AI vs Reality: Classifying Synthetic Faces with a Fine-Tuned ResNet50 Neural Network

Nizamuddin Naeem Mandekar[1]

[1]Department of Computer Science, B. K. Birla College of Arts, Science & Commerce (Autonomous) Kalyan, 421301, Maharashtra, India.

Publication Date: 2025/07/16

Abstract: with the rise of generative technologies, distinguishing between real and AI-generated images has become increasingly challenging. Advanced generative frameworks such as Generative Adversarial Networks (GANs) and Latent Diffusion Models (LDMs) now generate highly convincing synthetic images that closely resemble genuine photographs. This phenomenon poses significant challenges for domains including cybersecurity, journalism, and social media platforms, where image authenticity verification is paramount. This study explores the application of ResNet50 deep learning architecture for distinguishing between AI-synthesized and authentic facial images. Our model underwent training using a comprehensive dataset containing 140,000 facial photographs, equally distributed between genuine and artificially generated samples. The ResNet50 architecture was enhanced through transfer learning methodologies to improve its capability in identifying subtle characteristics that differentiate authentic images from synthetic ones. Two distinct experimental approaches were employed: feature extraction methodology and comprehensive fine-tuning procedures. The optimized model demonstrated remarkable performance, achieving accuracy rates of up to 98%, validating its effectiveness in this domain. This investigation demonstrates the effectiveness of fine-tuned ResNet50 architecture in identifying AI-synthesized images. The research contributes to developing robust verification systems for image authentication, combating the proliferation of synthetic content, and maintaining the integrity of digital media platforms.

Keywords: *Artificial Intelligence, AI-Generated Images, Generative Adversarial Networks, Latent Diffusion Models, Image Verification, Deep Learning, ResNet50, Transfer Learning.*

How to Cite: Nizamuddin Naeem Mandekar; (2025) AI vs Reality: Classifying Synthetic Faces with a Fine-Tuned ResNet50 Neural Network. *International Journal of Innovative Science and Research Technology*, 10(7), 809-816. https://doi.org/10.38124/ijisrt/25jul706

I. INTRODUCTION

The contemporary landscape of Artificial Intelligence (AI) has witnessed unprecedented developments, transforming numerous sectors including medical sciences, entertainment industry, and cybersecurity. Among the most remarkable innovations are generative models - sophisticated AI frameworks capable of producing novel content, including photorealistic images that are virtually indistinguishable from authentic photographs. These technologies, particularly Generative Adversarial Networks (GANs) and Latent Diffusion Models (LDMs), have demonstrated exceptional capabilities in creative fields such as digital art, graphic design, and multimedia entertainment. Nevertheless, these advancements simultaneously introduce significant concerns regarding image authenticity verification.

The increasing sophistication of AI-generated imagery has created substantial challenges in distinguishing synthetic content from genuine photographs. This presents critical issues in professional domains such as investigative journalism, legal proceedings, and security operations, where image credibility is fundamental. Artificially generated images can facilitate the creation of deceptive content, including deepfake media and fabricated evidence, potentially causing substantial societal harm. Therefore, developing reliable methodologies for differentiating between authentic and AI-synthesized images has become imperative.

This investigation addresses these challenges by implementing deep learning methodologies to categorize images as either authentic or artificially generated. Our approach specifically utilizes a pre- trained deep learning architecture known as ResNet50. This robust model has undergone extensive training on comprehensive datasets and demonstrated exceptional performance in image classification applications. Our methodology involves fine-tuning this architecture to identify nuanced differences between genuine photographs and AI-generated counterparts.

The primary objective of this research is to develop a highly accurate system capable of reliably identifying AI-synthesized images. Through fine-tuning ResNet50, we aim to leverage its pre-existing knowledge base while adapting it to this specialized application. The ability to distinguish between authentic and synthetic imagery is becoming increasingly critical as AI-generated content proliferates across digital platforms. This research contributes to developing tools that address the challenges posed by deceptive and manipulated visual content, ensuring the trustworthiness of digital media.

The potential misuse of generative AI technologies represents an escalating concern due to their widespread accessibility. These synthetic images can facilitate identity concealment in online environments, enabling fraudulent activities. Additionally, they can compromise facial recognition systems, while AI-generated videos or audio content can be weaponized for extortion purposes. Deepfake technology can even fabricate evidence to implicate innocent individuals [2].

This paper proposes a comprehensive solution for classifying AI-generated and authentic images utilizing a fine-tuned ResNet50[3] architecture.

II. DATASETS

This study is grounded on the **140K Real and Fake Faces Dataset**, a benchmark dataset extensively used in synthetic image detection research. It contains a total of **140,000 facial images**, evenly split into **70,000 authentic human faces** and **70,000 AI-generated faces**. This balanced distribution is crucial for training binary classification models without class bias.

The **authentic face images** are derived from publicly available Flickr image collections curated by NVIDIA, featuring diverse lighting conditions, facial orientations, age groups, and ethnicities. Meanwhile, the **synthetic images** are generated using the advanced **StyleGAN architecture**, from which a representative subset of one million fake faces was filtered and selected to match the distribution of the real samples.

All images were **resized to 256×256 pixels** to standardize input dimensions and ensure compatibility with the ResNet50 architecture. No additional data augmentation was applied, as the dataset already includes sufficient visual diversity to avoid overfitting.

➢ *The Dataset was **Stratified and Split** into Three Parts:*

- **Training set**: 100,000 images (50,000 real, 50,000 fake)
- **Validation set**: 20,000 images (10,000 real, 10,000 fake)
- **Test set**: 20,000 images (10,000 real, 10,000 fake)

This configuration ensured that each subset retained an equal distribution of real and AI-generated samples. Such organization facilitates robust performance evaluation while minimizing the risk of data leakage during model training.

This dataset is publicly accessible on **Kaggle**, and it has been utilized in previous academic works focusing on deepfake detection and facial forgery classification, validating its reliability and applicability to the research community. The following examples demonstrate the dataset composition:

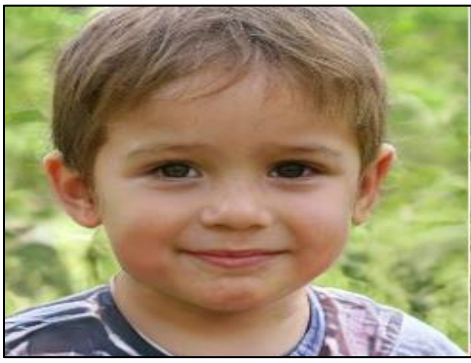
Fig 1 Fake 1

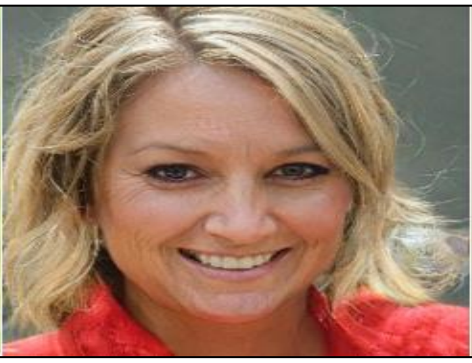
Fig 2 Fake 2

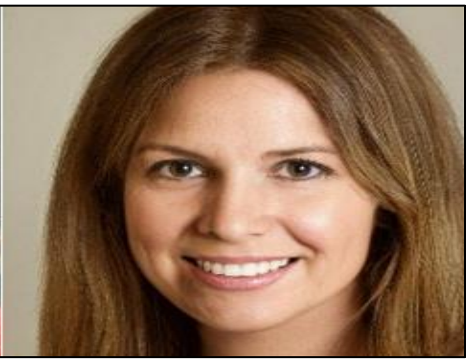
Fig 3 Fake 3

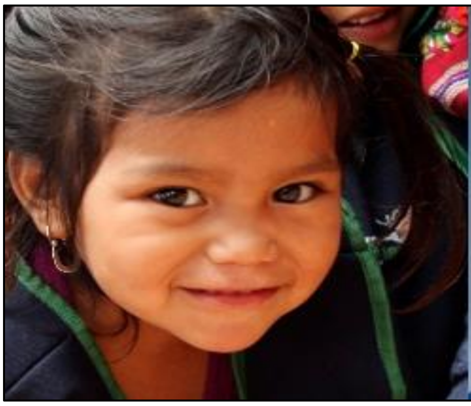
Fig 4 Real 4

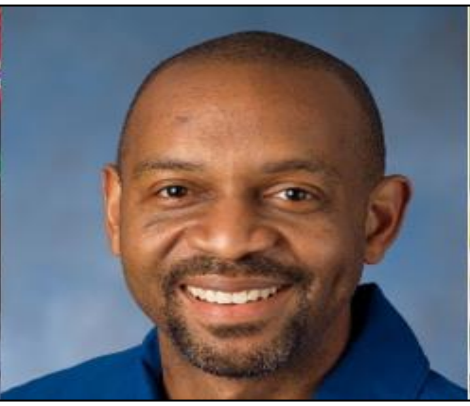
Fig 5 Real 5

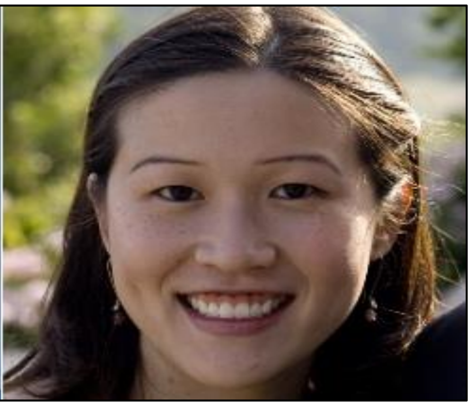
Fig 6 Real 6

III. METHODS

➢ *Transfer Learning Implementation:*

ResNet50 Architecture Transfer learning represents a sophisticated deep learning approach that leverages pre-existing, trained models to address novel problems sharing similar characteristics This methodology can be implemented through two primary strategies:

➢ *Fine-Tuning Approach:*

This technique maintains the model's architectural framework while allowing all layers to adapt to the new problem domain. The process requires significant computational resources and time investment as it updates all parameters through backpropagation mechanisms. In some implementations, selective layers remain trainable while others are frozen to optimize performance.

➢ *Feature Extraction Methodology:*

This approach freezes specific model layers while training only selected layers for the new application. This technique offers computational efficiency and reduced training time since only a subset of layers undergoes training. Our research employs the ResNet50 architecture, which incorporates "Residual Connections" – an innovative design element that enables the construction of deeper networks while avoiding vanishing gradient problems (where gradients become too small to effectively update parameters). These residual connections facilitate information flow across layers, maintaining gradient strength throughout the network. In certain implementations, these connections incorporate convolutional layers to adjust data dimensionality appropriately.

➢ *The ResNet50 Architecture Offers Several Advantages for our Application:*

- Pre-trained weights from extensive image datasets
- Proven performance in image classification tasks
- Robust feature extraction capabilities
- Efficient handling of complex visual patterns.

➢ *Experiment 1: Feature Extraction*

The initial experimental phase involved minimal modifications to the ResNet50 architecture. The fully connected (fc) layer underwent adjustment by modifying the output classification layer from 1000 classes to accommodate our binary classification requirement (2 classes). The ResNet50 structural organization:

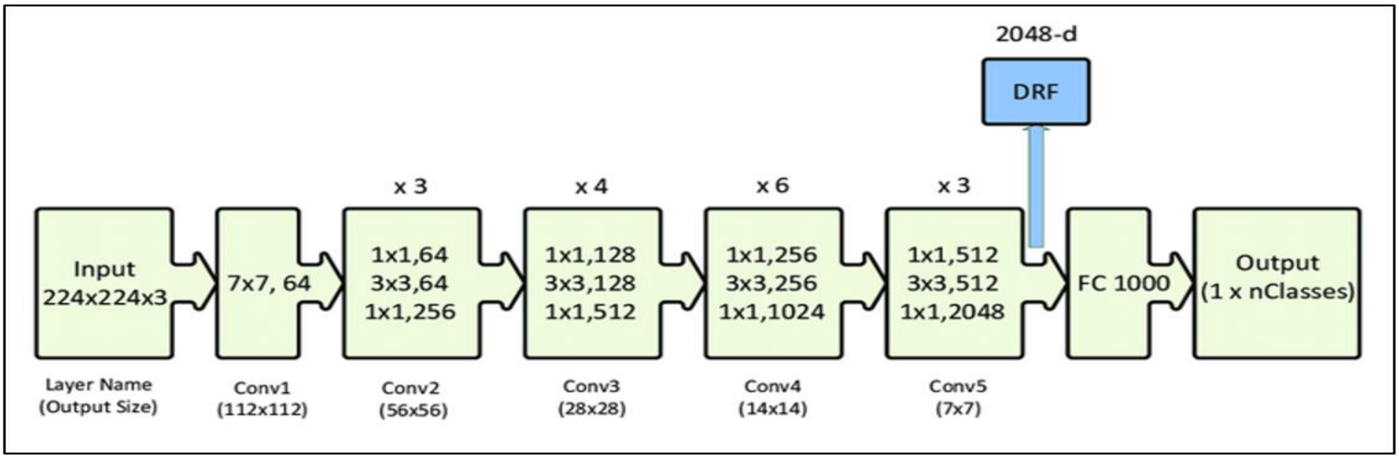

Fig 7 Feature Extraction

The code for Feature Extraction is: # Code to reconfigure the fully connected layer for binary classification *model.fc = torch.nn.Sequential(torch.nn.Linear(2048, 2)*

➢ *Experiment 2: Comprehensive Fine-Tuning:*

In the second experiment, additional changes were made to improve accuracy. A combination of Linear layers, ReLU() functions, and Dropout layers was used. The fully connected (fc) layer of ResNet50 was modified as follows:

Model.fc = torch.nn.Sequential(torch.nn.Linear(2048, 1000), torch.nn.ReLU(), torch.nn.Linear(1000, 500), torch.nn.Dropout(), torch.nn.Linear(500, 100), torch.nn.ReLU(),torch.nn.Dropout(), torch.nn.Linear(100, 2) # Number of classes)

➢ *This enhanced architecture incorporates:*

- Progressive dimension reduction through multiple linear transformations
- ReLU activation functions for non-linear processing
- Dropout layers for regularization and overfitting prevention
- Optimized layer sizes for balanced performance

➢ *Training Configuration and Parameters:*

Given the substantial dataset size of 50,000 images per category, additional data augmentation techniques were deemed unnecessary. All images underwent preprocessing to standardize dimensions at 256 pixels. The feature extraction model underwent training for 15 epochs, utilizing a learning rate of 0.01 with Adam optimization algorithm. Training employed a batch size of 32 samples. The fine-tuned model underwent training across multiple epoch configurations (5, 10, and 15 epochs) using identical parameters to evaluate optimal training duration.

IV. RESULTS AND ANALYSIS

Model performance evaluation was conducted through comprehensive loss and accuracy visualization. The following sections present detailed analysis of experimental outcomes.

- *Experiment 1: Feature Extraction Model Performance*

In Figure 8, demonstrates the training loss, validation loss, training accuracy, and validation accuracy progression throughout the training process. Both training and validation accuracy exhibited consistent improvement across epochs, achieving maximum performance of 89%. Similarly, training and validation loss demonstrated decreasing trends. However, beginning at the 12th epoch, the model exhibited signs of overfitting behaviour.

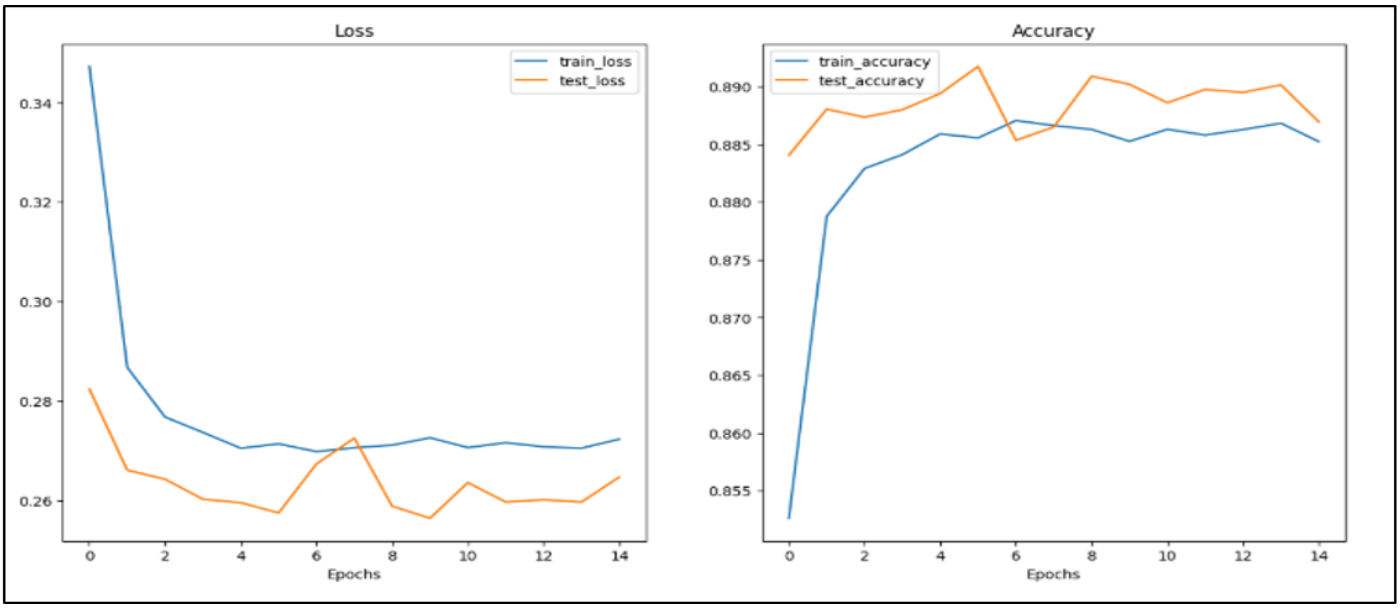

Fig 8 Training and Validation Accuracy/Loss Curve for Feature Extraction Model

- *Experiment 2: Fine-Tuning Performance Analysis*

Figure 9, illustrates the results for the fine-tuning experiment conducted over 5 epochs. Training loss exhibited consistent decrease throughout the process, while validation loss initially increased until the second epoch before declining alongside training loss. Training accuracy demonstrated steady improvement, while validation accuracy initially decreased before recovering after the first epoch, ultimately achieving 98% performance. However, experimental replication under identical conditions yielded validation accuracy not exceeding 94%, indicating some variability in results.

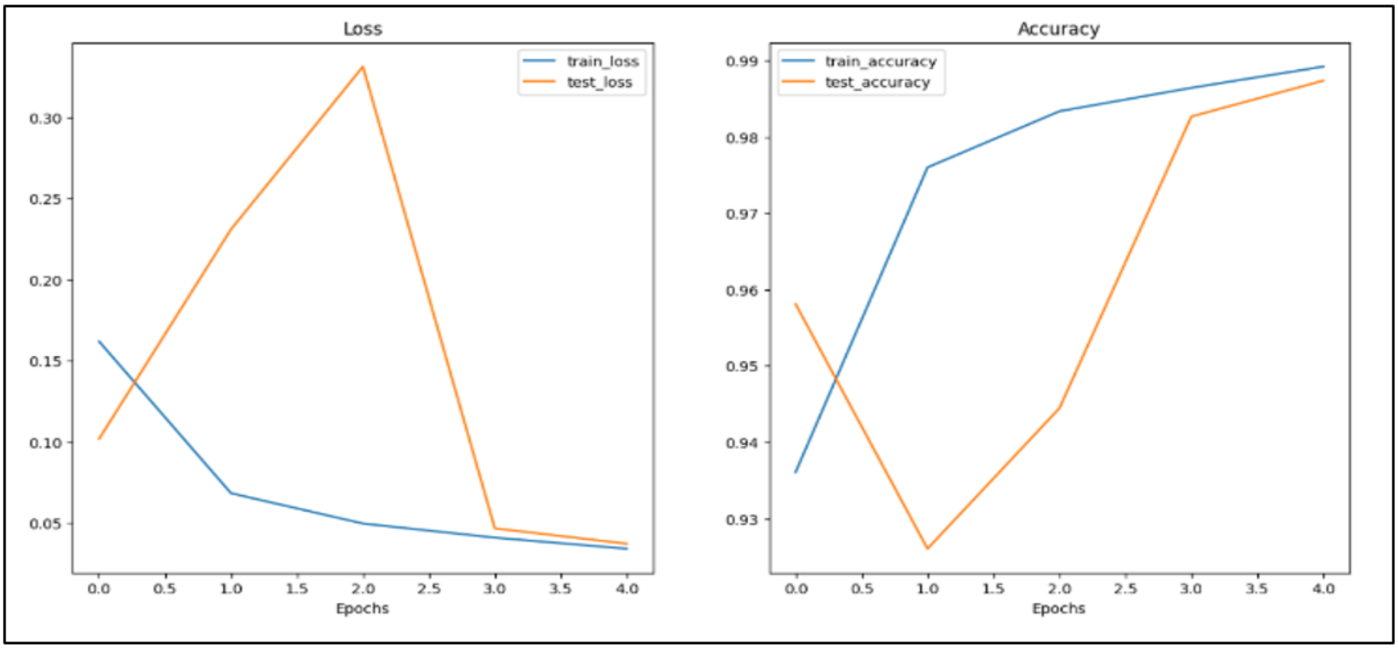

Fig 9 Performance Metrics for Fine-Tuned ResNet50 Model (5 Epochs)

Figure 10, presents results for the fine-tuning experiment extended to 10 epochs. Both training and validation loss demonstrated decreasing trends, with training loss exhibiting smooth decline while validation loss showed more irregular patterns. Accuracy metrics indicated improvement in both training and validation performance, with validation accuracy reaching 95% and training accuracy approaching 97%.

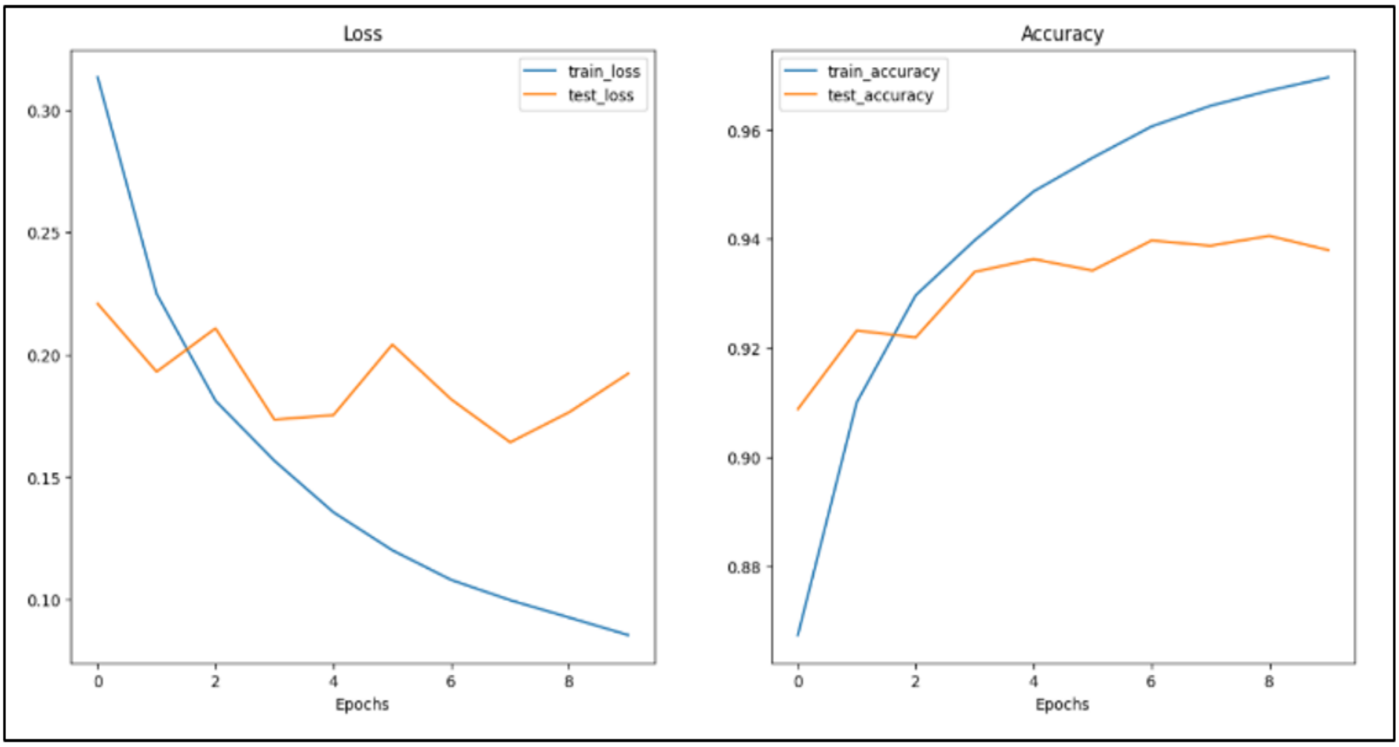

Fig 10 Training and Validation Trends for Fine-Tuned Model (10 Epochs)

Figure 11, demonstrates results for the extended 15-epoch fine-tuning experiment. Both training and validation loss continued decreasing, though a noticeable gap emerged between them. Training and validation accuracy initially improved, reaching peak performance of 96%, before exhibiting fluctuating patterns characteristic of potential overfitting.

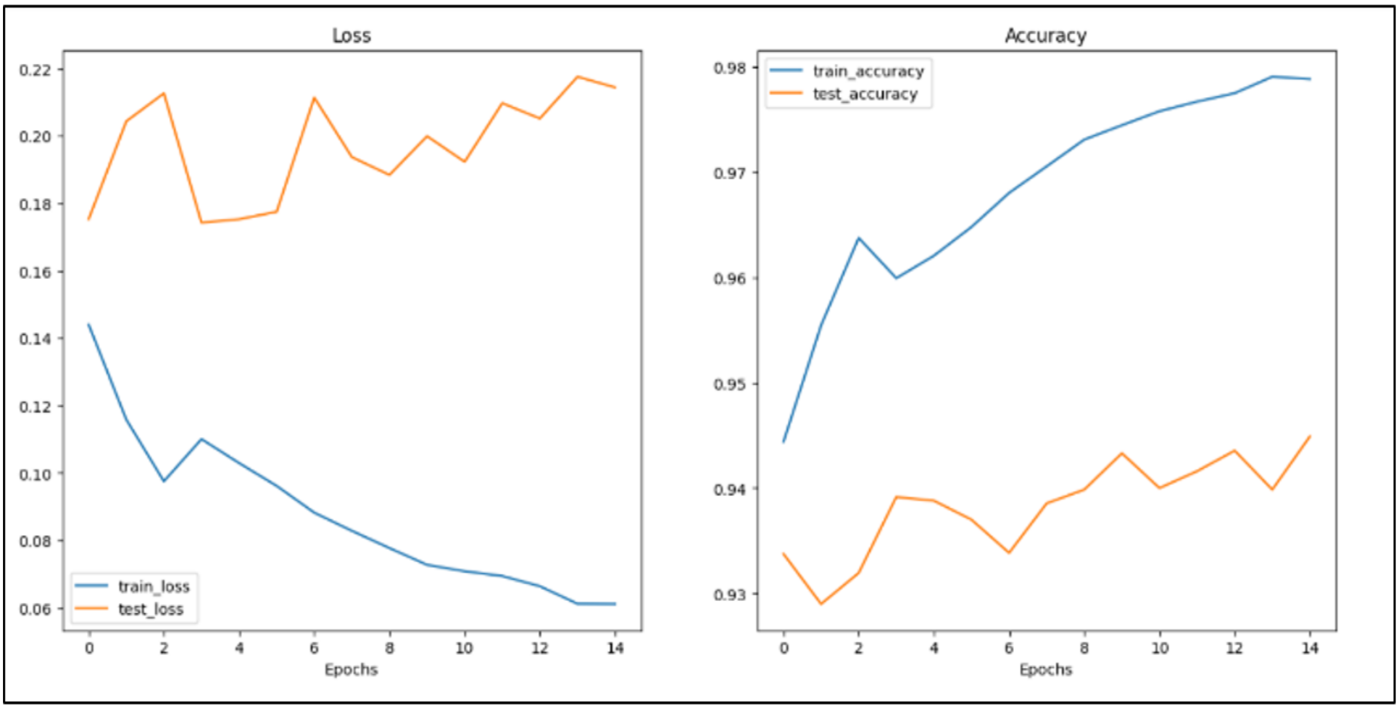

Fig 11 Accuracy and Loss Analysis for Fine-Tuned Model (15 Epochs)

➢ *Performance Evaluation Metrics:*

Model performance was assessed using standard classification evaluation criteria. These include:

- True Positives (TP): Instances where AI-generated images were correctly identified as synthetic.
- False Positives (FP): Instances where authentic images were mistakenly identified as synthetic.
- True Negatives (TN): Correct classifications of real images as genuine.
- False Negatives (FN): Synthetic images misclassified as real.

These values were used to compute accuracy, precision, recall, and F1-score, defined through the following formulas:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

$$\text{F1 Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

Table 1 The accuracy, precision, recall, and F1 score for all models are summarized in the table below:

Model	Accuracy	Precision	Recall	F1 Score
Feature Extraction Model	0.8997	0.9172	0.8505	0.8505
Fine Tuned Model (5 epoch)	0.9776	0.9803	0.9947	0.9875
Fine Tuned Model (10 epochs)	0.9542	0.9348	0.9425	0.9386
Fine Tuned Model (15 epochs)	0.9685	0.9475	0.9416	0.9445

V. DISCUSSION

Our experimental findings demonstrate that fine-tuning the ResNet50 architecture substantially enhances classification performance compared to feature extraction approaches. Initially, the feature extraction implementation achieved 89% accuracy, representing solid performance for a minimally modified pretrained model. However, implementing comprehensive fine-tuning with weight updates resulted in remarkable 98% accuracy after merely 5 training epochs. This improvement demonstrates that enabling the model to adapt to dataset-specific characteristics significantly enhances classification capabilities. The reproducibility of results showed variations across different experimental runs. While 5-epoch finetuning achieved high accuracy in initial experiments, subsequent repetitions yielded varying outcomes, suggesting some inherent instability. Furthermore, extending training to 10 or 15 epochs maintained stable performance but demonstrated slight accuracy reductions compared to 5 epoch fine-tuning. This performance degradation after extended training periods may indicate overfitting onset, where the model becomes excessively specialized to training data, reducing generalization capability to unseen samples. This could also indicate training process instability, were prolonged training causes performance fluctuations. Several enhancement strategies warrant future consideration:

➢ *Alternative Architecture Exploration:*

Investigating different deep learning architectures (including DenseNet, EfficientNet, or Vision Transformers) could provide insights into whether alternative model architectures yield improved stability or enhanced performance.

➢ *Advanced Fine-tuning Strategies:*

Extended fine-tuning with adjusted hyperparameters such as learning rate and batch size modifications could improve stability. Implementing gradual learning rate decay or advanced optimization techniques like learning rate warm-up could stabilize extended training procedures.

➢ *Data Augmentation Integration:*

Enhancing model generalization through data augmentation techniques including rotations, reflections, color modifications, and cropping operations could artificially expand training set diversity, reducing overfitting likelihood through varied example provision.

➢ *Cross-Validation and Regularization Implementation:*

Employing cross-validation during training could more effectively assess model generalization capabilities. Regularization techniques such as dropout or weight decay could help mitigate overfitting and enhance model robustness.

➢ *Learning Rate Scheduling:*

Implementing dynamic learning rate scheduling could optimize training stability and convergence behaviour.

➢ *Ensemble Methods:*

Combining multiple model predictions could improve overall performance and reduce prediction variance. Implementing these enhancements could achieve more stable and reliable performance, making the model more suitable for real-world applications where consistency is essential.

VI. FUTURE WORK

While our research demonstrates ResNet50's effectiveness in detecting AI-generated images, numerous opportunities exist for advancement and extension. The following areas represent promising directions for future investigation:

➢ *Advanced Model Architecture Investigation:*

Although ResNet50 provides solid performance, newer architectures may offer superior results. **EfficientNet** and **Vision Transformers (ViTs)** have demonstrated exceptional performance across various image processing tasks. Future research could evaluate these architectures for improved AI-generated image detection accuracy or processing efficiency. Additionally, specialized architectures designed specifically for synthetic content detection, particularly those trained on GAN-generated images, merit exploration.

- *Dataset Expansion and Diversification:*

A significant limitation of our current research involves dataset size and variety constraints. Enhancing detection system reliability requires larger, more diverse datasets incorporating higher-resolution images, samples from various AI models, and content from diverse sources including social media platforms, news outlets, and entertainment media. Expanded datasets would improve model performance across broader ranges of AI-generated content. **Real-Time Detection**: The current model might take some time to process large images or datasets. In the future, it would be useful to make the model faster so it can classify images in real time. This could help with practical uses, like detecting fake images on social media or during live TV broadcasts.

- *Include Other Types of Data:*

This research focused only on images, but AI-generated content often comes with other data, like **text** or **metadata** (such as the time and place an image was uploaded). Future work could combine different types of data into one model. For example, if an AI-generated image has a caption, analysing both the image and the text together could help the model detect fakes more accurately.

- *Handle Evolving AI Models:*

AI-generated images are getting better and harder to detect, with some newer AI tools specifically designed to fool detection systems. Future research could focus on making models stronger against these changes by using **adversarial training**. This means training the model to recognize fake images even when they have been altered to look more real.

- *Address Ethical Issues:*

AI-generated images raise serious concerns about **misinformation** and **manipulation**, especially when they are used to mislead people. Future research should look not only at improving detection but also at the ethical side of these technologies. It should consider how these technologies can be misused and how we can set rules to ensure they are used responsibly. This includes thinking about how fake images might impact **public trust**, **privacy**, and **security**.

- *Explore Other Applications:*

This research focused on detecting if an image is real or AI-generated, but there are other ways this technology could be used. For example, it could be helpful in areas like **medicine**, where fake medical images could harm patients. Detecting fake medical images would be crucial for correct diagnosis. It could also be useful in other fields like **journalism** and **art**, where fake images could mislead people.

- *Computational Efficiency Optimization:*

Future work should address model compression and optimization techniques to reduce computational requirements while maintaining performance.

- *Interpretability and Explainability:*

Developing methods to understand what features the model uses for classification could improve trust and help identify potential biases.

- *Cross-Domain Generalization:*

Research into how well models trained on one type of AI-generated content perform on images from different generative models.

VII. CONCLUSIONS

This research addresses the critical challenge of generative technology misuse, particularly focusing on AI-generated image detection through a comprehensive model development approach. Initially, our feature extraction methodology demonstrated limited effectiveness due to difficulties in distinguishing between authentic and AI-generated images. This limitation likely stemmed from the model's inability to adequately adapt to the distinctive characteristics of synthetic imagery. Model enhancement involved modifications to the fully connected layer architecture and training across various epoch configurations. These modifications resulted in substantial performance improvements, with accuracy increasing significantly following ResNet50 fine-tuning implementation. The fine-tuning approach enabled more effective learning and enhanced classification capabilities. Comparative analysis across different epoch configurations revealed that 5-epoch fine-tuning produced optimal accuracy, while extended training periods (10 or 15 epochs) yielded slightly reduced performance, likely due to overfitting where the model becomes overly specialized to training data and loses generalization capability. Our findings demonstrate that fine-tuning pre-trained models such as ResNet50 can substantially improve AI-generated image detection capabilities. However, results also indicate opportunities for further improvement, as model performance remains susceptible to factors such as overfitting and training instability. This research contributes to the growing body of work addressing synthetic media detection and provides a foundation for future advancements in this critical area. The implications of this work extend beyond technical achievement to address broader societal concerns about media authenticity and digital trust. As AI-generated content becomes increasingly sophisticated, robust detection methods become essential for maintaining information integrity across digital platforms.

REFERENCES

[1]. Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).

[2]. Jovanović, Radiša. "Convolutional Neural Networks for Real and Fake Face Classification." In Sinteza 2022-International Scientific Conference on Information Technology and Data Related Research, pp. 29-35. Singidunum University, 2022.

[3]. 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.

[4]. P. Datasets, "140k Real and Fake Faces," [Online]. Available: https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces. [Accessed 23 3 2022].

[5]. P. Datasets, "70k Real Faces," [Online]. Available: https://www.kaggle.com/c/deepfake-detection-challenge/discussion/122786. [Accessed 23 3 2022].

[6]. P. Datasets, "1 Million Fake Faces on Kaggle," [Online]. Available: https://www.kaggle.com/c/deepfake-detection-challenge/discussion/121173. [Accessed 23 3 2022].

[7]. Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434 (2015).

[8]. Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. "Image-to-image translation with conditional adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition (2017): 1125-1134.

[9]. Zhang, Richard, Phillip Isola, and Alexei A. Efros. "Colorful image colorization." European conference on computer vision (2016).

[10]. Odena, Augustus, Christopher Olah, and Jonathon Shlens. "Conditional image synthesis with auxiliary classifier GANs." Proceedings of the 34th International Conference on Machine Learning-Volume 70 (2017): 2642-2651.

[11]. Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition (2019): 4401-4410.

[12]. Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks." Proceedings of the IEEE international conference on computer vision (2017): 2223-2232.

[13]. Li, Xin, Jianchao Yang, Hongdong Li, and Haibin Ling. "Horizon: A scalable framework for learning deep generative models for 3D object modeling." IEEE Transactions on Pattern Analysis and Machine Intelligence 41.10 (2019): 2379-2392.

[14]. Kingma, D.P., and M. Welling. "Auto-Encoding Variational Bayes." International Conference on Learning Representations (ICLR), 2014.

[15]. Xie, L., and L. Ren. "Deepfake detection with GAN-based methods." Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2021): 3154-3162.

[16]. Choi, Yong-Hyun, et al. "Learning deep generative models for efficient image synthesis and generation." International Journal of Computer Vision (2020).

[17]. Wu, Y., and M. Zeng. "Exposing deepfakes with adaptive learning." IEEE Transactions on Image Processing 29 (2020): 741-755.

[18]. Rössler, Andreas, et al. "FaceForensics++: Learning to Detect Manipulated Facial Images." Proceedings of the IEEE International Conference on Computer Vision (2019): 1-11.

DATA SET AVAILABILITY

The dataset utilized in this research is accessible through the Kaggle platform at: P. Datasets, "140k Real and Fake Faces," [Online]. Available: https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces This dataset has been previously employed in the research conducted by Jovanović, Radiša. "Convolutional Neural Networks for Real and Fake Face Classification." In Sinteza 2022 International Scientific Conference on Information Technology and Data Related Research, pp. 29-35. Singidunum University, 2022.