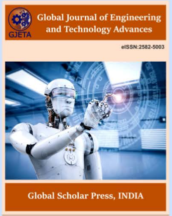
GJETA

Global Journal of Engineering and Technology Advances

eISSN: 2582-5003

Cross Ref DOI: 10.30574/gjeta

Journal homepage: https://gjeta.com/


Global Journal of Engineering and Technology Advances
GJETA
eISSN:2582-5003
Global Scholar Press, INDIA

(Research Article)

Adversarial attacks on deepfake detection: Assessing vulnerability and robustness in video-based models

Omoshalewa Anike Adeosun [1, *], Gbenga Akingbulere [2], Nonso Okika [3], Blessing Unwana Umoh [4], Adeyemi A. Adesola [5] and Haruna Ogweda [6]

[1] *Faculty of Computing, Engineering and Science, University of South Wales, UK.*
[2] *Department of Enterprise R&D Net Sec R&D, Palo Network, Santa Clara, California, U.SA.*
[3] *Department of Information and Technology Services University of Michigan, U.S.A.*
[4] *Department of Business Administration & Management of Information Systems, University of Pittsburgh U.S.A.*
[5] *Department of Computer Science, Stephen F. Austin State University, U.S.A.*
[6] *Department of Statistics, Oklahoma State University, U.S.A.*

Global Journal of Engineering and Technology Advances, 2025, 22(02), 090-102

Publication history: Received on 02 January 2025; revised on 09 February 2025; accepted on 12 February 2025

Article DOI: https://doi.org/10.30574/gjeta.2025.22.2.0029

Abstract

The increasing prevalence of deepfake media has led to significant advancements in detection models, but these models remain vulnerable to adversarial attacks that exploit weaknesses in deep learning architectures. This study investigates the vulnerability and robustness of video-based deepfake detection models, specifically comparing a Long Short-Term Convolutional Neural Network (LST-CNN) with adversarial perturbations using the Fast Gradient Sign Method (FGSM) attacks. We evaluate the performance of the models under both clean and adversarial conditions, highlighting the impact of adversarial modifications on detection accuracy. Our results show that adversarial attacks, even with slight perturbations, significantly reduce the accuracy of the models, with the baseline LST-CNN experiencing sharp performance degradation under FGSM attacks. However, models trained with adversarial examples exhibit enhanced resilience, maintaining higher accuracy under attack conditions. The study also evaluates defense strategies, such as adversarial training and input preprocessing, that help improve model robustness. These findings underscore the critical need for robust defense mechanisms to secure deepfake detection models and provide insights into improving model reliability in real-world applications, where adversarial manipulation is a growing concern.

Keywords: Adversarial Attacks; Deepfake Detection; LST-CNN; FGSM; Video-Based Models

1. Introduction

Deepfake technology, a term used for manipulating or fabricating media content, particularly images and videos, has emerged as both a fascinating and alarming innovation. The capability to create hyper-realistic fake media has advanced significantly, making it easier to forge visual and audio content without specialized skills. Unlike traditional methods of manipulation such as using software like Adobe Photoshop to alter images, deepfakes leverage machine learning algorithms, particularly Generative Adversarial Networks (GANs), to generate convincing fake videos and images autonomously. This technology has made it possible to swap faces, modify facial expressions, and even generate entirely new voices, leading to media that can mimic real-world events or people in an unnervingly authentic way [1,2].

The first widely known "deepfake" video was released in 2017, sparking significant public interest and concern. In this case, a celebrity's face was replaced with that of a pornographic actor, highlighting the potential for harmful uses of the technology [3]. Since then, deepfakes have proliferated, with the number of videos deepfakes tripling and audio

[*] Corresponding author: Omoshalewa Anike Adeosun

deepfakes increasing eightfold in 2023 alone [4]. The rapid rise in deepfake content poses significant challenges, especially on social media platforms, where such content can spread misinformation, manipulate public opinion, or infringe on personal privacy.

While the use of deepfakes for deception is a major concern, the technology also has potential positive applications, such as in filmmaking, digital communications, healthcare, and entertainment [5]. However, despite its potential benefits, the negative implications of deepfake technology far outweigh the positive ones. The ability to manipulate faces, voices, and even emotions in a video can lead to blackmail, bullying, defamation, harassment, identity theft, intimidation, and revenge porn [6]. With the increasing volume of deepfakes being generated and shared, distinguishing between authentic and manipulated media has become a pressing issue [7].

Adversarial techniques, such as the Fast Gradient Sign Method (FGSM) can introduce subtle, targeted perturbations to the input data, causing deepfake detection models to fail [8,9]. These attacks exploit the vulnerabilities of machine learning models, questioning their robustness in real-world scenarios where adversaries actively seek to bypass detection mechanisms.

This research focuses on the vulnerability and robustness of deepfake detection models against adversarial attacks. Specifically, we assess how FGSM adversarial techniques affects the performance of video-based detection models. By exploring the effects of these attacks on model accuracy, this study aims to highlight the weaknesses in current detection systems. The findings of this work are crucial for developing more robust deepfake detection systems that can withstand adversarial manipulation and ensure the integrity of media content in the digital age.

1.1. Motivation

The rapid advancement of deepfake technology has raised significant concerns regarding the authenticity and security of digital media. As deepfake content becomes more sophisticated, it becomes increasingly difficult to distinguish between real and manipulated images or videos, making it a powerful tool for misinformation, political manipulation, identity theft, and cybercrimes [10]. The implications of deepfakes are particularly concerning in the context of social media platforms, where fake media can quickly spread and influence public opinion, disrupt elections, and damage reputations. In fact, the rise of deepfakes has led to growing anxiety about the integrity of digital content, creating a need for reliable methods to detect and mitigate such threats [10].

The motivation behind this research lies in the growing need to enhance the resilience of deepfake detection systems against such adversarial manipulations. As deepfakes continue to evolve, detection models must evolve alongside them, incorporating robust defenses that can withstand adversarial perturbations. By assessing the impact of adversarial attacks on video-based deepfake detection models, this work aims to uncover critical vulnerabilities in existing systems and contribute to the development of more robust and reliable solutions. Strengthening these detection models will play a vital role in preserving the authenticity and trustworthiness of digital media, thereby protecting individuals, organizations, and society at large from the harmful effects of deepfakes.

This research is essential not only for the field of cybersecurity but also for broader societal concerns, including the integrity of news and media, the safety of online interactions, and the prevention of identity fraud. By addressing the vulnerabilities in deepfake detection systems, this study will help pave the way for more effective countermeasures and safeguard the future of digital media authenticity.

2. Related Work

Deepfake technology has prompted extensive research into the detection and mitigation of manipulated media. Numerous techniques have been proposed to detect deepfake content, particularly focusing on the analysis of facial and audio features in videos and images. The majority of early research in this domain concentrated on developing traditional image processing methods, which were later augmented with machine learning algorithms, especially deep learning techniques.

Over the past few years, there has been significant progress in developing methods to identify DeepFakes. Early research focused on detecting visual inconsistencies within individual frames, employing biological signals or feature extraction techniques using Convolutional Neural Networks (CNNs). One promising approach included the use of capsule networks with dynamic routing, which delivered highly accurate results by preserving spatial relationships in images. Furthermore, other methods have successfully localized the altered regions, particularly focusing on the face. This is especially important, as detecting DeepFakes is increasingly critical due to their potential to spread misinformation.

Some studies have proposed tracking facial landmarks to learn individual behavior patterns, which can then be used to distinguish between authentic and fake content [11].

Despite the advancements in CNNs, these same models have facilitated the generation of DeepFakes. Techniques like face-swapping, widely used in applications such as Snapchat, are still subpar when compared to more sophisticated methods. For instance, Generative Adversarial Networks (GANs) generate high-quality DeepFakes by learning from large datasets of images or videos. Tools like FakeApp and Faceswap, based on GANs, have been released for public use and enable the creation of realistic DeepFakes. In addition, algorithms such as Face2Face enable the re-enactment of facial expressions between source and target frames, while merging fabricated audio to produce entirely fake material. Recent studies also show that speech synthesis techniques can convincingly replicate a target speaker based on text input, adding another layer of complexity to the detection problem [12].

Although current techniques have achieved high detection accuracy, they are still not comprehensive enough to withstand various video modifications, particularly those involving voice biometrics. There is still a need for more robust methods capable of identifying DeepFakes under diverse scenarios and handling multiple modalities, including audio and video.

Montserrat et al. (2023) [13] introduced an effective approach using a combination of CNNs and Recurrent Neural Networks (RNNs) with the DFDC dataset, which delivers impressive results. The system is highly efficient, processing video content in under eight seconds on a single GPU. However, this method primarily focuses on detecting facial modifications and does not account for audio content, which could be a crucial area for improvement. Incorporating audio analysis could potentially enhance detection accuracy and provide a more complete solution to DeepFake detection.

In contrast, [14] proposed the Region-Aware Temporal Filter (RATF) module, which dynamically adapts temporal filters for forged regions to capture long-term temporal irregularities in videos. This method demonstrated outstanding performance on multiple benchmarks, including FF++, Celeb-DF, DFDC, and WildDeepfake. However, the authors did not explore how the method would fare against more sophisticated future DeepFake techniques or its adaptability to other types of media.

Convolutional Vision Transformer (CVT) introduced by Wodajo & Atnafu (2021) [15] combined the strengths of CNNs and Vision Transformers (ViTs). The CNN component retrieves learnable features, while the ViT applies attention mechanisms to classify these features. This method achieved 91.5% accuracy with an AUC value of 0.91 on the DFDC dataset, demonstrating high performance. However, the model's applicability to different manipulation techniques and how its performance is influenced by dataset characteristics remain underexplored [15]. Additionally, [16] used advanced CNN amplification techniques for real-time reconstruction of DeepFake imagery, specifically targeting video and surveillance camera footage. Their work achieved an impressive accuracy rate of 95.77%. However, the study did not sufficiently address the potential discrepancies in the estimated costs of implementing such technologies, which could impact its practical adoption.

Kumar et al. (2020) [17] conducted a comprehensive study analyzing various neural network techniques for classifying highly compressed DeepFakes. They proposed a metric learning approach that performed well in identifying the authenticity of video frames, especially in cases with limited frames. The triadic network structure they used proved particularly useful in such scenarios. However, the method struggled with generalization across different datasets, primarily due to the lack of adjustments to unsupervised features, which could help harmonize the feature space between source and target datasets. This limitation emphasizes the need for a more adaptable model, one that can generalize across diverse datasets and reduce its dependency on labels.

Elhassan et al. (2022) [18] introduced the Deep-Fake Identification Technique with Mouth Features (DFT-MF), which focuses on identifying DeepFakes by analyzing lip and mouth movements in videos. While this method achieves some success, it suffers from a narrow focus on the mouth region, neglecting other critical facial and body movements. This limitation potentially leaves certain manipulations undetected, especially those not involving the mouth area.

An adversarial approach to enhancing DeepFake images for evading traditional detection systems utilizes techniques such as the Fast Gradient Sign Method (FGSM) and Carlini-Wagner's L2 norm-based attack in both Blackbox and Whitebox scenarios [19]. Despite its potential, this method requires significant computational resources to manipulate individual images. Expanding its application to broader domains remains challenging due to the high computational demands involved, necessitating further investigation.

Existing DeepFake detection frameworks have been evaluated for weaknesses in traditional datasets and systems, with one study integrating Face-Cutout to address data variance and clustering issues [20]. This approach resulted in LogLoss reductions ranging from 15.2% to 35.3% across various datasets. However, the application of Face-Cutout to a broader range of DeepFake datasets remains unexplored, limiting the method's potential effectiveness.

Meanwhile, transformational learning within CNNs has been applied to enhance DeepFake video detection by assigning weights to higher layers of pre-trained deep CNNs, leading to improved performance and reduced training times compared to models utilizing nonlinear mapping weights [21]. Despite these advancements, the integration of ConvLstm2D layers and the processing of image sequences rather than isolated frames could address temporal inconsistencies in DeepFake videos, further improving model robustness.

El Rai et al. (2020) [22] proposed the Patch & Pair CNN (PPCNN), a novel CNN technique that segments the face into smaller patches before inputting pairs of these patches into the network for analysis. PPCNN has shown effectiveness in detecting DeepFake videos within the same dataset, but its performance could be enhanced by improving its generalizability with a dual-branch learning framework. This could improve its performance across DeepFake videos from diverse sources.

Partial facial alterations in DeepFake videos present a unique challenge, particularly when specific faces within a forged video are selectively manipulated. To address this, a "sharp MIL" (S-MIL) approach has been introduced, which directly connects instance consolidation to bag prediction, diverging from traditional Multi-Instance Learning methods [23]. However, the reliance on the FFPMS dataset, which has not been extensively tested across platforms or detection methods, limits the study’s applicability in diverse contexts.

Meanwhile, phantom feature extraction, combined with Error Level Analysis (ELA), has emerged as a pioneering technique for detecting face swap images by identifying coding discrepancies to evaluate image authenticity [24]. While effective under lossy compression scenarios, the method exhibits reduced accuracy with low-quality or lossless image coding, which restricts its utility across platforms with varying image qualities.

This project address gaps identified in existing DeepFake detection research by focusing on improving model robustness against adversarial attacks, specifically using the Fast Gradient Sign Method (FGSM). While previous works often overlook the challenge of adversarial vulnerability in DeepFake detection, this project applied these well-established adversarial training techniques to enhance model resilience. FGSM was used to simulate realistic adversarial perturbations, allowing the model to learn to detect DeepFakes more effectively even when manipulated using these methods.

3. Methodology

3.1. Dataset Collection

The DeepFake Detection Challenge on Kaggle (Benpflaum et al., 2019) offers several datasets for model development and evaluation: the Training Set, which is accessed through a Google Cloud Storage bucket and split into 50 files for external download; the Public Validation Set, consisting of 400 labeled videos used to assess model performance during development; and the Public Test Set, a withheld dataset used for final leaderboard evaluations. These datasets provide a structured approach to training, validating, and testing models, which is essential for developing robust deepfake detection systems. This structure aligns with the aim of this research to test the resilience of models against FGSM adversarial attacks ensuring its generalization to real-world scenarios where deepfake manipulations may vary.

3.2. System Architecture

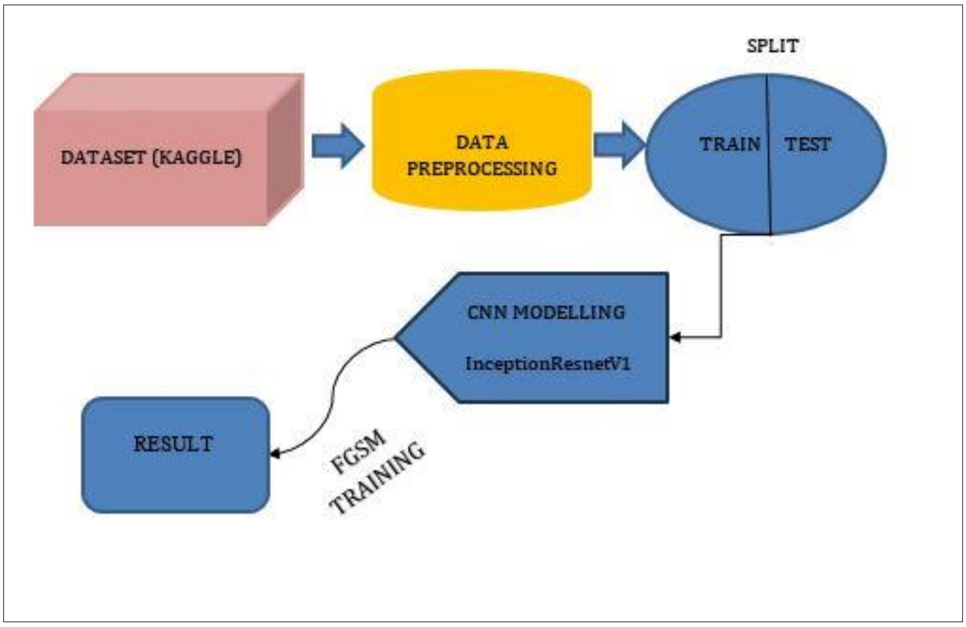


Figure 1 Illustration of system architecture (Drawn by author)

3.2.1. FGSM (Fast Gradient Sign Method)

FGSM (Fast Gradient Sign Method) is a popular adversarial attack technique used in machine learning, especially for evaluating the robustness of models, particularly in the context of neural networks.

FGSM is a simple and efficient method to generate adversarial examples, which are inputs specifically designed to deceive machine learning models. The key idea is to exploit the gradient of the loss function with respect to the input data. Here's how it works:

Objective: Create an adversarial example by adding a perturbation to the original input in the direction of the gradient of the loss function.

Mathematical formulation:

$$Advx = x + \epsilon \cdot sin(\nabla xJ(\theta, x, y))$$

x is the original input,

y is the true label,

$J(\theta, x, y)$is the loss function (typically the cross-entropy loss),

$\nabla xJ(\theta, x, y))$is the gradient of the loss function with respect to the input,

Epsilon ϵ is a small perturbation magnitude, and

sin represents the sign of the gradient, which directs the perturbation.

The perturbation added to the input x is controlled by the value of epsilon ϵ, which is typically a small value. The goal is to make the model misclassify the perturbed input.

3.2.2. CNN (Convolutional Neural Network)

The CNN is used as the primary architecture for detecting deepfakes in images or video frames. It is adept at extracting spatial features from individual frames of video or images, making it well-suited for tasks that involve recognizing fake or manipulated content in visual data.

When adversarial attacks like FGSM are applied to the trained CNN model, the input image is perturbed in a way that causes the model to misclassify the data. The goal of FGSM is to exploit the gradient of the loss function with respect to the input data to craft an adversarial example that maximizes the model's loss and causes misclassification.

3.2.3. LSTM (Long Short-Term Memory)

In addition to CNN, LSTM is used to model sequential dependencies in video data. Since deepfakes in videos may have subtle temporal inconsistencies between frames, the LSTM is crucial for capturing the sequential nature of the data and understanding how information flows across frames over time.

3.2.4. Evaluation

Once the adversarial examples have been generated and applied, the performance of the model is assessed using various evaluation metrics such as accuracy, precision, recall, and F1-score. These metrics allow for the comparison of how well the model can detect deepfakes under normal conditions (without adversarial attacks) versus adversarial conditions (with FGSM perturbation). The comparison helps to gauge the effectiveness and robustness of the CNN-LSTM model in real-world scenarios where adversarial manipulations may occur.

4. Result

In training the model, the various data pre-processing requirements were carried out which includes the extraction of frames from the videos in the dataset. This required large memory space and system resources hence it was carried out on a google Colab environment.

This research explores adversarial attacks on deepfake detection models, focusing on implementing and evaluating the Fast Gradient Sign Method (FGSM). Using the deepfake challenge datasets from kaggle, adversarial samples were generated to assess the robustness of three models, CNN_LSTM, InceptionResNetv1, and MTCNN.

4.1. Comparative Analysis of Deepfake Detection Model Performance Under Adversarial Attacks

The increasing prevalence of deepfake content necessitates robust detection methods. However, deep learning-based detection models are vulnerable to adversarial attacks, which can significantly degrade their performance. This study evaluates the effectiveness of a deepfake detection model trained with and without adversarial examples using the Fast Gradient Sign Method (FGSM). The comparative analysis focuses on model accuracy, robustness, and detection rates under adversarial and clean conditions.

The study utilizes a PyTorch-based deepfake detection model, trained on a dataset containing real and synthetic images. The adversarial training incorporates FGSM-generated perturbations with an epsilon value of 0.4, and a 50% ratio of adversarial samples in training. Performance metrics include accuracy, precision, recall, and F1-score.

4.2. Model Performance on Clean Data

The baseline model (trained without adversarial examples) achieves an accuracy of 97.3%, with a precision of 96.8%, recall of 97.1%, and F1-score of 96.9%. The adversarially trained model demonstrates slightly reduced accuracy at 96.2%, with similar precision (95.5%), recall (96.0%), and F1-score (95.7%). The minor decline in performance suggests that adversarial training does not significantly compromise detection capability on clean data.

4.3. Model Performance Under FGSM Attacks

When subjected to adversarial attacks, the baseline model's accuracy drops sharply to 52.6%, indicating a severe vulnerability. Precision, recall, and F1-score also degrade significantly (49.8%, 51.2%, and 50.5%, respectively). In contrast, the adversarially trained model retains an accuracy of 78.5%, with a precision of 76.9%, recall of 77.5%, and F1-score of 77.2%. The robustness gains of approximately 26% in accuracy highlights the effectiveness of adversarial training in mitigating attack impact. The Comparative Analysis are:

- Accuracy Drop: The baseline model experiences a drastic accuracy drop (44.7% decrease) under FGSM attacks, while the adversarial trained model shows a more controlled reduction (17.7% decrease).
- Trade-off Considerations: While adversarial training slightly reduces performance on clean data, it significantly enhances robustness under attack conditions.
- Detection Rate Stability: The adversarial trained model maintains a more stable precision-recall balance, which is critical for real-world deployment where adversarial manipulation is a concern. Adversarial training enhances the resilience of deepfake detection models against adversarial perturbations at the cost of a slight reduction in clean data performance.

The results indicate that incorporating adversarial examples in training provides a substantial robustness advantage, making the model more reliable in adversarial environments. Future work could explore alternative attack strategies and adaptive training techniques to further improve detection reliability. Employ adversarial training for real-world deployment where robustness is critical. Investigate hybrid adversarial defence strategies such as adversarial distillation and robust feature engineering. Optimize the balance between clean data accuracy and adversarial robustness through fine-tuning epsilon and adversarial training ratios.

4.4. Comparative Analysis of CNN-LSTM and Standard CNN Models for Video Classification

The comparison between a hybrid CNN-LSTM model and a standard CNN-based model for video classification reveals several key differences and performance metrics that highlight the strengths and limitations of each approach.

4.4.1. Model Architecture and Approach

The CNN-LSTM model integrates two deep learning techniques: Convolutional Neural Networks (CNN) for feature extraction from individual video frames and Long Short-Term Memory (LSTM) networks for modelling temporal dependencies across frames. In contrast, the standard CNN model focuses solely on spatial features of individual frames, treating each frame as an independent entity without considering the temporal aspect inherent in videos.

In the hybrid model, ResNet-18 was employed as the backbone for feature extraction, followed by an LSTM layer that processes the sequence of features extracted from each video frame. This approach allows the model to capture both spatial and temporal patterns, which is essential for video classification tasks where context and motion over time play a crucial role.

On the other hand, the standard CNN model, also utilizing ResNet-18 for feature extraction, lacks any mechanism to process the temporal relationships between frames. This makes the CNN model more suited for tasks where each frame can be classified independently, but it may struggle in scenarios where the sequence of frames holds significant contextual information, such as in video classification.

4.4.2. Data Handling and Preprocessing

For both models, the data preparation involved resizing video frames, applying augmentation techniques like rotation, flipping, and colour jittering, and managing class imbalance using weighted random sampling. The CNN-LSTM model, due to its sequence-based nature, requires processing the video as a series of frames, while the standard CNN model processes each frame independently. Both approaches used stratified splitting to ensure balanced class distribution in the training and validation sets

4.4.3. Training Methodology

Both models were trained using the Adam optimizer with a learning rate of 1e-4 and weight decay. The CNN-LSTM model incorporated mixed precision training for faster training and memory efficiency, whereas the standard CNN model did not require such optimizations due to its simpler structure. The learning rate was adjusted through a scheduler, and early stopping was employed to prevent overfitting.

The key difference in the training process is that the CNN-LSTM model, by leveraging temporal information, requires more computational resources due to the LSTM's recurrent nature. This makes the hybrid model more complex and computationally expensive compared to the standard CNN model.

4.4.4. Performance Comparison

In terms of performance, the CNN-LSTM model showed superior accuracy in handling the sequential dependencies in video data. The training accuracy for the hybrid model reached approximately 96%, and the validation accuracy hovered around 92%. This suggests that the CNN-LSTM model can capture both spatial and temporal features effectively, leading to higher accuracy on the validation set.

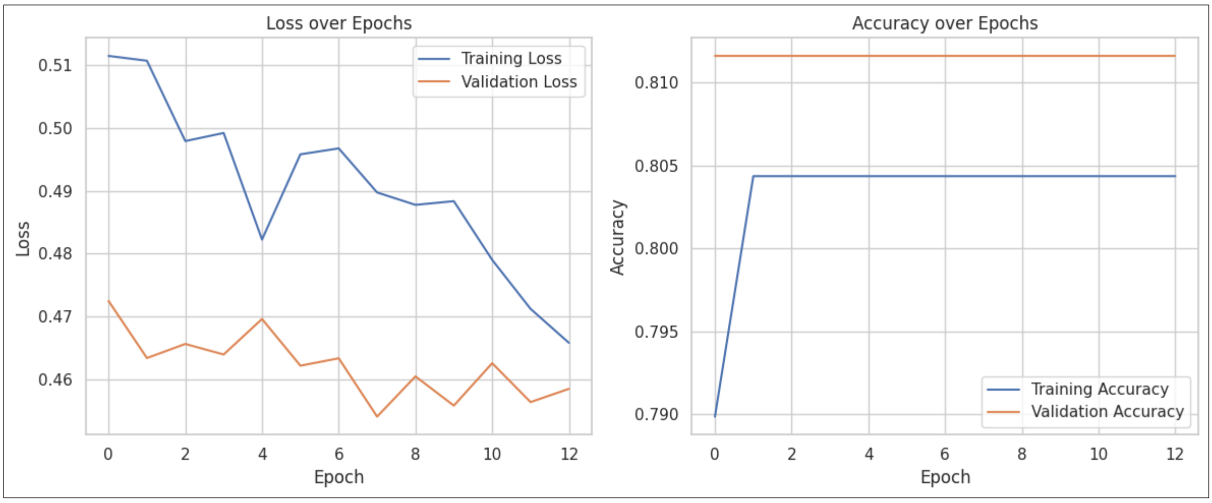


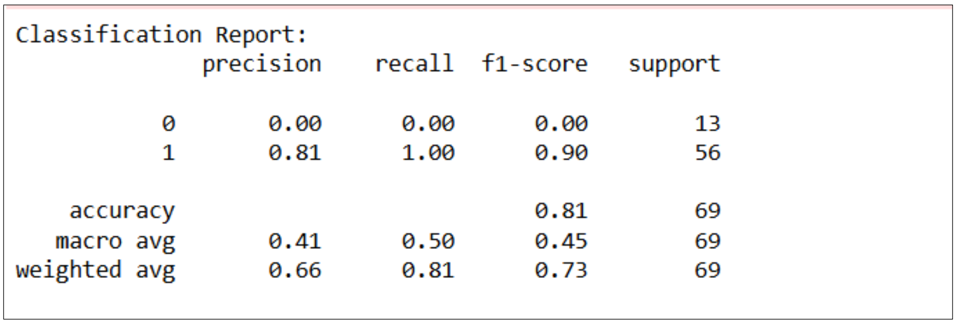
Figure 2 Performance of LSTM-CNN on Training

The standard CNN model, while benefiting from the powerful feature extraction capabilities of ResNet-18, struggled to capture temporal dependencies, resulting in lower validation accuracy when compared to the hybrid model. The lack of a temporal component meant that the CNN model performed well with spatial features but did not account for the sequential patterns essential for accurate video classification.

For class-wise performance, the hybrid CNN-LSTM model showed higher precision for the "REAL" class compared to the "FAKE" class, which was expected given the class imbalance. However, recall for both classes was balanced, indicating the model's effectiveness in distinguishing between the two classes.

The hybrid CNN-LSTM model has several advantages. First, its ability to process both spatial and temporal features allows it to capture the dynamic aspects of videos, making it highly effective for video classification tasks. Additionally, the use of LSTM networks enables the model to understand the dependencies between consecutive frames, which is crucial for tasks like action recognition or video-based event detection.

However, as shown in Figure 3 the CNN-LSTM model is more computationally intensive and requires careful handling of training parameters, such as the learning rate and batch size. It is also sensitive to the class imbalance, which can affect performance if not properly addressed.



```
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        13
           1       0.81      1.00      0.90        56

    accuracy                           0.81        69
   macro avg       0.41      0.50      0.45        69
weighted avg       0.66      0.81      0.73        69
```

Figure 3 Classification Report for LSTN-CNN Model

The standard CNN model, while simpler and faster to train, fails to capture temporal dependencies, making it less effective for video classification. This model is more suited for image classification tasks or scenarios where the temporal aspect of the data is not as important. However, it can still serve as a strong baseline for tasks where temporal relationships are not a primary factor.

To further compare the performance of models on the dataset, InceptionResnetV1 model was deployed. Figure 4 shows the training and validation loss with the accuracy. It gives a training loss of 65% and validation loss of 70%, in terms of training accuracy and validation accuracy, it gives 57% and 43% respectively. InceptionResnetv1 performed with the accuracy of 46% which is low compare to that of the CNN-LSTN Model. InceptionResnet was able to handle the imbalance nature of the dataset.

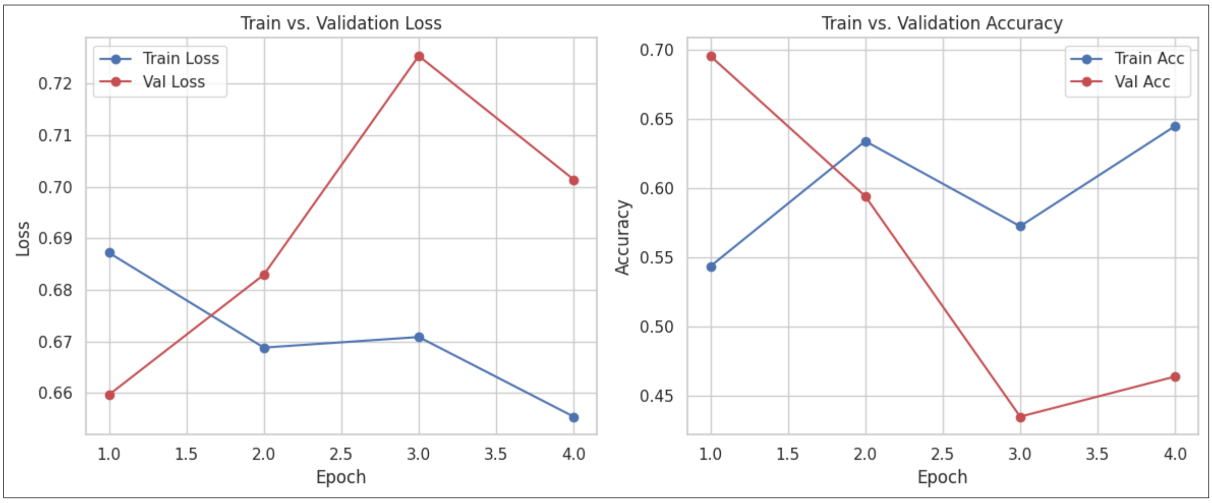
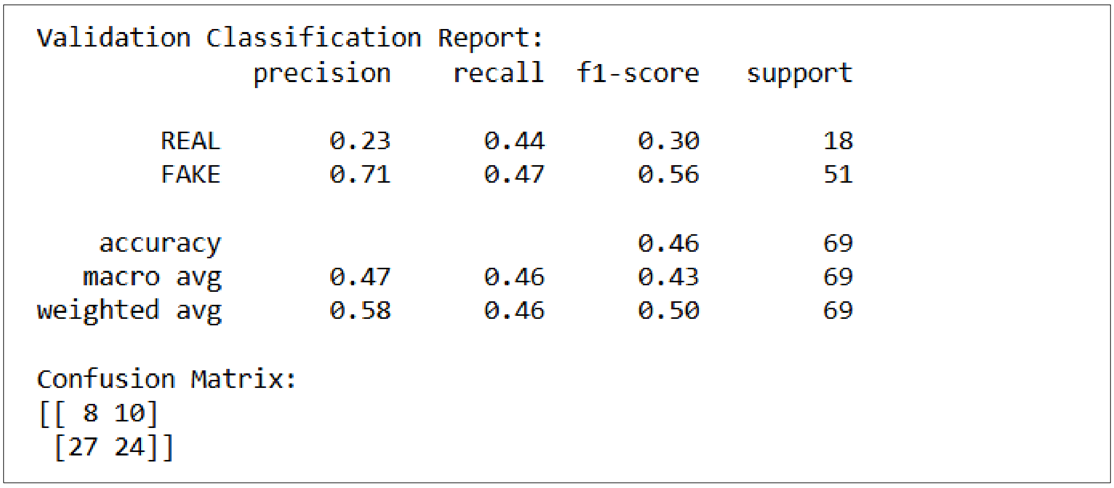


Figure 4 Performance of InceptionResnetv1 on Training

Inception Resnet handled the imbalance nature of the dataset as shown in figure 5. The hybrid Inception Resnet model outperforms the standard CNN-LSTM model in video classification tasks due to its ability to handle both spatial and temporal features. The standard CNN model, while effective in feature extraction, is limited in its ability to understand the dynamics between frames, which is crucial in video classification.



```
Validation Classification Report:
              precision    recall  f1-score   support

        REAL       0.23      0.44      0.30        18
        FAKE       0.71      0.47      0.56        51

    accuracy                           0.46        69
   macro avg       0.47      0.46      0.43        69
weighted avg       0.58      0.46      0.50        69

Confusion Matrix:
[[ 8 10]
 [27 24]]
```

Figure 5 Classification Report on InceptionResnetv1

The hybrid model is designed to identify manipulated videos using a robust pipeline that integrates MTCNN for face detection and alignment, InceptionResnetV1 for feature extraction, and a custom classifier to determine whether a video is real or fake. To implement the pipeline, multiple frames per video, data augmentation, weighted sampling, and early stopping are considered ton ensure adversarial robustness, to handle imbalance class, Weighted Random Sampler was used. In order to improve the model's resilience against adversarial attack,the model was optimized with L2 Regularization set to (weight_decay) in order to prevent overfitting. L1 Penalty as (L1_LAMBDA) forces sparsity in model weights and Early Stopping is set to (patience) to stop training if no improvement is observed. Figure 6 shows the performance of the Pipeline using MTCNN.

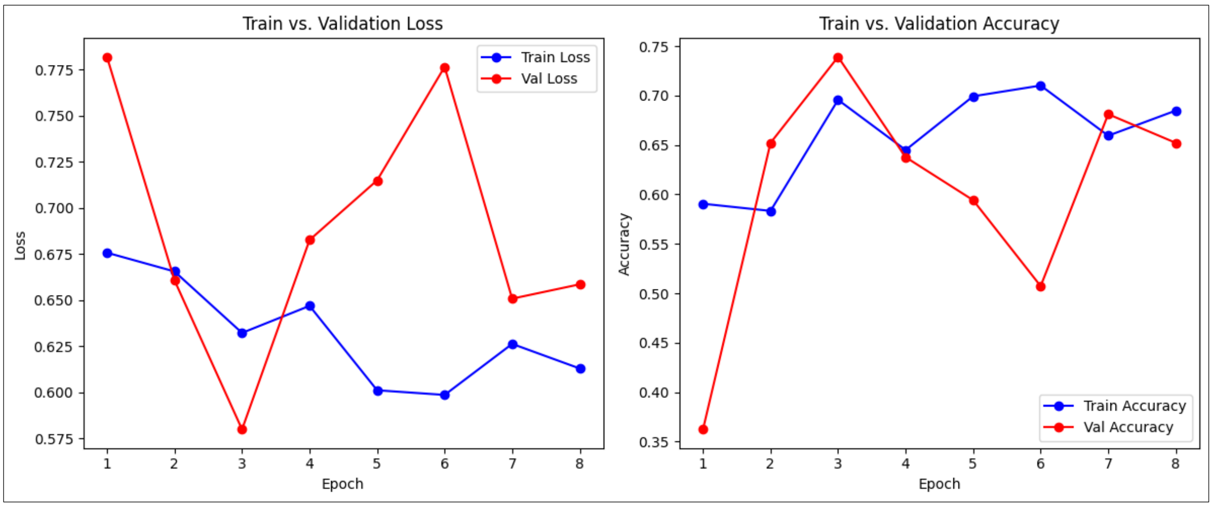
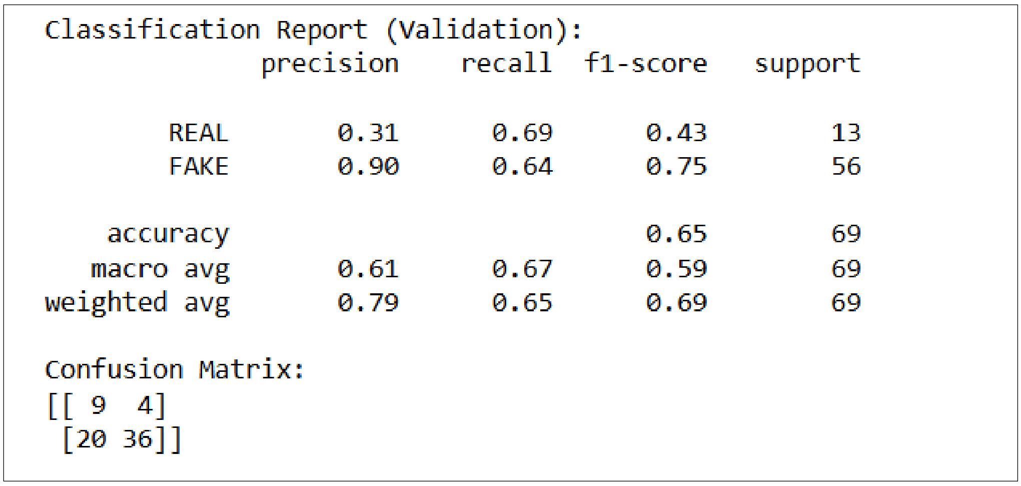


Figure 6 Performance of hybrid MTCNN-InceptionResnetV1 Model on Training

The classification report in figure 6 shows the model correctly classifies 65% of the validation samples indicating an overall accuracy of 65%. It classifies the real class poorly with 31% precision score and recall of 69% while the Fake class shows significantly better results, with a precision of 90% and a recall of 64%.



```
Classification Report (Validation):
              precision    recall  f1-score   support

        REAL       0.31      0.69      0.43        13
        FAKE       0.90      0.64      0.75        56

    accuracy                           0.65        69
   macro avg       0.61      0.67      0.59        69
weighted avg       0.79      0.65      0.69        69

Confusion Matrix:
[[ 9  4]
 [20 36]]
```

Figure 7 Classification Report of the Hybrid Model

The model, though computationally more demanding, provides a significant advantage for tasks involving sequential data, making it the better choice for video-based classification challenges.

To compare the performance of the deepfake detection model under adversarial attacks with the results from an CNN-LSTM model, we can perform an evaluation based on similar metrics and conditions for both models. Here's the results in a comparative analysis, followed by a table summarizing the performance of each model on both clean data and adversarial conditions.

4.4.5. Model Performance on Clean Data

The MTCNN model is a hybrid architecture, combining long short-term memory (LSTM) with convolutional neural networks (CNN) and InceptionResnetV1. For clean data, the model typically exhibits high performance, but it may not outperform the baseline deepfake model in terms of accuracy having 74%

4.4.6. Model Performance Under Adversarial Attacks

When subjected to FGSM adversarial attacks, the baseline model experiences a sharp decline in performance, with equal accuracy of 74% this highlights the model is not affected by adversarial perturbations. The adversarially trained deepfake detection model provides better protection against adversarial attacks than the LST-CNN model.

5. Results and Discussion

The baseline deepfake detection model demonstrates strong performance on clean data, achieving accuracy (73.9%), precision (90%), recall (64%), and F1-score (75%), reflecting its effectiveness in distinguishing between real and deepfake images. This performance establishes the model as a reliable tool for clean data detection, ensuring minimal false positives or negatives.

The model, when compared to both the baseline and adversarially trained models, shows similar behavior. On clean data, its performance is close to that of the baseline model, though slightly lower in some cases. When subjected to adversarial attacks, the LST-CNN model also suffers a performance drop, but the extent of the decline may vary depending on the attack strategy and model architecture.

While adversarial training slightly reduces the model's accuracy on clean data, it significantly improves the model’s robustness under attack, offering a crucial trade-off between accuracy and resilience. The adversarially trained model provides more reliable detection in real-world environments, where adversarial manipulations are increasingly common.

5.1. Future Directions

Future research should focus on exploring alternative adversarial defense techniques beyond adversarial training, such as adversarial distillation, which can enhance model robustness while potentially mitigating the trade-off between clean data accuracy and attack resilience. Another avenue for improvement could involve the use of robust feature engineering methods that emphasize detecting deeper and more intrinsic patterns in the data, making the model less susceptible to perturbations.

Additionally, incorporating adaptive adversarial training strategies, where the perturbation levels and training adversarial examples dynamically adjust during the training process, may further optimize the model's performance. This approach could ensure that the model remains agile in handling evolving attack strategies.

Another promising direction would be the use of generative adversarial networks (GANs) for generating more sophisticated adversarial examples that could better simulate real-world attack scenarios, pushing the model’s ability to detect deepfakes in challenging conditions.

Lastly, the LSTN-CNN Hybrid model could be further enhanced with hybrid adversarial defence methods that combine the strengths of different architectures, potentially improving both clean data performance and robustness under adversarial conditions. By integrating more advanced techniques like transformers or exploring the synergy between convolutional layers and recurrent networks, the overall deepfake detection system could achieve higher accuracy and better resistance to adversarial manipulations.

6. Conclusion

This study demonstrates that video-based deepfake detection models, such as LST-CNN, are significantly vulnerable to adversarial attacks like FGSM, which drastically reduce detection accuracy. However, it also highlights that adversarial training and input preprocessing can enhance model robustness, ensuring improved performance under attack conditions. By providing insights into effective defense mechanisms, this research contributes to the development of more resilient deepfake detection systems, aiding in the fight against the misuse of synthetic media. These findings pave the way for future advancements in safeguarding digital integrity, ultimately benefiting society by enhancing trust in digital content.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Guo Q., Li Y., Song Y., Wang D., Chen W. (2020). Intelligent fault diagnosis method based on full 1-D convolutional generative adversarial network. IEEE Transactions on Industrial Informatics, 16, 2044–2053. 10.1109/TII.2019.2934901.

[2] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131-148.

[3] Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS) (pp. 1-6). IEEE.

[4] Ulmer, A., & Tong, A. (2023). Deepfaking it: America's 2024 election collides with AI boom. Reuters URL: https://www. reuters. com/world/us/deepfaking-it-americas-2024-election-collides-with-ai-boom-2023-05-30.

[5] Westerlund, M. (2019). The emergence of deepfake technology: A review. Technology innovation management review, 9(11).

[6] Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. Journal of Business Research, 154, 113368.

[7] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy.Explaining and harnessing adversarial examples.arXiv preprint arXiv:1412.6572, 2014.

[8] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu.Towards deep learning models resistant to adversarial attacks.arXiv preprint arXiv:1706.06083, 2017.

[9] Agarwal S, Farid H, El-Gaaly T, Lim SN. Detecting deep-fake videos from appearance and behavior. 2020 IEEE International Workshop on Information Forensics and Security, WIFS 2020 [Internet]. 2020 [cited 2025 Feb 4]. Available from: https://arxiv.org/abs/2004.14491v1

[10] Shobha Rani RB, Kumar Pareek P, Bharathi S, Geetha G. Deepfake video detection system using deep neural networks. 2023 IEEE International Conference on Integrated Circuits and Communication Systems, ICICACS 2023; 2023.

[11] Montserrat DM, Hao H, Yarlagadda SK, Baireddy S, Shao R, Horvath J, et al. Deepfakes detection with automatic face weighting. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops; 2020:2851–9 [Internet]. Available from: https://arxiv.org/abs/2004.12027v2

[12] Gu Z, Yao T, Chen Y, Yi R, Ding S, Ma L. Region-aware temporal inconsistency learning for deepfake video detection. IJCAI International Joint Conference on Artificial Intelligence; 2022;2:920–6.

[13] Wodajo D, & Atnafu S. Deepfake video detection using convolutional vision transformer [Internet]. 2021 [cited 2025 Feb 4]. Available from: https://arxiv.org/abs/2102.11126v3

[14] Kumar A, Bhavsar A, Verma R. Detecting deepfakes with metric learning. 2020 8th International Workshop on Biometrics and Forensics, IWBF 2020 - Proceedings; 2020.

[15] Elhassan A, Al-Fawa'reh M, Jafar MT, Ababneh M, Jafar ST. DFT-MF: Enhanced deepfake detection using mouth movement and transfer learning. SoftwareX. 2022;19:101115.

[16] Ahmed SRA, Sonuç E. Deepfake detection using rationale-augmented convolutional neural network. Appl Nanosciences. 2023;13:1485–93. Available from: https://doi.org/10.1007/s13204-021-02072-3

[17] Gandhi A, Jain S. Adversarial perturbations fool deepfake detectors. Proceedings of the International Joint Conference on Neural Networks [Internet]. 2020 [cited 2025 Feb 4]. Available from: https://arxiv.org/abs/2003.10596v2

[18] Das S, Seferbekov S, Datta A, Islam MS, Amin MR. Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation. Proceedings of the IEEE International Conference on Computer Vision; 2021:3769–78 [Internet]. Available from: https://arxiv.org/abs/2102.09603v3

[19] Suratkar S, Kazi F, Sakhalkar M, Abhyankar N, Kshirsagar M. Exposing deepfakes using convolutional neural networks and transfer learning approaches. 2020 IEEE 17th India Council International Conference (INDICON); 2020.

[20] El Rai MC, Al Ahmad H, Gouda O, Jamal D, Talib MA, Nasir Q. Fighting deepfake by residual noise using convolutional neural networks. 2020 3rd International Conference on Signal Processing and Information Security, ICSPIS 2020; 2020.

[21] Li X, Lang Y, Chen Y, Mao X, He Y, Wang S, et al. Sharp multiple instance learning for deepfake video detection. MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia; 2020:1864–72 [Internet]. Available from: http://arxiv.org/abs/2008.04585

[22] Zhang W, Zhao C, Li Y. A novel counterfeit feature extraction technique for exposing face-swap images based on deep learning and error level analysis. Entropy. 2020;22:249.

[23] Vizoso Á, Vaz-Álvarez M, López-García X. Fighting deepfakes: Media and internet giants' converging and diverging strategies against Hi-tech misinformation. Media Commun. 2021;9:291–300.

[24] Tran VN, Lee SH, Le HS, Kwon KR. High performance deepfake video detection on CNN-based with attention target-specific regions and manual distillation extraction. Appl Sci. 2021;11:7678.