Explainable Detection and Classification of AI-Generated Images Using Deep Learning Techniques

MAADA BHAVANI [1], Mr FIROZ[2]

[1]PG Scholar, Dept. of AI&ML, St. Marys Group of Institutions Guntur for Women, Guntur.

[2]Asst Professor, Dept. of AI&ML, St. Marys Group of Institutions Guntur for Women, Guntur.

Abstract—: The rapid advancement of artificial intelligence has led to the widespread creation of synthetic and manipulated images, raising significant concerns regarding digital media authenticity and reliability. The increasing availability of AI-generated content poses challenges for journalism, social media platforms, digital forensics, and legal investigations, where image credibility is crucial. Conventional image authentication methods often exhibit limited effectiveness in identifying sophisticated manipulations produced by modern generative models. To address this issue, this study presents an intelligent image forensic framework for the detection and classification of real and AI-generated images. The proposed approach employs a Vision Transformer (ViT) model to extract discriminative spatial and contextual features from input images. These deep features are subsequently processed using an Extreme Gradient Boosting (XG Boost) classifier to achieve accurate classification performance. Furthermore, explainable artificial intelligence (XAI) techniques are integrated to enhance transparency by providing visual interpretations of the model's predictions and highlighting influential image regions. The proposed framework improves both detection accuracy and interpretability, making it suitable for practical deployment in content verification systems. Experimental results demonstrate that the hybrid ViT–XG Boost model effectively distinguishes authentic images from manipulated content, outperforming conventional approaches. The developed system contributes to strengthening digital trust and supports reliable media verification in the era of AI-generated visual content.

***Key Words**:* Artificial Intelligence (AI), Image Forgery Detection, Vision Transformer (ViT), XG Boost, Deep Learning, Explainable Artificial Intelligence (XAI), Synthetic Image Detection, Digital Image Forensics.

1.Introduction

The rapid evolution of Artificial Intelligence (AI) and deep generative models has revolutionized digital content creation, enabling the generation of highly realistic synthetic images. Advanced techniques such as Generative Adversarial Networks (GANs) and diffusion-based models can produce visual content that closely resembles authentic photographs. Although these technologies offer significant benefits in areas such as entertainment, design, and media production, they also introduce serious challenges related to misinformation, digital forgery, identity manipulation, and cybersecurity. Consequently, the development of reliable methods for distinguishing real images from AI-generated content has become a critical research problem in digital forensics and content verification.

Traditional image forensic techniques often struggle to identify sophisticated synthetic images because modern generative models effectively mimic natural image characteristics. Recent advances in deep learning have demonstrated promising results for image authenticity verification by automatically learning discriminative visual patterns. In particular, Vision Transformers (ViTs) have gained considerable attention due to their ability to capture long-range dependencies and global contextual information within images. These capabilities make transformer-based architectures highly effective for detecting subtle artifacts and inconsistencies present in AI-generated content.

This work proposes a hybrid framework that combines Vision Transformer-based feature extraction with the XG Boost classifier for accurate detection of synthetic images. To improve transparency and user trust, Explainable Artificial Intelligence (XAI) techniques are incorporated to provide visual interpretations of model predictions and highlight influential image regions. The proposed approach aims to enhance both detection performance and interpretability, contributing to the development of robust and trustworthy AI-driven image forensic systems for secure digital media verification.

2.Literature Survey

Recent studies have utilized deep learning models, particularly CNNs, to detect AI-generated images by identifying subtle artifacts and visual inconsistencies. To improve model transparency and user trust, Explainable AI (XAI) techniques such as SHAP, LIME, and Grad-CAM are increasingly integrated to provide interpretable explanations for classification decisions.

Mansoor et al. [1] addressed the challenge of interpretability in deep learning-based forensic systems by incorporating network dissection techniques with CNN architectures such as ResNet-50, InceptionV3, and VGG-16. Their study analyzed internal feature representations and associated them with human-understandable visual concepts, providing valuable insights into the decision-making process of image classification models. The results demonstrated that explainable AI methods can significantly enhance the reliability and practical adoption of AI-generated image detection systems.

Bale et al. [2] presented a comprehensive survey of deep fake detection approaches, including machine learning and CNN-based techniques. The review highlighted key challenges such as rapidly evolving generative models, dataset limitations, and poor cross-domain generalization. The authors concluded that while CNN-based methods remain effective for synthetic image detection, future research should focus on improving robustness, explain ability, and adaptability to emerging AI-generated media.

3.Proposed System

The purpose of this thesis is to develop an effective framework for detecting AI-generated images using advanced deep learning and machine learning techniques. The proposed system addresses the growing challenge of verifying digital image authenticity in areas such as digital forensics, social media, and cybersecurity. A Vision Transformer (ViT) is employed to extract rich visual features from images, while XGBoost is used for accurate classification of real and synthetic content. To enhance transparency, Explainable Artificial Intelligence (XAI) techniques are incorporated to provide visual explanations of model predictions. The framework aims to achieve high detection accuracy, robustness, and interpretability. By combining transformer-based feature extraction with ensemble learning and explain ability, the system improves trust in automated image verification. The research contributes to reliable digital media authentication and supports efforts to combat misinformation and image manipulation.

4.System Architecture

The proposed AI-generated image detection system consists of image pre-processing, feature extraction, classification, and explanation modules. Input images are resized and normalized before being processed by a Vision Transformer (ViT) to extract discriminative features.

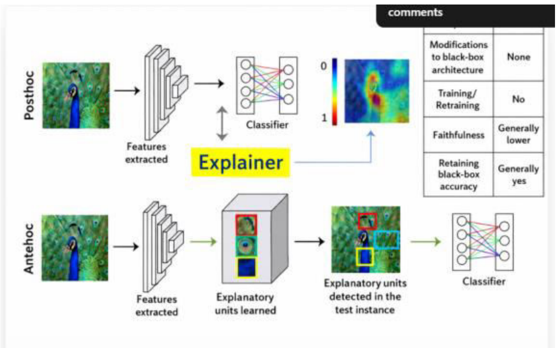

Fig 4.1: System Architecture

These features are classified using the XGBoost algorithm to identify whether an image is real or AI-generated. Explainable AI techniques are then applied to highlight important image regions influencing the prediction. The system provides both accurate classification results and interpretable visual explanations, ensuring reliability and transparency.

5.Methodology

The Vision Transformer is used as a feature extraction model that divides input images into patches and processes them through self-attention mechanisms to learn both local and global visual representations. The extracted high-level feature vectors capture complex image patterns and dependencies, providing rich information for accurate classification of real and AI-generated images.

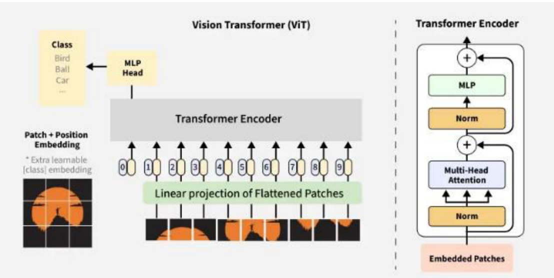

Fig 5.2: Generalised diagram for ViT

XGBoost is employed to classify the feature vectors extracted by the Vision Transformer (ViT). It utilizes an ensemble of gradient-boosted decision trees to learn complex patterns and effectively distinguish between real and AI-generated images. The combination of ViT features and XGBoost improves classification accuracy, robustness, and generalization while reducing the risk of overfitting through regularization techniques.

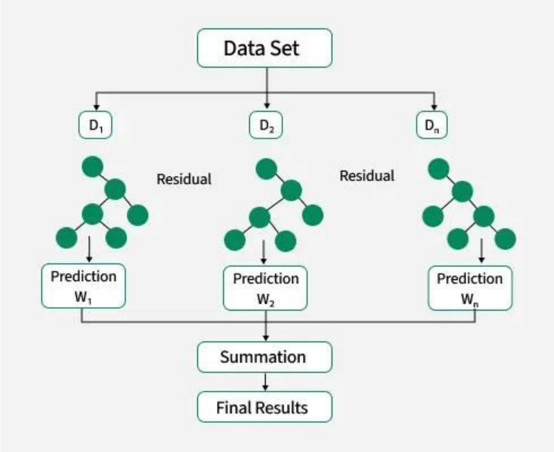

Fig 5.3: XG boost features extraction

Additionally, the system leverages explainable AI tools like SHAP to visualize the contribution of each feature in the prediction process, providing interpretability. This thorough evaluation ensures that the XG Boost classifier is not only accurate but also reliable and explainable in distinguishing synthetic images from authentic ones.

6. Design and construction

The proposed system is designed using a Vision Transformer (ViT) for feature extraction and XGBoost for image classification. The framework is constructed with modules for image pre-processing, feature extraction, classification, and explainable AI analysis, ensuring accurate detection of AI-generated images with interpretable results.

i). Image Dataset Collection Module: This module collects and organizes real and AI-generated images from various datasets. The images are labelled and stored for training and testing purposes.

ii). Image Pre-processing Module: This module performs resizing, normalization, and augmentation to prepare images in a consistent format and improve model performance.

iii). ViT-XG Boost Classification Module: The Vision Transformer extracts deep image features, and XG Boost classifies the images as real or AI-generated based on these features.

iv). Explainable AI (XAI) Module: This module provides visual explanations of model predictions using XAI techniques, highlighting important regions influencing the classification.

v). Performance Evaluation Module: This module evaluates the system using metrics such as accuracy, precision, recall, F1-score, and confusion matrix to measure detection performance.

The performance of the XGBoost model is evaluated using metrics such as accuracy, precision, recall, F1-score, sensitivity, and specificity. A confusion matrix is generated to show correct and incorrect classifications between AI-generated and real images. ROC curves and AUC scores can also be calculated to assess the model's discriminative ability.

7.Results and Discussion

The experimental results show that the proposed Vision Transformer (ViT) with XG Boost outperforms the existing DenseNet121 with Perceptron model in detecting AI-generated images. By capturing global contextual features and long-range dependencies, the proposed approach achieves higher accuracy, better generalization, and improved robustness. Furthermore, Explainable AI techniques enhance transparency by providing visual insights into the model's predictions.

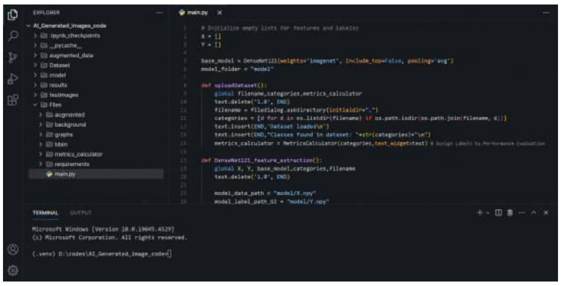

Fig 7.1: Deployments of The Project

Fig. 7.1 shows the Py Charm-based project development environment organized with a Python virtual environment and modules for dataset processing, model training, testing, and result evaluation.

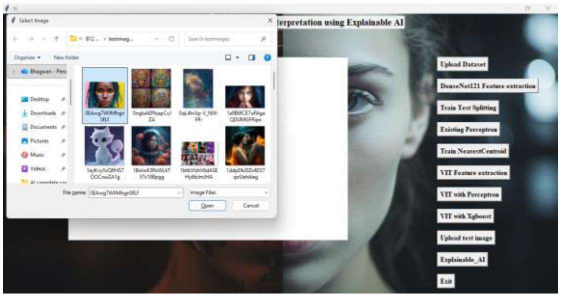

Fig 7.2: Uploading Test Image for Prediction

Fig 7.2 illustrates the process of uploading a test image for prediction through the graphical user interface (GUI). When the user clicks the Upload Test Image button, a file selection dialog is displayed, allowing the user to browse and choose an image from the test images folder

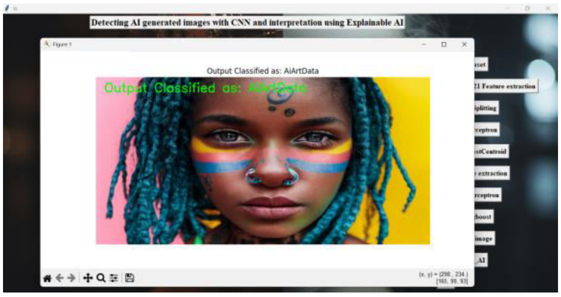

Fig 7.3: AI-Generated Image Classification Result

Fig. 7.3 shows the prediction result of the proposed XG Boost–ViT model for a test image. The model successfully classifies the uploaded image as Ai Art Data, demonstrating its effectiveness in accurately detecting AI-generated images through a user-friendly interface.

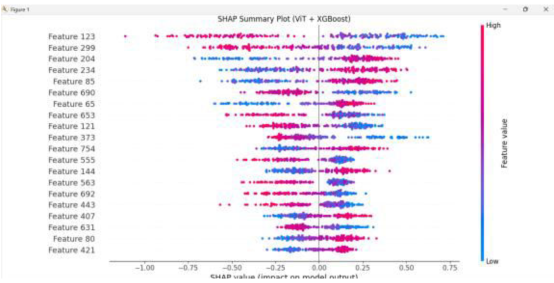

Fig 7.4: SHAP Summary Proposed XGBoost–ViT Model

Fig. 6.4 illustrates the SHAP summary plot for the proposed XG Boost–ViT model, highlighting the most influential features contributing to image classification. The plot shows how individual feature values impact model predictions, improving the interpretability and transparency of the classification process

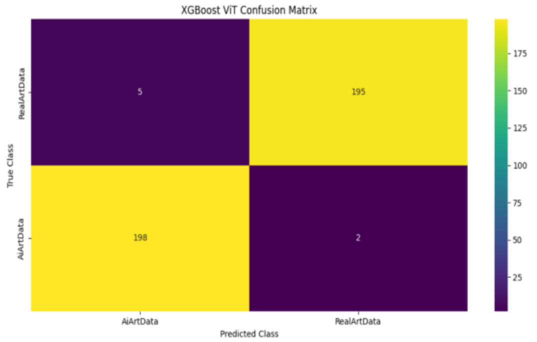

Fig 7.5: Confusion Matrix of the Proposed XG Boost–ViT Model

Fig 6.5 illustrates the confusion matrix of the proposed XG Boost–ViT model, showing that most AI-generated and real images are correctly classified with very few misclassifications. The results confirm the model's high accuracy **(98.25%),** robustness, and effectiveness in distinguishing AI-generated images from real images.

8.Conclusion and Future Scope

This research presents an effective framework for detecting AI-generated images using Vision Transformer (ViT) and XGBoost. The proposed model successfully captures complex visual patterns and distinguishes real images from synthetic content with high accuracy. By combining transformer-based feature extraction with ensemble classification, the system achieves improved robustness and classification performance. The integration of Explainable Artificial Intelligence (XAI) techniques provides transparency by highlighting the factors influencing model predictions. Experimental results demonstrate the reliability and effectiveness of the proposed approach for digital image forensics. Overall, the framework contributes to trustworthy and accurate AI-generated image detection in modern digital environments.

Future Enhancement: Future work can focus on real-time detection, multimodal analysis, and larger diverse datasets to improve robustness and generalization. Lightweight implementations can also enable deployment on mobile and edge devices for practical applications.

References

[1] H. Bayar and M. C. Stamm, "A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer," in *Proc. ACM Workshop Information Hiding and Multimedia Security*, Portland, OR, USA, 2016, pp. 5–10.

[2] J. Fridrich and J. Kodovský, "Rich Models for Steganalysis of Digital Images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 868–882, Jun. 2012.

[3] S. Nataraj, F. Bunyak, B. S. Manjunath, and T. E. Boult, "Detecting GAN-Generated Fake Images Using Co-Occurrence Matrices," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, Taipei, Taiwan, 2019, pp. 1547–1551.

[4] Y. Li and S. Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artifacts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Long Beach, CA, USA, 2019, pp. 46–52.

[5] A. Cozzolino, D. Gragnaniello, and L. Verdoliva, "Image Forgery Localization Through the Fusion of CNNs," in *Proc. IEEE Int. Workshop Information Forensics and Security (WIFS)*, Hong Kong, China, 2017, pp. 1–6.

[6] R. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, Seoul, South Korea, 2019, pp. 1–11.

[7] A. Dosovitskiy et al., "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021, pp. 1–21.

[8] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, Montreal, Canada, 2021, pp. 10012–10022.

[9] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794.

[10] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model

Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1135–1144.

[12] N. Mansoor, F. Baig, and M. I. Azam, "Explainable CNN Models for DeepFake Detection Using Network Dissection," *Applied Sciences*, vol. 15, no. 2, p. 725, Jan. 2025.

[13] D. L. T. Bale, L. C. Ochei, and C. Ugwu, "A Survey on Deepfake Detection Techniques: From Traditional to Deep Learning Methods," *International Journal of Innovative Information Systems*, vol. 13, no. 2, pp. 11–28, Apr. 2024.

[14] A. L. Lađević, "Detection of AI-Generated Synthetic Images Using Deep Convolutional Features," *Applied Sciences*, vol. 12, no. 3, pp. 1–18, 2022.