Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2024) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

9th International Conference on Computer Science and Computational Intelligence 2024 (ICCSCI 2024)

Comparison of Deepfake Detection using CNN and Hybrid Models

Fransiskus Triyanto Winata[a], Nicholas Justin Tanuwijaya[a*], Dr. Reina, S.Kom., M.M.[a], Reinert Yosua Rumagit, S.Kom, M.TI[a]

School of Computer Science, Computer Science, Bina Nusantara University, Jakarta, Indonesia

Abstract

In today's digital era, manipulated videos known as deepfakes are becoming more common and harder to detect. These deepfake contents can be misused by people to spread false information, damage someone's reputation, or even harm many people. To address this issue, we conducted a study to compare the performance of three deep learning models Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and a hybrid CNN-LSTM model to detect deepfake content. In this study, we used the DeepFake Detection Challenge (DFDC) dataset, which includes a variety of real and fake videos, then we will evaluate each model using accuracy, precision, recall, F1-score, AUC, and loss values. The results of our research show that the RNN model provides the best overall performance with an accuracy of 77.5%, F1-score of 77.5%, and AUC of 0.928, while CNN followed closely with 75.5% accuracy and AUC of 0.930. The CNN-LSTM model showed lower performance with 67.5% accuracy and AUC of 0.886. The findings demonstrate the strength of RNNs in modeling temporal patterns in video data, which is crucial for effective deepfake detection. Then CNN model also have good performance and shows strong results in detecting spatial features in individual frames. Meanwhile, CNN-LSTM model does not perform as well as the other models. This can be happen due to hybrid model having more complex structure, which makes it more difficult to train effectively. We hope this research helps in building better tools to detect deepfakes in real-world situations.

Keywords: Type your keywords here, separated by semicolons ;

1877-0509 © 2024 The Authors. Published by ELSEVIER B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 9th International Conference on Computer Science and Computational Intelligence 2024

1. Introduction

Deepfakes over the last two years have been a worldwide and significant menace under digitization's regime that can potentially generate synthetic yet realistic-appearing media using the deep learning methodology. However, it can also become a dangerous instrument for fake news propaganda, identity theft, as well as political deception too. Deepfake's availability has been a cause for concern regarding its potential misuses. Therefore, the shortness of successful detection solutions has been a zealous research interest area.

Deepfake detection mainly relies on deep learning techniques, such as Convolutional Neural Networks (CNN), a widely used architecture due to its ability to distinguish between real faces and faces modified by Deepfakes as it is successful in examining visual features between fake images and videos [8]. However, CNN models are not yet capable of being applied to large deep fake data and to keep up with the rapidly evolving manipulation algorithms.

To improve the performance of deepfake detection tools, it is recommended to combine CNN with RNN or LSTM. Therefore, this combination is expected to be more accurate in revealing the differences in structure between original content and deepfakes, thus providing more benefits compared to a single CNN model or a single LSTM.

In addition, we also found several papers explaining deepfake detection. However, papers such as paper [1] only explain the CNN model and paper [2] only explain the LSTM model. Based on both papers, we can conclude that both papers only explain one of the models in its application for deepfake detection.

Therefore, our aim for this paper is to present comparative results of deepfake detection by using all the models, which are CNN model, hybrid (LSTM) model and the fusion of CNN-LSTM model. This is done to compare the accuracy, robustness, and efficiency of the three models against deepfake datasets. Through this research by evaluating their strengths and vulnerabilities, it is possible to make recommendations on the most appropriate approach to detect deepfakes in real-world applications.

2. Literature review

2.1. Deepfakes

Deepfakes is a video that can be used to make an individual seem to have said a thing he or she never said, or performed on a thing he or she never did. It is realized by uploading videos of two individuals to a deep learning algorithm which in turn learns how to switch faces. Thus, deepfakes utilize facial mapping technology and AI so that they can replace a person's face with someone else's face [3].

Deepfake is a process involving the use of deep learning models, specifically the Generative Adversarial Networks (GANs) for generating realistic content [4]. GANs is a model that is accountable for the generation of deepfake media. GANs is applied for training the generative models independently by solving the issue with supervision as well as generating realistic fake faces on images/videos. Some ML-based solutions tend to accentuate certain anomalies that can be observed on fake images/videos generated using GANs [5].

Still, the emergence of deepfakes has made it even more difficult to distinguish between synthetic (fake) and real content, requiring improvement to develop detection systems. This is due to the fact that even if deep learning is successful, it is at times difficult to clarify how this learning approach comes to be inferred and interpreted for forensic use [6]. Models (i.e. CNN and Hybrid models) thus need to be developed, where these models can detect and evaluate deepfake images/videos.

2.2. CNN

Convolutional Neural Networks (CNN) were widely applied for the detection of deepfakes as they were able to detect certain features from videos and images. A comparison is then performed by a study [7] of how various architectures of CNNs perform on datasets. The architecture of CNN involves three major elements, which are a facial features extractor, a biometric anomaly detector using temporal networks, and generative adversarial networks

(GANs) that try to predict a person's specific movements. A study even established that the efficiency of various architectures of CNNs for DeepFake video detection revealed that networks that are very deep work better for the task, especially with low-quality videos.[8]

These researches are aimed at the effectiveness of CNNs for Deepfakes detection and facial inconsistency detection. The CNN model performs two tasks for detecting Deepfakes. It first extracts some features between fake and real images by concatenating some dense units, where each dense unit is a sequence of dense blocks representing fake images. It then trains the proposed CNN using these features for classifying the given images as fake or real [5]. Therefore, CNN also has the ability to learn very powerful image features [9]. Then, the performance of CNN scales with the dataset size, where CNN can extract a patch image size (i.e., 2 x 2) from the input image and apply full attention on it, thus outperforming the vanilla transformer [10]. Besides that, CNN works on learning from spatial data too well and CNN feature representations display their ability of learning from spatial data [2].

And then Zhou [11] suggested a face classification network on the basis of CNN to analyze evidence of manipulated artifacts, a steganalysis feature-based triplet network to analyze faces' local residues' evidence of tampering. Besides that, Optical Flow (OF) fields can also be used to differentiate between Deepfake and original images through the use of CNN by extracting them to exploit inter-frame correlations and passing them as input to CNN in the form of chunks [12]. Therefore, Convolutional Neural Network (CNN) image classifiers as well as models for image recognition were proved to be trainable to distinguish between true images and manipulated images [13].

2.3. DFDC Dataset

Then it also proposed another dataset, named DeepFake Detection Challenge (DFDC) Dataset, largest-scale recently published deepfake detection dataset, and most challenging dataset for deepfake detection owing to high-quality fake videos of this dataset [14]. The DFDC dataset has a large number of clips of varying quality and is a good representation of existing face swapping approaches. DFDC dataset comprises 5.244 videos with 66 actors, which are 4.464 training videos and 780 test videos, among which are 1.131 real videos and 4.113 synthetic videos generated with two unknown face swap approaches [15].

Apart from that, the DFDC dataset includes video data of 3,426 paid speakers who are actresses whose faces can be manipulatable using computer vision algorithms [16]. Therefore, CNN-based models, being responsible for fake video detection, can perform well on the DFDC dataset [1].

2.4. LSTM

LSTM is a recurrent neural network able to learn long-term dependencies. The network is introduced to address the long-term dependency problem of RNN. The 3 sets of input, output, and forget gates of the architecture of_lstm protect and keep its internal cell state [2]. Subsequently, for a reference baseline, a straightforward LSTM model is employed, comprising three layers of LSTMs followed by a linear layer whose output is averaged on the time axis to get a single embedding vector [17]. Additionally, the MFN model, for guiding the embedding of facial and speech modalities, is a recurrent neural network architecture with three important constituents embodied by an LSTM system, a Memory Attention Network, and a Gated Memory component, where the LSTM system is being fed different views of the input data [18].

2.5. CNN + LSTM

Researchers attempt to overcome the limitation of these models, thus they create a new model that combines CNN with hybrid models, known as CNN-LSTM model for more advanced tasks. This model is proposed where CNN is combined with LSTM for more advanced tasks typically. CNN extracts a set of features for each frame of a sequence of images and passes it to LSTM for processing [1]. Subsequently, researchers named Guera and Delp present a temporal-aware methodology for auto detection of deepfake movies. These researchers present a robust trainable recurrent model for deepfake detection of videos. For frame processing as a sequence, the proposed system makes use of a convolutional LSTM architecture. Convolutional LSTM consists of two principal parts, which are CNN for frame feature extraction as well as sequence modeling over a period using LSTM [19]. Optical flow features are utilized as features for a hybrid system of a classifier that is a mixture of CNN as well as LSTM [20].

2.5.1. CNN + LSTM Architecture

After that, it introduced a new architecture, named CNN-LSTM architecture that combines CNN with RNN for efficiency and accuracy. The system proposed is basically a hybrid of a convolutional long-short-term memory (LSTM) architecture to address the sequence of frames. The two major factors are CNN and LSTM model, of which CNN is applied for feature extraction of frames while analysis of the sequence of time is conducted by LSTM. The algorithm has proven to predict well even if the video duration is not more than 2 seconds [21]. Aside from that, a researcher by the name Chan put forth a strategy designed by using a set of LSTM-CNNs for content analysis of the video as well as sound content of the digital media, and generate descriptive caption descriptions to accompany the extracted feature information to generate a representative unique hash for the content [19]. Additionally, research by David Guera has proven the efficacy of Long Short-Term Memory (LSTM) networks paired with CNN models in detecting deepfakes [20]. Then, Daniel Mas Montserrat et al. [22] proposed a system that is able to extract visual and temporal features from faces in a video by combining a CNN and RNN architecture to be able to detect Deepfake videos. Besides that, David and Edward combined two deep learning models for effectively spotting fake videos, where each frame of the video undergoes two phases of checking and analysis. The first is a Convolutional Neural Network (CNN) for extracting features from forged video frames, the second is a Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM) for forged video detection after training [23].

2.6. Related Work

Besides that, Afchar who introduced a CNN model for real and Deepfake classification, suggested a double CNN-based approach for detection of face forgery at detection rate of 98% and 95% on deepfake images. Hasil awal menunjukkan bahwa metode terbaik (XceptionNet) atau approach CNN memberikan presisi 93,0% dan mencapai akurasi 98,4%. In addition, the average accuracy of the proposed detection approach on the DFDC dataset is only 65.18%, proving that the current detection approach is still far from meeting practical needs [24]. Then, there is a comparison graph of the performance of hybrid recurrent models with base stacked models, where the multilayer LSTM accuracy is 0.8076 and GRU is 0.8086, Hybrid LSTM-GRU is 0.8152 and GRU-LSTM is 0.8165 [1]. In addition, a research found that CNN achieves accuracy of 91.5% on DFDC, AUC 0.91 (91%), loss 0.32 [11]. On the other hand, other research found that CNN achieves accuracy of 66.26% on DFDC [13].

3. Methodology

3.1 Research Design

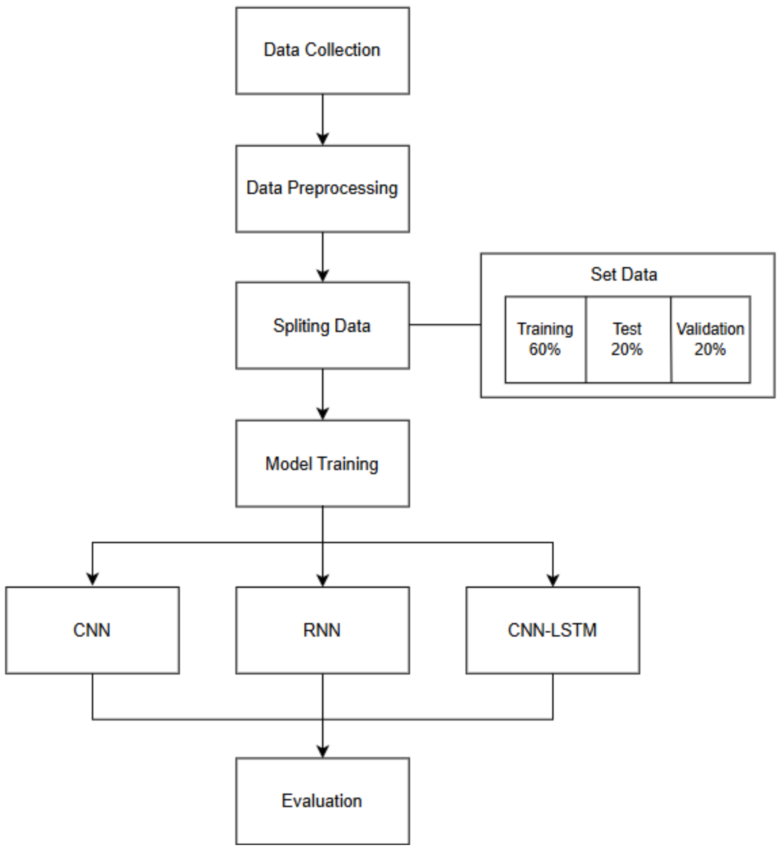

Fig. 1. **Deepfake Detection Experiment Flow with CNN, RNN, and CNN-LSTM**

This study uses an experimental approach to compare the performance of Convolutional Neural Networks (CNN) and hybrid deep models in detecting deepfake content. All model will be trained using the same dataset and preprocessing techniques to ensure fairness in the comparison. The overall workflow can be seen in Figure 1, the process begins with data collection and preprocessing, after that the dataset will be divided into three parts: 60% for training, 20% for testing, and 20% for validation. After the data is separated, the training process will be carried out separately for each model starting from CNN, RNN, and CNN-LSTM. Finally, the trained models are evaluated using several performance metrics to determine their accuracy and effectiveness in detecting fake video content.

3.2 Data Preparation

The dataset used in this study serves to evaluate how well the selected model performs. For this research, the dataset we use is DeepFake Detection Challenge (DFDC) dataset. This dataset consists of thousands of video clips, both real and fake. DFDC is one of the best datasets for testing deepfake detection systems due to its diversity and high-quality data at large scale. The total dataset used for this paper is 5365 data which consists of 2667 real images and 2698 fake images. These 5365 data will be divided into 3218 for training, 1073 for validation, and 1074 for testing.

3.3 Data Preprocessing

Before we could train the models, the video data from the DFDC dataset needed to be converted into a suitable format for processing. The first thing we did was extract individual frames from each video clip. From every clip, we will picked a set number of frames using a uniform sampling method to keep things consistent and make sure we had a balanced mix of real and fake data. After that, we resized all the frames to fit the input size of CNN and hybrid models. This resizing ensure that all the data goes through the same processing steps and makes the models easier to train. After frame extraction we will normalized the pixel values so that they were between 0 and 1, which helped the training run more efficient. We also applied simple augmentation techniques like flips and rotations to the images, so that the model can generalize better. Lastly, we label each frame as real or fake based on the labels of the original video, then we split the data into three parts: 60% for training, 20% for validation, and 20% for testing.

3.4 Model Training

Our deepfake detection approach uses a sequential frame-based technique that includes three different architectures: CNN, RNN, and a hybrid CNN-LSTM model, where each architecture can examine a set of 10 consecutive frames resized to 128×128 px. We also use weighted sampling with inverse frequency weights with all models trained using the Adam optimizer (lr=0.001, weight_decay=1e-5) at the initial stage along with a binary cross-entropy loss to address the class imbalance between real and fake samples. The performance of this optimization method encompasses mixed-precision training, effective gradient calculation, and enhanced data loading utilizing parallel workers and prefetching. Subsequently, we implement early stopping with 5 epochs of patience to avoid overfitting while observing validation metrics. Concerning the three models, the CNN design features three convolutional blocks with ascending filter counts (16 => 32 => 64), the RNN employs a streamlined CNN frontend together with LSTM layers (hidden_size=128, num_layers=2), whereas our hybrid CNN-LSTM merges a more complex CNN encoder with temporal sequence processing for superior feature extraction in both spatial and temporal areas.

3.5 Metric Evaluation

Then recall is the number of fake videos the model accurately identified. To provide a balanced picture of both precision and recall, the F1-score is taken. It combines both precision and recall into a single figure, which allows one to see whether the model not only makes accurate fake predictions but also identifies most of the fake videos. This is very useful if the number of real and fake videos are not equal. This study also uses AUC (Area Under the Curve), AUC shows how well the model can distinguish between real and fake videos. The higher the AUC value the better the model is at separating the two classes. Lastly, Loss to shows how far the model's predictions are from the correct answers. A smaller loss value means the model is making better predictions.

4. Result and Discussion

This research conducted to compare the effectiveness deep learning algorithms such as CNN, RNN, and Hybrid models in detecting deepfake content. Each of this model is trained using the DeepFake Detection Challenge (DFDC) dataset. The evaluation metrics used include accuracy, precision, recall, F1-score, AUC, and loss. Here we include the training results of the three models, ROC Curve and also the confusion matrix.

Table 1. Model performance metrics

Model	Accuracy	Precision	Recall	F1-Score	AUC	Loss
CNN	0.7550	0.7592	0.7550	0.7559	0.9302	0.6228
RNN	0.7750	0.7762	0.7750	0.7754	0.9282	0.6501
CNN-LSTM	0.6750	0.6747	0.6750	0.6741	0.8860	0.8264

As can be seen from the Table 1, the results obtained are that the RNN model has the highest overall performance, with an accuracy of 0.7750, precision of 0.7762, recall of 0.7750, F1 score of 0.7754, AUC of 0.9282, and Loss of 0.6501. Then followed by CNN with the second best performance where CNN has higher AUC and Loss scores but loses in the other four metrics. CNN can be slightly superior in two metrics, possibly due to its effectiveness in extracting spatial features from individual frames. Both CNN and RNN have almost equal scores, while CNN-LSTM shows the weakest performance on all metrics. This can be happen because CNN-LSTM is designed to combine the strengths of spatial and temporal modelling, but in practice it actually increases its complexity and causing bad performance. The superior performance of RNN compared to the other two models is likely due to RNN strength in modeling temporal relationships between video frames, which is essential for detecting dynamic inconsistencies in deepfakes.

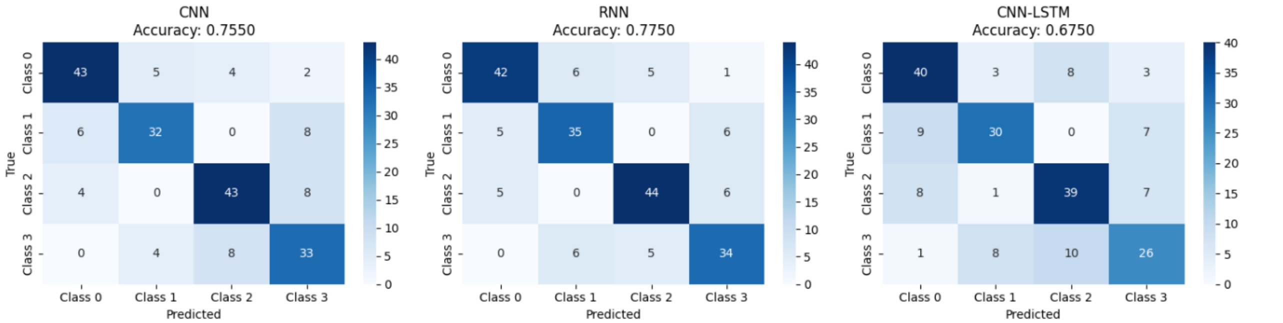

Fig. 2. Confusion Matrix

To support our result, we insert confusion matrix that can be seen in Fig. 2 so that it can help to provide an overview of how the model classifies each class by showing both correct and incorrect predictions. Each matrix displays the number of videos that were correctly identified (diagonal elements) and those that were misclassified (non-diagonal elements).

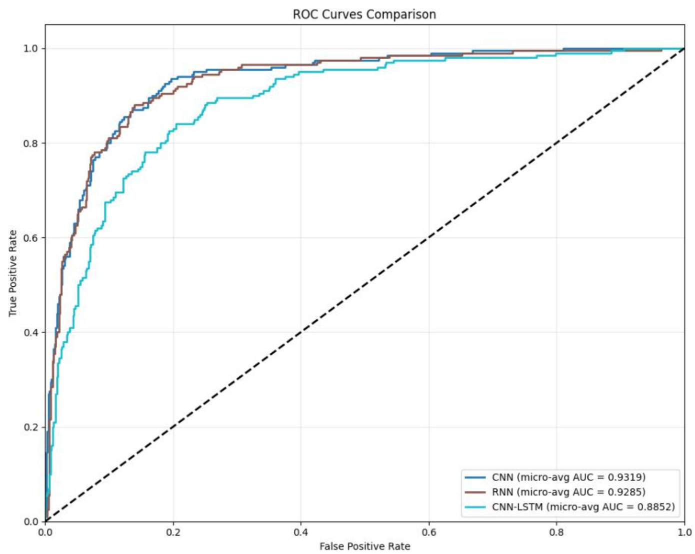

Fig. 3. ROC Curves

We also include the ROC curve, as shown in Fig. 3, to visualize how well each model can distinguish between real and fake classes at different threshold values. From the figure, we can see that the RNN model has the highest AUC score (0.9282), indicating that it is the most capable in separating the real and fake classes overall. CNN follows closely with an AUC of 0.9191, while CNN-LSTM has the lowest AUC at 0.8852.

5. Conclusion

Detecting deepfakes manually is very time-consuming and requires a high level of expertise. With the help of deep learning models, we can make this process faster, more efficient, and more accurate. In this study, we compared three different models CNN, RNN, and a hybrid model (CNN-LSTM) to identify which model performs best for deepfake video detection. The results showed that the RNN model demonstrated the most consistent and reliable performance in detecting deepfakes. The CNN model also performed well, slightly behind the RNN. However, the hybrid CNN-LSTM model, which was expected to perform best by combining the two approaches, showed the weakest performance and a high loss value of 0.826, indicating instability during training and testing. This may be due to the increased complexity of the hybrid model, which makes training and optimization more difficult. These findings suggest that while CNN and RNN models are effective for deepfake detection, further research is needed to optimize hybrid models and improve their training stability for real world deployment.

References

1. Ikram ST, Chambial S, Sood D, V P, V A. A Performance Enhancement of Deepfake Video Detection through the use of a Hybrid CNN Deep Learning Model. Int J Electr Comput Eng Syst. 2023 Feb 27;14(2):169–78.
2. Jaiswal G. Hybrid Recurrent Deep Learning Model for DeepFake Video Detection. In: 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) [Internet]. 2021 [cited 2025 Apr 20]. p. 1–5. Available from: https://ieeexplore.ieee.org/abstract/document/9667632
3. Westerlund M. The Emergence of Deepfake Technology: A Review. Technol Innov Manag Rev. 2019;9(11):40–53.
4. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM. 2020 Oct 22;63(11):139–44.
5. Rana MS, Nobi MN, Murali B, Sung AH. Deepfake Detection: A Systematic Literature Review. IEEE Access. 2022;10:25494–513.
6. Classification of Deepfake Images Using a Novel Explanatory Hybrid Model | CommIT (Communication and Information Technology) Journal [Internet]. [cited 2025 Apr 20]. Available from: https://journal.binus.ac.id/index.php/commit/article/view/8761
7. Korshunov P, Marcel S. DeepFakes: a New Threat to Face Recognition? Assessment and Detection [Internet]. arXiv; 2018 [cited 2025 Apr 19]. Available from: http://arxiv.org/abs/1812.08685
8. Cozzolino D, Rössler A, Thies J, Nießner M, Verdoliva L. ID-Reveal: Identity-Aware DeepFake Video Detection. In 2021 [cited 2025 Apr 20]. p. 15108–17. Available from: https://openaccess.thecvf.com/content/ICCV2021/html/Cozzolino_ID-Reveal_Identity-Aware_DeepFake_Video_Detection_ICCV_2021_paper.html
9. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Niessner M. FaceForensics++: Learning to Detect Manipulated Facial Images. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV) [Internet]. 2019 [cited 2025 Apr 19]. p. 1–11. Available from: https://ieeexplore.ieee.org/document/9010912
10. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale [Internet]. arXiv; 2021 [cited 2025 Apr 19]. Available from: http://arxiv.org/abs/2010.11929
11. Rana MdS, Murali B, Sung AH. Deepfake Detection Using Machine Learning Algorithms. In: 2021 10th

International Congress on Advanced Applied Informatics (IIAI-AAI) [Internet]. 2021 [cited 2025 Apr 20]. p. 458–63. Available from: https://ieeexplore.ieee.org/abstract/document/9790940

12. Caldelli R, Galteri L, Amerini I, Del Bimbo A. Optical Flow based CNN for detection of unlearnt deepfake manipulations. Pattern Recognit Lett. 2021 Jun 1;146:31–7.
13. Taeb M, Chi H. Comparison of Deepfake Detection Techniques through Deep Learning. J Cybersecurity Priv. 2022 Mar;2(1):89–106.
14. Zhao H, Zhou W, Chen D, Wei T, Zhang W, Yu N. Multi-Attentional Deepfake Detection. In 2021 [cited 2025 Apr 20]. p. 2185–94. Available from: https://openaccess.thecvf.com/content/CVPR2021/html/Zhao_Multi-Attentional_Deepfake_Detection_CVPR_2021_paper.html?ref=https://githubhelp.com
15. Nirkin Y, Wolf L, Keller Y, Hassner T. DeepFake Detection Based on Discrepancies Between Faces and Their Context. IEEE Trans Pattern Anal Mach Intell. 2022 Oct;44(10):6111–21.
16. Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, et al. The DeepFake Detection Challenge (DFDC) Dataset [Internet]. arXiv; 2020 [cited 2025 Apr 19]. Available from: http://arxiv.org/abs/2006.07397
17. Müller NM, Czempin P, Dieckmann F, Froghyar A, Böttinger K. Does Audio Deepfake Detection Generalize? [Internet]. arXiv; 2024 [cited 2025 Apr 20]. Available from: http://arxiv.org/abs/2203.16263
18. ACM Conferences [Internet]. [cited 2025 Apr 20]. Emotions Don't Lie | Proceedings of the 28th ACM International Conference on Multimedia. Available from: https://dl.acm.org/doi/10.1145/3394171.3413570
19. Heidari A, Jafari Navimipour N, Dag H, Unal M. Deepfake detection using deep learning methods: A systematic and comprehensive review. WIREs Data Min Knowl Discov. 2024;14(2):e1520.
20. Saikia P, Dholaria D, Yadav P, Patel V, Roy M. A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow Features. In: 2022 International Joint Conference on Neural Networks (IJCNN) [Internet]. 2022 [cited 2025 Apr 20]. p. 1–7. Available from: https://ieeexplore.ieee.org/abstract/document/9892905
21. A Survey on Deepfake Video Detection - Yu - 2021 - IET Biometrics - Wiley Online Library [Internet]. [cited 2025 Apr 19]. Available from: https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/bme2.12031
22. Wodajo D, Atnafu S. Deepfake Video Detection Using Convolutional Vision Transformer [Internet]. arXiv; 2021 [cited 2025 Apr 20]. Available from: http://arxiv.org/abs/2102.11126
23. Jafar MT, Ababneh M, Al-Zoube M, Elhassan A. Forensics and Analysis of Deepfake Videos. In: 2020 11th International Conference on Information and Communication Systems (ICICS) [Internet]. Irbid, Jordan: IEEE; 2020 [cited 2025 Apr 20]. p. 053–8. Available from: https://ieeexplore.ieee.org/document/9078989/
24. Narayan K, Agarwal H, Mittal S, Thakral K, Kundu S, Vatsa M, et al. DeSI: Deepfake Source Identifier for Social Media. In 2022 [cited 2025 Apr 20]. p. 2858–67. Available from: https://openaccess.thecvf.com/content/CVPR2022W/FaDE-TCV/html/Narayan_DeSI_Deepfake_Source_Identifier_for_Social_Media_CVPRW_2022_paper.html