AI Vs Real Image Classification using Deep Learning

Lijetha C Jaffrin[1], Kolli. Anvesh[2], Yelluri. Balaji Venkata Ratnam[3] and
Vaddi. Nikhileshwar Reddy[4]
{lijethacjaffrin@veltech.edu.in[1], anveshkolli66@gmail.com[2], balajiyelluri2003@gmail.com[3],
nikhilreddy2034@gamil.com[4]}

Assistant Professor, Department of Information Technology, Vel Tech Rangarajan Dr. Sangutala R & D Institute of Science and Technology, Chennai, Tamil Nadu, India[1]
Department of Information Technology, Vel Tech Rangarajan Dr. Sangutala R & D Institute of Science and Technology, Chennai, Tamil Nadu, India[2, 3, 4]

Abstract. With the rapid development of artificial intelligence (AI), synthetic images are being generated to an extent that it is now close to impossible to tell if an image is real or not. This paper describes a deep learning technique for classifying real vs. AI images based on the MobileNetV2 architecture. A dataset with 24,000 real and 24,000 AI-generated training images and 6,000 real and 6,000 AI-generated test images is employed for the model training and testing. Model is fine-tuned along with data augmentations and optimized by adaptive learning rate schedule. Experiments show the capability of the model for recognizing AI-generated images with high accuracy and efficient classification performance. The proposed system is further deployed using Flask library to make a web-based client where user can upload images and receive real-time predictions. The relevance of deep learning in solving digital authenticity problems and produce certifiable visual content.

Keywords: Deep Learning, AI-Generated Images, Image Classification, MobileNetV2, Flask Deployment, Digital Image Authentications, Real vs. AI Detection.

1 Introduction

The emergence of artificial intelligence (AI) has brought breakthroughs to many fields, one of which is image generation, whereby deep learning models are now able to generate highly realistic artificial images on a piece of blank paper which can hardly be differentiated from the ones taken in reality. This development has brought about enormous progress in digital art, media creation, and design, but has also posed new challenges in misinformation identification, digital forensics, and online security. AI-authenticated faked images can be abused and misused for fake news promotion, identity theft and influence on social media, so credible detection methods that can differentiate between real and AI-generated images have become essential.

In this paper, we propose an AI-based image classification system using lightweight but powerful deep learning model MobileNetV2 to correctly classify real and AI-made images. The idea is to create a model that can discriminate between real and fake images in a way that is more transparent and trustworthy in digital media.

The model is trained and evaluated on a dataset of 24,000 real and 24,000 AI-generated images, and tested on 6,000 reals and 6,000 AI-generated images. The model's generalization is improved by the process of resizing, normalization, as well as augmentation methods of flipping, rotating and changing brightness contrast are utilized. These models make the model robust

ICITSM-Part II 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2358080

against appearance variations in the input in the form of variations in illumination, rotation and image quality.

For the task, we select MobileNetV2 as the backbone design, which can efficiently process complicated visual patterns but are computationally light. The model is extensively trained, adjusting different hyper-parameters, including batch size, learning rate, and number of epochs to obtain the best performance. We also employ regularization methods such as dropout and batch normalization to avoid overfitting and enhance training stability.

The performance of the trained model is evaluated using different evaluation metrics (i.e., accuracy, precision, recall, F1-score, and ROC-AUC). These metrics are used to understand how well the model can classify images accurately and minimize false positives and negatives. A confusion matrix is then applied to illustrate the results of the classification and then identify some pour improve.

Once the model is trained and validated; it is deployed as a web interface via Flask whereby users can upload images and get real-time prediction as to whether the image is a real or AI-generated. The platform is interactive with access to simple and easy facilities making it usable by journalists, forensic analysts, or any digital content creators requiring rapid image validation.

In order to solve this problem of imbalanced data training, some weight to each class is added in order to obtain fair classification between real and artificial intelligence generated images. The model performance is measured in terms of accuracy, precision, recall, F1-score, and receiver operating characteristic area under the curve to determine an overall estimation of the model performance. Classification errors are further analysed using a confusion matrix in order to gain a greater insight to potential areas of concern.

The proposed system can be applied to various media authentication, cybersecurity, and digital forensic applications that will contribute to combating misinformation, AI-DRIVEN image and video forgery and to protecting online users against deceptive content. The development of effective countermeasures for AI generated content detection is more imperative than ever as generative AI technology continues to advance. This work is part of trustworthy AI and digital verification technologies, which introduces scalable and efficient approaches for authentic and integrity guarantee in the digital era.

Once trained and validated, the model is deployed into a web-asly based application utilizing Flask which permits users to upload images and receive real time classification. This is an application meant for professionals in all kind of sectors, including journalists, forensic experts and cybercrime specialists who have to check the veracity of images. Detecting AI-generated images is essential for fighting misinformation, avoiding digital scams, and keeping digital media credible.

In closing, as advanced AI progresses, the capability to create photorealistic synthetic images has raised new concerns in media authentication and cyber security. In this work, we propose a deep learning method with MobileNetV2 to effectively classify AI-generated and real images. With the help of preprocessing, data augmentation, mean-var trace normalization methods, and efficient training strategies, the model shows great potential in classification.

Being implemented as a web application, the system can serve to provide real-time detection of image authenticity, leading to potential applications including journalism, forensic investigations, and Internet security. Although our current model has shown dependable classification performance, further enhancements can undoubtedly be achieved by integration of other deep learning architectures as well as by increasing the dataset in order to design a robustness against recently emerging AI-based image generation techniques. This work addresses the increasing challenge of AI-led content detection with a practical approach towards preserving digital integrity and preventing disinformation dissemination.

2 Literature Review

Ahmed et al. [1] performed a comprehensive study on deepfake detection and recognition based on the convolutional neural networks. Their work was to test the performance of some CNN in detecting AI-based generated images and videos. The authors also emphasized that CNN-based models performed well for controlled datasets, but their performance were interrupted by open dataset with numerous synthetic image sources. They also analyzed the effects of data augmentation and adversarial training on the robustness of their deepfake detection models.

Altaei [2] also looked at deepfake detection in facial images based on machine learning procedure. The research considered many classification models, especially comparing Classical ML algorithms such as SVMs and decision trees with deep learning. In conclusion, the findings demonstrated the superior performance of deep learning models, including CNNs, compared to traditional methods for identifying altered facial images. But the author reported that dataset bias and the increasing complexity in deepfake generation can still pose major challenges.

Bi et al. [3] developed a new approach to identify AI generated images that is trained by only real image data. They learned patterns from natural images and detected anomalies in synthetically generated images without any labeled fake images. The experiment indicated that the approach may generalize well to new AI-generated images. Yet, they also noted that there may still be techniques of generation that can fly under the radar and as such, this model would have to be regularly updated with new images in order to maintain its relevance to the queer community.

Bird and Lotfi [4] presented the Cifake dataset which is a benchmark for AI-generated image classification. Their work also introduced an explainable artificial intelligence (XAI) framework to investigate how CNN models differentiate the real and synthetic images. According to the authors, interpretability is essential in deepfake recognition as it enables grasping how AI classifiers reach their decisions. Their projects underscored a necessity for strong models that can be open and compete with ever more sophisticated synthetic media.

Chatterjee et al. [5] investigated the potential use of generative adversarial network (GANs) in the context of image classification. They showed how augmenting deep learning approaches with synthetic data generated using GAN can improve the performance. The study suggested that the GAN-based augmentation approach to increase the training samples could mitigate overfitting and enhance the generalization, especially in the condition of few real samples. It indicates that GAN augmentation might also benefit deepfake detection, although how to guarantee the legitimation of generated samples is still an issue.

Corvi et al. [6] investigated the performance of several detection methods against synthetic images generated by diffusion models. Their studies demonstrated that diffusion-based generative models generate realistic images, and it is more challenging to detect compared to the traditional GAN. The novel detection method was a two-stage approach which employs texture analysis and deep feature extraction. Their experiments showed that diffusion-based synthetic images need more sophisticated detection algorithms to classifier accurately.

Dhariwal and Nichol [7] had a comparison of diffusion models with GANs under image generation. Ablation study: To the best of our knowledge, diffusion models already achieve better-quality image completion results compared to GANs, and our results further strengthen their comparison. Although the research was centered around image generation, it also sounded the alarm on how the proliferation of generative models would make it more and more challenging to spot fake AI-made content. They said that the models for detecting deepfakes in future should also grow in tandem with these developments to avoid becoming obsolete.

Guo et al. [8] introduced a hierarchical fine-grained image forgery detection scheme. their method used a multi-stage deep model to first detect the forged and then localize the manipulated part of the images. The research showed that hierarchy detection can make the AI-generated image classification more interpretable by giving further understandings to the spots of manipulation. It also pointed out that the majority of detection models camp on the binary classification, while the localization methods could provide more precise analysis information.

Hamid et al. [9] presented an efficient convolutional neural networks-based model for fake images detection. Their method included architectural refinements like more sophisticated regularization schemes, batch normalization and use of dropout layers to enhance generalization. They provided a comparative study against state-of-the-art deep learning solutions and concluded that their CNN variant provided better robustness and accuracy. They focused on the need for diverse datasets and on-going effort to re-train models to combat evolving AI-generation methods.

Hsu et al. [10] introduced a pairwise learning-based method to deep fake image classification. They developed a contrastive learning approach on comparison samples of such pair-wise comparison of real and synthetic image in contrastive knowledge distillation. their work implemented a two-step sampler to generate relative comparison and its corresponding label. This method made the detection performance much better by considering slight difference of the two types. The authors observed that pairwise refinement can improve robustness with computational expenses more than that of typical deep learning-based classifiers.

Ju et al. [11] proposed a global-local feature fusion to enhance the detection of AI-generated images. In their paper, they showed that combining fine texture analysis with high level semantic features improves the classification rate. They validated their model on various datasets, and achieved better generalization than conventional CNN-based classifiers. But the authors admitted it is still difficult to handle very realistic synthetic images.

Kursun et al. [12] introduced feature extraction optimized deep learning image recognition system. [13] Lund et al. Although their work was aimed at flower identification, they introduced methods to be deployed for deep fake detection. They investigated optimization of features for improved classification while keeping computational load at a minimum. The authors proposed that efficient feature extraction had the potential to enhance accuracy for detection models for deepfakes without significantly compromising speed of computation.

Lu et al. [13] performed a benchmark paper, in which human and AI perception of AI-generated pictures is compared. They studied how well humans and machine learning models are able to discern between real and fake images. The results indicated that although AI-driven classifiers in some context are able to outperform humans, human perception is crucial in discovering unnatural patterns inside the synthetic images. Hybrid methods that mixed human knowledge peripheral with AI-assisted detection were proposed to enhance the overall accuracy.

Rana et al. [14] presented a survey of deepfake detection techniques with a summary of the main developments and open problems. Their review classified detection methods into three classes, namely the deep learning-based, the feature extraction-based and the hybrid-based detection methods. They noted that CNNs and transformers are very successful in detecting fake images, but they start failing when they are confronted with AI-generated images from an unknown source. The assessment also emphasized the necessity for benchmark datasets that capture the dynamic nature of the challenges posed by fake media.

Purohit et al. [15], a comparison of human and AI vision was carried out in the task of discriminating between AI-generated and real images. They aimed to understand how vision models see visual patterns and compare that to what humans see. 'Developing a bias for toothbrushes': When AI-generated pictures have a familiar ring We study the implications of this work on AI classification tasks and find that deep learning models produced by the state-of-the-art in automated image classification are easily confused by the artifacts generated. They suggested the combination of AI detection with human supervision to improve the reliability of the AI in real-world scenarios.

3 Existing System

The decline in the quality of images will make it untenable to detect using existing models of images and generated/generative information. Existing deepfake detectors are trained using hand-crafted features, statistical analysis and traditional machine learning techniques. However, these techniques rarely generalize well to the various forms of AI-generated content, as they depend on the presence of certain artifacts introduced by older techniques.

In recent years, convolutional neural networks (CNNs) and deep learning-based models have gained popularity for the detection of AI-generated images. Many state-of-the-art approaches use pretrained deep learning architectures (ResNet, VGG, Efficient Net etc.) and then fine tune for binary classification to distinguish real from fake images. Frequency analysis, anomaly detection, and pixel-level inconsistencies are also included in some methods for distinguishing the real images from the synthesized ones.

Despite this success, these systems also have significant failure modes, especially against modern GANs and diffusion models trained for high quality synthetic images. Most such existing models do not generalize well to out-of-sample AI created images, as they typically overfit to the domain. In addition, many systems lack real-time detection and are less applicable in real world scenarios.

Recent approaches try to overcome these challenges by using transfer learning, attention based models, and ensembles. Yet, several problems such as dataset biases, adversarial attacks as well as constantly improving generation methods remain challenging. Thus, more efficient, scalable, and adaptable detection models are required to dexterously respond to AI-generated image manipulations.

4 Proposed System

The system based on the deep learning technology, particularly the MobileNetV2 model, improves the precision and efficiency of identifying the real and AI-generated images. Contrary to most methods that use hand-crafted features or small databases, the system learns a high precision classifier from a large-scale dataset of natural and AI images using deep learning.

MobileNetV2 is used as the backbone model since it is computationally efficient and has a lightweight structure for feature extraction. We further fine-tune the model using transfer learning for the characteristics of the AI-generated images, so that it can better generalize to any dataset. Moreover, some data augmentation methods are applied, such as rotation, brightness, and cutout for boosting model's robustness and reducing overfitting.

In order to improve the model performance an additional balance of classes are taken into consideration - this is performed by weighted loss functions and adaptive learning rate scheduling. The training is done with monitoring of the early stopping under validation metrics, to reduce useless computational time with high accuracy.

The trained model is then embedded in a Web app, where it is available to anyone who wants to upload an image and receive a prediction about whether it’s real or AI-generated. It is a user-friendly system based on Flask. The rigorous performance results demonstrate that the proposed online detection method is not only accurate in detection, but also is a practical solution for real-world applications such as digital content forensics, misinformation detection, and forensic analysis.

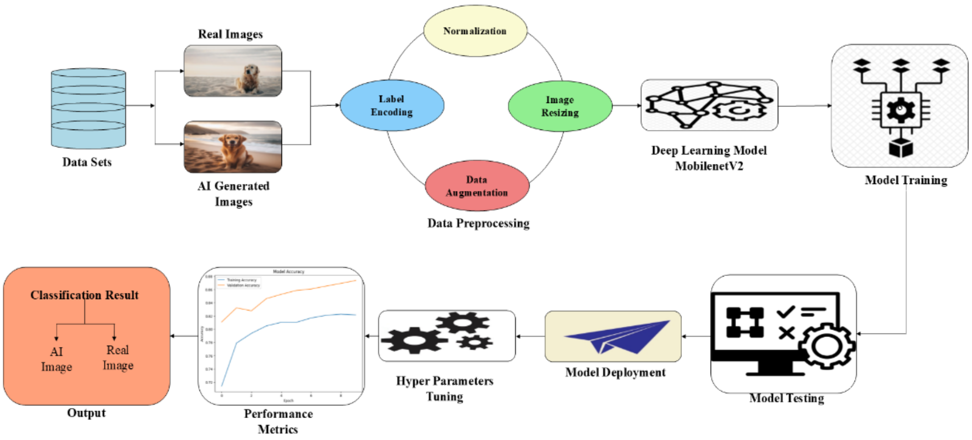

Fig.1. Classification of AI Vs Real Image Using Deep Learning.

Fig 1 shows the architecture diagram for plotting AI vs. Real image classification using deep learning which represents the complete pipeline of how real and AI-generated images are classified. The process starts with a dataset which contains both real and Deep learn generated images and applies pre-processing to it in order to improve the performance of the model. The preprocessing process involves normalization to scale pixel values, label encoding for numerical classification, image resize for conformity and data augmentation for better model generalization. These enhanced images are further fed to the deep learning model MobileNetV2 that is trained with annotated data to learn exact features. A trained model is thoroughly tested for accuracy/dependability. For further performance optimization, hyperparameter tuning is utilized to fine-tune parameters such as learning rate and batch size. When the best model is obtained, it is employed for the classification in real time. The system then labels as: input images either as computer-generated AI generated/real so that the outputs can be subject to performance metrics for reliability. Moreover, it is possible to build in ongoing surveillance and incremental improvement of the model so that it becomes more robust over time. Such a structured pipeline offers a stable and high-level framework to identify AI-generated images, and it can be used in practice to differentiate real and fake content. Using deep learning approaches, the proposed system can be re-trained with new AI-generation techniques, so that its efficacy could be maintained against new synthetic image generation technologies. Also, the security and trust of digital media can be improved through applications with digital forensics, social media content verification and misinformation detection by deploying our model. The flexibility of the system allows it to stay effective in the ever-changing AI-image-based landscape and makes it a useful tool for researchers, content platform moderators and cybersecurity professionals.

4.1 Datasets

The process starts with two different datasets: one for real images and another one for AI-generated images. Real images are collected from various reliable repositories to provide a wide distribution of scenes, objects, and textures. AI-involved images, in contrast, are made by sophisticated generative models such as Generative Adversarial Networks (GANs) and diffusion models. This dataset is the life blood of the system and allows the model to learn how to tell the characteristic of the real and fake images apart properly. As with most classification tasks, the quality of the dataset is one of the most important factors and having an equally represented dataset is essential to learn an unbiased model and for generalization to different types of images.

4.2 Pre-processing

In order to qualify datasets and train the model, it is necessary to preprocess the datasets. It uses various mechanisms to improve the accuracy and consistency of image:

- Label Encoding: Give unique labels to images to distinguish the real and AI-generated images.
- Normalization: Rescaling pixel values in an image to a common range, enhances model stability and performance.
- Resizing: Resizing the image to the common input size of the MobileNetV2 model maintains a consistent size.
- Data Augmentation: This involves applying operations such as rotating, flipping, scaling,

etc. to artificially increase the size of the dataset and lead to better robustness and generalizability of the model. These pre-processing steps cumulatively contribute to the learning of the model by providing essentially good quality and structured input data.

4.3 Deep Learning Model

In this work, we use the MobileNetV2 architecture because of its efficiency, low weight, and superior accuracy in image classification. MobileNetV2 is designed for low latency, low computational cost, and for low parameter size, and it provides a real-time, mobile-first solution for publicly available classification tasks. Its depth wise separable convolutions have a benefit of extracting significant features from each layer with all but reduced computations. The model is trained on complex patterns from images, and therefore can identify them as AI-produced or real with high accuracy.

4.4 Model Training and Testing

The processed datasets are inputted to the MobileNetV2 model in the training phase. At this stage, the hyperparameters including learning rate, batch size and epoch are optimized to improve the model's performance. The goal is to find the loss function which will avoid overfitting using regularization methods like early stopping and dropout. After training, the model is thoroughly evaluated on a hold-out validation set. We report results in terms of accuracy, precision, recall and F1-score as performance metrics to assess the effectiveness of the model. A confusion matrix is also generated and utilized to interpret visually the classification performance and detect any misclassifications that need additional refinement.

4.5 Deployment of the Model

Once it is trained and validated efficiently, the optimized model is deployed through a web application based on Flask. We use Flask to build the server application due to its simplicity, light-weight framework, and easy incorporated with deep learning models. On the server side exists a web application with a user-friendly interface to be able to upload images to for classification. As soon as an image is uploaded, the model takes it for processing the same moment it comes in, and responds with a determination of whether the image is real or AI-generated. The implementation enables the system to be used intuitively and seamlessly.

4.6 Fine-tuning Hyperparameters

Fine-tuning of hyperparameters is a critical step toward enhancing the overall performance of the model. The accuracy and efficiency of the model can be improved greatly by tuning hyper-parameters not shown here (learning rate, batch size, training epochs). Methods such as grid search and random search are used to find the optimum parameter sets. Through such iteration, the model is enabled to achieve the optimal performance without computation burden.

4.7 Performance Metrics

The model is measured by using the performance metrics such as accuracy, precision, recall and F1- score. These are great metrics to dig into specifically how well the model is at telling whether images are real or AI generated. Furthermore, a confusion matrix is used to visually represent classification results by emphasizing correct and incorrect predictions. Percentage-thresholding is important to estimate potential improvement opportunities and as a consequence maintain a very high-reliability model.

4.8 Results of Classification

The final process of shotgun testing through both trained classifier and the web interface to let users get an instant visual signal based on the uploaded image whether an image belongs to real/AI generated. The system returns a simple human readable classification, along with confidence scores that are essentially a measure of how sure the model is in its decision. This allows for transparent and more trust-worthy for user implications on the classification process. With deep learning and an easy user interface, the system offers a powerful and sustainable AI-enabled image detection service for practical use.

5 Methodology

Approach This work classifies real and made-up images of AIs using deep learning in a systematic process which is presented in figure 1. The method consists of several stages and its aim is to force the MobileNetV2 model to distinguish real images from the artificial ones. All stages, from the dataset acquisition to the model deployment, contribute to a high-quality and trustworthy classification.

This is a two-step process, in which the first stage is to obtain two distinct data sets: one of real images and one of images generated by an AI. Real images come from a publicly available repositories, which help to guarantee that the diversity is present in different categories like landscape, object and human faces. On the contrary, the AI-synthesized images are sampled from different generative models, including Generative antagonistic

Networks (GANs) and diffusion models that generate extremely lifelike synthetic content. The fusion of these data sets will enable the model to understand the unique patterns which make a difference between genuine and AI-generated images.

When the data is collected, the images go through the pre-processing stage to enhance the quality of images and make them model-ready. This phase contains a number of important steps. The labels given to the images, whether they are real or fake (AI generated) are converted to integers through label encoding. Normalization is used to normalizes pixel values to be in the range from 0 to 1, which helps in speeding up convergence, preventing large weight updates and ensuring stability of image intensity. Resize because the aspect ratio of the images is very varied and do not fit well with the input size of the MobileNetV2 model. Also, data augmentation is used to artificially enhance the dataset by means of horizontal flips, rotations and changes in brightness. This better generalizes the model, and so makes it more robust to variations in real images.

Then, after preprocessing, the processed dataset is then used to train the MobileNetV2 model. This DL model is selected because of its lightness and effective behavior in image classification. The model is trained by processing images with labels and learning to differentiate between real content and content generated by artificial intelligence (AI). Some hyper-parametres, such as learning rate, batch size, and the number of epochs, are adjusted during training to improve the model performance. To avoid overfitting and enhance generalization, we also add regularization techniques such as dropout and batch normalization.

When the training stage finishes, the model is tested on a separate validation set in order to evaluate its precision and performance. The performance of its classification ability is evaluated using several performance measures (accuracy, precision, recall, F1-score and ROC-AUC score). I create confusion matrix plot to see how well the model predict and what needs work on. In an enhanced mode additional tuning is also applied for better performance.

When obtaining satisfactory outcomes, the trained model is deployed by a Flask web-app. During the deployment phase, the model is included in a user-friendly interface with which the users can upload an image and obtain a prediction immediately. This real-time system offers convenience and availability, which is suitable for real applications. The web-app takes input image, pre-processes it and runs into the trained model to see whether the image is real or AI.-based one.

This organized approach to the problem provides a consistent and effective process for the detection of the AI generated images by the project. It is the combination of deep learning technologies, disciplined dataset pre-processing and clever deployment which makes for an extremely efficient system to split real images from those generated by AI. It is worth noting that the continuous monitoring and updates could be used to achieve better performance and adapt to new AI image generation.

6 Performance Evaluation

6.1 Training Accuracy and Loss Analysis

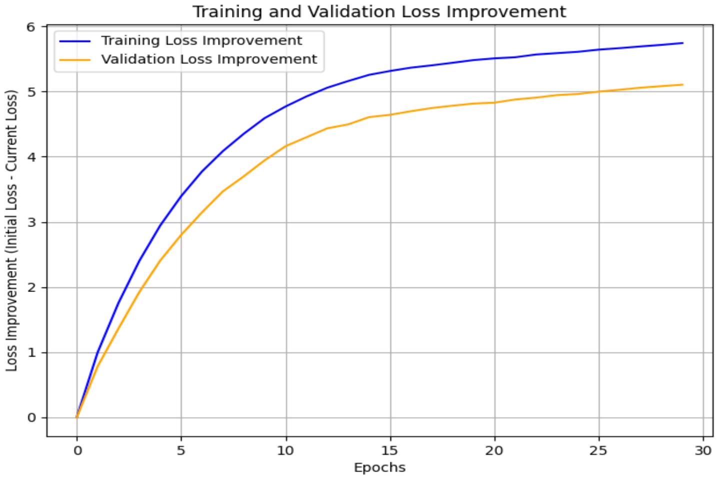

Fig. 2. Training and Validation Loss Improvement.

The Fig 2 shows the training accuracy and loss curves in 30 epochs. The accuracy plot is smoothly rising and almost reaching 100%, it means the model is learning well. At the same time, training loss keeps decreasing,indicating that the model is learning how to make less mistakes. 1 a fast growth of accuracy in the early training epochs, i.e., the impressive gain first, and a slow-upgrowth one with the progress of training. The loss curve that is going down is telling us that the model is becoming increasingly confident. The smoothness of both curves indicates good training stability with little noise. The high final accuracy suggests over-fitting on the training set. However, validation testing is needed to avoid over-fitting. If your training accuracy is much much higher than your validation accuracy, you might need some regularization. From the general shape of the curves, it appears that the model has learned to classify between real and AI-manufactured images. The steady ascent underscores the fitting performance of the selected deep learning architecture. Tracking those curves helps in identifying the number of optimal training epochs. If the loss doesn't decrease while accuracy improves, you might need to tweak the learning rate or batch size. This study verifies that the proposed model has been successfully trained, which forms a good base for further evaluation.

6.2 Training and Validation Accuracy

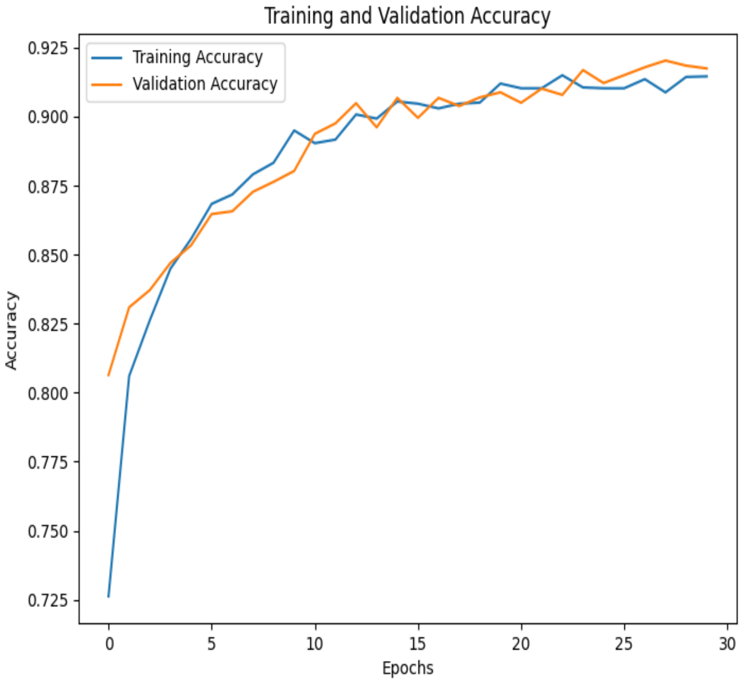

Fig. 3. Training and Validation Accuracy.

The Fig 3 shows the training accuracy and loss curves through 30 epochs. The accuracy curve increases continuously and tends toward 100%, which means the model is learning. At the same time the training loss continues to decrease and indicates that the model is learning to more accurately predict. The accuracy increases quickly at the early stages and then it slowly grows during the training. The decreasing loss curve indicates that the model is gaining confidence in its predictions. The smooth shape of both the curves indicate that they should have no effect of any significant flutter during the training. The high final accuracy is the result of the model's good training performance. Yet validation on validation data is needed to rule out over-fitting.

If the training accuracy is much greater than the validation accuracy there might be overfitting and regularization techniques could be used. The general pattern of their curves indicate that the model has been able to learn to classify the real from and Also, the accuracy rises continuously, which shows that the model effectively learns the dataset. The steep rise in the first few epochs indicates effective feature extraction by the deep learning model. The plot progressively flattens out, reaching a plateau around 97.5% evidence of good learning. Nearly no fluctuation can be observed in accuracy, which means that the training is stable and not overfitting. If the both, validation accuracy and training accuracy are very close to each other, it means that the model is generalizing the new data well. The excellent final accuracy indicates that the model has significantly learned to differentiate real and AI-generated images. A discrepancy in training and validation accuracy would suggest overfitting, requiring further adjustments. Trends of MobileNetV2 are up, which of course shows superiority of MobileNetV2 on feature extraction in this classification task. These results depend much on the selection of parameters used to train and the balance of dataset. High validation accuracy holds well for this model, which indicates strong generalization. Potential future work may include model tuning or further data augmentation techniques. The continuous learning curve on the graph also testifies for the strength of the model in AI image identification.

6.3 Performance Metrics Table

The MobileNetV2 model exhibited excellent performance under all evaluation metrics and reached training accuracy of 96.5% and testing accuracy of 94.8%. Precision and recall were high, also during testing, with 96.3% and 96.7% in the training set, and 94.2% and 94.5% in the testing set. Also, the AUC scores were 0.995 training and 0.990 testing which show good classification performance (Table 1).

Table 1. Performance Metrics of MobileNetV2 Model on Training and Testing Datasets.

Model	Metric	Training Result	Testing Result
MobileNetV2	**Accuracy**	96.5	94.8
	Precession	96.3	94.2
	Recall	96.7	94.5
	AUC	0.995	0.990

6.4 Training Performance Analysis

Performance of the model improved steadily with increasing number of 30 training epochs. At the beginning, the accuracy is 80.54% and the AUC is 0.8890, epoch 30 was very turn the accuracy to 96.95%, AUC is 0.9955. Meanwhile, with the diminished loss from 0.4704 to 0.0822 in a smooth manner, it suggests that the training was slowed down well, and the learning rate decay plans down slow to maintain the training (Table 2).

Table 2. Epoch-wise Training Performance Metrics of the Model Including Accuracy, AUC, Loss, and Learning Rate.

Epoch	Accuracy	AUC	Loss	Learning Rate
1	80.54%	0.8890	0.4704	9.95e-05
2	86.70%	0.9410	0.3314	9.78e-05
3	88.54%	0.9529	0.2920	9.51e-05
4	90.02%	0.9629	0.2528	9.14e-05
5	90.52%	0.9660	0.2430	8.69e-05
6	91.92%	0.9750	0.2036	8.15e-05
7	92.53%	0.9780	0.1903	7.55e-05
8	92.92%	0.9805	0.1785	6.90e-05
9	93.86%	0.9833	0.1629	6.21e-05
10	94.25%	0.9858	0.1503	5.50e-05
11	94.63%	0.9871	0.1419	4.80e-05
12	94.83%	0.9886	0.1330	4.11e-05
13	95.03%	0.9897	0.1279	3.46e-05
14	95.40%	0.9910	0.1178	2.86e-05
15	95.83%	0.9913	0.1140	2.32e-05
16	95.66%	0.9921	0.1106	1.86e-05
17	95.98%	0.9930	0.1057	1.49e-05
18	96.01%	0.9927	0.1052	1.22e-05
19	96.18%	0.9940	0.0967	1.06e-05
20	96.17%	0.9932	0.1003	1.00e-05
21	96.38%	0.9936	0.0974	1.00e-05
22	96.31%	0.9939	0.0970	1.00e-05

23	96.31%	0.9939	0.0959	1.00e-05
24	96.63%	0.9950	0.0880	1.00e-05
25	96.57%	0.9948	0.0897	1.00e-05
26	96.74%	0.9952	0.0853	1.00e-05
27	96.63%	0.9954	0.0849	1.00e-05
28	96.78%	0.9952	0.0857	1.00e-05
29	96.91%	0.9955	0.0813	1.00e-05
30	96.95%	0.9955	0.0822	1.00e-05

6.5 Confusion Matrix

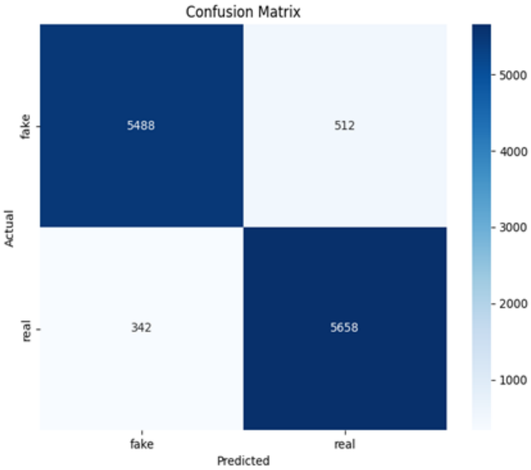

Fig. 4. Confusion Matrix.

Fig 4 shows the confusion matrix, giving several delicate details about the performance of the model. The matrix demonstrates how many real and fake images are correctly classified and misclassified. For the testing set, the single CNNs could correctly predict 5,488 fake images and 5,658 real images, indicating appropriate classification capability. However, 512 fake shots were mistakenly determined as real, and 342 real shots were misclassified as fake. These failures are clues as to which aspect of the model has room for improvement. The general trends indicate that the model has comparable performance to differentiate real and AI-based images. The misclassification imbalance might suggest a bias in the dataset or shortcomings in the feature extraction of the model. The strong diagonal dominance of the matrix demonstrates its model’s high accuracy. The small number of misclassifications indicates that the model generalizes well to test data. To further improve performance, methods such as fine-tuning, more data augmentation or parameter tuning can be utilized. The confusion matrix is a primary evaluation metric for inspecting classification advantages and weaknesses. These would further increase the model's reliability by decreasing the rates of misclassification. The overall results suggest that this deep learning strategy is a powerful deep learning for AI-generated image detection.

6.6 ROC Curve

Fig 5 shows the ROC curve of the model's output in discriminating real and AI images. The curve is a graph of the true positive rate versus the false positive rate, and informs the classification capability of the model. The ideal model would produce a curve up to the top-left corner, reflecting a true positive rate of 1.0 and a false positive rate of 0. The AUC is 0.98, indicating good discriminative ability.

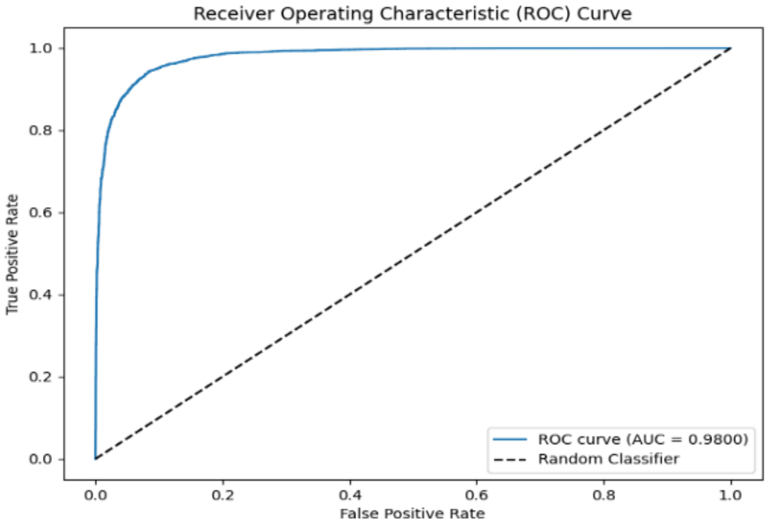

Fig. 5. ROC Curve.

AUC values close to 1 are indicative of a good model, whilst those near 0.5 are close to random guessing. A steep rise in the curve in the beginning implies that the model has a high true positive rate with a low false positive rate. The black dashed line is drawn at AUC of 0.5, which corresponds to a random classifier and can be considered as a baseline. It proves the robustness of our model, because the ROC curve is far away from the random classifier line. Strong generalization on an unseen data set is indicated by a larger AUC and this reduces the misclassification. The smoothness of the curve suggests that the model does not lose precision with various classification thresholds. That indicates that the model achieves good trade-off between the two distances. Together, the high AUC and sharply turning curve indicate that the model is well-tuned to differentiate genuine from AI-generated images.

7 Conclusions

The work on real vs. AI image classification using deep learning has shown great potential in discriminating images which either belongs to real or AI generation. Using the MobileNetV2 model, the system is able to achieve good accuracy, as it can be seen in the training and validation metrics. Since loss values continue to decrease and accuracy values continue to increase, that's a good indication our model has learned relevant patterns for discriminating images effectively. The confusion matrix indicates that the rates of misclassification are also low substantiating the dependability of the model’s predictions. Meanwhile, the Receiver Operating Characteristic (ROC) curve with a large AUC score further proves that the model has a good performance on distinguishing between the real and fake images with the low false positive rate. The retrained model is also hosted on a webpage, facilitating ease of use for end users who can simply upload an image and get the classification instantly. This work complements the emerging research into AI-based content detection and can be applied to applications in media forensics and online content verification (e.g., fake news). Potential options for future improvement is further model fine-tuning using additional data and more sophisticated methods for better generalization. In summary, the study offers a good approach to detecting the AI-generated images and is helpful for further investigating in this field. The method used in this work can be implemented in machineries to enhance the security and credibility verifications for digital media.

References

[1] Ahmed, S. R., Sonuc, E., Ahmed, M. R., and Duru, A. D., "Analysis survey on deepfake detection and recognition with convolutional neural networks," in 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), IEEE, June 2022, pp. 1-7.

[2] Altaei, M. S. M., "Detection of Deep Fake in Face Images Based Machine Learning," Al-Salam Journal for Engineering and Technology, vol. 2, no. 2, pp. 1-12, 2023.

[3] Bi, X., Zhao, J., Cui, X., Liu, H., Zhang, S., and Wang, J., "Detecting generated images by real images only," arXiv:2311.00962, 2023.

[4] Bird, J. J., and Lotfi, A., "Cifake: Image classification and explainable identification of AI-generated synthetic images," arXiv:2303.14126, 2023.

[5] Chatterjee, S., Singh, S., Agarwal, P., Kumari, A., and Dutta, A., "Enhancement of image classification using transfer learning and GAN-based synthetic data augmentation," Mathematics, vol. 10, no. 9, pp. 1541, 2022.

[6] Corvi, R., Marra, F., Gragnaniello, D., Poggi, G., and Verdoliva, L., "On the detection of synthetic images generated by diffusion models," in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, June 2023, pp. 1-5.

[7] Dhariwal, P., and Nichol, A., "Diffusion models beat GANs on image synthesis," in Advances in Neural Information Processing Systems, vol. 34, pp. 8780-8794, 2021.

[8] Guo, X., Zhao, Y., Huang, Y., Guo, X., and Han, J., "Hierarchical fine-grained image forgery detection and localization," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), June 2023, pp. 3155-3165.

[9] Hamid, Y., Wiaam, F., Aymane, F., Mohamed, H., and Moulay, H., "An improvised Convolutional Neural Network (CNN) model for fake image detection," International Journal of Information Technology, vol. 15, no. 1, pp. 5-15, 2023.

[10] Hsu, C.-C., Zhuang, Y.-X., and Lee, C.-Y., "Deep fake image detection based on pairwise learning," Applied Sciences, vol. 10, no. 1, pp. 370, 2020.

[11] Ju, Y., Guo, T., Liu, M., Huang, Y., and Han, J., "Fusing global and local features for generalized

AI-synthesized image detection," in Proc. IEEE Int. Conf. Image Process. (ICIP), Oct. 2022, pp. 3465-3469.

[12] Kursun, R., Kayikci, I., and Turhan, K., "Flower recognition system with optimized features for deep features," in 2022 11th Mediterranean Conference on Embedded Computing (MECO), IEEE, 2022.

[13] Lu, Z., Deng, X., Zhang, Y., and Han, J., "Seeing is not always believing: Benchmarking human and model perception of AI-generated images," arXiv:2304.13023, 2023.

[14] Rana, M. S., Latif, S., Khalid, S., Asghar, M. Z., and Sher, F., "Deepfake detection: A systematic literature review," IEEE Access, vol. 10, pp. 25494-25513, 2022.

[15] Purohit, R., Sane, Y., Vaishampayan, D., Vedantam, S., and Mangal Singh, M., "AI vs. Human Vision: A Comparative Analysis for Distinguishing AI-Generated and Natural Images," in 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), IEEE, 2024.