*Grenze International Journal of Engineering and Technology, June Issue*

# Enhancing Resilience of Deepfake Detection using Adversarial Frequency Masking

Nilesh Seth[1], Nitin Patel[2] and Dr. Rajni Jindal[3]
[1-2]Department of Computer Science and Engineering, Delhi Technological University, Delhi, India
Email: {sethnilesh07, np68175}@gmail.com
[3]Professor, Department of Computer Science and Engineering, Delhi Technological University, Delhi, India
Email: rajnijindal@dce.ac.in

***Abstract*—This research presents an advanced approach to deepfake detection that combines frequency domain masking with adversarial training techniques to enhance the precision and resilience of the current models against adversarial attacks. The approach consists of advanced preprocessing of the images with Fourier and Wavelet transforms for feature extraction along with the creative adversarial training techniques. The approach's usefulness is demonstrated by experimental findings, especially in identifying difficult and realistic looking deepfakes. The study presents a comparative analysis between multiple models when trained using this technique and trained without using this technique. This research can be used to counteract deepfake threats and what future research should focus on in order to improve deepfake detection capabilities.**

***Index Terms*— Fake content, deepfake detection, frequency masking, adversarial training, deep learning.**

## I. INTRODUCTION

Deepfake technology has become a major problem in recent years. They negatively affect a number of industries including misinformation, cybersecurity, and privacy [1]. These alarming lifelike artificial media which are produced by deep learning algorithms have the ability to mislead people and influence public opinion. Thus, an increasing demand for trustworthy and dependable deepfake detection techniques is foreseeable. Recent developments in deep learning, computer vision etc. have contributed to AI revolution. Current deep learning techniques have shown great potential in improving precision and efficacy of deepfake detection systems [2]. The prominent methods like convolutional neural networks (CNNs) and Generative Adversarial Networks (GANs) have produced great results but further research is needed to determine how vulnerable they are to attacks. The main concern of current deepfake detection techniques is their limited capacity to identify more complex and lifelike [3].

We combine advanced frequency domain masking with adversarial training techniques to provide a comprehensive solution to the problem of identifying deepfakes. Through detailed frequency analysis and efficient adversarial training techniques, our method enhances robustness and accuracy in detecting sophisticated deepfakes. Conventional techniques might miss little but significant frequency-based patterns in deepfakes, on the other hand our method successfully detects and utilizes these patterns to discriminate between real and fake media. This new integration increases the model's capacity to generalize across many deepfake variations. Apart

*Grenze ID: 01.GIJET.10.2.566_2*

from the current deepfakes, our approach makes the model resilient against deepfakes that are made to outwit detection techniques while simultaneously maintaining a competitive detection accuracy.

## II. Related Works

Deepfake technology has raised multiple concerns due to its potential to deceive and manipulate individuals by creating realistic but fabricated content [4][5][6]. Researchers are actively introducing methods to detect deepfakes to reduce the risks involved. Multiple strategies have been put forth to deal with the problem of deepfake detection.

Early approaches like [7] focused on detecting AI-generated fake videos by analyzing eye blinking patterns. This method highlights the importance of minute details like eye movements in identifying deepfake content. [8] presented MesoNet, a compact facial video forgery detection network that focuses on mesoscopic properties of images. This method signifies the importance of considering distinct characteristics of deepfake content for accurate detection. Similarly, [9] utilized body language analysis to enhance detection accuracy by demonstrating superior results compared to standard approaches when evaluated with GAN-based deepfake datasets. [10] suggested a deepfake attack prevention strategy using steganography GANs. They combined blockchain and watermarking techniques for improved detection.

Recently, [11] highlighted the significance of manual identification of flaws in existing deepfake methods as a foundation for detection. They emphasized the role of neural networks in spotting anomalies in deepfake content. Similar approach was developed earlier in which an audiovisual method was used to detect inconsistencies between visual lip movements and speech in audio for deepfake identification. [12]. The importance of audio-visual synchronization was clearly evident in recognizing manipulated media content. Furthermore, [13] introduced an ensemble-based approach to increase defense robustness against GAN-based deepfake variations. This showcases the enhanced resilience against adversarial attacks.

Frequency domain masking has been utilized to improve deepfake detection by focusing on high-frequency components in images. [14] proposed a method that combines color features with frequency-domain-related features to enhance deepfake detection. [15] introduced a new method called frequency-based masking. It involves hiding parts of the image in the frequency domain before training a deep learning model to detect fake images. This approach helped in increasing the resilience of the existing SOTA models [16]. Additionally, frequency domain learning and optimal transport theory was applied in knowledge distillation to enhance the detection of low-quality compressed deepfake images [17]. These researches show that focusing mainly on high-frequency components helps to improve the resistance of detection models against manipulation.

With rapid growth in the generative models, the need for robust detectors have also increased significantly. Hence, performing the classification of image as real or fake without learning the existing patterns was proposed [18]. They use a feature space which is not explicitly trained to differentiate the real images from the fake ones. This adversarial training of the models is a strategy to enhance the robustness of deepfake detection [19]. Later it was discovered that training the model on corrupted data helps in increasing model's resistance to adversarial attacks. In [20], adversarial perturbations were used along with the model to enhance deepfake images and deceive common detection systems. Afterwords, a unified approach was proposed to leverage orthogonal gradients in order to prevent first-order adversaries and improve the robustness of deepfake detection models [21].

In summary, the research on deepfake detection has seen advancements in utilizing frequency domain masking and adversarial training techniques to enhance detection accuracy and robustness. Adversarial training techniques as claimed by researchers aim to make detection systems more resilient to adversarial attacks and improve overall detection performance.

## III. Methodology

### *A. Pre-processing*

Preprocessing steps as mentioned in Fig. 1 include resizing images to a fixed resolution (e.g., 224x224 for ResNet-50), normalization to ensure pixel values fall within a certain range (e.g., [0, 1]) along with augmentation techniques such as random crops, flips, and rotations. It increases the variability of the training data and prevent overfitting. Pre-processed images are then passed to AFM to generated masked images.

### *B. Adversarial Frequency Masking*

Adversarial Frequency Masking (AFM) is a technique to train the models on unseen data by masking certain frequency components in input images. Earlier, adversarial training methods operated in the spatial domain. But,

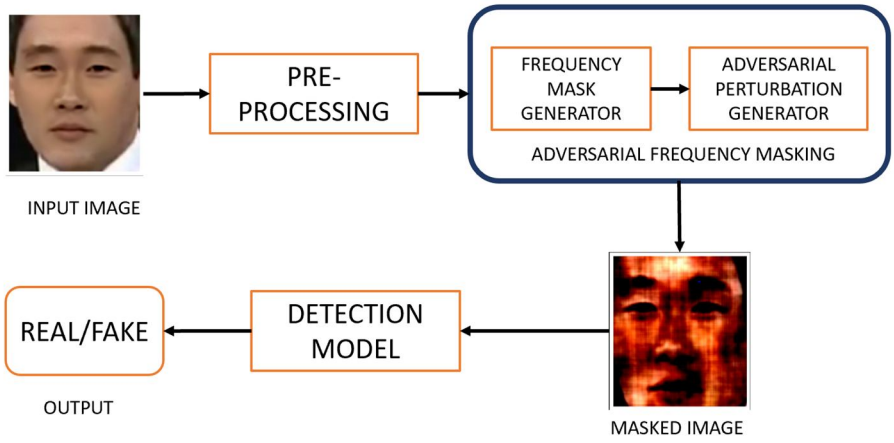


Figure 1: Deepfake Detection using Adversarial Frequency Masking

AFM utilizes the Fourier domain to introduce perturbations in input image I(x,y) and convert it to F(u,v). This disrupts adversarial attacks while preserving image content and interpretable quality simultaneously.

$$F(u,v) = \int\int I(x,y)e^{-i2\pi(ux+vy)}dxdy \qquad (1)$$

The Adversarial Frequency Masking framework consists of two main components: the Frequency Mask Generator and the Adversarial Perturbation Generator. The Frequency Mask Generator is utilized for generating frequency masks. Certain frequency components are removed from the Fourier transformed input images while generating their frequency masks. As a result, a dynamic mask M(u,v,α) is generated which adapts based on the frequency content of the image. This function will adjust the frequency mask based on the magnitude at each pixel in the Fourier-transformed image.

$$M(u,v,\alpha) = \frac{1}{1 + e^{-\alpha.|F(u,v)|}} \qquad (2)$$

In simpler words, it applies a stronger masking effect to pixels with higher frequency magnitudes and a relatively weaker effect to those with lower magnitudes. An adaptive approach like this enhances the model's capability to distinguish between fine variations in frequency content across the set of images. It potentially boosts its efficacy in detecting deepfakes with greater accuracy and robustness. The mask is then combined with the modified image *F(u, v)*. This new image is to Adversarial Perturbation Generator.

$$Fmasked(u,v) = F(u,v) \times M(u,v,\alpha) \qquad (3)$$

The Adversarial Perturbation Generator adds fine changes to the frequency domain. These changes disrupt attacks from adversaries more effectively. However, the changes have minimal effect on image content. The Generator introduces these perturbations carefully to resist adversaries while preserving image quality.

$$P(u,v) = \epsilon \qquad (4)$$

*C. Model Architecture*

Our training module utilizes the ResNet [22] architecture. Particularly when it comes to tasks like image recognition, ResNet is renowned for its effectiveness at processing large amounts of data and organizing it neatly. We are experimenting with three different ResNet versions: an 18-layer version, a 50-layer version, and an even 101-layer version. This allows us to see how the number of layers we add impacts the system's ability to discriminate between objects in our tests.

ResNet-50 is constructed with 50 layers. These- layers have convolutions, pooling, and fully connected sections. ResNet-50 employs residual blocks to overcome vanishing gradie-nts during the training process. This innovation enable-s training very deep neural networks.

The architecture has shortcut connections bypassing one or more layers. These shortcuts allow direct gradient flow during backpropagation. They reduce the degradation issue caused by increasing network depth. Thanks to these features, ResNet-50 achieves top performance on image classification benchmarks.
We customized our base ResNet architecture by adding below mentioned modifications that suits our experimental setup:

- Prediction Layer Addition: A linear layer is added at the end to the output of ResNet for binary classification tasks. This additional layer enables the model to make binary predictions based on the extracted features from the ResNet backbone.
- Integration with Frequency Masking: The frequency masking mechanism is integrated into the preprocessing pipeline of the model. Images undergo Fourier transformation and are masked using the generated frequency masks before feeding them into ResNet for training or inference. This process perturbs the input images in the frequency domain and introduces adversarial robustness into the training process.

When we add Adversarial Frequency Masking to ResNet, we're essentially making it more aware of tricky situations. This is like giving it special glasses to see through potential tricks in the images it's looking at, which is important for training it to be smarter.
We chose ResNet because it's good at handling lots of information and can learn from examples really well. Its ability to stay focused and not get confused easily makes it a good fit for our project, especially when we're training it to be extra careful about spotting deceptive things in pictures.

## IV. Experimental Setup

### *A. Dataset*

The standard image classification datasets are used in the experimental setup, such as CIFAR-10 and CIFAR-100 [23]. These datasets contain a large number of natural images across multiple classes. Hence these datasets are suitable for training and evaluating deep learning models. Before training, the images undergo standard preprocessing steps such as resizing to a uniform size, normalization, and augmentation to increase the diversity of the training data.

### *B. Model Setup*

We utilized the chosen Resnet model variations and trained the model on CIFAR dataset. Training the model involves two distinct processes: training it once without Adversarial Frequency Masking (AFM) and then training it again with AFM. In the first scenario, the model is set up without incorporating AFM. For each variation of Resnet i.e. Resnet 18, 50 and 101, we follow the same procedure of training them in two different scenarios. This is done to showcase the significance of our approach in increasing the precision and robustness of the models.

### *C. Hyperparameters*

Following hyperparameters were chosen while training the models:

- Learning Rate: Chosen empirically through hyperparameter tuning or based on previous studies, typically ranging from 0.001 to 0.01.
- Batch Size: Determined based on available memory resources and computational efficiency, commonly ranging from 32 to 128.
- Number of Epochs: Typically set to a fixed value or determined using early stopping criteria.

### *D. Early stopping and Model Checkpointing*

In order to prevent overfitting, early stopping is used to track the validation loss during training and stop the training process when the validation loss stops decreasing. In the event of disruptions or failures, the model can be recovered to its optimal performance state thanks to model checkpointing, which saves the model parameters on a regular basis during training.

## V. Training Procedure

### *A. Forward and Backward Passes*

During each training iteration, input images are fed through the model, and predictions are generated. The loss between the predicted labels and ground truth labels is then computed using a chosen loss function. This loss is

used to perform backpropagation, where gradients are computed with respect to the model parameters, and optimization algorithms update the parameters to minimize the loss.

### *B. Optimization Strategy*

The $Adam_W$ optimizer [24], an extension of the Adam optimizer, is commonly used for training deep learning models. $Adam_W$ has a built-in weight decay mechanism, which helps prevent overfitting by penalizing large parameter values.

### *C. Loss Function*

For binary classification tasks (e.g., presence or absence of a certain attribute in images), the Binary Cross-Entropy with Logits Loss ($BCE_{WithLogitsLoss}$) [25] is commonly used. BCE function uses a pair of sigmoid activation function and binary cross-entropy loss which is suitable for binary classification tasks.

### *D. Evaluation Metrics*

Our deepfake detection model undergoes careful evaluation. Following model training, we validate its effectiveness using a separate validation dataset. This is while carefully monitoring performance to prevent overfitting. The testing phase involves assessing the model on an entirely new testing dataset. The evaluation of model on unseen data will help in assessing the model's robustness and its unbiased evaluation. We evaluate mean Average Precision (mAP) to provide a metric to estimate our model's ability to detect deepfakes accurately and reliably.

## VI. Results and Discussion

Table I shows the evaluated mAP of various Resnet Models which were trained with and without using Adversarial Frequency Masking on the input images. The experimental findings clearly indicate the degree to which our method is effective in improving the precision and robustness of the models that are trained without it.

TABLE I. EXPERIMENTAL RESULTS

| **Model** | | **Mean Average Precision (mAP)** |
|---|---|---|
| ResNet 18 | without AFM | 0.52 |
| | with AFM | 0.57 **(+9.6%)** |
| Resnet 34 | without AFM | 0.67 |
| | with AFM | 0.72 **(+7.5%)** |
| Resnet 50 | without AFM | 0.79 |
| | with AFM | **0.83 (+5.1%)** |

Starting with ResNet 18 model, which has least depth, we observe that, in comparison to the scenario where AFM wasn't used its mean average precision took a serious growth of 9.6% when trained using AFM. Similar growth can be observed in the next model, i.e., ResNet 34 which outperformed its precision by 7.5% whereas ResNet 50 improved its precision by 5.1% while also achieving the best performance in comparison to other two models. The incorporation of enhanced frequency domain analysis and adversarial techniques significantly enhances the model's ability to detect subtle alterations characteristic of deepfake imagery.

Furthermore, the adversarial training component contributes to the model's resilience against adversarial attacks. Thus, it ensures robustness in real-world scenarios where sophisticated deepfake techniques are employed to evade detection. However, it is crucial to acknowledge the constraints of our research, such as the requirement for ongoing modifications and adjustments as deepfake methodologies advance. With proper resources and equipment for research we should get even better performance which would be comparable to state of the art methods.

## VII. Conclusion

Our research findings demonstrate the considerable advantages of employing adversarial frequency masking in deepfake detection models. Our models showcase improved resilience against adversarial attacks after

integration of this approach. The results display a significant improvement in their ability to detect and combat fake content. This is a critical step forward in the ongoing battle against the spread of misinformation and digital deception. This advancement will surely improve the systems to more effectively identify and label potentially harmful deepfakes.

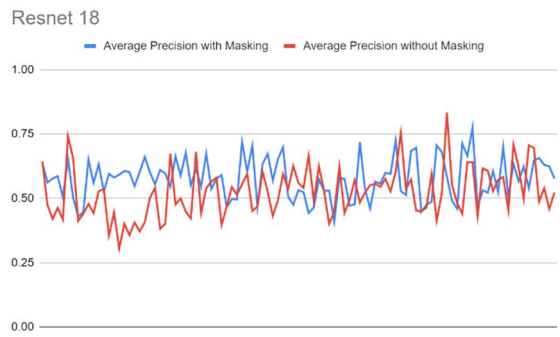


Figure 2: Variation of Average Precision over multiple validation epochs for Resnet 18 when trained with AFM and without AFM

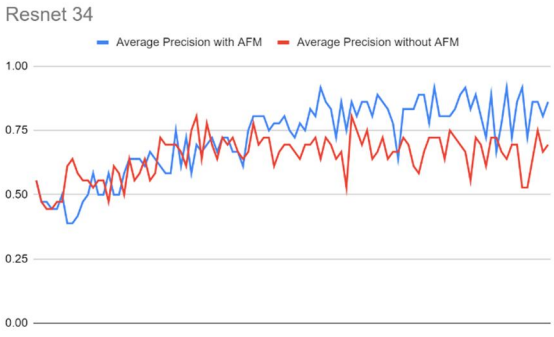


Figure 3: Variation of Average Precision over multiple validation epochs for Resnet 34 when trained with AFM and without AFM

Apart from increasing resilience, our approach maintains competitive precision levels when dealing with clean, unaltered data. Models trained with adversarial frequency masking display higher precision in identifying authentic media content. This balance between precision and robustness is essential for building trust in deepfake detection technologies and strengthen their credibility for widespread adoption.

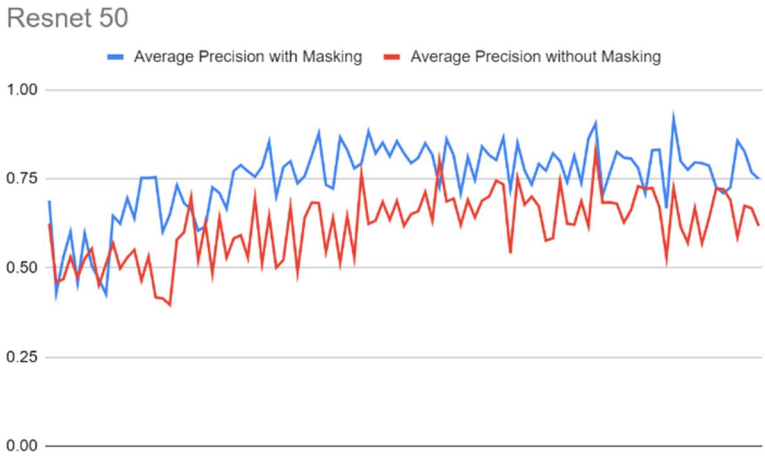


Figure 3 : Variation of Average Precision over multiple validation epochs for Resnet 50 when trained with AFM and without AFM

Our research showcased major progress in detecting deepfakes through adversarial frequency masking. However, we must consider certain constraints. Applying frequency masking on large datasets or real time detection would increase the computational resources load. This might affect scalability and implementation feasibility in resource limited settings. Furthermore, how well this approach works depends on distinctive and specific deepfakes traits. Highly sophisticated deepfakes that mimic real media may still puzzle detectors.
Despite these challenges, our approach represents a major milestone in combating the spread of misinformation while also maintaining digital integrity. Future work will focus on addressing these limitations of our approach, how can we improve their scalability factor, and developing enhancements to increase the adaptability of our model to emerging deepfake threats. Through continued refinement and innovation, we aim to strengthen the reliability and practicality of deepfake detection technologies in real-world applications.


## ACKNOWLEDGMENT

This project was carried out as the Major Project of our Bachelor's Degree in Computer Engineering. We would like to express our sincere gratitude to Delhi Technological University for their support throughout the research.

We also extend our appreciation to Dr. Rajni Jindal for their invaluable guidance and support throughout the research process. Lastly, we thank the participants who contributed their time and data for this study.

## REFERENCES

[1] C. -C. Chang, H. H. Nguyen, J. Yamagishi and I. Echizen, "Cyber Vaccine for Deepfake Immunity," in IEEE Access, vol. 11, pp. 105027-105039, 2023, doi: 10.1109/ACCESS.2023.3311461.
[2] Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. WIREs Data Mining and Knowledge Discovery, 14(2), e1520. https://doi.org/10.1002/widm.1520
[3] C. Liu, H. Chen, T. Zhu, J. Zhang and W. Zhou, "Making DeepFakes More Spurious: Evading Deep Face Forgery Detection via Trace Removal Attack," in IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 6, pp. 5182-5196, Nov.-Dec. 2023, doi: 10.1109/TDSC.2023.3241604.
[4] Dr. Pawan Singh , Dr. Bharat Dhiman . Exploding AI-Generated Deepfakes and Misinformation: A Threat to Global Concern in the 21st Century. TechRxiv. December 05, 2023.
[5] Dr. A. Shaji George, and A. S. Hovan George. 2023. “Deepfakes: The Evolution of Hyper Realistic Media Manipulation”. Partners Universal Innovative Research Publication 1 (2):58-74. https://doi.org/10.5281/zenodo.10148558
[6] Mekhail Mustak, Joni Salminen, Matti Mäntymäki, Arafat Rahman, Yogesh K. Dwivedi, Deepfakes: Deceptions, mitigations, and opportunities, Journal of Business Research, Volume 154, 2023, 113368, ISSN 0148-2963, https://doi.org/10.1016/j.jbusres.2022.113368
[7] Li, Yuezun & Chang, Ming-Ching & Lyu, Siwei. (2018). In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. 1-7.
[8] Afchar, Darius & Nozick, Vincent & Yamagishi, Junichi & Echizen, I.. (2018). MesoNet: a Compact Facial Video Forgery Detection Network. 1-7. 10.1109/WIFS.2018.8630761.
[9] Yasrab, Robail, Wanqi Jiang, and Adnan Riaz. 2021. "Fighting Deepfakes Using Body Language Analysis" Forecasting 3, no. 2: 303-321. https://doi.org/10.3390/forecast3020020.
[10] Noreen I, Muneer MS, Gillani S. 2022. Deepfake attack prevention using steganography GANs. PeerJ Computer Science 8:e1125 https://doi.org/10.7717/peerj-cs.1125
[11] Gupta, Gourav, Kiran Raja, Manish Gupta, Tony Jan, Scott Thompson Whiteside, and Mukesh Prasad. 2024. "A Comprehensive Review of DeepFake Detection Using Advanced Machine Learning and Fusion Methods" Electronics 13, no. 1: 95. https://doi.org/10.3390/electronics13010095.
[12] Korshunov, Pavel & Marcel, Sébastien. (2019). Vulnerability assessment and detection of Deepfake videos. 1-6. 10.1109/ICB45273.2019.8987375.
[13] Yang, C., Ding, L., Chen, Y., & Li, H. (2020). Defending against GAN-based Deepfake Attacks via Transformation-aware Adversarial Faces. *ArXiv*. /abs/2006.07421
[14] Fang Sun, Niuniu Zhang, Pan Xu, Zengren Song, "Deepfake Detection Method Based on Cross-Domain Fusion", Security and Communication Networks, vol. 2021, Article ID 2482942, 11 pages, 2021. https://doi.org/10.1155/2021/2482942
[15] Doloriel, Chandler Timm, and Ngai-Man Cheung. "Frequency Masking for Universal Deepfake Detection." arXiv preprint arXiv:2401.06506 (2024).
[16] Liu, Decheng, Tao Chen, Chunlei Peng, Nannan Wang, Ruimin Hu, and Xinbo Gao. "Attention Consistency Refined Masked Frequency Forgery Representation for Generalizing Face Forgery Detection." arXiv preprint arXiv:2307.11438 (2023).
[17] Le, B. M., & Woo, S. S. (2021). ADD: Frequency Attention and Multi-View based Knowledge Distillation to Detect Low-Quality Compressed Deepfake Images. *ArXiv*. /abs/2112.03553
[18] U. Ojha, Y. Li, and Y. Lee, “Towards universal fake image detectors that generalize across generative models,” in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24480–24489.
[19] Zhiyuan Yan, Yong Zhang, Yanbo Fan, Baoyuan Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22412-22423
[20] Gandhi, Apurva, and Shomik Jain. "Adversarial perturbations fool deepfake detectors." In 2020 International joint conference on neural networks (IJCNN), pp. 1-8. IEEE, 2020.
[21] Hooda, Ashish & Mangaokar, Neal & Feng, Ryan & Fawaz, Kassem & Jha, Somesh & Prakash, Atul. (2022). Towards Adversarially Robust Deepfake Detection: An Ensemble Approach.
[22] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.
[23] Krizhevsky, A., Nair, V. and Hinton, G. (2014) The CIFAR-10 Dataset. https://www.cs.toronto.edu/~kriz/cifar.html
[24] Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." arXiv preprint arXiv:1711.05101 (2017).
[25] Ruby, Usha & Yendapalli, Vamsidhar. (2020). Binary cross entropy with deep learning technique for Image classification. International Journal of Advanced Trends in Computer Science and Engineering. 9. 10.30534/ijatcse/2020/175942020.