ILKOM Jurnal Ilmiah Vol. 14, No. 3, December 2022, pp. 178-185
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article Open Access (CC–BY-SA)

Deepfake detection in videos using Long Short-Term Memory and CNN ResNext

Muhammad Indra Abidin [a,1,*]**; Ingrid Nurtanio** [a,2]**; Andani Achmad**[a,3]
[a] *Hasanuddin University, Jl. Perintis Kemerdekaan No.KM.10, Makassar and 90245, Indonesia*
[1] *muhindraabidin@gmail.com;* [2] *ingrid@unhas.ac.id;* [3]*andani@unhas.ac.id.*

Article history: Received August 02, 2022; Revised August 11, 2022; Accepted December 15, 2022; Available online December 20, 2022

Abstract

Deep-fake in videos is a video synthesis technique by changing the people's face in the video with others' face. Deep-fake technology in videos has been used to manipulate information, therefore it is necessary to detect deep-fakes in videos. This paper aimed to detect deep-fakes in videos using the ResNext Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) algorithms. The video data was divided into 4 types, namely video with 10 frames, 20 frames, 40 frames and 60 frames. Furthermore, face detection was used to crop the image to 100 x 100 pixels and then the pictures were processed using ResNext CNN and LSTM. The confusion matrix was employed to measure the performance of the ResNext CNN-LSTM algorithm. The indicators used were accuracy, precision, and recall. The results of data classification showed that the highest accuracy value was 90% for data with 40 and 60 frames. While data with 10 frames had the lowest accuracy with 52% only. ResNext CNN-LSTM was able to detect deep-fakes in videos well even though the size of the image was small.

Keywords: Deepfake; ResNext CNN; LSTM

Introduction

Computer vision is a computer science branch leading to object detection by a computer. The development of computer science with the computer vision concept is based on how humans see objects with their eyes, then the objects are interpreted by the brain to be recognized. The process of detection and tracking is an important area in computer vision in different objects [1]. The implementation of this concept can be done for human object detection, image segmentation, estimation of human poses, video classification, and object movement tracking. The success of computer vision will provide impetus for new research focusing on finding methods that provide more efficient performance [2].

Face detection is one implementation of computer vision [3]. Face can be recognized by getting the coordinates of the face in the image. By performing face detection, we can obtain the position and size of the face in an image. The feature extraction subsystem functions to extract the features in the face area. The face recognition subsystem works for comparing the input face image with a group of faces in a database, so that finally we can determine the level of recognition of the face image. An issue that can be overcome by performing facial recognition is identifying facial fakes or deep fakes [4].

Deepfake is a technique for image synthesis in which a person in an image or video is replaced with another person's face [5]. Although the act of counterfeiting content is not new, there are many new technologies that make it difficult to distinguish between original and fake contents [6]. Deepfake utilizes machine learning and artificial intelligence techniques to manipulate or generate visual and audio content that is used to spread false information. Deepfake video creation also uses deep learning and other artificial neural network methods such as autoencoders or generative adversarial networks (GANs). In recent years, deepfakes have been used widely so that a system is needed to detect deepfakes to minimize news with forgery and hoax[7].

Research on deepfakes has been carried out by various researchers, using the CNN method and the matrix learning function. In this study, the method used would require large-scale video data [8]. Another research used the Faster Region based Convolutional Neural Network (Faster R-CNN) method to identify facial fakes [9]. Pranjal

http://dx.doi.org/10.33096/ilkom.v14i3.1254.178-185

Ranjan et al [10] detected deepfakes in videos using CNN and Transfer Learning (TL). The dataset used came from video extracted into an image measuring 128*128 pixels. Furthermore, the data was classified in two ways, namely using CNN alone, and using CNN combined with TL. The detection results showed that the highest accuracy reached 86.49% for data classification using CNN and TL.

This current research has differences from previous research in many ways. The deepfake identification process was carried out by combining several methods for the deepfake detection process. The feature extraction process was carried out using the ResNext Convolution neural network method and the classification process was carried out using the Long Short-Term Memory (LSTM) method. The system testing analysis process was carried out using two methods, namely blackbox and confusion matrix. With this research, various forgeries and hoax information circulating can be minimized. Besides that, it provides good accuracy analysis of the ResNext CNN and LSTM methods in identifying deepfakes.

Method

This research utilized two methods in the deepfake detection process, the feature extraction process using ResNext CNN. Furthermore, from the previous process, information was obtained that can be used as a reference in the classification process with the LSTM method. The evaluation process was carried out on two ways, namely from the functional way of the application and the performance of the methods used in the detection process. Each of these testing methods were blackbox and confusion matrix.

A. Dataset

The data used in this study was digital image. In deepfake detection, data wass obtained through data provider sites that have been widely used by previous researchers. The data would be used in making the model, besides that it was also used to evaluate the build system model. In general, research data was divided into 2, first was training data that was data to train and to build models and second was data testing which was the data sets used in testing models after the training process completed. The average length of all videos was 10 seconds with a standard frame rate of 30 frames per second see in **Figure 1**.

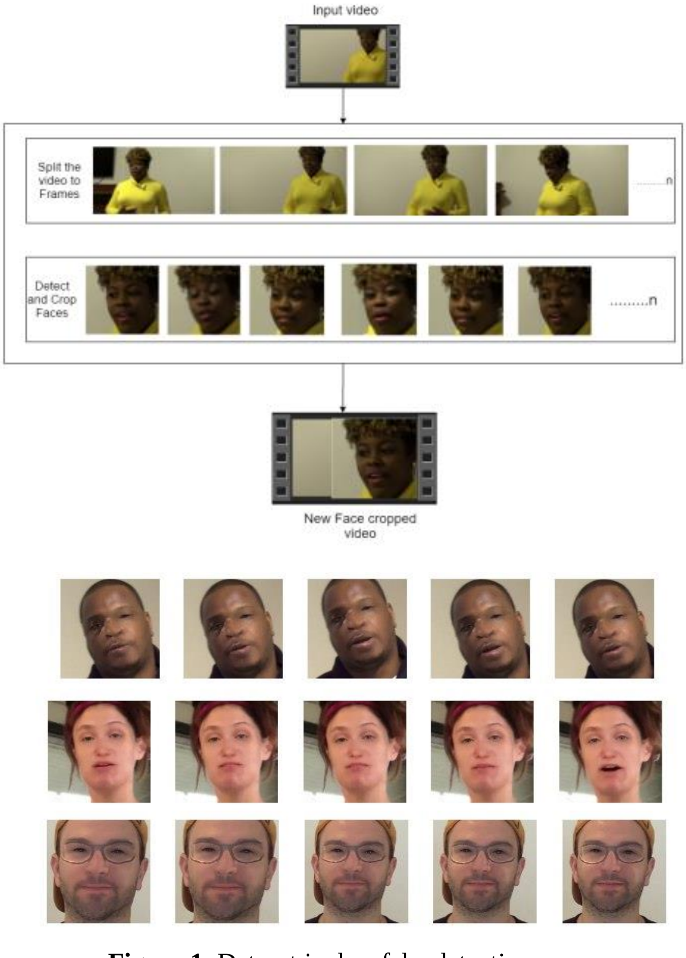


Figure 1. Dataset in deepfake detection

B. Research stages

The research started by collecting data in the form of a video. The obtained data would go through the preprocessing stage. This process changed the data into a dataset form that can be accepted by the system. And then all unnecessary things and noise in the video were removed. Only the required portion of the video were saved. The results of the preprocessing would be stored and labeled as detected face markers. Furthermore, the labeled data would be converted into a binary image with the feature extraction process using the ResNext CNN method. After performing data extraction, the data was classified into categories of fake data or real data. The final step would be to carry out the process of analyzing the accuracy of the detection process with the confusion matrix method. Based on **Figure 2** is research stages.

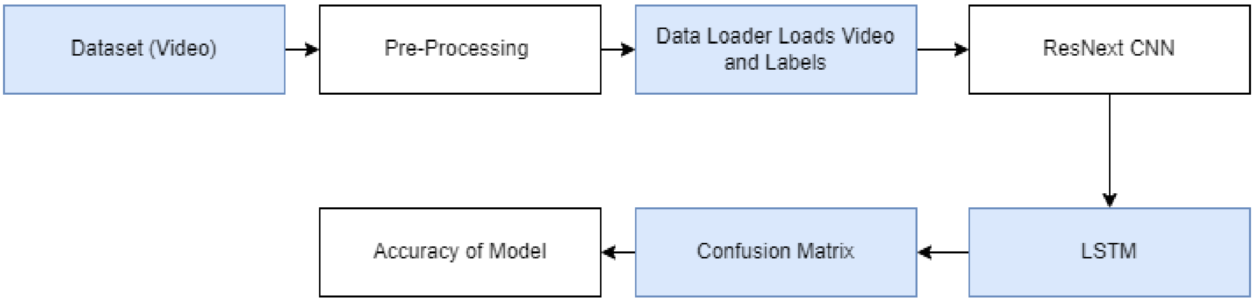


Figure 2. Research stages

C. Planning the Model

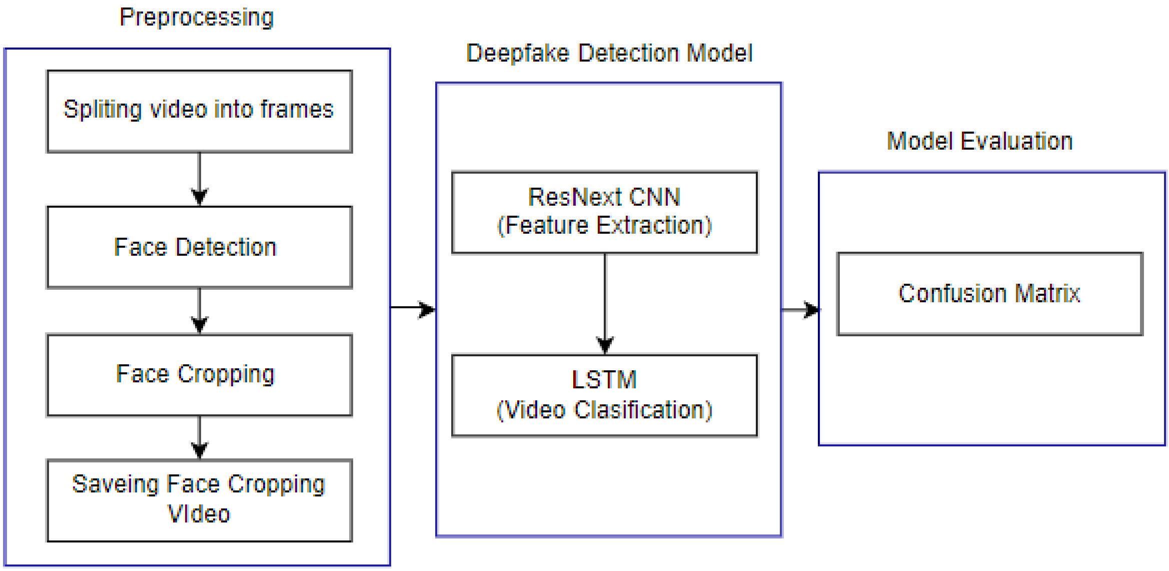

Preprocessing
Spliting video into frames
Face Detection
Face Cropping
Saveing Face Cropping VIdeo
Deepfake Detection Model
ResNext CNN (Feature Extraction)
LSTM (Video Clasification)
Model Evaluation
Confusion Matrix

Figure 3. Deepfake Model Planning

1) Preprocessing

Based on **Figure 3** is Deepfake model planning. Preprocessing is the initial stage in object recognition by combining the concepts of digital image, pattern recognition, mathematics, and statistics. In the pre-processing stage, the captured image was processed first to adjust in size and converted into a scale form that was suitable for training data and trial data focusing on the face. Frames without faces were ignored. The final process in preprocessing stage was storing the results of the faces that had been processed in the previous stage. Each video had an average duration of 10 seconds and was divided into 30 frames per second with a resolution of 112*112.

2) Deepfake Detection Model

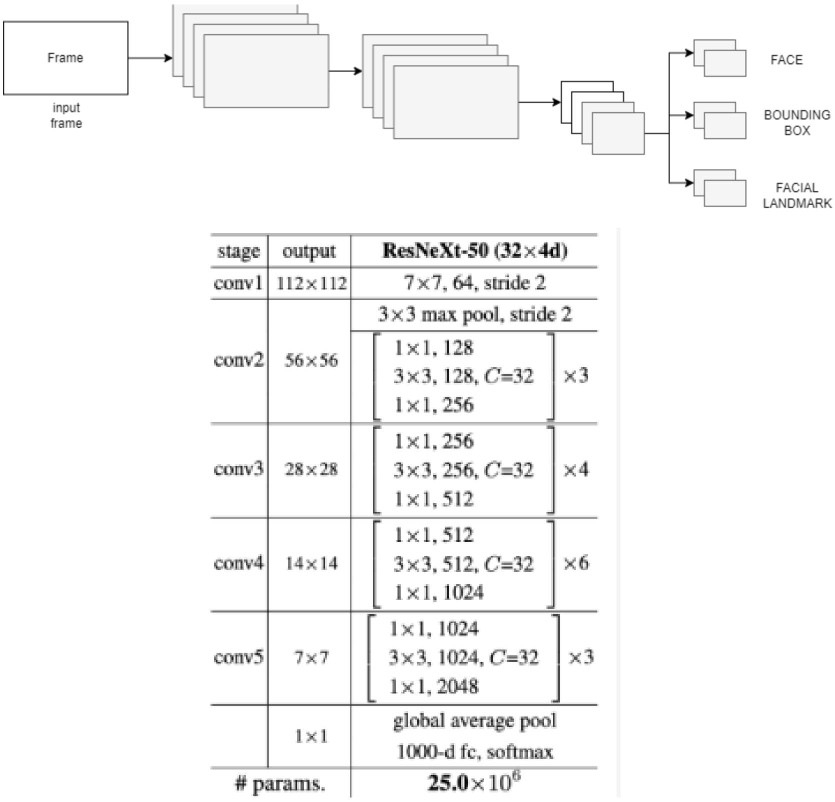
In this study, a combination of methods would be used when performing deepfake detection. The combined methods ware the ResNext CNN and LSTM methods. The feature extraction process would be carried out by the CNN ResNext method while LSTM performed the classification. Data that has gone through the preprocessing process would be inputed into the model.

a) ResNext Convolutional Neural Network (ResNext CNN)

Convolutional Neural Network (CNN) is a development of Multilayer Perceptron (MLP) designed to process two-dimensional data. CNN is included in the type of Deep Neural Network because of its high network depth and is widely applied to image data. In general, the convolutional layer detects edge features in the image, then the subsampling layer would reduce the dimensions of the features obtained from the convolutional layer, and finally passed them on to the output node through the forward propagation process [11][12][13].

$$c_i = f(W.b_c^{t} + b) \tag{1}$$

ResNext CNN is one of the optimizations of CNN in getting more optimal results. ResNext CNN is a method that has an advantage in maintaining computational complexity. It has good performance in reducing 6.7% errors in data processing [14]. This method has a simple design in extraction [15][16]. In this study the ResNext CNN method was used in the feature extraction process, the results of the extraction would be classified using the LSTM method. Feature Extraction process using ResNext CNN see in **Figure 4**.



stage	output	ResNeXt-50 (32×4d)
conv1	112×112	7×7, 64, stride 2
conv2	56×56	3×3 max pool, stride 2 $\begin{bmatrix}1\times1, 128\\ 3\times3, 128, C{=}32\\ 1\times1, 256\end{bmatrix}\times3$
conv3	28×28	$\begin{bmatrix}1\times1, 256\\ 3\times3, 256, C{=}32\\ 1\times1, 512\end{bmatrix}\times4$
conv4	14×14	$\begin{bmatrix}1\times1, 512\\ 3\times3, 512, C{=}32\\ 1\times1, 1024\end{bmatrix}\times6$
conv5	7×7	$\begin{bmatrix}1\times1, 1024\\ 3\times3, 1024, C{=}32\\ 1\times1, 2048\end{bmatrix}\times3$
	1×1	global average pool 1000-d fc, softmax
# params.		**25.0**×10^6

Figure 4. Feature Extraction process using ResNext CNN

b) Long Short-Time Memory(LSTM)

LSTM algorithm is a component widely used in the Recurrent Neural Networks (RNN). The algorithm can rely on its learning sequence on pattern recognition problems [17]. It has three states that help the network to reduce long term dependency data. It has three steps, forget step, input step and output step. The forget state was used to remove redundant or useless data. The input state functioned to process new data while the output step was to process input data with cell status. The formulas of the three steps can be seen below [18]:

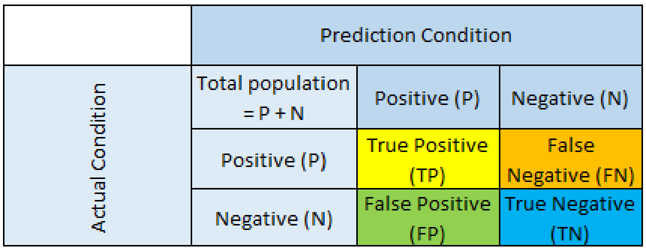
$$f_i^{(t)} = \sigma(b_i^f + \textstyle\sum_j U_i^f,\ j\, x_j^t + \sum_j W_i^f,\ j\, h_j^{(t-1)}) \tag{2}$$

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma(b_i + \textstyle\sum_j U_i^f,\ j\, x_j^t + \sum_j W_i^f,\ j\, h_j^{(t-1)}) \tag{3}$$

$$q_i^{(t)} = \sigma(b_i^o + \textstyle\sum_j U_i^f,\ j\, x_j^t + \sum_j W_i^f,\ j\, h_j^{(t-1)}) \tag{4}$$

3) Model Evaluation

Classification algorithm performance can be measured using a confusion matrix. There are several evaluation indicators that can be used. In this research, the indicators used were accuracy, precision, and recall. Accuracy described the percentage of correct classification of deep-fake video detection presented follows[19]:



		Prediction Condition	
Actual Condition	Total population = P + N	Positive (P)	Negative (N)
	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

$$Akurasi = \frac{TP+TN}{TP+FP+FN+TN} \qquad (2)$$

$$Presisi = \frac{TP}{TP+FP} \qquad (3)$$

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

The Confusion Matrix contains all the raw information about the predictions made by a classification model on a given data set. To evaluate the accuracy of model generalizations, it is common to use test data sets that are not used during the learning process of the model [20].

Results and Discussion

Deepfake detection in this study was carried out using the ResNext CNN method with facial images extraction taken from the frame. Frame captures were done in stages with a 10-second video data set. Each frame was intended to get frames with different facial expressions or positions. In this study each video was analyzed based on the number of frames. The number of frames taken started from 10, 20, 40, 60. Each frame had different results which would be categorized into, true positive, false positive, false negative, true negative see in **Figure 5**, **Figure 6** and **Figure 7**.

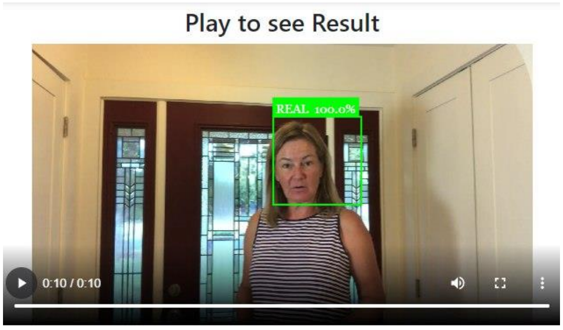


Figure 5. True Positive Detector

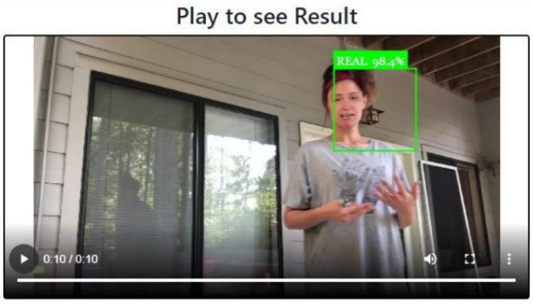


Figure 6. False Negative Detector

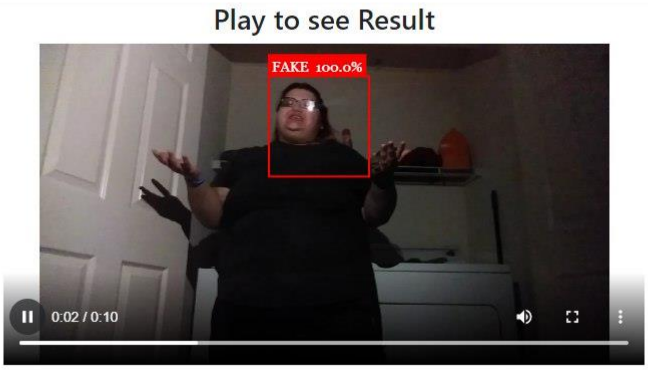


Figure 7. True Negative Detector

Table 2. Table of Deepfake System Testing

No	**Nama File**	**10 frames**	**20 frames**	**40 frames**	**60 frames**
		Label	**Label**	**Label**	**Label**
1	Video_1	Real	Fake	Fake	Real
2	Video_2	Fake	Fake	Real	Fake
3	Video_3	Real	Fake	Fake	Fake
4	Video_4	Real	Real	Fake	Fake
5	Video_5	Fake	Fake	Fake	Fake
6	Video_6	Real	Fake	Fake	Fake
7	Video_7	Fake	Fake	Fake	Fake
8	Video_8	Real	Fake	Fake	Fake
9	Video_9	Real	Fake	Fake	Real
10	Video_10	Real	Real	Real	Fake
…	…	…	…	…	…
99	Video_99	Fake	Fake	Real	Fake
100	Video_100	Real	Fake	Real	Fake

Based on system testing on 100 tested data, the TP, TN, FP, and FN values for each test were obtained, see in **Table 2**. In a video with 10 data frames, the accuracy was 52%, while precision and recall were 52% and 50% respectively. This value was obtained from TP of 25, TN of 27, FP of 23 and FN of 25.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{25+27}{100} = 52\%$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{25}{25+23} = 52\%$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{25}{25+25} = 50\%$$

Table 3. Table of Testing Confusion Matrix

Number of Frames	**Classification Results (%)**		
	Accuracy	**Precision**	**Recall**
10 Frames	52%	52%	50%
20 Frames	88%	88%	97%
40 Frames	90%	90%	97%
60 Frames	90%	100%	97%`

Based on the classification results in **Table 1**, the performance of CNN and LSTM was lowest on video data with 10 frames with 52% accuracy. The accuracy of the ResNext CNN and LSTM methods increased on video with 20 frames reaching 88% and achieving stability on data video with 40 frames and 60 frames with 90% accuracy. It was similar with the value of precision and recall. Deepfake classification for data with 10 frames indicated lower accuracy, precision, and recall values than data with a larger number of frames, see in **Table 3**.

Conclusion

Deep-fake video detection highly recommended to be done to find out whether information is genuine or fake. It can prevent mistakes in making decisions. Deep-fake video classification was conducted by processing video data, training the classification algorithm, evaluating, and comparing the performance of the algorithm based on video data on deep-fake video and real video. ResNext CNN was used to process video data. In detecting deep-fake videos, the videos should be extracted into an image so that the image can be cropped to a smaller size, a size of 100 x 100 pixels. This was to make the classification process easier. LSTM was the model used to classify image data and model performance needs to be tested using indicators of accuracy, precision, and recall. Classification results using ResNext CNN-LSTM showed that the highest accuracy was 90%, precision reached 100% and recall reached 97% for data from 60 frames. While data from 10 frames of video had lower performance with 52% accuracy, 52% precision and 50% recall. Based on the experimental results, it can be concluded that the combination of ResNext CNN and LSTM produced high performance even though the image size was only 100 x 100 pixels. In addition, ResNext CNN-LSTM also had good performance on video data with 20 frames.

For further research, it is better to explore the function of the LSTM to have more optimal performance. Some researchers have also explored certain facial features as datasets to include in models, for example eyes, nose, ears, or mouth so it is interesting to compare model performance between models trained with the entire face and models trained with partial facial features.

References

[1] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, “Computer vision and deep learning techniques for pedestrian detection and tracking: A survey,” *Neurocomputing*, vol. 300, pp. 17–33, 2018.

[2] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 2818–2826, 2016.

[3] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, and J. Zhou, “CosFace: Large Margin Cosine Loss for Deep Face Recognition,” *Cvpr*, pp. 5265–5274, 2018.

[4] O. Victoria and I. P. Solihin, “Pendeteksi wajah secara realtime menggunakan Metode Eigenface,” *SEINASI-KESI (Seminar Nas. Inform. Sist. Inf. Dan Keamanan Siber)*, pp. 126–131, 2018.

[5] D. Pan, L. Sun, R. Wang, X. Zhang, and R. O. Sinnott, “Deepfake Detection through Deep Learning,” *Proc. - 2020 IEEE/ACM Int. Conf. Big Data Comput. Appl. Technol. BDCAT 2020*, pp. 134–143, 2020.

[6] M. Westerlund, “The emergence of deepfake technology: A review,” *Technol. Innov. Manag. Rev.*, vol. 9, no. 11, pp. 39–52, 2019.

[7] S. T. Suganthi *et al.*, “Deep learning model for deep fake face recognition and detection,” *PeerJ Comput. Sci.*, vol. 8, pp. 1–20, 2022.

[8] S. Agarwal, H. Farid, T. El-Gaaly, and S. N. Lim, “Detecting Deep-Fake Videos from Appearance and Behavior,” *2020 IEEE Int. Work. Inf. Forensics Secur. WIFS 2020*, 2020.

[9] S. Megawan, W. S. Lestari, and A. Halim, “Deteksi Non-Spoofing Wajah pada Video secara Real Time Menggunakan Faster R-CNN,” vol. 3, no. 3, 2022.

[10] P. Ranjan, S. Patil, and F. Kazi, “Improved generalizability of deep-fakes detection using transfer learning based CNN framework,” *Proc. - 3rd Int. Conf. Inf. Comput. Technol. ICICT 2020*, pp. 86–90, 2020.

[11] I. A. Anjani, Y. R. Pratiwi, and S. Norfa Bagas Nurhuda, “Implementation of Deep Learning Using Convolutional Neural Network Algorithm for Classification Rose Flower,” *J. Phys. Conf. Ser.*, vol. 1842, no. 1, 2021.

[12] Q. Zhang, M. Zhang, T. Chen, Z. Sun, Y. Ma, and B. Yu, “Recent advances in convolutional neural network acceleration,” *Neurocomputing*, vol. 323, pp. 37–51, 2019.

[13] M. A. Saleem, N. Senan, F. Wahid, M. Aamir, A. Samad, and M. Khan, "Comparative Analysis of Recent Architecture of Convolutional Neural Network," *Math. Probl. Eng.*, vol. 2022, 2022.

[14] K. Annapurani and D. Ravilla, "CNN based image classification model," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 11 Special Issue, pp. 1106–1114, 2019.

[15] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 5987–5995, 2017.

[16] G. Pant, D. P. Yadav, and A. Gaur, "ResNext convolution neural network topology-based deep learning model for identification and classification of Pediastrum," *Algal Res.*, vol. 48, no. April, p. 101932, 2020.

[17] P. N. Srinivasu, J. G. Sivasai, M. F. Ijaz, A. K. Bhoi, W. Kim, and J. J. Kang, "Classification of skin disease using deep learning neural networks with mobilenet v2 and lstm," *Sensors*, vol. 21, no. 8, pp. 1–27, 2021.

[18] C. I. Garcia, F. Grasso, A. Luchetta, M. C. Piccirilli, L. Paolucci, and G. Talluri, "A comparison of power quality disturbance detection and classification methods using CNN, LSTM and CNN-LSTM," *Appl. Sci.*, vol. 10, no. 19, pp. 1–22, 2020.

[19] X. Li, S. Li, J. Li, J. Yao, and X. Xiao, "Detection of fake-video uploaders on social media using Naive Bayesian model with social cues," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, 2021.

[20] O. Caelen, "A Bayesian interpretation of the confusion matrix," *Ann. Math. Artif. Intell.*, vol. 81, no. 3–4, pp. 429–450, 2017.