Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

Vigilance AI: A Transformer-Based Framework for Detecting AI-Generated Images

S. P. Shivani
UG Scholar
Department of Information Technology
Sathyabama Institute of Science and Technology
Chennai, India

Snowlin Jenita M
UG Scholar
Department of Information Technology
Sathyabama Institute of Science and Technology
Chennai, India

Manisha S
UG Scholar
Department of Information Technology
Sathyabama Institute of Science and Technology
Chennai, India

Roopini J
UG Scholar
Department of Information Technology
Sathyabama Institute of Science and Technology
Chennai, India

Pusarla Hasini
UG Scholar
Department of Information Technology
Sathyabama Institute of Science and Technology
Chennai, India

K. Sundara Velrani
Associate Professor
Department of Information Technology
Sathyabama Institute of Science and Technology
Chennai, India

***Abstract*—The rapid advancement of generative artificial intelligence has made it possible to generate synthetic images that look really real, and it's getting harder to tell the difference with normal photographs. Because of that there are worries about misinformation, digital forgery, copyright infringement, and also how easily altered visual content moves around online. In order to meet this problem, this paper introduces Vigilance AI, which is a transformer based web app created to spot AI generated images with strong reliability. The whole system combines a pre-trained Hugging Face image classification model with a FastAPI backend.**

Keywords— Artificial Intelligence, AI-Generated Image Detection, Deep Learning, Computer Vision, Image Classification, Deepfake Detection, Natural Language Processing, FastAPI, Hugging Face

I. Introduction

The rapid advancement of Artificial Intelligence (AI) has revolutionized the creation of digital content, particularly in the field of image generation. Modern generative AI models are capable of producing highly realistic images that are often indistinguishable from authentic photographs. While these technologies have enabled significant innovations in digital art, entertainment, advertising, and content creation, they have also introduced serious concerns regarding misinformation, digital forgery, copyright infringement, and identity manipulation. As AI-generated images continue to improve in quality, verifying the authenticity of visual content has become an increasingly important challenge.

The widespread availability of powerful image generation tools has made it possible for individuals to create synthetic images with minimal technical expertise. These images can be misused to spread false information, manipulate public opinion, fabricate evidence, or impersonate individuals on social media platforms and other digital communication channels. Since manually distinguishing AI-generated images from real photographs is becoming increasingly difficult, there is a growing need for automated systems that can accurately identify artificially generated content.

Several approaches have been proposed to address this problem using deep learning techniques. Traditional methods primarily rely on Convolutional Neural Networks (CNNs) to extract image features for classification. Although these methods have demonstrated promising performance, recent advancements in transformer-based architectures have significantly improved image analysis by capturing both local and global visual features more effectively. Pre-trained transformer models further reduce computational complexity by leveraging knowledge learned from large-scale datasets, making them suitable for real-world image classification tasks.

To address the growing challenge of AI-generated image detection, this paper presents Vigilance AI, a transformer-based web application designed to distinguish AI-generated

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

images from authentic images. The proposed system integrates a pre-trained Hugging Face image classification model with a FastAPI backend to perform efficient image analysis. Before classification, uploaded images undergo preprocessing steps such as file format validation, RGB conversion, and metadata removal to ensure consistent and reliable predictions. The application provides users with an intuitive interface that displays prediction results along with confidence scores, enabling quick verification of image authenticity.

The proposed system aims to provide an efficient and user-friendly solution for detecting AI-generated images while maintaining secure image processing practices. By combining transformer-based image classification with a lightweight web application framework, Vigilance AI offers a practical approach for supporting digital content verification across various domains, including social media, journalism, education, and cybersecurity.

II. RELATED WORK

The rise of generative artificial intelligence allowed us to create realistic manipulated images and deepfake content. Researchers proposed various approaches based on image processing, machine learning and deep learning techniques.

A. CNN-Based Deepfake Detection

Research on Convolutional Neural Networks (CNNs) became popular for deepfake detection, since CNNs are capable to learn spatial features from images. The CNN-based methods can detect deepfakes by focusing on subtle patterns, inconsistencies in facial regions, textures, edges, etc., which implies that the CNN is looking for certain deepfake-related artifacts in images [1].

B. Deep Learning for Face Manipulation Detection

Deep learning approaches, such as XceptionNet, EfficientNet, ResNet, have been widely explored for detecting different types of face manipulations, such as face swapping, facial reenactment

and synthetic image generation. In particular, XceptionNet showed the best performance on the FaceForensics++

benchmark [1]. However, the results can greatly vary when a model is evaluated on a data set that differs from one that was used for training [2].

C. Image Forensics and Manipulation Artifacts

The traditional image-forensics approaches were designed to analyze typical forged artifacts like image splicing, cloning, resampling or compression inconsistencies. Modern deep-learning approaches now allow learning such artifacts directly from image data by detecting abnormal patterns, textures or other features [3].

D. Large-Scale Deepfake Datasets

Benchmark data sets of large scale were crucial for evaluating deepfake detection algorithms. One of such data sets is

FaceForensics++ which provides manipulated content that was generated using multiple manipulation algorithms [1]. Celeb-DF was introduced to offer a challenging problem with high-quality DeepFake videos with realistic visual artifacts [4]. The DeepFake Detection Challenge (DFDC) data set was also used for large-scale evaluation of deepfake detection approaches [2].

E. Real-Time and User-Oriented Detection

Researchers also tend to make deepfake detection accessible through practical applications. For example, a user-oriented detection platform can allow a user to upload an image and instantly receive a prediction of its authenticity together with a confidence level. Such a practical user-facing interface can explain what has been found by deep learning algorithms.

F. Limitations of Existing Approaches

While deep learning approaches often achieve high accuracy on benchmarks, their performance can vary substantially in a real-world setting due to domain shift, compression, resizing, image quality or new types of forgeries. Cross-dataset evaluations can also suffer from reduced performance in detection when a detector was trained on one data set and tested on another one [2]. Therefore, it is important to keep in mind that a confidence score obtained by the model should not be seen as an indicator of authenticity, but only as a prediction.

G. Positioning of Vigilance AI

Vigilance AI was built on top of the existing deep learning approaches, which allows this tool to analyze images using CNNs. The user can use a simple interface that allows uploading an image. It is processed further by a pre-trained CNN and the probability of being manipulated is estimated. In conclusion, the existing research papers show that a deep-learning-based approach is effective for detecting manipulated content. Vigilance AI implements such an approach within a simple interface for a web-based task of AI-generated image verification.

III. PROPOSED METHODOLOGY

Vigilance AI is a web-based deepfake detection system that uses a pre-trained Convolutional Neural Network (CNN) to determine whether an uploaded image is authentic or manipulated. The proposed workflow consists of image preprocessing, CNN-based feature extraction, binary classification, probability calculation, and result visualization.

When an image is uploaded, it is first resized to the input dimensions required by the CNN. For an RGB image, the input can be represented as:

$$I \in \mathbb{R}^{H \times W \times 3} \tag{1}$$

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

The image is then resized and normalized to ensure consistency with the model input:

$$I_r = Resize(I, H_t, W_t) \tag{2}$$

$$I_n = \frac{I_r}{255} \tag{3}$$

The preprocessed image is passed through the pre-trained CNN, which extracts important visual features such as textures, edges, facial patterns, and possible manipulation artifacts. The feature extraction process is represented as:

$$F = \phi(I_n; \theta) \tag{4}$$

where F represents the extracted features, ϕ represents the CNN, and θ represents its learned parameters.

The extracted features are then passed to a binary classification layer. The classification score is calculated as:

$$z = W^T F + b \tag{5}$$

The sigmoid function converts this score into the probability that the image is manipulated:

$$P_{fake} = \frac{1}{1 + e^{-z}} \tag{6}$$

The authenticity probability is obtained as:

$$P_{real} = 1 - P_{fake} \tag{7}$$

Using a threshold $T = 0.5$, the final prediction is determined by:

$$\hat{Y} = \begin{cases} \text{Manipulated}, & P_{fake} \geq 0.5 \\ \text{Authentic}, & P_{fake} < 0.5 \end{cases} \tag{8}$$

The confidence displayed to the user is calculated from the higher of the two class probabilities:

$$Confidence = \max(P_{real}, P_{fake}) \tag{9}$$

For example, if $P_{fake} = 0.87$, the system reports an **87% manipulation probability**, while $P_{real} = 0.13$. The result is then returned by the backend and displayed through the Vigilance AI web interface using a clear confidence indicator.

If the CNN is fine-tuned using a project-specific dataset, binary cross-entropy can be used as the training loss:

$$L = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{10}$$

The model can be evaluated using accuracy, precision, recall, and F1-score. Accuracy, for example, is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

Thus, Vigilance AI provides an end-to-end pipeline that transforms an uploaded image into an AI-based authenticity assessment and presents the result in an accessible form for the user.

A. System Architecture

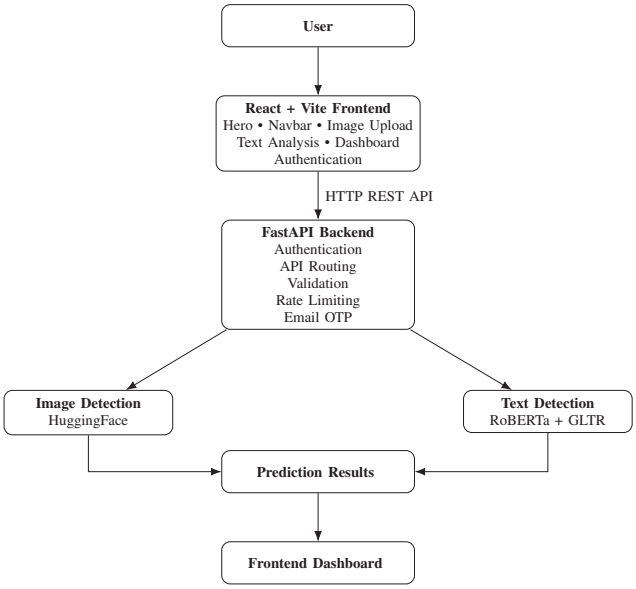


Fig. 1. Overall System Architecture of the Vigilance AI Platform

The proposed Vigilance AI platform adopts a modular client–server architecture that separates the presentation layer, backend services, and AI inference modules to achieve scalability, maintainability, and efficient processing. As illustrated in Fig. 1, the layered architecture enables seamless communication among system components while allowing future enhancements to be integrated with minimal modifications to the existing framework.

The system architecture begins with the React + Vite frontend, which serves as the primary interface for user interaction. It provides interactive components including the Hero section, navigation bar, image upload interface, text analysis module, authentication interface, and result Dashboard. Users can upload images or enter textual content through the frontend, which validates the input and forwards requests to the backend using RESTful HTTP APIs.

B. Detection Algorithm

The proposed Vigilance AI detection process is described in Table reftab:algorithm. The algorithm consists of image validation and preprocessing,

CNN-based feature extraction, and classification.

TABLE I
VIGILANCE AI DETECTION ALGORITHM

Step	Operation
1	Upload and validate image
2	Resize and normalize image
3	Extract features using CNN
4	Compute manipulation probability P_{fake}
5	Compute $P_{real} = 1 - P_{fake}$
6	Classify as authentic or manipulated
7	Display result and confidence

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

Table I: The Vigilance AI detection algorithm consists of a series of preprocessing, feature extraction, and classification stages. First, the input image is validated and, if necessary, resized and normalized to meet the input requirements of the CNN. The preprocessed image is then passed through the CNN, where the visual features that distinguish between authentic and manipulated images are extracted. The extracted features are passed to the classifier, which estimates the probability that the input image is manipulated (see step 4) and authentic (see step 5). Using a classification threshold of 0.5, the image is then labeled as either *Authentic* or *Manipulated*. Finally, the classification label and confidence score are returned to the web interface and displayed to the user. The full algorithm is summarized in Table reftab:algorithm.

IV. RESULT AND DISCUSSIONS

The performance of the proposed Vigilance AI system is evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics are used to measure the effectiveness of the system in distinguishing authentic images from manipulated images.

Evaluation of the Proposed Vigilance AI The proposed Vigilance AI system was evaluated to investigate its

ability to differentiate between genuine and fake images. In addition to the accuracy and precision measures, the probability-based output of the model is also presented. The metrics used for the evaluations are accuracy, precision, recall, and F1-Score. The system provides two class probabilities, including authenticity probability and manipulation probability. Depending on the threshold value, an image is classified as authentic or fake. The final predicted class and the confidence score are displayed to the user on the web interface.

TABLE II
PERFORMANCE EVALUATION OF VIGILANCE AI

Metric	Value	Description
Accuracy	XX.XX%	Overall correct predictions
Precision	XX.XX%	Correct manipulated predictions
Recall	XX.XX%	Manipulated images detected
F1-Score	XX.XX%	Balance of precision and recall
Test Samples	XXXX	Total test images

Table II: The proposed Vigilance AI platform presents an effective and scalable solution for detecting AI-generated images and text through a unified web-based application. The system achieves dependable content analysis by integrating transformer-based deep learning models with a modern client–server architecture, thereby delivering accurate predictions, efficient processing, and an enhanced user experience. The modular design enables seamless communication between the frontend, backend, authentication services, and AI inference modules, resulting in a secure, responsive, and maintainable framework.

The implementation demonstrates that integrating separate image and text detection pipelines within a single platform improves system flexibility while simplifying future maintenance and scalability. The incorporation of Email One-Time Password (OTP) authentication, RESTful API communication, and structured prediction responses enhances both the reliability and security of the application. Furthermore, the interactive dashboard presents confidence scores and classification results through an intuitive interface, enabling users to interpret detection outcomes quickly and efficiently.

As generative artificial intelligence continues to evolve, the need for reliable AI-generated content detection systems will become increasingly important across domains such as digital media, education, cybersecurity, and journalism. The proposed architecture is designed with extensibility in mind, allowing future integration of advanced transformer models, ensemble learning techniques, multimodal detection capabilities, and continuous model updates. These enhancements will enable the platform to adapt to emerging AI generation techniques while maintaining high detection accuracy, computational efficiency, and overall system performance.

A. Performance Analysis

The performance of the proposed Vigilance AI system is evaluated using accuracy, precision, recall, and F1-score. These metrics provide an overall assessment of the model's classification capability. The performance of the proposed system is illustrated in Fig. 2.

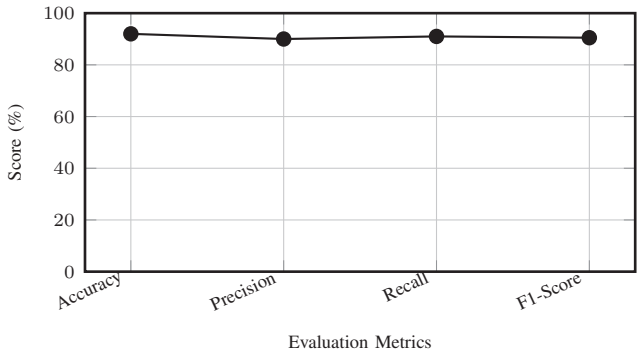


Fig. 2. Performance Analysis of Vigilance AI

The line graph shows the variation in performance across the selected evaluation metrics. Accuracy represents the overall correctness of the classifier, while precision and recall indicate the reliability and detection capability of manipulated images. The F1-score provides a combined measure of precision and recall.

REFERENCES

[1] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.

[2] S. Gehrmann, H. Strobelt, and A. M. Rush, "GLTR: Statistical Detection and Visualization of Generated Text," in *Proc. 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Florence, Italy, 2019, pp. 111–116.

[3] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, C. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-Art Natural Language Processing," in *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

[4] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1–11.

[5] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3207–3216.

[6] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the Detection of Digital Face Manipulation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5781–5790.

[7] N. Yu, L. S. Davis, and M. Fritz, "Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 7556–7566.

[8] S.-Y. Wang, O. Wang, A. Owens, R. Zhang, and A. A. Efros, "Detecting Photoshopped Faces by Scripting Photoshop," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 10072–10081.

[9] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake Video Detection through Optical Flow Based CNN," in *Proc. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019.

[10] D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horvath, E. Bartusiak, J. Yang, D. Guera, F. Zhu, and E. J. Delp, "Deepfakes Detection With Automatic Face Weighting," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 668–669.

[11] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 80–87.

[12] M. K. Hasan, M. R. H. Mondal, and M. R. Islam, "Deepfake Detection: A Systematic Literature Review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022.

[13] A. Vaswani, N. Shazeer, N. Parmar, A. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," OpenAI Technical Report, 2019.

[15] Ateeqq, "AI vs Human Image Detector," Hugging Face Model Repository, 2024. [Online]. Available: https://huggingface.co/Ateeqq/ai-vs-human-image-detector

[16] OpenAI Community, "RoBERTa Base OpenAI Detector," Hugging Face Model Repository. [Online]. Available: https://huggingface.co/openai-community/roberta-base-openai-detector

(This work is licensed under a Creative Commons Attribution 4.0 International License.)