COMPUTER VISION AND PATTERN RECOGNITION
ІКОМП'ЮТЕРНИЙ ЗІР ТА РОЗПІЗНАВАННЯ ОБРАЗІВ

https://doi.org/10.15407/intechsys.2026.04.099
UDC 004.92

K.K. DUKHNOVSKA, PhD (Engineering), Associate Professor,
Department of Software Systems and Technologies
Taras Shevchenko National University of Kyiv,
24, Bohdana Havrylyshyna St., Kyiv, 02000, Ukraine
https://orcid.org/0000-0002-4539-159X
duchnov@ukr.net
T.I. HOLUB, Student,
Department of Software Systems and Technologies
Taras Shevchenko National University of Kyiv,
24, Bohdana Havrylyshyna St., Kyiv, 02000, Ukraine
https://orcid.org/0009-0004-1178-7264
tetiana_holub@knu.ua

AI IMAGE DETECTION USING AN ENSEMBLE OF EFFICIENTNET, RESNET, AND VISION TRANSFORMER

Modern generative models, such as GANs and diffusion models, produce highly realistic synthetic images that are virtually indistinguishable from authentic photographs, posing significant risks in forensics, journalism and data verification. To address the need for a fast and accessible detection tool, this paper proposes an ensemble approach combining EfficientNet-B0, ResNet-50 and ViT-B/16 architectures. Fine-tuned via transfer learning on the CIFAKE and AI-vs-Real datasets, the ensemble leverages CNNs for local texture anomaly detection and the Vision Transformer for global patch dependencies. Evaluated on a 2,000-image test set using soft voting, the model achieved an accuracy of 89.60%, precision of 92.94%, recall of 85.70%, an F1-score of 89.17% and an AUC-ROC of 0.9641. The notably high precision minimizes false accusations of tampering. The system is deployed as a scalable web application utilizing a microservices architecture (React, Node.js, FastAPI, PostgreSQL, Docker).

Keywords: *synthetic image detection, deepfake detection, ensemble learning, EfficientNet, Vision Transformer, transfer learning, CIFAKE.*

Cite: Dukhnovska K.K., Holub T.I. AI Image Detection Using an Ensemble of Efficientnet, Resnet, and Vision Transformer. *Information Technologies and Systems*. 4 (10). 2026. 99—111. https://doi.org/10.15407/intechsys.2026.04.099

© Publisher PH "Akademperiodyka" of the NAS of Ukraine, 2025. This is an Open Access article under the CC BY-NC-ND 4.0 license (https://creativecommons.org/licenses/by-nc-nd/4.0/

Introduction

Synthetic images generated by neural networks no longer exhibit characteristic defects. Modern generative adversarial networks and diffusion models reproduce faces, scenes and objects with a level of quality that prevents even an experienced observer from identifying the forgery. Forged photographs have already been introduced into legal proceedings, electoral campaigns and training datasets of other systems.

The scale of the problem has shifted with the advent of diffusion models. While early generative adversarial networks left characteristic artifacts such as deformed ears, an incorrect number of fingers and blurred background details, systems like Stable Diffusion, Midjourney and DALL-E have eliminated most of them. Detectors trained to recognize these legacy features fail when applied to novel content.

The task of synthetic image detection is formulated as a binary classification problem where each input image is categorized as either real or generated. Existing approaches are generally divided into two main directions. Convolutional neural network methods capture local textural artifacts but are constrained by the size of their receptive field. Transformer architectures model the global structure of the image, although they require more substantial computational resources.

An ensemble combination of these architectures compensates for their individual weaknesses. EfficientNet-B0 adeptly detects low-frequency textural anomalies. ResNet-50 provides robust generalization due to its residual connections. The ViT-B/16 model analyzes the patch structure on a global scale. Together, these three networks cover a significantly broader spectrum of forgery features than each model operating independently.

Problem Statement

The rapid advancement of generative models — particularly GANs and diffusion architectures — has outpaced the development of reliable detection methods. Existing detectors are predominantly trained and evaluated on narrow, homogeneous datasets, limiting their generalization to novel generative models and real-world conditions such as image compression or resizing. Furthermore, the majority of current approaches focus on facial deepfake video content, leaving the domain of general-purpose synthetic images – objects, scenes, and graphics — underexplored. The absence of accessible, ready-to-deploy detection tools compounds the problem for end users. This study addresses the need for a robust, ensemble-based classifier capable of operating across heterogeneous image sources and deployable as a practical web application.

Literature Review

An analysis of contemporary scientific literature on the problem of synthetic image detection demonstrates a conceptual shift from the exclusive

use of convolutional neural networks to hybrid and ensemble approaches based on attention mechanisms.

The authors of the study [1] systematized CNN-based methods for digital image forensics, distinguishing between four types of manipulations: copy-move, image splicing, object removal and GAN generation. On the FaceForensics++ dataset, XceptionNet and InceptionNet demonstrated the highest accuracy among the evaluated architectures. Early forensic techniques predominantly relied on frequency-domain features, utilizing the discrete cosine transform and wavelet transforms. While these features effectively detected specific types of manipulations, they failed to generalize to novel ones. Data-driven neural network approaches partially mitigated this issue by learning to detect a broader spectrum of artifacts; however, this came at the cost of reduced accuracy when encountering shifts in the input data distribution. The authors highlight a systemic limitation: the training and testing of most reviewed methods are conducted on the identical data distribution. Although ResNet and EfficientNet exhibit greater robustness to such shifts due to residual connections and compound scaling, their generalization to diffusion models — which were only emerging at the time of publication — remained untested. Consequently, the challenge of developing detectors capable of identifying images from heterogeneous sources without the need for additional training remains unresolved [1].

In study [2], the specialized CIFAKE dataset is presented, comprising 120,000 images designed for binary classification tasks. Its authentic subset is formed from 60,000 images from the CIFAR-10 dataset, which covers ten distinct classes including airplanes, automobiles and birds. Conversely, the synthetic segment includes 60,000 images generated using the Stable Diffusion 1.4 model. The fixed resolution of 32×32 pixels distinguishes CIFAKE from similar datasets focused on facial recognition, ensuring high computational speed and enabling rapid model training even with limited hardware resources. The generation process for the artificial content was based on a 50-step reverse diffusion process utilizing text prompts for each semantic class; notably, additional thematic modifiers were applied to increase the diversity of the airplane class sample.

The authors conducted a comparative analysis of 36 convolutional neural network topologies, identifying an optimal configuration of two layers with 128 filters each, which achieved an accuracy of 92.98% and an F1-score of 0.936. A key scientific finding of the study is that the classifier relies not on the features of the central object, but on the background artifacts of the generation process. This assertion was verified using the Grad-CAM method: activation heatmaps concentrate on peripheral zones and background textural anomalies rather than on the target object. The CIFAKE dataset has been published in open access, leading to its extensive application in subsequent research. However, the study did not investigate the generalization capability of the developed methods to other generative architectures, nor did it explore the effectiveness of ensemble approaches and transformer-based models.

The authors of the study [3] compared MesoNet, XceptionNet, and DenseNet, proposing a modified Dense CNN architecture with densely connected convolutional blocks. This architecture demonstrated superior precision and F1-scores compared to MesoNet on the FaceForensics++ dataset.

Dense connections between the layers enable each layer to receive feature maps from all preceding layers. This mitigates the vanishing gradient problem and enhances feature reuse during the training process. Furthermore, the study investigated the impact of video compression: transitioning from an uncompressed format to a highly compressed one significantly degraded the accuracy of all evaluated methods — a result consistent with the findings in [1].

However, the scope of the study is limited to facial video datasets. The efficacy of the Dense CNN for general-purpose static synthetic images — such as objects, landscapes and graphics — was not evaluated. Consequently, the applicability of this approach beyond the domain of deepfake videos, as well as its robustness when encountering generative models absent from the training set, remain unresolved.

The authors of [4] investigated the performance gap of deepfake detectors between laboratory and real-world conditions. The vast majority of existing systems exhibit acceptable accuracy on academic datasets but experience significant degradation when encountering compressed or scaled images, which are typical of social media platforms.

The authors introduced a standard set of transformations: JPEG compression, resizing, Gaussian blur, cropping and their combinations. Evaluating XceptionNet and EfficientNet-B4 against these transformations revealed an average drop in AUC of 15–30 percentage points, even under moderate compression with a quality factor of 70.

The authors concluded that data augmentation during training is not merely an option, but a necessary prerequisite for robustness. Architectural solutions alone are insufficient to achieve this. However, the automatic selection of a classification threshold based on environmental conditions, as well as ensemble methods designed to compensate for the vulnerabilities of individual models under specific circumstances, were not addressed in the study.

The study [5] is devoted to the development of FG-TEFusionNet, a system that combines an inter-frame temporal module (an RNN with an SENet attention mechanism) and a spatial module (EfficientNet integrated with a Gram texture module). The final decision is determined through a voting mechanism between them. Preprocessing involves IMTCNN for face extraction and a diffusion model for artifact removal.

The architecture was evaluated on four subsets of the FaceForensics++ and Celeb-DF datasets. A distinct advantage of the proposed method is its robustness against adversarial attacks: the authors utilized I-FGSM to generate adversarial examples and applied adversarial training for defense. The vast majority of existing deepfake detectors overlook this aspect.

Acting as a spatial feature extractor, EfficientNet demonstrated competitive results, which corroborates the selection of this architecture in the present study as well. However, questions concerning the effectiveness of the approach for static synthetic images outside the context of facial videos, along with the viability of employing an ensemble lacking a temporal component, remained unanswered.

In study [6], DiffCoR is proposed as a fundamentally different approach to detection. Instead of feature classification, the method utilizes a diffusion model for reconstruction: the image is corrupted with noise and then restored, after which it is compared to the original based on correlation characteristics. Authentic photographs are reconstructed less accurately due to the unique noise characteristics of camera sensors — synthetic images lack such properties.

Among the baseline approaches used for comparison, Tran et al. evaluated EfficientNet-B0 and ViT in a configuration similar to the one employed in the present study. On specific subtypes of synthetic images, both models were outperformed by DiffCoR in terms of AUC. This underscores the limitations of approaches relying solely on feature classification, while simultaneously confirming that EfficientNet-B0 and ViT serve as viable foundational components.

However, DiffCoR necessitates executing the diffusion process for every single input image. This is computationally expensive and significantly hinders its application in real-time scenarios. The potential of integrating DiffCoR with classification-based approaches to achieve a more optimal balance between accuracy and inference speed was not addressed in the study.

The authors of [7] developed an eight-layer CNN to detect AI-generated images. A key distinction from the majority of studies is its evaluation across three diverse domains: the CIFAKE dataset comprising 32 by 32 pixel diffusion images, a Midjourney v6 dataset at a 64 by 64 resolution, and a proprietary Sentinel-2 satellite imagery dataset generated via StyleGANv3 at a 128 by 128 resolution.

On the CIFAKE dataset, the model achieved an accuracy of 96.60%, on Midjourney v6 it reached 99.94%, and on Sentinel-2 it attained 98.69%. To evaluate the quality of the synthetic satellite images, the authors employed the FID metric, noting that a value below 25 at the conclusion of training indicates a sufficient level of realism within the synthetic sample. The application of the Grad-CAM++ method corroborated the findings of [2], confirming that activations are concentrated on the background and peripheral artifacts rather than on the primary object.

The study [7] demonstrates that a compact CNN with a limited number of parameters can yield competitive results. Nevertheless, all three experiments were conducted using entirely separate models. The potential of integrating CNNs with transformer-based approaches in an ensemble framework, alongside the issue of stability during cross-domain transfer, remained unexplored.

In study [8], a two-stream system based on EfficientNet-B0 and ViT was developed. A fundamental distinction from a conventional ensemble is that both networks process not only RGB images but also noise maps generated by Steganalysis Rich Model filters. Feature fusion is achieved through an Efficient-to-Former block, where the local features extracted by EfficientNet-B0 are combined with the global dependencies of the ViT via cross-attention.

The authors explicitly justify their selection of EfficientNet-B0 over more advanced versions of the family, noting that it provides a superior balance between accuracy and computational complexity compared to ResNet-50 and MobileNetV3. The system achieved accuracy scores of 0.845, 0.999 and 0.815 on the FaceForensics, Celeb-DF-V2 and Wild Data datasets, respectively. Notably, on the Wild Data dataset, which consists of images sourced from the Internet, the proposed framework outperformed both baseline approaches.

However, the system is specifically tailored for facial video analysis. Evaluation on general synthetic image datasets, namely CIFAKE and AI-vs-Real, was not performed. Consequently, whether the advantage of the noise stream is maintained beyond the domain of video content remains an open question.

In study [9], the authors developed the Patch-Based Vision Transformer for Identifying Synthetic Media, denoted as PV-ISM, which is a ViT architecture modified specifically for the detection task. Instead of a standard sigmoid output, PV-ISM generates two logits, allowing for more flexible threshold tuning. For regularization purposes, dense layers and dropout were introduced between the transformer blocks. Notably, the model was trained entirely from scratch without relying on pre-trained weights.

Following a comprehensive hyperparameter search encompassing patch size, image resolution, the number of blocks and the number of epochs, PV-ISM achieved an accuracy of 96.60% and an F1-score of 96.61% on the CIFAKE dataset. Under identical conditions, a baseline ResNet-50 model reached an accuracy of 93.32%. Additional experiments conducted on the RVF-10K dataset, comprising facial images from StyleGAN and FFHQ and the Art Images dataset yielded accuracies of 81.70% and 93.16%, respectively. This demonstrates the model's capacity to adapt to diverse domains without the need for retraining.

However, the authors of [9] did not explore an ensemble combination of PV-ISM with CNN architectures. Furthermore, the model's performance when evaluated on combined datasets — such as integrating CIFAKE and AI-vs-Real — was not investigated.

Study [10] describes a deepfake image detection competition organized at the ICIAP 2021 conference. Task I required the detection of forgeries subjected to various post-processing attacks, specifically JPEG compression, resizing, blurring, cropping and combinations thereof. The dataset was compiled using authentic images from CelebA and FFHQ,

alongside synthetic images produced by six distinct GAN architectures: StarGAN, StarGAN-v2, StyleGAN, StyleGAN2, AttGAN, and GDWCT.

The outcomes proved to be highly revealing. The winning team, VisionLabs, utilized EfficientNet-B3 pre-trained on ImageNet weights, achieving an accuracy of 93.61% in the presence of these attacks. The runner-up employed EfficientNet-B5, attaining an AUC of 0.9674. The third-place solution featured a hybrid model combining EfficientNet and CrossViT. Notably, all three top positions were secured by solutions based on the EfficientNet family, with the best-performing approaches integrating EfficientNet with transformer elements.

However, the authors of [10] did not aim to compare EfficientNet with ResNet and ViT within an equal-weight ensemble framework. Furthermore, the performance of such a combination on general synthetic image datasets, outside the domain of facial imagery, was not analyzed in the study.

Studies [3, 5, 8, 10] evaluate detectors on videos or facial images. The classification of general-purpose synthetic images — specifically objects and scenes, such as those in the CIFAKE and AI-vs-Real datasets — remains under-investigated. Studies [2, 7, 9] focus exclusively on individual architectures. An ensemble combination of EfficientNet-B0, ResNet-50, and ViT-B/16 applied to general datasets utilizing an equal-weight soft voting configuration has not been examined by any of the analyzed works. Furthermore, the authors of [4] demonstrated that the vast majority of detectors degrade under real-world conditions and the systems themselves are rarely implemented as accessible applications for end users.

The Purpose of the Paper

The objective is to develop an ensemble software system for synthetic image detection based on EfficientNet-B0, ResNet-50 and ViT-B/16, which provides greater classification stability compared to each model operating independently.

To achieve this objective, the following tasks were formulated and accomplished:

- Train EfficientNet-B0, ResNet-50 and ViT-B/16 using a transfer learning approach on the CIFAKE and AI-vs-Real datasets, and construct an ensemble by averaging the predicted probabilities.
- Evaluate the ensemble using the Accuracy, Precision, Recall, F1-score and AUC-ROC metrics and compare the obtained performance with the results of the individual models.

Materials and Methods

The research was conducted using two open-source datasets, ensuring the representativeness of the sample. The CIFAKE dataset comprises 120,000 images with a spatial resolution of 32×32 pixels, of which 60,000 are authentic images from the CIFAR-10 database and the remaining 60,000 are synthetic images generated using the Stable Diffusion 1.4 dif-

fusion model, uniformly distributed across ten semantic classes [2]. The AI-vs-Real (Kaggle) dataset contains higher-resolution images and encompasses a broader spectrum of generative stylistic features. The integration of these datasets enabled the training of the models on both low-resolution baseline images and highly realistic samples with complex structures.

Three complementary neural network architectures were utilized as baseline classifiers. The EfficientNet-B0 model implements the compound scaling paradigm (simultaneous proportional scaling of network depth, width and resolution), which provides higher classification accuracy with fewer parameters compared to predecessor architectures; the feasibility of its application is corroborated by the results of study [8]. The ResNet-50 architecture mitigates the vanishing gradient problem through the use of residual connections between layers, demonstrating a high generalization capacity on heterogeneous datasets [1]. The Vision Transformer (ViT-B/16) model decomposes the input image into 16×16 pixel patches and models the global contextual dependencies among them via a self-attention mechanism — an approach that has proven effective on the CIFAKE dataset in study [9].

The models were trained utilizing a transfer learning approach with pre-initialized weights derived from the ImageNet dataset. In the initial stage, the base feature extraction layers were frozen and only the classification blocks were subjected to optimization. In the second stage, the upper layers were unfrozen to perform the fine-tuning procedure. The original classification heads of the models were replaced with fully connected linear layers for binary classification featuring a sigmoid activation function.

The training process was implemented using the AdamW optimization algorithm and the CosineAnnealingLR learning rate scheduler. Binary Cross-Entropy was selected as the loss function. Computational experiments were conducted in the Google Colab cloud environment using the PyTorch framework and the torchvision library. To enhance the robustness of the detectors, in accordance with the recommendations of study [4], data augmentation was applied, encompassing horizontal flipping, random cropping and standardized normalization using ImageNet parameters.

The final decision of the system is formed based on an ensemble method via averaging the predicted probabilities (soft voting) from all three equally weighted models. Classification is executed utilizing a threshold value of 0.5. The applied approach allows for the effective aggregation of the local sensitivity to textural artifacts characteristic of convolutional neural networks (EfficientNet-B0 and ResNet-50) with the global spatial analysis inherent to the ViT-B/16 architecture.

The evaluation of the developed method's effectiveness was conducted on a balanced test set consisting of 2,000 images (1,000 authentic and 1,000 synthetic), utilizing the following metrics: Accuracy, Precision, Recall, F1-score and the AUC-ROC.

The software implementation of the system was executed as a multi-tier web application with a microservices architecture, comprising four key abstractions:

- A client-side frontend based on React/TypeScript using the Nginx web server (port 8080);
- An API gateway on the Node.js/Express platform implementing JWT authentication and a caching mechanism based on the SHA-256 hash function;
- A machine learning model inference module utilizing Python/FastAPI;
- A PostgreSQL 16 relational database management system.

Results

Based on the results of the experimental evaluation of the three baseline neural network architectures and their ensemble on an independent test set of 2,000 images, the proposed comprehensive approach demonstrated the highest performance.

An analysis of the overall metrics (Table 1) reveals that the ensemble achieved the highest general accuracy of 89.60%, outperforming the best individual model, ResNet-50, which yielded an accuracy of 88.35%. The precision of the developed system reached 92.94%, demonstrating the high reliability of its positive predictions and the minimization of false labeling of authentic photographs. In terms of recall, the ensemble model identified 85.70% of all synthetic images, ensuring an optimal balance of metrics with an F1-score of 89.17%. The AUC-ROC for the ensemble was 96.41%, confirming its exceptional predictive capability. Among the baseline classifiers, the ResNet-50 architecture proved to be the most effective, followed by the ViT-B/16 transformer with an accuracy of 86.90%, whereas EfficientNet-B0 exhibited the weakest results on this dataset with an accuracy of 84.85%.

A detailed analysis of the confusion matrices reveals the specific nature of the misclassifications (Fig. 1). The ensembling approach significantly reduced the number of false-positive predictions: only 65 instances of authentic images being mistakenly labeled as synthetic were recorded, whereas this figure was 83 for ResNet-50, 98 for ViT-B/16 and 118 for EfficientNet-B0. Owing to this reduction in the error rate, the integrated

Table 1. **Metrics of individual models and the ensemble (n = 2000)**

Model	Accyracy, %	Precission, %	Recall, %	F1-score, %	AUC-ROC
EfficientNet-B0	84.85	87.37	81.50	84.34	91.43
ResNet-50	88.35	91.10	85.00	87.95	94.12
ViT-B/16	86.90	89.50	83.60	86.45	93.02
Ensemble	89.60	92.94	85.70	89.17	96.41

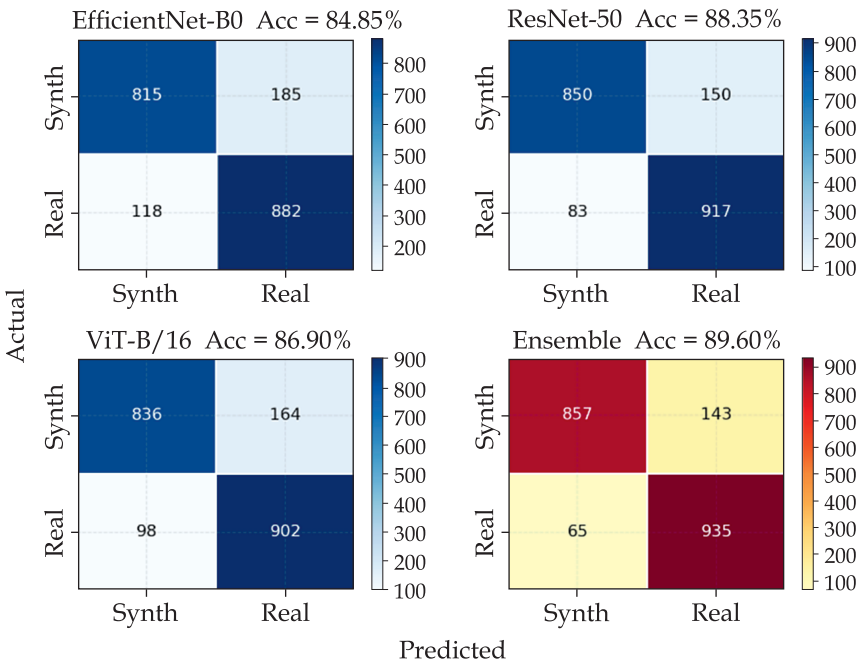


Fig. 1. Ensemble discrepancy matrix (n = 2000)

system correctly classified 935 out of 1,000 authentic images. Regarding the detection of generated content, the ensemble correctly identified 857 out of 1,000 synthetic samples. Although 143 synthetic images evaded detection, this result is notably better compared to the ViT-B/16 (164 misses) and EfficientNet-B0 (185 misses) models, performing on par with the best individual model.

Overall, the integration of the models by averaging their predictions successfully compensated for the individual weaknesses of the baseline architectures, resulting in a substantial reduction in prediction variance and a significant enhancement in the overall reliability of the visual content detector.

EfficientNet-B0 and ResNet-50 detect local textural anomalies, whereas ViT-B/16 analyzes the global patch structure. Their combined averaging reduces prediction variance, as each network compensates for the blind spots of the other. This aligns with the findings of study [8], where a similar combination of EfficientNet-B0 and ViT yielded superior results on the Wild Data dataset.

The accuracy of 89.60% is lower than the 96.60% achieved in study [9] for PV-ISM on the CIFAKE dataset. The reason lies not in the quality of the ensemble, but in the testing conditions: here, the ensemble was evaluated on a combined sample of two datasets, including more realistic images from AI-vs-Real, whereas [9] tested exclusively on low-resolution 32×32 images from CIFAKE. The authors of study [7] documented a similar effect when transitioning between datasets of varying sizes.

Both utilized datasets cover a limited spectrum of generative models; in particular, the CIFAKE database is formed exclusively using the Stable

Diffusion 1.4 architecture. At the same time, as demonstrated by the results of study [4], the effectiveness of detection systems suffers significant degradation when the base generator is changed or image compression algorithms are applied. This vulnerability constitutes a systemic problem for the entire domain, which is also corroborated by the conclusions of the review paper [1].

The precision of 92.94% exceeds the recall of 85.70%. The detector rarely labels authentic photos as synthetic — an essential property for verification applications where a false accusation of forgery is unacceptable. Study [10] also emphasizes the necessity of a careful balance between these metrics depending on the scenario.

The AUC-ROC of 0.9641 is higher than the average level of CNN detectors reported in review [1] and is comparable to systems [6] and [8], which utilize significantly greater resources. Further enhancement is possible through weighted ensembles, expanding the training data with images from next-generation generators, and domain adaptation methods. Furthermore, the authors of studies [2] and [7] highlight Grad-CAM as a promising tool for analyzing which specific forgery features the detector actually utilizes.

Conclusions

In this study, an ensemble software system for detecting artificially generated images was developed and experimentally verified, based on the integration of three complementary neural network architectures: EfficientNet-B0, ResNet-50 and ViT-B/16. The proposed ensemble approach, based on probability averaging, demonstrated its undeniable superiority over the individual models. The system achieved the highest overall accuracy of 89.60% and an AUC-ROC of 96.41%. Furthermore, the ensemble demonstrated high precision (92.94%), substantially reducing the number of false-positive predictions: only 65 instances of mistakenly labeling an authentic image as synthetic were recorded. This confirms the overall reliability of the developed system and its safety when analyzing genuine photographic content.

Such a strong performance was made possible by the synergistic effect of combining fundamentally different architectures. The integration successfully merged the local sensitivity to textural generation artifacts, inherent to the EfficientNet and ResNet convolutional networks, with the global contextual analysis of spatial dependencies provided by the ViT transformer model. This allowed for compensating the individual shortcomings of the baseline classifiers and significantly reducing prediction variance. The practical value of the work is corroborated by the implementation of the developed method as a ready-to-use multi-tier web application with a microservices architecture, addressing the pressing issue of the accessibility of modern detection systems for end users.

Perspectives for future research include expanding the experimental base by integrating images from the latest diffusion models, evaluating the

robustness of the ensemble system against adversarial attacks and complex post-processing methods, as well as optimizing the computational complexity of inference to ensure the system's efficient operation in real time.

DECLARATIONS

Author Contributions. *Holub T.: Conceptualization, Software, Investigation, Data Curation, Writing – Original Draft, Visualization. Dukhnovska K.K.: Methodology, Supervision, Writing – Review & Editing.*

AI and Automated Tools Usage. *During the preparation of this manuscript, the authors used Claude (Anthropic) to assist with text editing and language refinement. All scientific content, experimental results, and conclusions were produced and verified by the authors. The authors take full responsibility for the content of this publication.*

Conflict of Interest. *The authors declare no conflict of interest.*

Grant or Financial Support. *This research received no external funding.*

REFERENCES

1. Castillo Camacho I., Wang K. A Comprehensive Review of Deep-Learning-Based Methods for Image Forensics. *Journal of Imaging*, 2021, Vol. 7 (7), Article 69. https://doi.org/10.3390/jimaging7070069
2. Bird J.J., Lotfi A. CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. *IEEE Access*, 2024, Vol. 12, 15642–15650. https://doi.org/10.1109/ACCESS.2024.3356122
3. Patel Y., Tanwar S., Bhattacharya P., et al. An Improved Dense CNN Architecture for Deepfake Image Detection. *IEEE Access*, 2023, Vol. 11, 22081–22095. https://doi.org/10.1109/ACCESS.2023.3251753
4. Lu S., Ebrahimi M. Assessment Framework for Deepfake Detection in Real-World Situations. *EURASIP Journal on Image and Video Processing*, 2024, Vol. 2024, Article 9. https://doi.org/10.1186/s13640-024-00624-5
5. Lei S., Song J., Feng F., et al. Deepfake Face Detection and Adversarial Attack Defense Method Based on Multi-Feature Decision Fusion. *Applied Sciences*, 2025, Vol. 15 (12), Article 6588. https://doi.org/10.3390/app15126588
6. Tran V.-N., Lee S.-H., Le H.-S., Kwon K.-R. DiffCoR: Exposing AI-Generated Image by Using Stable Diffusion Model. *IEEE Open Journal of Computer Society*, 2025, vol. 6, 1353–1365. https://doi.org/10.1109/OJCS.2025.3575507
7. Lokner Lađević A., Kramberger T., Kramberger R., Vlahek D. Detection of AI-Generated Synthetic Images with a Lightweight CNN. *AI*, 2024, Vol. 5 (3), 1575–1593. https://doi.org/10.3390/ai5030076
8. Yang Q., Cai Y., Yan R., Zhang G. Detecting Deepfakes Using Two-Stream Framework Based on EfficientNet and Vision Transformer. *Journal of Electrical and Computer Engineering*, 2026, Vol. 2026, Article 9686924. https://doi.org/10.1155/jece/9686924
9. Çınar O., Doğan Y. Novel Deepfake Image Detection with PV-ISM: Patch-Based Vision Transformer for Identifying Synthetic Media. *Applied Sciences*, 2025, Vol. 15 (12), Article 6429. https://doi.org/10.3390/app15126429
10. Guarnera L., Giudice O., Guarnera F., et al. The Face Deepfake Detection Challenge. *Journal of Imaging*, 2022, Vol. 8 (10), Article 263. https://doi.org/10.3390/jimaging8100263

Received 11.06.2026
Accepted 21.08.2026
Published 23.10.2026

К.К. ДУХНОВСЬКА, канд. техн. наук, доцент,
каф. програмних систем і технологій,
Київський національний університет імені Тараса Шевченка,
вул. Богдана Гаврилишина, 24, Київ, 02000, Україна
https://orcid.org/0000-0002-4539-159X
SCOPUS ID: 57433704700
duchnov@ukr.net
Т.І. ГОЛУБ, студентка,
каф. програмних систем і технологій,
Київський національний університет імені Тараса Шевченка,
Богдана Гаврилишина, 24, Київ, 02000, Україна
https://orcid.org/0009-0004-1178-7264
tetiana_holub@knu.ua

ВИЯВЛЕННЯ ЗОБРАЖЕНЬ, ЗГЕНЕРОВАНИХ ШІ, ЗА ДОПОМОГОЮ АНСАМБЛЮ *EFFICIENTNET, RESNET І VISION TRANSFORMER*

Вступ. Інтенсивний розвиток архітектур генеративно-змагальних мереж та дифузійних моделей зумовив появу алгоритмів синтезу візуальних даних, які за своїми перцептивними характеристиками практично не відрізняються від автентичних цифрових зображень, отриманих за допомогою оптичних сенсорів. Такий контент уже фіксується у судових матеріалах, новинних публікаціях і тренувальних датасетах — на відміну від ранніх *GAN*-підробок, де артефакти були помітні без спеціальних інструментів.

Мета статті. Ця обставина формує актуальну науково-прикладну задачу: розроблення автоматизованих методів ідентифікації згенерованих зображень, що здатні забезпечити високу точність детектування в масштабі реального часу без жорстких вимог щодо наявності спеціалізованих апаратних прискорювачів на стороні клієнта.

Методи. У межах цієї роботи запропоновано та програмно реалізовано метод класифікації візуальних даних, в основу якого покладено ансамблювання трьох комплементарних нейромережевих архітектур: згорткових мереж *EfficientNet-B0* і *ResNet-50* та моделі на базі механізму уваги *Vision Transformer*. Адаптація моделей до предметної області здійснювалася із застосуванням парадигми трансферного навчання. Використання вагових коефіцієнтів, попередньо оптимізованих на масиві даних *ImageNet*, забезпечило ефективну екстракцію базових візуальних ознак на початкових етапах аналізу. Запропонований підхід дає змогу нівелювати вплив локальних похибок окремих моделей та суттєво підвищити загальну стійкість системи до вхідних шумів.

Результати. Емпірична оцінка ефективності запропонованого методу проводилася на репрезентативній тестовій вибірці обсягом 2000 зображень, які не використовувалися на етапі навчання. За результатами експериментальних досліджень класифікатор продемонстрував такі метрики якості: точність на рівні 89,60 %, прецизійність — 92,94 %, повнота — 85,70 %, F_1-міра — 89,17 %, а також *AUC-ROC* 0,9641. Отримані результати підтверджують наукову гіпотезу щодо доцільності інтеграції різнорідних архітектур.

Висновки. Ансамблевий детектор на основі *EfficientNet-B0*, *ResNet-50* і *ViT-B/16* перевершує кожну модель окремо і придатний для практичного застосування у веб-середовищі.

Ключові слова: *виявлення синтетичних зображень, deepfake, ансамблева модель, EfficientNet, Vision Transformer, transfer learning, CIFAKE.*