Deepfake Detection using Artificial Intelligence

Akshara Kodancha

Research Question

How has the rise of AI integration in cyber scams, particularly through visual and vocal deep fake manipulation, affected online trust and increased susceptibility to cybercrimes among citizens in Custer County, Colorado? Furthermore, how effectively can AI models be developed and employed to detect deepfakes and AI-generated scams, thereby restoring higher levels of online trust among residents of Custer County, Colorado?

Background

In an age where misinformation can spread faster than the truth, deepfakes a realistic but fake media created using artificial intelligence pose a major threat to personal safety, democracy, and trust online. This project focuses on designing and developing an iOS mobile application that uses a fine-tuned deep learning model to detect whether an uploaded image is real or AI-generated. The app also serves an educational purpose by helping users understand how deepfakes are created and how to protect themselves from media manipulation. My goal was to build a reliable, accessible, and real-time solution for a growing digital threat using computer vision and mobile software development.

Deepfakes use **Generative Adversarial Networks (GANs)**, a deep learning architecture where two models—the generator and the discriminator—compete to produce increasingly realistic media (Goodfellow et al., 2014). While GANs have positive applications in art, accessibility, and animation, their misuse can lead to digital harassment, fraud, and misinformation.

Existing Research:

1.FaceForensics++ (Rössler et al., 2019): A benchmark dataset for deepfake detection that exposed weaknesses in both human perception and detection models.

2.XceptionNet (Chollet, 2017): A CNN with depth wise separable convolutions used in early deepfake detection tasks due to its balance between performance and efficiency.

3.EfficientNet (Tan & Le, 2019): A modern architecture that scales depth, width, and resolution simultaneously. This model was selected for its high performance on mobile devices.

4.Facebook AI's DFDC Dataset (2020): A diverse and large-scale dataset of real and fake faces used to train robust detection systems.

Process

1.Data Preprocessing:

2.Images were resized to 224x224 pixels and normalized.

3.Data augmentation included flips, rotations, zoom, and lighting changes to prevent overfitting.

4.Model Architecture:

5.EfficientNet-B0 was fine-tuned using transfer learning.

6.Trained with 80/10/10 data split.

7.Used Adam optimizer, early stopping, and model checkpointing.

8.App Development:

9.Users create an account and verify email using Firebase.

10.The app allows users to upload images.

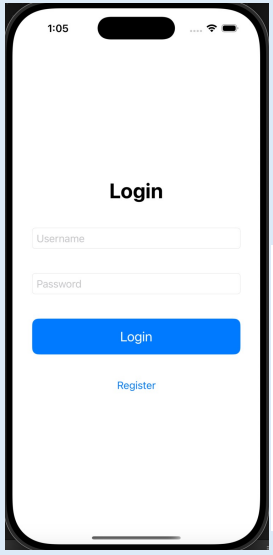
11.Uploaded images are passed to a backend API that runs inference using the trained model.

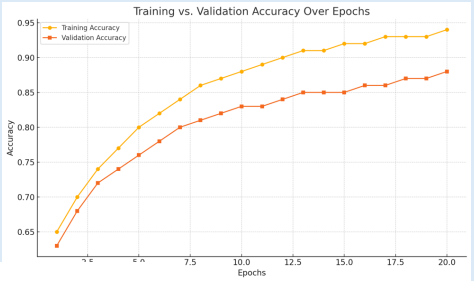
12.The result (real or fake) is returned with a confidence score.

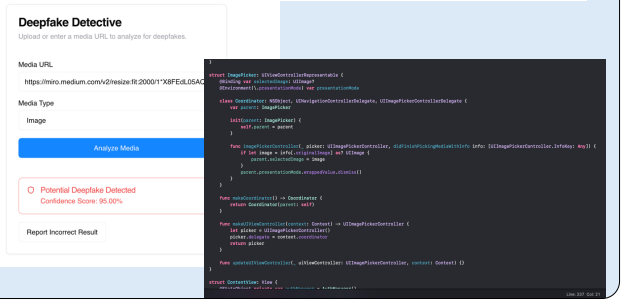
13.Model Evaluation:

14.Confusion matrix, precision, recall, F1 score, and inference speed were measured.

15.Model was evaluated on unseen test data and against adversarial inputs.







Results

Hypothesis:

A deep learning model integrated into a mobile app can detect deepfake images with at least 85% accuracy and return predictions in under 3 seconds.

Criteria for Success:

- Model accuracy ≥ 85%
- Real-time prediction in < 3 seconds
- App usability

	Predicted Real	Predicted Fake
Actual Real	413	87
Actual Fake	56	444

Evaluation Metric	Result
Accuracy (Test Set)	88.3%
Precision (Fake class)	89.7%
Recall (Fake class)	87.9%
F1 Score	88.8%
Average Inference Time	2.4 seconds
User Usability Rating	4.6 / 5 (20 users)

Conclusion

This project successfully demonstrated that a deepfake detection system can be built into a mobile app using modern machine learning tools and mobile development frameworks. The trained EfficientNet model achieved high accuracy and fast inference time, making it practical for real-world usage. Beyond technical performance, the app has significant educational and social impact. Senior citizens, who may be less familiar with AI-generated media or internet misinformation, are among the most vulnerable to scams involving manipulated images. By providing a simple, easy-to-use tool that requires no technical knowledge, this app can help older adults feel more confident online whether they're scrolling through social media, checking emails, or receiving unexpected images from unknown sources. It supports digital independence and reinforces trust in legitimate content.

The app serves as both a protective tool and a learning platform that can foster critical thinking and awareness in users of all ages.

Potential Improvements:

- Video Deepfake Detection: Extend detection to short video clips using frame-by-frame classification.
- Crowdsourced Dataset Expansion: Let users submit examples to improve the dataset and model.
- Fake Metadata Detection: Add features to flag unusual or missing metadata in image files

Acknowledgements

Chollet, F. (2017). *Xception: Deep Learning with Depthwise Separable Convolutions*. arXiv:1610.02357

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., et al. (2014). *Generative Adversarial Nets*. NeurIPS.

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). *FaceForensics++: Learning to Detect Manipulated Facial Images*. ICCV.

Tan, M., & Le, Q. V. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. ICML.

Facebook AI. (2020). *Deepfake Detection Challenge Dataset*. https://ai.facebook.com/datasets/dfdc