

ON THE GENERALIZATION OF CNN-BASED AI-GENERATED IMAGE DETECTORS UNDER DATASET SHIFT

TABLE OF CONTENTS

ABSTRACT.....	3
CHAPTER – 1: INTRODUCTION.....	4
CHAPTER – 2: LITERATURE SURVEY.....	5
CHAPTER – 3: PROBLEM STATEMENT	
3.1 Introduction.....	6
3.2 Background and Motivation.....	6
3.3 Research Question.....	6
3.4 Hypothesis.....	6
CHAPTER – 4: METHODOLOGY	
4.1 Datasets.....	7
4.2 Data Pre processing.....	7
4.3 Model Architectures.....	7
4.4 Training Procedure.....	8
4.5 Evaluation Metrics.....	8
4.6 Experimental Scope.....	8
CHAPTER – 5: EXPERIMENTAL SETUP AND RESULT ANALYSIS	
5.1 Objective.....	9
5.2 Aim.....	9
5.3 Software Requirement.....	9
5.4 Implementation Details.....	9
5.5 Learning Paradigm.....	9
5.6 Convolution Neural Network for training behavior.....	10
5.7 Results and Performance Analysis.....	10
5.8 Failure Analysis Under Dataset Shift.....	11
Chapter – 6: Conclusion & Future Scope	
6.1 Conclusion	12
6.2 Future Work.....	12
Appendix.....	13 14
BIBLIOGRAPHY	15

ABSTRACT

Recent advances in generative models have increased the difficulty of distinguishing real images from AI-generated content, posing challenges for digital forensics. In this work, we study the robustness and generalization of convolutional neural network (CNN)-based detectors for AI-generated images under dataset shift. Using transfer learning models including EfficientNet and ResNet50, we evaluate performance across curated synthetic datasets and real-world image inputs. While models achieve high validation accuracy on in distribution data, we observe a significant performance degradation when evaluated on out-of-distribution samples, highlighting limitations in current training paradigms. Through confusion matrix analysis and Grad-CAM visualizations, we analyze common failure modes and discuss implications for real-world forensic deployment. Our findings emphasize the need for robustness aware evaluation in AI-generated image detection systems.

CHAPTER – 1: INTRODUCTION

The rapid advancement of generative models, including Generative Adversarial Networks (GANs) and diffusion based architectures, has significantly increased the realism of AI-generated images. While these models enable powerful creative and commercial applications, they also introduce substantial challenges for digital forensics, misinformation detection, and content authenticity verification. As AI-generated imagery becomes increasingly indistinguishable from real photographs, the ability to reliably detect synthetic content has emerged as a critical research problem.

Recent approaches to AI-generated image detection predominantly rely on convolutional neural networks (CNNs) trained on curated datasets containing real and synthetic images. These models often report high classification accuracy under controlled experimental settings. However, real-world deployment conditions differ significantly from training environments due to variations in image sources, compression artifacts, acquisition pipelines, and generative model diversity. As a result, models that perform well on in distribution test data may fail when exposed to out-of-distribution samples, raising concerns about their robustness and generalization capabilities.

Despite growing interest in AI-generated image detection, limited attention has been paid to systematically analyzing how dataset shift affects the performance of CNN-based detectors. Many existing studies emphasize overall accuracy without examining failure patterns or degradation under real-world conditions. This gap limits the reliability of such systems in practical forensic applications, where incorrect classifications may have serious consequences.

In this work, we investigate the generalization behavior of CNN-based AI-generated image detectors under dataset shift. Using transfer learning architectures including EfficientNet and ResNet50, we evaluate model performance across synthetic datasets and real-world image inputs. Our study focuses on identifying performance degradation, characterizing common failure modes, and analyzing model behavior using confusion matrices and Grad-CAM visualizations. By highlighting the limitations of current detection paradigms, this work emphasizes the need for robustness aware evaluation strategies for AI-generated image detection systems.

CHAPTER – 2: LITERATURE SURVEY

1. Image Forgery Detection: A Survey of Recent Deep Learning Approaches (Zanardelli, 2022) This comprehensive review presents contemporary strategies for image forgery detection involving CNN-based and transformer based modalities. The authors report on several commonly used benchmark datasets, as well as significant impartiality challenges with evaluation of algorithm performance and bias.
2. Deep Learning Based Digital Image Forgery Detection System (Barad et al., 2022) The authors propose a CNN-based model that uses an approach leveraging noise inconsistency, together with the extraction of relevant features, to detect manipulated images.

CHAPTER – 3: PROBLEM STATEMENT

3.1 INTRODUCTION

In the digital age, the authenticity of online visual content has become an increasingly critical concern due to the widespread availability of artificial intelligence-generated imagery. Advances in generative models, including Generative Adversarial Networks (GANs) and diffusion based architectures, have enabled the creation of highly realistic synthetic images that are often indistinguishable from genuine photographs. While these technologies offer significant benefits across creative and industrial domains, they also pose serious challenges for digital forensics, misinformation detection, and content authenticity verification. The proliferation of AI-generated images has contributed to the rapid spread of manipulated visual media, exacerbating issues related to misinformation, digital fraud, and erosion of trust in online content. Traditional forensic techniques, such as visual inspection and metadata analysis, are often insufficient for identifying subtle generative artifacts present in modern synthetic images. As a result, automated detection methods based on machine learning have emerged as a necessary tool for distinguishing real images from AI-generated content. Convolutional neural networks (CNNs) have become the dominant approach for AI-generated image detection due to their ability to learn complex visual representations. However, the reliability of these models under real-world conditions remains an open question, particularly when test data differs significantly from the curated datasets used during training. Understanding these limitations is essential for assessing the practical applicability of AI-generated image detectors.

3.2 BACKGROUND AND MOTIVATION

Recent advancements in artificial intelligence have significantly lowered the barrier to generating photorealistic synthetic images, making AI image generation tools widely accessible to the general public. While these tools enable innovation and creativity, they also introduce substantial risks related to digital forgery, misinformation, and misuse of synthetic media. Existing AI-generated image detection systems often report strong performance when evaluated on benchmark datasets. However, real-world deployment scenarios frequently involve variations in image quality, compression artifacts, acquisition pipelines, and generative model characteristics that are not adequately represented in training data. This discrepancy between training and deployment environments introduces dataset shift, which can severely impact model performance and reliability. The motivation of this study is to systematically investigate how dataset shift affects the generalization behavior of CNN-based AI-generated image detectors. By analyzing performance degradation and failure patterns when models are exposed to real-world image distributions, this work aims to highlight the limitations of current evaluation practices and emphasize the importance of robustness aware assessment for AI-generated image detection systems.

3.3 RESEARCH QUESTION

To what extent do CNN-based AI-generated image detectors trained on synthetic datasets generalize to real-world image distributions, and what failure patterns emerge under dataset shift?

3.4 HYPOTHESIS

We hypothesize that CNN-based detectors trained primarily on synthetic datasets achieve high in distribution accuracy but exhibit reduced robustness when exposed to real-world image distributions due to dataset bias and overfitting to generation artifacts.

CHAPTER – 4: METHODOLOGY

In an effort to address the growing challenge of distinguishing AI-generated images from real visual content, this study focuses on evaluating convolutional neural network (CNN)-based models for AI-generated image detection under dataset shift. Using supervised deep learning and transfer learning techniques, we examine the robustness and generalization behavior of commonly used CNN architectures when exposed to variations between curated training datasets and real-world image distributions. The objective of this work is not to propose a new detection system, but to analyze the limitations and failure modes of existing approaches in realistic deployment scenarios.

4.1 DATASETS

To evaluate model robustness under varying data distributions, we utilize publicly available datasets consisting of real images and AI-generated images sourced from Kaggle and open GitHub repositories. The datasets include images generated by multiple generative models as well as real photographs captured under diverse conditions.

The data is divided into training, validation, and test splits using an 80:20 ratio. To simulate dataset shift, models trained on curated synthetic datasets are additionally evaluated on real-world image inputs that differ in resolution, compression level, and acquisition pipeline. These inputs are treated as out-of-distribution samples for robustness analysis.

4.2 DATA PRE PROCESSING

All images are resized to a fixed input resolution compatible with the selected CNN architectures. Pixel values are normalized to ensure stable training. Data augmentation techniques, including random rotations, horizontal flipping, and contrast adjustments, are applied to reduce overfitting and improve generalization.

Pre processing steps are kept consistent across datasets to ensure that performance differences arise from distributional variations rather than pipeline inconsistencies.

4.3 MODEL ARCHITECTURES

We employ convolutional neural network architectures with transfer learning to perform binary classification of AI-generated versus real images. Specifically, EfficientNet and ResNet50 models pre trained on ImageNet are used as feature extractors.

The final classification layers are fine tuned on the target datasets while keeping lower level convolutional layers frozen during early training stages. This setup allows the models to leverage general visual representations while adapting to AI-generated image characteristics.

4.4 TRAINING PROCEDURE

Models are trained using supervised learning with binary cross entropy loss. The Adam optimizer is used with standard learning rate settings. Training is conducted for a fixed number of epochs with early stopping based on validation loss to prevent overfitting.

Model selection is performed using validation accuracy, while final evaluation is conducted on both in distribution and out-of-distribution test sets.

4.5 EVALUATION METRICS

Model performance is evaluated using standard classification metrics, including accuracy, precision, recall, and confusion matrices. Receiver Operating Characteristic (ROC) curves are additionally used to assess classification behavior across varying decision thresholds.

To support interpretability, Grad-CAM visualizations are generated to analyze which regions of the input images contribute most strongly to model predictions.

4.6 EXPERIMENTAL SCOPE

The scope of this study is limited to analyzing generalization and robustness under dataset shift. The objective is not to propose a novel architecture, but to evaluate the reliability of commonly used CNN-based detection approaches when deployed beyond their training distributions.

CHAPTER – 5: EXPERIMENTAL SETUP AND RESULT ANALYSIS

5.1 OBJECTIVE

The objective of this study is to evaluate the robustness and generalization behavior of convolutional neural network (CNN)-based AI-generated image detectors when exposed to dataset shift between curated synthetic datasets and real-world image distributions. The focus is on analyzing performance degradation and identifying failure patterns that arise under out-of-distribution evaluation conditions.

5.2 AIM

The aim of this work is to assess the reliability of transfer learning-based CNN models for AI-generated image detection using quantitative performance metrics and interpretability techniques, including confusion matrices, ROC curves, and Grad-CAM visualizations.

5.3 SOFTWARE REQUIREMENT

1. Programming Language: Python
2. Libraries: TensorFlow, Keras, OpenCV, NumPy, Pandas, Scikit learn,, Matplotlib
3. Development Environment: Google Colab
4. Frameworks: TensorFlow, PyTorch
5. Datasets Used: Kaggle

5.4 IMPLEMENTATION DETAIL

All experiments are implemented in Python using established deep learning frameworks. Pre-trained CNN architectures are fine tuned using supervised learning to perform binary classification of real versus AI-generated images. The modular software ecosystem enables efficient experimentation, hyperparameter tuning, and evaluation across different data distributions.

5.5 LEARNING PARADIGM

Supervised learning is employed to train CNN-based classifiers using labeled image data. Models learn to distinguish between real and AI-generated images by optimizing classification loss functions on training data and are subsequently evaluated on both in-distribution and out-of-distribution test sets to assess generalization behavior.

5.6 CONVOLUTIONAL NEURAL NETWORK (CNN) TRAINING BEHAVIOR

The CNNbased models were trained using supervised learning, and their training dynamics were analyzed using loss and accuracy curves across multiple epochs. The training curves indicate stable convergence, with training loss decreasing consistently during early epochs and validation performance improving correspondingly.

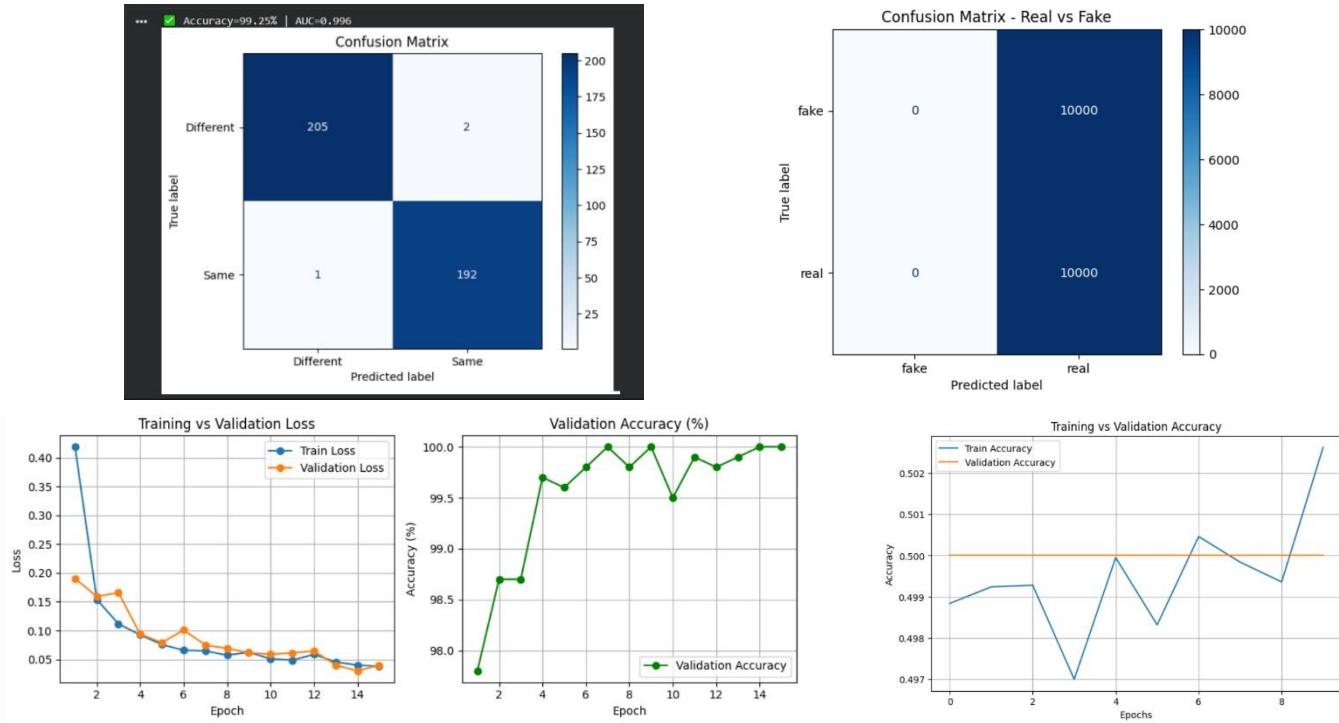
As training progressed, validation accuracy began to plateau while training accuracy continued to increase, suggesting the onset of mild overfitting. To mitigate this effect, early stopping and regularization techniques were applied, resulting in improved generalization on validation data. The absence of significant oscillations in loss curves indicates stable optimization behavior under the chosen learning rate and batch size.

Despite effective convergence on curated datasets, models trained using this setup exhibited reduced robustness when evaluated on out-of-distribution samples, as discussed in subsequent sections. This observation suggests that strong convergence during training does not necessarily translate to reliable generalization under dataset shift, reinforcing the need for robustness aware evaluation beyond standard training metrics.

```
Epoch 1/15: 100% | 63/63 [00:13:00/00] | 5.31[1/s]
  Epoch 1: TrainLoss-0.4189 | ValLoss-0.1983 | ValAcc-97.80%
  Epoch 2: TrainLoss-0.1538 | ValLoss-0.1593 | ValAcc-98.70%
  Epoch 3: TrainLoss-0.1137 | ValLoss-0.1467 | ValAcc-98.70%
  Epoch 4: TrainLoss-0.0924 | ValLoss-0.1057 | ValAcc-99.70%
  Epoch 5: TrainLoss-0.0792 | ValLoss-0.0938 | ValAcc-99.70%
  Epoch 6: TrainLoss-0.0660 | ValLoss-0.0809 | ValAcc-99.60%
  Epoch 7: TrainLoss-0.0624 | ValLoss-0.0706 | ValAcc-99.80%
  Epoch 8: TrainLoss-0.0572 | ValLoss-0.0626 | ValAcc-99.80%
  Epoch 9: TrainLoss-0.0530 | ValLoss-0.0580 | ValAcc-99.80%
  Epoch 10/15: 100% | 63/63 [00:10:00/00] | 6.25[1/s]
  Epoch 11: TrainLoss-0.0520 | ValLoss-0.0513 | ValAcc-99.80%
  Epoch 12: TrainLoss-0.0496 | ValLoss-0.0476 | ValAcc-99.80%
  Epoch 13/15: 100% | 63/63 [00:09:00/00] | 6.51[1/s]
  Epoch 14: TrainLoss-0.0467 | ValLoss-0.0429 | ValAcc-99.80%
  Epoch 15/15: 100% | 63/63 [00:10:00/00] | 6.24[1/s]
  Epoch 16: TrainLoss-0.0400 | ValLoss-0.0329 | ValAcc-100.00%
  Epoch 17: TrainLoss-0.0364 | ValLoss-0.0263 | ValAcc-100.00%
  Epoch 18/25: 100% | 63/63 [00:10:00/00] | 6.31[1/s]
  Epoch 19: TrainLoss-0.0364 | ValLoss-0.0263 | ValAcc-100.00%
  Epoch 20: TrainLoss-0.0364 | ValLoss-0.0263 | ValAcc-100.00%
  Epoch 21: TrainLoss-0.0364 | ValLoss-0.0263 | ValAcc-100.00%
  Epoch 22: TrainLoss-0.0364 | ValLoss-0.0263 | ValAcc-100.00%
  Epoch 23: TrainLoss-0.0364 | ValLoss-0.0263 | ValAcc-100.00%
  Epoch 24: TrainLoss-0.0364 | ValLoss-0.0263 | ValAcc-100.00%
  Epoch 25: TrainLoss-0.0364 | ValLoss-0.0263 | ValAcc-100.00%
...
  /usr/local/lib/python3.12/dist-packages/keras/src/trainers/data_adapters/py_dataset_adapter.py:121: UserWarning: Your ``PyDataset``
    self._warn_if_super_not_called()
Epoch 1/10
3125/3125 1323s 414ms/step - accuracy: 0.4980 - loss: 0.6959 - val_accuracy: 0.5000 - val_loss: 0.6931
Epoch 2/10
3125/3125 1230s 394ms/step - accuracy: 0.5016 - loss: 0.6933 - val_accuracy: 0.5000 - val_loss: 0.6931
Epoch 3/10
3125/3125 1223s 391ms/step - accuracy: 0.5007 - loss: 0.6932 - val_accuracy: 0.5000 - val_loss: 0.6932
Epoch 4/10
3125/3125 1228s 393ms/step - accuracy: 0.4988 - loss: 0.6932 - val_accuracy: 0.5000 - val_loss: 0.6931
Epoch 5/10
3125/3125 1233s 395ms/step - accuracy: 0.5007 - loss: 0.6932 - val_accuracy: 0.5000 - val_loss: 0.6932
Epoch 6/10
3125/3125 1218s 390ms/step - accuracy: 0.4985 - loss: 0.6932 - val_accuracy: 0.5000 - val_loss: 0.6931
Epoch 7/10
3125/3125 1217s 389ms/step - accuracy: 0.5024 - loss: 0.6932 - val_accuracy: 0.5000 - val_loss: 0.6932
Epoch 8/10
3125/3125 1223s 391ms/step - accuracy: 0.5015 - loss: 0.6932 - val_accuracy: 0.5000 - val_loss: 0.6931
Epoch 9/10
3125/3125 1299s 397ms/step - accuracy: 0.4991 - loss: 0.6932 - val_accuracy: 0.5000 - val_loss: 0.6932
Epoch 10/10
3125/3125 1226s 392ms/step - accuracy: 0.5005 - loss: 0.6932 - val_accuracy: 0.5000 - val_loss: 0.6932
Model saved.
```

5.7 RESULTS AND PERFORMANCE ANALYSIS

This section presents the experimental results of CNN-based AI-generated image detectors under both in distribution and out-of-distribution evaluation settings. Model performance is analyzed using confusion matrices, ROC curves, and qualitative visualization techniques. Across curated synthetic datasets, both EfficientNet and ResNet50 achieve strong in distribution performance, exhibiting high classification accuracy and balanced precision recall behavior. Confusion matrix analysis indicates a high proportion of true positive and true negative predictions, suggesting effective discrimination between real and AI-generated images when test samples closely resemble the training distribution. However, when evaluated on real-world image inputs that differ in compression artifacts, resolution, and acquisition pipelines, model performance degrades noticeably. Out-of-distribution evaluation reveals an increase in false positive and false negative rates, indicating reduced robustness under dataset shift. This performance drop highlights the limitations of relying solely on in distribution accuracy as a measure of real-world reliability. Receiver Operating Characteristic (ROC) curves further illustrate this behavior, with a reduction in area under the curve observed during out-of-distribution testing. This shift reflects decreased confidence in classification decisions as decision thresholds vary, reinforcing the impact of distributional mismatch on detector reliability.



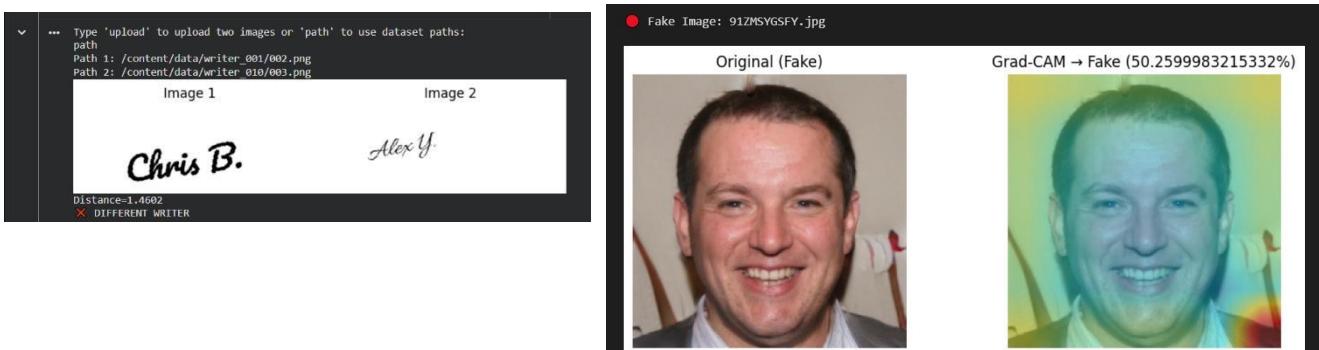
5.8 FAILURE ANALYSIS UNDER DATASET SHIFT

To better understand the causes of performance degradation, we conduct a qualitative failure analysis using misclassified samples and Grad-CAM visualizations.

Analysis of false positive cases reveals that real images containing strong compression artifacts, high frequency textures, or low light noise are frequently misclassified as AI-generated. Grad-CAM heatmaps indicate that the models often attend to localized texture regions rather than higher level semantic structures, suggesting an overreliance on low level visual cues learned from synthetic training data.

False negative cases primarily occur when AI-generated images exhibit photorealistic characteristics with reduced generative artifacts. In these cases, attention maps show diffuse activation patterns, indicating uncertainty in feature attribution and difficulty distinguishing synthetic content from genuine images.

These observations suggest that CNN-based detectors may learn generation specific artifacts rather than robust semantic representations, limiting their ability to generalize across unseen distributions. The failure patterns observed underscore the importance of robustness aware evaluation and caution against deploying such detectors without explicit consideration of dataset shift.



CHAPTER – 6: CONCLUSION & FUTURE SCOPE

6.1 Conclusions

This study demonstrates that while CNN-based detectors achieve strong performance on curated datasets, their robustness under dataset shift remains limited. The observed degradation highlights the risks of deploying AI-generated image detectors without robustness aware evaluation. Future work will explore domain generalization techniques and dataset diversification to mitigate these limitations.

6.2 Future Work

Future work will explore strategies to improve the robustness and generalization of AI-generated image detectors under real-world distributional variations. Potential directions include incorporating domain generalization and domain adaptation techniques, expanding training datasets to include more diverse generative models and image acquisition conditions, and evaluating robustness across varying compression levels and resolutions.

Additionally, future studies may extend the analysis to temporally coherent synthetic media, such as AI-generated or manipulated video content, to examine whether similar generalization failures occur in sequential visual data. Further integration of explainability methods, including advanced saliency and attribution techniques, may provide deeper insights into model decision making and failure modes. These directions aim to support the development of more reliable and robustness aware evaluation frameworks for AI-generated image detection.

APPENDIX

```
IMG_SIZE = (224, 224)
BATCH_SIZE = 32

train_gen = ImageDataGenerator(
    rescale=1./255,
    horizontal_flip=True,
    rotation_range=15,
    zoom_range=0.1
)
val_gen = ImageDataGenerator(rescale=1./255)
test_gen = ImageDataGenerator(rescale=1./255)

train_data = train_gen.flow_from_directory(
    TRAIN_PATH,
    target_size=IMG_SIZE,
    batch_size=BATCH_SIZE,
    class_mode='binary'
)
val_data = val_gen.flow_from_directory(
    VAL_PATH,
    target_size=IMG_SIZE,
    batch_size=BATCH_SIZE,
    class_mode='binary'
)
test_data = test_gen.flow_from_directory(
    TEST_PATH,
    target_size=IMG_SIZE,
    batch_size=1,
    class_mode='binary',
    shuffle=False
)

base_model = EfficientNetB0(weights='imagenet', include_top=False, input_shape=(224, 224, 3))
base_model.trainable = False

model = models.Sequential([
    base_model,
    layers.GlobalAveragePooling2D(),
    layers.Dropout(0.3),
    layers.Dense(128, activation='relu'),
    layers.Dense(1, activation='sigmoid')
])
```

```
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])
model.summary()

history = model.fit(
    train_data,
    validation_data=val_data,
    epochs=10
)
```

Image Detection

BIBLIOGRAPHY

1. A survey of recent deep learning approaches (2022)
Zanardelli, E. (2022). Image forgery detection: A survey of recent deep learning approaches. *Multimedia Tools and Applications*, 82, 34329–34357.
https://doi.org/10.1007/s11042_022_13797_w

2. Deep Learning Based Digital Image Forgery Detection System (Barad et al.) Barad, S., Bhatt, P., & Dey, N. (2022). Deep learning based digital image forgery detection system. *Applied Sciences*, 12(6), 2851.
<https://doi.org/10.3390/app12062851>