University of Michigan-Dearborn

CIS 583 – Deep Learning

Deepfake Detection System: A Comparative Analysis of ResNet-18 and Xception for Image-Based Classification

Authors: Bhavika Solao, Reshma Murali, Aniket Khedkar, Tega Ejughemre **Affiliation:** University of Michigan-Dearborn, CIS 583 – Deep Learning

Abstract

The proliferation of generative models, particularly Generative Adversarial Networks (GANs) and diffusion networks, has democratized the synthesis of hyper-realistic facial imagery. While these advancements foster creativity, they pose significant societal risks, including misinformation and identity theft. This study addresses the critical need for reliable automated detection tools by evaluating two Convolutional Neural Network (CNN) architectures: ResNet-18 and Xception. Unlike video-based approaches that utilize temporal cues, this research focuses exclusively on single-image classification. The models were trained and tested on a dataset of real and GAN-generated faces using a rigorous preprocessing pipeline. Experimental results indicate that ResNet-18 achieves a superior accuracy of 97.33% with stable convergence, whereas Xception, despite its theoretical advantages, underperformed with approximately 65% accuracy due to computational and dataset constraints. This paper discusses the trade-offs between architectural complexity and

resource availability, offering insights for deploying deepfake detectors in resource-constrained environments.

1. Introduction

The digital media landscape has been transformed by the rise of deep generative models. Techniques such as GANs and diffusion networks have advanced to the point where synthesized facial images—commonly termed "deepfakes"—are frequently indistinguishable from authentic photographs by human observers. While these technologies have legitimate applications in entertainment and art, their potential for misuse in political manipulation, fraud, and the erosion of trust in digital media is a growing concern.

As these threats escalate, the development of robust computational detection methods has become imperative. This project investigates the efficacy of deep learning in identifying manipulated facial images. The scope of this work is restricted to image-based detection, intentionally excluding video frame sequencing to focus on spatial artifacts and texture irregularities.

The primary objective is to compare two distinct CNN architectures: ResNet-18, known for its efficiency, and Xception, renowned for capturing fine-grained features. By evaluating these models on a balanced dataset of real and fake images, this study aims to determine which architecture offers the optimal balance of performance and efficiency under limited computational resources.

2. Related Work

Deepfake detection has evolved rapidly in parallel with generative techniques. This section reviews pivotal studies that influenced the design of this project.

A. Artifact Analysis Guarnera et al. (2023) focused on the residual artifacts inherent in synthetic images. Their analysis revealed that GAN and diffusion-generated images possess unique statistical noise distributions that differ from real photographs, even when visually convincing. They demonstrated that CNNs trained to detect these inconsistencies can achieve high detection rates.

B. Robustness and Compression Real-world deployment often involves image compression. Mo et al. (2023) investigated the resilience of detectors against JPEG compression, which can mask manipulation cues. Their findings suggest that complex architectures like Xception exhibit greater resilience to compression artifacts compared to simpler networks.

C. Facial Region Attention Xue et al. (2023) proposed a global-local facial fusion network. Their research indicates that manipulation traces are often concentrated in specific local regions, such as the eyes and mouth. Incorporating both holistic facial structure and local details significantly enhances accuracy.

D. Cross-Generator Generalization Ricker et al. (2024) highlighted the "generalization gap" where detectors trained on GAN images fail to detect diffusion-based deepfakes. This underscores the dependency of current detectors on the specific generative method used for training.

E. Reconstruction Approaches Liu et al. (2023) and Sarkar et al. (2025) explored reconstruction-based methods. Liu et al. utilized masked region reconstruction, finding that real images are reconstructed with lower error rates than fakes. Similarly, Sarkar et al. proposed an unsupervised approach using Singular Value Decomposition (SVD), identifying fakes as statistical anomalies without requiring labeled fake samples.

3. Dataset and Preprocessing

A. Dataset Description

The study utilized the "Deepfake and Real Images" dataset from Kaggle, partitioned into Train, Validation, and Test subsets. The data comprises authentic human faces and GAN-generated synthetic faces. The dataset is characterized by high variability in lighting, pose, and age, ensuring it represents diverse real-world scenarios. Labels were inferred automatically from the directory structure (Real vs. Fake).

B. Data Cleaning

To ensure training stability, a strict cleaning protocol was implemented. A verification script iterated through the dataset to identify and remove corrupted files that failed to open or contained incomplete pixel data, preventing runtime crashes.

C. Preprocessing Pipeline

1. **Channel Standardization:** All images were converted to RGB format to handle inconsistencies where some source files were grayscale or contained alpha channels, ensuring compatibility with standard CNN input requirements.
2. **Resizing:** Images were resized to pixels. This reduction from the original resolution was necessary to mitigate GPU memory limitations and speed up processing.
3. **Normalization:** Pixel values were normalized to a stable numerical range to facilitate faster convergence during gradient descent.
4. **Augmentation:** To combat overfitting, the training pipeline included random geometric transformations, including horizontal flips, rotations, and color jittering.
5. **Batching:** PyTorch DataLoaders were employed to manage memory efficiently, utilizing a batch size of 32

4. Methodology

A. Architecture 1: ResNet-18

ResNet-18 is a residual learning framework designed to enable the training of deep networks by utilizing skip connections.

- **Mechanism:** It mitigates the vanishing gradient problem, allowing gradients to flow through the network effectively.
- **Configuration:** The model was initialized with ImageNet pretrained weights to leverage transfer learning. The final fully connected layer was modified to output two classes (Real/Fake).
- **Optimization:** Training utilized the Adam optimizer and Cross-Entropy Loss.

B. Architecture 2: Xception

Xception stands for "Extreme Inception" and relies on depthwise separable convolutions.

- **Mechanism:** This architecture decouples cross-channel correlations from spatial correlations, theoretically allowing it to capture fine-grained texture irregularities common in deepfakes.
- **Configuration:** The model was trained from scratch. It typically requires an input size of , but was adapted to the smaller preprocessed size due to resource constraints.

5. Experimental Setup

The experiments were conducted using the Google Colab environment, leveraging GPU acceleration for training.

- **Framework:** PyTorch.
- **Hyperparameters:**
  - Batch Size: 32.
  - Epochs: 10–20, varying based on convergence speed and resource availability.
  - Learning Rate: Optimized for Adam.
- **Metrics:** Performance was evaluated using Accuracy, Training Loss, Validation Loss, and Confusion Matrices

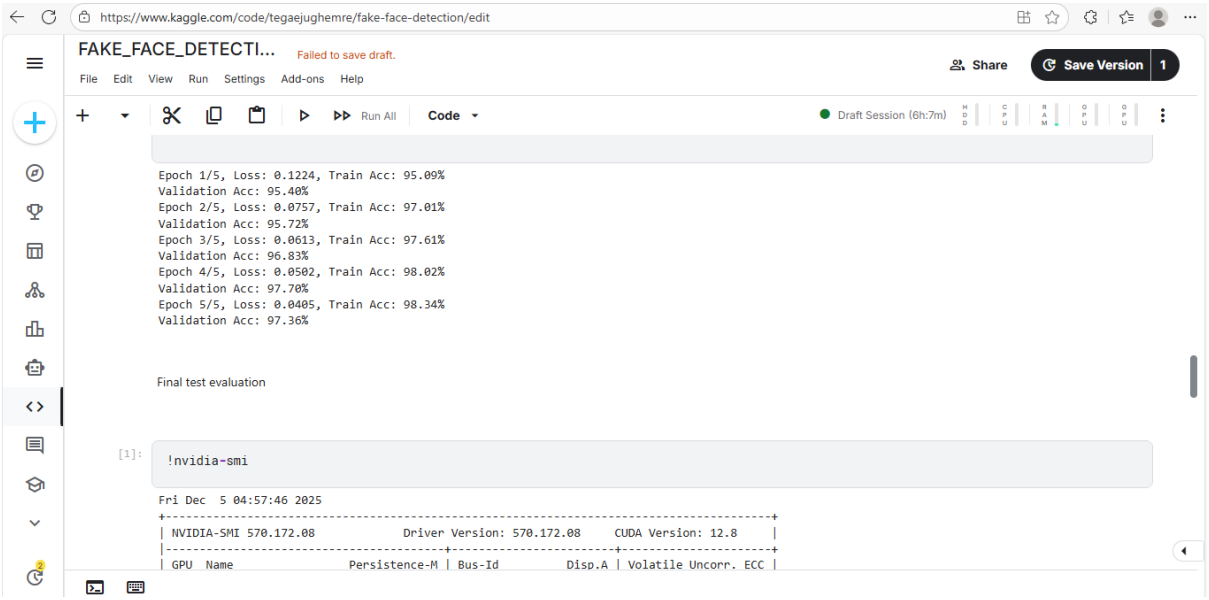
6. Results

A. ResNet-18 Performance

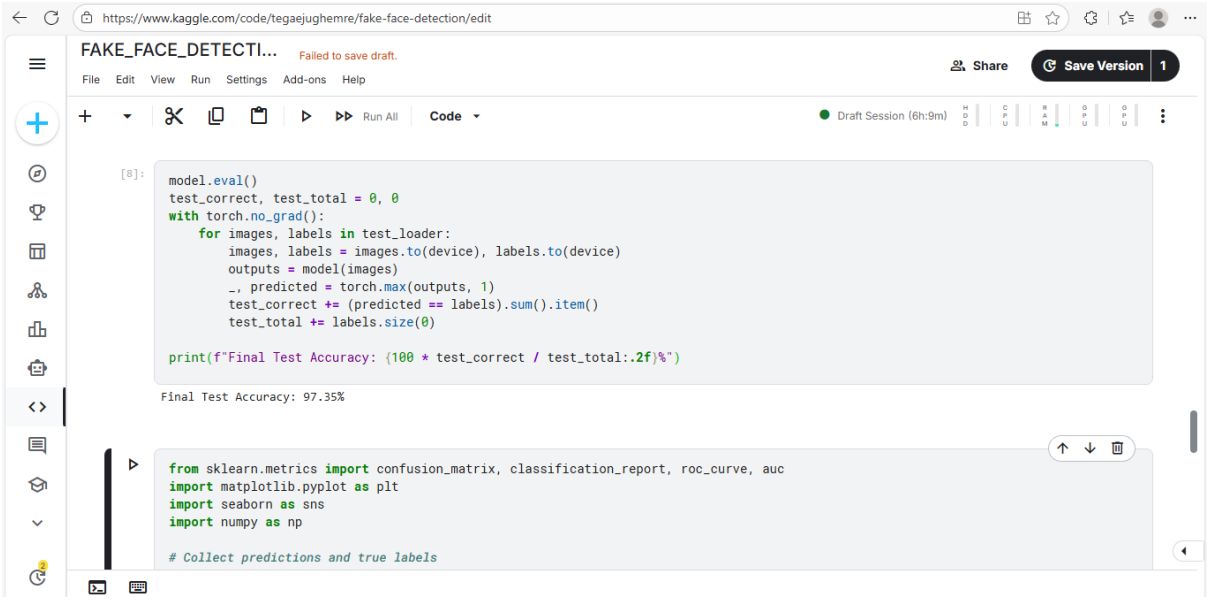
ResNet-18 demonstrated high efficacy, achieving a final test accuracy of **97.33%**. The training process was stable, with validation accuracy consistently tracking training accuracy, indicating minimal overfitting.

- **Confusion Matrix Analysis:** The model exhibited a balanced detection capability, with very few false positives or false negatives

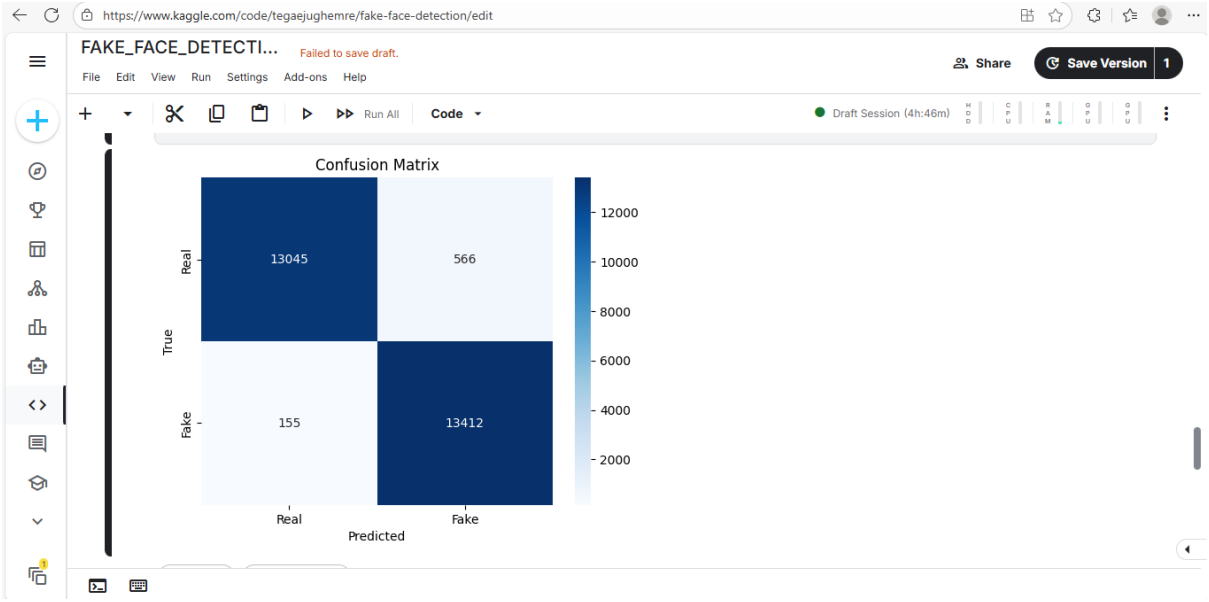
TRAINING FLOW



MODEL EVALUATION (FINAL ACCURACY)

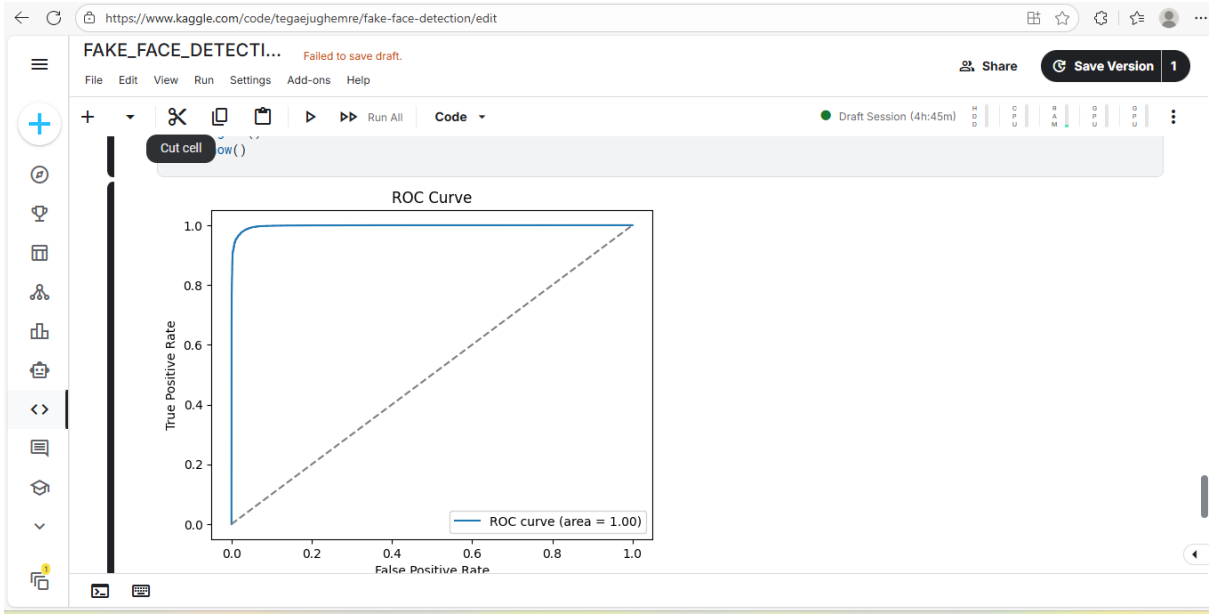


EVALUATION (CONFUSION MATRIX)



ROC Curve: The Receiver Operating Characteristic (ROC) curve showed an Area Under the Curve (AUC) of 1.00, confirming near-perfect separation between classes on the test set

EVALUATION (ROC-CURVE)



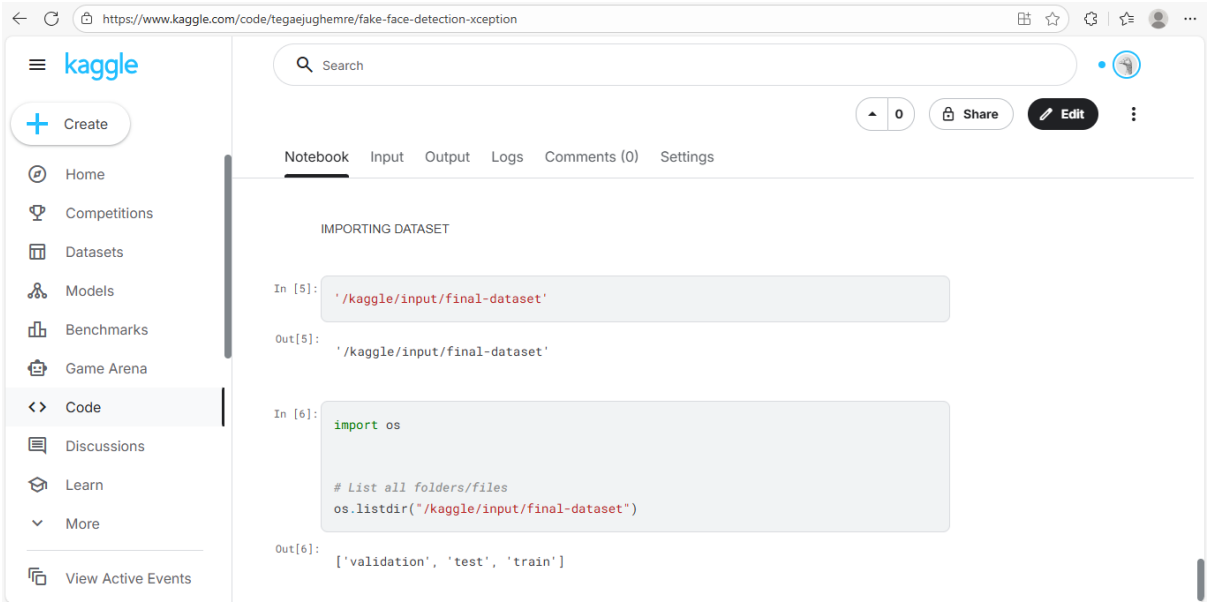
B. Xception Performance

In contrast, the Xception model achieved approximately **65%** accuracy.

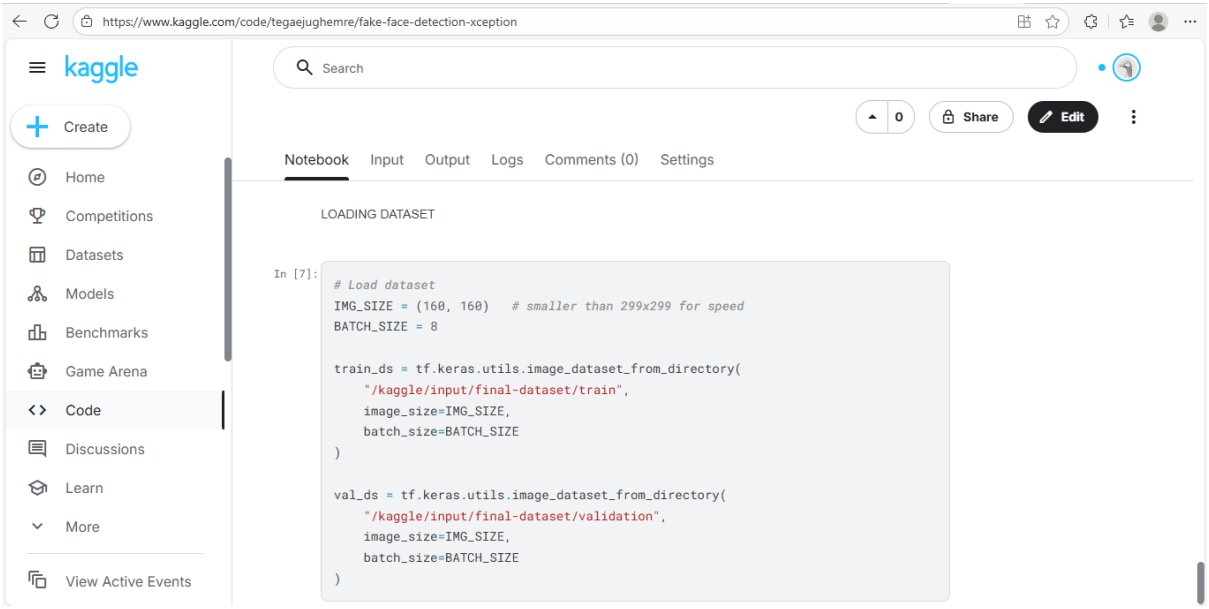
- **Instability:** The loss curves were unstable, suggesting the model struggled to converge

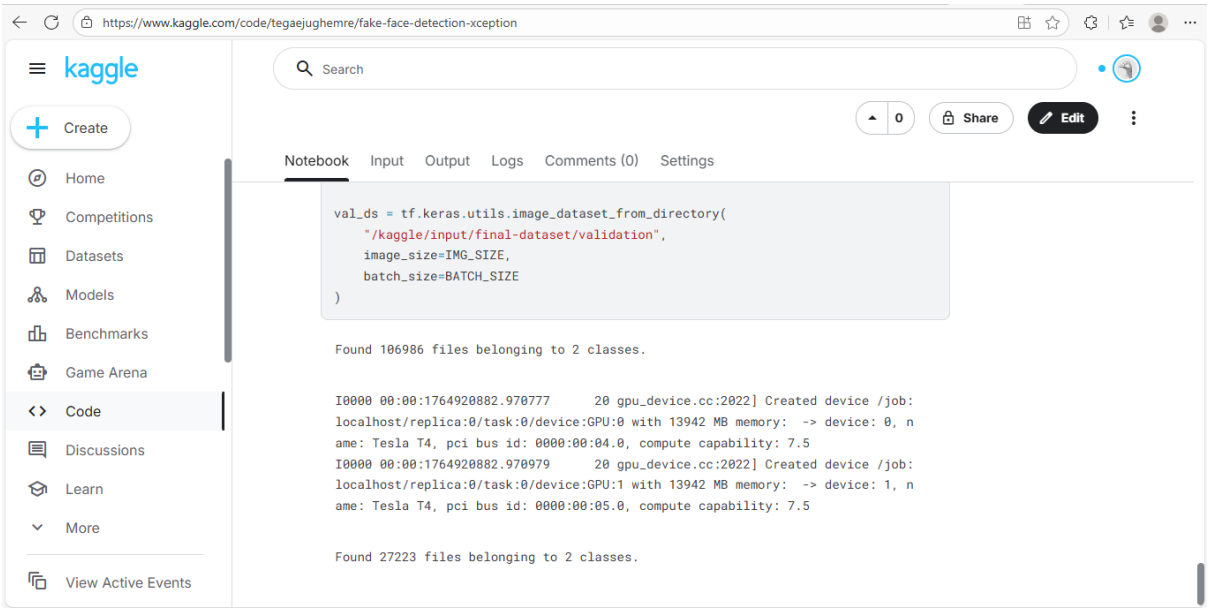
Resource Constraints: Xception requires significantly more data and computational power to train from scratch. The reduction in input size and dataset subset likely severely hampered its ability to learn robust features

SETTING UP THE ENVIRONMENT AND IMPORTING THE DATASET

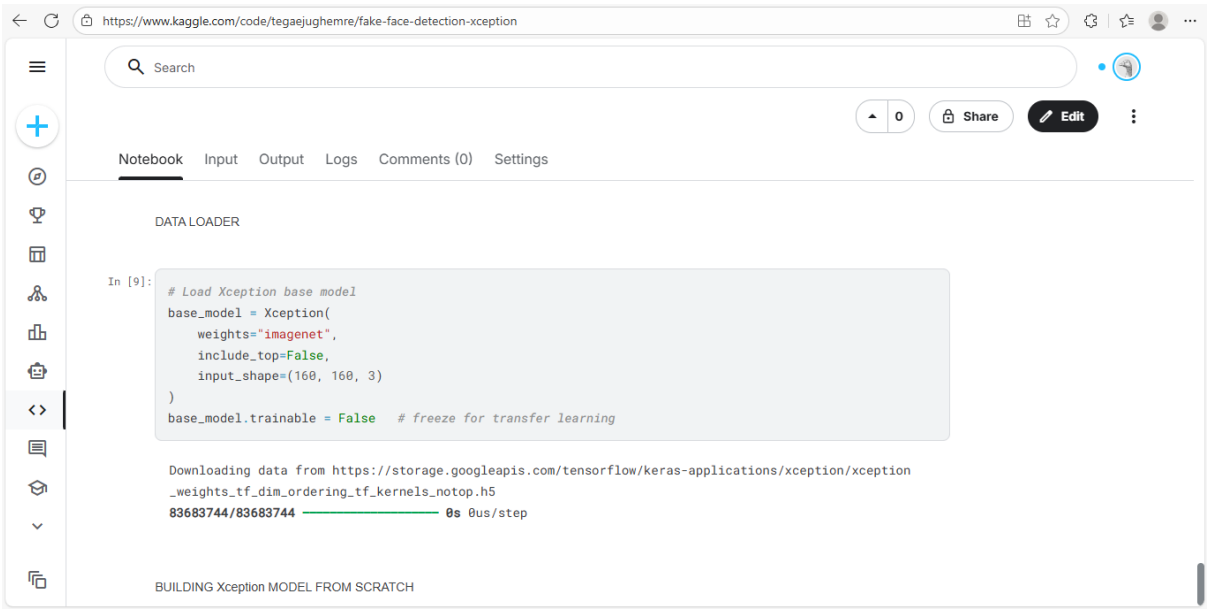


LOADING DATASET





MODEL



7. Discussion

The comparison reveals a clear trade-off between theoretical capability and practical deployability.

1. **Resource Efficiency:** ResNet-18 is superior for scenarios with limited compute. Its use of pretrained weights allowed it to generalize well despite the moderate dataset size.
2. **Complexity vs. Data:** Xception is architecturally stronger for texture analysis but failed to perform because it demands a larger training corpus to exploit its depthwise separable convolutions.
3. **Challenges:**
    - **Generalization:** Models trained on specific GANs may struggle with unseen deepfakes (e.g., Diffusion models).
    - **Data Quality:** Real-world data often suffers from compression, which degrades detection accuracy.
    - **Interpretability:** Both models function as "black boxes," lacking transparency in their decision-making process.

8. Future Directions

Future research should focus on enhancing robustness and explainability:

- **Cross-Model Generalization:** Incorporating diffusion-generated images into the training set to handle modern deepfakes.
- **Attention Mechanisms:** Integrating attention layers to focus the network on manipulated regions (eyes, mouth) automatically.
- **Mobile Optimization:** Developing lightweight models for real-time detection on mobile devices.
- **Explainability:** Implementing tools to visualize *where* the model detects manipulation, increasing user trust.

9. Conclusion

This project successfully implemented and evaluated deep learning models for deepfake detection. ResNet-18 emerged as the optimal solution for the given constraints, achieving 97.33% accuracy. While Xception underperformed, the results highlight the critical importance of aligning model complexity with available data and computational resources. As generative techniques evolve, continuous adaptation of these detection systems is essential to maintain digital trust.

10. References

[1] L. Guarnera, O. Giudice, and S. Battiato, “Level Up the Deepfake Detection: Discriminating Images Generated by GAN and Diffusion Models,” 2023. [2] H. Mo, F. Chen, and X. Lyu, “Fake Faces Under JPEG Compression: A Study on Deepfake Detection,” 2023. [3] Z. Xue, X. Jiang, Q. Liu, and Z. Wei, “Global–Local Facial Fusion for Deepfake Detection,” 2023. [4] J. Ricker, S. Damm, T. Holz, and A. Fischer, “Towards the Detection of Diffusion Model Deepfakes,” 2024. [5] C. Liu, T. Zhu, S. Shen, and W. Zhou, “Robust

Deepfake Detection Using Multi-View Reconstruction," 2023. [6] S. Sarkar, R. Bora, S. George, and K. Raja, "Unsupervised Deepfake Detection Using SVD-Based Reconstruction," 2025.