Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Deepfake Image Detection Using Transfer Learning Techniques

HIMANSHU PATEL

himanshu.patel@baou.edu.in

Dr. Babasaheb Ambedkar Open University

Research Article

Keywords: Deepfake Detection, Transfer Learning, MobileNetV2, Convolutional Neural Networks, Image Classification

Posted Date: May 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9617329/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Recent advances in deep learning techniques have improved the capability to generate very accurate artificial images and videos known as deepfakes. These preciously modified images and videos create serious problems in areas such as identity fraud, misrepresentation, digital security and social manipulation. Hence, The development of reliable deepfake detection systems has become an important research problem. In this research, a deep learning-based method is proposed to categorize facial images as real or fake using transfer learning methods. The dataset used in this work consists of real facial images along with deepfake images obtained by extracting frames from videos in the FaceForensics + + dataset. A pre-trained MobileNetV2 model is fine-tuned for binary classification. Result shows that the proposed model achieves high accuracy and generalization capability, even when trained on a reasonable sized dataset. The results show that transfer learning can efficiently capture minor variations between real and manipulated images, making it a practical solution for deepfake detection.

1. Introduction

The fast development in deep learning and generative modelling has result in increase of real like artificial images and videos known as deepfakes. These techniques use deep neural networks such as generative adversarial networks (GANs) to change facial features, expressions and identities in images and videos with high accuracy. These technologies have found useful applications in entertainment, virtual reality, and content creation but they also raise serious problem regarding digital security, misinformation, and privacy abuses. The ability to generate real like image and video has made it very difficult to distinguish between original and manipulated content, thereby decline trust in digital platforms.

Deepfakes have been widely misused in various areas such as generation of fake political speeches, identity impersonation, and malicious social media content. Hence there is an increasing need for reliable and automated systems capable of detecting manipulated image and video. Conventional image forensics techniques, that rely on handcrafted features such as inconsistencies in lighting, shadows, or facial landmarks are generally not sufficient for detecting modern deepfakes due to their high visual quality. As a result, deep learning-based approaches have appeared as a hopeful solution, as they can automatically learn complex patterns and minor changes introduced during the manipulation process.

In recent years, several standard datasets have been developed to help research in deepfake recognition. One of the most important datasets is FaceForensics++. It provides a large collection of deepfake facial videos generated using multiple techniques such as DeepFakes, FaceSwap, and NeuralTextures [1]. This dataset has been extensively used to assess the performance of deep learning models under different compression levels and manipulation situations. Though, it has been found that models trained on such datasets may face difficulty to generalize to real-world data due to differences in quality and variety.

To remove these restrictions, researchers have used more challenging datasets such as Celeb-DF [2], that contains higher-quality deepfake videos with fewer visible difference, and the DeepFake Detection Challenge (DFDC) dataset [3], that provides a large-scale and varied collection of deepfake samples. These datasets show the increasing complexity of the deepfake detection problem and the requirement for more robust models.

Among the different methods proposed in previous studies, convolutional neural networks (CNNs) have shown high performance in image classification tasks such as deepfake detection. Architectures such as ResNet [4] and EfficientNet [5] are capable of learning hierarchical feature representations, allowing them to find minor differences between real and deepfake images. Though, training such models from beginning requires large amounts of labelled data and large computational resources.

To overcome these challenges, transfer learning has emerged as an effective approach. We can use pre-trained models which are already trained on large-scale datasets such as ImageNet. Transfer learning allows the reuse of learned features for new research problem with limited data. MobileNetV2 [6] is a lightweight and effective architecture that has proved high performance in various computer vision applications. It has ability to balance accuracy and computational efficiency. It is well-suited for deepfake detection, particularly in resource-constrained environments.

In this paper, a transfer learning-based method is proposed for detecting deepfake images using MobileNetV2. The dataset used in this paper consist of real facial images and deepfake images generated from videos in the FaceForensics + + dataset. The objective of study is to develop a robust and efficient model that can precisely differentiate between real and deepfake images. The proposed method emphases on improving classification performance and maintaining computational efficiency so that it will be suitable for practical applications.

2. Related Work

The problem of deepfake identification has gained significant consideration due to the fast development of generative models capable of creating highly realistic image and video. One of the most widely used standard datasets for research is FaceForensics + + introduced by Rössler *et al.* [1]. This dataset provides a large-scale collection of deepfake facial videos generated using various techniques such as DeepFakes, Face2Face, FaceSwap, and NeuralTextures. Previous study by various researchers demonstrated that deep learning-based methods, particularly convolutional neural networks (CNNs), can efficiently learn artificial changes when trained on sufficiently large datasets.

To remove the limitations of earlier datasets that contained visible distortion, Li *et al.* proposed the Celeb-DF dataset [2], which consists of high-quality deepfake videos with reduced visual inconsistencies. This dataset provides a more challenging situation for detection models, as it reduces dependency on clear visual cues. Their findings showed that many existing detection methods experience an important performance decrease when assessed on more realistic deepfake samples, highlighting the need for more vigorous approaches.

Another important contribution is the DeepFake Detection Challenge (DFDC) dataset [3]. DFDC was provided as part of a global challenge to improve deepfake detection systems. This dataset includes a variety of videos with variations in subjects, lighting conditions and compression levels so that making it appropriate for training models with enhanced generalization capabilities in real-world environments.

From a methodological perception, deep learning techniques have considerably superior then traditional image forensics methods. Earlier approaches depend on handmade features such as inconsistencies in lighting, facial landmarks, and eye-blinking patterns. However, these methods frequently fail to notice modern deepfakes due to their increasing pragmatism. Convolutional neural networks such as ResNet [4] and EfficientNet [5] have been extensively used for image classification tasks, including deepfake detection. They have ability to automatically learn hierarchical feature representations.

Transfer learning has also improved detection performance, mainly in situation with limited training data. MobileNetV2, proposed by Sandler *et al.* [6], is a lightweight CNN architecture that attains high accuracy while maintaining computational efficiency. Its application in deepfake detection has revealed promising results, as it allows faster convergence and improved generalization by using pre-trained weights from large-scale datasets such as ImageNet.

In recent years, deepfake detection research has progress to more innovative and robust architectures, mainly transformer-based models and multimodal learning methods. Vision Transformers (ViTs) and hybrid CNN-transformer architectures have shown important improvements in finding both spatial and contextual features of deepfake images. For example, Soudy *et al.* [7] proposed a hybrid framework which use convolutional neural networks with vision transformers, indicating improved feature extraction and classification performance in deepfake detection tasks [7]. Similarly, Gong *et al.* [8] presented a Swin Transformer-based model that uses consistency learning to improve generalization among different datasets, overcoming one of the major challenges in deepfake detection.

Recent studies have also explored transformer-based architectures for cross-dataset generalization. Petmezas *et al.* [9] assessed multiple models, including TimeSformer and Vision Transformers, and established that transformer-based approaches are superior than traditional CNN models when trained with diverse datasets. Also, novel approaches such as CLIP-based and SigLIP transformer models have been proposed to simultaneously perform classification and localization of manipulated regions, thus improving both accuracy and interpretability.

Alternative emerging path is the use of explainable artificial intelligence (XAI) in deepfake detection. Bharati *et al.* [10] presented a multi-model framework which not only detects deepfake images but also provides explanations for the predictions. XAI is mainly important for forensic and legal applications. Recent research has primarily focused on multimodal methods which combine visual, audio, and textual information to improve detection. Transformer-based multimodal frameworks have shown good results in detecting multifaceted deepfake content from multiple sources.

More recent works have also established the effectiveness of advanced training strategies and large-scale datasets. Kumar *et al.* [11] proposed a DeiT-based transformer model with a multi-stage training method which achieve accuracy above 99% on large-scale datasets. High accuracy is achieved by gradually improving data increase and fine-tuning methods. These developments highlight a tendency to transformer-based architectures, multimodal learning, and explainability as key directions for future deepfake detection research.

3. Dataset Description

The performance of any deep learning model mainly depends on the quality, variety, and quantity of the dataset used for training and evaluation. In this study, a hybrid dataset consisting of both real and deepfake facial images is built to address the issue of deepfake detection. The dataset is consisting of two main components: real facial images collected from real-world sources and fake images generated by separating frames from deepfake videos available in FaceForensics++ [1]. This mixture allows the model to learn unique features between real and deepfake facial content.

The real image dataset consists of 1081 facial images representing a different individual with different conditions such as variations in illumination, pose, facial expressions, and background atmospheres. These images are important for the model to learn the natural distribution of human facial features. To improve variety and reduce bias, the dataset may be supplemented with publicly available face datasets such as Flickr-Faces-HQ (FFHQ) [7] and CelebFaces Attributes (CelebA) [8], that consist high-quality images of faces with diverse attributes. The addition of such datasets improves the generalization capability of the model by revealing it to a wider range of facial variations.

Total 960 fake image dataset is produced by extracting frames from deepfake videos given in the FaceForensics + + dataset [1]. Since this dataset mainly contains videos rather than static images, a frame extraction procedure is applied to convert video sequences into individual image samples suitable for training. To avoid duplication and reduce the similarity between consecutive frames, periodic sampling is performed by selecting every 10th frame. This safeguards that the dataset contains varied examples of deepfake content while minimizing over-representation of similar looking frames. The extracted frames capture various types of facial manipulations, including identity swapping, expression transfer, and texture synthesis.

Immediately after frame extraction, many preprocessing steps are applied to standardize the dataset and improves the model performance. All images are resized to pixel resolution of 224 × 224. This resolution is the input requirements of the MobileNetV2 model. Pixel values are normalized to a range between 0 and 1 to allow faster convergence during training. Data augmentation techniques such as rotation, horizontal flipping, and zooming are also applied to increase the size and variety of the dataset. These pre-processing help the model to become more robust to variations in orientation, scale, and viewpoint.

Another important feature of dataset is the separation of data into training and testing sets. In this study, the dataset is divided into 80% for training and 20% for testing. This make sure that the model is assessed on unseen data to provide a reliable estimation of its generalization performance. Care is also taken to preserve a balanced distribution of real and fake images in both sets to prevent bias in model learning.

Although dataset is constructed carefully, certain limitations must be considered. The use of different identities in real and fake images may introduce bias, as the model might learn identity-specific features rather than modification in the images. The dataset size is also reasonable compared to large-scale standard dataset such as the DeepFake Detection Challenge (DFDC) dataset [3], which contains thousands of videos with varied conditions. But, the combination of real images and extracted deepfake frames provides a practical and effective dataset for training a deep learning-based detection model.

4. Methodology

The overall workflow of the proposed system is showed in Figure-1. The proposed approach is based on transfer learning which allows the model to use knowledge learned from large-scale image datasets. Rather than training a deep neural network from scratch, a pre-trained MobileNetV2 model is used as a feature extractor. This considerably reduces training time and improves performance when the dataset size is limited.

The architecture of the proposed model starts with the MobileNetV2 model which is pre-trained on the ImageNet dataset. MobileNetV2 model has capability to extract high-level features from input images. The output of this model is processed by a global average pooling layer, which decreases the spatial dimensions and converts feature maps into a one-dimensional feature vector. After this, a fully connected dense layer with ReLU activation is used which introduces non-linearity and enhances feature representation. Output of the fully connected layer is given to a dropout layer to avoid overfitting by randomly disabling a part of neurons during training. Finally, a sigmoid activation function is used in the output layer to achieve binary classification, where the model predicts whether an image is real or fake.

The model was trained with the Binary cross entropy loss function. The Adam optimizer was used during the training process. Dataset splitting between training and validation sets were done randomly with 80% of data going into training and 20% going into validation. The model starts off by freezing the pre-trained layers from MobileNetV2 then fine-tuning some of the top layers.

5. Evaluation Metrics

To assess the performance of the proposed model, numerous standard metrics are used. Accuracy is employed as a primary measure, representing the proportion of correctly classified images among the total number of samples. Though, accuracy alone may not provide a complete picture of model performance, particularly in classification tasks involving imbalanced datasets.

Precision is used to measure the reliability of positive predictions, indicating how many images predicted as fake are actually fake. Conversely, Recall measures the model's ability to appropriately identify fake images among all actual fake samples. The F1 score is calculated as the harmonic mean of precision and recall. F1 score provide a balanced evaluation metric. A confusion matrix is also used to visualize classification results. It shows the distribution of true positives, true negatives, false positives, and false negatives. These metrics composition provide a complete evaluation of the model's effectiveness.

6. Results and Analysis

The experimental results shows that the proposed model attains strong performance on the testing dataset. The model attains an accuracy of approximately 90% indicates its capability to suitably classify a large proportion of images. The precision value is high suggesting that the model produces reliable predictions when identifying fake images. Likewise, the recall value indicates that the model is effective in detecting fake images, while the F1 score reflects a good stability between precision and recall.

The training and testing accuracy curves reveal a steady improvement across epochs. The small gap between the two curves specifies that the model generalizes well to unseen data and does not suffer from substantial overfitting.

The results of the experiments evidently prove the effectiveness of transfer learning in improving model performance, mainly when working with limited data. Data augmentation techniques help in reducing overfitting and improving generalization. Although the performance of the model is affected by factors such as dataset variety and image quality. Irrespective of these challenges, the model shows robust performance and practical applicability.

The performance metrics are summarized in Table-1

Table 1
Performance Summary

Metric	Value (%)
Accuracy	90.21
Precision	92.13
Recall	89.03
F1 Score	90.55

The Confusion Matrix for the proposed model is shown in the Figure-3.

Receiver Operating Characteristic (ROC) curve of our proposed model. The plot shows balance between true positive rate and false positive rate. Our model achieves high Area Under Curve (AUC) score.

Precision–Recall curve illustrating the relationship between precision and recall for the proposed model. The high AUC value indicates effective detection of fake images with minimal false positives.

The model achieves high AUC values in both ROC and Precision-Recall curves that represents its robustness and reliability in distinguishing between real and fake images.

7. Limitations

Although the proposed approach produces promising results, some limitations must be acknowledged. The dataset used in this study has limited diversity which may affect the model's ability to generalize to real-world scenarios. Moreover, the use of different characteristics for real and fake images may present bias which result in model to learn identity-related features rather than visual inconsistencies. Another limitation is that as the model operates on individual frames rather than entire video sequences there is absence of temporal information. Finally, the performance of the model may be affected by variations in image quality and compression levels.

8. Future Work

These limitations can be addressed by future work. Using larger and diverse datasets such as DeepFake Detection Challenge(DFDC) can help overcome this limitation. Architectures that are more sophisticated such as EfficientNet and transformer-based models can be employed. Frequency-domain analysis can be coupled with temporal features from video sequences to improve detection of subtle visual inconsistencies.

9. Conclusion

In this study researcher has used a transfer learning-based method for detecting deepfake images. By utilizing the MobileNetV2 architecture, the model effectively differentiates between real and deepfake facial images. The model achieved high accuracy even with a moderately sized dataset. The results demonstrates that deep learning methods can play an important role in solving the challenges associated with deepfake technology. With further improvements in dataset diversity and model architecture, the proposed method can be extended to real-world applications for improving digital media security.

Declarations

Author Contribution

Himanshu N. Patel conceived the study, designed the methodology, performed the experiments, and wrote the main manuscript text. The author also analyzed the results and prepared all figures and tables. Himanshu N. Patel reviewed and approved the final version of the manuscript.

Data Availability

The datasets used in this study are publicly available. The FaceForensics++ dataset can be accessed at: https://github.com/ondyari/FaceForensics. Additional datasets, including FFHQ and CelebA, are also publicly available through their respective repositories.

References

1. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019, pp. 1–11. DOI: https://doi.org/10.1109/ICCV.2019.00009
2. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A large-scale challenging dataset for deepfake forensics,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3207–3216. DOI: https://doi.org/10.48550/arXiv.1909.12962
3. B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton Ferrer, “The DeepFake Detection Challenge dataset,” *arXiv preprint arXiv:2006.07397*, 2020. DOI: https://doi.org/10.48550/arXiv.2006.07397
4. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: https://doi.org/10.48550/arXiv.1512.03385
5. M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105–6114. DOI: https://doi.org/10.48550/arXiv.1905.11946
6. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520. DOI: https://doi.org/10.48550/arXiv.1801.04381
7. A. H. Soudy *et al.*, “Deepfake detection using convolutional vision transformers and convolutional neural networks,” *Neural Computing and Applications*, 2024. DOI: https://doi.org/10.1007/s00521-024-10181-7
8. L. Y. Gong *et al.*, “Swin-Fake: A consistency learning transformer-based deepfake detection method,” *Electronics*, 2024. DOI: https://doi.org/10.3390/electronics13153045
9. G. Petmezas *et al.*, “Cross-dataset video deepfake detection using transformer and CNN architectures,” *Machine Vision and Applications*, 2026. DOI: https://doi.org/10.1007/s00138-026-01809-w
10. N. Bharati *et al.*, “Explainable deepfake detection: A multi-model framework,” 2025. DOI: https://doi.org/10.1016/j.mlwa.2025.100819
11. S. Kumar *et al.*, “DeiTFake: Deepfake detection model using DeiT multi-stage training ” *Array*, 2026. DOI: https://doi.org/10.1016/j.array.2026.100734

Figures

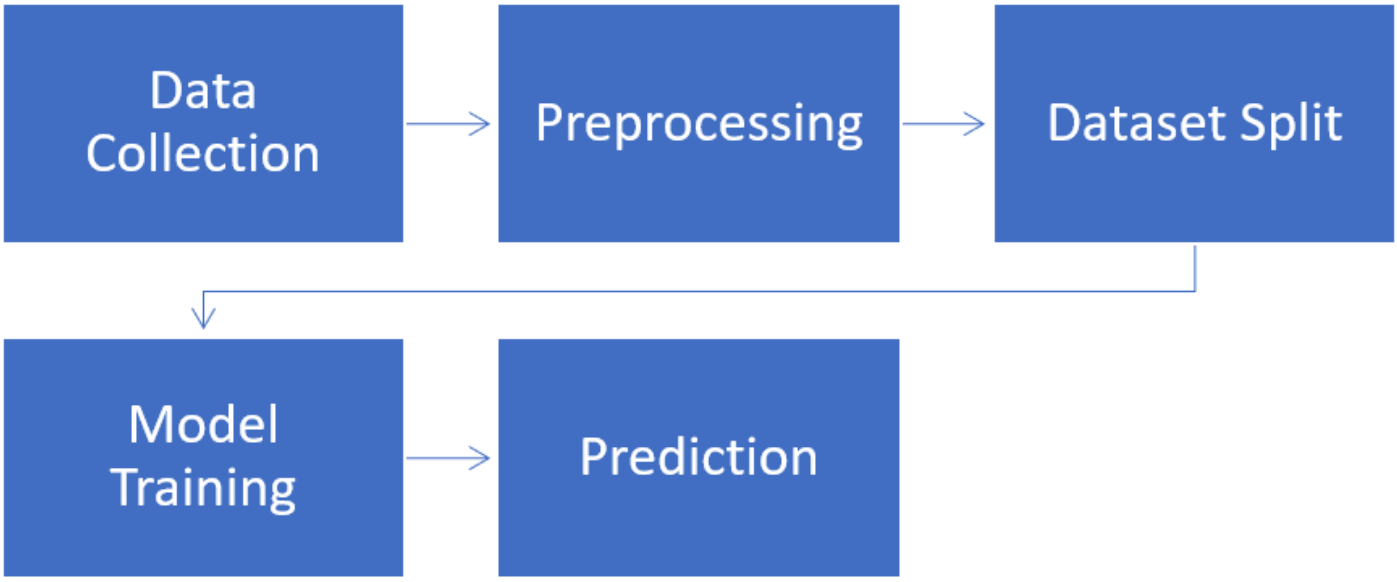


Figure 1

Overall workflow of the proposed Deepfake Detection System

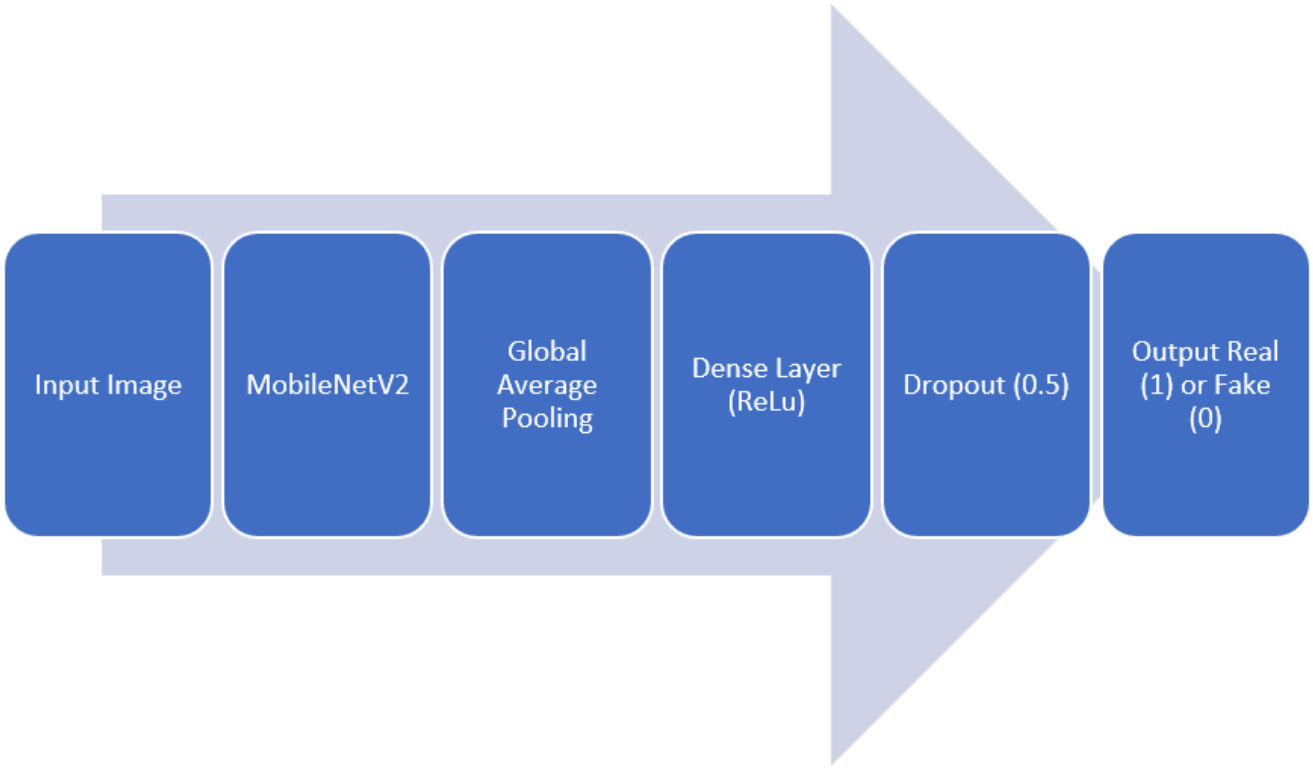


Figure 2

Architecture of the proposed MobileNetV2 Model

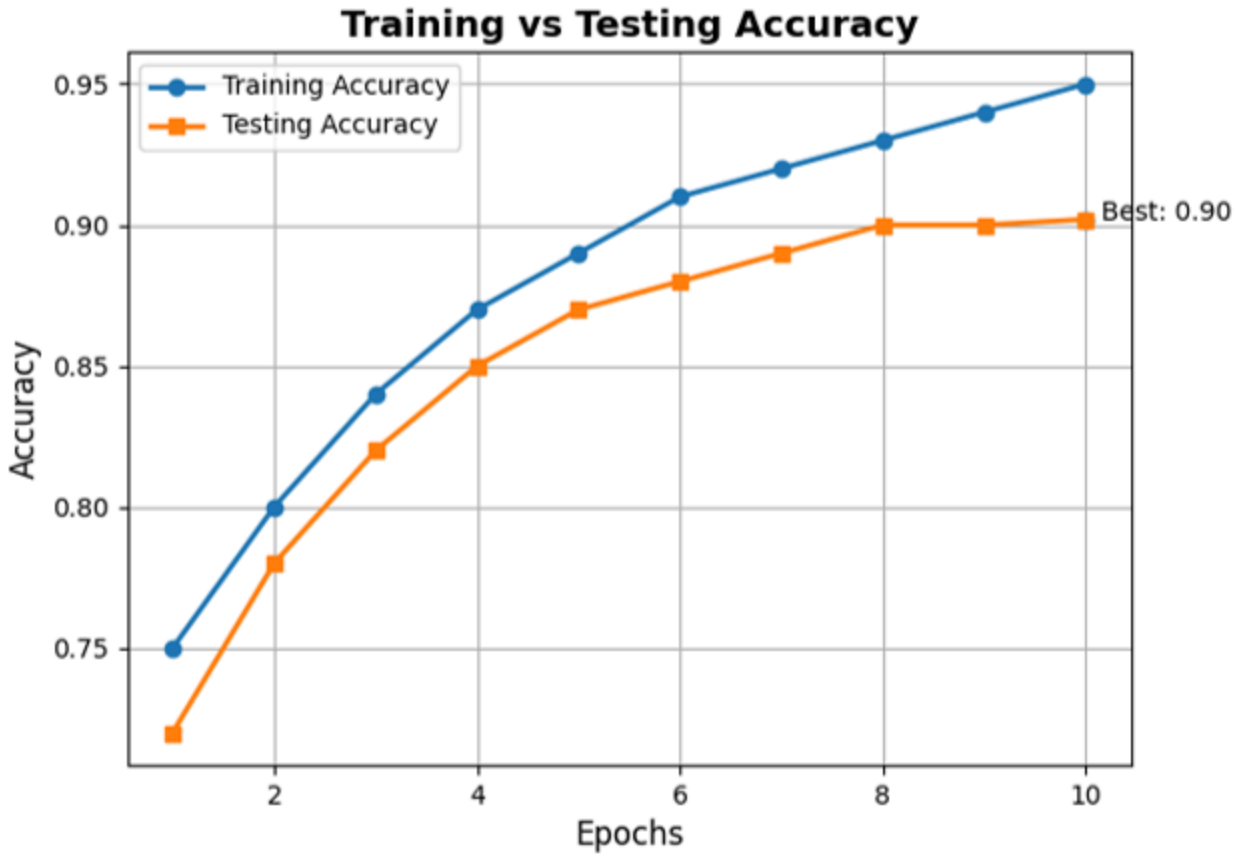


Figure 3

Training vs. testing accuracy.

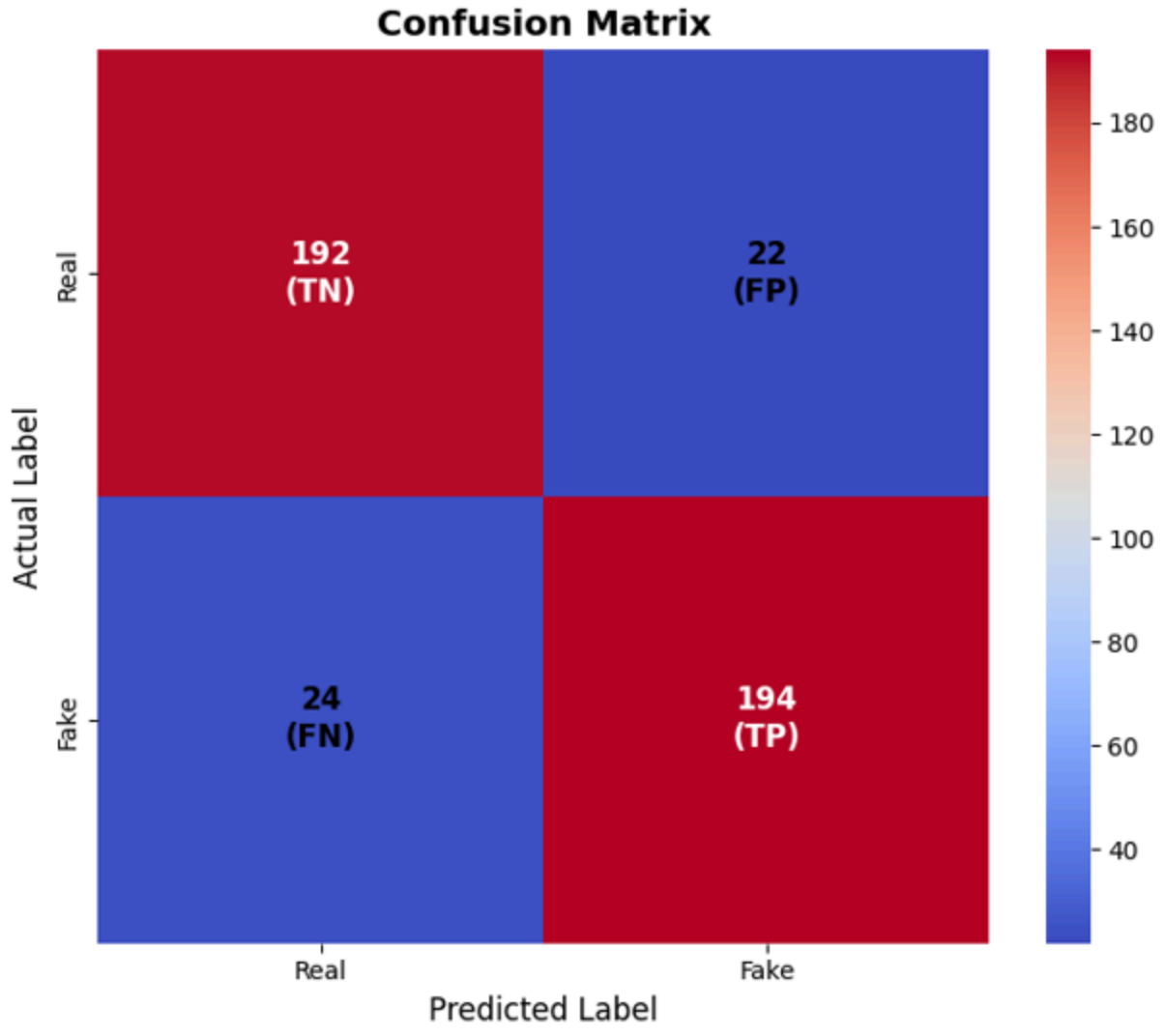


Figure 4

Confusion Matrix

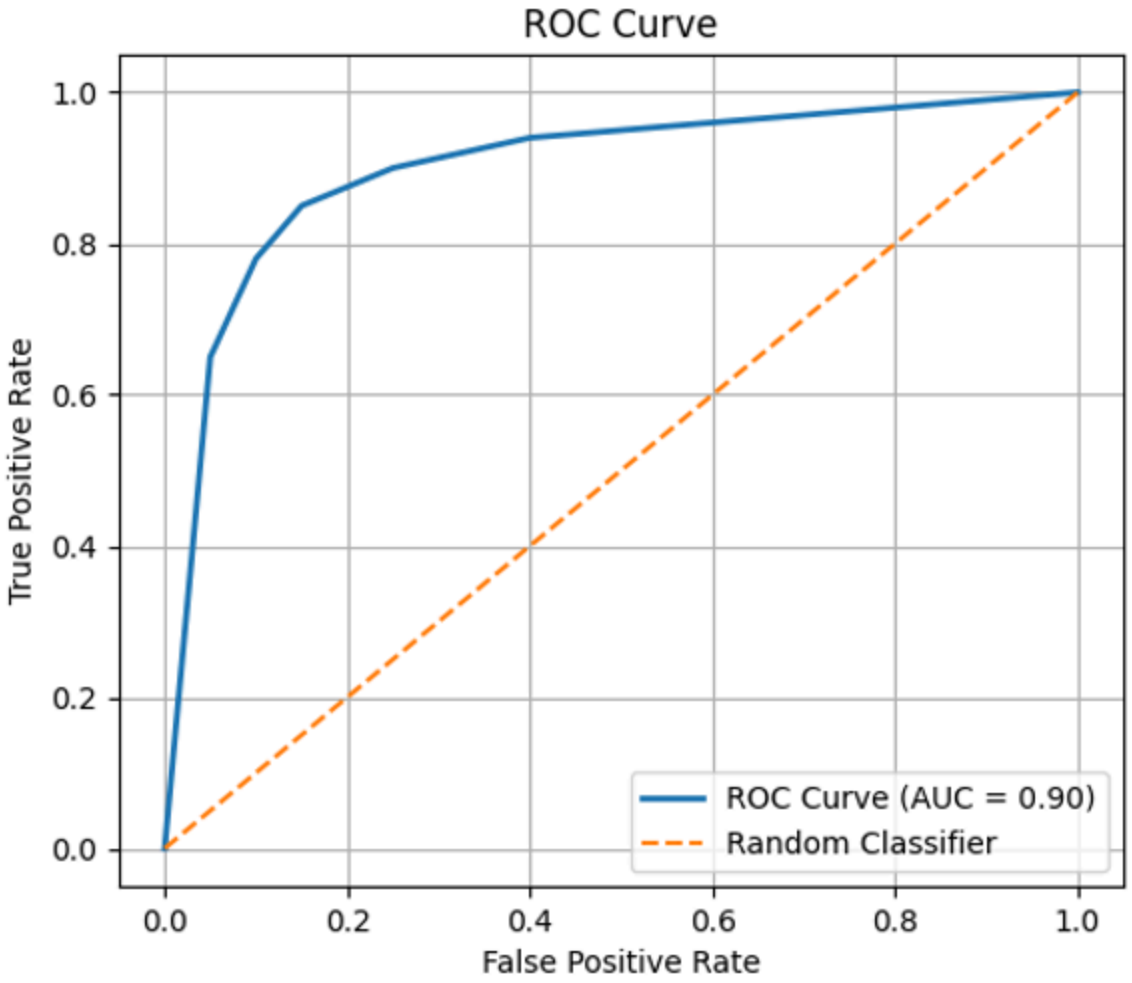


Figure 5

ROC Curve

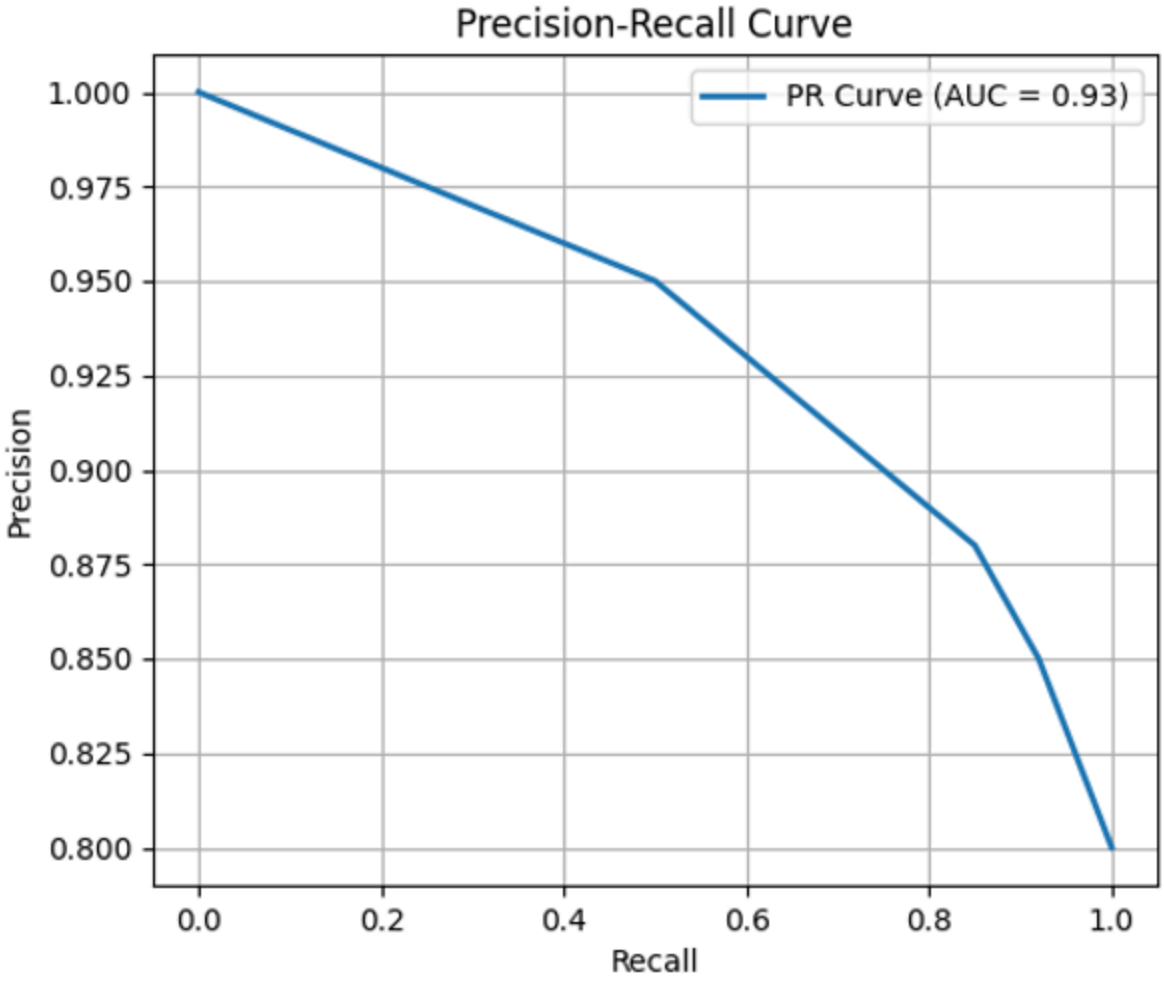


Figure 6

Precision-Recall Curve