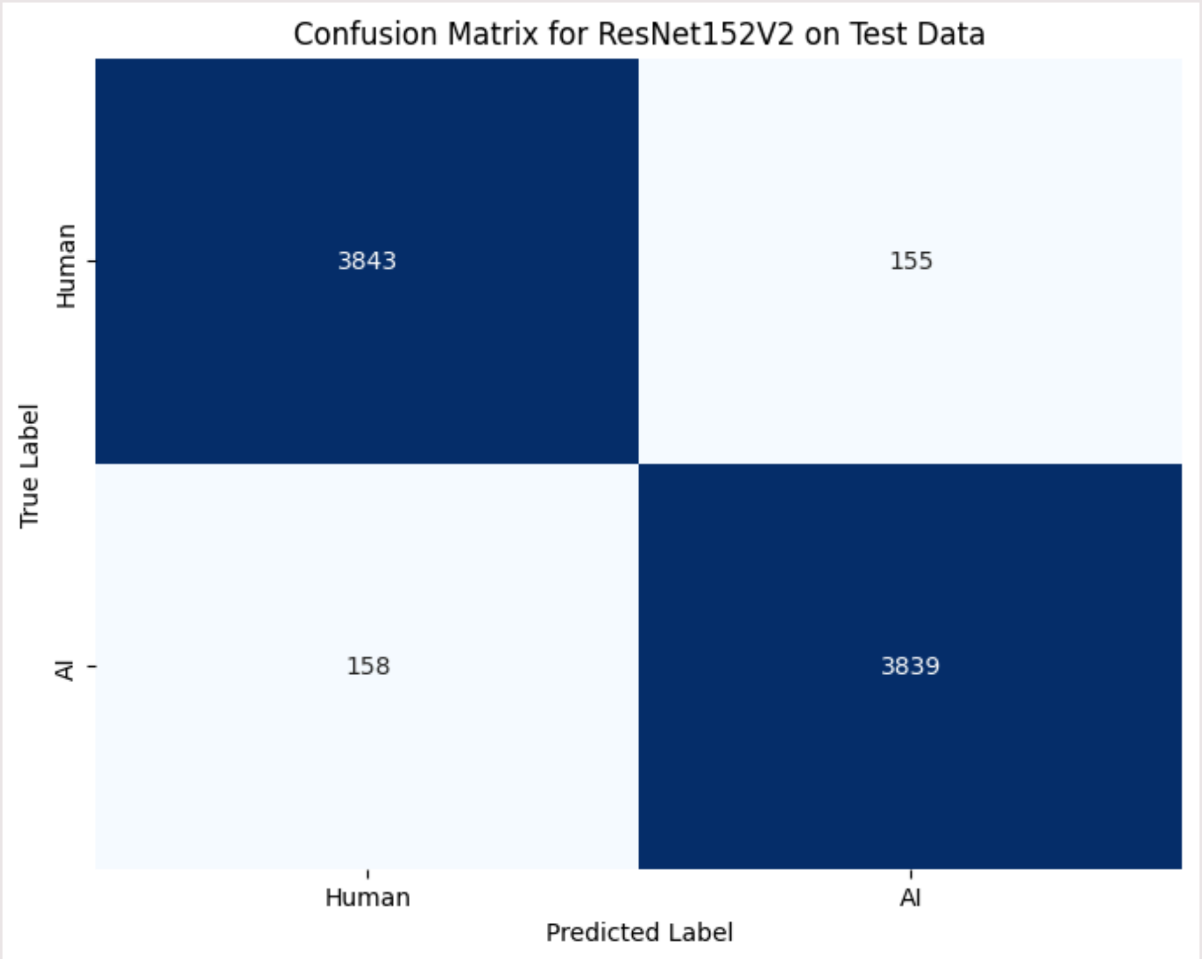
Random Forest



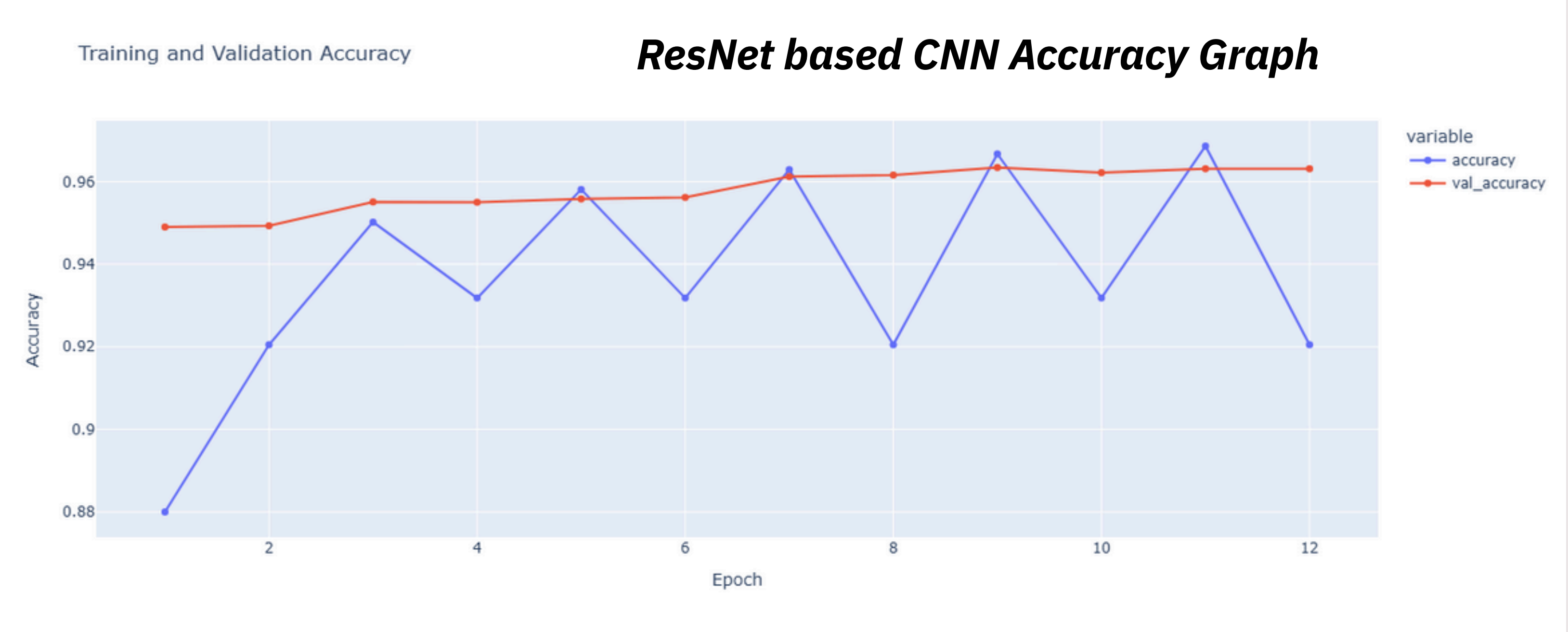
KNN



Logistic Regression



ResNet based CNN



ResNet based CNN Accuracy Graph

Results

Models	***Accuracy with the old dataset***	***Accuracy with the new dataset***
CNN	***72.00%***	***95%***
Logistic Regression	***49.7%***	***92.3%***
KNN	***58%***	***47%***
Random Forest	***46%***	***50.2%***

Conclusion

Ultimately, the ResNet152V2 CNN-based approach with added fine-tuning was able to produce 95.97% accuracy on the testing data, with minimal overfitting. Since CNN architectures are especially well-suited for image data, this result is expected, and we also observed that other, more lightweight models were not good enough for this image-detection task. Logistic regression built on top of CNN features did give us quite good accuracy, but still could not match the fully fine-tuned CNN. As generative models continue improving, it is possible that automated detection will become infeasible without introducing explicit fingerprinting capabilities into generative models.

STEVENS
INSTITUTE OF TECHNOLOGY