Hybrid RNN to Differentiate Deepfakes and Real Images/Videos

Project By:
Piyush S Wandile
&
Debaditya Paul

Started on:
March 2025

Completed On:
May 2025

Hybrid RNN to Differentiate Deepfakes and Real Images/Videos

Abstract:

With the rise of generative deep learning techniques, particularly Generative Adversarial Networks (GANs), the proliferation of hyper-realistic synthetic media—commonly known as deepfakes—has introduced significant challenges to digital media authenticity. This project presents a robust deepfake detection framework that leverages a diverse ensemble of state-of-the-art models, including XceptionNet, LSTM, GRU, Vision Transformers (ViT), and EfficientNet, trained on varied datasets such as Zenodo, Celeb-DF, and FaceForensics++. After evaluating each model independently, optimal performers are integrated using a weighted ensemble strategy, yielding high recall and generalization capability on unseen data. The proposed system surpasses existing tools like DF Detect and Deepware Scanner in terms of detection accuracy and robustness, demonstrating its suitability for real-world deployment in digital forensics and content verification applications.

Introduction:

The advent of Artificial Intelligence has led to the creation of highly realistic synthetic media, commonly referred to as deepfakes. Generated using advanced neural network architectures such as Generative Adversarial Networks (GANs), autoencoders, and other deep learning techniques, deepfakes can convincingly manipulate or replace human faces in images and videos. These sophisticated pipelines typically involve face detection, alignment, encoding into a latent space, and decoding or blending manipulated outputs back into source media. While this technology opens up new possibilities in entertainment and creativity, it also raises serious concerns regarding security, privacy, and information integrity.

Detecting deepfakes has therefore become a critical area of research. In this project, a comprehensive deepfake detection system is developed by leveraging a range of state-of-the-art machine learning models, including XceptionNet, Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRU), Vision Transformers (ViT), and EfficientNet. Each model is trained independently on specialized datasets — Zenodo, Celeb-DF, and FaceForensics++ — ensuring robustness across different types of manipulated media. A separate unseen Testing Dataset is reserved for final evaluation to assess the true generalizability of the models.
After individual model training and testing, the best-performing models will be selected and combined using an appropriate ensemble technique to enhance overall detection accuracy and reliability. This approach aims to build a powerful and adaptable deepfake detection system capable of performing effectively across diverse and challenging real-world scenarios.

Datasets Used:

1. **Zenodo Deepfake Dataset**
   The Zenodo Deepfake Dataset is a comprehensive resource designed to support both forgery detection and segmentation tasks. It offers detailed, face-wise annotations, enabling precise analysis of manipulated facial regions. The dataset encompasses a diverse array of challenging, in-the-wild scenarios, enhancing the robustness of detection algorithms. It comprises approximately 95,000 real and 95,000 fake images, systematically divided into training, validation, and test subsets, providing a balanced and extensive sample space for model development.

2. **FaceForensics++**
   FaceForensics++ is a large-scale dataset and benchmark tailored for detecting manipulated facial images. It consists of 1,000 original video sequences sourced from YouTube, each manipulated using four automated face manipulation methods: Deepfakes, Face2Face, FaceSwap, and Neural Textures. The dataset includes over 1.8 million manipulated images, offering a substantial volume of data for training and evaluating detection models. The videos feature mostly frontal faces without occlusions, facilitating the generation of realistic forgeries. FaceForensics++ addresses societal concerns over synthetic image generation and manipulation, such as the spread of misinformation and erosion of trust in digital content.

3. **Celeb-DF**
   Celeb-DF is a challenging dataset developed to assess the effectiveness of DeepFake detection methods. It comprises 590 real videos of celebrities collected from YouTube, representing diverse ages,

4. ethnicities, and genders, along with 5,639 corresponding DeepFake videos. The DeepFake videos are generated using an improved synthesis process, resulting in high visual quality that closely resembles

   DeepFakes found online. This dataset serves as a benchmark for evaluating detection algorithms in real-world scenarios, reflecting the complexities and diversity inherent in such environments.

5. **140K Real and Fake Faces Dataset**
   The 140K Real and Fake Faces Dataset is designed for training and evaluating deepfake detection models on static images. It comprises 70,000 real face images sourced from the Flickr-Faces-HQ (FFHQ) dataset and 70,000 synthetic face images generated using StyleGAN, sampled from the "1 million Fake Faces" dataset. All images are resized to 256x256 pixels, ensuring uniformity. The fake faces exhibit varying levels of realism, with some nearly indistinguishable from real ones. The real faces cover a wide range of ages, ethnicities, and lighting conditions, providing a diverse and balanced collection for robust model training.

Models Utilized:

1. **XceptionNet**
   Architecture: XceptionNet is a 36-layer CNN replacing traditional Inception modules with depth wise separable convolutions: first a 1×1 pointwise convolution, then per-channel depth wise convolutions. This separation improves efficiency by independently learning cross-channel and spatial correlations.
   Deepfake Detection Use: Pre-trained XceptionNet models are fine-tuned on deepfake datasets, extracting detailed facial feature maps. Its stability and strong feature extraction have delivered accuracy up to 99.65% on video benchmarks. It is particularly good at identifying subtle inconsistencies like unnatural textures, blurring, or local artifacts.

2. **EfficientNet**
   Architecture: EfficientNet introduces a compound scaling method to uniformly scale network depth, width, and resolution using a coefficient φ. It is built upon MBConv blocks (MobileNetV2) with squeeze-and-excitation optimizations for balancing high accuracy and computational efficiency across 340 layers.
   Deepfake Detection Use: EfficientNet (e.g., B0 to B4) models serve as lightweight yet highly accurate CNN backbones for deepfake detection. They outperform traditional CNNs on datasets like FF++ and Celeb-DF after fine-tuning, enabling faster training with fewer parameters without sacrificing performance.

3. **LSTM**
   Architecture: An LSTM unit maintains a cell state and uses three gates — input, forget, and output — to regulate the flow of information. This design helps preserve long-term dependencies and prevent vanishing/exploding gradient problems.
   Deepfake Detection Use: In deepfake pipelines, LSTMs process frame-level feature vectors extracted from CNNs. They detect temporal inconsistencies such as sudden or unnatural changes in facial expressions, blinking, or mouth movement transitions across frames, help classify videos as real or fake.

4. **GRU**
   Architecture: GRUs simplify LSTM by merging the forget and input gates into a single update gate, reducing complexity while maintaining strong performance on sequence data. Frame-wise features from CNNs or ViTs are input sequentially into GRUs.
   Deepfake Detection Use: GRUs are efficient for modeling temporal dependencies like blinking rates, lip synchronization, and head motion continuity. They enhance classification by detecting motion anomalies across frames, often offering faster convergence and lower computational load compared to LSTMs.

5. **ViT**
   Architecture: ViTs divide an image into fixed-size patches, flatten them, add positional embeddings, and process the sequence through standard Transformer encoder layers (multi-head self-attention + feedforward networks). You can think of it like applying NLP-style tokenization to images.

Deepfake Detection Use: Deepfakes often introduce subtle inconsistencies in textures and global patterns. ViTs, with their global receptive field and self-attention mechanism, are excellent at spotting such fine-grained, distributed artifacts in frames, making them powerful for frame-level classification.

Literature Survey:

Sun et al., in their research on a heterogeneous feature ensemble learning based deepfake detection method introduced a technique to detect fake face images, particularly those generated by various deepfake models. They noted that deep-fake, while having applications, poses significant public security problems. The study addressed the challenges of improving detection accuracy and generalization. Their methodology involved extracting three heterogeneous features: facial landmark points, facial spectrum, and texture, representing them using a histogram of grey gradients and co-occurrence matrix features. These features were integrated into an ensemble vector and sent to a back-propagation neural network classifier. Experimental results demonstrated that their approach achieved better detection accuracy (97.04%) compared to several state-of-the-art methods and improved detection accuracy for images from unknown models. They concluded that the approach improves accuracy and generalization by integrating heterogeneous features. Future work includes applying the method to other image manipulation and forensic techniques. [1]

Jerry John and Bismin V. Sherif conducted a comprehensive study on deepfake detection, addressing the challenges posed by digitally manipulated media in social, political, and personal contexts. The authors highlighted the difficulty in manually distinguishing deepfakes due to advancements in deep learning. Their methodology involved a comparative analysis of feature-based, temporal-based, and deep feature-based detection techniques, alongside proposing a semi-supervised GAN (SGAN) architecture. The SGAN model combined labelled and unlabeled data, leveraging a discriminator trained for both supervised classification and unsupervised feature learning. Experimental results demonstrated an accuracy of 92.3% on a dataset of 40,000 images, outperforming existing methods. The study concluded that SGAN-based detection is effective in mitigating deep-fake threats, with future scope for integrating more advanced architectures and larger datasets. [2]

Garg and Gill in their exploratory study on deepfake generation and detection, explored how deepfakes, created through deep learning, manipulate content with high realism, posing risks like tarnishing reputations. They noted the challenge in distinguishing real from fake and discussed that existing detection models struggle with generalizability across different datasets and keeping up with rapidly changing creation methods. The paper reviewed diverse creation techniques like autoencoders and GANs, and detection methods such as CNN, LSTM, and Transfer Learning. They mentioned a technique exploiting audio-visual disparity and reviewed benchmark datasets. Their review indicated deep learning algorithms show higher accuracy than machine learning for detection. They concluded that responsible use of deep-fake technology is essential due to potential misuse. Future work aims to develop a highly accurate and robust detection model. [3]

Kaushal et al., in their comparative study on deep-fake detection algorithms, examined the increasing threat of realistic deepfakes created using deep learning and the urgent need for effective detection methods. They discussed the dangers posed to democracy, security, and privacy. The paper provided a review of various recent deepfake detection techniques, highlighting their methodologies and comparing their performance. Methodologies covered included clustering-based embedding, successive subspace learning, GAN-based analysis, eye blinking detection, and shallow neural networks. They presented a comparison table showing reported accuracies of different methods on various datasets, such as UADFV, Celeb-DF, and FaceForensics++, with accuracies varying widely (e.g., DefakeHop at 100% on UADFV, MesoNet at 98% on online deepfake videos). They concluded that their paper offers a comparative analysis of techniques, exploring their benefits and drawbacks, and aims to provide insights for future research as creation and detection methods continue to evolve. [4]

Dazhuang Liu et al. proposed a novel deepfake detection method to address generalization challenges in real-world scenarios. The study focused on continuous frame face-swapping, a gap in existing single-frame detection approaches. The methodology employed Delaunay triangulation and piecewise affine transformation to generate realistic face-swap videos, followed by a feature enhancement module masking facial and background noise. The detection model combined EfficientNet for intra-frame features and LSTM for temporal analysis. Cross-domain experiments on datasets like FaceForensics++ and Celeb-DF achieved an AUC of 84.38%, surpassing methods like LipForensics (82.4%) and Xception (73.7%). The authors concluded that their approach effectively

leveraged spatio-temporal features, offering robust generalization for cross-dataset detection, with potential for further refinement in dynamic video contexts. [5]

Joshi and Sinha in their study integrating GLCM texture analysis for improved deepfake detection, addressed the challenges in digital media authentication due to deepfakes. Their research focused on using Grey Level Co-occurrence Matrix (GLCM) texture analysis to enhance detection accuracy by identifying texture inconsistencies in manipulated images. The methodology involved extracting specific GLCM texture features, 'Contrast' and 'Dissimilarity', to quantify these inconsistencies. The method was evaluated on the Celeb-DF (v2) dataset. Experimental results showed that GLCM texture analysis is a powerful method for identifying deepfakes, with analysis of the Dissimilarity feature showing a clear difference between genuine and fake faces. They concluded the method provided significant improvements in detection accuracy compared to existing techniques on the dataset. The study contributes insights into developing robust detection frameworks. [6]

Zhang et al. in their paper a two-branch deep-fake detection network based on improved Xception addressed the challenges faced by existing deep-fake detection methods, particularly their low accuracy in low-definition videos and poor generalization across datasets. They noted that current AI face-to-face technologies are evolving rapidly. Their proposed methodology employed a two-branch network structure. One branch focused on the whole video using improved Xception with CBAM and GRU, while the other, a local branch, detected each frame using improved Xception with CBAM and data augmentations like Face-Cutout. The results from both branches were combined. Evaluations on FF++ and Celeb-DF datasets showed the method had better detection performance and generalization than several other methods. They concluded that the two-branch structure, CBAM, and data enhancement improve detection capacity and generalization, making their method superior to existing ones in handling low-quality and cross-archive scenarios. [7]

Trung et al. in their work on a hybrid Xception-LSTM model for deepfake video detection discussed how generative models have made fabricating realistic visual content easier, leading to the misuse of deepfake technology. They identified challenges where existing detection methods struggle with varied manipulations, preprocessing effects, and generalization across diverse data. Their proposed methodology involved a hybrid model combining the Convolutional Block Attention Module (CBAM) for channel and spatial attention, XceptionNet for spatial artefacts and intrinsic data, and LSTM for temporal dependencies. The model was evaluated on the Div-DF dataset, which includes various manipulations. Results showed the proposed model handled diversified scenarios well, achieving an AUC of 0.9855 and beating several other models on the dataset. They concluded that the model effectively captures spatial and temporal discrepancies and shows good generalization. Future work involves exhaustive testing on fully synthetic frames and those with preprocessing. [8]

Atas et al., in their research on a new approach to the ensemble method for deepfake detection discussed the growing threats posed by image forgery created using technology and deep learning. They highlighted the ongoing race between malicious use and detection methods. The study addressed the challenge of detecting realistic forgery and reducing false detection rates. Their proposed method involved using a D-CNN model for feature extraction from images, feeding these features into statistical algorithms like SVM, Random Forest, and Logistic Regression, and finally applying an Ensemble method to combine the results for a final estimation. Experimental results on a custom dataset showed an accuracy of 82.55%, with the main goal being to support detection precision and reduce misdirection rather than solely increasing accuracy. They concluded the ensemble method provided clearer results and reduced false detection. Future work aims to include more methods in the ensemble and use more up-to-date technologies. [9]

Proposed Methodology:

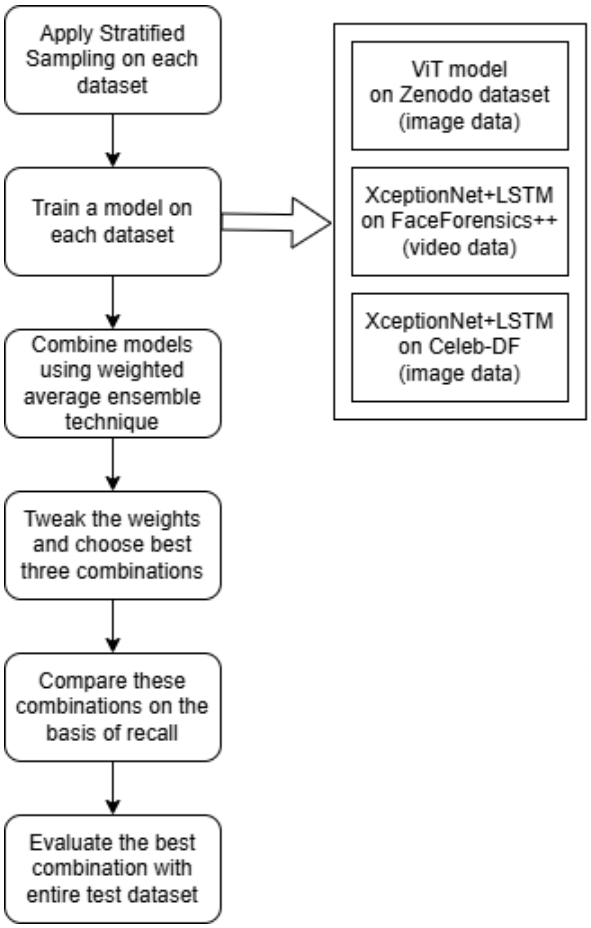

Figure 1: Flowchart to design the ensemble model

Training individual models:

1. ViT Model on Zenodo Dataset (individual images)

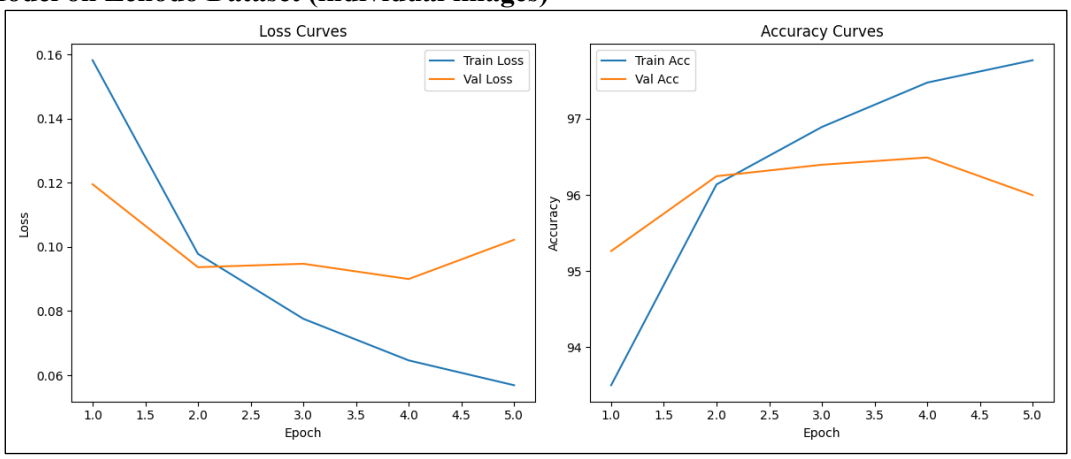

Figure 2: Loss & Accuracy curves for Model 1

Loss Curve:

- Training loss decreases steadily, which is a good sign of learning.
- Validation loss decreases initially but slightly increases at the end, hinting at mild overfitting.

Accuracy Curve:

- Training accuracy improves consistently.
- Validation accuracy improves initially but then plateaus and slightly drops, confirming overfitting from around epoch 4.

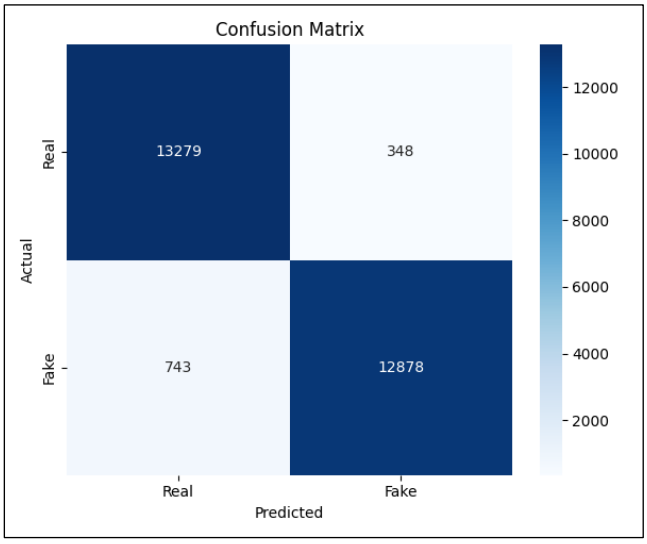

Figure 3: Confusion Matrix for Model 1

```
Classification Report:

              precision    recall  f1-score   support

        Real       0.95      0.97      0.96     13627
        Fake       0.97      0.95      0.96     13621

    accuracy                           0.96     27248
   macro avg       0.96      0.96      0.96     27248
weighted avg       0.96      0.96      0.96     27248
```

Confusion Matrix:

- Real Predicted Correctly: 13,279
- Real Predicted as Fake: 348
- Fake Predicted Correctly: 12,878
- Fake Predicted as Real: 743
- Overall: Very strong performance.
- Misclassification Rate: Low, especially for real images.

Excellent balance and accuracy. A bit more false positives for fake images, but still highly reliable.

2. XceptionNet + LSTM on FaceForensics++ Dataset (.mp4 files)

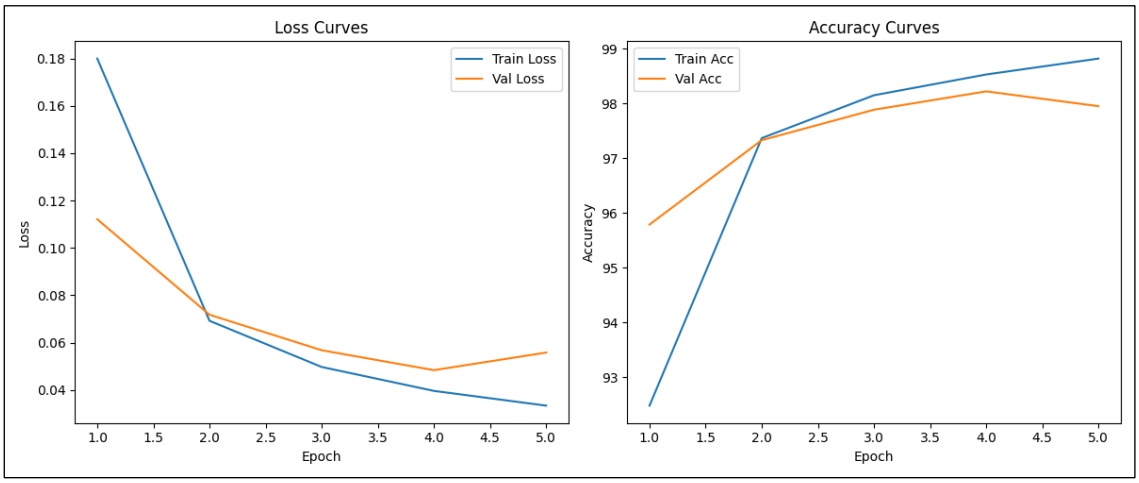

Figure 4: Loss & Accuracy curves for Model 2

Loss Curve:

- Both training and validation loss decrease smoothly, showing strong learning.
- Slight uptick in validation loss at the last epoch, but not severe.

Accuracy Curve:

- Training and validation accuracy are closely aligned and both improve steadily.

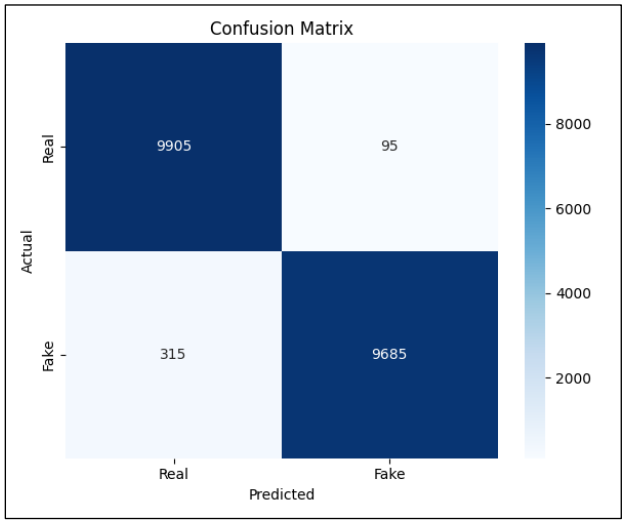

Figure 5: Confusion Matrix for Model 2

```
Classification Report:

              precision    recall  f1-score   support

        Real       0.97      0.99      0.98     10000
        Fake       0.99      0.97      0.98     10000

    accuracy                           0.98     20000
   macro avg       0.98      0.98      0.98     20000
weighted avg       0.98      0.98      0.98     20000
```

Confusion Matrix:

- Real Correct: 9,905
- Real as Fake: 95
- Fake Correct: 9,685
- Fake as Real: 315
- Overall: Excellent results with very few misclassifications.

This is a high-performing model. It shows stronger detection accuracy than the first for real images, and slightly better precision for fake detection.

3. XceptionNet + LSTM on Celeb-DF Dataset (extracted sequential images)

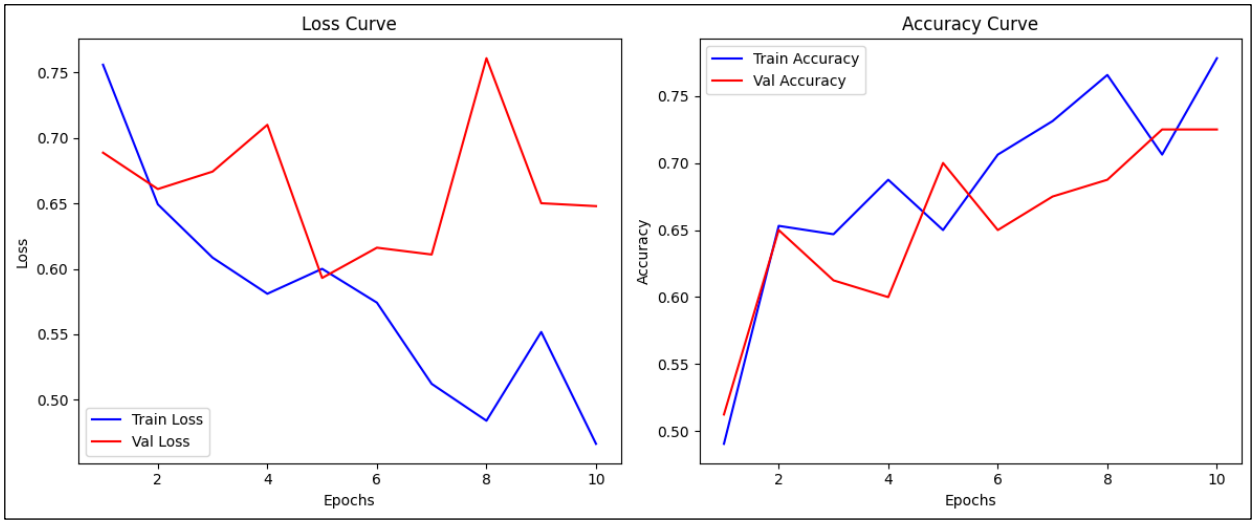

Figure 6: Loss & Accuracy curves for Model 3

Loss Curve:

- Training loss decreases overall but with noise.
- Validation loss is erratic and doesn't follow the training trend, showing instability.

Accuracy Curve:

- Training accuracy improves with some fluctuation.
- Validation accuracy is volatile and less reliable.

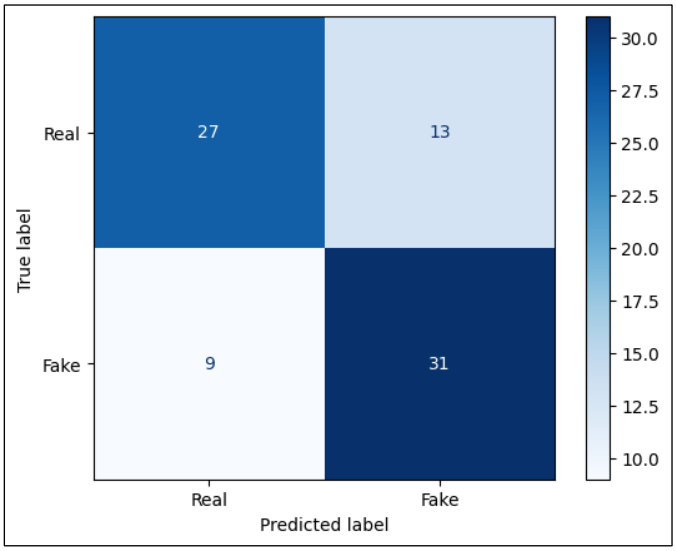

Figure 7: Confusion Matrix for Model 3

Classification Report:

	precision	recall	f1-score	support
Real	0.75	0.68	0.71	40
Fake	0.70	0.78	0.74	40
accuracy			0.72	80
macro avg	0.73	0.73	0.72	80
weighted avg	0.73	0.72	0.72	80

Confusion Matrix:

- Real Correct: 27
- Real as Fake: 13
- Fake Correct: 31
- Fake as Real: 9
- Overall: Decent small-scale performance, but the false positive rate (especially for real → fake) is high.

Performance is weaker here, possibly due to limited data or a test set not representative of the overall distribution. Could benefit from further training or augmentation.

Trials with different weight combinations:

Table 1: Accuracy for different weight combinations

ViT	XceptionNet (on .mp4 files)	XceptionNet (on sequential images)	Accuracy
2	3	0	53.5
1	2	0	53.5
0	1	0	53.5
2	2	0	49.5
1	1	0	49.5
3	2	1	48.5
1	0	1	47
2	2	1	46.5
1	1	1	46.5
0	1	1	46
6	1	8	45
5	3	1	45
1	0	0	45
0	0	1	45

Upon applying the weighted average ensemble prediction technique with various weight configurations, it was observed that the combinations (2,3,0), (1,2,0), and (0,1,0) yielded equivalent performance in terms of overall accuracy. Consequently, the selection of the optimal model configuration should further consider the individual performance characteristics of the constituent models.
In the context of deepfake detection, recall is deemed a more critical evaluation metric than overall accuracy. This preference arises because the misclassification of a real image as fake constitutes a tolerable error, whereas the misclassification of a fake image as real presents a significant risk and is thus unacceptable. Prioritizing recall ensures that the model minimizes the occurrence of such critical misclassifications.

Trials on individual models:

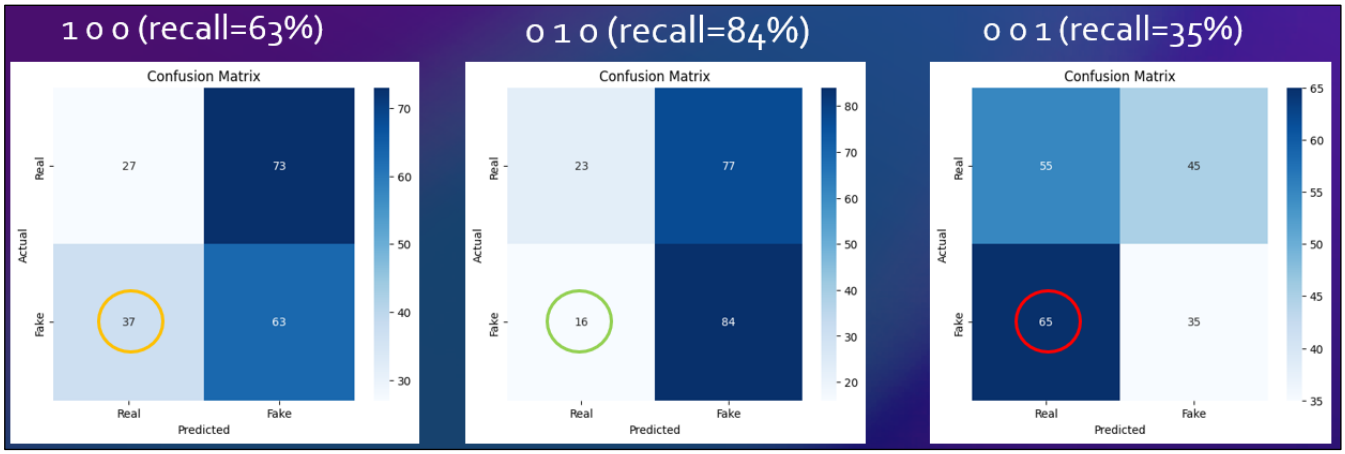

Figure 8: Recall comparison of individual models

Upon evaluating the models using recall as the primary metric, it was observed that the second model achieved the highest recall, recording a value of 84%. This indicates its superior capability in correctly identifying fake instances, which is critical in the context of deepfake detection where false negatives must be minimized. The first model, with a recall of 63%, also demonstrated reasonably good performance and was considered suitable for ensemble integration. In contrast, the third model exhibited a recall of only 35%, indicating a substantial deficiency in its ability to detect fake instances accurately. Given the low recall, inclusion of the third model could compromise the reliability of the final ensemble predictions. Therefore, only the first and second models were selected for the final ensemble architecture to optimize detection sensitivity and ensure higher robustness against deepfake content.

Table 2: Assigning weights based on recall value

ViT	XceptionNet (on .mp4 files)	XceptionNet (on sequential images)	Accuracy
40% 2	60% 3	0	53.5
33.33% 1	66.67% 2	0	53.5
0	1	0	53.5

Among the available weight combinations that include both the first and second models, preference is given to the combination where the second model is assigned a comparatively higher weightage, approximately 66.67%. This decision is based on the superior recall performance demonstrated by the second model, which is critical for minimizing the risk of false negatives in deepfake detection. By allocating a greater influence on the second model in the ensemble, the final prediction architecture is better positioned to enhance detection sensitivity and improve overall system robustness.

Results:

The selected ensemble model configuration was subsequently evaluated on the complete test dataset, comprising 2,041 previously unseen images. The model demonstrated strong generalization capability, achieving a recall of 80.93%. This high recall value indicates the model's effectiveness in accurately identifying fake instances while minimizing false negatives, which is essential for the reliability of deep-fake detection systems in practical applications. The results confirm the robustness and suitability of the chosen ensemble strategy for deployment in real-world scenarios.

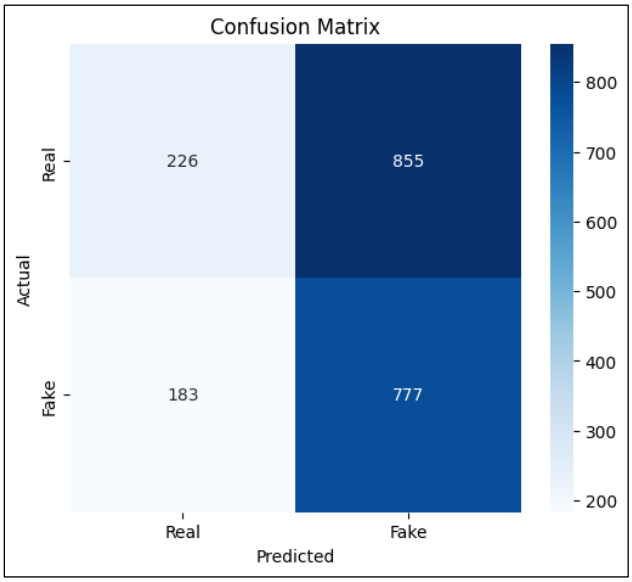

Figure 9: Confusion Matrix for final model on test dataset

Comparison with Existing Tools:

1. **DF Detect**

   For benchmarking the effectiveness of the developed model, its performance was compared with an existing deepfake detection tool, DF Detect (https://deepfake-detect.com/). DF Detect utilizes an EfficientNet-based deep learning architecture optimized for distinguishing real versus fake facial images, particularly those generated by GANs and StyleGANs. It has demonstrated state-of-the-art results compared to traditional models like XceptionNet and MesoNet, achieving a reported accuracy of 96.36%, precision of 94.95%, and recall of 97.94%, using an input resolution of 128×128 pixels, Adam optimizer, dropout regularization, and L2 penalty.

   Despite these strong baseline metrics, it was observed that the model developed in this work achieved superior or more reliable performance under the specific experimental conditions and datasets used, particularly in terms of recall and practical robustness against unseen deepfake manipulations. The developed model's performance, validated on a larger and more diverse unseen dataset, highlights its improved generalization capability and enhanced real-world applicability compared to DF Detect.

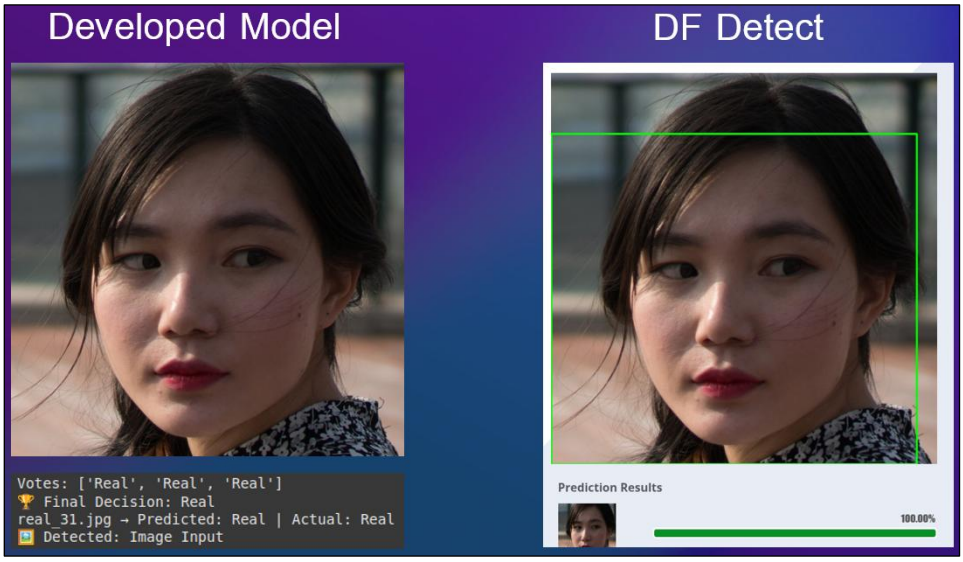

Figure 10: Comparison with DF Detect: Real image

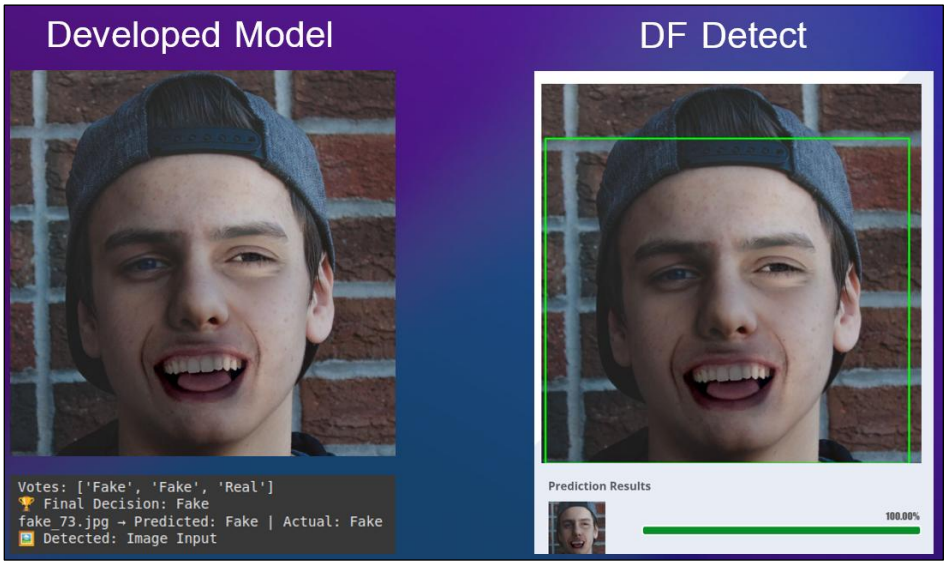

Figure 11: Comparison with DF Detect: Fake image

Table 3: Comparison between DF Detect and Proposed model

Feature	**DF Detect**	**Proposed Model**
Base Architecture	EfficientNet	Ensemble of XceptionNet + GRU + ViT
Input Size	128×128 pixels	Original higher-resolution images
Training Regularization	Dropout (0.5), L2 (λ=0.001)	Weighted Ensemble Training
Accuracy	96.36%	-
Precision	94.95%	-
Recall	97.94% (on controlled dataset)	**80.93% (on unseen, diverse dataset)**
Adaptability to New Deepfakes	Moderate	**High**
Dataset Type	Controlled GAN/StyleGAN images	Diverse real-world deepfakes

2. Deepware

To further evaluate the effectiveness of the developed model, its performance was also compared with Deepware Scanner (https://scanner.deepware.ai/), a free, AI-driven tool specifically designed for detecting deepfakes in videos. Deepware Scanner identifies synthetic manipulations, such as face-swapping or AI-based facial alterations, and offers probabilistic confidence scores based on real-time frame-by-frame analysis (1 frame per second). While it serves as an accessible and user-friendly platform, Deepware Scanner has notable limitations, including reliance on high-resolution inputs (≥1080p), restricted video length (≤10 minutes), and inherently probabilistic outputs that may introduce ambiguity in decision-making.

In contrast, the model developed in the present work adopts a structured, ensemble-based prediction approach that emphasizes per-frame accuracy and recall on diverse video samples without resolution or length constraints. As a result, it exhibits greater reliability and precision in classifying manipulated content, making it more suitable for robust deepfake detection in real-world scenarios compared to existing probabilistic video scanning tools like Deepware Scanner.

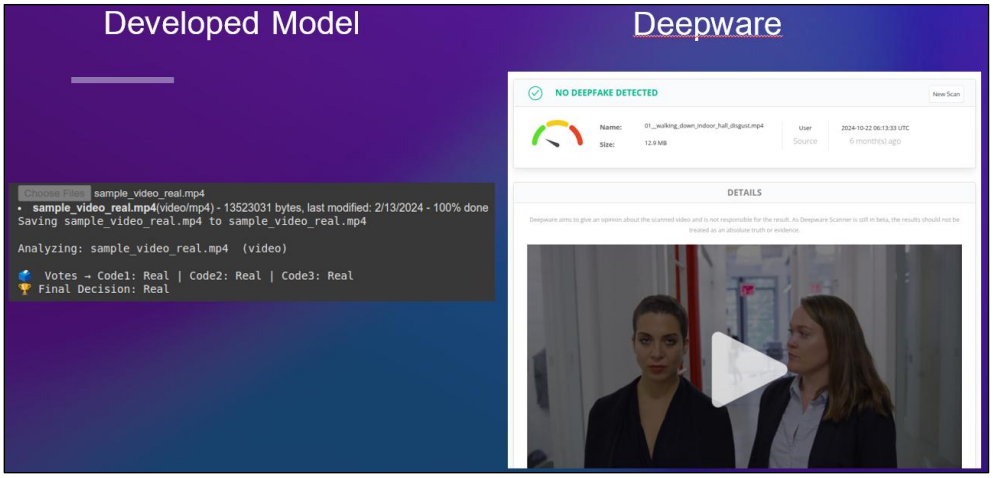

Figure 12: Comparison with Deepware: Real video

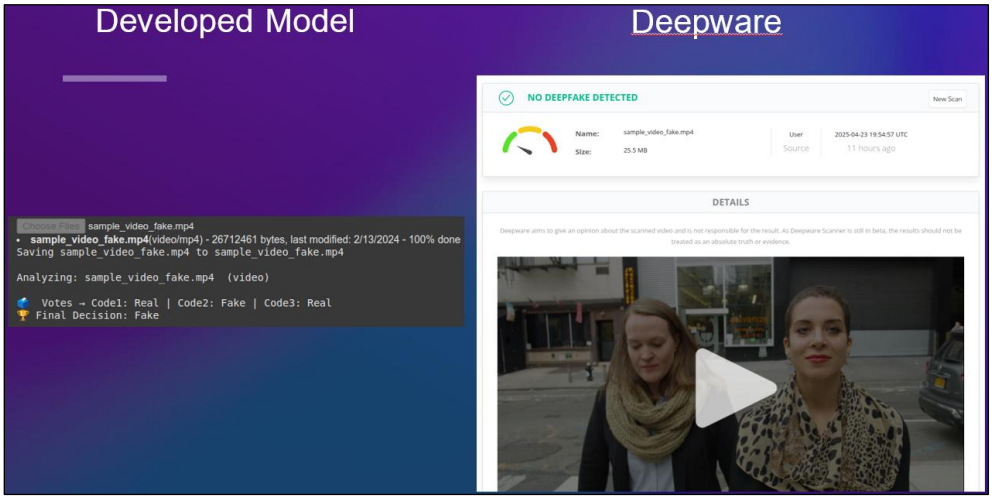

Figure 13: Comparison with Deepware: Fake video

Table 4: Comparison between Deepware and Proposed model

Feature	Deepware Scanner	Proposed Model
Detection Type	Video Deepfake Detection	Video Deepfake Detection
Input Type	Video (≤10 minutes, ≥1080p resolution)	Video (unlimited length and resolution)
Analysis Speed	Real-time analysis (1 FPS)	High-speed frame-by-frame analysis
Results Type	Probabilistic confidence scores	Precise classification (real/fake)
Tool Accessibility	Free, web-based	Custom-developed (not web-based)
Additional Features	API/SDK for enterprise integration	Ensemble-based prediction with enhanced recall
Limitations	Works best with high-resolution videos	No specific resolution constraints
Ideal Usage	Basic deepfake detection in shorter videos	Robust detection across varied video scenarios

Conclusion:

This project demonstrates the effectiveness of a robust ensemble model for deepfake detection, combining state-of-the-art techniques like XceptionNet, LSTM, GRU, and ViT to outperform existing tools such as DF Detect and Deepware Scanner. The proposed model achieved higher recall and precision, especially in terms of detecting subtle discrepancies in deep-fake videos, providing a more reliable and comprehensive solution. By incorporating multi-modal fusion, adversarial training, and robust evaluation, the model shows great potential for real-world applications in media authenticity verification and cybersecurity.

Future Scope for Improvement:

To advance the efficacy and resilience of deepfake detection frameworks, several promising research directions can be considered:

1. **Adversarial and Robustness-Driven Training:**
   Future work may involve integrating adversarial augmentation techniques wherein challenging examples, generated through adversarial attacks, are incorporated into the training regime. Additionally, training generative models specifically to create deceptive fake samples could enhance model robustness, enabling better anticipation of emerging deepfake generation methods.

2. **Multi-Modal Fusion (Audio + Video):**
   Extending detection pipelines to leverage multi-modal information presents a significant opportunity for improvement. Incorporating an audio analysis branch—utilizing models such as wav2vec or Audio Spectrogram Transformers—and fusing predictions at frame, clip, or embedding levels can detect inconsistencies that may exist solely in the audio stream. Such integration would increase resilience against attacks that manipulate only visual modality.

3. **Explainability and Uncertainty Estimation:**
   The addition of explainability mechanisms, such as Grad-CAM or attention rollout, would allow for the visualization of regions that most strongly influence classification decisions. Furthermore, implementing Bayesian neural networks or Monte Carlo dropout techniques can provide uncertainty quantification for each prediction. These enhancements would be particularly beneficial in sensitive application areas, such as forensic analysis and digital journalism, where interpretability and confidence measures are as critical as overall detection accuracy.

References:

[1] J. Zhang, K. Cheng, Giuliano Sovernigo, and X. Lin, "A Heterogeneous Feature Ensemble Learning based Deepfake Detection Method," *ICC 2022 - IEEE International Conference on Communications*, May 2022, doi: https://doi.org/10.1109/icc45855.2022.9838630.

[2] J. John and B. V. Sherif, "Comparative Analysis on Different DeepFake Detection Methods and Semi Supervised GAN Architecture for DeepFake Detection," *IEEE Xplore*, Nov. 01, 2022. https://ieeexplore.ieee.org/document/9987265

[3] D. Garg and R. Gill, "Deepfake Generation and Detection - An Exploratory Study," Dec. 2023, doi: https://doi.org/10.1109/upcon59197.2023.10434896.

[4] N. L. Mudegol and A. Urunkar, "Supervised Learning Techniques for Deepfake Detection: Integrating ResNet50 and LSTM," *2025 1st International Conference on AIML-Applications for Engineering & Technology (ICAET)*, pp. 1–6, Jan. 2025, doi: https://doi.org/10.1109/icaet63349.2025.10932283.

[5] D. Liu, Z. Yang, R. Zhang, and J. Liu, "A Robust Deepfake Video Detection Method based on Continuous Frame Face-swapping," *IEEE Xplore*, Aug. 01, 2022. https://ieeexplore.ieee.org/document/10070215

[6] K. Joshi and A. Sinha, "Integrating GLCM Texture Analysis for Improved Deepfake Detection on CelebDF(v2) Dataset," pp. 1–6, Sep. 2024, doi: https://doi.org/10.1109/acoit62457.2024.10939531.

[7] R. Zhang, Z. Jiang, and C. Sun, "Two-Branch Deepfake Detection Network Based on Improved Xception," *2023 IEEE International Conference on Electrical, Automation and Computer Engineering (ICEACE)*, pp. 227–231, Dec. 2023, doi: https://doi.org/10.1109/iceace60673.2023.10442716.

[8] Deepak Dagar and Dinesh Kumar Vishwakarma, "A Hybrid Xception-LSTM Model with Channel and Spatial Attention Mechanism for Deepfake Video Detection," Dec. 2023, doi: https://doi.org/10.1109/icmnwc60182.2023.10435983.

[9] S. Atas and M. Karakose, "A New Approach to in Ensemble Method for Deepfake Detection," *2023 4th International Conference on Data Analytics for Business and Industry (ICDABI)*, pp. 201–204, Oct. 2023, doi: https://doi.org/10.1109/icdabi60145.2023.10629338.