IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Identifying Synthetic Media with Vision Transformers

[1]Ms. Virginia P. Gonsalves, [2]Ms. Muskan A. Pathan, [3]Ms. Tahira H. Shaikh

[4]Plasin Francis Dias

[123]Student, [4]Assistant Professor
[1234]Department of Electronics & Communication Engineering,
[1234]KLS VDIT, Haliyal, Karnataka, India

Abstract: Artificial Intelligence (AI) has advanced to the point where it can generate hyper-realistic images that closely resemble genuine visuals. These synthetic images, while a testament to technological progress, raise significant concerns regarding misinformation, digital deception, and media manipulation. To address these challenges, this study proposes a novel image authenticity detection framework leveraging Vision Transformers (ViT)—a state-of-the-art deep learning architecture known for its superior performance in image classification tasks. A custom-curated dataset containing both real and AI-generated images is used to train and evaluate the model. By analyzing global and local image features through the self-attention mechanism of ViTs, the system effectively classifies images based on their authenticity. This approach aims to enhance digital media integrity, contributing to the responsible and ethical use of AI-generated content.

***Keywords* – Vision Transformers (ViT), Artificial Intelligence, Real images, AI-generated images**

I. Introduction

The rapid advancement and accessibility of Artificial Intelligence (AI) technologies, particularly through generative models such as Generative Adversarial Networks (GANs) and diffusion models, has led to the proliferation of highly realistic synthetic images that are often indistinguishable from genuine visual content. While these technologies open up new horizons in creative industries, reduce production time, and fuel innovation, they simultaneously introduce significant risks. The misuse of AI-generated content poses severe challenges such as misinformation propagation, synthetic media in legal or journalistic contexts, fake product imagery in e-commerce, and deepfakes that deceive individuals and institutions alike. As AI-generated visuals grow in quality and complexity, it becomes increasingly difficult—if not impossible—for the human eye alone to detect subtle falsifications, thereby undermining trust in visual media and necessitating the development of automated authenticity detection systems.

To address these concerns, this paper presents a novel approach to image authenticity detection using Vision Transformers (ViTs)—a cutting-edge deep learning architecture inspired by the success of transformers in natural language processing. Unlike traditional Convolutional Neural Networks (CNNs), which focus on localized spatial features using filters and pooling operations, Vision Transformers leverage self-attention mechanisms to capture both global and local dependencies across the entire image. This design enables ViTs to model long-range relationships between pixels and extract high-level abstract features more effectively, which is particularly valuable for identifying subtle irregularities that distinguish synthetic images from real ones.

Vision Transformers represent a paradigm shift in computer vision. Rather than using hand-engineered kernels and hierarchical convolutional layers, ViTs treat image patches as sequences of tokens—similar to words in a sentence—and apply multi-head self-attention to learn contextual relationships among them. This capability allows them to attend to multiple regions of an image simultaneously, adaptively weighing their importance based on the learned features. As a result, ViTs can detect nuanced anomalies in texture consistency, unnatural shading, repetitive artifacts, or semantic mismatches that are often overlooked by conventional models.

In this study, we design a ViT-based image classification system trained on a meticulously curated dataset consisting of 5,392 authentic and synthetic images. The dataset includes a diverse range of categories such as human faces, landscapes, objects, and digital illustrations, ensuring robustness and generalizability across different types of media. During training, the Vision Transformer learns high-level image representations through positional encodings, linear projections, and self-attention blocks—thereby automating the feature extraction process without relying on manual heuristics. This comprehensive approach improves the model's ability to recognize the minute distinctions that differentiate AI-generated content from real-world imagery.

To make this solution accessible to a broader audience, the trained ViT model is deployed through a user-friendly web application. This interface allows users—regardless of technical expertise—to upload images and receive immediate feedback on their authenticity. The backend integrates the Vision Transformer’s classification pipeline, ensuring efficient inference and intuitive results presentation. The integration of ViTs into this system enhances its transparency, scalability, and adaptability, aligning with modern expectations for real-time, AI-powered tools in public-facing applications.

Beyond its technical achievements, the proposed system addresses a growing societal need for digital trust and accountability. In an age where synthetic media can influence public opinion, manipulate perception, and even disrupt democratic processes, having reliable detection mechanisms is essential. This work contributes to safeguarding the integrity of digital content, promotes responsible AI usage, and supports ethical innovation by offering a practical, accurate, and scalable solution for image authenticity verification. It empowers individuals, organizations, and platforms to critically assess visual content, reinforcing credibility and trust in an AI-dominated digital landscape.

II. Literature Review and Problem Statement

With the ever-increasing sophistication of AI-generated imagery, The advent of AI image generation here is detection and has really changed since then. They can offer so many ways of detection purposes. Precision Agriculture is one of them. Precision agriculture has greatly become a relevant tool to change pleasured practices, to better production, and to ensure sustainability due to new technological advances

One of the latest innovations in this space increases Explainable AI (XAI) in machine recommendation systems. Turgut et al. (2024) presented AgroXAI, an approach based on edge computing, to bring XAI techniques like SHAP and LIME. It provides clarity about the reasoning behind the recommendations made by the system using different XAI methods. With this provision of transparency in decision-making in these crop recommendation systems, farmers would understand the rationale for recommending specific crops [1].

Simultaneous to that, there has been promising progress in applying Long Short-Term Memory (LSTM) networks for prediction of future crop harvests, keeping at bay expected climate-related uncertainty. This has a study for the area of Maharashtra, which developed the LSTM-based crop recommendation and forecasting system, using data from the years 2001 to 2022. The LSTM-based recommendation and forecasting system was coupled with expectation-maximization techniques to predict the future-cultivatable crops according to what has transacted through the past weather facts in the recommendation and forecasting system of the states. This approach enables farmers to take data-supported decisions, so the access to better crop productivity is realized from the vagaries of the unpredictable weather pattern, which ultimately leads to a more stable ('secure') crop yield [2].

Real-time production crop recommendation systems using IoT and Machine Learning have been the center of much recent research, promising to give farmers a lot more useful recommendations than what they can manage in their own fields. Among many such studies found in the IJERT Journal, the one that used IoT sensors along with some machine learning algorithms like Random forest has been discussed. Such a system gathers data on soil moisture, temperature, and humidity in real-time handling the important parameters of the environment and recommends the best crops for such conditions. This integration promises to give access to sustainable agriculture through improvized efficient use of resources, minimized wastage, and increased productivity [3].

The application of IoT sensors has advanced beyond crop advisory systems to provide automated irrigation systems. One of the research propositions is an IoT sensor-based as well as ML algorithms irrigation system that recommends crops in addition to automation according to a real-time measurement of soil moisture level. This will, by relying on such sources of data, drastically reduce agricultural laborers while improving the efficiency of resource uses much more sustainable farming practices [4].

Blockchain technology has even landed a place in precision agriculture. Patel et al. (2023) further ensured the acquisition of a crop recommendation system using blockchain technology in the IoT environment. In this: Data integrity and security were provided through the blockchain, making the whole process trustworthy in precision farming decisions. Thus, not only was the crop recommendation based on this data authentic, it was also a safeguard against data fraud. Thus, farmers can rely on farm decisions based on accurate data that has true tamper-proof integrity [5].

Furthermore, soil health monitoring systems with fertilizer recommendations are now becoming core components of precision farming. In fact, one such IoT and ML-based system developed can monitor nutrient levels from soils NPK, pH and moisture content. This system provides crop-specific fertilizer recommendations that optimize use of fertilizer resource and thereby, enhance yield performance. Integrating such systems with real-time data from IoT sensors could also help farmers in much effective resource management, thereby improving crop production as well as sustainability [6].

There is also involvement of AI and XAI techniques using crop yield prediction models in precision agriculture. Jagan Mohan et al. (2025) will explore how AI and SHAP or LIME can be used to make precision crop yield prediction models. These interpretable models will provide farmers with insights into the roles of weather, soil and crop types in determining yield outcomes. The insights provided by these models will aid farmers in making better crop management decisions toward higher productivity and sustainable aspects [7].

The core problem addressed by this project is the detection of synthetic or AI-generated images that are visually indistinguishable from real ones. As generative models continue to improve, they increasingly produce high-resolution, contextually accurate, and photorealistic outputs. These outputs can be misused in ways that cause social, political, financial, and psychological harm.

The problem becomes more complex due to the variety of generative techniques, each with different patterns and noise characteristics. A model trained on one type of synthetic image may fail to detect others, unless it is properly generalized. Furthermore, human ability to distinguish such images is unreliable and subjective, especially at scale.

Thus, the problem statement is defined as, to design and develop an automated image authenticity detection system that can
classify images as real or AI-generated using deep learning-based binary classification techniques. The system must operate with a high degree of accuracy and generalization, be scalable and efficient in handling large volumes of images, perform reliably across diverse sources and content types. Solving this problem is vital for maintaining the integrity of visual information in online platforms, forensic investigations, and digital archiving.

III. Methodology

This section details the methodology employed to develop and evaluate a Vision Transformer (ViT) model for binary classification of images as authentic or fake. The approach encompasses dataset preparation, preprocessing, model architecture, training procedures, and evaluation metrics, ensuring a comprehensive framework for image authenticity detection.

3.1. Dataset

The dataset consists of 5,393 images, carefully curated to include both authentic and AI-generated samples across diverse categories such as human portraits, landscapes, and product photography. Some of the images from the dataset used for training the model are given in fig.3.1. This diversity enables the model to generalize across various image types encountered in real-world applications. The dataset is divided into three subsets: a training set of 3,964 images for learning discriminative features, a validation set of 714 images for monitoring training progress and tuning hyperparameters, and a test set of 714 images for final evaluation. Each subset contains subfolders for the two classes, "Authentic" and "Fake," facilitating binary classification. The balanced distribution of classes across subsets ensures unbiased training and evaluation, critical for achieving reliable performance metrics.

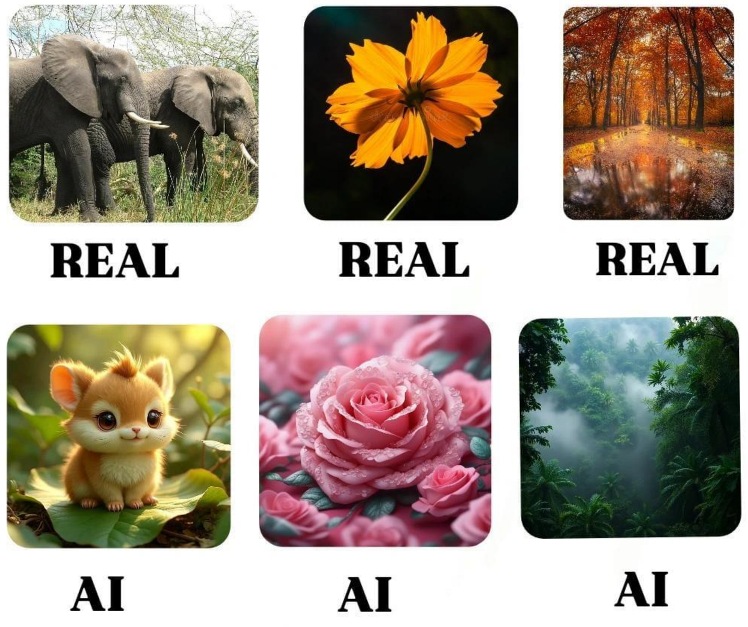

Fig. 3.1. Images from the dataset

3.2. Preprocessing

Preprocessing is essential to standardize input data and enhance model robustness. To achieve this, the following steps are applied with variations between training, validation, and test sets.

For the training set, extensive data augmentation is applied to improve generalization and prevent overfitting. This includes rescaling pixel values from [0, 255] to [0, 1], random rotation up to 15 degrees, width and height shifts up to 10%, shear transformation up to 10%, random zooming up to 10%, and horizontal flipping. Additionally, brightness adjustment within the range [0.9, 1.1] is applied to simulate different lighting conditions. The ‘nearest’ fill mode is used to fill in pixels created during transformations, ensuring smooth image boundaries. These augmentations are implemented in real-time during training, generating diverse variations of each image.

In contrast, validation and test images undergo only rescaling (division by 255) to maintain consistency with the training data without introducing artificial variability. This ensures that evaluation reflects the model’s true performance on unaltered data.

All images are resized to a uniform resolution of 256x256 pixels to meet the model’s input requirements. A data generator is used to load images from their respective directories, automatically assigning labels based on subfolder structure (“Authentic” or “Fake”). The training data is shuffled to randomize sample order, while validation and test data maintain a fixed order to ensure consistent evaluation.

These preprocessing steps create a robust pipeline that prepares the data for effective model training and evaluation, enabling the model to handle real-world variability and accurately classify images.

3.3. Model Architecture

The Vision Transformer (ViT) model is designed to leverage transformer-based processing for image classification, treating images as sequences of patches to capture both local and global features as shown in fig. 3.2. The architecture is implemented using TensorFlow and includes custom layers tailored for the task.

The process begins with patch extraction, where input images of size 256x256 pixels with 3 color channels (RGB) are divided into non-overlapping patches of size 16x16 pixels. This results in a grid of patches, which are flattened into a sequence of vectors. Each patch represents a small region of the image, enabling the model to process images as sequences, similar to tokens in natural language processing.

Each patch vector is then projected into a 256-dimensional embedding space using a dense layer, followed by the addition of positional embeddings to preserve spatial information. This step ensures that the model understands the relative positions of patches within the image, as transformers are inherently permutation-invariant.

The model includes 8 transformer blocks, each comprising multi-head self-attention with 8 attention heads, allowing the model to focus on different parts of the image simultaneously and capture diverse feature interactions. This mechanism is critical for detecting subtle artifacts in AI-generated images. Each transformer block also includes a feed-forward network (FFN) with a hidden dimension of 512 and GeLU activation, dropout with a rate of 0.1, and layer normalization to stabilize training.

After processing through the transformer blocks, the sequence of patch embeddings is aggregated into a single vector using global average pooling, reducing dimensionality and preparing the data for classification. The aggregated vector is then passed through two dense layers for further processing: the first with 512 units and ReLU activation, and the second with 256 units and ReLU activation, both followed by batch normalization and dropout with a rate of 0.5.

Finally, a final dense layer with a single unit and sigmoid activation produces a probability score between 0 and 1, indicating the likelihood of an image being "Authentic" or "Fake". The model's input shape is (256, 256, 3), accommodating resized RGB images. The architecture's use of transformers enables it to model long-range dependencies and global context, making it effective for detecting subtle differences between authentic and AI-generated images.

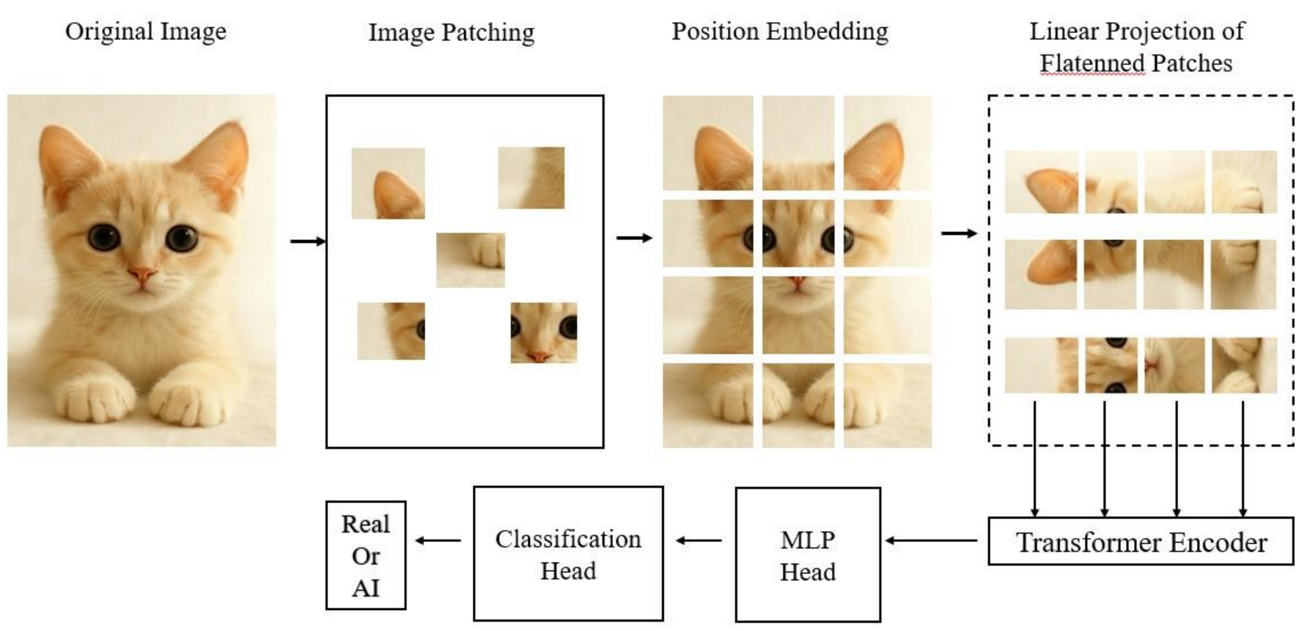

Fig. 3.2. ViT Architecture for Image Classification

3.4. Training Procedure

The training process is designed to optimize the model's performance while ensuring generalization to unseen data. The Adam optimizer is used with an initial learning rate of 0.0001, chosen for its adaptive learning rate properties and effectiveness in deep learning tasks. Binary cross-entropy is employed as the loss function, suitable for binary classification.

The model is monitored using key metrics: accuracy, precision, recall, and Area Under the Curve (AUC). Accuracy measures overall performance, precision minimizes false positives, recall minimizes false negatives, and AUC evaluates the model's ability to distinguish between classes.

The model is trained for up to 100 epochs with a batch size of 32, allowing sufficient time for convergence. Approximately 124 steps per epoch are calculated based on the training images, and 23 validation steps are calculated based on the validation images. To enhance training efficiency and prevent overfitting, callbacks such as EarlyStopping, ModelCheckpoint, and ReduceLROnPlateau are implemented.

EarlyStopping monitors validation loss and stops training if no improvement is observed for 10 consecutive epochs, restoring the best-performing epoch's weights. ModelCheckpoint saves the model with the highest validation accuracy, and ReduceLROnPlateau adaptively adjusts the learning rate when progress stalls.

Random seeds are set for reproducibility, ensuring consistent results across training runs. This training procedure enables the model to learn effectively while maintaining robust generalization through validation monitoring and adaptive optimization techniques.

3.5. Evaluation Metrics

The model's performance is rigorously evaluated on the test set, demonstrating its effectiveness in detecting image authenticity. The model achieves a test accuracy of 97.06%, correctly classifying 97.06% of the test images as either Authentic or Fake. Additionally, it achieves a precision of 96.15%, indicating that most images predicted as Authentic are indeed Authentic.

The model also exhibits a high recall of 98.04%, correctly identifying 98.04% of actual Authentic images and minimizing false negatives. The F1-score, calculated as the harmonic mean of precision and recall, is 97.09%, providing a balanced measure of the model's performance. Furthermore, the area under the Receiver Operating Characteristic (ROC) curve is 99.46%, demonstrating the model's excellent ability to distinguish between Authentic and Fake images.

A detailed evaluation using a confusion matrix reveals that out of 357 Authentic images, 350 are correctly classified, with 7 misclassified as Fake. Out of 357 Fake images, 344 are correctly classified, with 13 misclassified as Authentic. This balanced error distribution confirms the model's robustness. A classification report provides further insights into class-specific performance, while the ROC curve illustrates the trade-off between true positive rate and false positive rate, with the high AUC score confirming strong discriminative power.

IV. Results

The Vision Transformer (ViT) model exhibits outstanding performance on the test set, achieving high metrics across multiple evaluation criteria, which validate its effectiveness in distinguishing authentic images from AI-generated ones. The key performance metrics—accuracy, — accuracy (97.06%), precision (96.15%), recall (98.04%), F1-score (97.09%), and AUC-ROC (99.46%)—demonstrate — demonstrate the model’s capability to accurately classify images with minimal errors. These metrics provide a comprehensive assessment of the model’s reliability and robustness, and their implications are further explored through detailed analyses of the confusion matrix, training and validation dynamics, and the Receiver Operating Characteristic (ROC) curve. The results confirm the ViT model’s suitability for real-world applications in image authenticity detection, where high accuracy and discriminative power are essential for maintaining trust in digital media manipulation.

4.1. Confusion Matrix

The confusion matrix, illustrated in Fig. 4.1., offers a detailed breakdown of the model's classification performance by comparing predicted labels against true labels for the 714 images in the test set. Out of 357 authentic images, 350 are correctly classified as authentic (true positives), while 7 are incorrectly labeled as fake (false negatives). Similarly, of the 357 fake images, 344 are correctly identified as fake (true negatives), with 13 misclassified as authentic (false positives). This balanced error distribution of errors—7 errors—7 false negatives and 13 false positives—indicates positives indicates that the model performs consistently across both classes without significant bias toward either authentic or fake images. The confusion matrix is a critical tool for understanding the model's strengths and weaknesses, as it quantifies the exact number of correct and incorrect predictions. The relatively low number of misclassifications (20 out of 714, or approximately 2.8%) underscores the model's ability to differentiate between authentic and AI-generated images, which is vital for applications such as media verification, where even small errors can have significant consequences. The visual representation of the confusion matrix (Fig. 4.1.) aids in interpreting these results, providing a clear snapshot of the model's classification accuracy and potential areas for improvement, such as reducing false positives for fake images.

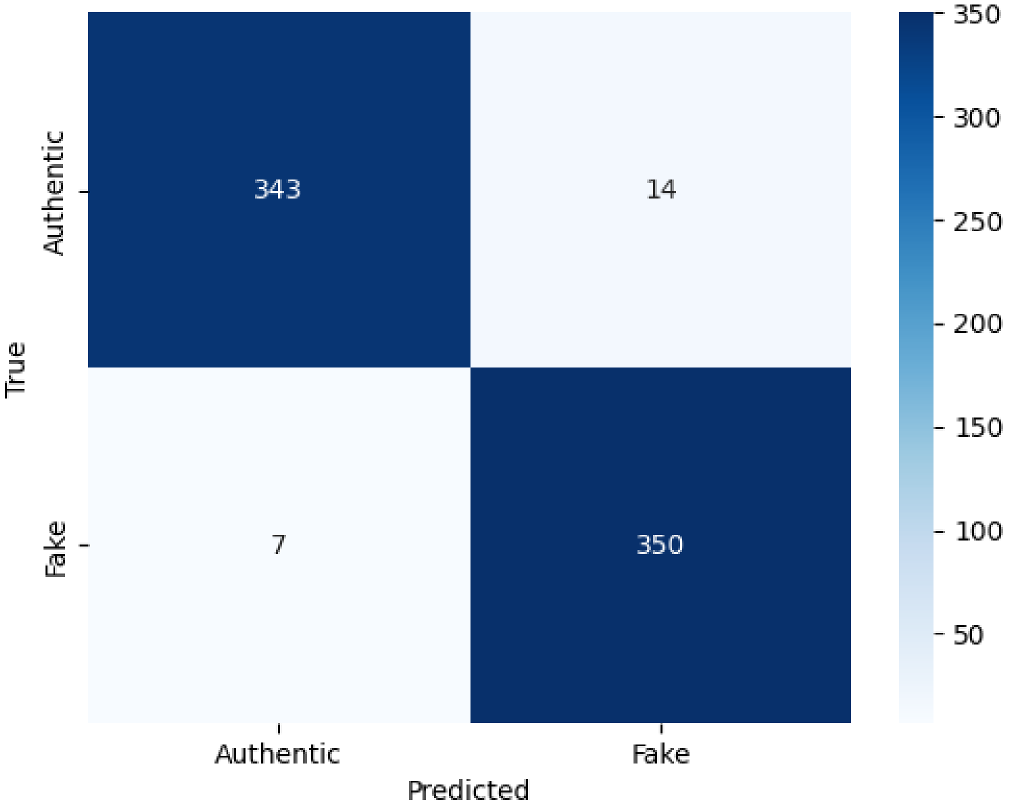

Fig. 4.2. Confusion matrix

4.2. Training and Validation Dynamics

The training and validation dynamics, depicted in Fig. 4.2, illustrate the model's learning behavior over the training epochs, providing insights into its convergence and generalization capabilities. The training accuracy stabilizes at approximately 98%, indicating that the model effectively learns to classify the training data. The validation accuracy peaks at around 97%, closely aligning with the training accuracy, which suggests that the model generalizes well to unseen data. The training loss converges to around 0.05, reflecting a low error rate during training, while the validation loss reaches approximately 4, indicating stable performance on the validation set. The small gap between training and validation accuracy (1%) and loss (0.02) suggests minimal overfitting, a positive outcome indicating that the model has not memorized the training data, but instead learned generalizable features. These dynamics are visualized in Fig. 4.2. through accuracy and loss curves, which plot the metrics over epochs, offering a clear view of the training process's stability. The use of regularization techniques, such as dropout, and callbacks like EarlyStopping, likely contributed to this balance, ensuring the model's robustness. This stability is crucial for real-world deployment, where the model must handle diverse image types without performance degradation.

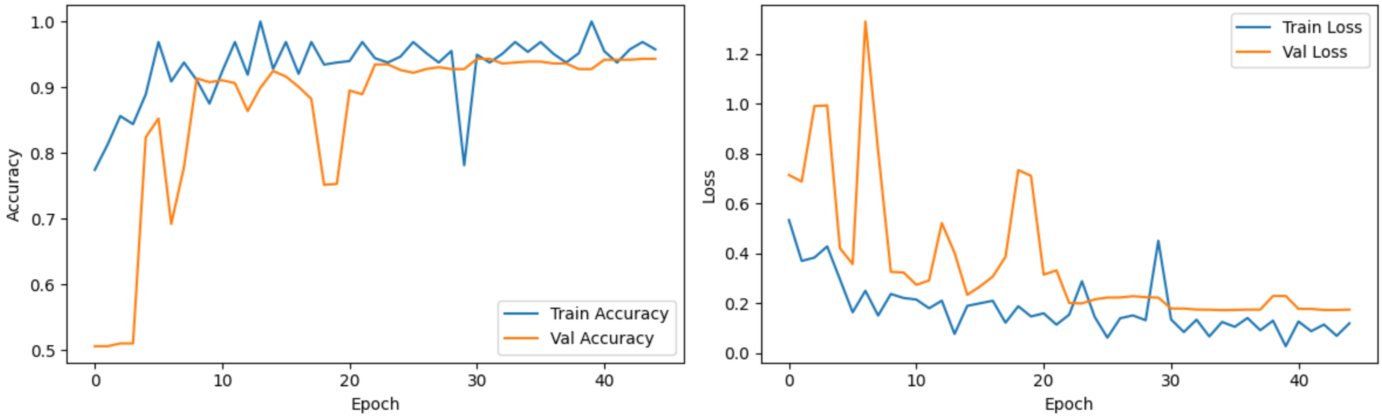

Fig. 4.3. Training and validation accuracy and loss curves

4.3. Receiver Operating Characteristic (ROC) Curve

The Receiver Operating Characteristic (ROC) curve, shown in Fig. 4.3., evaluates the model’s ability to discriminate between authentic and fake images across various classification thresholds. The curve plots the true positive rate (TPR, or recall, sensitivity) against the false positive rate (FPR) at different probability thresholds, with an Area Under the Curve (AUC) of 99.46%. The TPR represents the proportion of authentic images correctly identified, while the FPR indicates the proportion of fake images incorrectly classified as authentic. An AUC of 99.46% signifies exceptional discriminative power, meaning the model can nearly perfectly separate the two classes, with minimal overlap in their predicted scores. This high AUC is particularly significant for image authenticity detection, where distinguishing subtle differences between real and AI-generated images is challenging. The ROC curve (Fig. 4.3.) visually confirms this performance, showing a curve that closely hugs the top-left corner, indicating high TPR with low FPR across thresholds. This result highlights the model’s model’s suitability for applications requiring high confidence in classification, such as forensic analysis or digital media verification, where false positives and negatives must be minimized.

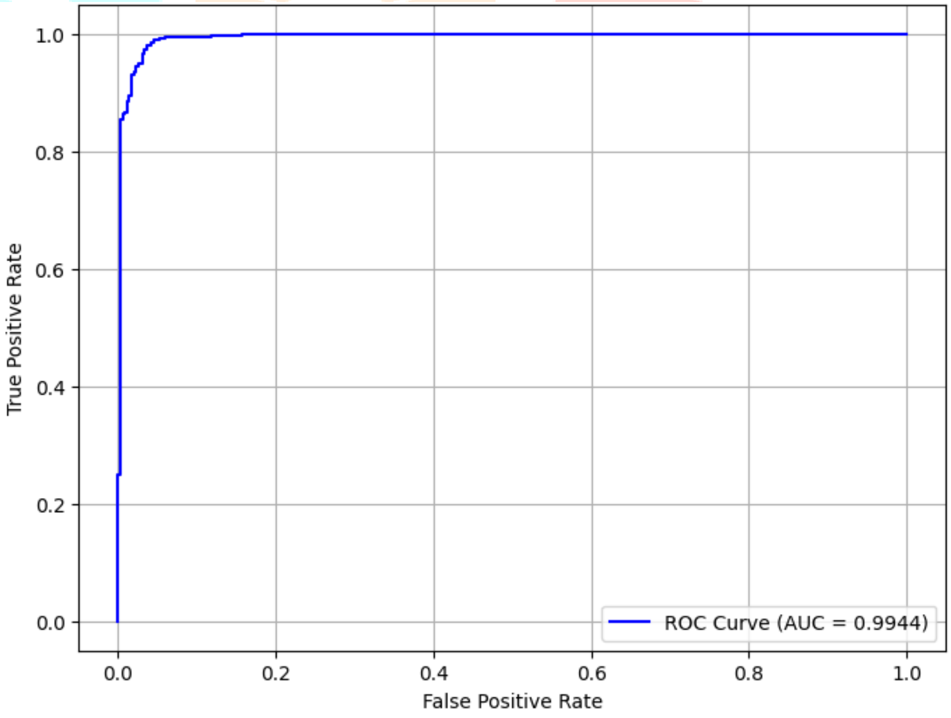

Fig. 4.3. ROC curve

V. Discussion

The Vision Transformer model’s high accuracy (97.06%) and AUC (99.46%) underscore its effectiveness in detecting image authenticity. The use of self-attention mechanisms enables the model to capture global context and long-range dependencies, outperforming traditional CNNs in identifying subtle artifacts in AI-generated images. Data augmentation plays a crucial role in enhancing generalization, allowing the model to handle diverse image variations.

However, the dataset size (5,393 images) could be expanded to include more challenging samples, such as those generated by advanced AI models, to further improve robustness. Incorporating explainability techniques, such as Grad-CAM, could enhance transparency by visualizing the model's decision-making process. Future work may explore hybrid CNN-ViT architectures or deployment on edge devices for real-time applications, aligning with trends in explainable AI and efficient computing.

VI. Conclusion

This paper presents a Vision Transformer-based approach for image authenticity detection, achieving a test accuracy of 97.06% and an AUC of 99.46% on a curated dataset. By leveraging transformer architectures, this work provides a robust solution for verifying digital visual content, contributing to trust and credibility in the era of AI-generated media. Future enhancements could include larger datasets and explainability features to further advance the field.

References

[1] Zicong Hu, Jian Cao, Weichen Xu, Ruilong Ren, Tianhao Fu, Xinxin Xu and Xing Zhang, "EMPIRICAL RESEARCH ON QUANTIZATION FOR 3D MULTI-MODAL VIT MODELS," 2024 IEEE International Conference on Image Processing (ICIP), vol. 1, no. 1, pp. 3606-3616, Oct. 2024.

[2] Rachid Bousaid, Mohamed EL Hajji and Youssef ES-SAADY, "Facial Emotions Recognition Using Vit and Transfer Learning," 2022 5th International Conference on Advanced Communication Technologies and Networking (CommNet), vol. 1, no. 1, pp. 1-6, Dec. 2022.

[3] Renhe Zhang, Qian Zhang and Guixu Zhang, "SDSC-UNet: Dual Skip Connection ViT-Based U-Shaped Model for Building Extraction," IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, vol. 20, no. 1, pp. 1-5, Apr. 2023.

[4] Reza Akbarian Bafghi, Nidhin Harilal, Claire Monteleoni and Maziar Raissi, "Parameter Efficient Fine-tuning of Self-supervised ViTs without Catastrophic Forgetting," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), vol. 1, no. 1, pp. 3679-3684, June. 2024.

[5] J.Benita and M.Vijay, "Implementation of Vision Transformers (ViTs) based Advanced Iris Image Analysis for NonInvasive Detection of Diabetic Conditions," Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS-2025), vol. 1, no. 1, pp. 1451-1457, March. 2025.

[6] Pi-Chuan Chen and Tzi-Dar Chiueh, "TPC-NAS for ViTs: A Systematic Approach to Improve Vision Transformer Performance with Total Path Count," 2024 International Joint Conference on Neural Networks (IJCNN), vol. 1, no. 1, pp. 1-7, June. 2024.

[7] Minhao Ding, Guangxin Dongye, Ping Lv and Yipeng Ding, "FML-Vit: A Lightweight Vision Transformer Algorithm for Human Activity Recognition Using FMCW Radar," IEEE SENSORS JOURNAL, vol. 24, no. 22, pp. 38518-38526, Nov. 2024.

[8] Qiulong Yu, Zhiqiang Wang, Lei Ju, Sicheng Yuan and Ying Zhang, "Android Malware Detection Technology Based on SC-ViT and Multi-Feature Fusion," 2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), vol. 1, no. 1, pp. 1105-1114, Dec. 2024

[9] ArunaDevi Karuppasamy, "Recent ViT based models for Breast Cancer Histopathology Image Classification," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), vol. 1, no. 1, pp. 1-5, July. 2023.

[10] Tong Su, Shuo Ye, Chengqun Song and Jun Cheng," MASK-VIT: AN OBJECT MASK EMBEDDING IN VISION TRANSFORMER FOR FINE-GRAINED VISUAL CLASSIFICATION," 2022 IEEE International Conference on Image Processing (ICIP), vol. 1, no. 1, pp. 1626-1630, Oct. 2024.