Deepfake Image Detection Using Vision Transformers

Vasanthkumar Ramamoorthi - 2171863
vramamoo@depaul.edu
Arul Michael Antony Felix Raja - 2133065
afelixra@depaul.edu
Praveen Kumar Jayakumar - 2131794
pjayakum@depaul.edu

June 13, 2025

Abstract

The rapid advancement of deepfake technology has raised serious concerns about the credibility of visual content, making it harder than ever to tell apart real images from manipulated ones. In response to this challenge, the DeepfakeDetect project focuses on building a reliable image classification system powered by Vision Transformers (ViT) to effectively distinguish authentic images from deepfakes. Using cutting-edge deep learning and image processing methods, this system aims to improve both detection accuracy and generalization. The approach involves training a ViT model on a carefully curated and balanced dataset of real and fake images, with techniques like data augmentation and oversampling employed to manage class imbalance. The resulting model achieved a remarkable accuracy of 99.35.

1 Introduction

Deepfake technology has evolved rapidly, enabling the creation of highly convincing fake images and videos. This development poses serious threats to the integrity of digital media, as it's becoming more difficult to tell real content from fabricated ones [Mirsky and Lee, 2021]. Detecting deepfakes accurately is essential to maintaining trust across sectors such as journalism, cybersecurity, and social media platforms.

To address this growing concern, there's a clear need for a robust and efficient system capable of identifying deepfake images. Traditional detection techniques often fall short, especially when it comes to subtle manipulations, and they typically rely on extensive manual feature extraction. In contrast, Vision Transformers (ViT) have shown strong capabilities in learning complex visual patterns [Khan et al., 2022], making them an ideal solution for this challenge.

The key difficulty lies in developing a model that not only achieves high accuracy but also generalizes effectively across different types of deepfake images. This report presents the goals, approach, and outcomes of the DeepfakeDetect project, which harnesses the power of Vision Transformers to improve the detection and classification of deepfake content.

2 Related Works

Deepfake detection continues to be a highly active research field, with many methods traditionally relying on convolutional neural networks (CNNs) [Sun et al., 2022; Wang et al., 2022]. While CNN-based techniques have delivered solid results, they often involve complex architectures and require extensive datasets to perform effectively.

In recent years, transformer-based models particularly Vision Transformers (ViT) have emerged as powerful alternatives in image classification, thanks to their ability to capture long-range dependencies and fine-grained visual details [Khan et al., 2022].

Several studies have explored the use of ViTs for deepfake detection with encouraging outcomes. For example, Zhao et al. [2022] introduced a multi-scale transformer architecture specifically designed to detect deepfakes. Their method leveraged ViTs to identify subtle visual inconsistencies, achieving state-of-the-art results and outperforming many CNN-based approaches.

Building on these advancements, this project adopts a Vision Transformer-based model specifically tailored for deepfake image classification. The goal is to further improve detection accuracy and ensure strong generalization across different types of manipulated content.

3 Preliminary/Background

3.1 Deepfake Technology

Deepfakes are artificially generated media in which a person's face or likeness is seamlessly swapped into existing images or videos. These convincing manipulations are made possible by deep learning, especially through the use of Generative Adversarial Networks (GANs). GANs operate through a pair of neural networks—a generator and a discriminator—that are trained together in a competitive setup. The generator is responsible for creating synthetic content, while the discriminator tries to distinguish between real and fake inputs. Over time, this adversarial process leads to the creation of highly realistic deepfakes that are increasingly difficult to detect.

3.2 Vision Transformers (ViT)

Vision Transformers (ViT) are a type of deep learning model that adapt the transformer architecture—originally developed for natural language processing—to work with image data. Instead of analyzing entire images directly, ViTs break an image down into fixed-size patches, treating each patch as a token, similar to how words are treated in text processing. This allows the model to use self-attention mechanisms to learn long-range dependencies and contextual relationships across the image.

The image is processed in the following pipeline:
Image → Patch Embedding → Transformer Encoder → Classification Head.

Here's how it works: the image is divided into N patches, each of size P × P. These patches are then flattened and mapped into a higher-dimensional space to create patch embeddings. To preserve the spatial structure, positional embeddings are added. The resulting sequence of embeddings is then passed through multiple layers of a transformer encoder, which ultimately leads to the classification output.

3.3 Evaluation Metrics

To assess the performance of the deepfake detection model, the following metrics are employed:
Final Project Report Deepfake Detection Using Vision Transformers

Accuracy: The proportion of correctly classified images.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

F1 Score: The harmonic means of precision and recall.
F1 Score = 2 × (Precision × Recall / Precision + Recall)

ROC AUC Score: Measures the ability of the model to distinguish between classes.

Confusion Matrix: A matrix that visualizes the performance of the model in terms of true positives, true negatives, false positives, and false negatives.
These metrics provide a comprehensive evaluation of the model's classification performance, especially in handling imbalanced datasets.

4 The Methodology

The DeepfakeDetect system utilizes Vision Transformers for image classification through a well-structured pipeline that includes data preprocessing, model setup, and strategic training techniques.

4.1 Data Collection and Preprocessing

A balanced dataset of real and deepfake images was compiled from publicly available sources, including the Celeb-DF (v2) dataset [Li et al., 2020] and recent contributions on Kaggle [Kaggle, 2023]. The final dataset contains approximately 190,402 images, evenly distributed between real and fake classes.

To handle any class imbalance, Random OverSampling was applied [Buda et al., 2018]. To further improve the model's robustness and generalization, various data augmentation techniques were used, such as random rotations, sharpness enhancement, and horizontal flipping [Shorten and Khoshgoftaar, 2021].

The preprocessing workflow can be summarized as:

$$\begin{aligned}
\textbf{Original Image} &\rightarrow \text{Resize to } 224 \times 224 && (4)\\
&\rightarrow \text{Normalize pixel values} && (5)\\
&\rightarrow \text{Apply data augmentation} && (6)\\
&\rightarrow \text{Oversample minority class} && (7)
\end{aligned}$$

4.2 Model Architecture

A pre-trained Vision Transformer (ViT) model [Khan et al., 2022] is fine-tuned using the curated dataset. The ViT architecture processes input images by dividing them into fixed-size patches, embedding those patches, and then feeding them through multiple transformer layers to capture complex visual features and relationships. For this project, the model is specifically configured for a binary classification task—distinguishing between real and deepfake images.

The high-level architecture of the ViT can be represented as:

Mathematically, the ViT architecture can be described as:

$$\text{Input Image} \xrightarrow{\text{Patch Embedding}} \text{Transformer Encoder} \xrightarrow{\text{Classifcation Head}} \text{Output Probability} \qquad (8)$$

4.3 Training Strategy

The model is trained using the cross-entropy loss function, which measures the difference between the true labels and the predicted probabilities. It is defined as:

The model is trained using cross-entropy loss, defined as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \qquad (9)$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability for each sample i.

For optimization, the model uses the AdamW optimizer [Loshchilov and Hutter, 2019], which introduces weight decay to reduce overfitting. The parameter update rule is expressed as:

$$\theta = \theta - \eta(\partial L/\partial \theta) + \lambda\theta \qquad (10)$$

Here, θ represents the model parameters, η is the learning rate, and λ is the weight decay coefficient.

Key hyperparameters are carefully tuned based on validation performance:
Learning rate: $\eta = 5 \times 10^{-6}$
Batch size: 32 for training, 8 for validation
Epochs: 10

Additionally, early stopping is employed to prevent overfitting by halting training once the validation loss stops improving.

4.4 System Design

The system follows a modular architecture, consisting of distinct components for data management, model training, and prediction. This design promotes scalability and maintainability, enabling updates or modifications to individual modules without disrupting the entire system. A visual representation of the architecture is provided in Appendix??.

5 Numerical Experiments

5.1 Experimental Setup

The experiments were carried out on a dataset of 190,402 images, evenly divided between real and deepfake categories. The dataset was split into training (133,281 images), validation (28,561 images), and test (28,560 images) sets. The Vision Transformer model was fine-tuned using the defined hyperparameters over 10 epochs, with training accelerated by an NVIDIA RTX 3060 GPU to enhance computational efficiency.

5.2 Results

The ViT-based deepfake detection model demonstrated strong performance on the test set, achieving the following metrics:

Accuracy: ***99.35%***
F1 Score: ***99.30%***
ROC AUC Score: ***99.40%***

The confusion matrix indicated very few misclassifications, with only a small number of false positives and false negatives. This reflects the model's high precision and recall, confirming its effectiveness in accurately distinguishing between real and deepfake images.

5.3 Performance Analysis

The model’s high accuracy and F1 score highlight the effectiveness of Vision Transformers in capturing the subtle patterns and inconsistencies typical of deepfake images. Its strong performance across various image qualities and characteristics demonstrates a high degree of robustness and generalization. Additionally, the ROC AUC score reinforces the model’s excellent discriminatory capability, effectively distinguishing between real and fake content across varying classification thresholds.

5.4 Sample Predictions

Figure 1 presents sample predictions generated by the model, demonstrating its ability to accurately classify both real and deepfake images with high confidence scores.

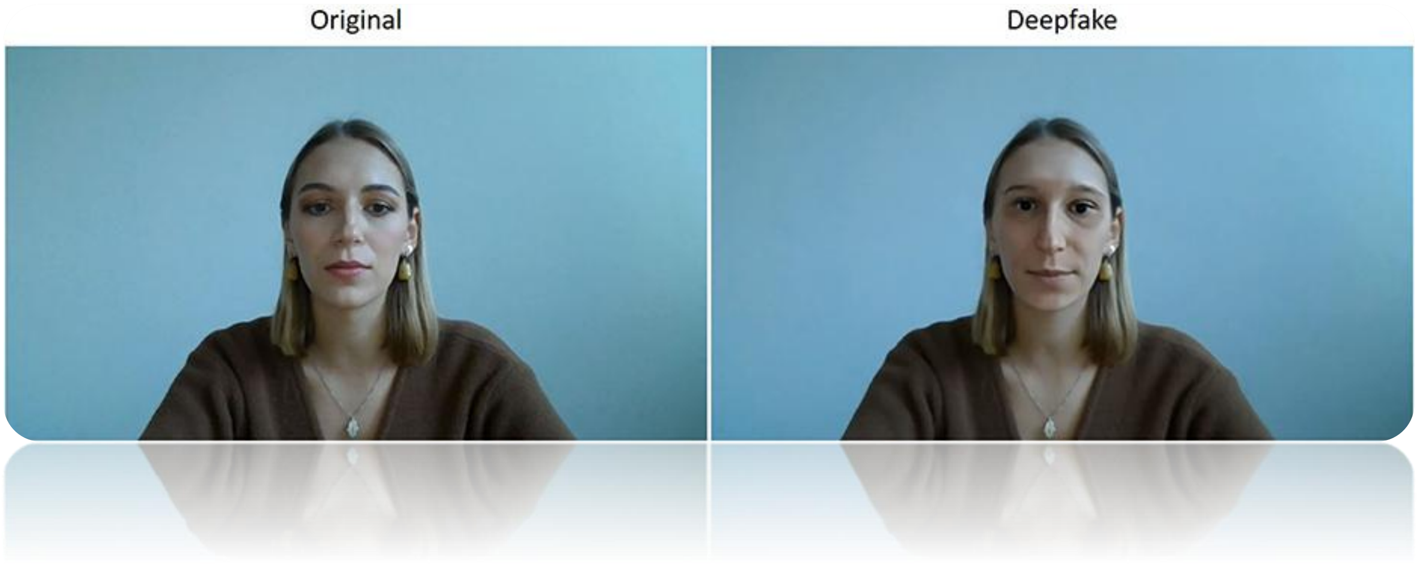

Figure 1: Sample Predictions – Left: Real image (Prediction: Real, Confidence Score: 0.9987); Right: Deepfake image (Prediction: Fake, Confidence Score: 0.9975). These examples highlight the model’s high accuracy and confidence in distinguishing real and manipulated content.

6 Conclusion

The DeepfakeDetect project successfully developed a Vision Transformer (ViT)-based deepfake detection system that delivers high accuracy and robust performance in differentiating real images from deepfakes. The model’s outstanding evaluation metrics—99.35% accuracy, 99.30% F1 score, and 99.40% ROC AUC score—underscore the effectiveness of transformer-based architectures in advancing the field of digital media authenticity.

References

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249–259.

Kaggle. (2023). Deepfake Detection Challenge Dataset. Retrieved from https://www.kaggle.com/competitions/deepfake-detection-challenge/data

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. ACM Computing Surveys, 54(10), 1–41.

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3207–3216).

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. ACM Computing Surveys, 54(1), 1–41.

Shorten, C., & Khoshgoftaar, T. M. (2021). Image data augmentation for deep learning: A survey. Journal of Big Data, 8(1), 1–48.

Sun, L., Ma, Y., Liu, Y., Ding, Y., & Wang, X. (2022). Domain generalization for deepfake detection: A survey. arXiv preprint arXiv:2203.08807.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2022). GAN-generated faces detection: A survey and new perspectives. arXiv preprint arXiv:2205.06903.

Zhao, J., Wang, X., Han, X., Xu, X., & Wang, Y. (2022). A multi-scale transformer network for deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1215–1224).

A Appendix

A.1 Model Architecture Diagram: Vision Transformer (ViT)

A.2 Sequence Diagram: Sequence of Operations in the System

A.3 Data Metrics: Confusion matrix

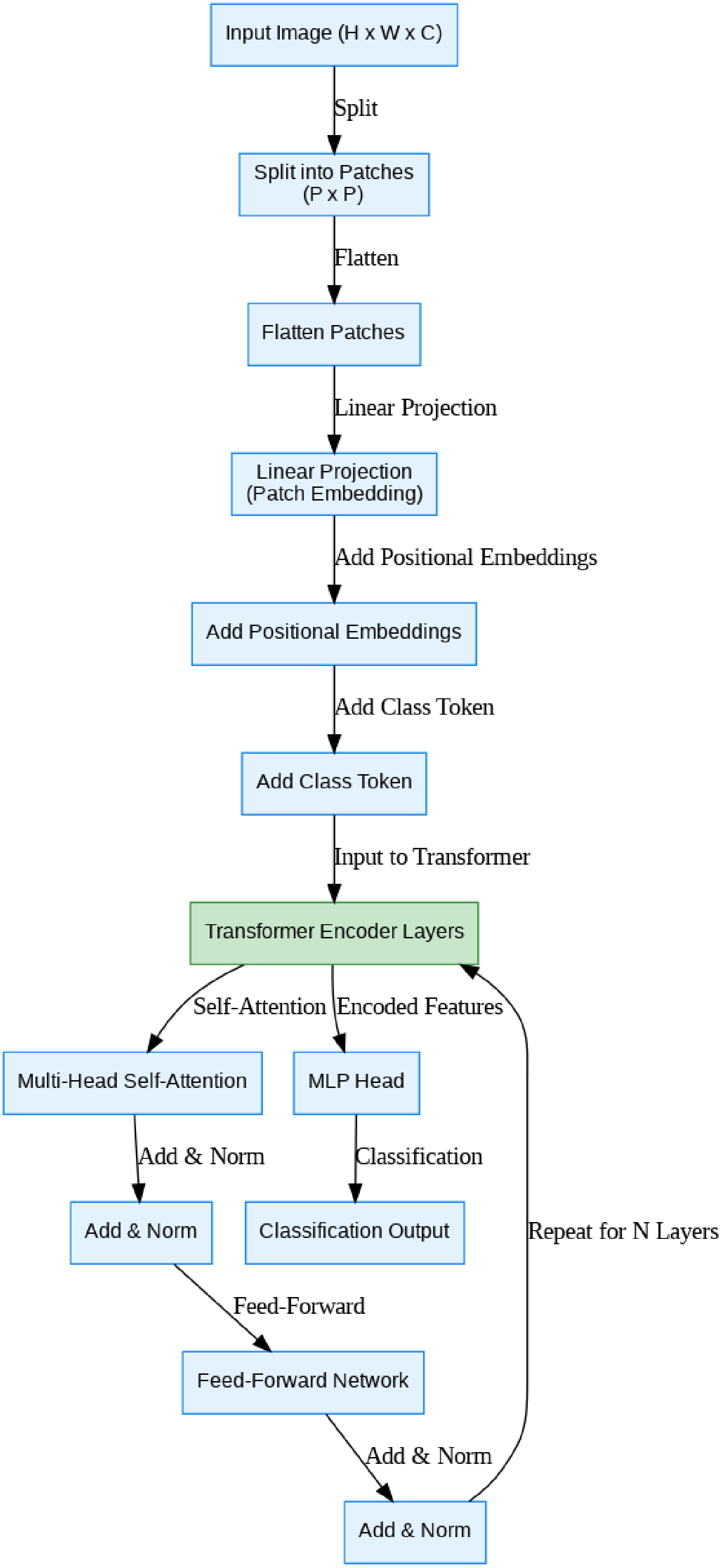

Figure 2: Vision Transformer (ViT) Architecture Diagram

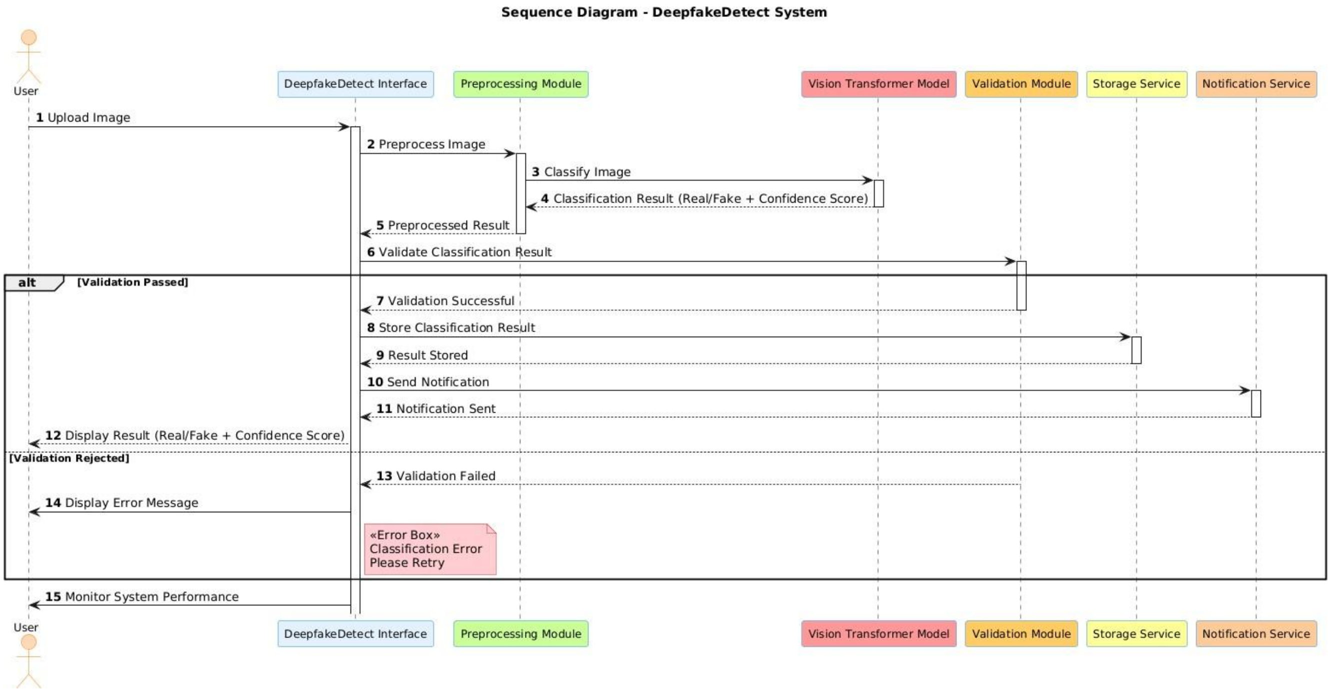

Figure 3: Sequence of Operations in the System

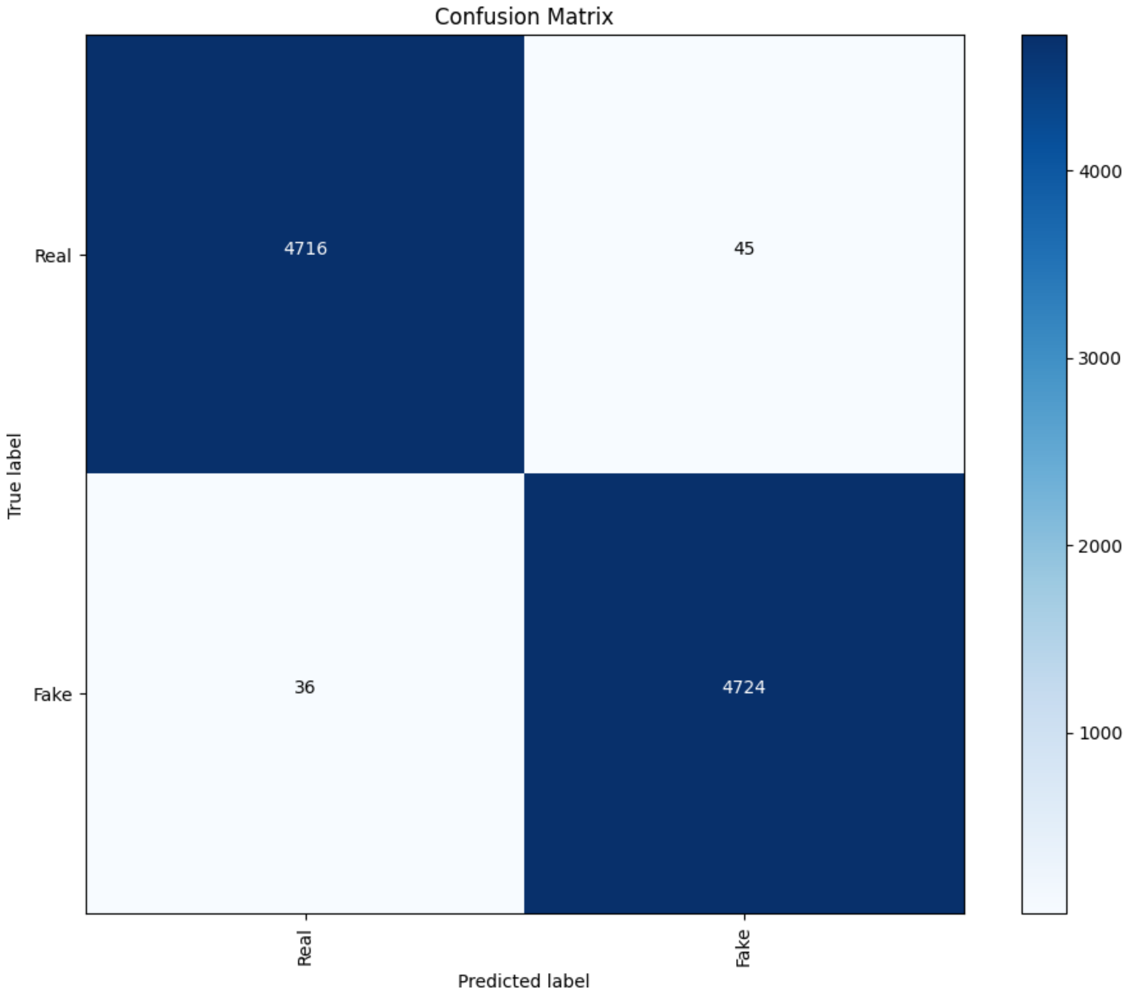

Figure 4: Confusion Matrix