



Deepfoax Image Detection based on Deep Learning using Convolutional Neural Network Architectures

SY Yuliani¹, Nasywa Naura Aulia²

^{1,2}Informatics, Departement of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia

*sy.yuliani@umn.ac.id

Abstract

The rapid growth of generative artificial intelligence has increased the spread of deepfoax images, creating significant challenges for digital security, public trust, and the reliability of online information. This study evaluates the performance of MobileNet and Xception architectures for classifying facial images into fact and hoax categories while analyzing different feature representation levels within the Xception architecture through an ablation-based fine-tuning strategy. The proposed framework consists of data preparation, model architecture design, training procedures, and performance evaluation using confusion matrix analysis and metrics such as accuracy, precision, recall, and F1-score. Two publicly available datasets, DeepDetect2025 and FF-GenAI were utilized to evaluate the robustness of the proposed models. Experimental results show that Xception consistently outperformed MobileNet across all evaluation metrics, with the Middle Level Layer configuration achieving the best performance at 99.71% accuracy and F1-score. The findings indicate that intermediate feature representations are the most effective for capturing structural inconsistencies, texture irregularities, and synthetic blending artifacts commonly found in AI-generated facial images. In contrast, low-level representations were less discriminative, while highly abstract semantic representations slightly reduced sensitivity to localized manipulation artifacts. Overall, this study demonstrates the effectiveness of Xception-based feature refinement for deepfoax image detection and contributes to AI-based approaches for digital content verification, cybersecurity, and visual misinformation mitigation.

Keywords: convolutional neural network; deepfoax detection; image classification; mobilenet; Xception

How to Cite: S. Yuliani and N. N. . Aulia, "Deepfoax Image Detection based on Deep Learning using Convolutional Neural Network Architectures", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 4, pp. 1117–1124, Aug. 2026, doi: [10.29207/resti.v10i7403](https://doi.org/10.29207/resti.v10i7403).
Permalink/DOI: <https://doi.org/10.29207/resti.v10i7403>

Received: December 20, 2025

Accepted: April 19, 2026

Available Online: August 20, 2026

This is an open-access article under the CC BY 4.0 License

Published by Ikatan Ahli Informatika Indonesia

1. Introduction

The development of digital technology and the internet have revolutionized many aspects of life, but one major drawback is that it has created additional cybersecurity issues. The content spreads quickly on social media, hate speech, black campaigns, and hacks are becoming easier [1]. According to several studies, hoax identification is a significant challenge in cyber security. The growing amount of digital information and the ease of access to digital content can be shared on social media was the cause of fake information can spread rapidly, and it was hard to stop [2], [3].

Among these threats, the spread of hoaxes or false information has become one of the most serious issues because it can stir up panic or twist what people believe, and that erodes trust in online stuff. As artificial intelligence technology continues to advance, a new phenomenon known as deepfake or deepfoax has

emerged. This refers to visual manipulation using deep learning techniques, which are basically fake images or videos made with deep learning to look highly realistic. Findings in the field of cyber security show that deepfoax technology increases the complexity of digital threats because AI-generated manipulative content can bypass traditional detection systems and mislead the public through highly convincing visual alterations [4], [5]. According to the Oxford English Dictionary, the term "hoax" is defined as a malicious deception or lies created with harmful intent [6]. Therefore, a deepfoax can be understood as an advanced form of digital hoax that uses sophisticated technology to deceive visually. A recent example is illustrated in Figure 1, which shows the hoax "Prabowo Beri Dana Bantuan untuk Lunasi Utang dan Bangun Rumah," an AI-manipulated content that was officially verified as false by Tribrata News Polda Jawa Barat.



Figure 1. Example of Deephoax Prabowo Subianto

The spread of deephoax content presents serious risks. This type of manipulative media can damage a person's reputation, generate political disinformation, and disrupt social stability. In practice, the technology is commonly produced using generative models such as Generative Adversarial Networks (GANs), which are generative models that modify faces in ways hard to notice [7], [8]. One case that received national attention was a fraud incident uncovered in October 2025. In this case, an individual used deepfake or deephoax technology to create a fabricated video that appeared to show President Prabowo Subianto handing out cash assistance to the public. According to reports from Kompas and Liputan6, the deephoax was distributed to convince and entice potential victims, making it easier to deceive [9], [10]. Even though it was video, the face swapping idea is the same for images, where AI creates phony pictures of faces that look realistic. The same technology can be used to generate convincing hoax images, which means that the risk of misuse extends beyond video and affects image-based content shared on social media platforms. This situation demonstrates the increasing importance of early detection of deephoax images, especially for preventing digital fraud, public manipulation, and identity misuse across various forms of visual content.

As these threats continue to grow, research on deephoax image detection has become increasingly important. Unlike conventional manual image editing, deephoax content is automatically generated using deep learning techniques. Deep learning-based detection models have demonstrated strong capability in recognizing subtle

manipulation patterns [11]. Furthermore, using pre-trained models such as MobileNet and Xception has become popular, as it is more efficient and does not require training from scratch [12]. Among the available architectures, Xception (Extreme Inception) is considered one of the most effective. Xception applies depthwise separable convolution, which separates spatial convolution and channel convolution, making the model computationally lighter while enabling deeper visual feature extraction [13], [14]. Studies using the FaceForensics++ dataset show that Xception can detect deepfake video frames with up to 98% accuracy due to its ability to capture compression issues, fake face patterns, and micro-patterns artifacts [15]. In contrast, MobileNet offers a lightweight architecture optimized for computational efficiency and faster inference, making it suitable for real-time and resource-constrained applications [16].

The main contribution of this study is the evaluation of Xception and MobileNet architectures for deephoax facial image detection, combined with an ablation-based analysis of feature representation levels within the Xception architecture (Low, Middle, and High level layer) through ablation experiments. While most previous studies primarily focus on detecting facial deepfake or specific image manipulations, this study targets hoax images spread on social media that combine visual manipulation with misleading contextual information. By analyzing how different representation levels contribute to detection performance, this study contributes to the development of an artificial intelligence framework for visual misinformation mitigation, digital content verification, and cybersecurity applications.

2. Methods

The purpose of this study is to determine which component of the Xception architecture has the greatest impact on categorization performance. Using the FF-GenAI dataset, this method assesses the model's capacity to discriminate between real facial images and AI-generated images.

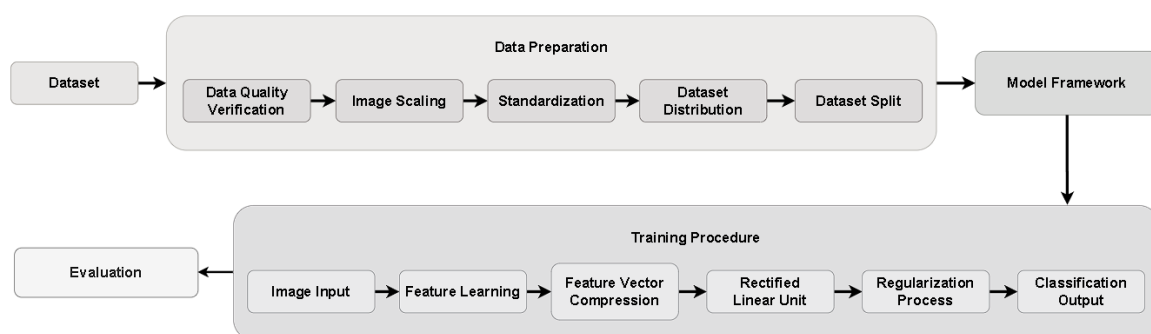


Figure 2. Research Methodology for Deephoax Detection

Figure 2 illustrates the general research methodology, which consists of several main stages: dataset collection, data preparation, model selection (pre-

trained Xception), training and validation using multiple fine-tuning scenarios, and model performance evaluation.

2.1 Facial Image Dataset

This study employs two datasets of human facial images divided into two categories: real (fact) images and AI-generated (hoax) images. The first dataset is the DeepDetect2025 Dataset, which consists of 30,000 facial images containing 15,000 real images and 15,000 hoax images [17] as shown in Figure 3. This dataset is utilized for training and evaluating the MobileNet model.

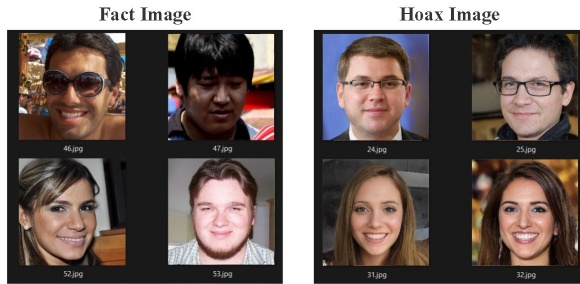


Figure 3. Fact and Hoax Images in the DeepDetect Dataset

The second dataset is FF-GenAI Dataset with a total of approximately 100,000 images, which consists of 50,000 real and 50,000 hoax images [18] as shown in Figure 4. This dataset is utilized for training and evaluating the Xception model.

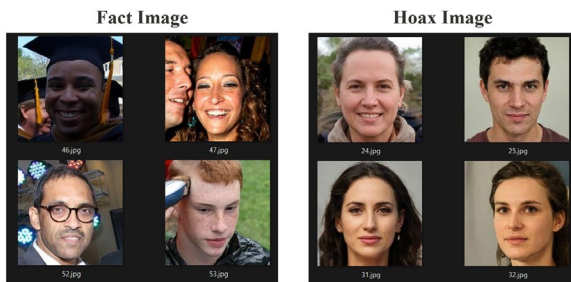


Figure 4. Fact and Hoax Images in the FF-GenAI Dataset

Both datasets are publicly accessible via Kaggle and contain facial images generated or manipulated using deep learning techniques such as Generative Adversarial Networks (GANs), alongside authentic facial images without digital manipulation. The use of two different datasets and model architectures allows this study to compare how different convolutional neural network architectures perform under different dataset characteristics and visual distributions.

2.2 Data Preparation

To guarantee that the input has a consistent format and features, all data are pre-processed before being utilized in the training process. This step ensures that the model receives consistent input, expedites training, and improves data quality as shown in Figure 5.

The preprocessing of data stage is performed to enhance data quality and consistency before the training process. First, an inspection of data quality is performed to identify images that are damaged, corrupted or unreadable. Images that do not fit the quality criteria are removed to avoid negative effects on the training

process and the overall performance of the model. Then the images are resized according to the input requirements of each architecture. The input size for MobileNet is 224×224 pixels and the input size Xception is 299×299 pixels. Standardizing input dimensions cuts down on unnecessary variability and improves computational efficiency during training [19].

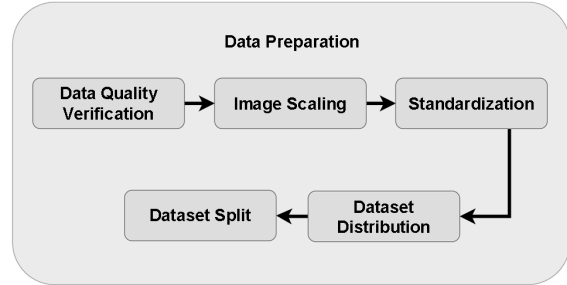


Figure 5. Data Preparation Diagram

The overall idea of resizing for image scaling is given by Equation 1.

$$Scaling(i) = \frac{target_size}{\max(width_i, height_i)} \quad (1)$$

In this implementation, images were resized directly using the default interpolation provided by the image processing library, without padding, cropping or explicit aspect-ratio preservation, thus producing a uniform tensor shape for the deep learning models.

After resizing, pixel values are standardized by normalizing the intensity range from $[0, 255]$ to $[0, 1]$. This normalization maintains the consistency of input scale, which helps stabilize gradient updates and speeds up model convergence [20]. The normalization procedure is given in Equation 2.

$$Normalization(i) = \frac{p}{255} \quad (2)$$

p represents the original pixel value.

The numbers of fact and hoax images are analyzed to evaluate the class distribution and to ensure that the data set is sufficiently balanced. ($N_{\{minor\}}$) and ($N_{\{major\}}$) are the number of the minority and majority classes, and the imbalance ratio is calculated using Equation 3. If the imbalance ratio is lower than a pre-defined threshold (e.g. 0.8), balancing techniques such as oversampling by data augmentation or undersampling are applied to preserve the class equilibrium.

$$ImbalanceRatio(i) = \frac{N_{minor}}{N_{major}} \quad (3)$$

ImbalanceRatio(i): represents the class imbalance ratio for the i -th dataset. i denotes the dataset (or data partition) being evaluated. $N_{\{minor\}}$ is the number of samples in the minority class. $N_{\{major\}}$ is the number of samples in the majority class.

Finally, the dataset is randomly shuffled and split into training, validation and testing subsets with a ratio of 70%, 20% and 10%, respectively. The number of

samples assigned to each of the subsets is calculated by Equations 4-6.

$$\begin{aligned} N_{train} &= [0.70 \times N] \\ N_{val} &= [0.20 \times N] \\ N_{test} &= [0.10 \times N] \end{aligned} \quad (4)$$

N is the total number of images in each class (fact or hoax), N_{train} is the number of training samples, N_{val} is the number of validation samples, N_{test} is the number of testing samples.

Finally, the data set is randomly shuffled and divided into training, validation and testing subsets with ratios of 70%, 20% and 10%, respectively. The number of samples assigned to each of the subsets are calculated by Equations 4-6.

2.3. Model Architecture

The model used in this research are MobileNet and Xception (Extreme Inception), both are Convolutional Neural Network (CNN) architecture. This study utilizes MobileNet as a comparative lightweight CNN architecture. MobileNet is specifically designed for computational efficiency and mobile-oriented deep learning applications. Like Xception, MobileNet employs depthwise separable convolution to reduce the number of parameters and computational cost while maintaining competitive classification performance [21]. Compared with conventional CNN architectures, MobileNet achieves faster inference and lower memory consumption, making it suitable for resource-constrained environments. The architecture extracts visual representations progressively from shallow to deeper layers, enabling the network to capture both local texture information and higher-level semantic patterns relevant to deepfake detection.

This study also utilized Xception depthwise separable convolution to separate the spatial convolution process from channel-wise feature merging, which combines multiple convolution kernels within a single block [22], [23]. The Xception architecture consists of 132 layers and is divided into three main stages. The architecture was adapted and redrawn based on the original Xception network. In general, the feature extraction process works as follows:

Low Level Representation. This stage is responsible for extracting low-level visual features. Through a combination of standard convolution layers and depthwise separable convolution layers, the Low-level layer captures basic patterns such as edges, contours, simple textures, and overall facial outlines. This stage establishes the foundational spatial representation that will be further refined in the deeper parts of the network.

Middle Level Abstraction. This stage is the core of the Xception network, where convolutional blocks are repeated eight times. Every block in the Middle level layer is built entirely using depthwise separable convolution layers, enabling highly efficient and fine-

grained feature extraction. In this stage, the network learns more complex and abstract patterns such as facial structure, skin texture variations, and subtle visual irregularities features.

High Semantic Feature. In this part of the network, advanced semantic features relevant to classification are emphasized. Similar to earlier flows, the High-level layer also utilizes depth wise separable convolution layers, followed by a final expansion in channel depth. The resulting feature maps are then condensed into a 2048-dimensional vector using Global Average Pooling, preserving only the most informative and discriminative features for the binary classification task.

In this research, a pre-trained Xception model based on ImageNet is utilized through transfer learning. The training process is conducted using three experimental configurations, in which each of the model's major architectural flows such as Low-level layer, Middle level layer, and High-level layer is independently fine-tuned and evaluated. This ablation-style comparison aims to investigate how different levels of feature abstraction within the Xception architecture influence the model's ability to recognize deepfake image patterns. By isolating and analyzing each flow, the study seeks to identify the most optimal configuration that provides the highest, most stable, and most computationally efficient performance.

2.4 Training

The training stage is conducted to enable the proposed deep learning architectures to classify facial images into two categories, namely fact and hoax. In this study, both Xception and MobileNet are implemented using transfer learning with ImageNet pre-trained weights. Transfer learning is utilized to accelerate convergence and improve feature learning capability by leveraging previously learned visual representations from large-scale image datasets.

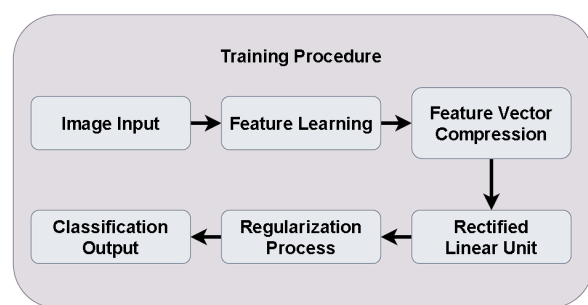


Figure 6. Model Training Workflow

As illustrated in Figure 6, the training process begins by feeding preprocessed facial images into the CNN backbone architecture. The input images are processed through multiple convolutional operations to extract facial features, which serve as the primary feature extractor [24], [25].

After passing through the base model, the resulting feature maps undergo feature vector compression via a Global Average Pooling layer. This operation averages

all values within each channel, producing a more compact one-dimensional feature vector [26]. In addition to reducing computing cost, this dimensionality reduction serves as a regularization strategy to avoid overfitting.

The pooled feature vector is then coupled to a Dense Layer via a ReLU (Rectified Linear Unit) activation function. A simple function that returns the identity function for positive input and zero for negative input [27]. The model can discover more intricate associations in the data by using this non-linear activation function. The definition of the ReLU function is shown in Equation 5.

$$f(x) = \max(0, x) \quad (5)$$

x denotes the input to the activation function, $f(x)$ is the corresponding output, and $\max(0, x)$ returns the greater value between 0 and x , allowing only positive activations to pass while setting negative values to zero.

Which has a non-saturating property that accelerates convergence during training.

To enhance regularization and prevent overfitting, a Dropout Layer with a rate of 30% is applied. This layer randomly deactivates 30% of the neurons in the preceding layer during training. This method compels the network to acquire more robust and generalizable properties by preventing the model from over-relying on specific neurons [28].

The Output Layer, which has a sigmoid activation function, is the model's last layer. The sigmoid function is ideal for producing an output probability value between 0 and 1 because this classification problem is binary (two classes). The definition of the sigmoid function is shown in Equation 6.

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (6)$$

$\sigma(z)$ is the output of the sigmoid activation function.

z is the input to the activation function, typically representing the weighted sum of the input features and bias from the previous layer. e is Euler's number (approximately 2.71828), the base of the natural logarithm. e^{-z} denotes the exponential function with exponent $-z$.

z is the input value to the output neuron [29]. This output value can be interpreted as the probability that an image belongs to the "hoax" category, while its complement is the probability for the "fact" category.

2.5 Evaluation

After training, the model is evaluated to assess how well it predicts new or previously unseen data. The evaluation employs a confusion matrix, a standard method in classification tasks that compares actual values to model predictions. Four evaluation metrics, including accuracy, precision, recall, and F1-score, can be computed from the confusion matrix shown in Equations 7-10.

Accuracy is the ratio of correct predictions to the total number of samples, whereas precision is the number of true positive predictions, which relates to the model's ability to avoid false positives [30]. Recall assesses the model's ability to accurately identify all positive samples and is related with fewer false negatives [31]. A balanced performance evaluation is provided by the F1-score, which is a harmonic mean of precision and recall, particularly in situations when the data distribution between classes is unbalanced [32].

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \quad (7)$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (8)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (9)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

The confusion matrix and these four metrics allow the model's performance to be evaluated more objectively and fully, not only by looking at the number of right predictions, but also by understanding how the model handles classification errors for each class, both fact and hoax.

3. Results and Discussions

3.1 Performance Result

This experiment evaluates the classification performance of different CNN architectures and fine-tuning configurations for detecting fact and hoax facial images. The experiments were conducted using two different datasets and model architectures, namely MobileNet trained on the DeepDetect2025 dataset and Xception trained on the FF-GenAI dataset.

For the Xception architecture, three fine-tuning configurations were evaluated based on different representation stages: low level layer, middle level layer, and high-level layer. The objective of this comparison was to analyze how different feature abstraction levels influence the model's capability to distinguish authentic and AI-generated facial images.

Table 1. Performance Comparison of CNN Architectures and Fine-Tuning Configurations

Model	Accuracy	Precision	Recall	F1-Score
MobileNet	96%	93.50%	98.86%	96.11%
Xception (Low level layer)	98.03%	98.78%	97.26%	98.03%
Xception (Middle level layer)	99.71%	99.68%	99.74%	99.71%
Xception (High level layer)	99.14%	99.14%	99.10%	99.14%

Based on Table 1, the results show that Xception consistently outperformed MobileNet across all evaluation metrics. The middle level layer fine-tuning achieved the best performance with 99.71% accuracy

and F1-score, indicating that intermediate feature representations are the most effective for detecting manipulation artifacts in AI-generated facial images. In comparison, the low-level layer captured only basic visual features, while the high-level layer focused more on global semantic information, resulting in slightly lower performance. Although MobileNet achieved lower accuracy, the model still demonstrated competitive performance with advantages in computational efficiency and faster inference.

Table 2. Performance Comparison with Previous Studies

Study	Model / Method	Dataset	Accuracy
Jagdale et al. (2025)	Meso4	Celeb-DF and FaceForensics++	92%
Demir et al. (2025)	DenseNet	OpenForensics dataset	93%
Taeb et al. (2022)	VGG-19	Custom Dataset	95%
Al-Dulaimi et al. (2024)	Hybrid CNN-LSTM Model	DFDC	98.21%
Jiang et al. (2025)	Custom CNN framework: RMLD-HFTF	DeepFake (DF)	99.56%
Proposed Method	Xception	FF-GenAI	99.71%

Table 2 presents a comparison between the proposed method and several previous deepfake or deepphoax detection studies. As shown in the table, the proposed Xception-based model achieves 99.71% accuracy, which is higher than several previous approaches such as Meso4, DenseNet, and VGG-19. The strong performance of the proposed method can be attributed to the effectiveness of depthwise separable convolution in extracting subtle facial manipulation artifacts while maintaining computational efficiency. Furthermore, the fine-tuning strategies applied to specific components of the Xception architecture enable the model to learn more discriminative representations for detecting manipulated facial images. Although the datasets and experimental settings vary across studies, the comparison shows that the proposed method exhibits competitive performance relative to existing deepfake and deepphoax image detection approaches.

3.2 Confusion Matrix Analysis

After the training process was completed, the model's performance in classifying real and fake facial images was evaluated using a confusion matrix. This method compares actual labels with predicted labels, allowing for a detailed analysis of classification performance, including correct predictions, false positives, and false negatives. The evaluation was conducted using a separate test dataset to ensure that the results reflected the model's performance on previously unseen data.

The confusion matrix for the low-level layer configuration in Figure 7 shows several misclassifications, particularly where fake images were predicted as real. This indicates that improvements to only the early layers of the Xception architecture are insufficient to capture the complex visual features needed to detect manipulated facial images.

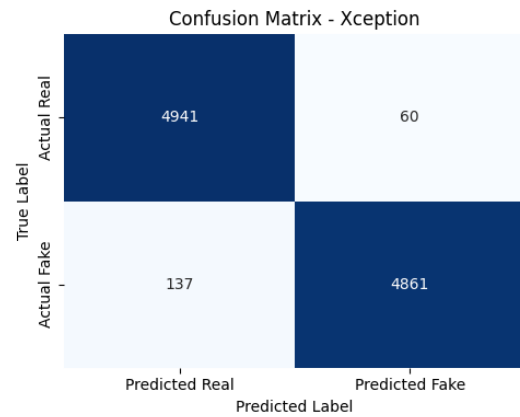


Figure 7. Confusion Matrix Low-level layer

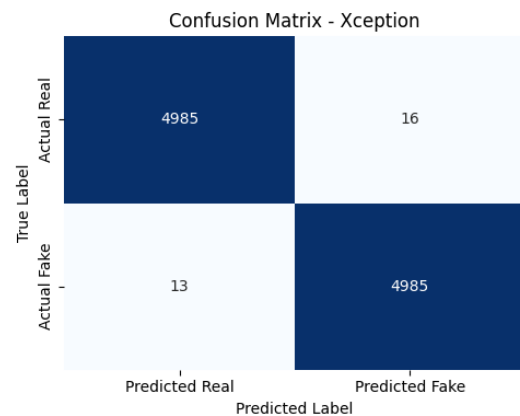


Figure 8. Confusion Matrix Middle level layer

In contrast, the Middle level layer configuration shown significant performance improvements in Figure 8, with more accurate predictions for both classes. This improvement can be attributed to the middle layers of Xception, which extract more discriminatory mid- to high-level features through deep-segregated convolution blocks.

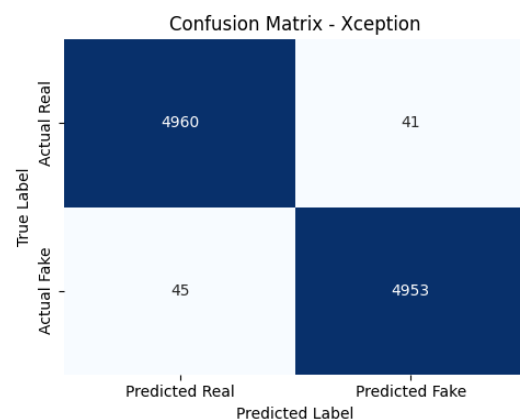


Figure 9. Confusion Matrix High level layer

The confusion matrix for the High-level layer configuration shown in Figure 9 also demonstrates strong performance with minimal misclassifications. The deeper layers of Xception can capture high-level features such as subtle textures, manipulation artifacts,

and abnormal pixel patterns commonly found in deep-hoax images.

Figure 10 presents the confusion matrix for MobileNet. Although MobileNet produced lower accuracy compared to Xception, the model still demonstrated relatively strong classification capability with high recall and a limited number of classification errors. These results indicate that lightweight architectures remain effective for deephoax detection while providing advantages in computational efficiency and faster inference speed.

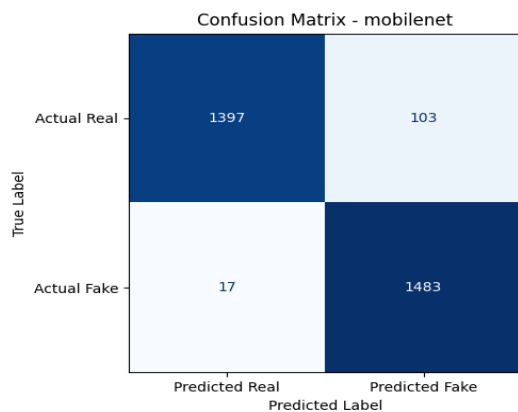


Figure 10. Confusion Matrix MobileNet

Overall, the results show that deeper layer refinement significantly improves classification performance. The Middle level layer configuration achieved the highest accuracy on internal test data, while the High-level layer configuration demonstrated better stability when evaluated on previously unseen external data. These findings demonstrate that Xception's deeper layer refinement provides a more robust and reliable approach for detecting both real and AI-generated facial images.

3.3 Limitations and Future Work

This study evaluated the effectiveness of Xception and MobileNet architectures for facial deephoax classification. Although the models achieved high performance, several limitations remain, including the use of controlled facial image datasets with limited environmental variation and evaluation restricted to static images from only two datasets, namely FF-GenAI and DeepDetect2025. These limitations may reduce the model's generalization capability when applied to real-world scenarios involving extreme poses, occlusion, motion blur, low-resolution images, or diverse AI-generated manipulation techniques.

Future work should expand the dataset to include more facial variations, environmental conditions, and diverse types of manipulation. Techniques such as data augmentation and adversarial training can also improve the model's robustness. Furthermore, extending the detection framework to video-based analysis and integrating the model into real-time or mobile applications would enable more practical evaluation in

real-world scenarios and improve the effectiveness of the deep hoax detection system.

4. Conclusions

Based on the experimental results, both Xception and MobileNet architectures demonstrated strong performance in classifying facial images into fact and hoax categories, although Xception consistently outperformed MobileNet across all evaluation metrics. Among the evaluated configurations, the middle level layer fine-tuning of Xception achieved the best performance with 99.71% accuracy and F1-score, indicating that intermediate feature representations are the most effective for capturing structural inconsistencies, texture irregularities, and synthetic blending artifacts commonly found in AI-generated facial images. The study also revealed that different representation levels within the Xception architecture contribute differently to deephoax detection, where low-level features were less discriminative and highly abstract semantic features slightly reduced sensitivity to localized manipulation artifacts. Overall, this research contributes to the development of an AI-based deephoax image detection framework through the comparative evaluation of Xception fine-tuning strategies and the implementation of a lightweight MobileNet baseline, supporting future applications in digital content verification, cybersecurity, and visual misinformation mitigation.

References

- [1] A. Mulahuwaish *et al.*, "A survey of social cybersecurity: Techniques for attack detection, evaluations, challenges, and future prospects," May 01, 2025, Elsevier B.V. doi: <https://doi.org/10.1016/j.chbr.2025.100668>.
- [2] S. Y. Yakub *et al.*, "Cyber Security for Hoax News Detection with Similarity Algorithm," in *2024 9th International Conference on Informatics and Computing, ICIC 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: <https://doi.org/10.1109/ICIC64337.2024.10956546>.
- [3] E. Broda and J. Strömbäck, "Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review," *Ann. Int. Commun. Assoc.*, vol. 48, no. 2, pp. 139–166, 2024. doi: <https://doi.org/10.1080/23808985.2024.2323736>.
- [4] S. Y. Yakub *et al.*, "Deep Learning for Cyber Security: Deephoax Detection," in *2025 2nd International Conference on Advanced Innovations in Smart Cities, ICAISC 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: <https://doi.org/10.1109/ICAISC64594.2025.10959570>.
- [5] S. Alanazi and S. Asif, "Exploring deepfake technology: creation, consequences and countermeasures," *Human-Intelligent Systems Integration*, vol. 6, no. 1, pp. 49–60, Dec. 2024. doi: <https://doi.org/10.1007/s42454-024-00054-8>.
- [6] A. T. Wibowo, "Hoax and Fake News by Saracen Syndicate and the Problems for National Cyber Security," *Indonesian Journal of Counter Terrorism and National Security*, vol. 1, no. 1, pp. 91–108, Jan. 2022. doi: <https://doi.org/10.15294/ijctns.v1i1.56732>.
- [7] H. Sun, T. Zhu, Z. Zhang, D. Jin, P. Xiong, and W. Zhou, "Adversarial Attacks Against Deep Generative Models on Data: A Survey," Dec. 2021. doi: <https://doi.org/10.1109/TKDE.2021.3130903>.
- [8] L. Zhang *et al.*, "X-Transfer: A Transfer Learning-Based Framework for GAN-Generated Fake Image Detection," Jan. 2024. doi: <https://doi.org/10.1109/IJCNN60899.2024.10650566>.

- [9] Tim Cek Fakta, "[HOAKS] Video Prabowo Menyebutkan NIK KTP Berisi Bantuan Rp 7 Juta," *Kompas*, Oct. 11, 2025. Accessed: Nov. 27, 2025. [Online]. Available: <https://www.kompas.com/cekfakta/read/2025/10/11/084800182/-hoaks-video-prabowo-menyebutkan-nik-ktp-berisi-bantuan-rp-7-juta>
- [10] M. Linawati, "Cek Fakta: Tidak Benar Video Prabowo Cairkan Bansos Rp 7 Juta per KTP," *Liputan6*, Jakarta, Sep. 09, 2025. Accessed: Nov. 27, 2025. [Online]. Available: https://www.liputan6.com/cek-fakta/read/6154574/cek-fakta-tidak-benar-video-prabowo-cairkan-bansos-rp-7-juta-per-ktp#google_vignette
- [11] Z. Xia, T. Qiao, M. Xu, X. Wu, L. Han, and Y. Chen, "Deepfake Video Detection Based on MesoNet with Preprocessing Module," *Symmetry (Basel)*, vol. 14, no. 5, May 2022, doi: <https://doi.org/10.3390/sym14050939>.
- [12] L. E. Demir and Y. Canbay, "Deepfake Image Detection with Transfer Learning Models," *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, vol. 14, no. 1, pp. 546–560, Mar. 2025, doi: <https://doi.org/10.17798/bitlisfen.1610300>.
- [13] Y. Xu and S. Y. Yayilgan, "When Handcrafted Features and Deep Features Meet Mismatched Training and Test Sets for Deepfake Detection," Sep. 2022, doi: <https://doi.org/10.48550/arXiv.2209.13289>.
- [14] H. Huo, Y. L. Yu, and Z. H. Liu, "Facial expression recognition based on improved depthwise separable convolutional network," *Multimed. Tools Appl.*, vol. 82, no. 12, pp. 18635–18652, May 2023, doi: <https://doi.org/10.1007/s11042-022-14066-6>.
- [15] Q. Jiang et al., "Robust manipulated media localization and detection based on high frequency and texture features," *Discover Computing*, vol. 28, no. 1, Dec. 2025, doi: <https://doi.org/10.1007/s10791-025-09500-w>.
- [16] B. Dhanalaxmi, B. N. Kumar, Y. Raju, and R. S. R. Channapragada, "MobileNetV3: an efficient deep learning-based feature selection and classification technique for cardiovascular disease," *Journal of Engineering and Applied Science*, vol. 72, no. 1, Dec. 2025, doi: <https://doi.org/10.1186/s44147-025-00654-4>.
- [17] "DeepDetect-2025," Kaggle. Accessed: Dec. 14, 2025. [Online]. Available: <https://www.kaggle.com/datasets/ayushmandatta1/deepdetect-2025>
- [18] "Flickr-Face-HQ and GenAI Dataset (FF-GenAI)," Kaggle. Accessed: Oct. 28, 2025. [Online]. Available: <https://www.kaggle.com/datasets/argonautex/flickr-face-hq-and-genai-dataset-ff-genai>
- [19] M. Bunde and G. M. Danciu, "Pneumonia Image Classification Using DenseNet Architecture," *Information (Switzerland)*, vol. 15, no. 10, Oct. 2024, doi: <https://doi.org/10.3390/info15100611>.
- [20] M. Abbasi, P. Váz, J. Silva, and P. Martins, "Comprehensive Evaluation of Deepfake Detection Models: Accuracy, Generalization, and Resilience to Adversarial Attacks," *Applied Sciences (Switzerland)*, vol. 15, no. 3, Feb. 2025, doi: <https://doi.org/10.3390/app15031225>.
- [21] H. Noprisson, V. Ayumi, M. Purba, and N. Ani, "MOBILENET PERFORMANCE IMPROVEMENTS FOR DEEPAKE IMAGE IDENTIFICATION USING ACTIVATION FUNCTION AND REGULARIZATION," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 10, no. 2, pp. 441–448, Nov. 2024, doi: <https://doi.org/10.33480/jitk.v10i2.5798>.
- [22] A. Jagdale, V. Kubde, R. Kortikar, and N. Rajgure, "Deepfake Image Detection using Deep Learning," *International Journal of Scientific Research in Engineering and Management*, 2025, doi: <https://doi.org/10.55041/IJSREM46579>.
- [23] H. M. Lala and H. S. Ahmed, "An impact on convolutional neural networks amelioration for early detection of skin cancer," *International Journal of Intelligent Computing and Information Sciences*, vol. 24, no. 2, pp. 1–17, Jun. 2024, doi: <https://doi.org/10.21608/ijicis.2024.273450.1325>.
- [24] R. Adityatama and A. T. Putra, "Image classification of Human Face Shapes Using Convolutional Neural Network Xception Architecture with Transfer Learning," *Recursive Journal of Informatics*, vol. 1, no. 2, pp. 102–109, Sep. 2023, doi: <https://doi.org/10.15294/rji.v1i2.70774>.
- [25] R. Yao, Z. Bai, J. Tong, and K. Rezaee, "Deepfake Detection in Image Sequences: A Temporal Approach for Anomaly Detection," *International Journal of Intelligent Systems*, vol. 2025, no. 1, 2025, doi: <https://doi.org/10.1155/int/8566328>.
- [26] B. Gülmez, "A novel deep neural network model based Xception and genetic algorithm for detection of COVID-19 from X-ray images," *Ann. Oper. Res.*, vol. 328, no. 1, pp. 617–641, Sep. 2023, doi: <https://doi.org/10.1007/s10479-022-05151-y>.
- [27] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark," Jun. 2022, doi: <https://doi.org/10.1016/j.neucom.2022.06.111>.
- [28] I. Salehin and D. K. Kang, "A Review on Dropout Regularization Approaches for Deep Neural Networks within the Scholarly Domain," Jul. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: <https://doi.org/10.3390/electronics12143106>.
- [29] H. Kamal and M. Mashaly, "Enhanced Hybrid Deep Learning Models-Based Anomaly Detection Method for Two-Stage Binary and Multi-Class Classification of Attacks in Intrusion Detection Systems," *Algorithms*, vol. 18, no. 2, Feb. 2025, doi: <https://doi.org/10.3390/a18020069>.
- [30] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: <https://doi.org/10.1038/s41598-024-56706-x>.
- [31] C. Miller, T. Portlock, D. M. Nyaga, and J. M. O'Sullivan, "A review of model evaluation metrics for machine learning in genetics and genomics," 2024, *Frontiers Media SA*. doi: <https://doi.org/10.3389/fbinf.2024.1457619>.
- [32] C. Sweeney, E. Ennis, M. Mulvenna, R. Bond, and S. O'Neill, "How Machine Learning Classification Accuracy Changes in a Happiness Dataset with Different Demographic Groups," *Computers*, vol. 11, no. 5, May 2022, doi: <https://doi.org/10.3390/computers11050083>.