Deepfake Image Detection Using Deep Learning Models

Name : Sajib Das
ID : 171-35-2061

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

APPROVAL

This thesis titled on "**Deepfake Image Detection Using Deep Learning Model**", submitted by **Sajib Das (ID: 171-35-2061)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

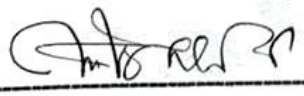
Dr. Imran Mahmud Professor & Head
Department of Software Engineering
Faculty of Science and Information Technology Daffodil International University

Internal Examiner 1

Md Shohel Arman Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology Daffodil International University

Internal Examiner 2

Md. Rajib Mia
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology Daffodil International University



External Examiner

Md Habibur Rahman Associate Professor
Department of Computer Science and Engineering Islamic University, Bangladesh

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and, in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science of Science.

(Supervisor's Signature)

Mr. Suprove Chandra Sarkar

Lecturer, Department of Software Engineering

Daffodil International University

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

Sajib Das

(Student's Signature)

Sajib Das
171-35-2061
Department of Software Engineering
Daffodil International University

ACKNOWLEDGEMENTS

By the grace of Almighty, I have come to this fur. My parents supported me throughout my journey. I am very grateful to my supervisor Lecturer Mr. Suprove Chandra Sarkar, for his guidance, mentorship and motivation, which has been a great help to complete this study. I would like to thank all the teachers who contributed to my academic growth and shaped the person who I am now. I am thankful to the Department of Software Engineering and Daffodil International University for providing the resources.

DEDICATION

To my parents, for their support, dedication and love.

ABSTRACT

Fake images are becoming a threat to our society. Recognizing fake images with the naked eye is a very difficult task. And in many cases it is impossible to recognize even after much effort. Due to which it is becoming easier to spread false news among people and misguide them with less effort. This problem will become even bigger in the future and will pose a threat to human safety. In this research paper I performed deep learning models one by one and from those I took the best performing models. With their help, I created a hybrid model using the best performing models. My goal is to create a hybrid model that will give high accuracy and high recall rate. And will also perform well on any custom dataset.

To this end, a custom hybrid model was created and implemented leveraging innovative factorized residual blocks to achieve efficient feature extraction with reduced parameter counts. Firstly I performed with single models MobileNetV2, InceptionV3, EffieientNetV2B0, ResNet50, VGG16. From these models I took 2 models with best accuracy, precision, recall and f1 score and combined them. Then I fine-tuned for binary classification of real versus AI-generated images. This hybrid model trained and validated on the CIFAKE dataset, which contains labeled samples of authentic and synthetic images. I made a custom dataset also to check how my hybrid model on unseen dataset and it performed pretty well. Training leveraged GPU acceleration within TensorFlow/Keras frameworks to optimize computational performance.

The model performance results are presented in the form of accuracy, precision, recall, f1 score, performance metrics and confusion metrics generated by the model. And I saw that Mobilentiv2 and ResNet50 are performing the same results on the CIFAKE dataset. And the rest of the models are performing almost close. So I created a hybrid model using MobilityTV2 and ResNet50 as the best performing single model. The hybrid model performs well. Robustness check is done by using a custom dataset and checking the unseen dataset. For the CIFAKE dataset my hybrid models accuracy was 0.98 and recall rate also increased to 0.98. And for the custom dataset my accuracy was 0.8875 and recall rate was 0.8875.

This study underscores the feasibility of using hybrid models for fake image detection, providing a practical pathway toward scalable, real-time systems capable of mitigating the spread of AI-generated misinformation.

Keywords: Fake Image Detection, Hybrid Models, MobileNetV2, ResNet50, VGG16, EfficientNetV2B0, InceptionV3 Hybrid Deep Learning Model, CIFAKE Dataset, Custom dataset, Synthetic Media, Deepfake Detection, Real-Time Inference, Model Evaluation

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
CNN	Convolutional Neural Network
ReLU	Rectified Linear Unit
ML	Machine Learning
DL	Deep Learning
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
SGD	Stochastic Gradient Descent
Adam	Adaptive Moment Estimation
F1-score	Harmonic Mean of Precision and Recall
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
TPU	Tensor Processing Unit

CHAPTER 1

INTRODUCTION

1.1 Overview

The rapid progress in artificial intelligence (AI) has revolutionized synthetic media creation, with generative models—especially Generative Adversarial Networks (GANs)—capable of producing highly realistic images that are often indistinguishable from authentic photographs. These advances have unlocked numerous creative and industrial applications, spanning entertainment, design, and scientific visualization. However, they have also amplified concerns regarding misinformation, digital forgery, and malicious manipulations of visual content. The proliferation of convincingly fabricated images threatens the integrity of journalism, legal systems, and public discourse, necessitating the development of reliable and efficient detection mechanisms to counteract such misuse (Nguyen et al., 2022).

Traditional deep learning techniques for detecting fake images have achieved promising results through the use of deep convolutional neural networks. Yet, many of these approaches rely on large-scale, computationally intensive architectures that require significant hardware resources and prolonged inference times (Zhou et al., 2021). This dependency severely restricts their applicability in real-time systems or on devices with limited processing capabilities, such as smartphones, embedded sensors, and surveillance units. Consequently, there is an escalating demand for lightweight deep learning models that retain high detection accuracy while minimizing computational overhead.

Previous studies have attempted on different individual model architectures including MobileNetV2, ResNet50, EfficientNetV2B0, VGG16, InceptionV3 etc. They are training from scratch and transfer learning strategies to identify fake or synthetic images. But no single architecture is performing well on diverse datasets like unseen dataset. Because they extracted features from a single model. But Hybrid models, which combine features from different models show better detection results and also robustness. Like ResNet50 models are adept at capturing hierarchical representations through residual connections, while MobileNetV2B0 architectures offer efficient feature extraction using depthwise separable convolutions, making them ideal for deployment in resource-restricted settings.

This research is a hybrid centric study, where multiple models are worked together. A hybrid approach is created by combining the features of different models to achieve greater success in fake images detection and to increase the recall rate. The model can identify almost all fake images as fake. The hybrid model combines the two best performing models. The learned elements of the model are combined and concatenated, which allows the model to maintain accuracy more efficiently.

Model training and test on the same dataset, naturally gives slightly better results. But to check whether the model is robust, it should be tested on an unseen dataset, which actually shows the model acceptability of the model. Diversity of the dataset is essential for real life testing. For combining features from different models, Hybrid models give better results in real life datasets. Using hybrid models increases the resilience of the model.

As synthetic image generation techniques evolve rapidly, detection frameworks must adapt. This research address is not only just better accuracy. Our goal is to achieve resilience. Our model achieved quite good accuracy and recall rates on the unseen dataset, indicating its reliability.

1.2 Background & Motivation

The world is currently in the midst of an internet revolution. More than ever before, people are more dependent on the internet and this dependence is increasing the growth of AI. It is now very easy to create fake images. Advanced photo editing software has made it more difficult to detect fake images. They are capable of creating very realistic images that are undetectable to the human eye. These images are often used for various crimes, spreading misinformation, inciting riots, and various cyber crimes. They are used to manipulate people's opinions. People are being scammed and blackmailed in various ways.

A robust analytical approach is needed to overcome these problems. Deep learning models are performing well in image classification in fake images in terms of image recognition.

Hybrid models bring breakthrough changes resulting in image recognition. By combining different features extraction strategies from different models and enhancing performance to classify fake or real images. Although hybrids need more computational cost than a single model, they achieve better recall rate also which is a great thing. It increases reliability.

The primary motivation of this research is making a hybrid model to enhance accuracy, recall, precision and f1 score. The research is aimed at improving the accuracy of fake image detection and contributing to the fight against fake images to decrease digital crime. And another thing is to create custom datasets and test them to get better results so that robustness can be achieved.

1.3 Research Question

This study aims to answer the following questions:

- What is the current state of fake image detection accuracy when using hybrid deep learning models compared to traditional or single models?
- Which specific hybrid deep learning architecture performs best in detecting fake images while maintaining high efficiency?
- Can a computational approach using hybrid models achieve comparable or better performance than conventional methods using the CIFAKE dataset?
- Proposed hybrid model performing well or not on custom dataset?

1.4 Research Objective

The key objectives of this research are:

- Assess the accuracy of existing models on CIFAKE dataset and look for better accuracy and recall rate.
- To design and implement a novel hybrid deep learning architecture to detect fake images and better classify results for fake and real both classes.
- To create a custom dataset and test on proposed hybrid models and get better results to check robustness.

1.5 Scope

This research concentrates on enhancing fake image detection through hybrid deep learning models. The primary dataset utilized is CIFAKE, encompassing diverse classes of real and synthetic images. Then also used a custom made dataset to check robustness.

The study covers model design, training, fine tune, augmentation, evaluation, and comparative analysis of multiple hybrid architectures, with a focus on balancing detection performance. While the research does not extend to adversarial robustness or multimodal forgery detection, it lays a robust foundation for practical, scalable fake image detection solutions with potential applications in security, forensic analysis, and media verification. There is also scope for making this model more lightweight to implement in real life application.

CHAPTER 2

Literature Review

The detection of fake images is highly challenged nowadays, because of Generative Adversarial Networks (GANs) and diffusion-based approaches. And these approaches are evolving rapidly day by day. These generative techniques have dramatically advanced in recent years. They create images in a way that makes them much tougher to detect. Which actually appears to be a very big threat to cyber security and the justice system. There is no way to tell by looking at them that they are actually fake. There is no way to tell with the naked eye that they are fake. We need a reliable, scalable, automated system that can protect us from this image crime and classify almost all real and fake images with high precision and recall.

2.1.1 Early Detection Approaches

In the early stages, fake image detection methods relied on hand-built engineering. These approaches analyzed specific artifacts introduced during image manipulation, leveraging domain knowledge to extract cues such as inconsistencies in color filter arrays, texture irregularities, or anomalies in frequency domain representations. All of these methods could detect images in gaps in a very small number of datasets, and even then not completely accurately. When faced with a raw dataset, these would have been a complete failure. They would not have been able to perform well outside of a specific dataset. As a result, their use has declined dramatically over time, especially since the development of image-making techniques.

2.1.2 Emergence of Deep Learning Techniques

With the development of deep learning methodologies, particularly convolutional neural networks, fake image detection has undergone a major transformation. In which it has been possible to analyze the image in detail in various ways, pixel by pixel, without any manual touch. CNNs and, more recently, vision transformers, are performing very well to find clues to those left behind synthetic image generators. They learn layered features that help to detect synthetic images. They can detect high manipulation. They are actually performing better than older handcrafted methods in both accuracy and adaptability.

Despite their effectiveness, many deep learning models have low recall rates. This is a challenge. This is causing reliability to decrease. And many models can not perform well on custom datasets and are not performing well, which is another challenge. We need to overcome these challenges.

2.1.3 Hybrid and Ensemble Approaches

However, single models are a bit lightweight, which is quite convenient for use on smaller devices. And they are quite good at detection. But still, if features taken from multiple models are combined to create a hybrid model, its performance surpasses that of a single model. The features extracted from MobileNetV2 and ResNet50 are the same, but the features combined are the same, and the concatenation is the same. However, its accuracy and recall, precision, and F1 scores all increase significantly. They also demonstrate considerable robustness in image classification.

2.1.4 Research Gap

Although deep learning models have improved significantly, we still need to make significant improvements due to ongoing challenges. Current works have achieved impressive results by leveraging powerful convolutional neural networks (CNNs) and transfer learning from large-scale pretrained models. However, closer examination reveals gaps that restrict their effectiveness and generalization in real-world applications.

1. **Limited Dataset Diversity**
   Many studies rely on a few well-curated benchmark datasets such as CIFAKE, CelebA, or FFHQ. While these datasets provide controlled experimental environments, they do not fully capture the wide variability of real-world images, particularly those generated by different and evolving generative models (e.g., StyleGAN3, Stable Diffusion, or future unseen architectures). As a result, the models may fail to generalize effectively to images outside their training distribution.

2. **Model Complexity vs. Practical Deployment**
   Most state-of-the-art approaches depend on large, computation-heavy architectures (e.g., ResNet50, VGG16, InceptionV3). Although these models achieve high accuracy, their computational overhead makes them unsuitable for deployment in resource-constrained environments such as mobile devices, edge systems, or real-time forensic analysis tools. There is a clear gap in exploring lightweight yet accurate hybrid models that balance performance and efficiency.

3. **Overemphasis on Accuracy Metrics**
   Prior studies frequently focus on reporting high accuracy while overlooking other critical evaluation metrics such as recall, precision, and F1-score. In fake

image detection, recall is particularly important because failing to detect a fake (false negative) could have severe consequences. However, this dimension remains underexplored, leaving a gap in designing models that optimize for comprehensive reliability rather than accuracy alone.

4. **Lack of Multimodal Approaches**
   Most of the research that currently exists is largely image-centric. But we are seeing that audio, video, and other forms of censorship are being used around us lately. This is a huge threat. We need to pay attention to this. We also see that the work of building hybrid models using multi-model ensembles is decreasing. This is very important for increasing security. In this case, high compression power is required, which is a big problem. We need to look at how we can make a hybrid lightweight model. So that the models can be easily used on mobile devices.

5. **Explainability and Trustworthiness**
   A major problem with deep learning models is that almost all of them operate in a black box, meaning that there is little chance that ordinary people will be able to see or understand anything. So, actually, there may be a trust issue here in important areas of journalism, law enforcement, and national security from their perspective. There needs to be improvement in the area of Explainable AI. So that people can get a clear idea about why this model is detecting an image as fake or real.

6. **Robustness Against Evasion Attacks**
   Recent research has shown that models can be fooled by small changes. In this case, various frequency tricks are being used. We need to make more improvements in these areas so that our models can catch any tricks and give accurate predictions.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter presents the hybrid model that is created for fake image detection or binary classification and how it is built. And it is stated exactly how the adopted method was evaluated. Here the aim is to build and use hybrid models from single models, and get best accuracy, recall, precision and F1 score. Also check robustness from tests on custom dataset.

Here we used different single models called MobileNetV2, ResNet50, VGG16, InceptionV3 and EfficientNetV2B0. This shows the actual difference of all models. We generate accuracy, recall, precision and F1 score also. This allows fair comparison about their ability to binary classification. Here also generate confusion metrics to show how many predictions are right and how many are wrong.

The aim of this study is making a hybrid model to enhance accuracy rate. And enhance recall rate to build reliability of this hybrid model. Also aim is to make a custom dataset and test on this model to check our hybrid models robustness.

Methodology starts with preprocessing with the Cifake dataset. The Cifake dataset has balanced images for every class. Here we resize the image, normalize and apply augmentation also to enhance consistency and improve generalization.

We trained our train set with MobileNetV2, ResNet50, VGG16, InceptionV3 and EfficientNetV2B0 for 20 epochs and 30 epochs. After getting the results we pick two best models based on their result and make a hybrid model. Here we extracted features of both models and concatenated them. Hybrid models are trained using the Adam Optimizer. Here it is tuned with hyperparameters, early stoppping, transfer learning when applicable. We used data augmentation with rotation, flip, brightness etc.

After testing we generate accuracy, recall, precision, F1 score and confusion metrics. We also generate training and validation curves.

The flowchart highlights the iterative nature of the experimental design, underscoring how insights gained from initial training phases inform subsequent refinement and combination of models.

Overall, this systematic and rigorous methodology enables the identification of lightweight deep learning models that not only achieve high accuracy in fake image

detection but also maintain computational efficiency, making them suitable candidates for real-time deployment in practical security and forensic applications.

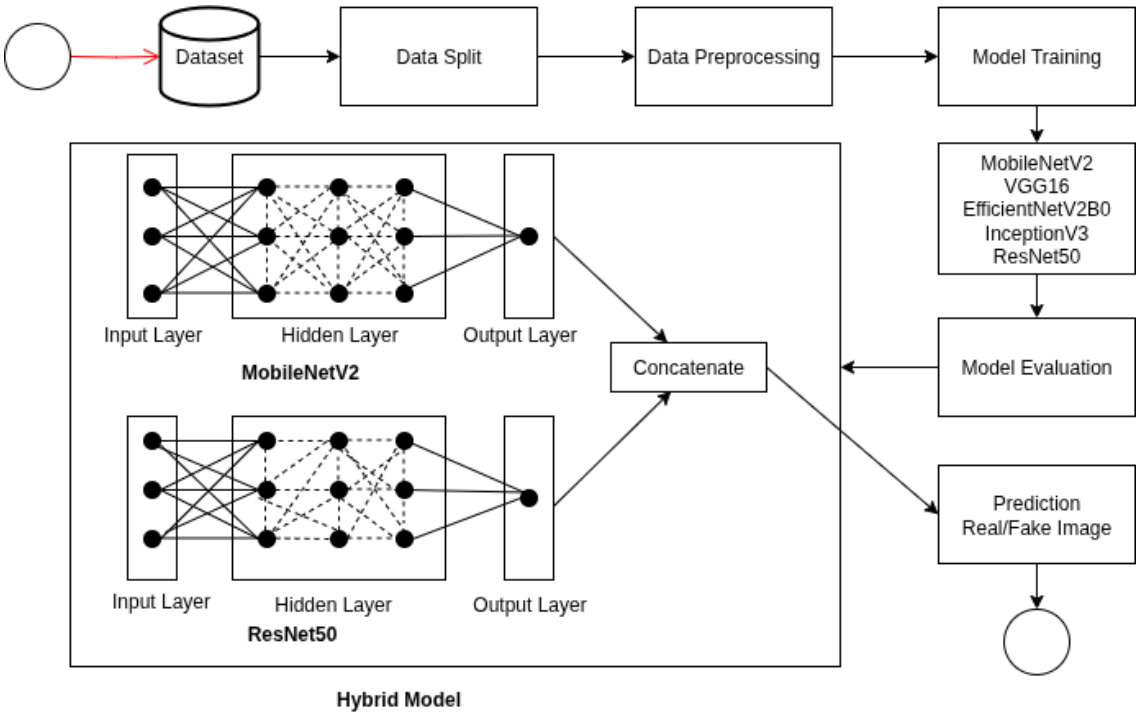


Figure 3.1 The architecture of methodology

3.2 Dataset Description

The primary dataset utilized in this study is the CIFAKE dataset, a publicly available benchmark specifically curated for fake image detection tasks. CIFAKE comprises 120,000 images, split evenly between authentic natural images sourced from CIFAR-10 and synthetically generated images produced by Stable Diffusion v1.4. Each synthetic image is designed to mimic the visual characteristics of CIFAR-10 categories, providing a challenging testbed for classification models.

The dataset is divided into training and testing subsets with a perfect class balance to prevent bias during model learning: 100,000 images (50,000 real and 50,000 fake) for training, and 20,000 images (10,000 real and 10,000 fake) reserved for testing. Images are resized to 224 X 224. And normalize to [0,1] scale. We used data augmentation like brightness, flip, rotation during training to improve robustness. Here only used public and synthetic data.

We created a custom dataset. We generated fake images from several sites and we collected real images from the Cifar-10 dataset.

3.3 Computational Model Development

Experiments were run on Kaggle using Python 3.11 and an NVIDIA Tesla P100 GPU for faster training on the large image dataset. We used Pandas for data handling, Scikit-learn for metrics, Matplotlib & Seaborn for visualization and Tesorflow/keras for model development and training.

3.3.1 Workflow Overview

The methodology followed a four-stage pipeline:

1. **Data Acquisition and Preprocessing**
   The CIFAKE images were loaded directly from the Kaggle repository and split into training, validation, and testing sets. Images were resized to a uniform resolution (typically 128×128 pixels) and normalized to scale pixel intensities. Data augmentation techniques—such as random rotation, horizontal flipping, and zooming—were applied to the training set to increase diversity and reduce overfitting.
2. **Model Selection and Architecture Design**
   A lightweight CNN architecture was custom-designed to serve as a baseline, incorporating depthwise separable convolutions to minimize parameter count while retaining feature extraction power. Additionally, transfer learning experiments were conducted using established lightweight pretrained models such as MobileNetV2, EfficientNetV2, ResNet50, InceptionV3, and VGG16, fine-tuned on the CIFAKE dataset.
3. **Model Training**
   Models were trained using the Adam optimizer with categorical cross-entropy loss, employing learning rate scheduling and early stopping based on validation loss to optimize performance and prevent overfitting. Batch sizes and epoch numbers were carefully selected to balance training efficiency and result stability.
4. **Evaluation and Performance Analysis**
   Models were evaluated using classification metrics (accuracy, precision, recall, F1-score) computed on the test set. Confusion matrices and ROC curves were generated for interpretability. Training and validation losses were monitored to assess learning behavior and generalization.

3.3.2 Model Architecture Components

The proposed lightweight CNN architecture comprises the following layers (see Figure 3.2):

- **Input Layer:** Accepts RGB images standardized to 224×224 pixels to align with pretrained models' input requirements.
- **Convolutional Layers:** Stacked lightweight convolution blocks utilizing depthwise separable convolutions reduce computational complexity while extracting spatial hierarchies.
- **Batch Normalization and ReLU Activation:** Applied after convolutions to accelerate convergence and improve generalization.
- **Dropout Layers:** Included strategically to prevent overfitting by randomly deactivating neurons during training.
- **Flatten Layer:** Converts feature maps into a one-dimensional vector suitable for classification.
- **Fully Connected Dense Layers:** Learn complex, high-level representations distinguishing real and synthetic images.
- **Output Layer:** A single neuron with sigmoid activation providing binary classification probability.

This architecture is specifically optimized to achieve a high detection accuracy while maintaining a low parameter count and computational demand, suitable for deployment in real-time and resource-constrained environments.

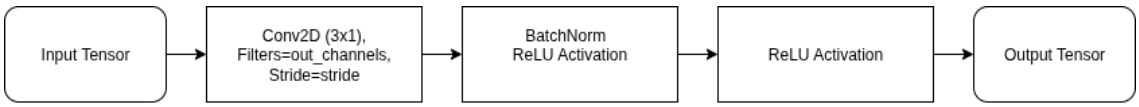


Figure 3.2 Architecture A

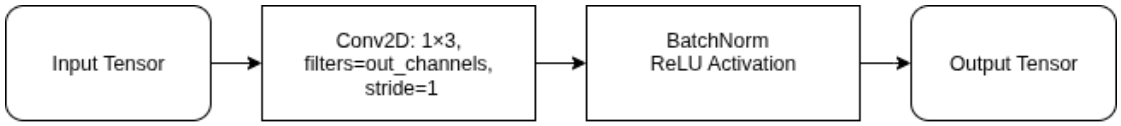


Figure 3.3 Architecture B

3.3.3 Mathematical Formulation

The binary classification task is formalized as learning a function fθ parameterized by θ, mapping input images xi∈RH×W×C to predicted labels y^i∈(0,1), where:

y^i=σ(Wdzi+bd)

Here, zi=fθ(xi) denotes the feature representation extracted by the convolutional layers, σ is the sigmoid activation function, and Wd, bd are weights and biases of the dense classification layer.

The model parameters are optimized by minimizing the binary cross-entropy loss function:

L(θ)=−N1i=1∑N[yilog(y^i)+(1−yi)log(1−y^i)]

where yi are the true labels indicating real (0) or fake (1) images.

A decision threshold τ=0.5 is applied to convert the output probabilities into binary predictions.

3.4 Model Evaluation

The effectiveness of the lightweight CNN and pretrained models was systematically evaluated to ascertain their ability to discriminate between authentic and AI-generated images.

3.4.1 Evaluation Metrics

The following metrics were selected for comprehensive assessment:

- **Accuracy:** Proportion of correctly classified images among all test samples, reflecting overall performance.
- **Precision:** The fraction of correctly identified fake images among all images predicted as fake, indicating the model's ability to reduce false positives.
- **Recall (Sensitivity):** The ratio of detected fake images to the total actual fake images, measuring detection completeness.
- **F1-Score:** The harmonic mean of precision and recall, balancing false positives and false negatives to provide a single robust performance measure.

3.4.2 Confusion Matrix and Visualization

Confusion matrices showed all correct prediction and wrong prediction also. From it we can be clear about the model performance. Here also plotted accuracy and loss curves. From it we can check learning progress, overfitting and model stability.

3.4.3 Validation Strategy

Here the dataset is split by 3 sets. Training, Testing and Validation. All sets are balanced. I used early stopping to prevent overfitting. And here checkpoints saved best model weights.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Classification Performance Overview

This research uses the CIFAKE dataset. This dataset is a balanced dataset for binary classification of deep learning models. Here are 5 models that have been used. And they are MobileANetV2, ResNet50, InceptionV3, EfficientNetV2B0 and VGG16. First, a single model was used. 20 and 30 epochs were run on it for the train. Then the test was performed on the test data. From here, a hybrid model was created using the two models that gave the best results. Then, after training the hybrid model with the A-train data, it was tested again and the expected results were obtained. Also, a custom dataset was created and testing was run on that custom dataset. The expected results were obtained there too.

Our model performed quite well. We generated metrics for the model's results. We calculated the accuracy, precision, recall and F1 score. Here accuracy defined the proportion of all classifications that were correct. It could be real and fake also. Recall means the true positive rate. That means how many fake images do we have and how many gaps was the model able to detect. Precision means is the proportion of all the model's positive classifications that are actually positive.

Among the models, the Mobilnetv2 Resnet50 and Efficientnetv2B0 almost always result in the same results for the 20 epochs. And when I used 30 epochs EfficientNetV2B0 had a slightly loose accuracy rate and recall rate. So here we actually get an idea about the acceptability of the models. Models that perform quite well. And the models are quite good at image recognition or classification.

In comparison, the efficient TV2B0 is showing less power. Which means its compatibility with the CIFAKE dataset is probably a little less. We saw that B16 performed remarkably well, but it was lower than the others. Which indicates that despite its large parameters, it is able to learn less during training. The same applies to InceptionV3.

We see that in the case of the best performing two models, even though the duration was increased from 20 to 30 epochs, it had no impact on performance. The accuracy, precision, recall and F1 score was the same as it was for the 20 epochs. This is pointing out that using 20 epochs, it also avoids overfitting.

So from here, a hybrid model was created using ResNet50 with MobilNetV2 to further improve the classification power. MobilNetV2 has been added to the powerful architecture of ResNet50. Features from both have been extracted and combined. So that the gap image detection e-mean classification power is further increased. And our hybrid model performed as expected. It is capable of delivering high accuracy, precision, recall and F1 scores.

4.1.1 Detailed Evaluation: MobileNetV2 at 20 Epochs

In the case of MobileNetV2 after 20 epochs, we see how a lightweight model learns efficiently. We see that this is a high accuracy without any overfitting. And there loss also decreased.

The confusion matrix (Figures 4.1 and 4.2) highlights a high concentration of true positives and true negatives, with very few false positives or false negatives. This balanced distribution underscores the model's capability to accurately discriminate between real and synthetic images. The minimal misclassification rates suggest MobileNetV2's architecture, which leverages depthwise separable convolutions, is well-suited for capturing the subtle artifacts present in AI-generated images within CIFAKE.

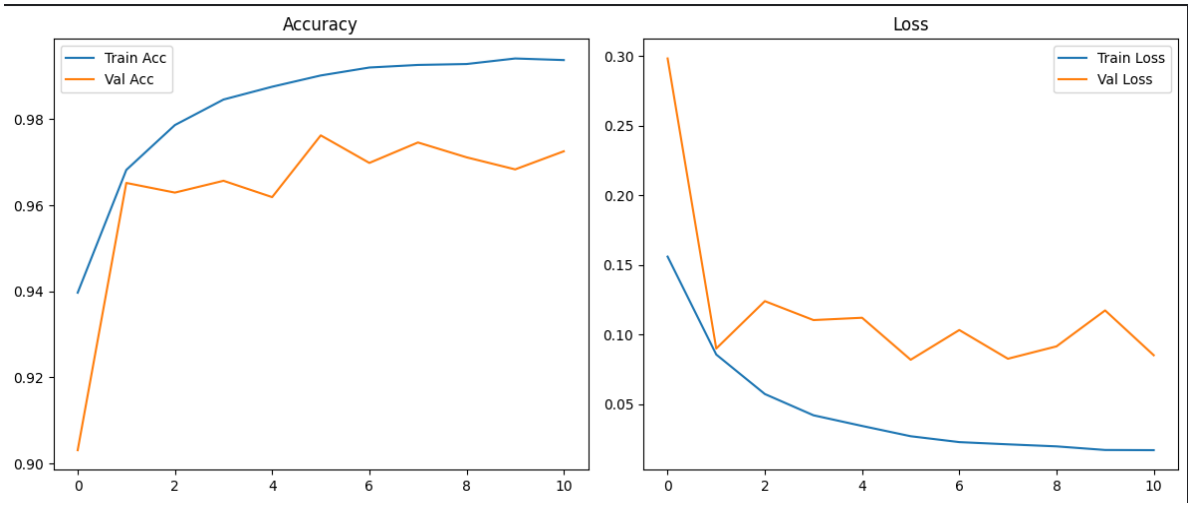


Figure 4.1 Accuracy of **MobileNetV2** Model at 20 Epochs

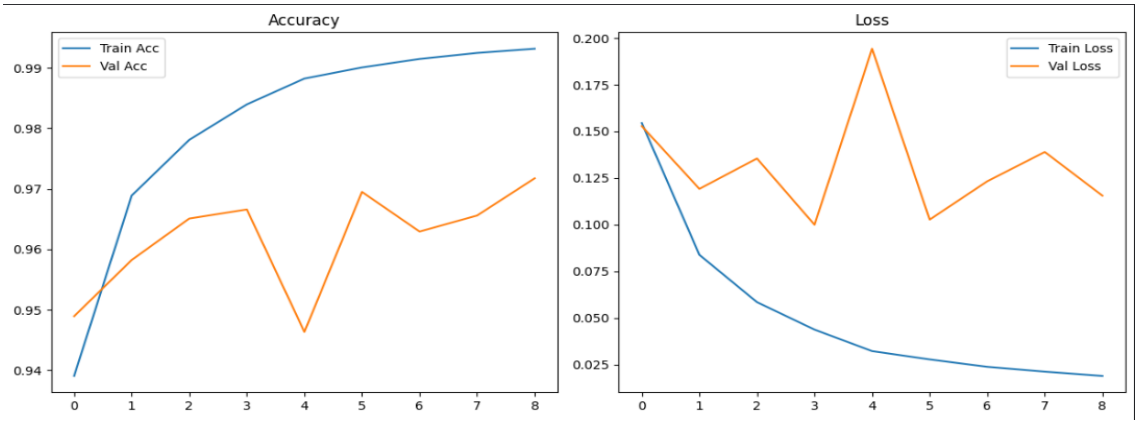


Figure 4.2 Accuracy of **MobileNetV2** Model at 30 Epochs

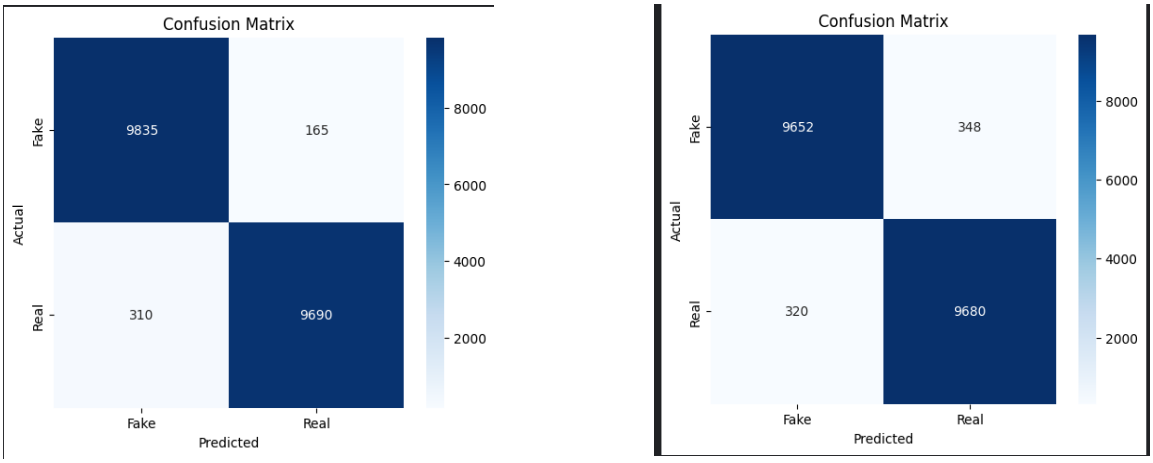


Figure 4.3 Confusion matrix of **MobileNetV2** Model at 20 and 30 Epochs

4.1.2 Summary of Other Models' Performance

- **EfficientNetV2B0:** Despite its design for efficiency, EfficientNetV2B0 achieved only around 90% accuracy at both 20 and 30 epochs, indicating moderate success in the fake image detection task. Precision and recall were correspondingly lower, suggesting some challenges in consistently identifying fake images or avoiding false alarms.

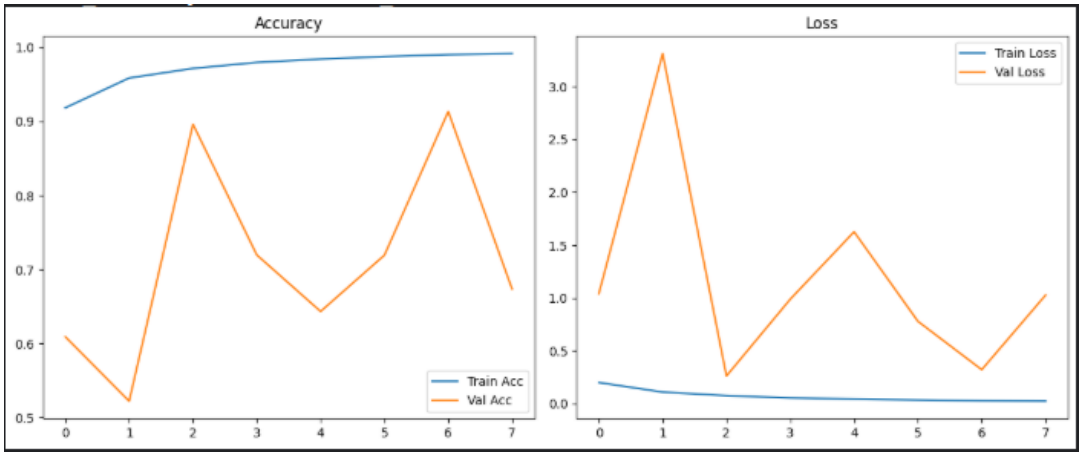


Figure 4.4 Accuracy of **EfficientNetV2B0** Model at 20 Epochs

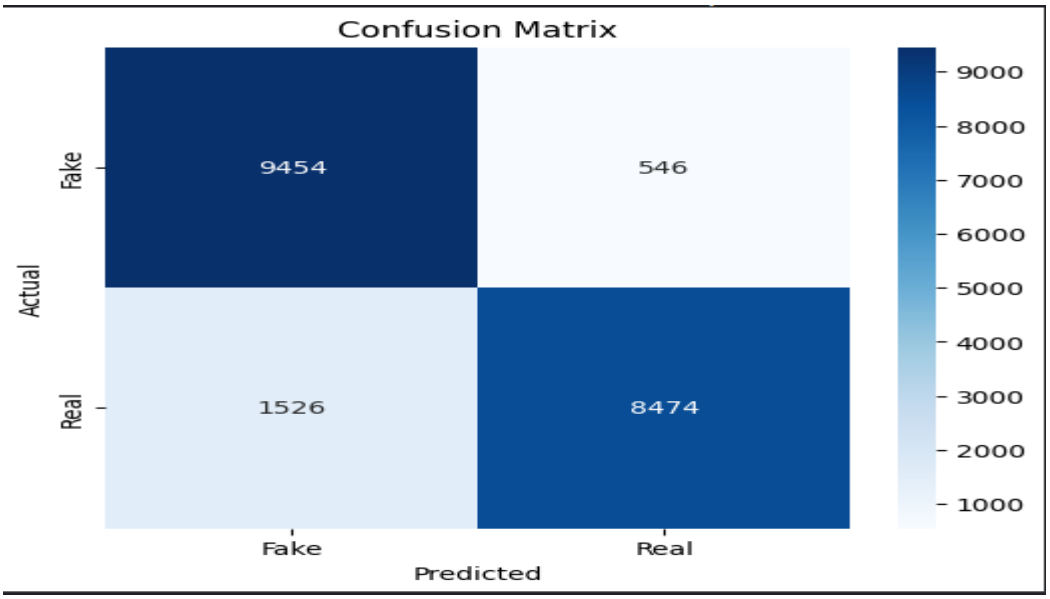


Figure 4.5 Confusion matrix of **MobileNetV2** Model at 20 Epochs

- **InceptionV3:** Exhibited excellent performance, reaching 98% accuracy across both training durations. Precision and recall metrics mirrored this success, confirming the architecture's suitability for capturing diverse image features critical for detection.

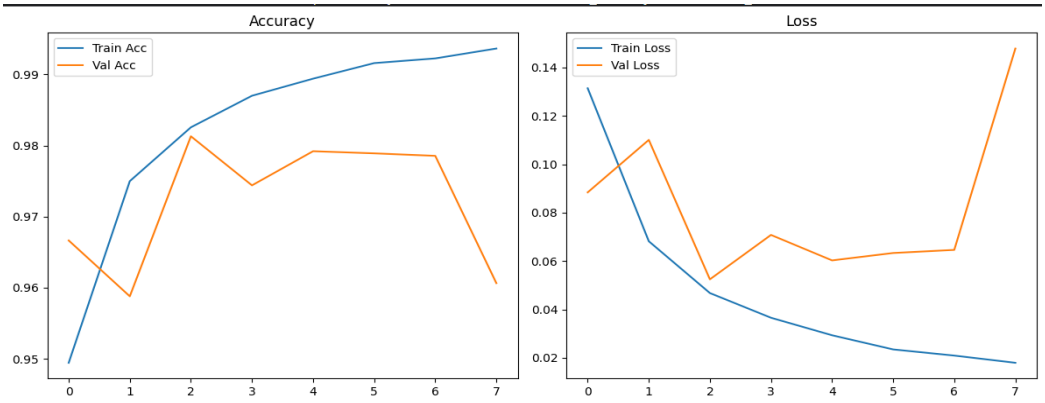


Figure 4.6 Accuracy of **InceptionV3** Model at 20 Epochs

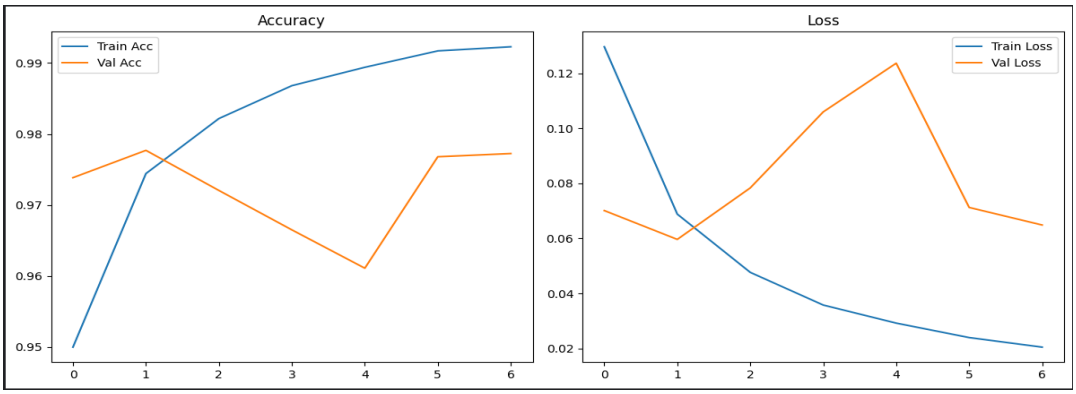


Figure 4.7 Accuracy of **InceptionV3** Model at 30 Epochs

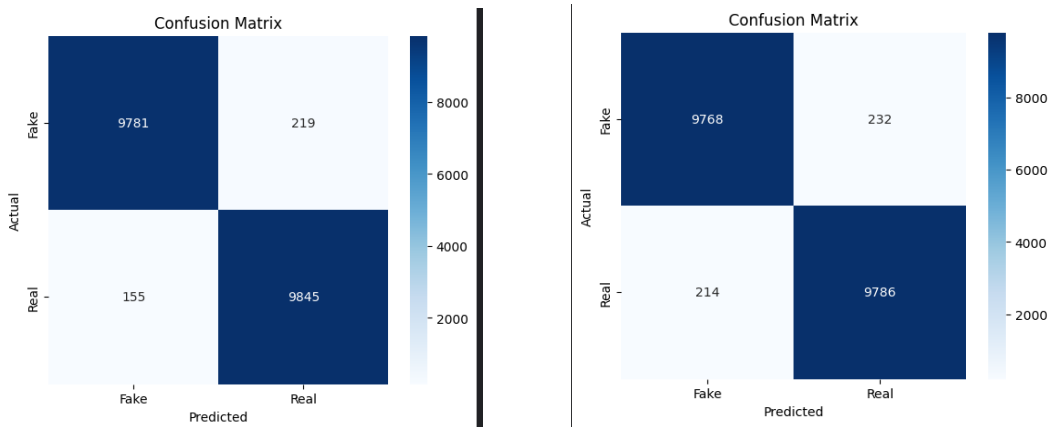


Figure 4.8 Confusion matrix of **InceptionV3** Model at 20 and 30 Epochs

- **ResNet50:** Delivered high and stable accuracy near 98%, comparable to MobileNetV2, with consistent results across training epochs. Its residual learning framework likely contributes to its robustness in learning discriminative features for the classification task.

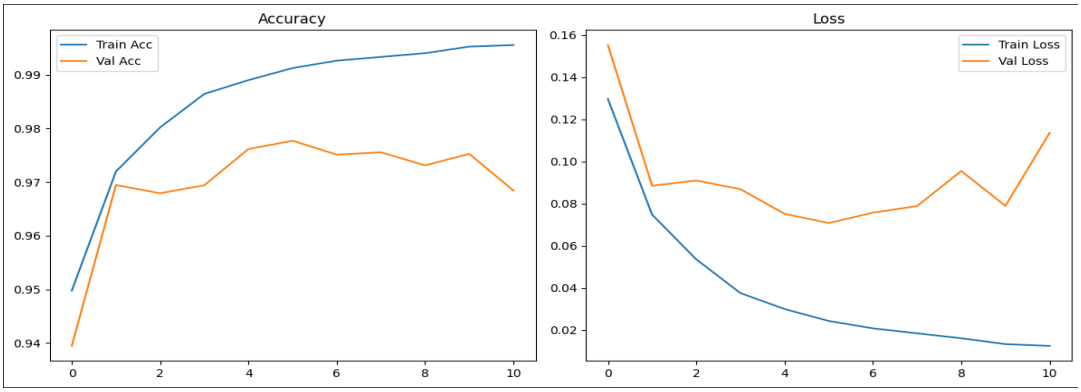


Figure 4.9 Accuracy of **ResNet50** Model at 20 Epochs

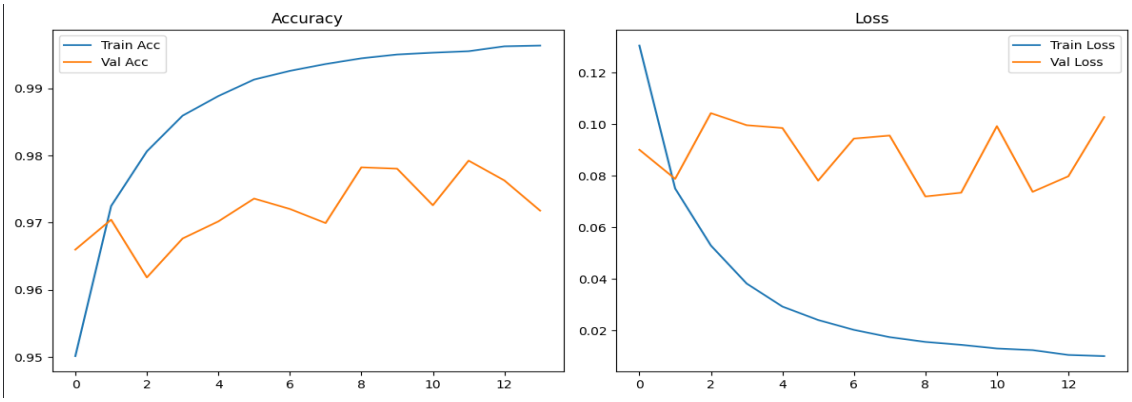


Figure 4.10 Accuracy of **ResNet50** Model at 30 Epochs

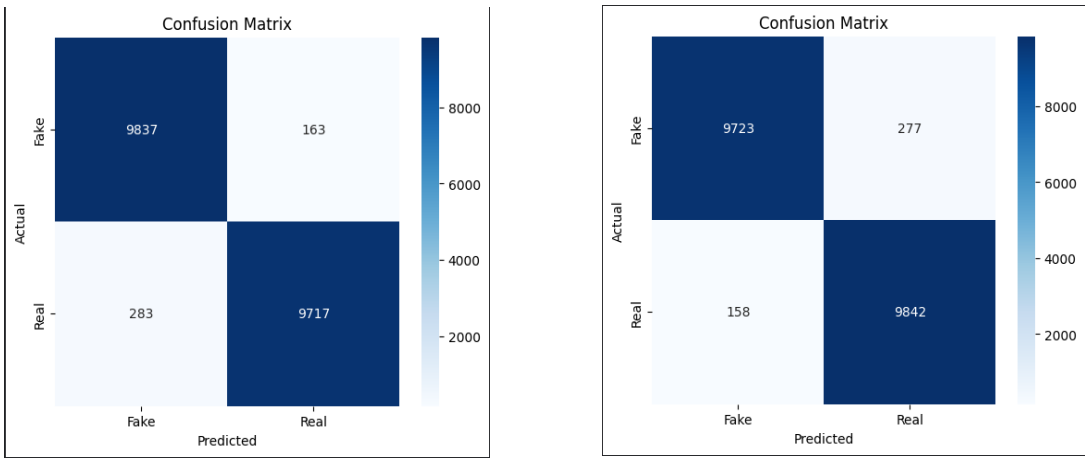


Figure 4.11 Confusion matrix of **ResNet50** Model at 20 and 30 Epochs

- **VGG16:** While slightly behind the top models, VGG16 still maintained strong performance (~97% accuracy), reflecting its status as a deeper, yet older, CNN architecture.

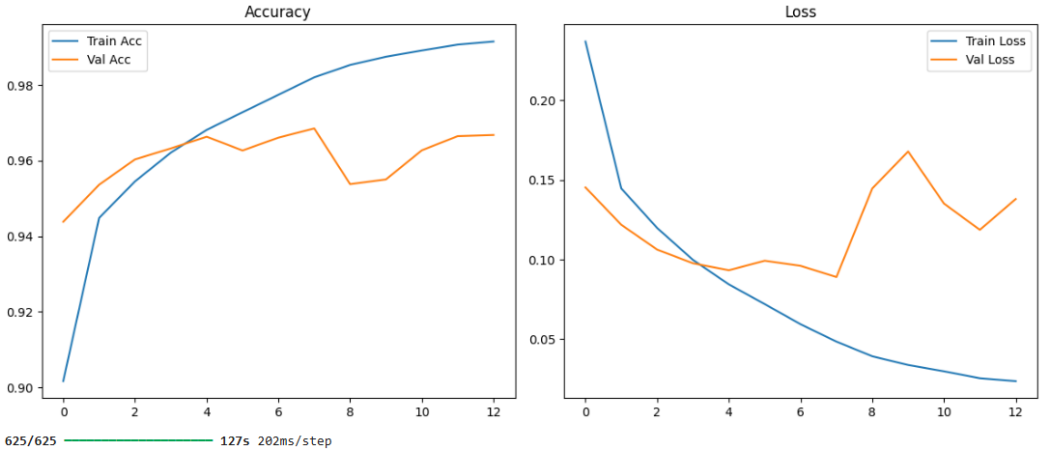


Figure 4.12 Accuracy of **VGG16** Model at 20 Epochs

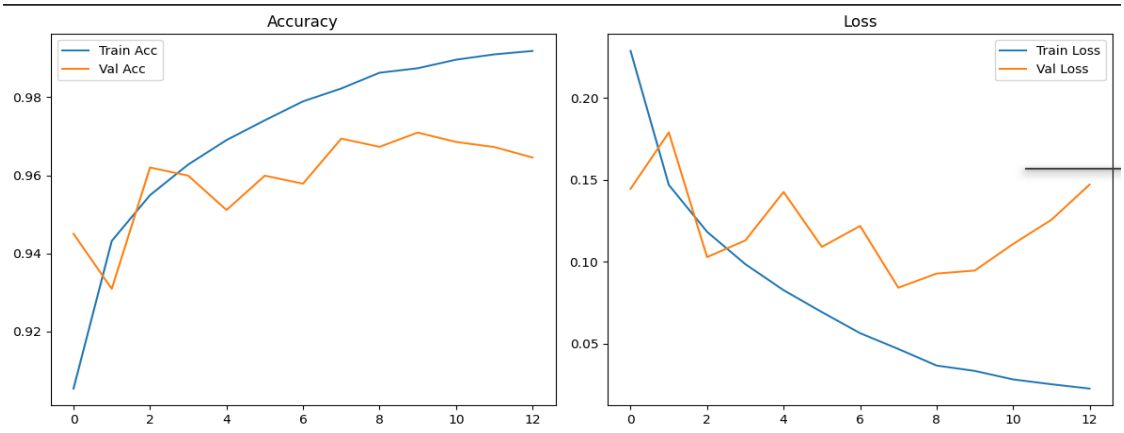


Figure 4.*13* Accuracy of **VGG16** Model at *30* Epochs

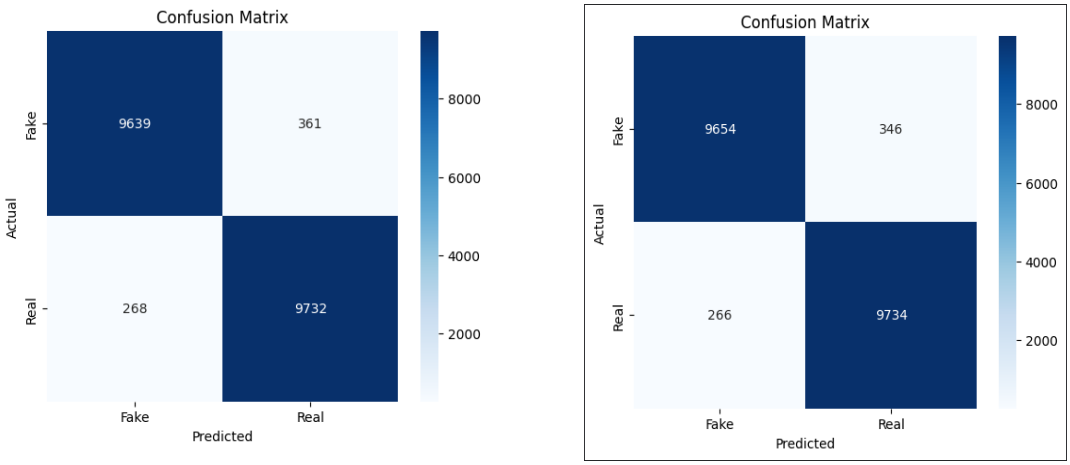


Figure 4.14 Confusion matrix of **VGG16** Model at 20 and 30 Epochs

4.2 Hybrid Model Evaluation

The hybrid model, constructed by combining MobileNetV2 and ResNet50 feature extractors, was rigorously evaluated on a held-out test set to confirm its generalization capacity beyond the training data. This evaluation is critical for validating the model's practical viability in real-world scenarios where unseen images may vary in style and quality.

The hybrid architecture maintained a strong classification accuracy of 98%, with precision, recall, and F1-score values similarly high at 0.98 across both macro and weighted averages. This consistent performance highlights the model's ability to reduce both false positives (misclassifying real images as fake) and false negatives (failing to detect fake images), an essential quality for trustworthy fake image detection systems.

The confusion matrix is showing the model classified images. It is quite reliable. Because here we can see it gives almost all correct predictions. But there are very few mistakes. It is so important for media checks and foresics. Hybrid model is working well because of MobileNetV2 and ResNet50 models features combination.

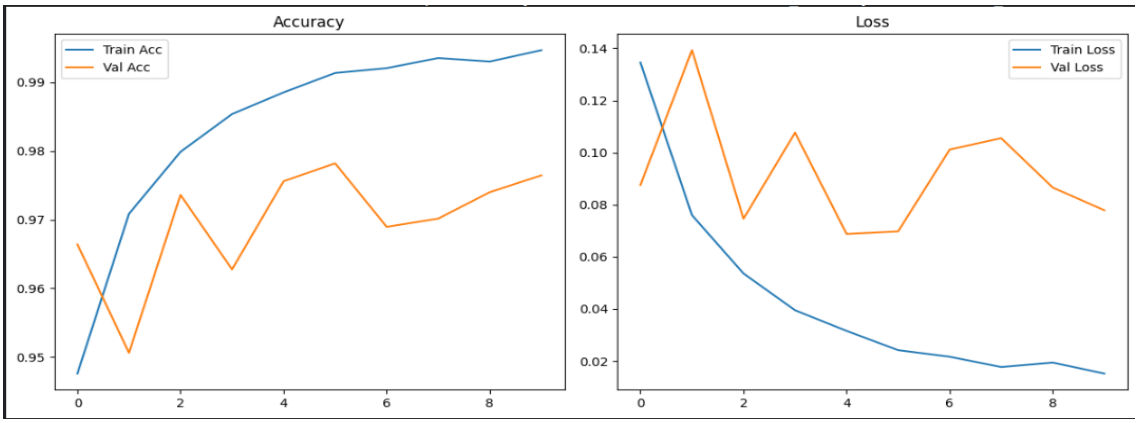


Figure 4.15 Accuracy of **Hybrid Model** at 20 epochs

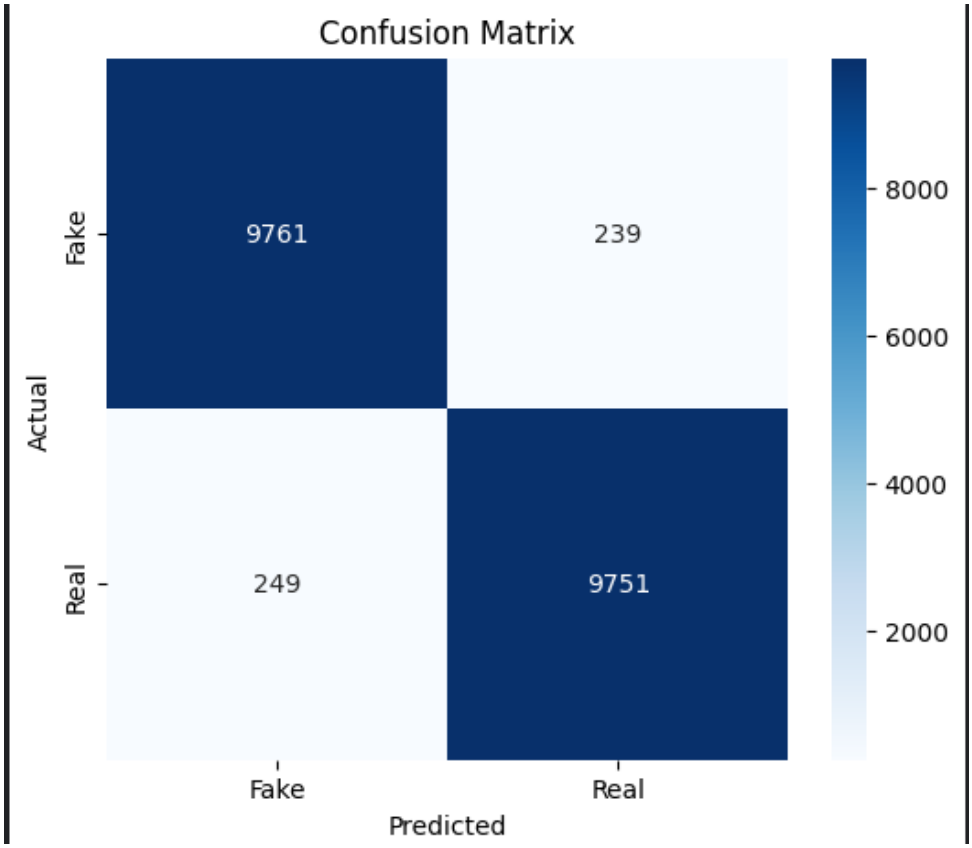


Figure 4.16 **Hybrid Model** Confusion Matrix

4.3 Discussion

The results show our hybrid model is performing very well to detect fake images or binary classification. It achieved 0.98 accuracy after 20 epochs. The hybrid model has proved that extracting features from multiple models and combining them is actually better. It is reliable.

Table 4.2 provides a comparative summary of all models' performance metrics, highlighting the superiority of newer CNN designs over legacy models like VGG16. It also reveals the limitations of EfficientNetV2B0 in this context, suggesting that architecture suitability depends heavily on the data characteristics and task specificity. The balance between accuracy and computational efficiency is thus a key consideration for model selection.

Table 4.3 Comparison of different study methodology and findings

Model	**Epochs**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
MobileNetV2	20	0.98	0.98	0.98	0.98
EfficientNetV2B0	20	0.90	0.90	0.90	0.90
InceptionV3	20	0.98	0.98	0.98	0.98
ResNet50	20	0.98	0.98	0.98	0.98
VGG16	20	0.97	0.97	0.97	0.97
Hybrid Model	20	0.98	0.98	0.98	0.98

While the results are promising, several limitations must be acknowledged. The CIFAKE dataset, although balanced and well-curated, represents a specific subset of AI-generated images derived from Stable Diffusion v1.4. Real-world fake image distributions are likely to be more diverse, including manipulations from different generative models or domains, which could challenge model generalization.

Additionally, despite the hybrid model's improved performance, its increased architectural complexity may require more computational resources than single lightweight models, potentially limiting its deployment in highly resource-constrained environments. Future work could explore model compression techniques such as pruning, quantization, or knowledge distillation to alleviate these constraints.

Furthermore, explainability and interpretability of these CNN models remain open research questions. Understanding which image features contribute most to classification decisions would be valuable for trust and accountability, especially in forensic and legal contexts.

To build on this work, future research might include:

- Though we use a custom dataset, we need to use a larger custom dataset to generalize and real life edits.
- We need to apply compression and acceleration for faster execution and edge friendly models.
- We need to build explainability tools to show how models make decesion.

CHAPTER 5

CONCLUSION

5.1 Conclusion

This study shows that hybrid models, when well designed, trained, and tested, provide better accuracy rate than single models in the area of fake image detection. And after using the hybrid model we can achieve a better recall rate than before, which will be capable of gaining significant trust of people. We also created a custom dataset and tested our hybrid model on it. And it has the best results, which is pretty good.

Key insights derived from this work include the following:

- **Impact of Training Duration:** After using several epochs, I conclude that 20 epochs is the best to be considered ideal for most of the models. After 20 epochs a model can learn enough and train well to detect. It helps avoid overfitting. And 20 epochs are enough for a model to produce the best result.
- **Advantages of Hybrid Modeling:** We have created a hybrid model by combining the strengths of two models, MobileNetV2 and RaceNet50. So, by concatenating the features of the two models, the detection generalization also gave very good results. Recall and accuracy in particular have increased significantly. Also increase robustness.
- **Reliability of the CIFAKE Dataset:** The CIFAKE dataset can be considered a balanced dataset for real-time image detection and binary result classification. It provides a nice controlled environment for model trains. This dataset can be considered ideal for model train and test.

Overall, this research supports the answer that using a hybrid model provides better accuracy than using a single model. This provides a good recall rate which can strengthen reliability. In addition, a custom dataset was created and tested in this research. And it was found that the hybrid model gave the best results in terms of accuracy, precision, recall and F1 score. We can say that the hybrid model is capable of providing good results even on unseen datasets because it takes features from multiple models.

5.2 Limitations

Although the results of our research are promising, it is essential to provide some guidelines for further improvement in the future :

- **Dataset Scope and Diversity:** Here in this research for training relying on CIFAKE dataset. This could be a trust issue. CIFAKE collected fake images from Stable Diffusion v1.4 and real images from CIFAR-10. But there is no cover for DeepFakes, different GANs, or advanced edited images. Though we test on Custom dataset after training on Cifake dataset. I think if we train based on a diverse dataset, the result will be better.
- **Binary Classification Framework:** This research is only for image classification. That is, it will only predict the real or fake image. But in real life, some more detailed research may be needed, and that's normal. For example, let's find the difference between a GAN-created image and a Photoshop-created image. And present them beautifully. And describe why and how the difference was found.
- **Computational Resource Requirements:** We can see that for using hybrid models, it requires more computational cost. Here we relied on GPU acceleration (e.g., NVIDIA Tesla P100 on Kaggle) for training and testing. So next work should be to make this model lightweight to deploy and use in mobile devices where it requires less computational cost.
- **Hybrid Model Complexity:** The hybrid model, while enhancing classification performance, introduced additional layers and processing steps, increasing computational overhead compared to individual lightweight models. This trade-off may impact its feasibility on extremely constrained platforms without further optimization.

5.3 Future Work

Building upon the foundation established in this study, several promising research directions can be pursued to advance the field of fake image detection:

- **Dataset Expansion and Diversification:** We can train the model on several larger datasets and some more custom datasets, that can increase robustness. This is a good way to test the accuracy of unseen datasets.

- **Multi-Class and Multi-Source Classification:** More analysis is needed beyond binary classification. This will allow us to distinguish between images created by different models and images created by Photoshop. These will provide greater acceptance. It is possible to expand our model in this way in the future.
- **Model Compression and Optimization:** We can use techniques like pruning, quantization, knowledge distillation. We can also use neural architecture search to make the hybrid model smaller and faster. This will give us the ability to use mobile devices without losing accuracy rate.
- **Explainable Artificial Intelligence (XAI):** We can add tools like Grad-CAM, LIME, or SHAP. They can help us to visualize and understand how the model makes decisions. It will increase trust in this study and forensics analysis.
- **Advanced Ensemble Techniques:** We need to explore more fusion strategies like late fusion, weighted ensemble voting, or stacking classifiers to enhance detection accuracy.
- **Real-Time Application Development:** We need to take the next step to make this model lightweight. We need to achieve the ability to integrate the model in web applications or mobile apps. This will allow people to use our system on their own.
- **Robustness Against Adversarial Attacks:** In the future, we need to make the model more sophisticated by thinking about what other features image generators can bring to fool our system.

5.4 Summary of the Study

The main objective of our research was to accurately predict the image gap between real visas and provide better results. And also increase the recall rate so that it can detect images in the maximum number of gaps. And gain trust. Creating a custom dataset and testing on it yielded good results. This checked the robustness of the hybrid model.

First, we start with the Cifake dataset and use a single model to train it. I use pre-processing and augmentation during training. I had to use various apps like 20/30 while training. Five CNNs were tested (MobileNetV2, InceptionV3, EfficientNetV2B0, ResNet50, and VGG16). Then I create a hybrid model using the best 2 models found here.From the result we can see that MobileNetV2 and ResNet50 worked best and achieved the best result. They achieved best accuracy, recall, precision and F1 score. Combine features from 2 models to make a model train. And test the test data on this hybrid model.

The study shows that hybrid models combined from the best two models and can provide the most accurate results like accuracy, precision, recall and f1 score. The model is built from joining features layers from two best models. This approach can boost results from fake image detection and classification.

REFERENCES

Fatoni, F., Kurniawan, T. B., Dewi, D. A., Zakaria, M. Z., & Muhayeddin, A. M. M. (2025). Fake vs Real Image Detection Using Deep Learning Algorithm. *Journal of Applied Data Sciences*, *6*(1), 366-376.

Islam, M. T., Lee, I. H., Alzahrani, A. I., & Muhammad, K. (2025). MEXFIC: A meta ensemble eXplainable approach for AI-synthesized fake image classification. *Alexandria Engineering Journal*, *116*, 351-363.

Raza, S. A., Habib, U., Usman, M., Cheema, A. A., & Khan, M. S. (2024). MMGANGuard: A Robust Approach for Detecting Fake Images Generated by GANs using Multi-Model Techniques. *IEEE Access*.

Alrusaini, O. A. (2024). Sustainable Artificial Intelligence: Assessing Performance in Detecting Fake Images. *International Journal of Advanced Computer Science & Applications*, *15*(4).

Namani, Y., Reghioua, I., Bendiab, G., Labiod, M. A., & Shiaeles, S. (2025). DeepGuard: Identification and Attribution of AI-Generated Synthetic Images. *Electronics*, *14*(4), 665.

Diao, Y., Zhai, N., Miao, C., Yang, X., & Wang, M. (2024). Vulnerabilities in AI-generated Image Detection: The Challenge of Adversarial Attacks. *arXiv preprint arXiv:2407.20836*.

Petrzelkova, N., & Cech, J. (2024). Detection of Synthetic Face Images: Accuracy, Robustness, Generalization. *arXiv preprint arXiv:2406.17547*.

Chen, J., Yao, J., & Niu, L. (2024). A single simple patch is all you need for ai-generated image detection. *arXiv preprint arXiv:2402.01123*.

Azzeh, M., Qusef, A., & Alabboushi, O. (2025). Arabic fake news detection in social media context using word embeddings and pre-trained transformers. *Arabian Journal for Science and Engineering*, *50*(2), 923-936.

Prachi, N. N., Habibullah, M., Rafi, M. E. H., Alam, E., & Khan, R. (2022). Detection of fake news using machine learning and natural language processing algorithms. *Journal of Advances in Information Technology*, *13*(6).

Dhiman, P., Kaur, A., Gupta, D., Juneja, S., Nauman, A., & Muhammad, G. (2024). GBERT: A hybrid deep learning model based on GPT-BERT for fake news detection. *Heliyon*, *10*(16).

Yan, F., Zhang, M., Wei, B., Ren, K., & Jiang, W. (2024). FMC: Multimodal fake news detection based on multi-granularity feature fusion and contrastive learning. *Alexandria Engineering Journal*, *109*, 376-393.

Cavus, N., Goksu, M., & Oktekin, B. (). Real-time fake news detection in online social networks: FANDC Cloud-2024based system. *Scientific Reports*, *14*(1), 25954.

Sharma, D. K., & Garg, S. (2023). IFND: a benchmark dataset for fake news detection. *Complex & intelligent systems*, *9*(3), 2843-2863.

Villela, H. F., Corrêa, F., Ribeiro, J. S. D. A. N., Rabelo, A., & Carvalho, D. B. F. (2023). Fake news detection: a systematic literature review of machine learning algorithms and datasets. *Journal on Interactive Systems*, *14*(1), 47-58.

Shu, K., Wang, S., & Liu, H. (2019, January). Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining* (pp. 312-320).

Rahman, M. R., Karim, R., Arefin, M. S., Dhar, P. K., Hossain, G., & Shimamura, T. (2025). Facilitating automated fact-checking: a machine learning based weighted ensemble technique for claim detection. *Discover Applied Sciences*, *7*(1), 73.

Raza, S., Paulen-Patterson, D., & Ding, C. (2025). Fake news detection: comparative evaluation of BERT-like models and large language models with generative AI-annotated data. *Knowledge and Information Systems*, *67*(4), 3267-3292.

Cui, W., & Shang, M. (2025). MIGCL: Fake news detection with multimodal interaction and graph contrastive learning networks. *Applied Intelligence*, *55*(1), 78.

Li, Y., Liu, X., Wang, X., Lee, B. S., Wang, S., Rocha, A., & Lin, W. (2024). Fakebench: Probing explainable fake image detection via large multimodal models. *arXiv preprint arXiv:2404.13306*.

ACCOUNTS CLEARANCE