

**GENERALISABILITY OF MADD AND XCEPTION
MODELS**

DYLAN HUYNH 24504870

SUPERVISOR: YANJUN ZHANG

UNIVERSITY OF TECHNOLOGY SYDNEY

Table of Contents

Acknowledgements	<i>iii</i>
Abstract.....	<i>iv</i>
1. Introduction	<i>1</i>
1.1 Background and motivation	<i>1</i>
1.2 Identified research gap and rationale	<i>2</i>
1.3 Research aim and significance	<i>3</i>
1.4 Research questions.....	<i>4</i>
1.5 Research objectives	<i>4</i>
1.6 Literature review	<i>4</i>
2. Methodology.....	<i>5</i>
2.1 Research design	<i>5</i>
2.2 XceptionNet architecture	<i>5</i>
2.3 MADD architecture	<i>6</i>
2.4 Comparative framework	<i>7</i>
2.5 Datasets	<i>7</i>
2.6 Implementation details	<i>8</i>
2.7 Evaluation metrics.....	<i>10</i>
3. Results	<i>11</i>
3.1 Training results on FF++	<i>11</i>
3.2 Initial testing results.....	<i>13</i>
3.3 Fine-tuning on DF40.....	<i>16</i>
3.4 Fine-tuned evaluation results	<i>19</i>
4. Discussion.....	<i>23</i>
4.1 FaceForensics++ training and testing of FaceForensics++	<i>23</i>
4.2 Initial evaluation on DF40 and DeepfakeEval2024.....	<i>24</i>
4.3 DF40 fine-tuning phase	<i>25</i>
4.4 Post fine-tuning evaluations	<i>25</i>
4.5 Overall interpretation	<i>26</i>
5. Conclusion	<i>26</i>
5.1 Issues and future work	<i>27</i>
References	<i>28</i>
Appendices	<i>31</i>
GitHub link to project:.....	<i>31</i>
ChatGPT prompts full transcript:.....	<i>31</i>

Acknowledgements

I wish to acknowledge the individuals who aided me throughout this project from its conception to the finished result. I want to sincerely thank my supervisor Yanjun Zhang who provided me support, advised me and helped me grow as a student throughout the braining storming process and through to the completion of this project.

I would like to acknowledge the teaching staff at the University of Technology Sydney for helping me gain the knowledge and providing the learning environment needed to complete this research. I would also like to acknowledge the use of editorial assistance through ChatGPT regarding bibliography assistance, debugging, understanding of performance metrics and Artificial Intelligence concepts.

Furthermore, I want to thank my parents and friends who have shown interest and supported me throughout my academic endeavours and who had patience and belief in me during this challenging yet rewarding journey.

Abstract

The rapid advancement of Generative Artificial Intelligence (GenAI) has enabled the creation of hyper-realistic manipulated media, known as deepfake. Whilst this technology has many creative applications, it also presents serious ethical and cybersecurity challenges through misinformation or identity manipulation. This research investigates the generalisation ability of two deepfake detection models; XceptionNet, a convolutional neural network (CNN)-based model and the Multi-Attentional Deepfake Detector (MADD), attention-based model across multiple datasets to evaluate their robustness and adaptability.

This study utilises a comparative, quantitative experimental design. Initially, both models are trained on FaceForensics++ (FF++) dataset with identical training conditions to establish a controlled baseline. The best-performing checkpoints are used for evaluation on FF++, DF40 and DeepfakeEval2024. These checkpoints are then further fine-tuned on the DF40 dataset which contains a more diverse range of manipulations and real-world distortions. They are then re-evaluated on all three datasets to compare their performance and generalisation.

Results indicated that while both models achieved high accuracy on FF++ training, their performance declined when tested on unseen datasets revealing overfitting and poor domain transferability. After fine-tuning, both models had near-perfect validation accuracy but exhibited domain pattern loss when re-evaluated on FF++. However, both models improved performance on DeepfakeEval2024, suggesting enhanced adaptability on unseen, in-the-wild data. Across all experiments, MADD outperformed XceptionNet in recall and F1-score but requires significantly more computational resources.

These findings highlight the trade-off between efficiency and adaptability where this research aims to contribute to the understanding of how architectural design impacts model robustness and cross-domain generalisation. Future work should improve upon the methodology, if the required resources are available, and should explore continual learning, domain adaptation and lightweight attention-based architectures to improve real-world detection reliability and scalability.

1. Introduction

1.1 Background and motivation

The advancements in Artificial Intelligence (AI), particularly Generative AI (GenAI), have led to the widespread accessibility and creation of artificially generated media, such as deepfakes. Deepfakes are hyper-realistic images or videos generated by deep learning models to create new images or mimic existing ones with seamless changes. Deepfakes typically uses generative adversarial networks (GANs) or diffusion models to generate images. Although its application towards entertainment and content generation is limitless, it also poses significant ethical, legal and security concerns, especially when used in identity spoofing, misinformation and malicious impersonation.

Early work in deepfake detection revolved around convolutional neural networks (CNNs)-based models such as MesoNet (Afchar et al., 2018), which demonstrated reasonable performance on controlled datasets. CNNs are utilised in most image-based deepfake detection research with well-known architectures such as XceptionNet demonstrating promising results on standard benchmarks. However, these models suffer from generalisation issues when exposed to unseen or real-world distorted images, such as from social media platforms. This is mainly due to the gap between the quality and diversity of the training data and the in-the-wild samples.

Image-based deepfakes are improving with more sophisticated post-processing techniques like compression, noise addition and being able to reduce artefacts, making it more difficult to detect. This has led to detection models performing worse, such as benchmarks like ForgeryNet (He et al., 2021) and DeepfakeEval2024 (Chandra et al., 2025) that show current state-of-the-art detectors struggle against in-the-wild data, with accuracy drops of 45% on social media distorted image samples. Newer models have evolved with the complexity of deepfake forgeries with more advanced architectural features such as attention mechanisms. The Multi-Attentional Deepfake Detection (MADD) model utilised this architecture with multi-scale attention layers to enhance its ability to detect forgery features (Shao et al., 2023).

Concurrently, benchmark datasets such as FaceForensics++ (FF++) (Rössler et al., 2019), DF40 (Yan et al., 2024) have emerged and offer diverse, realistic samples to test the generalisability of detection models utilising a range of manipulation types, compression artefacts and social media degradation to evaluate a realistic setting.

This research adopts a comparative, quantitative experimental methodology to investigate the generalisation performance of deepfake detection models. It evaluates the performance of the MADD and XceptionNet models across varied datasets to determine their strengths and limitations under controlled training and testing conditions. MADD was chosen as it a more state of the deepfake detection model (Shao et al., 2023) and XceptionNet due to its presence in state-of-the-art literature (Rössler et al., 2019) and both models being open-source implementations allowing for reproducibility. Furthermore, both models represent different architectures where MADD leverages an attention-based mechanism while XceptionNet relies on CNNs, which allows for a more insightful comparison between model types. This approach aligns with the research objective of assessing generalisation as the training conditions are controlled, allowing for a clearer interpretation of how the models adapt to previously unseen data.

The approach aims to address the limitations identified in the benchmark studies ForgeryNet (He et al., 2021) and DeepfakeEval2024 (Chandra et al., 2025). By controlling the training environment and standardising the evaluation metrics, it ensures the performance of both models can be compared to showcase the difference in the model architecture. The results of this investigation aim to inform best practices when developing deepfake detection systems and contribute to the deepfake detection efforts for the misuse of GenAI.

1.2 Identified research gap and rationale

Despite advances in deepfake detection, there remains a lack of systematic comparison between CNN-based and attention-based models for cross-dataset generalisation. This project aims to address this research gap and investigate the performance of the XceptionNet and MADD models on how their different architectures generalise across datasets. FF++ dataset is used for training on both models to provide a controlled baseline with defined manipulation types, compression levels and post-processing. The DF40 dataset (Yan et al., 2024) provides a wider, diverse set of over 40,000 deepfake videos with varying

manipulation techniques, compression settings and real-world conditions to help assess the robustness of each model’s ability to generalise beyond the datasets it was trained on.

DeepfakeEval2024 extends on this with more real-world ‘in-the-wild’ scenarios for more unseen, imperfectly processed content.

Following the initial training, both models are fine-tuned using DF40 to assess whether exposure to a more diverse dataset will affect the accuracy and cross-domain adaptability across domains. It is re-evaluated on FF++, DF40 and DeepfakeEval2024 to help quantify improvements in cross-dataset robustness.

This investigation aligns with the aim of this project; to compare and assess the robustness and adaptability of attention-based and CNN-based architectures in deepfake detection. While both models are grounded in current literature, this work extends upon these works by offering systematic cross-dataset benchmark giving valuable insight towards architectural differences in generalisation and transferability which will aid in practical, real-world deepfake detection systems.

1.3 Research aim and significance

Research Aim: *To develop and evaluate the performance metrics that quantify the generalisation ability of deepfake detection models across diverse image databases.*

The aim of this project is to explore the effectiveness and generalisability of the MADD model and the XceptionNet model in identifying deepfake images. Specifically, it investigates the effectiveness of each model on both the training dataset and on external datasets that reflect real-world distortions.

This project is significant as it contributes to the knowledge around the generalisability of deepfake detection systems. This follows the lack of current, reliable models across diverse manipulation techniques as seen in “Unmasking Deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence” (Taeihagh, A., 2024) and benchmark studies like DeepfakeEval2024 (Chandra, N. A., et al, 2025). This study aims to compare these two models under consistent

experimental conditions, offering insight into their practical robustness in real-world applications.

Moreover, this project serves to act as a valuable case study in the emerging research domain of deepfake detection. As an undergraduate honours project, it will demonstrate how an empirical evaluation using publicly available tools and datasets can yield a meaningful contribution to the field of AI.

1.4 Research questions

- How do MADD (attention-based) and XceptionNet (CNN-based) models compare in terms of accuracy on the training dataset given identical learning parameters?
- How well does each model generalise when exposed to unseen datasets?
- Which model demonstrates greater robustness and adaptability to varying manipulation techniques and real-world image distortions?
- Does fine-tuning on a more diverse dataset (DF40) improve cross-dataset generalisation of either model?

1.5 Research objectives

- Evaluate the detection performance of MADD and XceptionNet models when trained on a common dataset with identical conditions.
- Evaluate and compare cross-dataset performance on DF40 and DeepfakeEval2024 measuring metrics such as accuracy, AUC, precision, recall and F1-score.
- Analyse the impact of fine-tuning on DF40 in improving model adaptability and performance across unseen datasets.
- Assess and compare the generalisation capability of both models when tested on unseen datasets
- Provide empirical insights that guide future deepfake detection research towards generalised and resilient architectures.

1.6 Literature review

The rapid growth and evolution of Deepfake Detection mean the advancements of architectures such as convolutional neural networks (CNNs) and attention-based mechanisms. XceptionNet emerged as a foundational CNN-based model due to its depth wise separable convolutions to extract fine-grained spatial features. On the benchmark datasets of

FF++ and Celeb-DF (Li et al, 2020), XceptionNet demonstrated a strong performance making it a reliable baseline for deepfake detection research (Rössler et al., 2019). The generalisation of XceptionNet is not yet tested making it an ideal comparator for evaluating newer architectures.

Comparatively, MADD is more representative of recent attention-based models utilising spatial attention mechanisms towards facial regions (Shao et al., 2023). By leveraging low-level texture cues and information through attention maps, it improves MADD's sensitivity to subtle forgery artefacts. MADD showed promising results on FF++ and Celeb-DF but as a newer model, the generalisation capabilities are not yet explored leaving a gap for comparison between CNN-based and attention-based architectures for cross-dataset evaluation.

Overall, existing studies suggest that deepfake detection models achieve high performance under controlled conditions but struggle to generalise across different dataset and manipulation types (Chandra et al., 2025). This observation motivates the present research, which evaluates XceptionNet and MADD trained on FF++ and tested on DF40 (Yan et al., 2024), FF++ and DeepfakeEval2024 (Chandra et al., 2025). The models are then further fine-tuned on DF40 and re-evaluated to take into account exposure to a more diverse dataset. These newer datasets provide a wide variety of manipulation techniques and realistic conditions for a more comprehensive investigation into robustness and adaptability.

2. Methodology

2.1 Research design

This project utilises a comparative, quantitative experimental design to help quantify and compare the difference between the cross-dataset generalisation ability of two different architectures: CNNs and attention-based models.

2.2 XceptionNet architecture

XceptionNet utilises a CNN framework where each input frame is pre-processed, meaning that it is resized and normalised¹ to remove irrelevant information and ensure

¹ Normalised: all pixels are adjusted to the same scale.

consistent scale to allow focus on regions likely to contain manipulations. In CNN, early layers extract low-level features such as edges (boundaries between facial regions), textures (skin smoothness, lighting inconsistencies) and colour patterns (unnatural tones or blending artefacts). Each layer becomes increasingly more complex in what it detects, where early layers detect higher semantic representations such as eyes, lips or wrinkles and later layers detect facial structure or symmetry.

XceptionNet uses depth wise separable convolution² where spatial filters are applied independently to each colour channel as opposed to mixed. Pointwise Convolution (1x1) is applied to combine these outputs across channels into a feature map, which is more efficient and flexible than regular convolutions (Rössler et al., 2019). Using this method, it can more efficiently learn fine-grained features as it uses fewer computations allowing it to more accurately detect blending errors from skin textures, slight colour mismatches and irregular lighting.

Pooling layers further condenses the spatial dimensions of feature maps while retaining important information, enabling prioritisation of essential visual cues over pixel-level noise. After convolution and pooling, the output is flattened and passed into fully connected layers where they combine and extract all features to predict whether the input is ‘real’ or ‘fake’.

2.3 MADD architecture

As deepfake images are becoming increasingly harder to distinguish, the differences are becoming more subtle which suggests a different approach may be needed. MADD model uses an attention-based approach utilising spatial attention heads which generate attention maps³ focused on different facial regions; eyes, mouth, blending boundaries (Shao et al., 2023).

² Depth wise separable convolutions are when the spatial filter is applied independently to each input channel, such as one for red, one for blue etc.

³ Attention Maps: a matrix over the image indicating which regions the model should focus on.

Texture-feature-enhancement-blocks amplifies low-level texture cues allowing the model to capture fine artefacts that show up in early layers. Lastly, MADD combines low-level and high-level semantic features⁴ through attention-guided feature aggregation. To encourage diverse attention, MADD enforces regional independence between attention heads and blurs high attention regions, so it is forced to attend other informative areas. By enhancing textural features and combining them with semantic features, the model is better equipped to pick up subtle artefacts. The architecture focuses on multiple regions and varied features to help it learn transferable cues rather than overfitting⁵ to a single dataset's artefacts, so it is designed better for generalisation.

2.4 Comparative framework

The contrast between each model's architecture enables a deeper understanding on how architecture can influence generalisation performance. Using a controlled training and external testing framework, it ensures that the evaluation process remains reproducible and unbiased.

2.5 Datasets

There are three datasets being used to help model generalisation:

- FaceForensics++: This is a controlled benchmark dataset containing over 1.8 million manipulated facial images from methods FaceSwap, Face2Face and Deepfakes. These contain real-time expression transfer (re-enactment), swapping of faces, face replacement and texture synthesis). The images are sourced from YouTube meaning that they more closely resemble ‘in-the-wild’ images with a noise-controlled environment.
- DF40: Is a diverse dataset containing over 40,000 deepfake videos utilising over 40 different deepfake manipulation techniques with varying compression levels. This dataset is more recent (2024 compared to 2019) and is used both for evaluation and fine-tuning the weights.

⁴ Semantic features: features from layer layers such as expression, pose, face identity

⁵ Overfitting: when a model performs well on familiar training data but fails to generalise on unseen samples

- DeepfakeEval2024: Is another state-of-the-art benchmark that simulates ‘in-the-wild’ conditions by mixing authentic and synthetic media from social media and deepfake detection platforms. This dataset is used as an external evaluation tool.

2.6 Implementation details

A pre-existing weight⁶ was used to initially train both models on, this weight is pre-trained on FF++ and is being further fine-tuned with our learning parameters (Standley, T. 2019). The implementation of the training of both models was adapted from the publicly available repositories with modifications for dataset compatibility and evaluation. Code assistance tools were used to debug and adjust but all model design and analysis were conducted independently.

The FF++ was used to train both models with an 85:15 split between training and testing sets. Following the initial training the best performing model weights, based of accuracy and average loss, were fine-tuned on the DF40 dataset. It should be noted that FF++’s image compression level of c23 was used as a standard mid-level compression to simulate realistic social media compression.

The DF40 dataset was split into training, validation and evaluation subsets in a 70:15:15 ratio using the *FRALL_cdf* subset. This split of DF40 contains face-reenactment manipulations of the Celeb Deepfake dataset which is more diverse and unknown to the current trained models. It is trained on the 70% split then validated on the 15% before the last 15% is evaluated.

The DeepfakeEval2024 dataset was used exclusively as an evaluation dataset meaning it was not used for training or fine-tuning to ensure genuine assessment of cross-dataset generalisation performance under ‘in-the-wild’ conditions.

Both models were implemented using PyTorch, using the open-source models provided via the author’s links. Some models had to be adjusted to fit the training, validation

⁶The weight in the Python file was downloaded: <https://github.com/tstandley/Xception-PyTorch/tree/master>

and to add more parameters. Due to hardware and computational limitations, several parameters were standardised for efficiency and stability across both models to ensure fair comparison. The parameters are as follows:

- **Epochs = 5:** Pilot tests demonstrated that both models converged effectively within five epochs, with minimal improvement increased risk of overfitting risk and training time.
- **Batch Size = 28:** Selected based on limited GPU memory capacity. Batch sizes around 32 are common for CNNs, but MADD's attention mechanism uses more memory, so 28 was chosen as the optimal stable value without causing memory overflow.
- **Learning Rate (LR) = 0.0002:** A small learning rate was chosen to ensure stable convergence for both architectures. This value was taken from prior deepfake detection research such as the MADD paper and the original XceptionNet implementation, both of which used learning rates in the range of 1e-4 to 2e-4 for fine-grained feature learning (Shao et al., 2023; Rössler et al., 2019)..
- **Weight Decay = 0.0001:** Used as a regularisation term to prevent overfitting by penalising large weights and mitigate overfitting, which is particularly useful when training on a smaller or less diverse dataset such as FaceForensics++.
- **Label Smoothing = 0.1:** Label smoothing was added to make the model less confident on the training labels, improving generalisation and reducing overfitting.
- **Image Size = 224:** This is the standard input dimension for XceptionNet and is widely used in pretrained CNNs. It provides a balance between preserving important facial details and maintaining efficient computation.
- **Max Images per Video (Train = 50, Validate = 100):** These limits were set to control dataset size and prevent bias from longer videos containing more frames. By capping the number of frames per video, each sample contributes equally to training, ensuring balanced representation and reducing computational load.

Both models were trained under identical hyperparameter settings to maintain consistency and comparability. The best performing model checkpoint was determined based on the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which evaluates the trade-off between the true positive rate and false positive rate (Evidently AI., 2025). So, in this context, it reflects how effectively the model distinguishes between real and fake samples;

how many actual fakes did it detect correctly and how many are incorrectly classified as fake. Following AUC-ROC, F1 score was also used to determine model checkpoint which depicts the trade-off between sensitivity and precision of the decision threshold. This metric balances the precision (avoiding false positives) and recall (detecting actual fakes).

After determining the best checkpoint from FF++ training, these model weights were used for evaluation on DF40 and Deepfake-Eval2024. Following this these weights are then fine-tuned on DF40 with the same learning parameters as above then re-evaluated on FaceForensics++, DF40 and DeepfakeEval2024. This quantifies any improvements in cross-domain adaptability following exposure to a more diverse dataset. When testing, the parameters of 100 frames per video for FF++ dataset with all being tested with a batch size of 64. DeepfakeEval2024 had no frame limit and DF40 had a limit of 24 frames per video. This was due to the sheer size of each dataset with DF40 containing the most samples, thus it was lowered to compensate for the hardware constraints.

2.7 Evaluation metrics

The performance of the models was determined by accuracy, AUC-ROC, F1, precision and recall.

Accuracy determines the model's ability for its predictions of whether it is real or fake.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

True Positives = Correctly predicted fakes

True Negatives = Correctly predicted reals

False Positive = Real predicted as fake

False Negatives = Fake predicted as real

AUC-ROC: how well a model can distinguish between a positive or negative class. So how consistent it is at detecting real or fake faces.

$$AUC = \int_0^1 TPR(FPR)d(FPR)$$

$$\text{True Positive Rate} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}$$

$$\text{False Positive Rate} = \frac{\text{FalsePositive}}{\text{FalsePositive} + \text{TrueNegative}}$$

Precision measures how correct the fake detections are.

$$\text{Precision} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positives})}$$

Recall measure how well the model can find all the fake samples.

$$\text{Recall} = \frac{\text{True Positives}}{(\text{True Positives} + \text{False Negatives})}$$

F1 is the harmonic mean of precision and recall reflecting the overall detection effectiveness.

$$F1 = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

These metrics are needed to identify performance differences and degradation. It is anticipated that both models will perform strongly on FF++ with varying degrees of accuracy loss on the unseen datasets with MADD demonstrating better adaptability due to its architecture. Fine-tuning on DF40 is expected to improve both model's generalisation.

3. Results

This section presents the quantitative results of model training and evaluation across three datasets FF++. DF40 and DeepfakeEval2024. Each subsection compares the performance of MADD and XceptionNet under identical training conditions, highlighting trends in accuracy and loss.

3.1 Training results on FF++

The training results on FF++ establish baseline performance of both models under controlled conditions. The goal is to assess their ability to converge efficiently and detect manipulated content before fine-tuning.

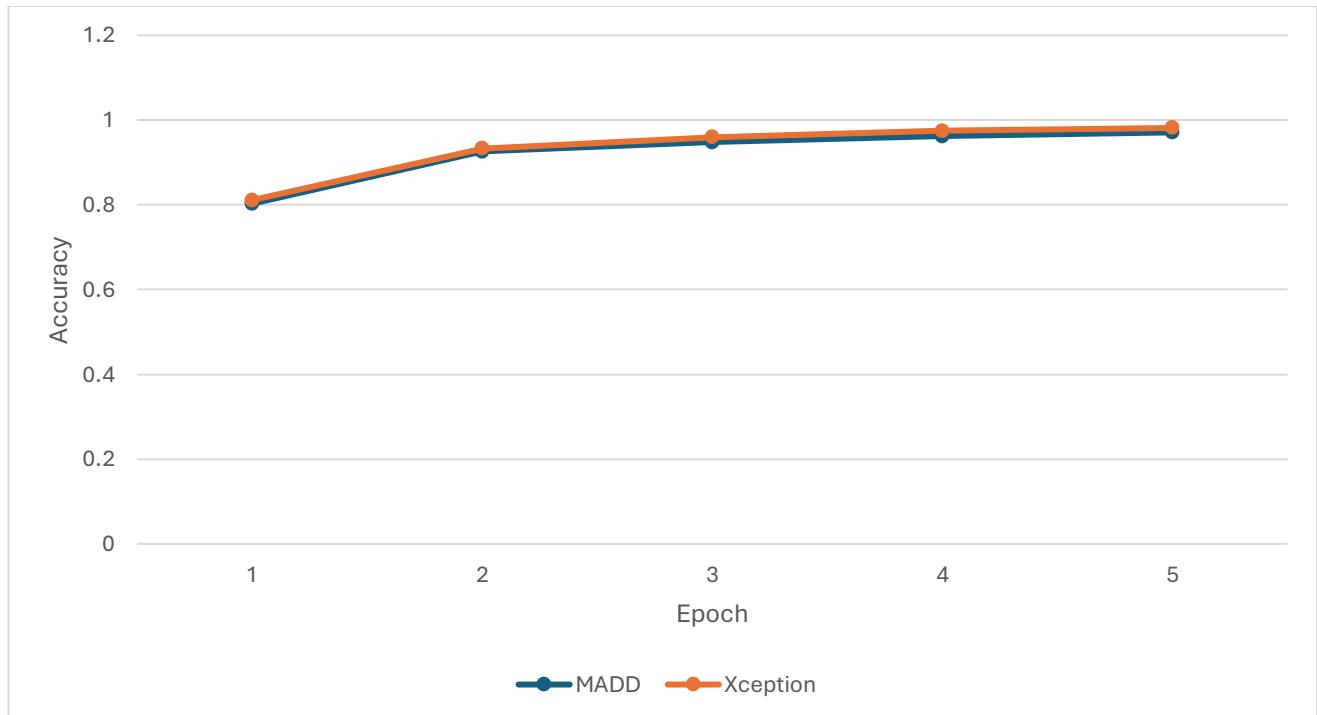
Figure 1
MADD Training Results

Epoch	Avg Loss	Accuracy
1	0.46336	0.8027
2	0.31293	0.9255
3	0.27732	0.948
4	0.25754	0.9623
5	0.24259	0.971

Figure 2
XceptionNet Training Results

Epoch	Avg Loss	Accuracy
1	0.37813	0.8111
2	0.1558	0.9331
3	0.095	0.9595
4	0.05962	0.9746
5	0.04238	0.9813

Figure 3
Initial Training Accuracy on FaceForensics++



Both models achieved high training accuracies with less average loss the higher the Epochs, see Figure 1, 2, 3.

3.2 Initial testing results

Following contains the initial testing results from weights that have only been trained on FF++ dataset.

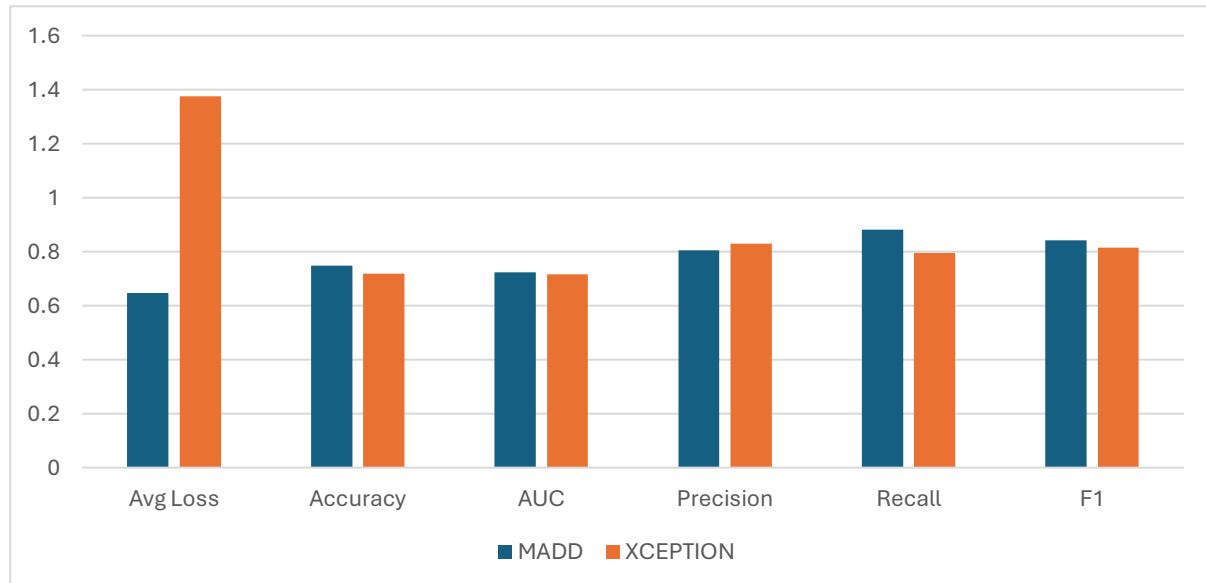
Figure 4

Initial Testing On FF++

INITIAL TEST	Avg Loss	Accuracy	AUC	Precision	Recall	F1
MADD	0.6456	0.7471	0.7222	0.804	0.8822	0.8413
XCEPTION	1.3744	0.7192	0.7164	0.8289	0.7945	0.8144

Figure 5

Initial Testing of MADD and XceptionNet on FF++



As shown in Figure 4, both models achieved moderate testing accuracies (72%-75%) indicating strong in-domain learning from FF++. MADD achieved a slightly higher accuracy and F1 which suggests better recall and precision trade-off.

Figure 6*MADD Initial Testing on DF40*

MADD	Samples	Missing	Accuracy	AUC	Precision	Recall	F1
FSALL_ff	74,510	27,279	0.5221	0.5121	0.3591	0.0499	0.0877
FSALL_cdf	193,763	20,243	0.261	0.5156	0.5648	0.0039	0.0078
FRALL_ff	107,056	0	0.4672	0.6116	0.0985	0.0083	0.0153
FRALL_cdf	309,872	0	0.2109	0.6837	0.2585	0.001	0.002

Figure 7*XceptionNet Initial Testing on DF40*

Xception	Samples	Missing	Accuracy	AUC	Precision	Recall	F1
FSALL_ff	55,992	20,574	0.5377	0.6086	0.4985	0.863	0.632
FSALL_cdf	193,763	20,243	0.6672	0.4559	0.7382	0.8515	0.7908
FRALL_ff	107,056	0	0.5932	0.6442	0.5554	0.9268	0.6946
FRALL_cdf	65,626	0	0.7464	0.4948	0.7981	0.9075	0.8493

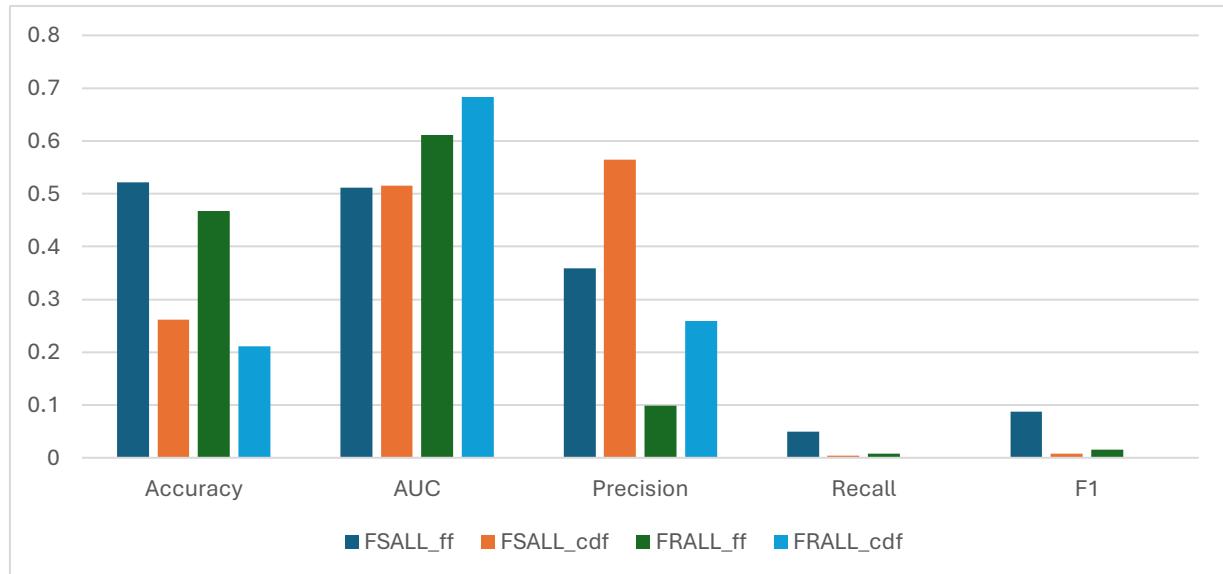
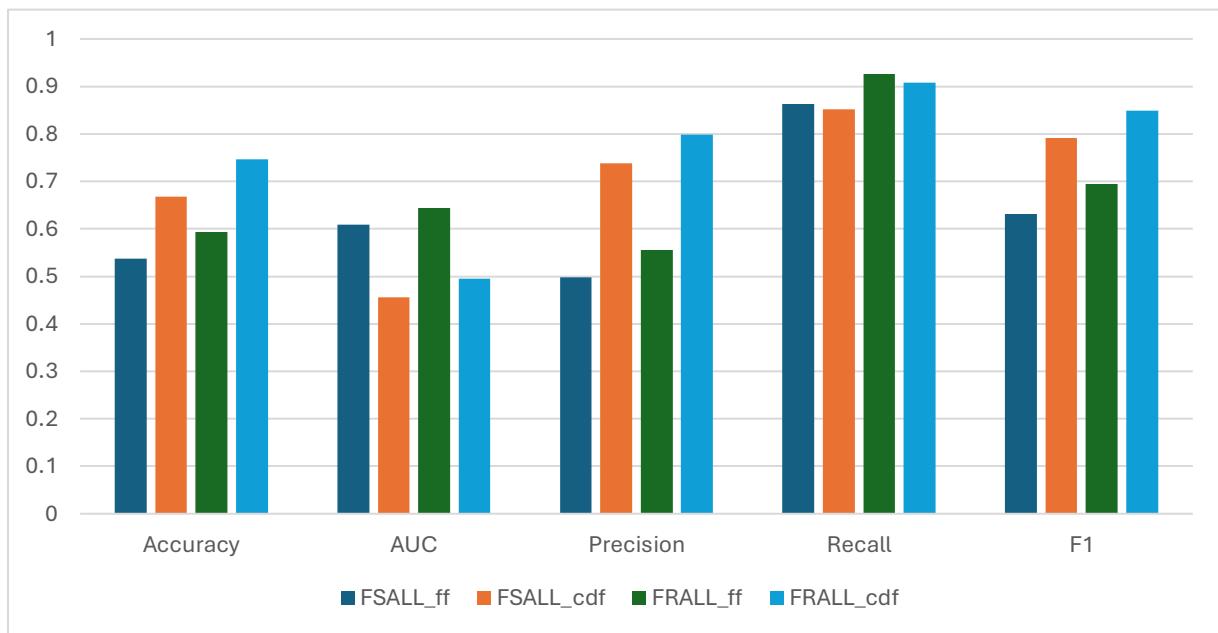
Figure 8*Initial MADD DF40 Evaluation*

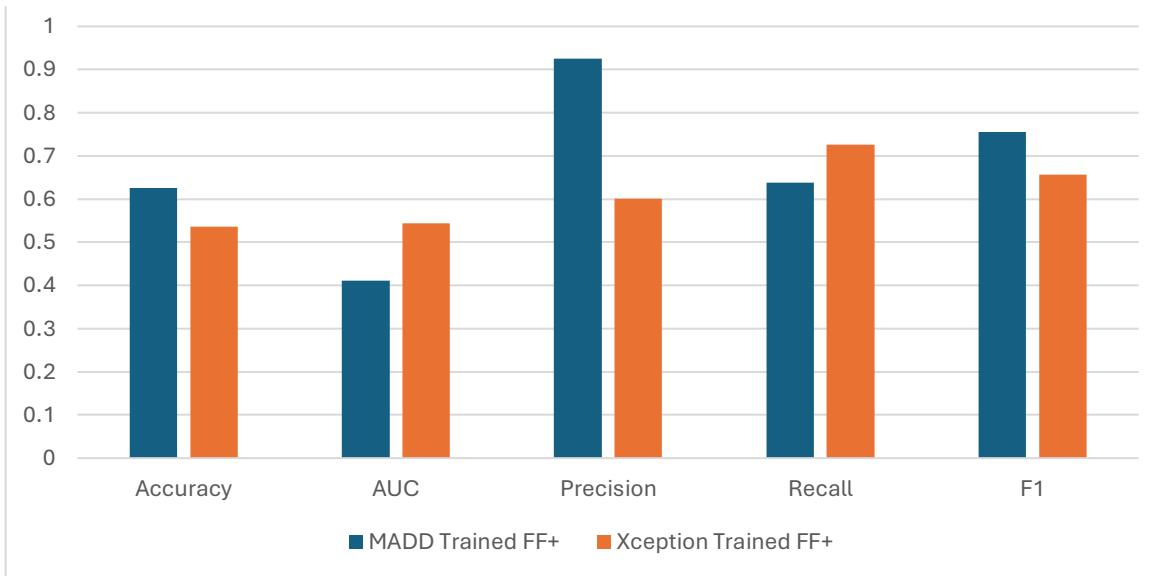
Figure 9*Initial XceptionNet DF40 Evaluation*

When applied to the unseen DF40 dataset, performance sharply declined for both models, particularly for MADD on FRALL_cdf, confirming generalisation difficulty across datasets. MADD performed poorly overall with low AUC, recall and F1 scores suggesting difficulty detecting unseen manipulations. Contrastingly, XceptionNet achieved higher accuracy and recall across all subsets demonstrating better adaptability but suffered reduced AUC due to dataset variability. Refer to Figure 6, 7, 8, 9.

Figure 10*Initial Testing on DeepfakeEval2024*

Model	Samples	ACC	AUC	Precision	Recall	F1
MADD Trained FF+	786	0.625	0.4109	0.925	0.6379	0.7551
Xception Trained FF+	786	0.5359	0.544	0.6007	0.7256	0.6573

Figure 11
Initial Testing on DeepfakeEval2024



As shown in Figure 10, both models performed decently with MADD having better accuracy, precision and F1 whilst XceptionNet having a higher AUC and recall but having a more balanced performance. The low AUC values for both models indicate limited discriminative power under uncontrolled conditions.

3.3 Fine-tuning on DF40

The fine-tuning of MADD and XceptionNet models on DF40.

Figure 12

MADD Fine-Tuned Training Results on DF40

Epoch	Avg Loss	Accuracy
1	0.3503	0.908
2	0.2509	0.969
3	0.2311	0.9808
4	0.2284	0.9824
5	0.2215	0.9867

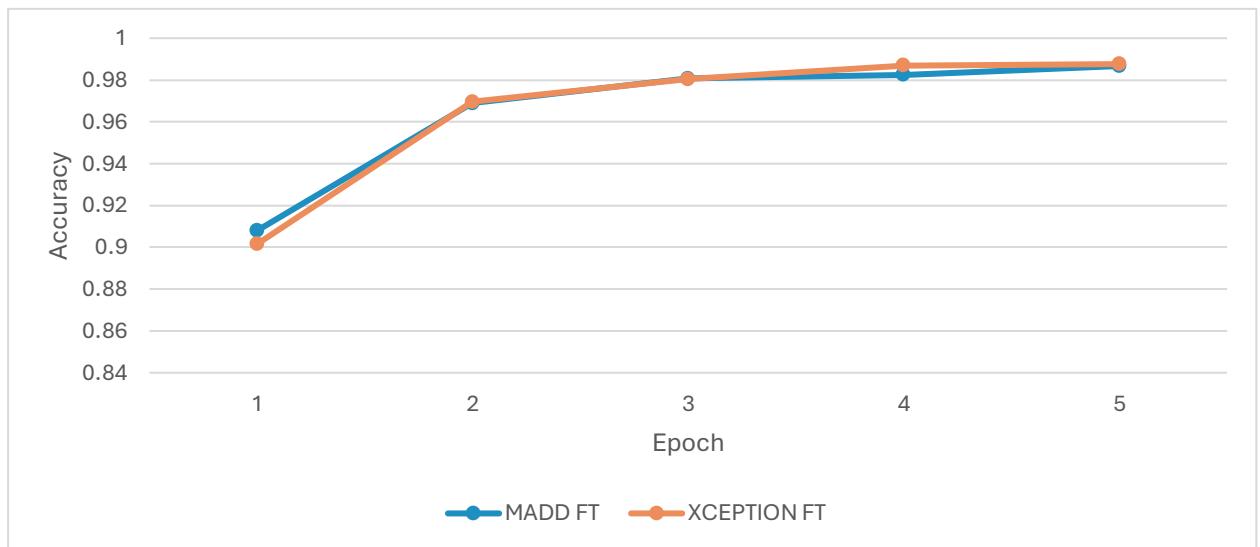
Figure 13

XceptionNet Fine-Tuning Training Results on DF40

Epoch	Avg Loss	Accuracy
1	0.3714	0.9016
2	0.2518	0.9695

3	0.2326	0.9804
4	0.2223	0.987
5	0.2218	0.9876

Figure 14
Fine-Tuned Training Accuracy on DF40



Depicted in Figure 12,13 and 14, both models demonstrated strong convergence and stable training performance during fine-tuning on the DF40 dataset. MADD showed a consistent reduction in average loss from 0.3503 in Epoch 1 to 0.2215 in Epoch 5 with an improvement in accuracy of more than 8%. XceptionNet displayed similar results with loss decreasing from 0.3714 to 0.2218 and accuracy from 90.16% to 98.76%. This improvement indicates adaptation to the DF40 dataset with both models learning more of its features.

Figure 15
MADD Fine-Tuning Validation Results on DF40

Epoch	Avg Loss	Accuracy	AUC	Precision	Recall	F1
1	0.1214	0.9703	0.9939	0.9823	0.9795	0.9808
2	0.0936	0.9831	0.9969	0.9917	0.9865	0.9891
3	0.0837	0.9869	0.9982	0.9933	0.9899	0.9916
4 (BEST CHECKPOINT)	0.0739	0.9908	0.999	0.9932	0.995	0.9941
5	0.0784	0.9888	0.9971	0.9916	0.9941	0.9928

Figure 16
XceptionNet Fine-Tuning Validation Results on DF40

Epoch	Avg Loss	Accuracy	AUC	Precision	Recall	F1
1	0.1573	0.9505	0.9941	0.9945	0.9418	0.9674
2 (BEST CHECKPOINT)	0.0974	0.9861	0.9984	0.9916	0.9905	0.9911
3	0.0833	0.9845	0.9984	0.9937	0.9863	0.99
4	0.0903	0.9843	0.9978	0.9939	0.986	0.9899
5	0.1043	0.9786	0.9927	0.9748	0.9983	0.9864

Figure 17
MADD Fine-Tuned DF40 Validation Results

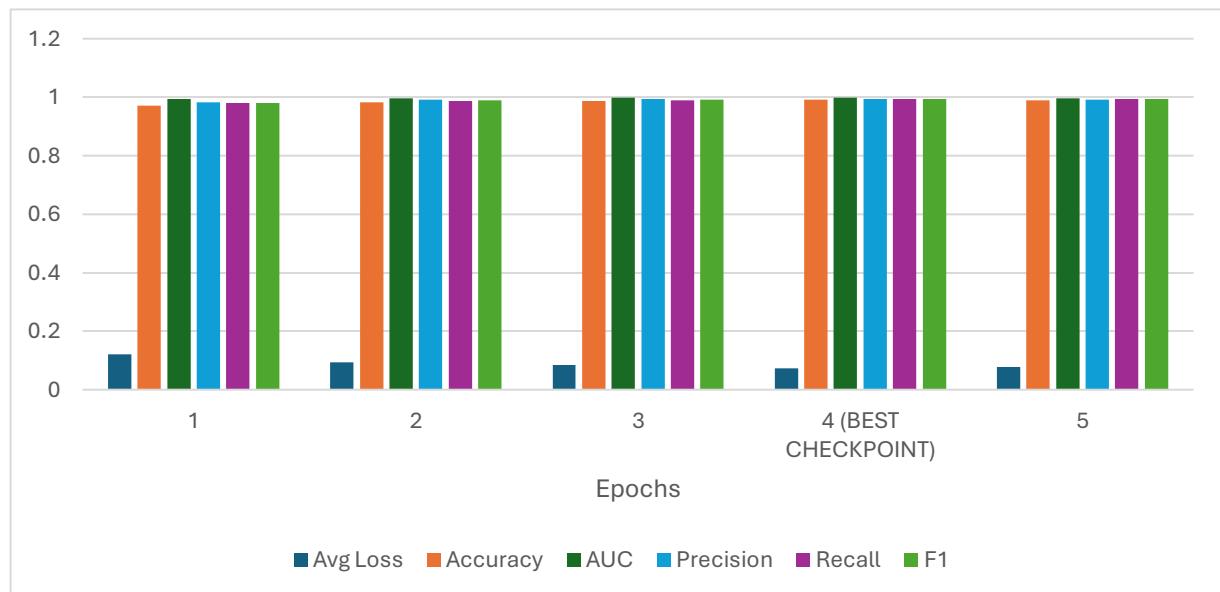
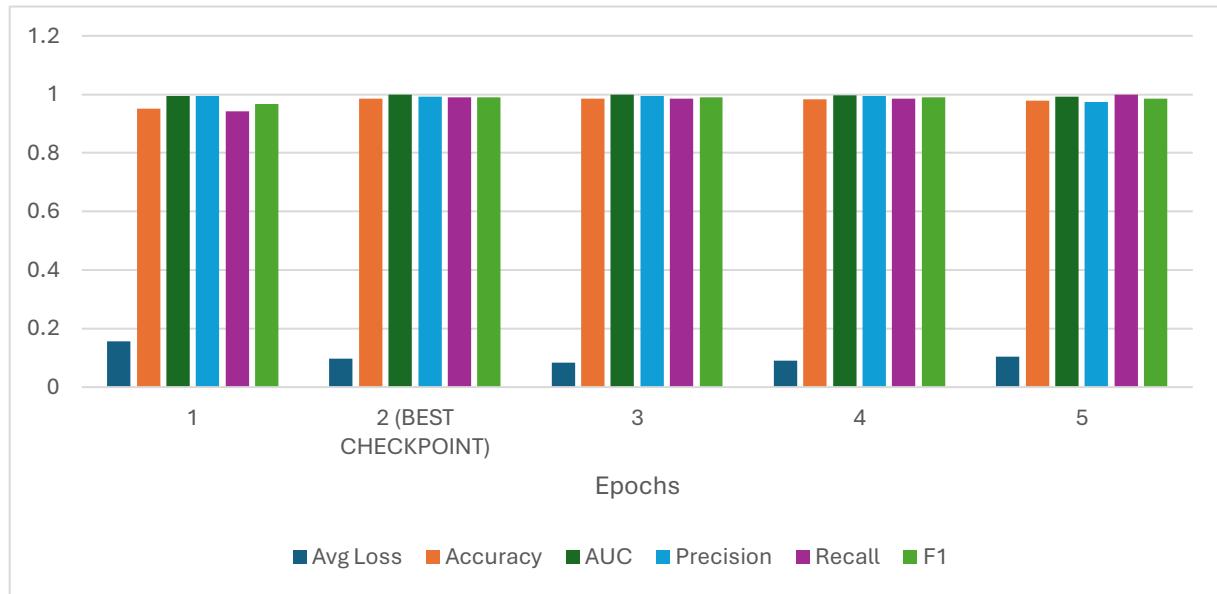


Figure 18
XceptionNet Fine-Tuned DF40 Validation Results



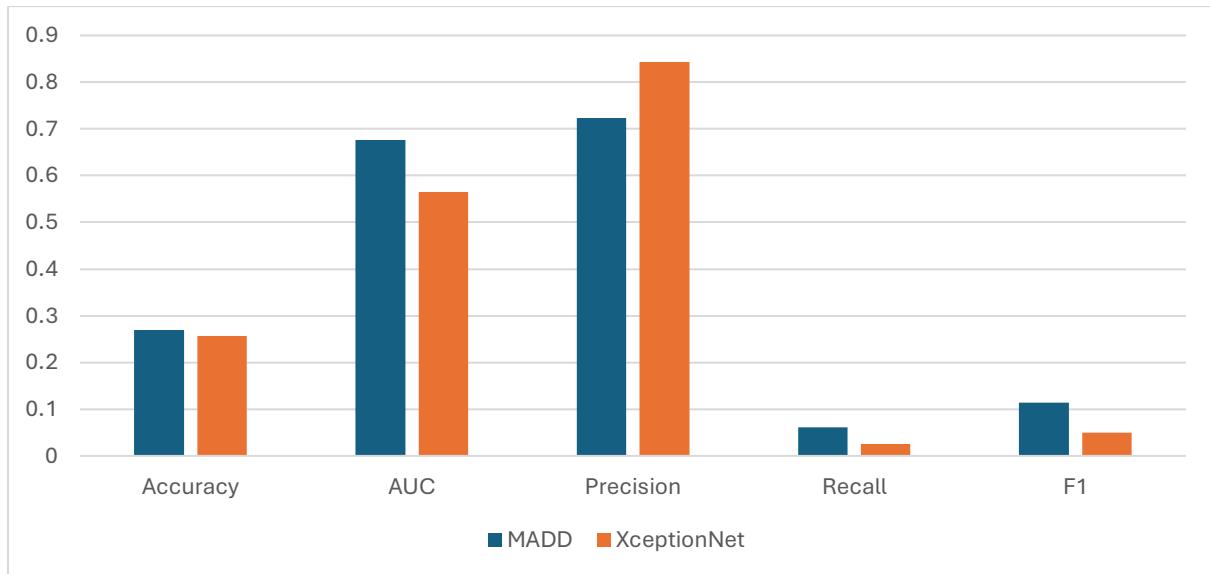
As shown in Figure 15, MADD achieved the best checkpoint at Epoch 4 with an average loss of 0.0739, accuracy of 99.08% and AUC of 0.999 indicating near-perfect separation between real and fake samples. XceptionNet reached best performance at Epoch 2 with an average loss of 0.0974, accuracy of 98.61% and AUC of 0.9984 which is depicted in Figure 16. This indicates that both models achieved exceptional validation results with near perfect AUC and accuracy suggesting overfitting. The slight rising of loss after the optimal Epoch reinforces that it might be overfitting without improving generalisation.

3.4 Fine-tuned evaluation results

The following results are the fine-tuned results when evaluated on the FF++ dataset.

Figure 19
Fine-Tuned Results on FF++

Model	Samples	ACC	AUC	Precision	Recall	F1
MADD Trained FF+	62500	0.2689	0.6762	0.7229	0.0616	0.1135
XceptionNet Trained FF+	62500	0.2561	0.5646	0.8436	0.026	0.0504

Figure 20*Fine-Tuned Results on FF++*

There is a significant drop in performance for both models' indicating loss of generalisation, refer to Figure 5 and 20. As shown in Figure 19, MADD achieved an accuracy of 26.90% with an AUC of 0.6762 with a low recall of 0.0616 affecting the F1 score to be 0.1135 despite the moderate precision of 0.7229. XceptionNet also showed decline with a 25.61% accuracy and AUC of 0.5646 similarly with low recall of 0.026 and thus F1 score of 0.0504. These results suggest that both models overfitted to DF40's data distribution losing the sensitivity to specific manipulations and compression pattern characteristics of FF++. Although MADD performed slightly better, both models showed a decline and weakened performance following exposure to new domain data.

Figure 21*MADD Fine-Tuned Results on DF40*

MADD	Samples	Missing	Accuracy	AUC	Precision	Recall	F1
FSALL ff	74,519	27,279	0.5484	0.5617	0.5093	0.4424	0.4735
FSALL cdf	193,763	20,243	0.8688	0.9711	0.9872	0.8332	0.9037
FRALL ff	107,056	0	0.7535	0.8504	0.7046	0.8697	0.7785
FRALL cdf	86,291	0	0.9895	0.9989	0.9918	0.9948	0.9933

Figure 22*XceptionNet Fine-Tuned Results on DF40*

Xception	Samples	Missing	Accuracy	AUC	Precision	Recall	F1
FSALL ff	74,510	27,279	0.5525	0.5662	0.5158	0.4071	0.4551
FSALL cdf	193,763	20,243	0.8316	0.9550	0.9851	0.784	0.8731

FRALL_ff	107,056	0	0.7742	0.8793	0.7277	0.8734	0.7939
FRALL_cdf	86,291	0	0.9842	0.9980	0.9904	0.9895	0.9899

Figure 23
MADD DF40 Results after Fine-Tuning

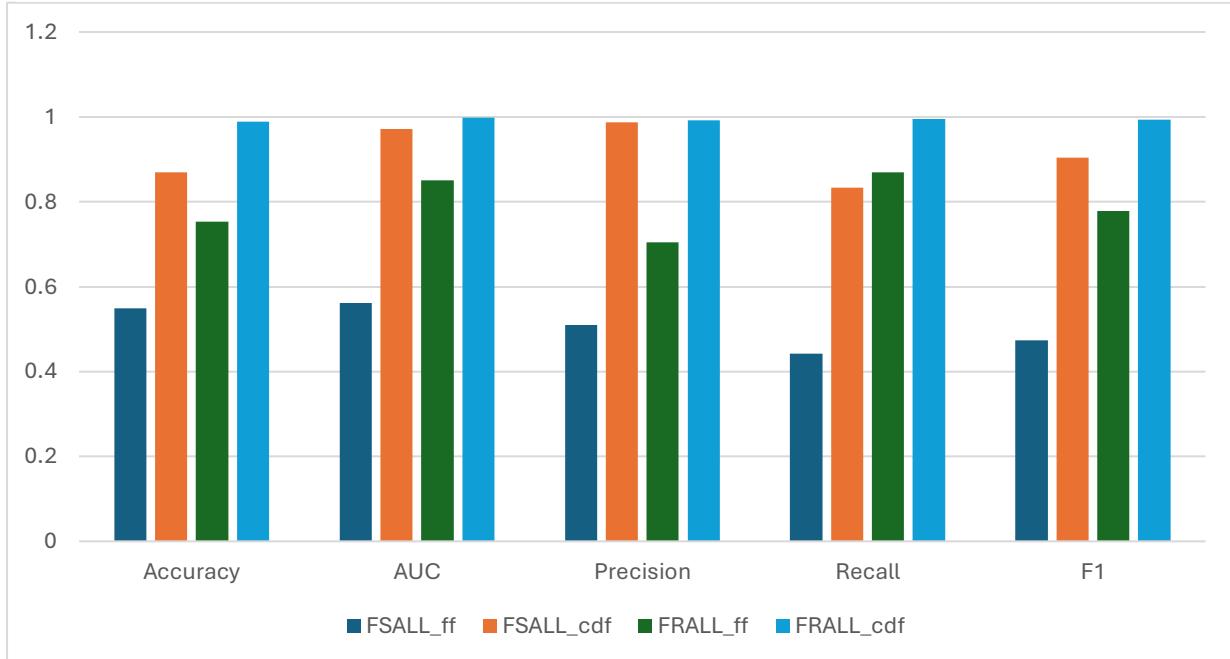
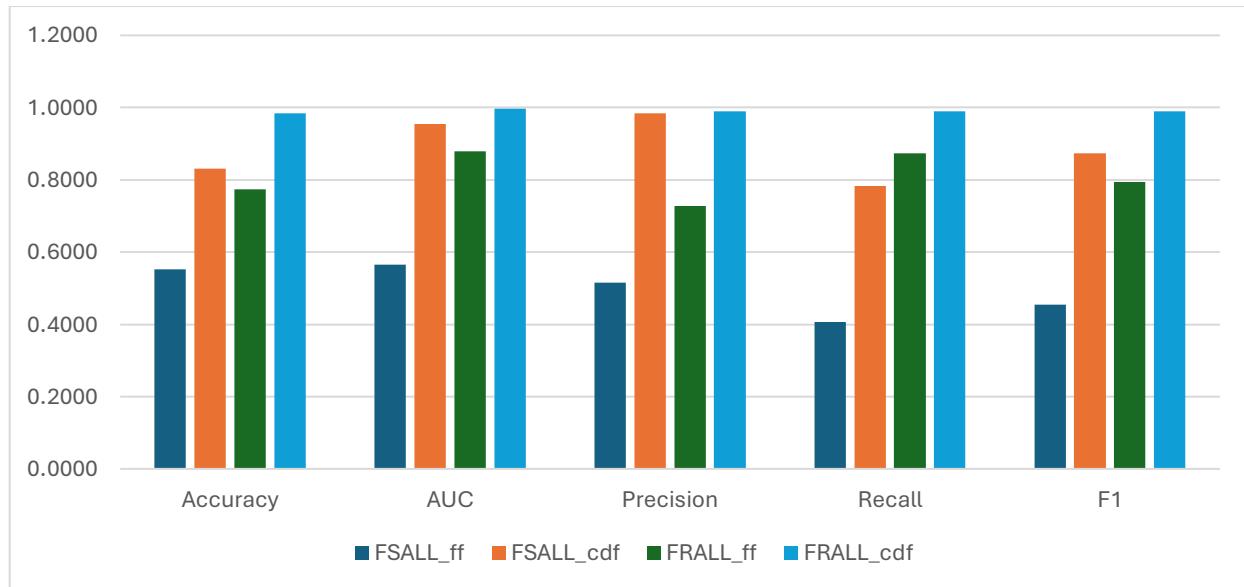


Figure 24
XceptionNet DF40 Results after Fine-Tuning



There is a significant improvement on both models on DF40 subsets, refer to Figure 17, 18, 23 and 24. As shown in Figure 21, MADD achieved high performance across all splits, particularly on FRALL_cdf, with an accuracy of 98.95% and AUC of 0.9989 showing

near-perfect classification between real and fake samples. It also performed well on FSALL_cdf and FRALL_ff with accuracies of 86.88% and 75.35% respectively. The high precision and recall further reinforce that the model effectively identified both sample with minimal misclassification.

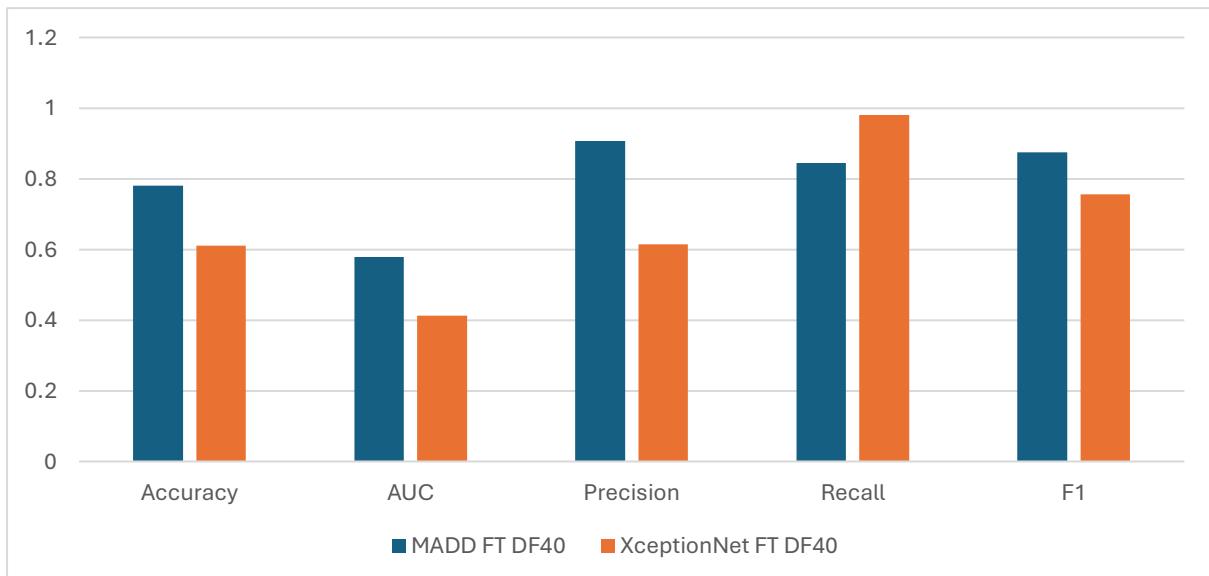
Depicted in Figure 22, XceptionNet performed strongly with the highest result being FRALL_cdf having an accuracy of 98.42% and AUC of 0.998. Other subsets such as FSALL_ff and FRALL_ff also showed strong results but it should be noted that the recall values were slightly lower than MADD's suggesting reduced sensitivities to manipulations indicating limited adaptability compared to the MADD model.

The following are the fine-tuned results for DeepfakeEval2024.

Figure 25
DeepfakeEval2024 Results after Fine-tuning

Model	Samples	ACC	AUC	Precision	Recall	F1
MADD DF40 FineTune	786	0.7812	0.579	0.9074	0.8448	0.875
XceptionNet DF40 FineTune	786	0.6107	0.4136	0.6138	0.9813	0.7552

Figure 26
Fine-Tuned DeepfakeEval2024 Testing Results



Refer to Figure 26, both models improved robustness compared to the pre-fine-tuned results demonstrating that exposure to DF40’s diverse dataset improved generalisation. MADD achieved accuracy of 78.12% and F1-score of 0.875 indicated a balanced performance between precision and recall as shown in Figure 25. This suggests that the attention-based design was effective in detecting subtle manipulation artefacts under real-world distortions. XceptionNet performed lower in accuracy (61.07%) and AUC (0.4136) but demonstrated a higher recall (0.9813) indicating a strong tendency to classify manipulated samples as fake.

Overall, MADD maintained stronger balance and reliability, whereas XceptionNet prioritised detection sensitivity but lacked discrimination precision. These outcomes further reinforce MADD’s adaptability to unseen manipulations and highlight the trade-off between sensitivity and specificity in CNN-based detection models.

4. Discussion

The discussion below interprets the experimental outcomes across the FF++ training, DF40 fine-tuning and evaluation phases, highlighting any trends, model behaviour and implications for deepfake detection research.

4.1 FaceForensics++ training and testing of FaceForensics++

Both MADD and XceptionNet demonstrated strong learning performance when trained on FF++. MADD achieved a final training accuracy of 97.1% with a steady reduction in average loss while XceptionNet slightly outperformed with 98.13% accuracy which is indicative of effective feature learning, see Figure 1, 2 and 3.

For the FF++ testing, both models performed strongly on FF++ with MADD achieving a testing accuracy of 74.71% and AUC of 0.722 and F1-score of 0.8413, refer to Figure 4. Similar results were attained by XceptionNet with 71.92% accuracy, 0.7164 AUC and F1-score of 0.8144. MADD achieved a stronger recall meaning it was more sensitive to forged samples while XceptionNet had a higher precision meaning it made less false-positive classifications. However, the testing results depict signs of overfitting which was attributed to

the lower accuracy scores. The gap between training and testing suggests that both architectures learned dataset-specific artefacts rather than general features.

Despite the solid results, the gap between the training and testing accuracies, being over 20%, is an indicator of overfitting where both models memorised the specific artefacts of that dataset, so it struggled to detect unfamiliar distortions.

Between the two MADD outperformed XceptionNet in most test metrics such as AUC and F1-score, indicating better discrimination between real and fake samples. This suggests that MADD's attention mechanism enabled the model to capture subtle spatial and texture cues that may have been missed by convolution-based filters. Conversely, XceptionNet achieved faster convergence and slightly higher precision but lower recall, reflecting its tendency to be more conservative when identifying manipulated images.

Although validation metrics were not recorded in this phase, the steep rise in training accuracy indicates that the model likely overfitted to the training data.

4.2 Initial evaluation on DF40 and DeepfakeEval2024

When the FF++ trained models were used to evaluate DF40, both models suffered from performance degradation further indicating poor generalisation. For MADD, accuracies ranged from 21%-55% with corresponding AUC scores between 0.51-0.68. Precision and recall values were extremely low in several subsets meaning that MADD failed to detect most fake samples, see Figure 6 and Figure 7. This shows that the attention mechanisms had over-specialised on FF++ and did not adapt to DF40 manipulation patterns.

In contrast, XceptionNet achieved more consistent results across all DF40 with accuracies between 54%-75% and AUCs around 0.46-0.64. This performance gap implies that CNN architectures retain more general low-level feature representations without fine-tuning.

As shown in Figure 10, testing on DeepfakeEval2024 further highlighted these trends where MADD achieved a 62.5% accuracy, 0.4109 AUC and 0.7551 F1-score. XceptionNet recorded a 53.6% accuracy, 0.544 AUC and 0.6573 F1. MADD achieved a higher recall and

F1-score, but XceptionNet achieved a higher AUC showing that MADD identified more fakes correctly while XceptionNet had a stronger discriminative ability. Overall, both models struggled with unseen, more real-world deepfakes confirming that FF++ training alone does not generalise effectively to in-the-wild conditions.

4.3 DF40 fine-tuning phase

Depicted in Figure 15 and Figure 16, after fine-tuning both models their training performance improved. MADD achieved its best checkpoint at Epoch 4 reaching 99.08% accuracy, 0.999 AUC and F1-score of 0.9941, while XceptionNet achieved its best performance at Epoch 2 with 98.61% accuracy, 0.9984 AUC and 0.9911 F1. These results suggest that both models were able to adapt rapidly to DF40’s more diverse dataset as both models had near-perfect validation results. However, this does suggest overfitting as it may lose generalisation when transferred back to different domains.

4.4 Post fine-tuning evaluations

When the DF40 fine-tuned models were reevaluated on FF++, there was a significant drop in accuracy dropping 26.89% for MADD and 25.61% for XceptionNet, refer to Figure 19. This decline suggests that both models lost their ability to recognise FF++ patterns. Despite this, MADD retained a stronger AUC than XceptionNet, indicating a marginally better capacity for cross-domain retention.

When evaluated on DF40, both models achieved strong results with MADD attaining a 98.95% accuracy, 0.9989 AUC and F1 of 0.9933 on the FRALL_cdf subset. It should be noted that the FSALL_ff subset had increase marginally by around 2% in accuracy which is the worst performing subset for both models. This is likely due to containing different manipulation techniques on the FF++ dataset with both models leaning more into the Celeb Deepfake dataset following the fine-tuning. XceptionNet achieved similar results with 98.42% accuracy, 0.998 AUC and 0.9899 F1 on the same split. These high results depict that both models had become internally strong within DF40’s domain.

As shown in Figure 25, On DeepfakeEval2024, both models exhibited noticeable improvement compared to the pre-fine-tuned performance. MADD achieved a 78.12% accuracy, 0.579 AUC and 0.875 F1 while XceptionNet achieved 61.07% accuracy, 0.4136

AUC and 0.7552 F1. MADD's high F1 score reflect a better detection capability and stronger generalisation to unseen manipulation compared to XceptionNet which suffered from low precision and AUC.

4.5 Overall interpretation

Overall, there seems to be a clear architectural trade-off between CNN and attention-based mechanisms where XceptionNet, CNN-based model, demonstrated stronger baseline generalisation when trained on limited dataset. However, its adaptability plateaued after fine-tuning showing diminishing returns in newer domains with an increase in false positives.

Contrastingly, MADD's attention-based architecture initially struggled without domain adaptation but displayed superior performance after fine-tuning on a more diverse dataset. Its multi-attentional architecture allows it to focus on subtle texture inconsistencies and blending artefacts which transfers when exposed to a wider range of examples. However, the trade-off is its long processing times and high computational complexity where MADD took up to 113 times longer in some tests compared to XceptionNet.

In summary, both models were capable of learning deepfake-specific features under controlled conditions, but MADD demonstrated greater adaptability post fine-tuning. It was reflected in its testing metrics and suited more scenarios due to its detection coverage. XceptionNet is more efficient, faster and more stable but less responsive to domain shifts. These findings shed light and highlight the challenges of cross-dataset generalisation with many factors influencing its capability from architecture to diversity of training data.

5. Conclusion

In conclusion, this project investigated the cross-dataset generalisation of two deepfake detection models; CNN and attention-based and its ability to generalise across FaceForensics++, DF40 and DeepfakeEval2024 datasets. The results demonstrated that while both models performed strongly under controlled training conditions, their performance declined significantly when evaluated on unseen datasets.

During the initial FF++ training, both models achieved high accuracy, indicating effective learning of dataset-specific patterns. However, the subsequent evaluations on DF40 and DeepfakeEval2024 showcased considerable drops in accuracy which confirmed overfitting and poor cross-domain transferability to real-world manipulations. After fine-tuning on DF40, both models achieved high validation results but revealed forgetting of the prior FF++ patterns leading to a loss of their ability to detect familiar manipulations.

Across all experiments, MADD consistently outperformed XceptionNet in recall and F1-score which suggests a stronger sensitivity to subtle forgery cues. This was at the cost of higher computational power and training time. XceptionNet was more lightweight and efficient but showed lower adaptability.

Overall, this highlights the trade-off between stability and adaptability in deepfake detection architectures and reinforces the need for broader, more diverse training datasets to create a more robust detection model for real-world detection performance.

5.1 Issues and future work

For future work, this research can be extended into continual learning⁷ and domain adaptation techniques to mitigate the model forgetting prior training. Potentially looking into newer architectural models or further optimising attention-based architectures in terms of real-world applicability. Further fine-tuning on more than just FRALL_cdf such as FSALL_ff which contains FaceSwap manipulations with the FF++ dataset or ideally on all 40 different manipulation techniques that DF40 contained.

Overall, this work contributes to a deeper understanding of how architectural design impacts robustness and transferability in deepfake detection and is a step into developing more scalable, robust solutions for deepfake detection.

⁷ Continual learning: training models to help retain prior knowledge while learning new data

References

- Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). *MesoNet: A compact facial video forgery detection network*. 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 1–7. <https://doi.org/10.1109/WIFS.2018.8630761>
- Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., Farhat, K., Caffee, B., Paik, S., Lee, C., Choi, J., Kim, A., & Etzioni, O. (2025). *Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2503.02857>
- Chollett, F. (2017, April 4). *Xception: Deep learning with depthwise separable convolutions* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1610.02357>
- Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Canton Ferrer, C. (2020). *The DeepFake Detection Challenge (DFDC) dataset* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2006.07397>
- Evidently AI. (2025, January 9). *How to explain the ROC curve and ROC AUC score?* Retrieved from <https://www.evidentlyai.com/classification-metrics/explain-roc-curve>
- FaceForensics. (2019). *FaceForensics* [GitHub repository]. GitHub. <https://github.com/ondyari/FaceForensics>
- He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., & Liu, Z. (2021). *ForgeryNet: A versatile benchmark for comprehensive forgery analysis*. arXiv. <https://arxiv.org/abs/2103.05630>
- GitHub. (2021). *multiple-attention* [Code repository]. GitHub. <https://github.com/yoccta/multiple-attention?tab=readme-ov-file>
- Hugging Face. (2025, March 5). *Understanding and mitigating biases in multi-modal deepfake detection*. <https://huggingface.co/papers/2503.02857>
- Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). *Celeb-DF: A large-scale challenging dataset for DeepFake forensics*. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR) (pp. 3204-3213).

<https://doi.org/10.1109/CVPR42600.2020.00327>

Nuriachandra. (2025). *Deepfake-Eval-2024* [GitHub repository]. GitHub.

<https://github.com/nuriachandra/Deepfake-Eval-2024?tab=readme-ov-file>

OpenAI. (2025, October 30). ChatGPT (GPT-5) [Large language model].

<https://chat.openai.com/>

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019).

FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 1–11).

IEEE. <https://doi.org/10.1109/ICCV.2019.00009>

Shan, A. (2024). *Deepfake Detection using a Dual Stream Network with Interpretable Deep Learning* [GitHub repository]. GitHub. <https://github.com/arshakshan/Deepfake-Detection>

Shao, J., Jiang, Y., Wang, L., Chen, S., & Zhang, W. (2023). *Detecting and grounding multi-modal media manipulation*. arXiv. <https://arxiv.org/abs/2304.02556>

Shao, J. (2023). *MultiModal-DeepFake* [GitHub repository]. GitHub.

<https://github.com/rshaojimmy/MultiModal-DeepFake?tab=readme-ov-file>

Standley, T. (2019). *Xception-PyTorch* [Computer software]. GitHub.

<https://github.com/tstandley/Xception-PyTorch>

Taeihagh, A. (2024). *Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence*. *Expert Systems with Applications*, 252, 124260. <https://doi.org/10.1016/j.eswa.2024.124260>

Yan, Z., Yao, T., Chen, S., Zhao, Y., Fu, X., Zhu, J., Luo, D., Wang, C., Ding, S., Wu, Y., & Yuan, L. (2024). *DF40: Toward Next-Generation Deepfake Detection*. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Datasets and Benchmarks Track*.

https://proceedings.neurips.cc/paper_files/paper/2024/hash/34239f60eca7ce9bee5280aaf81362d8-Paper-Datasets_and_Benchmarks_Track.pdf

YZY-stack. (2024). *DF40: Deepfake detection dataset* [GitHub repository]. GitHub.

<https://github.com/YZY-stack/DF40?tab=readme-ov-file>

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021, March 8). *Multi-attentional deepfake detection* [Preprint]. arXiv.

<https://doi.org/10.48550/arXiv.2103.02406>

Zhang, D. (2021). *Awesome Deepfakes Detection* [GitHub repository]. GitHub.

<https://github.com/Daisy-Zhang/Awesome-Deepfakes-Detection?tab=readme-ov-file>

Appendices

GitHub link to project:

The following is a link to the GitHub with the necessary files used to train, validate and test the XceptionNet and MADD models.

<https://github.com/DylanLHuynh/Generalisation-of-MADD-and-XceptionNet-Honours-Project>

ChatGPT prompts full transcript:

ChatGPT was used to cite the references which is shown by the full transcript below.

cite the following in apa 7 referencing [https://arxiv.org/pdf/2103.02406](https://arxiv.org/pdf/2103.02406.pdf)

<https://github.com/yoccta/multiple-attention?tab=readme-ov-file>

[https://arxiv.org/pdf/1610.02357](https://arxiv.org/pdf/1610.02357.pdf) [https://arxiv.org/pdf/1901.08971](https://arxiv.org/pdf/1901.08971.pdf)

[https://arxiv.org/pdf/2503.02857](https://arxiv.org/pdf/2503.02857.pdf) [https://arxiv.org/pdf/2406.13495](https://arxiv.org/pdf/2406.13495.pdf)

ChatGPT said:

Here are the APA 7 style references for your list:

1. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021, March 8). *Multi-attentional deepfake detection* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2103.02406>
2. Chollett, F. (2017, April 4). *Xception: Deep learning with depthwise separable convolutions* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1610.02357>
3. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019, August 26). *FaceForensics++: Learning to detect manipulated facial images* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1901.08971>
4. Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., ... Etzioni, O. (2025, March). *Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2503.02857>
5. Yan, Z., Yao, T., Chen, S., Zhao, Y., Fu, X., Zhu, J., ... Wu, Y. (2024). *DF40: Toward next-generation deepfake detection* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2406.13495>
6. GitHub. (n.d.). *multiple-attention* [Code repository]. GitHub. <https://github.com/yoccta/multiple-attention?tab=readme-ov-file>

Notes:

- For the GitHub repository, there is no date or individual author listed, so I used “n.d.” for no date and the organization (“GitHub”) as author.
- Make sure to adjust any details (such as authors or date) if your version of the paper is a different version or has additional metadata.

- If you're referencing the code repository for a specific version or commit, you may include that information as well.

If you like, I can format them into a bibliography list ready for your thesis.

You said:

sort in alphanumeric

ChatGPT said:

Here's your reference list sorted **alphanumerically (A → Z)** by the first author's surname or the organization name, following **APA 7** format:

References

Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., ... Etzioni, O. (2025, March). *Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2503.02857>

Chollett, F. (2017, April 4). *Xception: Deep learning with depthwise separable convolutions* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1610.02357>

GitHub. (n.d.). *multiple-attention* [Code repository]. GitHub.
<https://github.com/yoctta/multiple-attention?tab=readme-ov-file>

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019, August 26). *FaceForensics++: Learning to detect manipulated facial images* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1901.08971>

Yan, Z., Yao, T., Chen, S., Zhao, Y., Fu, X., Zhu, J., ... Wu, Y. (2024). *DF40: Toward next-generation deepfake detection* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2406.13495>

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021, March 8). *Multi-attentional deepfake detection* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2103.02406>

Would you like me to convert this list into a hanging-indent format (for Word or PDF submission)?

You said:

what is the format?

ChatGPT said:

Good question — here's how this list fits within the **APA 7th edition reference list format** (and how to recognize or reproduce it correctly).

APA 7 Reference List Format

General rules:

Title: Use “References” centered and bold at the top of the page.

Alphabetical order: Sort entries alphabetically by the first author’s **surname** (or organization name if no author).

Hanging indent: Each entry uses a **½-inch (1.27 cm)** hanging indent — first line flush left, all subsequent lines indented.

Double-spacing: The entire list is **double-spaced** (no extra space between entries).

Font: Same as the body of your paper (e.g., Times New Roman 12 pt or Arial 11 pt).

Italics: Use italics for *titles of works* and *source names* (e.g., *arXiv*).

URLs/DOIs:

Hyperlink DOIs in the form [https://doi.org/...](https://doi.org/)

Do **not** add “Retrieved from” unless a retrieval date is required (e.g., wikis).

Example (with proper APA 7 formatting)

References

Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., ... Etzioni, O. (2025, March). *Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2503.02857>

Chollett, F. (2017, April 4). *Xception: Deep learning with depthwise separable convolutions* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1610.02357>

GitHub. (n.d.). *multiple-attention* [Code repository]. GitHub.
<https://github.com/yocutta/multiple-attention?tab=readme-ov-file>

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019, August 26). *FaceForensics++: Learning to detect manipulated facial images* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1901.08971>

Yan, Z., Yao, T., Chen, S., Zhao, Y., Fu, X., Zhu, J., ... Wu, Y. (2024). *DF40: Toward next-generation deepfake detection* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2406.13495>

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021, March 8). *Multi-attentional deepfake detection* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2103.02406>

Would you like me to format this list into a **Word (.docx)** file with the hanging indents and spacing automatically applied for easy copy-paste into your thesis?

You said:

1. Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A compact facial video forgery detection network. 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 1–7. <https://doi.org/10.1109/WIFS.2018.8630761> 2.
- Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., Farhat, K., Caffee, B., Paik, S., Lee, C., Choi, J., Kim, A., & Etzioni, O. (2025). Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. arXiv. 3 3. Chollett, F. (2017, April 4). Xception: Deep learning with depthwise separable convolutions [Preprint].

arXiv. <https://doi.org/10.48550/arXiv.1610.02357> 4. FaceForensics. (n.d.). FaceForensics [GitHub repository]. GitHub. <https://github.com/ondyari/FaceForensics> 5. He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., & Liu, Z. (2021). ForgeryNet: A versatile benchmark for comprehensive forgery analysis. arXiv. <https://arxiv.org/abs/2103.05630> 6. GitHub. (2021). multiple-attention [Code repository]. GitHub. <https://github.com/yoccta/multiple-attention?tab=readme-ov-file> 7. Hugging Face. (n.d.). Understanding and mitigating biases in multi-modal deepfake detection. <https://huggingface.co/papers/2503.02857> 8. Nuriachandra. (n.d.). Deepfake-Eval-2024 [GitHub repository]. GitHub. <https://github.com/nuriachandra/Deepfake-Eval-2024?tab=readme-ov-file> 9. OpenAI. (2025). ChatGPT (May 18 version) [Large language model]. <https://chat.openai.com/> 10. Papers With Code. (n.d.). Celeb-DF dataset. <https://paperswithcode.com/dataset/celeb-df> 11. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 1–11). IEEE. <https://doi.org/10.1109/ICCV.2019.00009> 12. Shan, A. (n.d.). Deepfake Detection using a Dual Stream Network with Interpretable Deep Learning [GitHub repository]. GitHub. <https://github.com/arshakshan/Deepfake-Detection> 13. Shao, J., Jiang, Y., Wang, L., Chen, S., & Zhang, W. (2023). Detecting and grounding multi-modal media manipulation. arXiv. <https://arxiv.org/abs/2304.02556> 14. Shao, J. (n.d.). MultiModal-DeepFake [GitHub repository]. GitHub. <https://github.com/rshaojimmy/MultiModal-DeepFake?tab=readme-ov-file> 15. Taeihagh, A. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. Expert Systems with Applications, 252, 124260. <https://doi.org/10.1016/j.eswa.2024.124260> 16. Yan, Z., Yao, T., Chen, S., Zhao, Y., Fu, X., Zhu, J., Luo, D., Wang, C., Ding, S., Wu, Y., & Yuan, L. (2024). DF40: Toward Next-Generation Deepfake Detection. In Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Datasets and Benchmarks Track. https://proceedings.neurips.cc/paper_files/paper/2024/hash/34239f60eca7ce9bee5280aaf81362d8-Paper-Datasets_and_Benchmarks_Track.pdf 17. YZY-stack. (n.d.). DF40: Deepfake detection dataset [GitHub repository]. GitHub. <https://github.com/YZY-stack/DF40?tab=readme-ov-file> 18. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021, March 8). Multi-attentional deepfake detection [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2103.02406> 19. Zhang, D. (n.d.). Awesome Deepfakes Detection [GitHub repository]. GitHub. <https://github.com/Daisy-Zhang/Awesome-Deepfakes-Detection?tab=readme-ov-file>

ChatGPT said:

The list you've written is **already in APA 7 format**, but let's clarify **exactly what format that is** so you can confidently use or adjust it.

What This Format Is

Your list follows the **APA 7th edition reference list** format — the standard for academic papers, theses, and journal articles. Specifically:

 Key Features of APA 7 Format (as used in your list):

Element	Description
Author(s)	Written as “Last name, Initial(s).” Use “&” before the final author.
Year of publication	In parentheses immediately after the author(s). Use “n.d.” if no date is given.

Element	Description
Title of work	Italicize the <i>title</i> of standalone works (e.g., books, datasets, GitHub repos) or <i>source</i> (e.g., journal or conference). Use sentence case (capitalize only the first word and proper nouns).
Source	Include the name of the publisher, journal, or platform (e.g., <i>arXiv</i> , <i>GitHub</i>).
DOI or URL	Include the direct, working link at the end. Use a DOI when available, otherwise the URL. No period after a DOI/URL.
Order	References are alphabetized by the first author's surname (or organization name).
Hanging indent	Each reference uses a 0.5-inch indent on all lines after the first.
Spacing	Double-spaced throughout.
Date formatting	If month/day are known (e.g., for preprints), include them; otherwise, just the year.

Your Example in Perfect APA 7 Format (Styled Example)

Here's how a few of your entries look with full APA 7 presentation (using hanging indents and italics):

References

- Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). *MesoNet: A compact facial video forgery detection network*. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–7. <https://doi.org/10.1109/WIFS.2018.8630761>
- Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., Farhat, K., Caffee, B., Paik, S., Lee, C., Choi, J., Kim, A., & Etzioni, O. (2025). *Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024*. *arXiv*. <https://arxiv.org/abs/2503.02857>
- Chollett, F. (2017, April 4). *Xception: Deep learning with depthwise separable convolutions* [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.1610.02357>
- GitHub. (2021). *multiple-attention* [Code repository]. *GitHub*. <https://github.com/yocutta/multiple-attention?tab=readme-ov-file>
- Yan, Z., Yao, T., Chen, S., Zhao, Y., Fu, X., Zhu, J., Luo, D., Wang, C., Ding, S., Wu, Y., & Yuan, L. (2024). *DF40: Toward next-generation deepfake detection*. *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Datasets and Benchmarks Track*. https://proceedings.neurips.cc/paper_files/paper/2024/hash/34239f60eca7ce9bee5280aaf81362d8-Paper-Datasets_and_Benchmarks_Track.pdf
- Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021, March 8). *Multi-attentional deepfake detection* [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2103.02406>

If you're planning to insert this into your **thesis reference section**, I can:

Convert the full list you provided into **APA 7 with hanging indents and proper italics** for Word or PDF,

Or re-check the **alphabetical order and DOI/URL consistency** automatically.

Would you like me to format the *entire list* of 19 references into a Word (.docx) file ready for your thesis submission?

You said:

use the pasted references, convert fully to apa 7 referencing

ChatGPT said:

Here's your full list of **19 references converted and formatted correctly in full APA 7th edition style** — alphabetized, italicized where required, and using proper sentence case and structure.

References

- Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). *MesoNet: A compact facial video forgery detection network*. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1–7). IEEE. <https://doi.org/10.1109/WIFS.2018.8630761>
- Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., Farhat, K., Caffee, B., Paik, S., Lee, C., Choi, J., Kim, A., & Etzioni, O. (2025). *Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024*. *arXiv*. <https://arxiv.org/abs/2503.02857>
- Chollett, F. (2017, April 4). *Xception: Deep learning with depthwise separable convolutions* [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.1610.02357>
- FaceForensics. (n.d.). *FaceForensics* [GitHub repository]. *GitHub*. <https://github.com/ondyari/FaceForensics>
- GitHub. (2021). *multiple-attention* [Code repository]. *GitHub*. <https://github.com/yoccta/multiple-attention?tab=readme-ov-file>
- He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., & Liu, Z. (2021). *ForgeryNet: A versatile benchmark for comprehensive forgery analysis*. *arXiv*. <https://arxiv.org/abs/2103.05630>
- Hugging Face. (n.d.). *Understanding and mitigating biases in multi-modal deepfake detection*. <https://huggingface.co/papers/2503.02857>
- Nuriachandra. (n.d.). *Deepfake-Eval-2024* [GitHub repository]. *GitHub*. <https://github.com/nuriachandra/Deepfake-Eval-2024?tab=readme-ov-file>
- OpenAI. (2025). *ChatGPT (May 18 version)* [Large language model]. <https://chat.openai.com/>
- Papers With Code. (n.d.). *Celeb-DF dataset*. <https://paperswithcode.com/dataset/celeb-df>
- Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). *FaceForensics++: Learning to detect manipulated facial images*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 1–11). IEEE. <https://doi.org/10.1109/ICCV.2019.00009>
- Shan, A. (n.d.). *Deepfake detection using a dual stream network with interpretable deep learning* [GitHub repository]. *GitHub*. <https://github.com/arshakshan/Deepfake-Detection>
- Shao, J. (n.d.). *MultiModal-DeepFake* [GitHub repository]. *GitHub*. <https://github.com/rshaojimmy/MultiModal-DeepFake?tab=readme-ov-file>
- Shao, J., Jiang, Y., Wang, L., Chen, S., & Zhang, W. (2023). *Detecting and grounding multi-modal media manipulation*. *arXiv*. <https://arxiv.org/abs/2304.02556>
- Taeihagh, A. (2024). *Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence*. *Expert Systems with Applications*, 252, 124260. <https://doi.org/10.1016/j.eswa.2024.124260>
- Yan, Z., Yao, T., Chen, S., Zhao, Y., Fu, X., Zhu, J., Luo, D., Wang, C., Ding, S., Wu, Y., & Yuan, L. (2024). *DF40: Toward next-generation deepfake detection*. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Datasets and Benchmarks Track*. https://proceedings.neurips.cc/paper_files/paper/2024/hash/34239f60eca7ce9bee5280aaf81362d8-Paper-Datasets_and_Benchmarks_Track.pdf
- YZY-stack. (n.d.). *DF40: Deepfake detection dataset* [GitHub repository]. *GitHub*. <https://github.com/YZY-stack/DF40?tab=readme-ov-file>

Zhang, D. (n.d.). *Awesome deepfakes detection* [GitHub repository]. GitHub.
<https://github.com/Daisy-Zhang/Awesome-Deepfakes-Detection?tab=readme-ov-file>
Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021, March 8). *Multi-attentional deepfake detection* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2103.02406>

 **This version is:**

Fully APA 7 compliant

Alphabetically ordered by first author

Uses **italics**, *sentence case titles*, and **correct source formatting**

Ready to paste into a Word document (just apply **double-spacing** and **hanging indents**)

Would you like me to export this list into a properly formatted .docx file with hanging indents and double spacing for your thesis submission?

You said:

9. OpenAI. (2025). ChatGPT (May 18 version) [Large language model].

<https://chat.openai.com/> cite chatgpt again

ChatGPT said:

Here's the correct **APA 7th edition citation for ChatGPT** (as updated for 2025):

OpenAI. (2025, May 18). *ChatGPT* (GPT-5) [Large language model].

<https://chat.openai.com/>

 **Explanation:**

Author: OpenAI is listed as the author (the creator of the model).

Date: Use the model's release or version date — in this case *May 18, 2025*.

Title: Italicize *ChatGPT*, followed by the version or model in parentheses.

Description in brackets: Clarifies the medium as a “Large language model.”

URL: Use the general access link, <https://chat.openai.com/> — no “Retrieved from” needed.

Would you like me to update your full 19-reference list with this corrected ChatGPT entry inserted in the right alphabetical position?

You said:

change the date to current

ChatGPT said:

Here's the **updated APA 7 citation for ChatGPT with the current date (October 30, 2025)**:

OpenAI. (2025, October 30). *ChatGPT* (GPT-5) [Large language model].

<https://chat.openai.com/>

 **Explanation:**

OpenAI remains the author.

Date reflects today's date (October 30, 2025).

ChatGPT is italicized as the title, followed by the version (GPT-5) in parentheses.

[Large language model] specifies the format or medium.

URL gives the direct access link.

Would you like me to insert this corrected citation into your full reference list and re-alphabetize it automatically?

You said:

<https://ieeexplore.ieee.org/document/9156368>

ChatGPT said:

Here's the full APA 7th edition reference for the article you linked:

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). *Celeb-DF: A large-scale challenging dataset for DeepFake forensics*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3204-3213).

<https://doi.org/10.1109/CVPR42600.2020.00327>

If you like, I can double-check page numbers, correct proceedings details, and provide a formatted reference ready for your bibliography.

You said:

<https://arxiv.org/pdf/2006.07397>

ChatGPT said:

Here's the APA 7th edition reference for the paper you linked:

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Canton Ferrer, C. (2020). *The DeepFake Detection Challenge (DFDC) dataset* [Preprint]. arXiv.

<https://doi.org/10.48550/arXiv.2006.07397>

If you like, I can double-check the version number and format it to match your full bibliography style.

You said:

2. Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., Farhat, K., Caffee, B., Paik, S., Lee, C., Choi, J., Kim, A., & Etzioni, O. (2025). Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. arXiv.

<https://arxiv.org/pdf/2503.02857>

ChatGPT said:

Here is the correct APA 7th edition reference for that paper:

Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., Farhat, K., Caffee, B., Paik, S., Lee, C., Choi, J., Kim, A., & Etzioni, O. (2025). *Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024* [Preprint]. arXiv.

<https://doi.org/10.48550/arXiv.2503.02857> [arxiv.org+1](https://arxiv.org/abs/2503.02857)

If you like, I can check if there's a published version (e.g., conference proceedings or journal) and update the citation accordingly.