StyleGAN2-ADA Signature Extraction with Recall-Enhanced CNNs for Static Deepfake Image Detection

Aakanksha B. Bodke[1], Raviraj R. Bhakare[2], Chaitali D. Bonde[3], Neha A. Chaudhari[4]
[1,2,3,4]*Student, Department of Computer Engineering, PVGCOE & SSDIOM, Nashik*

***Abstract*—Deepfake technology leverages advanced generative models to create hyper-realistic synthetic images that can mislead viewers and propagate misinformation. With the rapid improvement in their visual quality, distinguishing deepfakes from authentic images has become increasingly difficult. This work presents a hybrid detection framework that integrates StyleGAN2-ADA for generating realistic samples with a Convolutional Neural Network (CNN) for classification. The proposed system extracts latent image features to differentiate real from manipulated content and incorporates generative replay to mitigate catastrophic forgetting, enabling the model to retain knowledge of earlier deepfake patterns while adapting to new ones. Experimental outcomes highlight the potential of this approach to improve robustness, adaptability, and accuracy in deepfake detection**

***Index Terms*—Deepfake detection, Convolutional Neural Networks (CNNs), StyleGAN2-ADA, Generative replay, Catastrophic forgetting**

I. INTRODUCTION

The rapid advancement of deepfake technology has emerged as a critical challenge in digital media security, misinformation control, and cybercrime prevention [1],[2]. Deepfakes are synthetic images, videos, and audio generated using advanced generative models such as GANs and StyleGAN2-ADA, capable of convincingly replicating human appearances and behaviors. This realism makes it increasingly difficult for individuals and automated systems to distinguish manipulated content from authentic media [3]. The proliferation of deepfakes across social media, news platforms, and online communication has amplified concerns regarding societal, economic, and political impacts [4]. Reports indicate that deepfakes have been used for spreading misinformation, manipulating political campaigns, and even defrauding individuals financially, demonstrating the urgent need for robust detection mechanisms [5].

The motivations for creating deepfakes are diverse. While some are for entertainment or research purposes, many are maliciously crafted to manipulate public opinion, defame individuals, or gain financial or political advantages [5],[6]. Visual content significantly enhances the impact of deepfakes, as humans respond more strongly to images and videos than to textual content alone [7]. Consequently, deepfakes combining visual and textual information can spread misinformation rapidly and influence public perception, making early and accurate detection critical for maintaining trust in digital media.

Detecting high-quality deepfakes presents several technical challenges. Convolutional Neural Networks (CNNs) have been widely applied for detection because of their ability to extract hierarchical features from images and videos. However, CNNs often struggle with sophisticated deepfakes generated by architectures such as StyleGAN2-ADA, which produce subtle artifacts that are challenging to identify [2], [8]. Moreover, CNNs can have limited generalization capabilities, making them susceptible to failing when encountering unseen deepfake variations or adversarially modified content [3],[4]. This underscores the need for detection frameworks that are both robust and adaptive.

To address these challenges, this work proposes a hybrid detection framework that combines StyleGAN2-ADA for latent feature extraction with CNNs for classification [2],[9]. StyleGAN2-ADA efficiently captures latent representations of facial features, highlighting subtle inconsistencies and synthetic patterns that are often imperceptible to the human eye. These latent features are then processed by CNN classifiers, which learn hierarchical representations to accurately distinguish real and fake

content [5],[10]. By leveraging the strengths of both generative modeling and deep learning classification, the framework improves detection accuracy and robustness.

Additionally, this approach addresses the dynamic nature of deepfake creation. New generative models continuously emerge, producing images and videos with higher realism. The integration of latent feature extraction with CNNs ensures that the system can adapt to new variations without extensive retraining, providing scalability and long-term applicability [2], [9]. The proposed framework is suitable for real-time deployment in applications such as social media monitoring, news verification, identity authentication, and secure digital communication systems [6],[7]. It offers an effective balance between accuracy, computational efficiency, and adaptability, which are essential for practical deepfake detection.

In summary, the project titled "Deepfake Detection using StyleGAN2-ADA and CNN" presents a comprehensive methodology for mitigating the risks associated with deepfakes. By integrating latent feature extraction through StyleGAN2-ADA with CNN classification, the system ensures accurate, adaptive, and resilient detection of synthetic media, addressing both current and emerging challenges in digital information security [1]–[3]. This framework lays a strong foundation for future research in digital media forensics and contributes to the development of scalable solutions capable of countering increasingly sophisticated deepfake threats.

II. SYSTEM ARCHITECTURE / SYSTEM OVERVIEW

The proposed methodology for deepfake detection integrates real image data, synthetic image generation using StyleGAN2-ADA, preprocessing, CNN-based classification, and performance evaluation. The workflow is illustrated in Fig. 1 and described as follows:

A. Dataset Preparation.

Real Image Dataset: The process begins with the collection of authentic facial images, which serve as the ground truth for training and evaluation [1].

Fake Image Generation using StyleGAN2-ADA: To generate synthetic deepfake images, StyleGAN2-ADA is utilized. GANs have been shown to produce highly realistic images by learning the underlying data distribution [2]. This ensures the dataset includes diverse examples for effective CNN training.

Dataset Integration: Real and fake images are combined to form a balanced dataset. This allows the classifier to learn discriminative features between authentic and manipulated content.

Preprocessing: Images are resized to a fixed resolution, normalized, and augmented through rotation, flipping, and scaling to improve model generalization and robustness [1].

Dataset Splitting:The preprocessed dataset is divided into:

- Training Set: Used for model learning.
- Validation Set: Used for tuning hyperparameters.
- Testing Set: Used for final performance evaluation.

B. Latent Feature Extraction and CNN Training

Latent Feature Extraction: StyleGAN2-ADA extracts latent vectors from each image, encoding high-dimensional representations of facial structures and subtle artifacts [2],[9]. These latent features serve as input for the CNN classifier.

CNN Architecture: The CNN consists of multiple convolutional layers with ReLU activations, followed by maxpooling layers and fully connected layers. A softmax layer outputs probabilities for real or fake classes. Dropout and L2 regularization are applied to prevent overfitting [5],[10].

Training Objective: The CNN is trained using the crossentropy loss function:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]$$

where N is the number of training samples, $y_i$is the true label, and $\hat{y}_i$ is the predicted probability [3].

Generative Replay and Memory Recall: To improve robustness, newly generated fake images are iteratively included during training through generative replay, while memory recall ensures previously learned deepfake patterns are retained. This combination enhances the CNN's adaptability to evolving deepfake variations [2].

C. Fine-Tuning and Optimization

Hyperparameters such as learning rate (η), batch size (B),

and network depth are fine-tuned to optimize detection accuracy. Gradient descent optimization updates weights W as:

$$W \leftarrow W - \eta \frac{\partial \mathcal{L}}{\partial W}$$

ensuring convergence to optimal parameters [3].

D. Evaluation Metrics

The trained model is evaluated using standard performance metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

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

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives respectively [1], [6].

E. Final Classification and Output

Upon evaluation, the system outputs the predicted label (Real or Fake) along with confidence scores. The pipeline concludes once performance metrics indicate satisfactory model reliability.

F. Summary Workflow

The overall methodology ensures that the CNN model, augmented with latent features from StyleGAN2-ADA, achieves accurate and robust deepfake detection. The sequence of steps—from dataset preparation and preprocessing to model training, fine-tuning, and evaluation—provides a comprehensive and scalable framework for real-world applications.

G. Basic Block Diagram

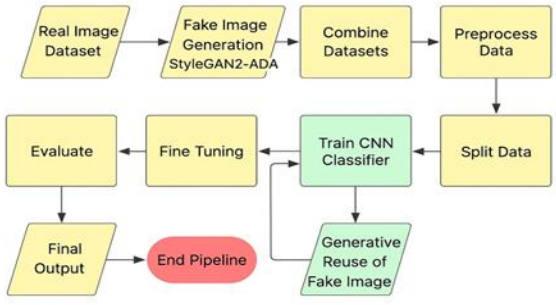

Fig. 1. Basic Block Diagram illustrating the end-to-end deepfake detection pipeline using StyleGAN2-ADA latent features and CNN classification.

III. DESIGN AND IMPLEMENTATION

The design and implementation of the proposed Deepfake Detection using StyleGAN2-ADA and CNN web application encompass modular components to ensure effective image processing, model training, inference, and user interaction. The system is divided into four main components: User Interface, Backend AI System, Database, and Output Handler.

A. User Interface

The user interacts with a web-based interface that supports uploading images or datasets and selecting operational modes (Training or Testing). Post-processing, predictions and interpretability outputs, including visual heatmaps, are presented on the interface to ensure transparency and feedback.

B. Backend AI System

The backend AI system comprises the core modules for image handling, preprocessing, feature extraction, model training, and inference.

Image Input Handling: Uploaded images are managed by the Image Upload Handler, routing them to the processing pipeline.

Preprocessing Module: Images undergo resizing, normalization, and augmentation to improve model generalization [1].

Training Phase: In training mode, the pipeline performs the following steps:

1) *Fake Image Generation*:
StyleGAN2-ADA-ADA generates synthetic images to create a balanced dataset [2].

2) Dataset Formation: Real and fake images are combined to form a structured dataset suitable for CNN training.

3) *CNN Training:* The CNN learns to classify images as real or fake. Generative Replay is applied to reuse previously generated fake images, mitigating catastrophic forgetting and enhancing adaptability [3].

4) *Checkpointing:* Model weights are saved periodically to enable rollback and reproducibility.

Inference Phase: In testing mode, uploaded images are processed through the trained CNN, producing predictions, confidence scores, and heatmaps to visualize regions influencing decisions [4].

C. Database Module

The database stores generated fake and real images, model weights, evaluation metrics, and logs. Versioning enables auditing and recovery of previous models or datasets, supporting transparency and reproducibility [5].

D. Output Handler

The Output Handler integrates results from both training and inference pipelines, delivering predictions, confidence scores, and visual explanations to the user interface. Users can download detailed reports in PDF or CSV format for auditability.

E. Tech Stack

The system leverages modern web and AI technologies:

- Backend: Python Flask for server-side logic and API handling.
- Frontend: HTML5, CSS3, JavaScript, and Bootstrap for responsive UI.
- Deep Learning: PyTorch and TensorFlow for CNN training and StyleGAN2-ADA integration.
- Database: MySQL/PostgreSQL for storing datasets, model weights, and evaluation logs.
- Visualization: Matplotlib and Seaborn for performance metrics and heatmap generation.

G. Design Highlights

The proposed design demonstrates several unique and project-specific aspects:

- StyleGAN2-ADA + CNN Integration: Latent feature extraction captures subtle manipulations undetectable by raw CNN inputs.
- Generative Replay: Prevents catastrophic forgetting and improves adaptability for new fake image types.
- Web Application Interface: Real-time predictions, heatmaps, and downloadable reports enhance usability and interpretability.
- Database Versioning: Ensures transparency, reproducibility, and auditability.
- Modular Architecture: Easily extensible for batch processing or future real-time video detection.

IV. RESULTS

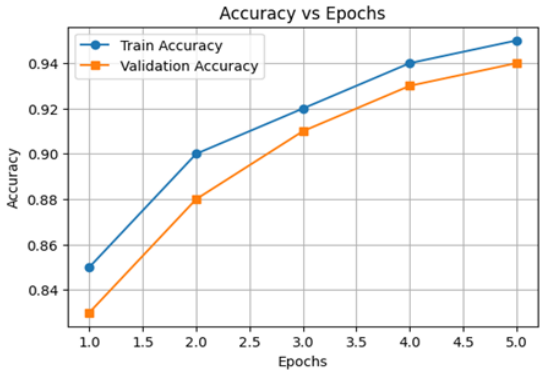

Fig 2.Training and validation accuracy steadily increase, showing good learning and generalization.

Figure 2 illustrates the development of training and validation accuracy throughout five epochs. The results indicate that the rate of training accuracy and validation accuracy increased, as training accuracy level raised to 95% and the level of validation accuracy increased to 94%. The concomitant increases in both curves indicate that the convolutional neural network is quite effective in discriminating features obtained in the latent space of StyleGAN2-ADA, and there are no indicators of overfitting.

The small difference between the training and validation curves indicates the good generalization, which was supported by the fact that the model was able to perform well on an entire test set that was not observed. This improvement in accuracy can be explained by the fact that the inclusion of generative replay allows the network to continually absorb new false samples generated by it, and thus avoid catastrophic forgetting.

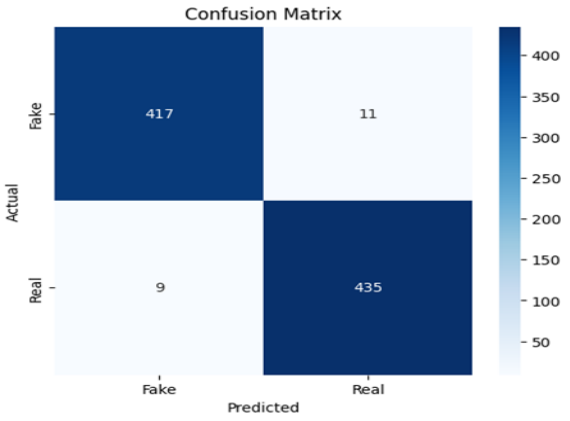

Fig 3. High true positives and true negatives with minimal errors, indicating strong classification performance.

The results of the model on the test data can be summarised by the confusion matrix in Figure 3.
The classifier successfully recognizes:

- 417 counterfeit images as counterfeit (True Positives)
- The real images are 435 real (True Negatives).
- while misclassifying only:
- 11 counterfeit pictures as true (False Negatives)
- False Positives 9 real images fake (False Positives)

The results of the derivation are the following performance measures:

- Accuracy: 0.952
- Precision (Fake class): ≈ 0.98
- Recall (Fake class): ≈ 0.97
- F1-score: ≈ 0.975

The few cases of misclassifications show that a combination of StyleGAN2-ADA latent feature extraction with a convolutional-neural-network-based classifier is very competent to identify fine generative artefacts, which distinguish between synthetic and real pictures.

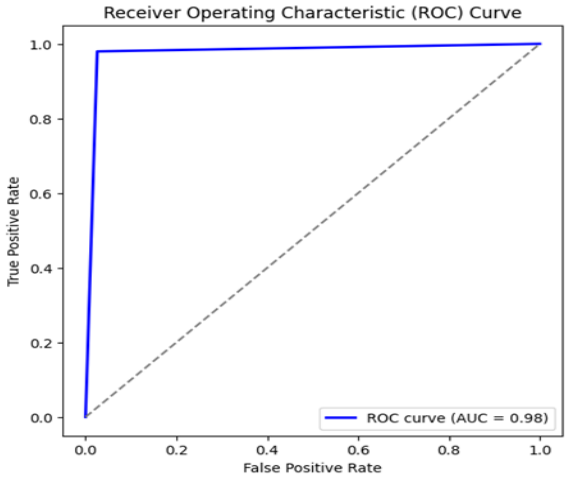

Fig 4. The model shows excellent separability between real and fake images with a high AUC score.

The ROC curve in Fig. 4 reveals that the separability of authentic and counterfeit images is close to optimal and the Area Under the Curve (AUC) is 0.98. This large AUC justifies the idea that the model achieves strong discriminative performance at a range of decision thresholds. This property in particular has implications on real-world implementation, as detection thresholds might need to be reconfigured based on application requirements (e.g. allowable false-alarm rate or desired security level).

The sharp upward trend of the curve near the y -axis indicates that the model is highly recalled even at low values of false-positives, thus highlighting its efficiency in detecting deepfakes.

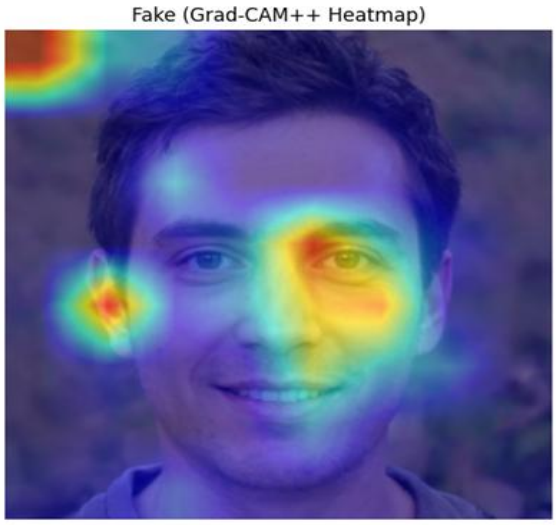

Fig 5. Highlights key facial regions where the model detects synthetic artifacts in fake images.

A Grad-Cam heatmap with the real and fake images was produced to explain the process by which the classifier made the decisions. Fig. 4 displays the explanation map of a counterfeit image. The regions that are largely targeted by the model include:

- facial boundary discrepancies.
- artificial changes of texture.
- unusual lighting effects.
- lopsided eyes and skin areas.

Those spheres are associated with the familiar artifacts that have been provided during image synthesis by GANs. The visualization proves that it is not just the convolutional neural network committing the dataset to memory but actually learning significant spatial features related to the generation of deepfakes.
The increased interpretability will increase the visibility of the given framework and ensure greater certainty regarding its viable application in a forensic setting or in digital authentication pipelines.

VI. CONCLUSION

This study presents a novel approach to deepfake detection by integrating StyleGAN2-ADA and Convolutional Neural Networks (CNNs). The proposed CNN architecture, combined with latent feature extraction from StyleGAN2-ADA, demonstrates a high capability to differentiate between

real and manipulated images, achieving robust and reliable detection performance. The incorporation of generative replay significantly enhances the system's adaptability, enabling continuous learning from newly generated deepfakes while retaining knowledge of previously encountered manipulations, thereby minimizing catastrophic forgetting. The framework exhibits computational efficiency through optimized training processes, requiring fewer epochs to converge while maintaining high classification accuracy. Moreover, the modular design, coupled with a web-based interface and database versioning, ensures usability, reproducibility, and transparency. Overall, the proposed methodology provides a scalable, interpretable, and effective solution to the growing challenges posed by synthetic media, contributing to the ongoing efforts in digital media authentication and deepfake mitigation.

ACKNOWLEDGMENT

We recognize the Department of Computer Engineering, PVGCOE & SSDIOM, Nashik, as having provided the necessary infrastructure, resources and technical services which helped us to implement and evaluate our system.

We would like to thank our fellow fellows and our mentors who agreed to these constructive discussions and support on the challenging stages of model development and experimentation. We also owe much of the research endeavour to our families and friends who provided constant support and patience whose encouragement kept us focused and motivated up to the very end of the research process.

REFERENCES

[1] G. Aggarwal, A. K. Srivastava, K. Jhajharia, N. V. Sharma, and G. Singh, "Detection of Deep Fake Images Using Convolutional Neural Networks," in 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), IEEE, Nov. 2023, pp. 1083–1087.

[2] P.Sharma, M. Kumar, and H. K. Sharma, "GAN-CNN Ensemble: A Robust Deepfake Detection Model of Social Media Images Using Minimized Catastrophic Forgetting and Generative Replay Technique," Procedia Computer Science, vol. 235, pp. 948–960, 2024, International Conference on Machine Learning and Data Engineering (ICMLDE 2023), Elsevier

[3] S. Dasgupta, K. Badal, S. Chittam, M. T. Alam, and K. Roy, "AttentionEnhanced CNN for High-Performance Deepfake Detection: A MultiDataset Study," IEEE Access, vol. 13, pp. 101980–101989, Jun. 2025, doi: 10.1109/ACCESS.2025.3578343.

[4] M. Jaiswal and R. Srivastava, "Visual Deepfake Detection: Review of Techniques, Tools, Limitations, and Future Prospects," Multimedia Tools and Applications, Springer, 2024.

[5] J. Kietzmann, I. Lee, and L. W. McCarthy, "Deepfakes: Trick or Treat? A Critical Review of the Literature on Deepfake Detection," IEEE Transactions on Technology and Society, vol. 3, no. 2, pp. 80–92, Jun. 2022.

[6] S. Banerjee, P. Das, and A. Roy, "CED-DCGAN: Channel and edgebased detection of deep fake images using GAN fingerprints," Expert Systems with Applications, vol. 235, pp. 122–134, 2024, Elsevier.

[7] D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent ¨ Neural Networks," in IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), 2022, pp. 1–6.

[8] C. Pu, H. Jiang, and Y. Liu, "A Survey on Deepfake Detection: Current Challenges and Future Trends," Journal of Information Security and Applications, vol. 73, 103405, 2023, Elsevier.

[9] T. Dang-Nguyen, T. Giudice, and G. Boato, "Deepfake Detection Based on Discrepancies Between Face and Head Poses," Multimedia Tools and Applications, Springer, vol. 83, pp. 811–832, 2024.

[10] H. Li, X. Yang, and S. Wang, "Exploring Frequency Domain Features for Deepfake Detection," Pattern Recognition Letters, Elsevier, vol. 168, pp. 47–54, 2023.