Comparative Analysis and Evaluation of CNN Models for Deepfake Detection

Pattrick Ritter
Computer Science Departement
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
pattrick.ritter@binus.ac.id

Devan Lucian
Computer Science Departement
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
devan.lucian@binus.ac.id

Anderies
Computer Science Departement
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
anderies@binus.ac.id

Andry Chowanda
Computer Science Departement
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
achowanda@binus.edu

2023 4th International Conference on Artificial Intelligence and Data Sciences (AiDAS) | 979-8-3503-1843-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/AIDAS60501.2023.10284611

***Abstract*—Deepfake technology has become a significant concern due to its ability to create highly realistic fake videos and images, leading to the potential deception of individuals. Detecting deepfakes has become a critical research area in computer vision and multimedia forensics. This paper presents a comparative analysis of deepfake detection models, focusing on evaluating their accuracy and robustness. Four CNN models, namely ResNet-152, MobilenetV3, Convnext Large, and EffecientNetB7, were implemented and trained using a custom dataset obtained from FaceForensics++. The models were evaluated based on training accuracy, average loss, and testing accuracy. An LSTM layer was also incorporated into each model's architecture to leverage sequential information. The results demonstrate varying performance among the models, with EfficientNet B7 achieving the highest testing accuracy of 75%. The findings of this study provide insights for future research in this critical area.**

***Index Terms*—deepfake detection, CNN models, comparative analysis, accuracy, LSTM layer**

I. Introduction

Deepfake refers to the alteration of digital media, such as photos and videos, through manipulations that replace the appearance of one person with that of another. Deepfake has raised significant concerns as it enables the creation of highly realistic fake videos and images that can deceive individuals. The detection of deepfake media has become a critical research area in computer vision and multimedia forensics.

Convolutional neural networks (CNN) have been widely used in deepfake detection methods because they effectively extract and analyze visual features from images and videos [1]. These methods utilize the hierarchical structure of CNN to identify patterns and anomalies in manipulated media, thus enhancing deepfake detection accuracy. However, the rapid evolution of deepfake generation techniques presents challenges for existing deepfake detection approaches [2].

Several studies have shown the effectiveness of CNN models in various computer vision tasks, including image classification [3] [4] [5]. In this study, we aim to evaluate the performance of these models, namely EfficientNetB7, MobileNetV3, and ConvNeXt, in the specific context of deepfake detection. As a baseline model, we will utilize ResNet-152, a widely adopted CNN model in deep-learning research [6]

The study objective is to assess the accuracy of these models in detecting various types of deepfakes. The findings of this research will contribute to mitigating the possible threats associated with deepfakes [7]. By conducting a comprehensive evaluation of state-of-the-art models, this study aims to advance deepfake detection techniques and provide guidance for future research in this critical area.

II. Literature Review

Deepfake technology has become a growing concern in recent years due to its potential to manipulate and deceive individuals through the creation of realistic fake videos or images. Initially developed for legitimate applications in the entertainment industry, deepfake technology utilizes deep learning techniques, particularly Convolutional Neural Networks (CNN) and Generative Adversarial Networks (GAN) [8], to generate highly convincing and manipulative media content. However, this technology's misuse for spreading misinformation, fake news, and malicious propaganda has raised significant alarms, necessitating research efforts to detect and combat deepfake media.

A. Deepfake Detection Methods

Deepfake Forensics, as one of the deepfake detection methods is a technique that analyzes various media features such as patterns, noise, and confusion matrix. This method involves the use of machine learning approaches, specifically deep learning and neural networks. The objective is to develop a model that can differentiate between real and fake media. After being trained on a set of real and fake media, the model is tested using a test set. If the accuracy falls below the expected level, the model is retrained until it achieves the desired accuracy.

One of the more common detection methods that use deep learning techniques involves binary classification. Binary classification is a task involving input determined by predictions based on the network according to two class labels. In the context of deepfake video detection, Binary classification is often used alongside two real and fake labels (0 or 1). While most deepfake detection methods often utilize deep-learning techniques, several methods have been proposed without using deep-learning techniques. Usage of successive subspace learning (SSL), extracted features that are distilled by using Spatial dimension reduction and Channel-wise Soft Classification, and the combination of Multi-Region and Multi-Frame ensemble have been tested to produce a light-weight, highly efficient deepfake detector model without using traditional deep-learning-based methods [9].

Various deepfake detection models, primarily based on Convolutional Neural Networks (CNN), have been proposed in the literature. EfficientNetB7, MobileNetV3, ResNet-152, and ConvNeXt are notable models with promising performance in deepfake detection. EfficientNetB7, a state-of-the-art CNN model, has demonstrated exceptional accuracy in various computer vision tasks [3]. MobileNetV3 also has shown promising results in deepfake detection while maintaining computational efficiency [4]. ResNet-152 is a widely adopted baseline CNN model in deepfake detection research [6]. ConvNeXt, designed to capture spatial and temporal features, has shown excellent performance in deepfake detection [5].

B. *Convolutional Neural Network (CNN)*

CNN, or Convolutional Neural Network, is a deep-learning architecture used to recognize features and patterns, including detecting deepfakes. The main advantage of CNN is that it can automatically learn valuable features from raw using the convolutional layers. These convolutional layers can collect spatial information from inputs and then be used for feature extraction. The features that have been extracted will then be used by the fully connected layers in the neural network to identify the deepfake

Although it has the benefit of self-learning, CNN is still vulnerable to attacks specifically created to avoid CNN-based model detection. Therefore, additional research is required to increase the generalization and robustness of CNN-based detection models, even to detect a deepfake created to avoid the CNN detection model.

While it can be used to detect deepfake, CNN can also be used for generating deepfake and sometimes leaving a trace. Research conducted by Luca et. Al [10] used Convolutional traces, a unique identifier, to detect deepfake media and even identify the GAN architecture that makes that deepfake. They used the Expectation-Maximization algorithm to extract the convolutional traces left by the CNN. Based on their research result, we can see that their proposed model has better accuracy than other models such as FakeSpotter [11] and AutoGAN [12].

C. *Long Short-Term Memory*

Long Short-Term Memory, or LSTM, is a modified version of a Recurrent Neural Network [13]. LSTM allows models to learn about temporal information that would otherwise be lost if regular RNNs were used instead due to their inability to preserve long-term dependencies by extending CEC by adding input and output gates connected to the input layer, which addresses the problem of conflicts during updating weights. [14].

LSTM has been proposed to be an excellent addition to most machine learning tasks due to LSTM's ability to preserve temporal information [15], which can be used to add more parameters that the model can learn and improve the outcomes of the model.

In 2019, Tzuu-Hseng S.LI et al. conducted an experiment to create a facial recognition model that detects human emotions, enhancing human-computer interaction for integrating robots into daily life [16]. The study highlights the superiority of LSTM and CNN-LSTM architectures in capturing temporal and contextual facial expression information compared to MLP and Singular CNN. LSTM's ability to retain temporal context makes it a valuable addition to deep-learning-based classification models.

D. *Vulnerabilities and Challenges*

Deepfake detection models are known to be vulnerable to adversarial attacks. Adversaries can manipulate deepfake videos to evade or even fool the detection models into misclassifying fake content as real [17] [18]. Developing generalized models capable of detecting different types of deepfakes remains a challenge. Deepfakes can vary in quality, manipulation techniques, and characteristics, making it challenging to develop a one-size-fits-all detection solution.

Most current deepfake detection methods often focus on analyzing the facial features contained in videos, prioritizing visual elements. The nature of current deepfake detectors relying on facial features of videos leads to potential concerns where implementation of strong Antiforensics measures on facially manipulated images alongside the usage of other non-visual deepfake media might cause current deepfake detection techniques becoming highly inefficient [19].

E. *Comparative Review*

In this section, we reviewed a comparative review to evaluate the performance of deep CNN in detecting distracted drivers. A comparative review involves analyzing and comparing multiple models or methods to determine their effectiveness. The goal of the comparative review is to identify the best-performing approach. Kathiravan et al. made a Comparison of deep convolutional neural networks in 2021 [20], in which three deep CNN models were given a set of pictures of distracted drivers, which the paper claims that several road accidents have happened due to humans not paying attention while driving. The paper suggests that developing a system capable of accurately predicting driver distraction can potentially reduce road accidents.

During this Comparative Study, the three chosen deep CNN models are Resnet, Xception, and VGG16 Model, and all models are evaluated based on precision, recall, and F1 score. After analyzing the results obtained from the experiments, it was concluded that Resnet was the best-performing deep CNN model, while VGG16 was the worst-performing model due to VGG16 being a primitive CNN model, even though the results are satisfactory, and Xception landing in between. It can be concluded that the Resnet model should be used as a baseline for our experiment and future experiments.

III. Methodology

The primary focus of this research is to investigate the performance of various convolutional neural network models in detecting deepfake videos. Thus, we used a qualitative approach to detect manipulated videos with equal parameters.

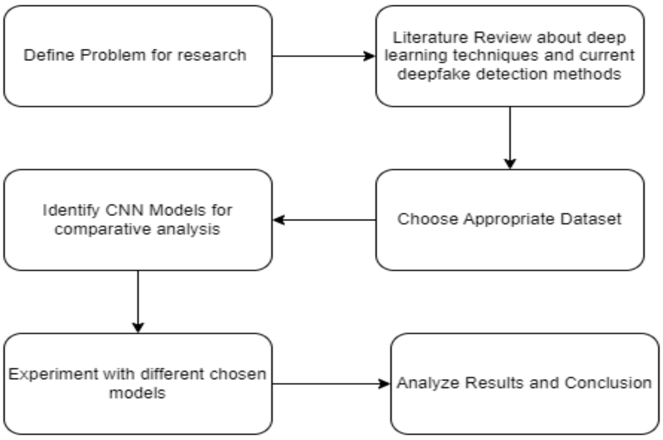
Define Problem for research
Literature Review about deep learning techniques and current deepfake detection methods
Choose Appropriate Dataset
Identify CNN Models for comparative analysis
Experiment with different chosen models
Analyze Results and Conclusion

Fig. 1. Research Methodology

A. Identifying Models for Experimentation

Based on the literature review and existing work, we imported our models from Pytorch libraries. First, Resnet-152 is chosen for the baseline, as it is one of the older models chosen from the roster. The rest of the models chosen, such as MobileNetV3, ConvNeXt Large, and EfficientNetB7, are much newer and considered much more efficient and accurate than the baseline model.

B. Collection and Analysis

We utilized the FaceForensics++ dataset [21]. FaceForensics++ is a comprehensive public set comprising 1000 original videos from the public internet and 1000 manipulated videos generated through advanced video editing techniques. By using this established set, we can ensure the inclusion of diverse and realistic deepfake scenarios. Fig. 2 shows an example of a FaceForensics++ Video.

To ensure a fair comparison, each deepfake detection model had the same set and underwent similar preprocessing methods. The preprocessing steps involved extracting frames from each randomly selected video and ensuring equal video dimensions and parameters for each video. This approach will create a level playing field for all models, allowing for a fair and unbiased evaluation of their performance.

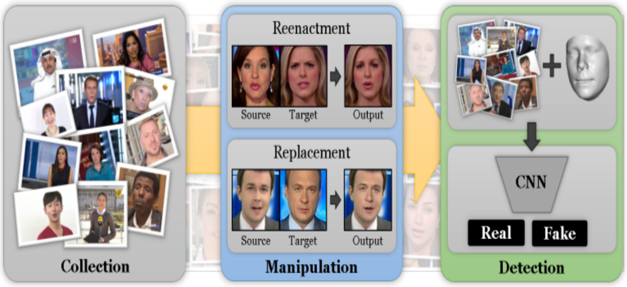

Fig. 2. FaceForensics++ Video Example Adapted from [21]

By leveraging the FaceForensics++ set and applying consistent preprocessing techniques, we aim to facilitate an objective comparison of the deepfake detection models and enable meaningful insights into their accuracy and robustness.

C. Preprocessing Methods

Each randomly chosen video was split into frames for the preprocessing methods used on the set. These frames were labeled based on meta to determine whether the video was real or fake. The dimension of the videos will be resized to 112x112 before preprocessing them. The mean and standard deviation were standardized for each video in the training and test sets.

D. Experimenting with Models

The experiments for this research paper were conducted in Google Colaboratory Pro using a preprocessed custom dataset obtained from the FaceForensics++ set, and the source code is a modified version from [22]. The set consisted of a ratio of real and fake videos 1:4 to ensure a balanced representation. The architecture of the model consisted of a Deep CNN Model accompanied by one LSTM layer incorporated to enhance efficiency and leverage sequential information, which comes after the data is put through the Deep CNN Model.

The experiments employed four CNN models: ResNet-152, MobileNet-V3 Large, ConvNeXt Large, and EfficientNetB7. Each model was trained separately to ensure independent training processes and reliable results. The Adam optimizer with a learning rate of 0.00001 was utilized, and the training was performed over 20 epochs. Following the training phase, the models were tested to evaluate their performance regarding training accuracy, average loss, and testing accuracy.

Finally, we made some additional code that is used to create a test dataset, and the newly trained model will be tested by using a test set to determine the precise accuracy of the model.

By employing this experimental setup, the research aimed to compare the performance of the selected models in deepfake detection. Using custom sets derived from FaceForensics++

and including an LSTM layer aimed to provide a fair evaluation and improve the models' effectiveness in detecting deepfakes.

IV. Result & Discussion

This section of our research is dedicated to discussing the results yielded from the research based on each of the four implemented models, which are ResNet-152, MobilenetV3, Convnext Large, and EffecientNetB7. The source code and the result could be found in GitHub Repositories https://github.com/DevanLucian15741/ComparisonDF_RM

Table 1. Accuracy Results of FaceForensics++ Dataset.

CNN Model	Training		Validation		Testing
	Accuracy	Loss	Accuracy	Loss	
EfficientNet_B7	87.88%	0.3877	67.5%	0.6260	**75%**
MobileNetV3	85.25%	0.4079	70.75%	0.6299	61.33%
ConvNeXt	**94.44%**	0.2077	73.25%	1.2783	67.17%
ResNet-152	91.19%	0.2993	80.75%	0.6678	61%

Table 2. F1-Score Results.

CNN Model	Precision	Recall	F1-Score
EfficientNet_B7	0.8623	0.7228	0.7853
MobileNetV3	0.8306	0.8022	0.8160
ConvNeXt	0.8025	0.8852	0.8414
ResNet-152	0.8266	0.9556	0.8852

A. *Resnet-152 Results*

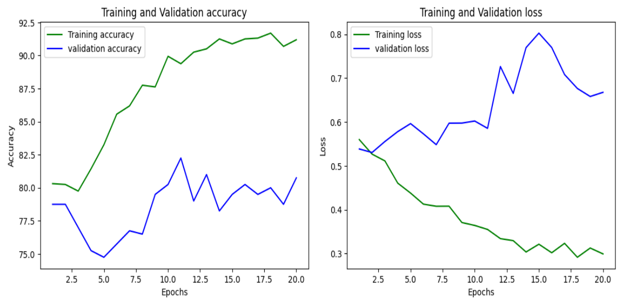

Fig. 3. Resnet-152 Training and Validation Graph

For the training results obtained from combining the ResNet-152 model with one layer of LSTM, as shown in Fig. 3. the model demonstrated a training accuracy of 91.19%. This high accuracy indicates excellent performance in classifying deepfake videos and distinguishing them from real videos. The model's training loss score of 0.2993 indicates that it made a few mistakes during the classification process, further affirming its strong performance.

Fig. 4 shows the true positive, false positives, false negatives, and true negative values that the model predicts, represented as a confusion matrix. From Fig. 4. we can determine that the precision and F1 Score of the model during training are 0.9556 and 0.8866, respectively, which proves that the model performed well on the training dataset.

Looking at the accuracy, the model performs relatively lower than the training and validation accuracy. It suggests the presence of some degree of overfitting within the model. However, despite this observation, the model still achieved satisfactory accuracy on the training set, and it can be considered to have performed reasonably well with a significant number of true positives.

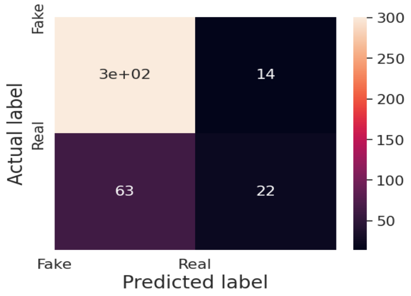

Fig. 4. Resnet-152 Confusion Matrix

B. *MobilenetV3 Results*

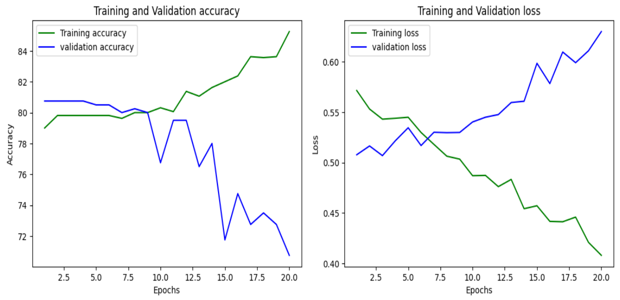

Fig. 5. MobilenetV3 Training and Validation Graph

In Fig. 5, the MobileNetV3 model achieved an overall training accuracy of 85.25% with a training loss of 0.407. These results indicate that the model performed well in training, demonstrating high accuracy and minimal errors.

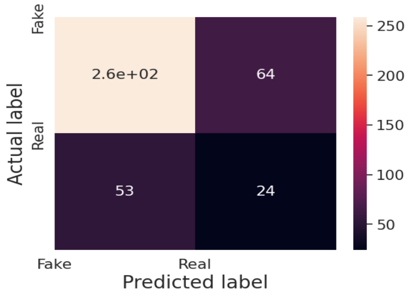

Fig. 6. MobilenetV3 Confusion Matrix

Fig. 6. shows the confusion matrix results for MobilenetV3 and shows that the Precision and F1 Score calculated based on the scores from the confusion matrix are 0.8019 and 0.8157, respectively. This further proves that the model performed well on the training set as it can maintain a good balance between precision and recall, even if Resnet-152 yielded better results during training.

However, when evaluated on both the validation and test sets from Table. 1, the model's performance was significantly worse at a testing accuracy of 61.33%, much lower than the training accuracy. This suggests the presence of overfitting, where the model may have memorized the training too well and struggled to generalize to the unseen. Despite this, the model still exhibited moderate performance.

Comparing the performance of MobileNetV3 to ResNet152, it appears that MobileNetV3 had a slightly higher testing accuracy. However, the difference in accuracy between the two models is minimal, making them practically identical in performance.

C. ConvNeXt Results

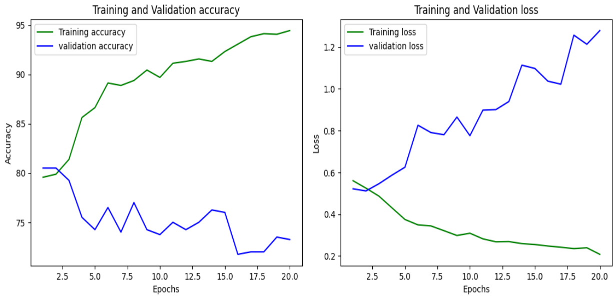

Fig. 7. ConvNeXt Training and Validation Graph

The ConvNeXt model used in this experiment is the Large version. It exhibited a notable symptom of overfitting, performing well on the training set with a high accuracy of 94.44% and a low loss of 0.207732 based on Fig. 7. These results indicate that the model was able to accurately classify the majority of the training, with minimal errors.

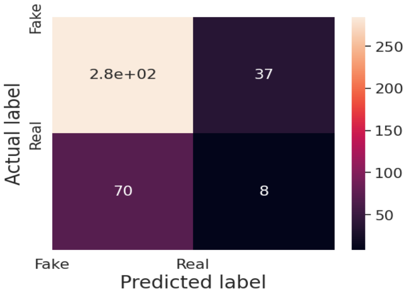

Fig. 8. ConvNext Confusion Matrix

Despite being the model that performed the best during training, the ConvNeXt model does not have the highest Precision nor F1 Score since both are 0.8851 and 0.8419, respectively, calculated from the confusion matrix shown in Fig. 8; these results are much lower when compared to Resnet152.

Interestingly, when evaluated on the testing set, the ConvNeXt model achieved significantly higher accuracy than ResNet-152 and MobileNetV3. From Table. 1, It achieved a testing accuracy of 67.17%, indicating that it could correctly classify over half of the test set. This performance surpassed the other two models, suggesting that ConvNeXt had a higher capability to generalize to the unseen.

D. EffecientNetB7 Results

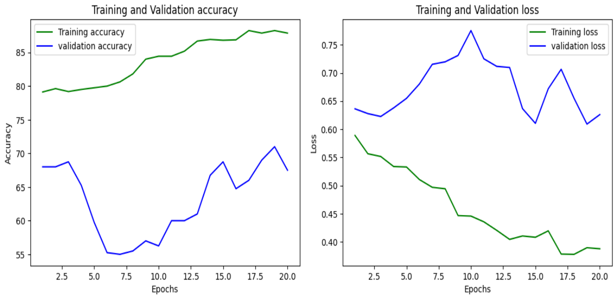

Fig. 9. EffecientNetB7 Training and Validation Graph

EfficientNetB7 was selected as the final model for this experiment. Based on Fig. 9, the EfficientNetB7 model achieved a training accuracy of 87.88% with a loss of 0.387696, indicating its strong performance on the training set, similar to the previous models. However, the model's accuracy on the validation set dropped slightly to 71%, which was still the highest accuracy among the models.

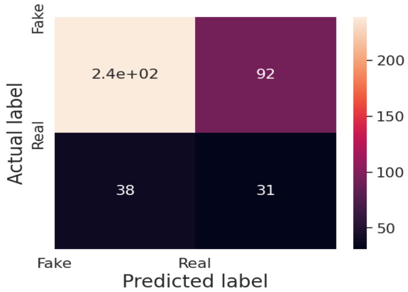

Fig. 10. EffecientNetB7 Confusion Matrix

From Fig. 10. It can be concluded that EfficientNetB7 yields the lowest Precision and F1 Score out of all the models tested, as both calculated scores are 0.7221 and 0.7862, respectively, significantly lower when compared to Resnet-152.

Table. 1. Shows that in terms of testing accuracy, EfficientNetB7 performed the best out of all four models, achieving a testing accuracy of 75%. EfficientNetB7 demonstrated the highest accuracy in classifying whether a video is deepfake or real. It is important to note that while EfficientNetB7 performed well, its testing accuracy was still lower than its training accuracy, suggesting some overfitting.

E. Discussion

The experiment results suggest that EfficientNetB7 is the most effective CNN model for detecting whether a video is real or manipulated based on its superior performance in binary classification. One factor that contributed to the success of EfficientNetB7 is its utilization of the compound scaling method, which played a role in enhancing its performance.

The second best-performing model in the experiment was ConvNeXt, which exhibited the second-highest testing accuracy. On the other hand, both Resnet-152 and MobileNet demonstrated similar results with comparable accuracies.

An important observation from the experiment is that all models exhibited signs of overfitting. This can be seen from the lower testing accuracy scores than the training accuracy scores. This indicates that the models struggled to generalize well beyond the training.

To address the issue of overfitting and improve the generalization ability of the models, further enhancements can be implemented. One approach could involve utilizing additional sets beyond FaceForensics++ to introduce more diverse features. Increasing the size of the training and testing sets can also be beneficial, as it provides more information for the models to learn from and improve their performance.

V. Conclusion

This study conducted a comparative analysis of deepfake detection models (ResNet-152, MobileNetV3, ConvNeXt, and EfficientNetB7) on the FaceForensics++ dataset. EfficientNetB7 achieved the highest testing accuracy, making it the best model to use when it comes to detecting deepfakes, outperforming the other models. However, all models showed signs of overfitting, indicating the need for further improvements in generalization ability. To enhance deepfake detection, future research should explore techniques such as data augmentation, regularization, and larger datasets. Additionally, incorporating advanced techniques like attention mechanisms, ensemble learning, and adversarial training can further improve the accuracy and robustness of deepfake detection systems. This study emphasizes the significance of deepfake detection and provides insights for selecting appropriate models and addressing challenges in the field.

References

[1] H. S. Shad, M. M. Rizvee, N. T. Roza, S. Hoq, M. Monirujjaman Khan, A. Singh, A. Zaguia, S. Bourouis, *et al.*, "Comparative analysis of deepfake image detection method using convolutional neural network," *Computational Intelligence and Neuroscience*, vol. 2021, 2021.

[2] A. Beckmann, A. Hilsmann, and P. Eisert, "Fooling state-of-the-art deepfake detection with high-quality deepfakes," 2023.

[3] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[4] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1314–1324, 2019.

[5] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[7] M. Westerlund, "The emergence of deepfake technology: A review," *Technology innovation management review*, vol. 9, no. 11, 2019.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[9] H. G. Hong-Shuo Chen, Mozhdeh Rouhsedaghat, "Defakehop: A light-weight high-performance deepfake detector," in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2021.

[10] L. Guarnera, O. Giudice, and S. Battiato, "Fighting deepfake by exposing the convolutional traces on images," *IEEE Access*, vol. 8, pp. 165085–165098, 2020.

[11] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, and Y. Liu, "Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces," 2020.

[12] X. Gong, S. Chang, Y. Jiang, and Z. Wang, "Autogan: Neural architecture search for generative adversarial networks," 2019.

[13] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, mar 2020.

[14] R. C. Staudemeyer and E. R. Morris, "Understanding lstm – a tutorial into long short-term memory recurrent neural networks," 2019.

[15] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, "A hybrid cnn-lstm model for video deepfake detection by leveraging optical flow features," in *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, IEEE, 2022.

[16] T.-H. S. Li, P.-H. Kuo, T.-N. Tsai, and P.-C. Luan, "Cnn and lstm based facial expression analysis model for a humanoid robot," *IEEE Access*, vol. 7, pp. 93998–94011, 2019.

[17] N. Carlini and H. Farid, "Evading deepfake-image detectors with white- and black-box attacks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 658–659, 2020.

[18] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, and J. McAuley, "Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3348–3357, 2021.

[19] S. Lyu, "Deepfake detection: Current challenges and next steps," in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 4–5, 2020.

[20] D. D. Kathiravan Srinivasan, Lalit Garg, "Performance comparison of deep cnn models for detecting driver's distraction," *Tech Science*, vol. 68, 2021.

[21] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," 2019.

[22] A. Jadhav, A. Patange, J. Patel, H. Patil, and M. Mahajan, "Deepfake video detection using neural networks," *International Journal for Scientific Research and Development*, vol. 8, no. 1, pp. 1016–1019, 2020.