SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

AI-Generated Facial Image Detection Using GAN-Based Synthetic Data and ResNet-50 Transfer Learning, with Real-Time Browser Deployment

Deepti Sharma[1], Omika Bhalla Saluja[2], Shalika Grace Das[3], Sana Jafar[4], Ranjana Tiwari[5]

[1]Department of AIML, Pranveer Singh Institute of Technology, Kanpur. E-mail: deepti.1307@gmail.com
[2]Department of Business Administration, Pranveer Singh Institute of Technology, Kanpur. E-mail: omikasaluja@gmail.com
[3]Department of Business Administration, Pranveer Singh Institute of Technology, Kanpur. E-mail: shalika.das@psit.ac.in
[4]Department of AI, Pranveer Singh Institute of Technology, Kanpur. E-mail: india.sana@gmail.com
[5]Department of Business Administration, Pranveer Singh Institute of Technology, Kanpur. E-mail: ranjana.tiwari@psit.ac.in

ABSTRACT

Generative Artificial Intelligence (Gen AI) has enabled the creation of clear, photo-realistic synthetic images that, in many cases, are indistinguishable from real photographs, allowing anyone to generate them. This has posed significant challenges in the area of digital media forensics, identity protection, cybercrime and fraud prevention, misinformation detection and image authenticity verification. In this work, a deep learning-based system consisting of a CNN model is designed to classify the facial images into two classes: AI-generated and Real. The CelebA dataset is used as the source of real images, while synthetic images are produced using a custom Generative Adversarial Network (GAN). The image classification is accomplished by fine-tuning a convolutional neural network (CNN) called ResNet-50. The proposed classifier attained the accuracy of 96.97% and Area under the ROC Curve (AUC) of 0.977 on the internally balanced test set of 4384 images, and maintained the accuracy of 94.0 % on the externally independent test set of 2041 images, thereby exhibiting strong generalization. This research offers a practical and efficient tool for the detection of images and verification of the authenticity of media, which is supported by Artificial Intelligence.

Keywords: Generative Artificial Intelligence (Gen AI); AI-Generated Image Detection; Facial Image Classification; ResNet-50; Generative Adversarial Network (GAN); Image Authenticity Verification

1. Introduction

In recent years, the automatic synthesis of human face images with photo-realistic qualities has become possible, at least at a glance, thanks to advances in deep generative modelling, specifically Generative Adversarial Networks (GANs) and diffusion models. These AI-generated images, known as “deepfakes” or “synthetic media”, are becoming more popular in the entertainment, digital marketing and content production industries. Simultaneously, the same technology is used for the malicious manipulation of identity, impersonation, content creation without consent and mass distribution of misinformation across social media. Though visual artifacts like lights that shift, textures that look unnatural, or faces that are asymmetric have been a hallmark of AI-generated content, increasingly they're too obscure to be seen by the naked eye, as synthetic image generators like StyleGAN, Progressive GAN and diffusion-based models get better. This has transformed the capabilities of automated, learning-based detection into a key part of digital forensics, cybersecurity, journalism and trust and safety on online platforms.

Despite prior efforts on handcrafted features, CNNs, and transformers largely working independently, few works have simultaneously used a controlled, self-generated GAN dataset and fine-tuned a ResNet-50 based classifier and validated their model on a truly independent external dataset to evaluate real-world generalization performance, not just in-distribution performance. In addition, most of the studies so far carried out end at the offline evaluation and fail to transfer the trained model to a convenient end-user tool available in real-time. This work aims to fill both of them by: (i) creating and balancing a CelebA-plus-GAN dataset, (ii) fine-tuning the ResNet-50 network using transfer learning and testing it on an external, unseen dataset, and (iii) running the fine-tuned network as an extension plug-in for the Chrome browser to verify authenticity while viewing the images in real time.

Table 1: Summary of existing related works

Author(s) & Year	Dataset	Method	Reported Accuracy	Limitation
Rana et al. [9] (2022)	Multiple deepfake benchmarks	Systematic literature review	N/A	Survey only; highlights lack of cross-dataset generalization in reviewed methods
Naitali et al. [10] (2023)	Multiple	Survey of generation & detection methods	N/A	Identifies open challenges; no single unified benchmark
Mukta et al. [11] (2023)	Multiple deepfake tools	Empirical investigation of DF tools/models	Varies	Effectiveness highly tool- and dataset-dependent
Alkishri et al. [16] (2023)	140K Real & Fake Faces	DenseNet-121 + VGG16 with DFT frequency analysis	~99%	Accuracy drops sharply on low-resolution images
Zhang et al. (2022)	CelebA + 5 GAN variants	Ensemble of gradient, frequency & texture features	97.04%	Feature engineering adds complexity; sensitive to GAN type
Patel et al. (2023)	Real/Fake image dataset	Dense CNN (D-CNN)	97.2% (94.67% on StyleGAN2 high-res)	Performance degrades on high-resolution StyleGAN2 images
Ghita et al. (2024)	40,000 Kaggle images	Vision Transformer (ViT)	89.91%	High computational cost; not tested across datasets
Usmani et al. [15] (2024)	RFF / RFFD datasets	Shallow Vision Transformer	92.15% / 88.52%	Lower accuracy than heavier CNN baselines despite efficiency gains
Chetan Kumar et al. (2024)	140K Real & Fake Faces (Kaggle)	Parallel Vision Transformer (PViT)	91.92% (AUC 0.979)	Requires multiple parallel transformer branches, increasing cost
Hybrid DCGAN + ResNet-50 studies (2024)	Custom face dataset	DCGAN + ResNet-50 hybrid	>83% (~10% above ResNet-50 alone)	Evaluated on a comparatively small custom dataset
This work (2026)	CelebA + custom GAN (internal); independent external set	GAN-augmented dataset + fine-tuned ResNet-50	96.97% internal / 94.0% external (AUC 0.977)	Evaluated on facial images; cross-generator (e.g., diffusion models) robustness left for future work

2. Material And Methods

2.1 Overall Framework

The end-to-end pipeline starts by collecting real facial images from CelebA and creating the synthetic images with a custom GAN. Both the images are first pre-processed and then mapped to a balanced binary-classification dataset. This data is then fed into a pre-trained ResNet-50 network that is then fine-tuned and tested on held out internal and external test sets. The model finally gets used in a real-time Chrome browser extension (fig. 1).

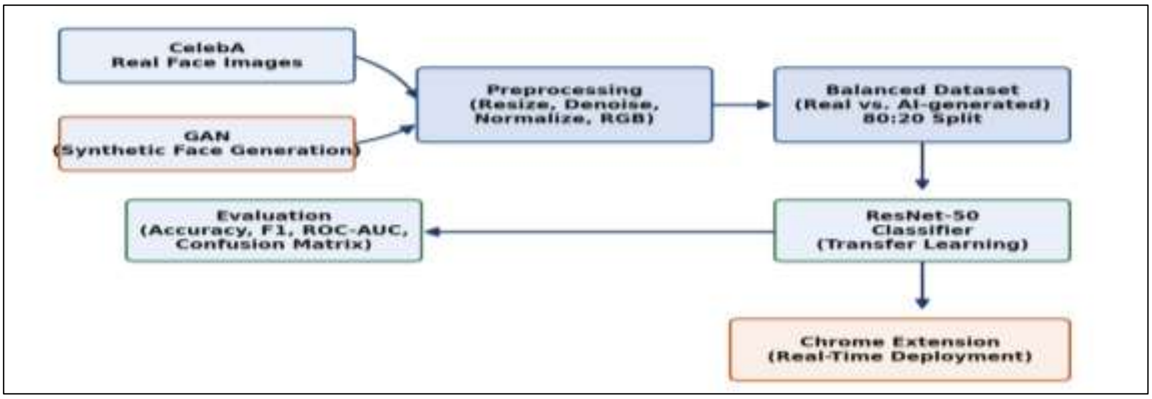

Fig 1. Over all Architecture

2.2 Dataset Used

In this study two image classes were employed. 1. Real images- from the CelebA (Celeb Faces Attributes) dataset, which has 200,000 images of celebrity faces with diverse poses, expressions, lighting, age, ethnicity and backgrounds. 2. AI-Generated Images- These are images created using a custom Generative Adversarial Network trained on the CelebA dataset to learn the distribution of the facial structure of the images, and then sampling images whose facial-structure distribution is similar to that of the images in the CelebA dataset.
Initially, 11,081 authentic images and 10,960 synthetic images were available. The real class was down-sampled to 10,960 images to make them equally sized, thus eliminating class imbalance. The balanced data set was then divided into an 80/20 training/testing data set, as shown in Table 2.

Table 2: Bifurcation of dataset

Dataset Split	Real Images	Fake Images	Total
Training	8,768	8,768	17,536
Testing	2,192	2,192	4,384

2.3Image Preprocessing

All pictures were pre-processed in a standardized way to achieve a better image quality and consistency before they were fed to the classifier. Firstly, the images were resized to 224x224 pixels to conform with the input size of ResNet-50. Then noise was removed using Gaussian filter, and re-Scale the Pixel Values to [0, 1] Range. Secondly, histogram equalization was done to enhance contrast and removing blurring or damage of samples. Finally, all pictures have been converted to RGB. This pipeline allows for the model to have consistent and clean inputs, aiding in the stability of the training process and the final classification accuracy.

2.4 GAN-Based Synthetic Image Generation

The AI-generated class was created using a Generative Adversarial Network to generate fake face images. A GAN is a pair of neural networks that are both trained at the same time in an adversarial manner:

- **Generator (G):** takes a random latent vector z as input and produces a synthetic face image G(z).
- **Discriminator (D):** Accepts real and generated images and discriminates between them giving a training signal to the generator.

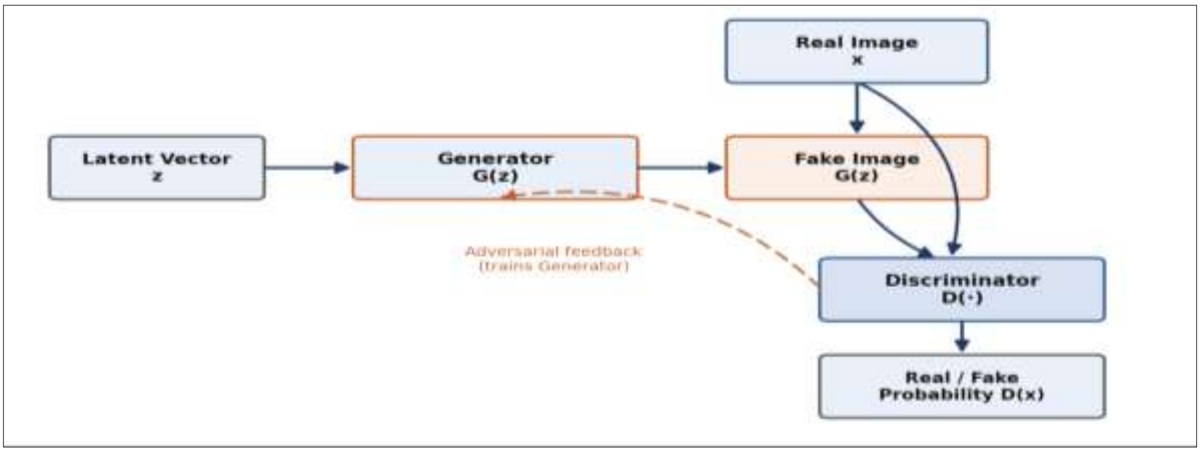

Fig. 2 GAN Architecture

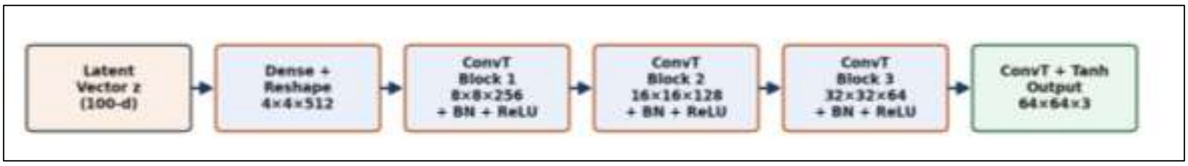

Fig. 3 Generator Architecture

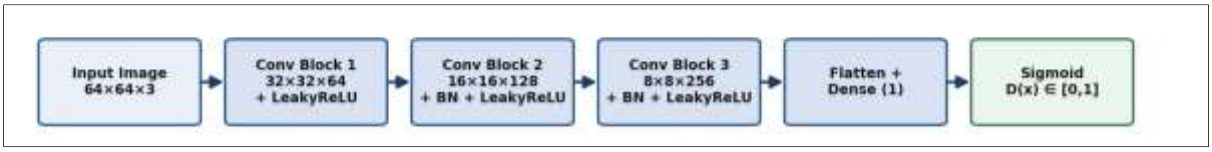

Fig. 4 Discriminator Architecture

The generator is then iteratively trained to generate more and more realistic faces, while the discriminator is trained to discriminate more and more between synthetic artifacts. The GAN was able to generate high-quality synthetic faces after a few epochs of training, which could be used to build the class of the classification set. The overall GAN architecture, generator and discriminator respectively are shown in figures 2, 3 and 4.

2.5 ResNet-50 Classifier and Transfer Learning

The backbone classifier ResNet-50 was chosen because it has deep residual connections, enabling the flow of gradients through very deep networks, good general purpose convolutional features extraction (pre-trained on ImageNet), and always performs well on binary and multi-class image classification tasks. For this task the following modifications were made:

- The last fully connected layer was modified to a single-logit output layer to produce a binary classification (Real vs. AI generated).
- The network was end-to-end fine-tuned with Adam optimizer and Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss).
- The training was carried out using 50 epochs, a learning rate of 1×10^{-4} and a batch size of 32.
- To ensure algorithm robustness and prevent overfitting, data augmentation methods such as random rotation, horizontal flipping, and brightness shifting were used.

By tuning to a degree, the model was able to capture subtle variations in local contrast, high-frequency patterns, texture and structural inconsistencies found in GAN-generated images. The ResNet-50 architecture for transfer learning implemented in this work is shown in Fig. 5.

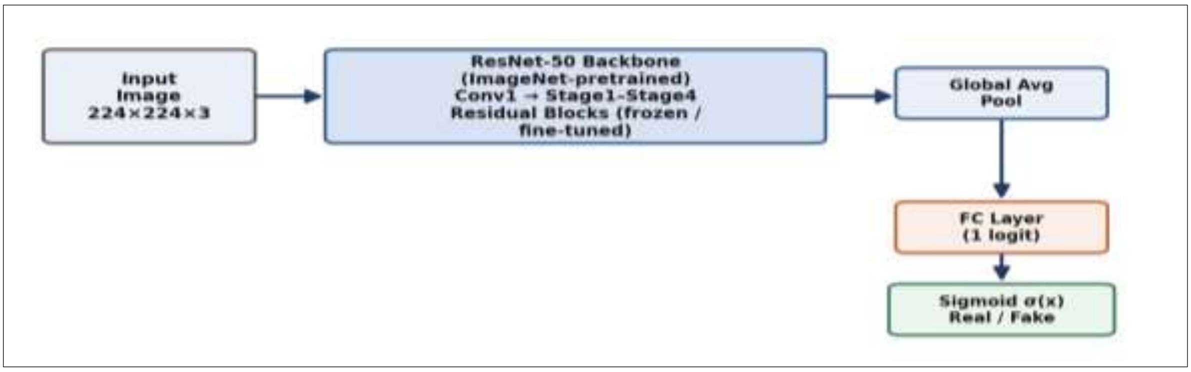

Fig. 5 ResNet-50 Transfer Learning Architecture

3. RESULTS

A test set created internally and an external test set were used to test the proposed framework to see its effectiveness in discriminating between the AI-generated facial images and real images. Accuracy, precision, recall, F1 score, confusion matrix and ROC/AUC analysis were used to measure the performance of the system.

3.1 Training Performance

The fine-tuned model was saved at regular checkpoints for 50 epochs with the BCEWithLogitsLoss and Adam optimizer. Initial epochs saw training loss drop sharply, from 0.0772 at epoch 1 to below 0.01 in epochs 10, and settling around 0.002 afterwards, suggesting good convergence and learning of discriminative image features. As seen in the training loss curve (Figure 7), there is no abrupt oscillation in the learning curve, indicating the stability of the optimization and generalization.

The classification performance on the Test Dataset is shown below. The proposed ResNet-50 classifier achieved an overall test accuracy of 96.97% on the internal test set of 4384 images which is balanced.

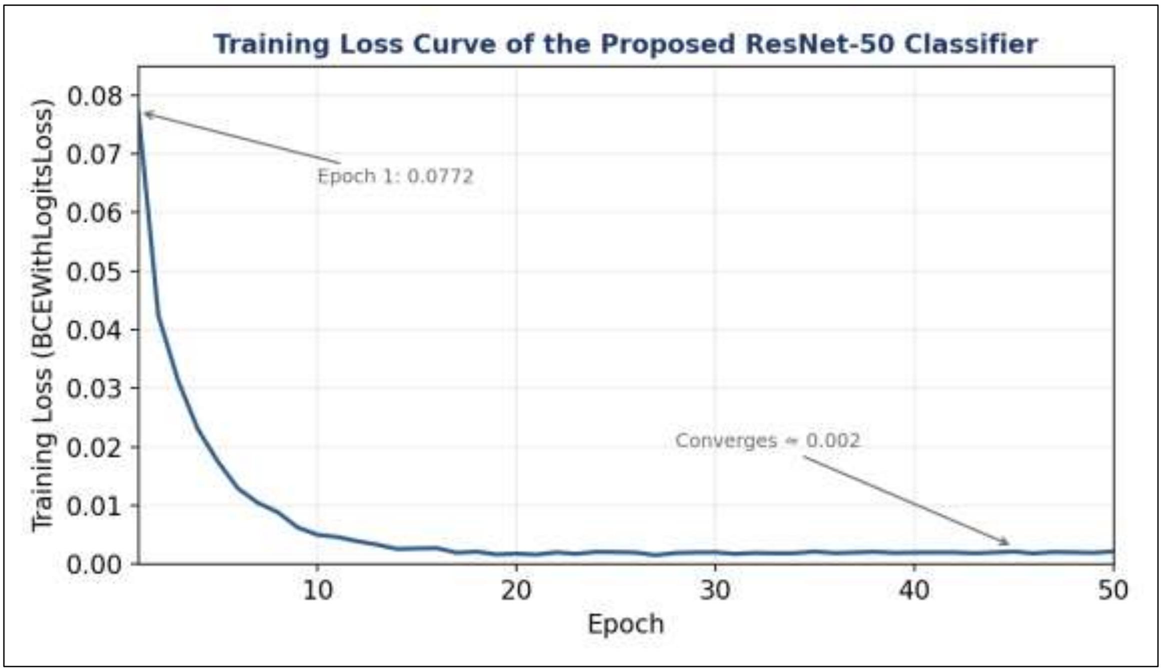

Fig. 6 Training Loss Curve of the Proposed ResNet-50 Classifier

The classification performance on the Test Dataset is presented as follows. For both classes, the classifier's precision, recall and F1-score values were 0.97, demonstrating balanced learning and ability to detect authentic images as well as AI-generated images (Table 3).

Table 3: Classification Performance on the Test Dataset

Class	Precision	Recall	F1-score
Fake	0.97	0.97	0.97
Real	0.97	0.97	0.97

Table 4: Classification Performance on the external Dataset

Class	Precise	Recall	F1
Fake	0.93	0.93	0.93
Real	0.94	0.94	0.94

3.2 Confusion Matrix Analysis

We get the confusion matrix on the internal test set as shown in Figure 8. The classifier was correct in identifying 2,132 fake images and 2,119 real images, and was wrong in identifying 60 fake images as real, and 73 real images as fake. The performance of this relatively small number of misclassifications is robust and confirms that the proposed classifier is effective at discriminating between authentic and synthetic facial images and the learned feature representations are effective at discriminating between the two.

3.3 ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve (Figure 9) was obtained by plotting the true-positive rate (TPR) and false-positive rate (FPR) of the classifier on a graph against the different decision thresholds:

$$TPR = TP / (TP + FN) \quad (1)$$

$$FPR = FP / (FP + TN) \quad (2)$$

The proposed framework has an Area Under the Curve (AUC) of 0.977, which results in excellent discriminative performance. The closer the AUC value is to 1, the more effective the classifier will be in separating facial images produced by AI from real images across a variety of decision thresholds.

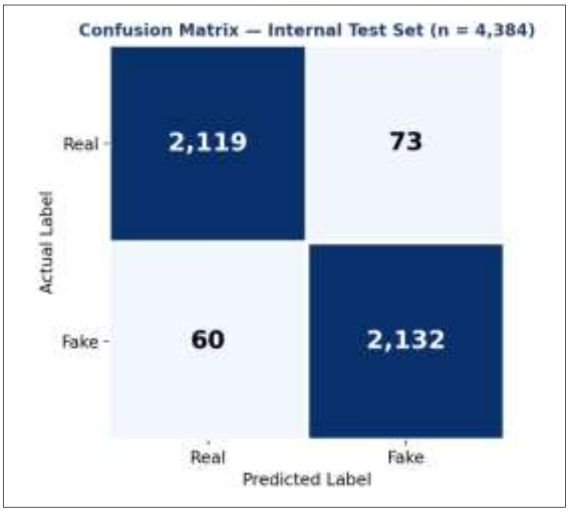

Fig. 7: Confusion matrix for the Proposed ResNet-50 Classifier (Internal Test Set)

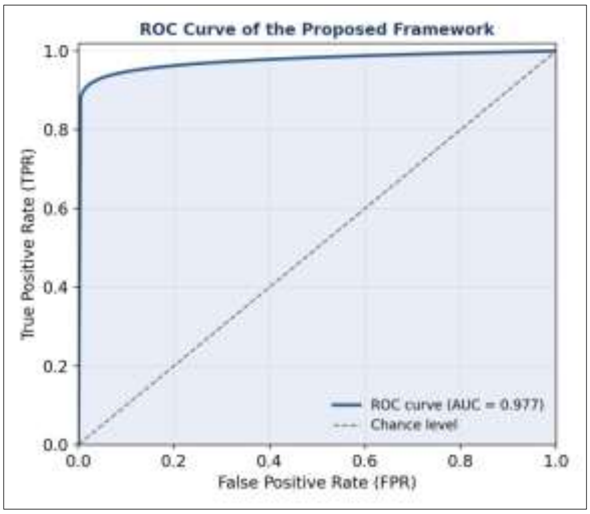

Fig. 8: ROC Curve of the Proposed Framework (AUC = 0.977)

3.4 External Dataset Validation

To measure the robustness against out-of-training distribution, the trained estimator was tested on an independent external test set that is not used during training or internal test, including 2041 images (960 fake & 1081 real images). The classification results are given in Table 4.

The confusion matrix on the external dataset indicated that the confusion matrix correctly classified 894 fake images and 1,017 real images, and failed to classify 66 fake images and 64 real images. The small degradation in accuracy when compared to the internal test set is anticipated as the external set was collected with different cameras (capture settings, generators, and demographics). However, the framework achieved high classification accuracy of 94.0%, which shows its good generalization ability on the other distribution.

3.5 Qualitative Analysis

Qualitative evaluation was carried out by visual examination of samples which were correctly classified and those classified incorrectly. Most images that were correctly classified have distinct facial features and texture attributes that are well captured by the ResNet-50 feature extractor. The misclassified samples are mostly highly realistic faces created by GAN or authentic photos with complex background content and/or extreme pose variation or illumination conditions. The observations indicate that the framework is effective for various facial images, but it would be beneficial to integrate attention mechanisms or Vision Transformer architectures to further exploit subtle synthetic artifacts, especially in challenging instances.

4. Discussion

The experimental results validate the effectiveness of the proposed framework, which combines the image synthesis capability of the GAN with transfer learning for AI image detection. The well-designed balanced dataset construction method and optimized ResNet-50 resulted in 96.97% accuracy in classification and 0.977 AUC on the in-house test dataset and 94.0% accuracy on the independent external test dataset, demonstrating the model's strength and generalization ability. The small number of mispredictions in both the confusion matrices shows that the proposed method can learn discriminative representations of authentic and synthetic facial images and the curve of the training loss show smooth convergence during the learning process,

reflecting stable optimization during learning. In summary, the proposed framework is a reliable and efficient scheme to detect the facial image generated by AI, and it could also be the basis of a practical, browser-based authenticity-verification system as shown in Section 5.
The research was compared with existing methods. Present research was compared with existing methods. Table 5 compares the performance of the proposed method with representative methods reported in the recent literature. These methods were tested on various datasets, with varying image resolutions and generator families, and the comparison is intended to be contextually interpreted, but demonstrates the competitiveness of the proposed ResNet-50 framework with and in some cases better than existing CNN and transformer-based methods, and also applying it to a truly independent external dataset that many previous studies have been absent from.

Table 5: Comparison with Existing Methods

Method	Dataset	Reported Accuracy
Dense CNN (D-CNN) — Patel et al. (2023)	Real/Fake image dataset	97.2%
DenseNet-121 + VGG16 (DFT-based) — Alkishri et al. (2023)	140K Real & Fake Faces	~99%
Ensemble (gradient + frequency + texture) — Zhang et al. (2022)	CelebA + 5 GAN variants	97.04%
Vision Transformer (ViT) — Ghita et al. (2024)	40,000 Kaggle images	89.91%
Shallow Vision Transformer — Usmani et al. (2024)	RFF / RFFD datasets	92.15% / 88.52%
Parallel Vision Transformer (PViT) — Chetan Kumar et al. (2024)	140K Real & Fake Faces (Kaggle)	91.92% (AUC 0.979)
Hybrid DCGAN + ResNet-50	Custom face dataset	>83%
Proposed: GAN-augmented dataset + ResNet-50 (internal)	CelebA + custom GAN	96.97% (AUC 0.977)
Proposed: GAN-augmented dataset + ResNet-50 (external)	Independent external set	94.0%

The proposed method has demonstrated high internal accuracy, a high AUC and proven good external-dataset generalization, which are comparable with the reported single-dataset evaluations from other studies found in the literature and indicates the proposed method is suitable for use as a detector that is practically ready for deployment.

5. CHROME EXTENSION: REAL-TIME DEPLOYMENT

The trained ResNet-50 classifer was embedded in a chrome browser extension to allow for real-time and easy-to-use AI-image detection. It enables users to see at a glance if an image on any webpage is real or AI-generated, without the need for downloading an app or coding expertise.

5.1System Overview

The Chrome extension provides a user interface and enables the extraction of images from the Web, which is handled by a client-side module, and then passes them to an image classification service deployed on the back-end, based on the trained ResNet-50 model. There are four parts to the architecture:

- Content Script — Taps the images shown on the current page, puts them in the background script for processing, and returns the classification result to the current page.
- The Background Script / Service Worker is responsible for communicating between the frontend and backend, as well as for forwarding the images that are extracted to the classifier API and for routing the results classification to the UI.
- Model Server (Flask / FastAPI): serves the trained ResNet-50 model, accepts image inputs in base64/ blob format from the extension, does base64 / blob pre-processing and returns a Real / AI-generated prediction.
- Chrome Popup UI — lets users pop up and manually review a selected image, shows what the classification confidence was, and gives a summary of the latest predictions.

5.2 Working Principle

Below is the process of using chrome extension for the real time usage of this work:
1.The user clicks on any web page with pictures.
2.The content script fetches the image that was selected (or all images on the page).
3.The image is then encoded as base64 data and submitted to the API on the back end.
4.The backend does preprocessing (noise removal and resizing) and passes the image to the ResNet-50 Classification.
5.The API provides classification results: Real Image or AI-Generated Image.

6.The outcome is shown in a pop-up window, as a badge or as a floating tooltip above the image.
This interaction in real time enables users to check the authenticity of images they face during the browsing process.

5.3 Chrome Extension Features

- One-click prediction on any page in the web.
- Live look-up of images throughout the web pages.
- Classifications with high accuracy, supported by the fine-tuned ResNet-50 model.
- Lightweight with clean and minimal user interface.
- Operates on any web page that has pictures.
- Cross browser support (Chrome, Edge, Brave, and other browsers based on Chromium).
- Scalability: The model for the backend can be modified at any time without needing to modify the extension itself.
- Automation: images on webpages can be automatically scanned for authenticity.
- Privacy means that user data is not saved by the extension: The images are only processed for a short time for classification.

5.4 System Integration

The data flow of the deployed system is shown in Figure 10. First, the content script takes out the images from the webpage, passes them to the ResNet-50 model in the Backend API, which produces the prediction, and the prediction is then returned through the background script to the Chrome popup UI to be displayed to the user.

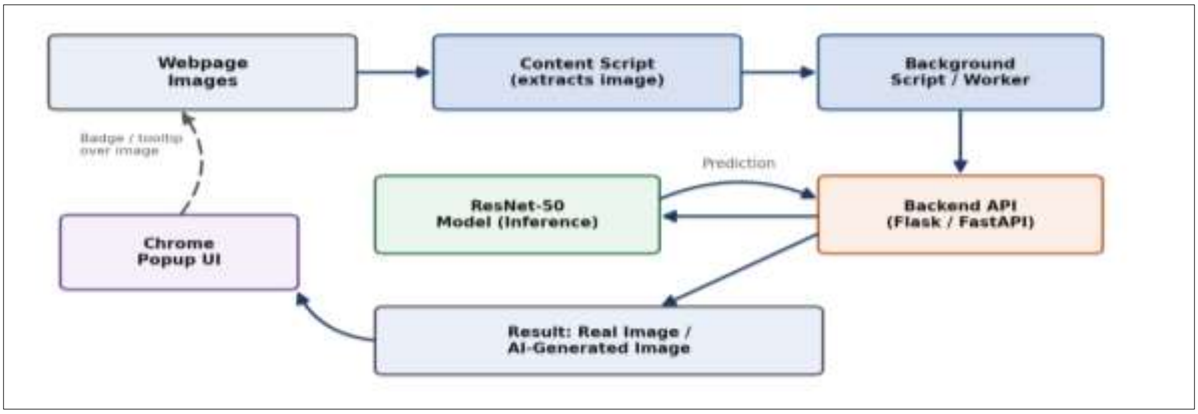

Fig. 9: Diagram of the Chrome Extension System Integration.

6. CONCLUSION AND FUTURE WORK

The proposed research provided a comprehensive solution for identifying AI-generated fake faces using both synthetic data generation techniques (GANs) and a fine-tuned ResNet-50 facial recognition model. The proposed classifier was evaluated on an internal test set, which was created using a balanced dataset from the CelebA dataset and a customized GAN, and was found to be highly accurate with 96.97% accuracy and an AUC of 0.977. It was also tested on an independent external dataset with 94.0% accuracy, indicating good generalization beyond the training distribution. Comparing with the recently proposed CNN, DenseNet, and Vision-Transformer-based models from existing literature, the proposed approach is competitive, and also has proven low correlation cross-dataset. The trained model was then ported to a Chrome browser extension, allowing for real-time in-browser authentication, and showcasing the practical use of this framework in digital forensics, media validation, and everyday online content authentication.
Future work involves several directions: First, using the Vision Transformer or hybrid CNN-transformer architecture to better capture synthetic artifacts that are hard to characterize but appear in highly realistic samples, alongside the use of GAN generated samples. Second, extending the training and evaluation datasets to include diffusion-model-generated images (e.g., Stable Diffusion, DALL·E, Midjourney) which are growing in prevalence alongside GAN generated images. Third, incorporating frequency-domain analysis as a complementary detection signal, and fourth, optimizing the inference pipeline within the Chrome extension (e.g., via model quantization or on-device inference), to reduce latency and server dependency for end users.

References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Commun. ACM 63(11), 139–144 (2020).
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016).
3. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434 (2015).

4. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proc. Int. Conf. Computer Vision (ICCV), pp. 3730–3738 (2015).
5. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR) (2019).
6. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics++: Learning to detect manipulated facial images. In: Proc. Int. Conf. Computer Vision (ICCV) (2019).
7. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proc. Int. Conf. Machine Learning (ICML) (2019).
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al.: An image is worth 16×16 words: Transformers for image recognition at scale. In: Int. Conf. Learning Representations (ICLR) (2021).
9. Rana, M.S., Nobi, M.N., Murali, B., Sung, A.H.: Deepfake detection: A systematic literature review. IEEE Access 10, 25494–25513 (2022). https://doi.org/10.1109/ACCESS.2022.3154404
10. Naitali, A., Ridouani, M., Salahdine, F., Kaabouch, N.: Deepfake attacks: Generation, detection, datasets, challenges, and research directions. Computers 12(10), 1–26 (2023).
11. Mukta, M.S.H., Ahmad, J., Raiaan, M.A.K., Islam, S., Azam, S., Ali, M.E., et al.: An investigation of the effectiveness of deepfake models and tools. J. Sensor Actuator Netw. 12(4), 1–43 (2023). https://doi.org/10.3390/jsan12040061
12. Wang, X., Guo, H., Hu, S., Chang, M.-C., Lyu, S.: GAN-generated faces detection: A survey and new perspectives. arXiv:2202.07145 (2022).
13. Hybrid DCGAN-ResNet50 model for fake face detection. In: Springer Nature Link Conference Proceedings, https://doi.org/10.1007/978-3-032-12834-8_13.
14. Hybrid deep learning model based on GAN and RESNET for detecting fake faces. IEEE Access (2024). https://doi.org/10.1109/ACCESS.2024.3416910
15. Usmani, S., Kumar, S., Sadhya, D.: Efficient deepfake detection using shallow vision transformer. Multimed. Tools Appl. 83, 12339–12362 (2024). https://doi.org/10.1007/s11042-023-15910-z
16. AI-powered deepfake detection using CNN and Vision Transformer architectures (literature review incl. Patel et al. 2023; Ghita et al. 2024; Alkishri et al. 2023; Zhang et al. 2022). arXiv:2601.01281.
17. Deepfake detection using parallel Vision Transformers. In: Proc. Creativity & AI Workshop (2024).
18. Wavelet-based GAN fingerprint detection using ResNet50. arXiv:2510.21822 (2025).