A three-way Decision-based Synthetic Image Detection

Zeshan Khan[1,*], Zunaira Zainab[2]

[1]National University of Computer and Emerging Sciences, Islamabad, Pakistan
[2]University of Engineering and Technology Taxila, Pakistan

Abstract

The sophistication of AI-generated content initiating a new challenge of distinguishing AI-generated visuals from real ones. In this research a three-way decision system is proposed for synthetic image detection, combining deep learning and texture-based analysis to improve decision reliability. The proposed method begins with a fine-tuned convolutional neural network (CNN) predict if image is real or fake. When the model's confidence exceeds 80%, the decision is finalized directly. For intermediate confidence levels (20%–80%), a TreeNet-based texture feature classifier is engaged to refine the decision using handcrafted statistical features. In the final stage, a threshold optimization algorithm determines the most effective decision boundary by analyzing probability distributions from prior stages.

The evaluation of the methodology on the MediaEval 2025 dataset demonstrates that the proposed approach achieves an accuracy of 41.3%, precision of 42.4%, recall of 48.7%, and an F1-score of 45.4%, with an ROC-AUC of 0.3764 and average precision of 0.4223.

1. Introduction

The rapid advancement of GenAI has transformed the media generation and transformation. From Generative Adversarial Networks (GANs) [1], Variational Autoencoders (VAEs) [2], to diffusion models [3], the generated images are getting more and more realistic. The photorealistic generation of the content is increasing in difficulty to distinguish from real ones. While useful for creative applications, these technologies raise concerns about misinformation, impersonation, and media authenticity [4, 5], making detection of AI-generated content a critical challenge [6, 7].

The detectors work well to detect some specific types of generated media while often failing to generalize to unseen generators. The detection becomes more challenging after post-processing content, as subtle cues in texture, frequency, or noise may be lost due to compression or scaling [8, 9]. Achieving a balance between precision and recall is also challenging, since high sensitivity can cause excessive false positives [10]. Furthermore, neural models often lack interpretability, motivating hybrid and explainable detection approaches [6, 11, 12].

To address these issues, **MediaEval 2025 Challenge on Synthetic Images** [13], providing a benchmark of real and synthetic images under real-world perturbations such as filtering, scaling, and compression [13].

Inspired by this challenge, we propose a **three-way decision framework** that integrates deep learning confidence, texture-based reasoning [14], and adaptive thresholding [14]. A fine-tuned CNN provides initial predictions, uncertain samples are re-evaluated with a TreeNet

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online
*Corresponding author.
zeshankhanalvi@gmail.com (Z. Khan)
https://orcid.org/0000-0001-9034-3602 (Z. Khan); https://orcid.org/0009-0004-2325-6022 (Z. Zainab)

[15] texture classifier, and an adaptive thresholding algorithm refines decision boundaries for improved detection reliability.

2. Related Work

Deepfake and synthetic image detection approaches started with the focus on visual artifacts and inconsistencies [16, 17]. Later the domain shifted to exploit the phase inconsistencies or spectral energy distribution for better generalization [18, 19]. Transformer-based models [20, 11, 21] and lightweight CNNs [22] improved accuracy, but robustness against new generative pipelines remains limited.

Some of the studies focused on hybrid strategies combining statistical texture features and deep embeddings [6, 23, 24, 14]. Diffusion-specific detection models explored reconstruction residuals [25]. Other approaches emphasize frequency fingerprints and color-space features for generalization across architectures [9].

3. Methodology

The proposed methodology employs a three-way decision strategy for detecting synthetic images, combining deep learning inference, texture-based reasoning, and adaptive thresholding. The overall workflow is illustrated in Figure 1.

3.1. Stage 1: CNN-based Fine-Tuning and Initial Decision

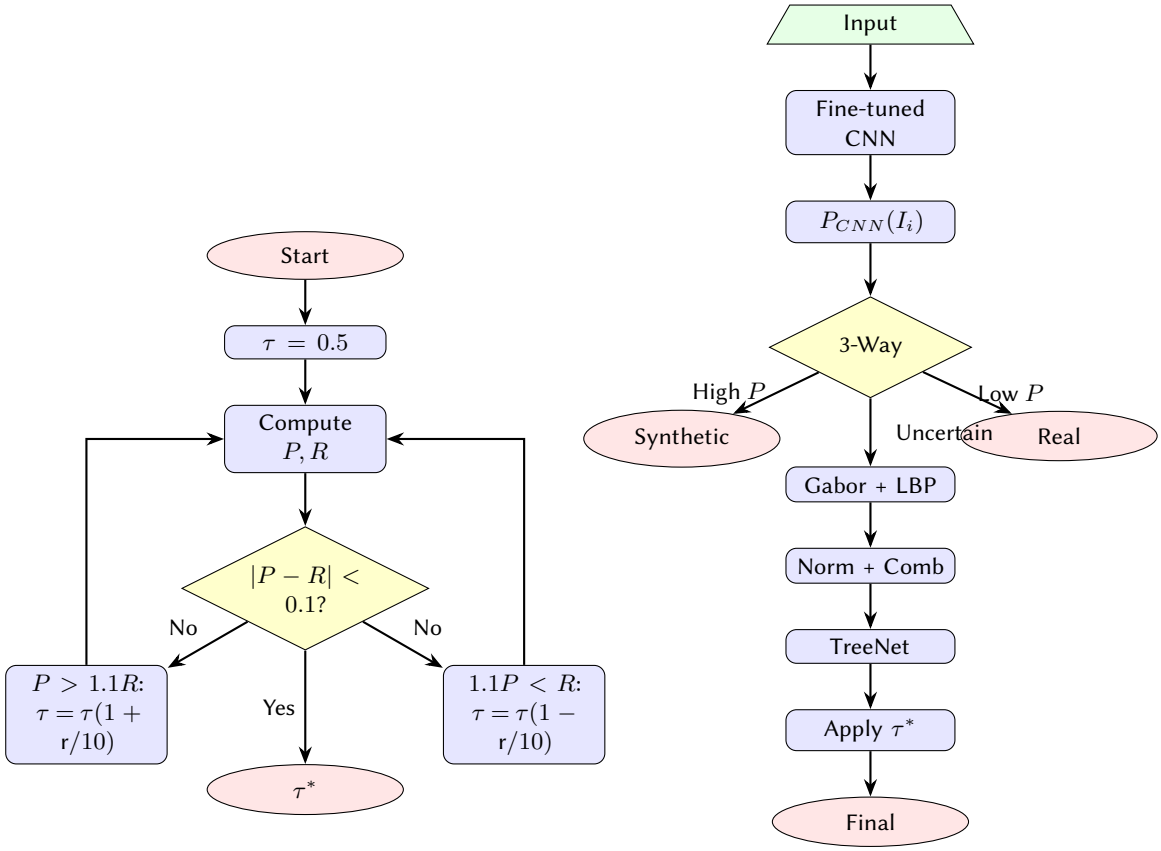
In the first stage, a Convolutional Neural Network (CNN) architecture adapted from Wang *et al.* [26] is fine-tuned on the MediaEval 2025 dataset to differentiate between real and generative images. Each image I is passed through the network to yield a probability score $P_{\text{CNN}}(I)$ representing the likelihood of being synthetic. Decisions with high confidence are immediately finalized: if $P_{\text{CNN}}(I) > 0.8$, the image is labeled as synthetic; if $P_{\text{CNN}}(I) < 0.2$, it is labeled as real. Samples falling in the uncertain range $(0.2 \leq P_{\text{CNN}}(I) \leq 0.8)$ are forwarded to the next stage. Based on validation analysis on 10% of the training set, predictions with CNN confidence above 0.8 or below 0.2 were correct about 61% of the time, whereas a 0.7/0.3 threshold band yielded only 48% accuracy. This empirical observation motivated the use of the 0.8/0.2 thresholds for defining high- and low-confidence regions.

3.2. Stage 2: Texture Feature Extraction and TreeNet Classification

For uncertain cases, the framework extracts interpretable texture features using Gabor filters and Local Binary Patterns (LBP). These features are normalized using a sigmoid function of $f_i' = \frac{1}{1+e^{-f_i}}$ and concatenated into a unified feature vector $\mathbf{F}(I) = [f_1', f_2', \ldots, f_n']$. The vector is then passed to a TreeNet classifier [15], configured with a depth of 3 and breadth of 2. TreeNet outputs a probability score $P_{\text{TreeNet}}(I)$ reflecting the synthetic likelihood based on texture cues.

3.3. Stage 3: Adaptive Threshold Optimization

The final decision stage applies a dynamic threshold selection algorithm (illustrated in Fig. 1a and Algorithm 1) that adaptively determines the optimal classification boundary. The algorithm begins with an initial threshold $t = 0.5$ and iteratively adjusts it based on precision-recall trade-offs from the training set. The pseudocode is shown in Algorithm 1.



(a) Adaptive threshold optimization for distinguishing synthetic and real images.

(b) Methodology combining CNN, three-way decision, feature fusion, and TreeNet classification.

Figure 1: Methodology combining CNN, three-way decision, feature fusion, TreeNet classification, and adaptive thresholding.

Algorithm 1 Adaptive Threshold Optimization

Input: Probabilities P, ground truths Y
Output: Optimal threshold t^*
Initialize $t \leftarrow 0.5$
while True **do**
    Compute Precision P_t and Recall R_t at threshold t
    if $|P_t - R_t| < 0.1$ **then**
        return $t^* = t$
    else if $P_t > 1.1 \times R_t$ **then**
        $t \leftarrow t \times (1 + \text{rand}(0, 1)/10)$
    else
        $t \leftarrow t \times (1 - \text{rand}(0, 1)/10)$
    end if
end while

4. Results and Analysis

The framework was evaluated on the **MediaEval 2025 Synthetic Images** dataset [13] using the code available at (https://github.com/zeshanalvi/mediaEval2025) which is using the CNN model of PeterWang provided at (https://github.com/zeshanalvi/SyntheticDetection). The model uses a TreeNet architecture which is available at (https://pypi.org/project/dtreenetwork/) and (https://github.com/zeshanalvi/TreeNet) with the features extraction at (https://github.com/zeshanalvi/Feature-Extraction/) [15]. The confusion matrix in Table 1 summarizes classification performance.

Table 1
Confusion Matrix and Metrics

	Pred Real	Pred Fake	Total
Real	1678	3322	5000
Fake	2566	2434	5000
Total	4244	5756	10000

The model achieved an accuracy of **41.1%**, precision of **42.3%**, recall of **48.6%**, and F1-score of **45.3%**. Probability-based metrics indicated a ROC-AUC of **0.3775**. Optimizing the global threshold to 0.001 improved F1-score to **0.6667** with recall = 1.0, showing the effectiveness of adaptive thresholding in uncertain domains.

5. Conclusion and Future Work

This study introduced a three-way decision model for synthetic image detection that integrates a fine-tuned CNN with a texture-based decision ensemble mechanism. The proposed approach leverages the CNN's probabilistic confidence for high-certainty predictions, while uncertain cases are re-evaluated through TreeNet [15] using handcrafted texture descriptors such as Gabor and Local Binary Patterns (LBP). This hybrid integration effectively combines the strengths of deep feature learning and interpretable texture analysis, resulting in a more balanced and explainable detection process. The model further uses an adaptive threshold learning approach to reduce the bias of the synthetic detection.

Experimental evaluation on the MediaEval 2025 dataset demonstrated that the model achieves moderate overall accuracy while offering high adaptability in handling uncertain predictions. The inclusion of the third decision path significantly improved robustness against ambiguous samples, illustrating the value of multi-level reasoning in distinguishing between real and generative imagery.

In future work, we plan to enhance the generalization capability of the system across unseen generative models and diverse synthesis pipelines by incorporating frequency and diffusion-aware features. Furthermore, integrating explainable AI (XAI) components will be explored to improve the interpretability of detection outcomes and to support forensic-level transparency in AI-generated content verification.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] I. Goodfellow, et al., Generative adversarial nets, NeurIPS (2014).
[2] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv:1312.6114 (2013).
[3] R. Rombach, et al., High-resolution image synthesis with latent diffusion models, in: CVPR, 2022.
[4] L. Verdoliva, Media forensics and deepfakes: An overview, IEEE JSTSP (2020).
[5] P. Korshunov, S. Marcel, The threat of deepfakes to computer and human visions, in: A. Mian, Y. Zubair, H. Ugail, X. Chen (Eds.), Deep Learning for Biometrics, Springer, 2022, pp. 71–92.
[6] L. Guarnera, et al., Deepfake detection by analyzing convolutional traces, IEEE Access (2020).
[7] Y. Zhang, et al., Detecting multimedia generated by large ai models: A survey, arXiv preprint arXiv:2403.00000 (2024).
[8] X. Dong, et al., Thinking in frequency: Deepfake detection by mining frequency-aware clues, in: Proceedings of ECCV (workshops), 2022.
[9] X. Li, et al., Frequency-based discrepancy detection for ai-generated images, arXiv preprint arXiv:2308.00000 (2023).
[10] H. Tariq, et al., A comprehensive review of deepfake detection methods, Journal of Information Security and Applications (2023).
[11] R. Gandhi, N. Yadav, A. Sinha, Towards robust detection of gan and diffusion generated images using transformer-based hybrid models, Pattern Recognition Letters 165 (2023) 45–56.
[12] L. Zhang, X. Xu, H. Li, Z. Wu, Hybridfusion: Cnn-vit based multi-scale feature fusion for generative image detection, IEEE Transactions on Artificial Intelligence (2025).
[13] O. Papadopoulou, M. Schinas, R. Corvi, D. Karageorgiou, C. Koutlis, F. Guillaro, E. Gavves, H. Mareen, L. Verdoliva, S. Papadopoulos, Synthetic images at mediaeval 2025: Advancing detection of generative ai in real-world online images, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.
[14] Z. Khan, M. A. Tahir, Real time anatomical landmarks and abnormalities detection in gastrointestinal tract, PeerJ Computer Science 9 (2023) e1685.
[15] Z. Khan, Treenet: Layered decision ensembles, arXiv preprint arXiv:2510.09654 (2025).
[16] Y. Li, S. Lyu, Exposing deepfake videos by detecting face warping artifacts, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018.
[17] S. Agarwal, et al., Protecting world leaders against deep fakes, in: CVPR Workshops, 2019.
[18] R. Durall, M. Keuper, J. Keuper, Watch your upconvolution: CNN-based generative deep neural networks are failing to reconstruct spectral distributions, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2020.
[19] J. Frank, T. Eisenhofer, U. Scherhag, C. Rathgeb, N. Damer, A. Kuijper, Leveraging frequency analysis for deep fake image recognition, in: International Conference on Machine Learning (ICML) Workshops, 2020.
[20] Y. Qian, H. Wang, S. Ren, Thinking in patches: Towards better generalization in deepfake detection with vision transformers, Neural Networks 153 (2022) 495–508.
[21] Y. Liu, C. Huang, F. Wang, L. Zhao, Diffguard: Detecting diffusion-generated images via cnn-transformer hybridization, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).
[22] Y. Qian, et al., Thinking in frequency: Face forgery detection by mining frequency-aware clues, in: ECCV, 2020.
[23] N. A. Chandra, R. Murtfeldt, L. Qiu, A. Karmakar, H. Lee, E. Tanumihardja, K. Farhat, B. Caffee, S. Paik, C. Lee, et al., Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024, arXiv preprint arXiv:2503.02857 (2025).
[24] Z. Khan, M. A. Tahir, Majority voting of heterogeneous classifiers for finding abnormalities in the gastro-intestinal tract., MediaEval 18 (2018) 29–31.
[25] J. Ricker, S. Damm, T. Holz, A. Fischer, Towards the detection of diffusion model deepfakes, in: VISIGRAPP 2024 / International Joint Conference on Vision, Imaging and Computer Graphics Theory and Applications, 2024.
[26] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A. A. Efros, Cnn-generated images are surprisingly easy to spot... for now, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.