

MesoNet Performance Analysis

Deepfakes Dataset

Dataset Information	
Total Videos Tested	2,000 (1,000 real + 1,000 fake)
Model Architecture	MesoNet Meso4
Weights File	Meso4_DF.h5
Analysis Date	December 28, 2025

Introduction

MesoNet is a convolutional neural network specifically designed to detect face tampering in videos, focusing on mesoscopic properties that are at an intermediate level between fine-grained (micro) and global (macro) features. This report presents a comprehensive analysis of the Meso4 model's performance on the Deepfakes dataset, which consists of 1,000 original videos and 1,000 manipulated videos. The analysis includes detailed metrics, visualizations, and actionable recommendations for improving detection accuracy.

Quick Summary

Correctly Classified	1,523 videos	Incorrectly Classified	477 videos
Overall Accuracy	76.1%	AUC Score	0.85

Visual Results

The following visualizations provide a comprehensive view of the model's performance, including the confusion matrix, key metrics, prediction distributions, confidence scores, and the ROC curve for discriminative ability assessment.

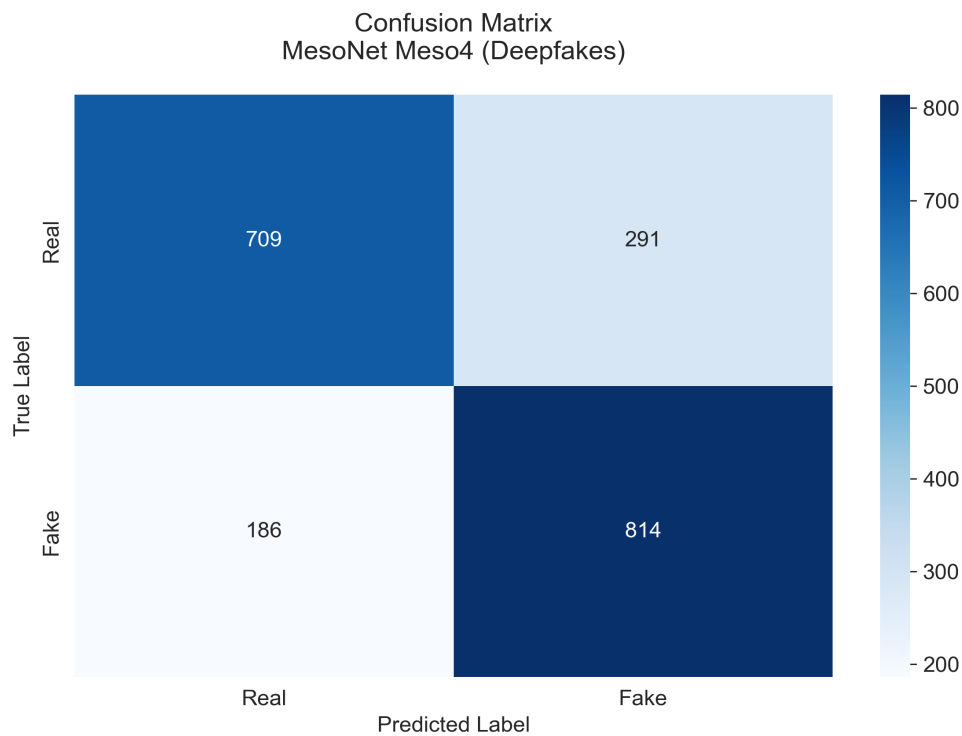


Figure 1: Confusion Matrix - Shows the distribution of correct and incorrect predictions.

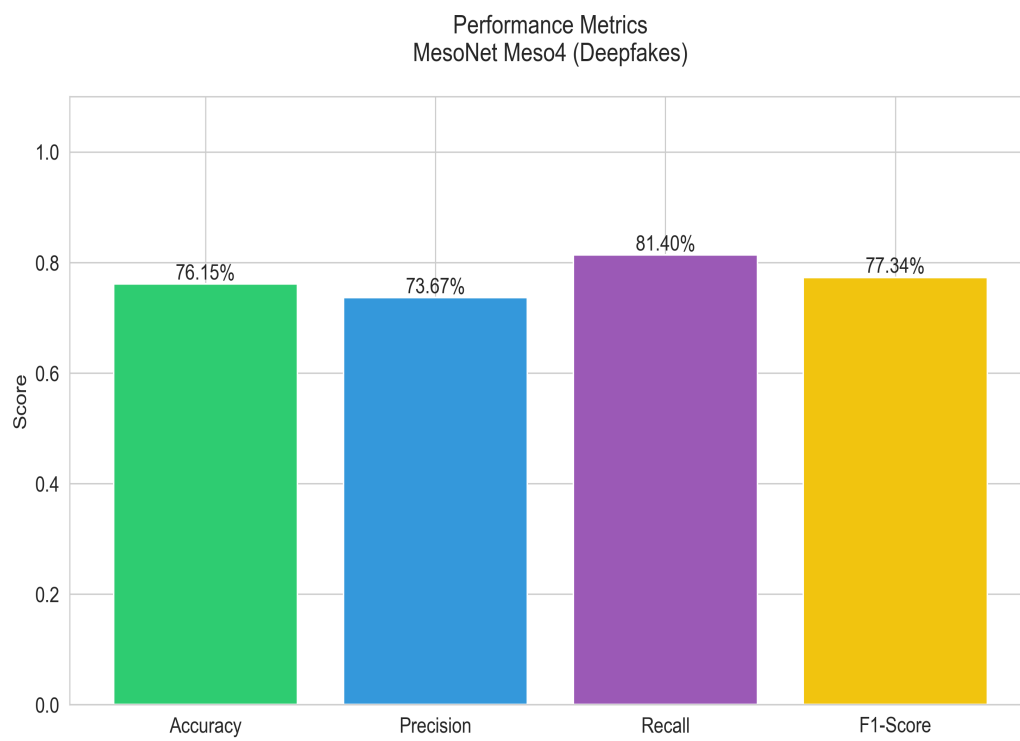


Figure 2: Performance Metrics - Accuracy, Precision, Recall, and F1-Score.

Visual Results (Continued)

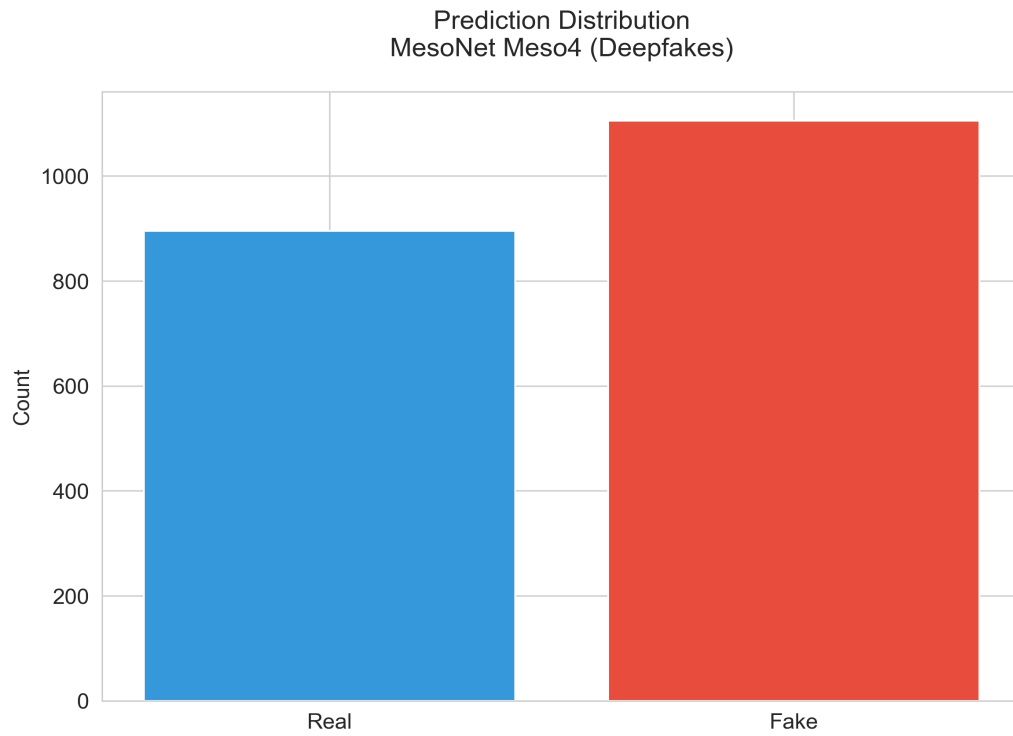


Figure 3: Prediction Distribution - Total count of Real vs Fake predictions.

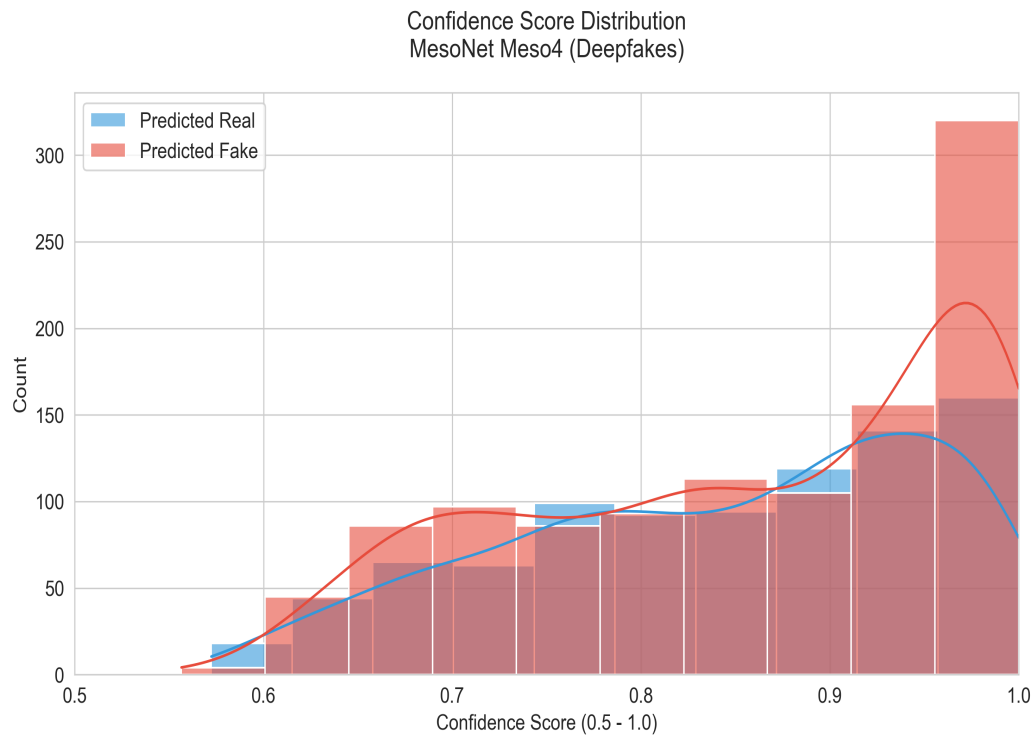


Figure 4: Confidence Score Distribution - Distribution of prediction confidence levels.

Visual Results (Continued)

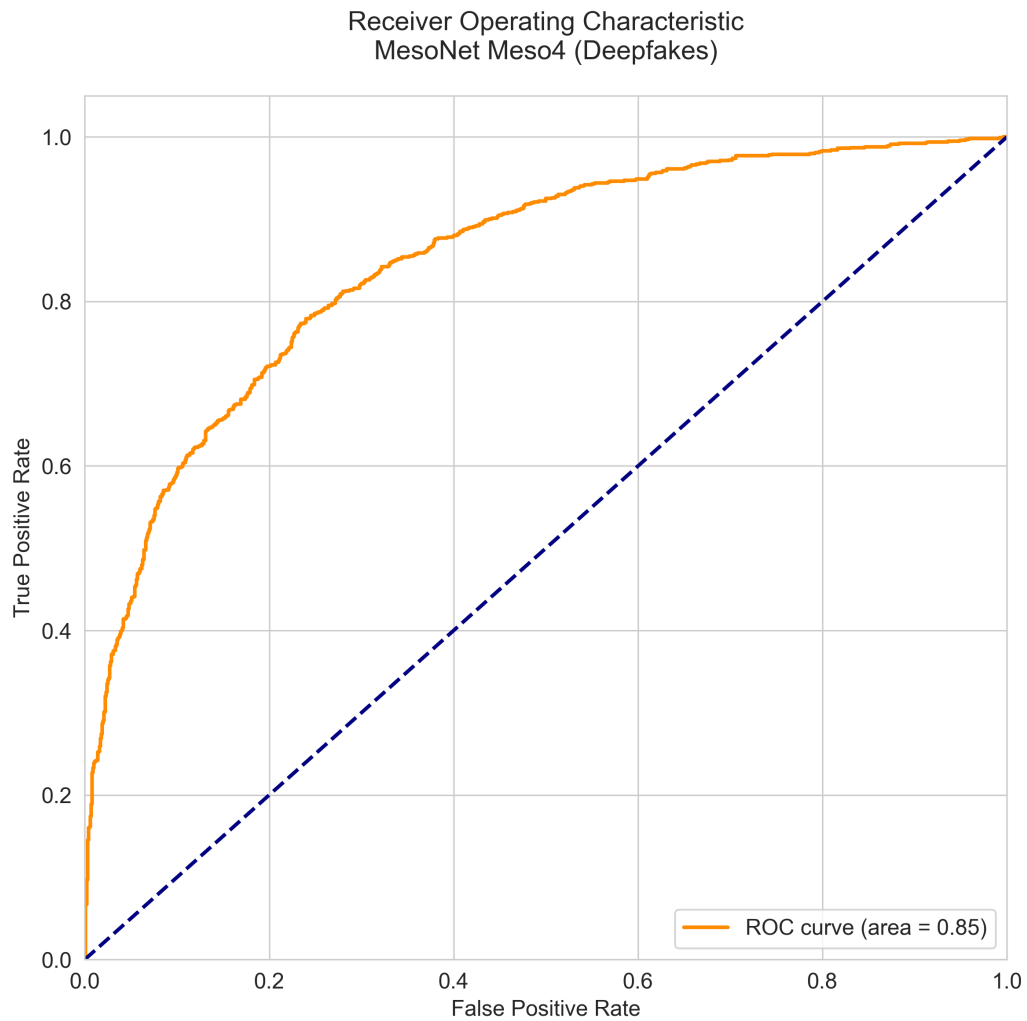


Figure 5: ROC Curve - Receiver Operating Characteristic showing discriminative ability.

Metrics Summary

This section provides a detailed breakdown of the key performance metrics used to evaluate the MesoNet model. Each metric offers unique insights into different aspects of the model's classification performance.

Metric	Value	Status	Interpretation
Accuracy	76.15%	Poor	Proportion of correct predictions
Precision	73.67%	Poor	Accuracy when predicting fake
Recall	81.40%	Moderate	Proportion of fakes detected
F1-Score	77.34%	Poor	Harmonic mean of precision & recall
AUC Score	0.85	Moderate	Area under ROC curve

Confusion Matrix Breakdown

	Predicted Real	Predicted Fake	Total
Actual Real	709	291	1,000
Actual Fake	186	814	1,000
Total	895	1,105	2,000

Metric Interpretation Guide

- **Accuracy:** The overall correctness of predictions. Values below 50% indicate the model is performing worse than random chance.
- **Precision:** When the model predicts a video is fake, how often is it correct? Low precision means many false alarms (real videos flagged as fake).
- **Recall (Sensitivity):** Of all actual fake videos, how many did the model detect? Low recall means fake videos are slipping through undetected.

- **F1-Score:** The harmonic mean of precision and recall. Useful when there's an imbalance between false positives and false negatives.

- **AUC (Area Under ROC Curve):** Measures the model's ability to distinguish between classes. AUC = 0.5 means no discrimination (random), AUC = 1.0 means perfect separation.

Detailed Analysis

What Worked Well

✓ Good recall (81.4%) shows the model catches most fake videos, though some sophisticated manipulations may go undetected.

Areas for Improvement

■ Low accuracy (76.1%) indicates significant classification difficulties. The model is performing worse than random chance (50%), suggesting a fundamental issue with the model configuration, weights, or prediction threshold.

■ Low precision (73.7%) indicates a high false positive rate. Many genuine videos are being incorrectly classified as fake, which could lead to false accusations in real-world applications.

■ Moderate AUC score (0.85) suggests room for improvement in the model's discriminative power.

■ The model has a strong bias toward predicting 'Fake'. Out of 477 total errors, 291 (61.0%) are false positives (real videos misclassified as fake). This suggests the classification threshold may be too low or the model has overfit to certain artifacts present in both real and fake videos.

Error Analysis

False Positives (Type I Error): 291 real videos were incorrectly classified as fake (29.1% of all real videos). In practical terms, this means legitimate content could be wrongly flagged as manipulated.

False Negatives (Type II Error): 186 fake videos were incorrectly classified as real (18.6% of all fake videos). These represent deepfakes that would slip through undetected.

Dominant Error Type: False Positives are the primary source of classification errors.

Error Type	Count	Percentage of Class
False Positives	291	29.1%
False Negatives	186	18.6%

Recommendations

Based on the analysis of the model's performance, the following recommendations are provided to improve detection accuracy and reliability.

1. Model Improvements

- Consider using an ensemble approach combining Meso4 with MesolInception for improved accuracy.
- Implement attention mechanisms to focus on manipulated facial regions.
- Add batch normalization layers to improve training stability and generalization.
- Experiment with different activation functions (e.g., Swish, GELU) for better feature learning.

2. Data & Preprocessing Improvements

- Apply face alignment and normalization to reduce pose variations.
- Implement data augmentation (rotation, scaling, color jittering) during training.
- Consider extracting more frames per video (20-30) for better temporal coverage.
- Add preprocessing for different video compression levels to improve robustness.

3. Threshold & Decision Logic

- Based on the current results showing high false positive rates, consider raising the classification threshold above 0.5 to reduce false alarms.
- Implement confidence-weighted voting when aggregating frame predictions.
- Use calibration techniques (Platt scaling, isotonic regression) to improve probability estimates.

4. Practical Deployment Considerations

- Given the current accuracy levels, this model should not be used as the sole decision-maker for deepfake detection in production systems.
- Implement human-in-the-loop verification for high-stakes decisions.
- Consider this model as one component in a multi-modal detection pipeline.
- Regularly update the model as new manipulation techniques emerge.
- Maintain separate confidence thresholds for different use cases (e.g., content moderation vs. forensic analysis).

5. Specific Improvements Based on Results

- Implement cross-validation testing on unseen manipulation methods.
- Consider data augmentation to improve robustness to compression artifacts.
- Test with different frame aggregation strategies (majority vote, average, max).

Technical Details

Model Architecture: Meso4

The Meso4 architecture consists of four convolutional blocks followed by fully connected layers. The network was specifically designed to focus on mesoscopic features of facial manipulations:

Layer	Configuration	Output Shape
Input	RGB Image	256 × 256 × 3
Conv Block 1	8 filters, 3×3, ReLU, BatchNorm, MaxPool	128 × 128 × 8
Conv Block 2	8 filters, 5×5, ReLU, BatchNorm, MaxPool	64 × 64 × 8
Conv Block 3	16 filters, 5×5, ReLU, BatchNorm, MaxPool	32 × 32 × 16
Conv Block 4	16 filters, 5×5, ReLU, BatchNorm, MaxPool	8 × 8 × 16
Flatten	-	1024
Dense 1	16 neurons, Dropout 0.5	16
Dense 2 (Output)	1 neuron, Sigmoid	1

Original Training Details (from paper)

The MesoNet model was originally trained by Afchar et al. (2018) with the following configuration:

- **Training Data:** Face2Face and Deepfakes datasets from FaceForensics
- **Optimizer:** Adam with default parameters
- **Loss Function:** Binary Cross-Entropy
- **Batch Size:** 75
- **Epochs:** 1000 (with early stopping)
- **Face Extraction:** Dlib face detector with margin expansion
- **Input Size:** 256×256 RGB images

Testing Methodology

The evaluation presented in this report was conducted using the following methodology:

- **Dataset:** Deepfakes dataset (1,000 real + 1,000 fake videos)
- **Weight File:** Meso4_DF.h5
- **Frame Extraction:** Multiple frames extracted uniformly from each video
- **Face Detection:** OpenCV's DNN face detector (ResNet-based)

- **Aggregation:** Per-frame predictions averaged for final video classification
- **Decision Threshold:** 0.5 (predictions ≥ 0.5 classified as fake)

Reference

D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," in IEEE International Workshop on Information Forensics and Security (WIFS), 2018.