Comparison of ResNet, EfficientNet, and Xception architectures for deepfake detection*

Khrystyna Lipianina-Honcharenko[1,†], Mykola Telka[1,*,†] and Nazar Melnyk[1,†]

[1] *West Ukrainian National University, Lvivska str., 11, Ternopil, 46000, Ukraine*

Abstract

This study presents a comparative analysis of three deep neural networks—ResNet, EfficientNet, and Xception—for deepfake video detection tasks. The primary goal was to identify the most effective architecture for classifying fake videos, as well as to explore additional mechanisms, such as Long Short-Term Memory (LSTM) and attention mechanisms, which could enhance the accuracy of the models. Using a dataset consisting of real and fake videos, each model was evaluated based on accuracy, precision, recall, and F1-score metrics. The results showed that the Xception model achieved the highest accuracy (87.7%), while EfficientNet also demonstrated high efficiency, particularly in resource-constrained tasks. ResNet showed stability but faced challenges in classifying underrepresented classes.

Keywords

deepfake, ResNet, EfficientNet, Xception

1. Introduction

The proliferation of deepfake videos poses a significant threat to digital security and information trust, creating challenges across various sectors, including media, politics, and legal systems. Deepfake technologies facilitate the creation of highly realistic yet fabricated videos, making their detection challenging for conventional methods. This contributes to the manipulation of public opinion [1], facilitates the dissemination of false information, and enables malicious activities such as deception and fraud. Therefore, developing effective systems for the automatic detection of deepfake videos is critically important for ensuring information security and combating disinformation [2].

This study aims to evaluate different deep neural network architectures, such as ResNet, EfficientNet, and Xception, for deepfake video detection. The primary focus is on identifying the most effective models and exploring the role of additional mechanisms, including LSTM and attention, in improving detection accuracy. The study assesses the performance of the models using metrics such as accuracy, recall, precision, and F1-score, ultimately identifying the best approaches for building reliable deepfake detection systems.

Future research could address current limitations by implementing advanced data augmentation techniques to balance datasets, exploring ensemble models to combine the strengths of multiple architectures, and optimizing computational efficiency for deploying lightweight models in real-world scenarios.

The paper is organized into several sections: Section 2 reviews existing deepfake detection methods and the neural network architectures commonly employed. Section 3 outlines the research methodology, detailing data preparation, model selection, and training processes. Section 4 provides

AdvAIT-2024: 1st International Workshop on Advanced Applied Information Technologies, December 5, 2024, Khmelnytskyi, Ukraine - Zilina, Slovakia

* Corresponding author.

† These authors contributed equally.

xrustya.com@gmail.com (K. Lipianina-Honcharenko); telkamikola@gmail.com (M. Telka); 88nazar88@gmail.com (N. Melnyk)

0000-0002-2441-6292 (K. Lipianina-Honcharenko); 0009-0002-4293-7515 (M. Telka); 0009-0000-5917-1099 (N. Melnyk)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

a comprehensive performance analysis of the models based on accuracy, recall, precision, and F1-score. Finally, Section 5 highlights the most effective approaches and proposes recommendations for advancing deepfake detection systems.

2. Related work

Recent studies in the field of deepfake video detection focus on utilizing deep neural networks, such as ResNet, EfficientNet, and Xception, to improve the accuracy of classifying real and fake videos. For instance, ResNet-50 is used for deepfake video detection by combining it with LSTM to account for both images and video frame sequences. This allows the model to consider temporal dependencies, significantly enhancing accuracy compared to methods that use only individual frames [3]. Additionally, other studies, such as those involving Inception-ResNet-V2, emphasize the necessity of developing effective deepfake detection methods due to security and privacy threats [4].

Other approaches concentrate on developing more complex architectures. Specifically, the Sequential-Parallel Networks (SPNet) model offers a novel method for processing deepfake videos, providing more efficient handling of spatiotemporal dependencies with a reduced number of parameters [5]. This architecture helps lower computational costs, which is a crucial factor when working with large volumes of video data. Furthermore, a five-layer convolutional neural network proposed in another study demonstrates a high accuracy of 98% compared to other models, such as Xception and EfficientNet-B0 [6].

In addition to these recent approaches, models that use attention mechanisms, such as channel and spatial attention, show significant improvements in deepfake detection accuracy compared to standard models. The use of attention mechanisms allows the model to focus on important features of the input data, which is particularly beneficial for complex detection tasks, such as identifying fake videos [7]. Other studies, including reviews of deepfake detection methods using ResNet, EfficientNet, and Xception, also confirm the effectiveness of these architectures in deep learning tasks [8]. Similarly, research involving the use of Xception and ResNet-50 in combination with Local Binary Pattern (LBP) for deepfake video classification demonstrates the effectiveness of image processing and the accuracy of these models [9].

In the work on deepfake detection using ResNext50 and LSTM, researchers significantly improved accuracy by integrating temporal dependency analysis. This approach enables not only the detection of individual frames but also the analysis of their interrelationships [10]. Finally, the use of Generative Adversarial Networks (GAN) combined with CNN has helped reduce computational costs by selecting key video frames to enhance results [11], making this approach promising in combating deepfake videos.

Most comparisons show that models like Xception and EfficientNet significantly outperform ResNet in deepfake detection tasks due to their ability to process textures and fine image details more effectively. Xception, with its architecture of deep separable convolutions, allows for a reduction in the number of parameters without sacrificing accuracy, making it particularly useful in resource-constrained environments. EfficientNet, in turn, offers optimal scaling of model depth, width, and resolution, leading to better performance compared to ResNet. However, ResNet remains an important foundational architecture, especially when used in combination with mechanisms like LSTM for handling temporal dependencies, making it effective in tasks that analyze both individual frames and video sequences [3][4][5].

Given this context, the aim of this study is to compare the effectiveness of different deep neural networks, such as ResNet, EfficientNet, and Xception, in deepfake video detection tasks. Special attention is given to how the architectural features of each model impact their ability to accurately classify fake videos and optimize their performance in resource-constrained conditions. Additionally, the study examines the role of supplementary mechanisms, such as LSTM and attention methods, which can enhance deepfake detection accuracy by combining the processing of both individual frames and temporal sequences.

3. Research methodology

3.1. Research architecture

The research architecture (Figure 1) for evaluating model accuracy in deepfake detection tasks is described below. The process begins with the initialization of the environment, including the import of necessary libraries and metadata loading. Next, data preprocessing is carried out, which involves reading metadata, randomly selecting a subset of videos, reading video files, extracting frames, and splitting the data into training and testing sets. Following this, model preparation takes place, where ResNet50, EfficientNet, and Xception are initialized and configured for binary classification. During the training phase, the models are trained on the training data, and their evaluation is conducted on the test data, with accuracy calculations. The process concludes with comparing the results of the three models based on the obtained accuracy metrics.

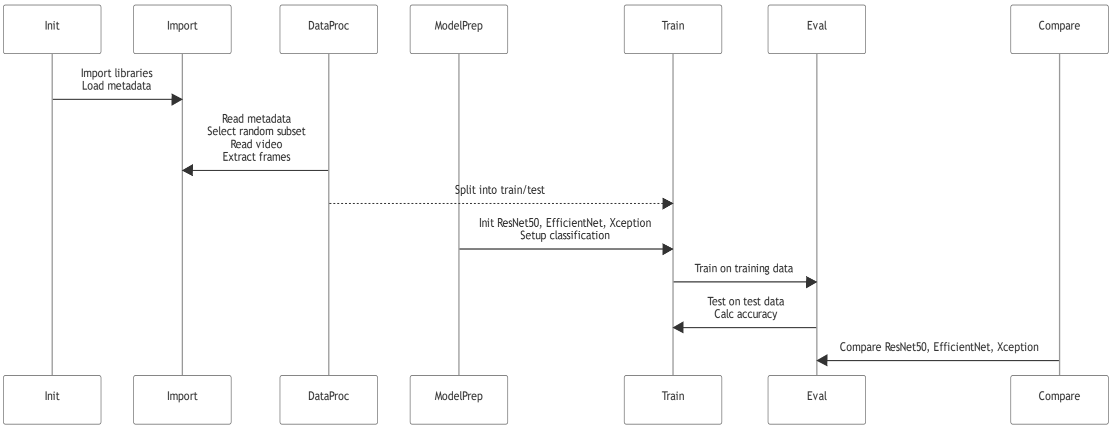


Figure 1: Research Architecture.

3.2. Model descriptions

ResNet50 [12], EfficientNetB0 [13], and Xception [14] are deep convolutional neural networks designed for feature extraction from images, each with unique characteristics in their approaches to scaling and optimization. All three models accept input tensors $X \in R^{H\times W\times C}$, where H, W, and C denote the image's height, width, and the number of channels, respectively. The output of the convolutional blocks in each model is a feature tensor $F \in R^{h\times w\times c}$, which is then transformed into a one-dimensional vector $f = Flatten(F)$ using a Flatten operation for further processing in dense layers.

ResNet50 [12] utilizes a convolutional layer architecture that includes "skip connections" to prevent gradient vanishing during the training of deep networks. Each ResNet block involves a sequence of convolutions, followed by adding the block input to its output before activation, which is mathematically described as

$$Fout = f(K, X) + X, \tag{1}$$

These skip connections help maintain information flow through the network and reduce problems related to network depth.

EfficientNetB0 optimizes its architecture using the composite scaling method, which simultaneously scales depth, width, and input size to balance accuracy and efficiency. The architecture employs depthwise separable convolutions, which reduce the number of parameters and computational operations by first applying depthwise convolutions independently on each channel and then using 1×11 convolutions to combine the channels.

Xception is an "extreme" version of the Inception architecture, where standard convolutions are entirely replaced by depthwise separable convolutions for each spatial point and channel. This approach not only reduces the number of parameters but also allows for more efficient feature extraction by utilizing a greater number of independent operations. The model uses a sequence of depthwise and pointwise convolutions in each layer, enabling better adaptation to diverse visual patterns in the data.

All three models use dense layers for further processing of the feature vector f and an output layer with sigmoid activation for classification, underscoring their versatility and effectiveness in modern computer vision tasks.

The integration of the Swish activation function [15] and the Dropout technique [16] into the ResNet50, EfficientNetB0, and Xception models can significantly enhance their performance and generalization capabilities. Swish is a smoothly varying nonlinear activation function defined as

$$Swish(x) = x \cdot \sigma(x), \tag{2}$$

where $\sigma(x)$ is the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$. This function has been proposed as an alternative to ReLU due to its ability to mitigate the issue of dead neurons, allowing smoother propagation of negative values and improving gradient flow in deep networks.

Dropout, on the other hand, is a regularization technique that helps prevent overfitting by randomly dropping out neurons during training. This forces the network to learn to be less reliant on specific features, thus enhancing its robustness and ability to generalize to new data. In the ResNet50, EfficientNetB0, and Xception models, applying Dropout in high-level dense layers can help manage model complexity, reducing the risk of overfitting the large number of weights these models have.

The combination of Swish and Dropout in these models can be particularly advantageous for tasks with large and complex datasets, where flexible activation and robust regularization are needed. Using Swish can improve the models' learning capability in deep layers, where traditional activation functions like ReLU may encounter limitations. Meanwhile, Dropout provides the additional benefit of encouraging the network to distribute useful information across a greater number of neurons, reducing the weight that any single neuron has on the model's decision.

A comparative table of ResNet, EfficientNet, and Xception models is presented below, highlighting their key characteristics, features, and advantages in the context of video data processing.

Table 1
Comparison of ResNet, EfficientNet, and Xception Models

Characteristic	ResNet	EfficientNet	Xception
Architecture	Residual network (residual blocks)	Balanced depth, width, and resolution	Depthwise separable convolution
Key Features	Contours, textures, objects, scenes	Contours, textures, object details	Fine details, textures, artifacts
Network Depth	Deep (up to 152 layers)	Balance of depth and performance	Very deep with efficient convolutions
Scalability	Difficult to scale	Efficient for different scales	Well-scaled but computationally intensive
Texture Processing	Good at detecting textures at high levels	Detects textures efficiently due to balanced scaling	Specializes in detailed textures and anomalies
Object and Scene Processing	Performs well with large objects and scenes	Optimized for various scenes and objects	Especially effective for detecting anomalies in objects

Video Processing Capabilities	Can handle temporal sequences (with LSTM [17])	Efficiently processes video frames due to flexible scaling	Detects artifacts, particularly useful for deepfake detection
Advantages	- Learns features at various levels: from simple to complex - Effective when combined with LSTM [17] for sequence analysis	- High efficiency due to balanced scaling - Suitable for large and complex images and videos	- Focused on artifact detection - Performs well in deepfake detection tasks
Disadvantages	- High computational resource requirements - Training very deep models is challenging	- May be less accurate without proper scaling	- Computationally intensive - Challenging to train due to deep convolutions
Use in Video Tasks	Extracts multi-level features (contours, objects, scenes); well-suited for analyzing temporal changes	Effective for processing videos with large or complex scenes; suitable for tasks requiring a balance of accuracy and efficiency	Excellent for detecting anomalies and artifacts in videos, especially useful for detecting fake videos (deepfake)

4. Research results

This study conducted a comparison of the performance of three deep learning models—ResNet50, EfficientNetB0, and Xception—in the task of image-based data classification. Each model was trained for ten epochs, and the results were evaluated using accuracy metrics [18], precision, recall, F1-score, as well as a confusion matrix for each model. Below is a detailed analysis of each model's performance.

The dataset used for this study was sourced from the Deepfake Detection Challenge on the Kaggle platform (Kaggle, 2020) [19]. It contains videos classified into two categories: "REAL" and "FAKE." After preprocessing, 480 samples were obtained, with 60 (12.5%) belonging to the "REAL" class and 420 (87.5%) to the "FAKE" class. The videos were standardized by frame size and used as input to pretrained neural networks for classification. The uneven class distribution reflects a real-world scenario, which is typical for deepfake detection tasks.

For each video, multiple frames were processed and converted into tensors for use in neural networks. All videos were standardized by frame size, and extracted features from these frames were fed into the pretrained models (ResNet50, EfficientNet, Xception).

Regarding the training process (Figure 2), all three models showed stable improvement in metrics on the training sets; however, significant fluctuations were observed during validation, indicating possible overfitting or sensitivity to parameter selection and data structures. Notably, in epochs 8-10, the models experienced some degradation in validation loss (val_loss), suggesting that the models began to overfit after a certain number of epochs.

The ResNet50 model demonstrated strong stability during training, achieving accuracy above 80%, but faced challenges in classifying the "REAL" class. This emphasizes the importance of further work on data balancing to improve the model's performance on minority classes.

EfficientNetB0, thanks to its efficient architecture, showed good performance in classifying both classes, maintaining high accuracy for the "FAKE" class while also delivering better results for the

"REAL" class compared to ResNet50. Its performance could be improved through more aggressive regularization to avoid the overfitting observed in later training stages.

Xception, with its use of depthwise separable convolutions, achieved the best results in overall accuracy and balance between classes. This suggests that its architecture is better suited for complex image classification tasks, with fewer parameters that reduce the risk of overfitting.

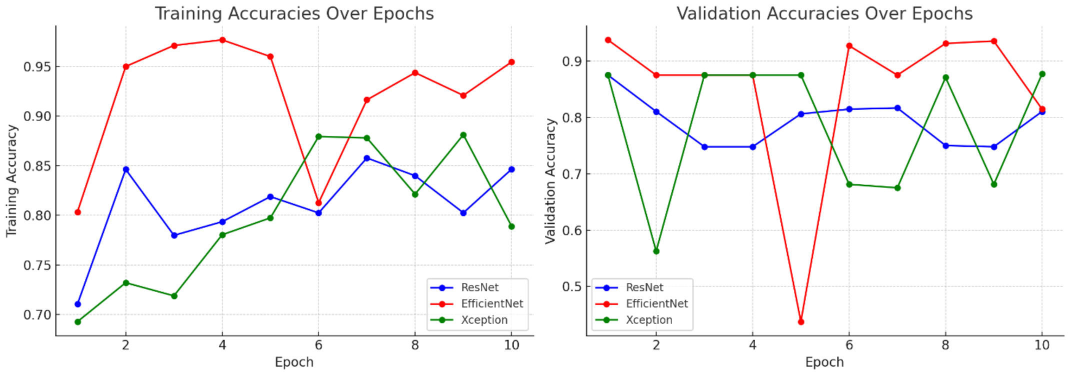
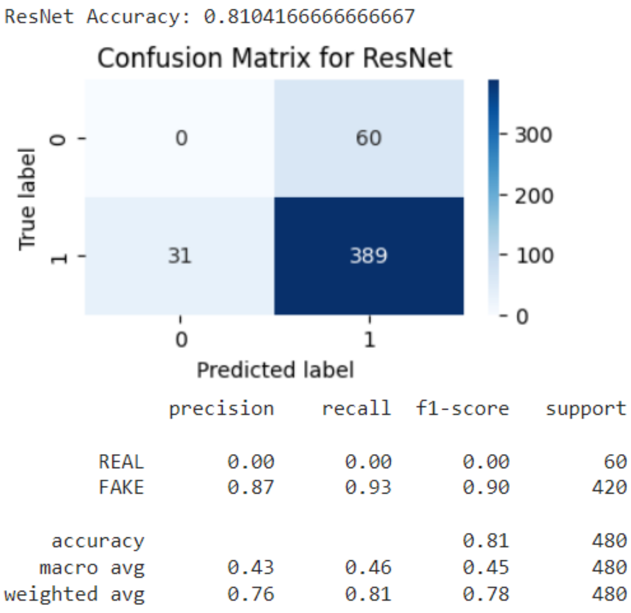


Figure 2: Training and Validation Accuracy Trends across 10 Epochs.

Next, the results were evaluated using accuracy, precision, recall, F1-score, and confusion matrices for each model.

ResNet50 (Figure 3) achieved an accuracy of 81.0% but faced difficulties in classifying the "REAL" class, failing to correctly classify any instances of this class, as clearly shown in the confusion matrix. This indicates an imbalance in model performance, which, despite high accuracy for the "FAKE" class (93% recall), could not accurately classify the "REAL" class. This limitation may be related to the network's depth and the need for additional regularization or data processing to balance the classes.



```
ResNet Accuracy: 0.8104166666666667
```

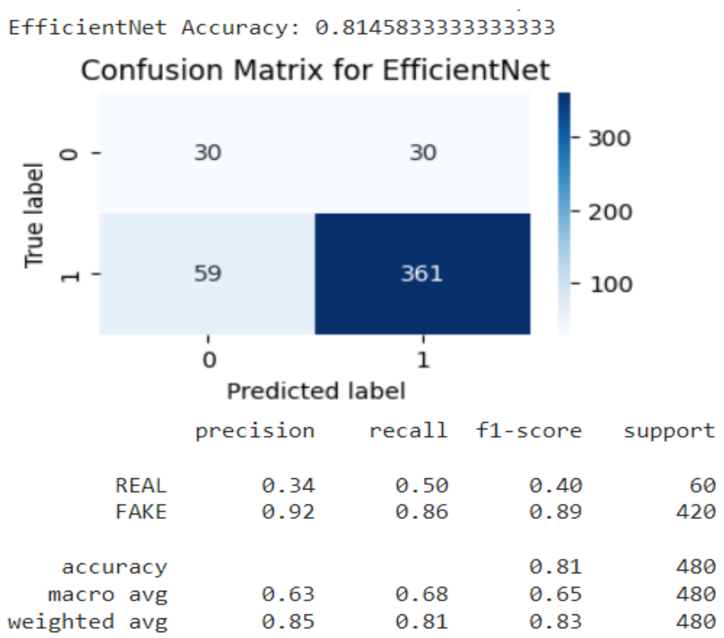
```
              precision    recall  f1-score   support

        REAL       0.00      0.00      0.00        60
        FAKE       0.87      0.93      0.90       420

    accuracy                           0.81       480
   macro avg       0.43      0.46      0.45       480
weighted avg       0.76      0.81      0.78       480
```

Figure 3: Confusion Matrix for ResNet.

EfficientNetB0 (Figure 4) reached an accuracy of 81.5%, slightly better than ResNet50. The model performed better in classifying the "REAL" class with a precision of 0.34 and recall of 0.50, representing a significant improvement compared to ResNet50. For the "FAKE" class, the model maintained high precision (0.92) and recall (0.86). This indicates that the EfficientNetB0 architecture is better optimized for resource-constrained tasks due to its scaling mechanism.



```
EfficientNet Accuracy: 0.8145833333333333
```

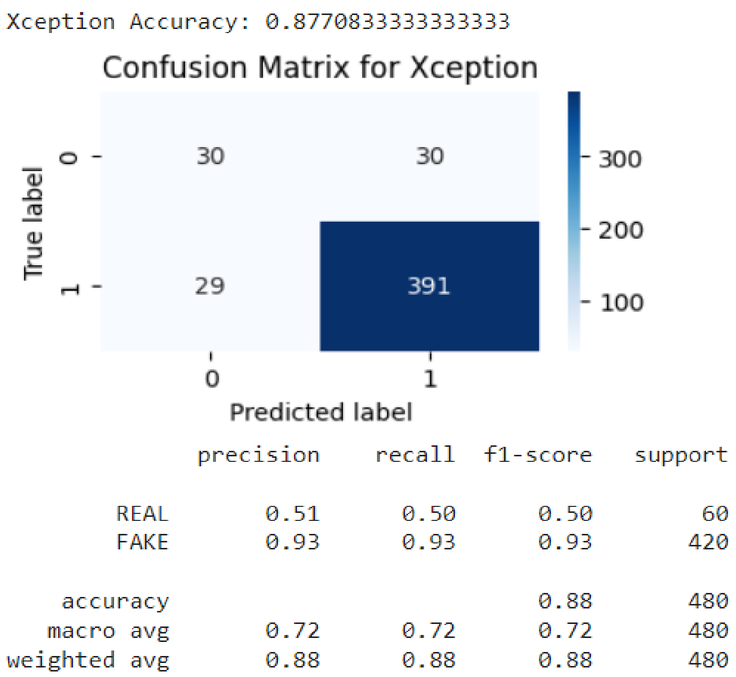
```
              precision    recall  f1-score   support

        REAL       0.34      0.50      0.40        60
        FAKE       0.92      0.86      0.89       420

    accuracy                           0.81       480
   macro avg       0.63      0.68      0.65       480
weighted avg       0.85      0.81      0.83       480
```

Figure 4: Confusion Matrix for EfficientNet.

Xception (Figure 5) achieved the highest accuracy among all models—87.7%. This model displayed balanced results for both classes, with precision and recall for the "REAL" class at 0.51 and 0.50, respectively, which is a significant improvement over the other models. For the "FAKE" class, precision and recall values were 0.93, demonstrating the model's strong ability to extract and utilize important features.



```
Xception Accuracy: 0.8770833333333333
```

```
              precision    recall  f1-score   support

        REAL       0.51      0.50      0.50        60
        FAKE       0.93      0.93      0.93       420

    accuracy                           0.88       480
   macro avg       0.72      0.72      0.72       480
weighted avg       0.88      0.88      0.88       480
```

Figure 5: Confusion Matrix for Xception.

Based on the obtained results, the idea of ensembling these three models—ResNet, EfficientNetB0, and Xception—could be a promising direction for further research. Using a combined approach, where the strengths of each model compensate for the weaknesses of others, could significantly improve the system's ability to generalize and its classification accuracy across a wide range of data. Xception, with its high accuracy and stability, could serve as the basis for accurate feature detection, while ResNet[20] and EfficientNetB0 could add robustness and computational efficiency, especially in resource-constrained environments. These initial observations encourage the development of a comprehensive ensemble model, which will be thoroughly analyzed and evaluated in future research projects aimed at optimizing detection and classification capabilities for modified images.

Conclusion

This study conducted a comparative analysis of three deep neural networks—ResNet, EfficientNet, and Xception—for deepfake video detection. The results indicate that Xception proved to be the most effective model for classifying fake videos, achieving an accuracy of 87.7% along with balanced precision and recall metrics for both classes. EfficientNet also demonstrated high performance, with an accuracy of 81.5% and superior results compared to ResNet in detecting the "REAL" class. ResNet, despite its stability in training and an accuracy of 81.0%, faced challenges in classifying videos of the "REAL" class, highlighting the need for further model improvements when working with imbalanced data.

The application of additional techniques, such as LSTM for handling temporal sequences, helped to enhance the accuracy of ResNet, demonstrating the importance of considering temporal dependencies in deepfake detection. However, models like Xception and EfficientNet, with their advanced architectures, significantly outperformed ResNet in deepfake detection tasks, providing more efficient feature extraction and better generalization capabilities.

For further improving deepfake video detection efficiency, a promising research direction is the use of model ensemble methods. Combining the strengths of different models, such as ResNet, EfficientNet, and Xception, could create a more robust and accurate detection system. Model ensembling can enhance the system's generalization ability and improve accuracy by reducing the risk of overfitting and compensating for the weaknesses of individual models. Future research should focus on developing effective ensemble approaches for deepfake detection, which could significantly improve results in real-world conditions.

Acknowledgments

This paper is supported by the EU Erasmus+ programme within the Capacity Building Project "WORK4CE" (619034-EPP-1–2020-1-UA-EPPKA2-CFHE-JP).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] K. Lipianina-Honcharenko, A. Melnychuk, K. Yurkiv, G. Hladiy, M. Telka (2024). Integrated Approach to the International Aspects of Online Dispute Resolution Formation. Proceedings of the First International Workshop of Young Scientists on Artificial Intelligence for Sustainable Development Ternopil, Ukraine, May 10-11, 2024. (pp. 88-98).

[2] K. Lipianina-Honcharenko, M. Soia, K. Yurkiv, A. Ivasechko (2024). Evaluation of the Effectiveness of Machine Learning Methods for Detecting Disinformation in Ukrainian Text Data. Proceedings of The Seventh International Workshop on Computer Modeling and Intelligent Systems (CMIS-2024), Zaporizhzhia, Ukraine, May 3, 2024. (pp. 97-109). https://ceur-ws.org/Vol-3702/paper9.pdf

[3] S. Rani, B. R., P. K. Pareek, B. S. (2023). Deepfake video detection system using deep neural networks. 2023 IEEE International Conference on Information and Communication Systems (ICICACS). https://dx.doi.org/10.1109/ICICACS57338.2023.10099618

[4] R. Rajalaxmi, P. S., A. M. Rithani, P. Dhivakar, G. E. (2023). Deepfake detection using Inception-ResNet-V2 network. 2023 International Conference on Computing and Communications (ICCMC). https://dx.doi.org/10.1109/ICCMC56507.2023.10083584

[5] R. Sun, Z. Zhao, L. Shen, Z. Zeng, Y. Li, B. Veeravalli, X. Yulei (2023). An efficient deep video model for deepfake detection. 2023 IEEE International Conference on Image Processing (ICIP). https://dx.doi.org/10.1109/ICIP49359.2023.10222682

[6] J. B. Awotunde, R. Jimoh, A. Imoize, A. T. Abdulrazaq, C. T. Li, C. C. Lee (2022). An enhanced deep learning-based deepfake video detection and classification system. Electronics, 12(1). https://dx.doi.org/10.3390/electronics12010087

[7] A. E. Bayar, C. Topal (2023). Deepfake detection via combining channel and spatial attention. 2023 IEEE Signal Processing Conference.

[8] A. Das, K. S. Viji, L. Sebastian (2022). A survey on deepfake video detection techniques using deep learning. 2022 International Conference on Next Generation Information Systems (ICNGIS). https://dx.doi.org/10.1109/ICNGIS54955.2022.10079802

[9] A. Arini, R. B. Bahaweres, J. Al Haq (2022). Quick classification of Xception and ResNet-50 models on deepfake video using Local Binary Pattern. IEEE Symposium on Artificial Intelligence and Multimedia (ISMODE).

[10] S. Z. Yunes Al-Dhabi (2021). Deepfake video detection by combining convolutional neural network (CNN) and recurrent neural network (RNN). 2021 IEEE International Conference on Artificial Intelligence and Smart Energy (CSAIEE). https://dx.doi.org/10.1109/CSAIEE54046.2021.9543264

[11] S. Lalitha, K. Sooda (2022). DeepFake detection through key video frame extraction using GAN. 2022 International Conference on Advanced Computing Research and Sustainability (ICACRS). https://dx.doi.org/10.1109/ICACRS55517.2022.10029095

[12] S. Targ, D. Almeida, K. Lyman (2016). Resnet in resnet: Generalizing residual architectures. arXiv preprint arXiv:1603.08029.

[13] K. Kansal, T. B. Chandra, A. Singh (2024). ResNet-50 vs. EfficientNet-B0: Multi-Centric Classification of Various Lung Abnormalities Using Deep Learning" Session id: ICMLDsE. 004". Procedia Computer Science, 235, 70-80.

[14] F. Chollet (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251-1258.

[15] R. Sudharsan, E. N. Ganesh (2022). A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy. Connection Science, 34(1), 1855-1876.

[16] S. Wager, S. Wang, P. S. Liang (2013). Dropout training as adaptive regularization. Advances in neural information processing systems, 26.

[17] H. Arab, I. Ghaffari, R. M. Evina, S. O. Tatu, S. Dufour (2022). A hybrid LSTM-ResNet deep neural network for noise reduction and classification of V-band receiver signals. IEEE Access, 10, 14797-14806.

[18] K. Lipianina-Honcharenko, V. Yarych, A. Ivasechko, A. Filinyuk, K. Yurkiv, T. Lebid, M. Soia (2024). Evaluating the Effectiveness of Attention-Gated-CNN-BGRU Models for Historical Manuscript Recognition in Ukraine. Proceedings of the First International Workshop of Young Scientists on Artificial Intelligence for Sustainable Development Ternopil, Ukraine, May 10-11, 2024. (pp. 99-108). https://ceur-ws.org/Vol-3716/paper8.pdf

[19] Kaggle (2020). Deepfake Detection Challenge. https://www.kaggle.com/competitions/deepfake-detection-challenge/data.

[20] S. Keerthana, N. Deepika, E. Pooja, I. Nandhini, M. Shanthalakshmi, G. R. Khanaghavalle (2024). An effective approach for detecting deepfake videos using Long Short-Term Memory and ResNet. 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT), 1–5. https://doi.org/10.1109/ic3iot60841.2024.10550265.