DOI: 10.24996/ijs.2025.66.2.17

ISSN: 0067-2904

A Hybrid Model of Deep Learning and Machine Learning Methods to Detect Deepfake Videos

Noor K. Alzurfi*, Mohammed S. Altaei

Department of Computer Science, College of Science, Al-Nahrain University, Baghdad, Iraq

Received: 20/9/2023 Accepted: 26/1/2024 Published: 28/2/2025

Abstract

The deepfake videos were spread in the last few years and were created by different deepfake techniques (i.e., faceswap, face2face, etc.). These techniques have a terrible impact on society and would give anyone a chance to create videos with fake faces. The objective of this paper was to develop a model that detects deepfake videos to reduce their negative impact. Two hybrid models of machine learning and deep learning were proposed. The first model used a convolutional neural network (CNN) for feature extraction and different machine learning classifiers (support vector machine (SVM), k-nearest neighbor (KNN), decision tree (DT), random forest (RF), logistic regression, and naive bayes (NB)). In contrast, the second model used a transfer learning concept developed by the VGG16 (Visual Geometry Group) pre-trained model and the same machine learning classifiers as the first model. Both models were evaluated on the FaceForensics++ video dataset, which includes four different deepfake techniques (Deepfake, Faceswap, Face2Face, and Neuraltexture). The results showed good accuracy, which proved the effectiveness of the proposed models, which may be used as a detection deepfake application. While the first model can obtain the highest accuracy with the SVM classifier on the four deepfake techniques sequentially: 0.96, 0.87, 0.90, and 0.64. In contrast, the second model achieved the highest accuracy with the KNN classifier on Deepfake and Face Swap techniques: 0.95 and 0.91, and with SVM on Face2Face and Neural Texture techniques: 0.86 and 0.77.

Keywords: Transfer learning (TL), Convolution neural network (CNN), Machine learning, Deepfake video, Deepfake detection

نموذج هجين بين التعلم العميق واساليب التعلم الالي للكشف عن مقاطع الفيديو ذات التزييف العميق

نور كاظم الزرفي, محمد صاحب الطائي

قسم علوم الحاسوب, كلية العلوم, جامعة النهرين, بغداد, العراق

الخلاصة

انتشرت في السنوات الأخيرة الفيديوهات ذات التزييف العميق التي أنتجت بواسطة مختلف تقنيات التزييف العميق (مثال: faceswap, face2face...إلخ). هذه التقنيات كان لها تأثير سلبي على المجتمع حيث تمنح الفرصة لأي شخص لتكوين فيديوهات بأوجه مزيفة. الهدف من هذا البحث هو تكوين نموذج لكشف

*Email: noor.kadem21@ced.nahrainuniv.edu.iq

الفيديوهات ذات التزييف العميق للحد من تأثيرها السلبي. لقد تم اقتراح بهذا البحث نموذجي هجينين بين التعلم العميق وأساليب التعلم الآلي. تم استعمال في النموذج الأول الشبكة العصبية اللافتة لاستخراج الخواص ومجموعة مختلفة من مصنفات التعليم الآلي(آلة المتجهات الداعمة, خوارزمية K أقرب الجيران , الانحدار اللوجستي ,خوارزمية شجرة القرار ,الغابة العشوائية ,المصنف البايزي الساذج). بينما في النموذج الثاني تم استعمال مفهوم نقل التعلم بواسطة النموذج المدرب مسبقا (VGG16) وذات المصنفات التعلم الالي المستخدمة في النموذج الاول. تم تقييم كلا النموذجين على البيانات الفديوية FaceForensics++ التي تتكون من اربع تقنيات مختلفة للتزييف العميق(Deepfake, Faceswap, Face2Face, and Neuraltexture). حيث اوضحت النتائج كفاءة النماذج المقترحة التي من الممكن استعمالها في تطبيقات الكشف عن الفيديوهات المزيفة. استطاع النموذج الاول الحصول على افضل نتائج باستعمال مصنف ال SVM بتطبيقه على التقنيات الاربعة للتزييف العميق تتابعا: 0.96 , 0.87, 0.90 و 0.64. اما النموذج الثاني حقق اعلى نتائج باستعمال مصنف ال KNN بتطبيقه على تقنية Deepfakeو Faceswap: 0.95 و 0.91 , ومع مصنف ال SVM بتطبيقه على تقنية Face2faceو Neuraltexture: 0.86 و 0.77.

1. Introduction

In recent years, deepfake multimedia has become a severe crisis in our society [1]. Deep learning made the applications and tools easier to use, and that would help users generate fake videos and images without the need for any experience in this field [2, 3]. The term ‘deepfake’ is a derivative of ‘deep learning’ and ‘fake’ terms. Deepfake algorithms help users create new images or videos that show people saying or doing something that they never did. This operation is done by swapping faces between target and original images or videos using an autoencoder or generative adversarial network (GAN) [1, 4]. In contrast, deepfake technology opens the door to productive possibilities such as movie productions, photography, Snapchat filters, and video games [4,5]. It has social, political, and legal issues such as spreading incorrect information, soiling celebrities' reputations, and blackmailing [2, 6]. Since the deepfake videos require just a few face photos to do the operation of the face swapping, some of the bad users use the available photos on the internet to generate fake videos, such as switching out pornographic heroes with female celebrities, creating phony movies for politicians, business leaders, and other powerful individuals, and utilizing those fake, artificial videos to lend money to other people [2].

There are several methods to create deepfake videos. The two most often used techniques are auto-encoders and generative adversarial networks (GANs) [5, 7, 8]. The encoder and decoder are the two parts of an auto-encoder. In the deepfake method, two auto-encoders are trained to pass latent faces between the source and the target video frames. By feeding these recovered characteristics to two decoders, the encoder could extract latent features from the picture and recreate faces. As a result, face A's created face will be given to decoder B. By using features from the face, decoder B would attempt to rebuild face B. Every frame in the video is created by repeating this technique [4]. A discriminator and a generator are the two neural networks that make up a GAN. The discriminator was trained to better differentiate between false and genuine images, while the generator was employed to create images that were more like the real ones [5, 9]. Many researchers have proposed and used different models to detect deepfake videos and tried to reach the best results to solve deepfake video problems. Most of them used deep learning methods in their models, and some of them used machine learning methods. This paper provides a combination of models between deep learning and machine learning methods to get an effective deepfake detection model that helps reduce the danger of deepfake videos. The paper covers the following sections: Section 2 presents the related works; Section 3 describes the methodology; Section 4 presents the results and discussion; and finally, Section 5 presents the conclusion.

2. RELATED WORK

Recently, the deepfake multimedia (images and videos) detection process has gained popularity. Numerous researchers have suggested various machine learning and deep learning methods for the detection and classification of deepfake videos and images. Here is a summary of some researchers who presented different models to detect deepfake videos:

1. Rana et al. [6] employed the conventional method of training and testing machine learning classifiers after using some feature extraction and selection strategies to identify Deepfakes. They obtained high accuracies for a few datasets: 99.84% for FaceForensics++, 99.66% for VDFD, 99.38% for the DeepFake Detection Challenge, and 99.43% for Celeb-DF datasets.
2. Mitra et al. [10] stated a neural network model consisting of different structures from a convolutional neural network (ResNet50, InceptionV3, Xception) for feature extraction and a classifier network for deepfake video detection. The Xception network achieved a high accuracy of about 96% and 93% for different compression levels on the FaceForensics++ dataset.
3. Cunha et al. [11] used the pre-trained EfficientNet-B4 model, one of the convolutional neural networks (CNN), to detect deepfake videos. They achieved 95% accuracy over the Celeb-DF (Celebrities-DeepFake) dataset.
4. Masood et al. [12] have used pre-trained convolutional neural network models to extract features and used an SVM classifier to classify the fake and real videos. The DFDC dataset was used and obtained the highest accuracy of 98% for the DenseNet-169 model and the lowest accuracy of 89% for VGG-16.
5. Nawaz et al. [13] have proposed a new technique to detect faceswap-based deepfake. They computed landmarks from input videos using the Dlib library and used them as features to train SVM and ANN (artificial neural network) classifiers. This method was done over five videos (Hillary Clinton, Elizabeth Warren, Donald Trump, Bernie Sanders, and Barack Obama) and achieved high accuracy, about 99% by SVM and about 98% by ANN.
6. Mallet et al. [14] have suggested support vector machine and convolutional neural network algorithms to detect deepfake images. Both algorithms applied the extracted features of CNN as the first step and obtained an accuracy of about 88% for CNN and 81% for SVM over 140k real and fake face image datasets.
7. Raza et al. [15] have proposed a new approach that consists of VGG16 and a convolutional neural network (CNN) to detect the deepfake content. They also used transfer learning by Xception, NAS-Net, Mobile Net, and VGG16 techniques for comparison. This approach achieved an accuracy of about 94% on the deepfake dataset.

In this paper, we have proposed two new hybrid models that combine ML and DL to detect deepfake video. The first hybrid model used CNN to extract features and different ML classifiers (support vector machine, k-nearest neighbor, decision tree, random forest, logistic regression, and naive bayes) to classify deepfake and real videos, while the second hybrid model used transfer learning (VGG16 pretrained model) for feature extraction and different ML classifiers (support vector machine, k-nearest neighbor, decision tree, random forest, logistic regression, and naive bayes) to classify the deepfake and real videos. The FaceForensics++ dataset has been used in both models, which have different types of deepfake methods, and it is one of the large video datasets.

3. METHOD

Convolutional neural networks, transfer learning, and different machine learning classifiers are used to detect deepfake videos. In this section, the first section (3.1) is a discussion of the FaceForensics++ video dataset. Then, Section 3.2 discusses the pre-processing phase to prepare the data for the models. Finally, in Section 3.3, the proposed models have been discussed. Figure 2 illustrates the steps of the proposed models.

3.1 Description of dataset

In this paper, the FaceForensics++ video dataset is used, which is one of the most widely used datasets in the deepfake detection field [16]. It consists of 5,000 videos with four different deepfake methods, which are Deepfake [17], Face2face [18], Faceswap [17], and Neuraltexture [17]. The dataset was obtained from the GitHub website [16] and is available on the Ka ggle website. Figure 1. shows the details of the dataset.

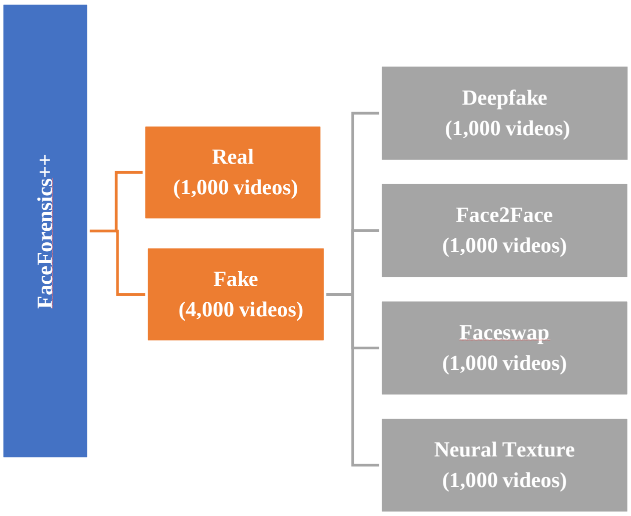


Figure 1: FaceForensics++ video dataset details

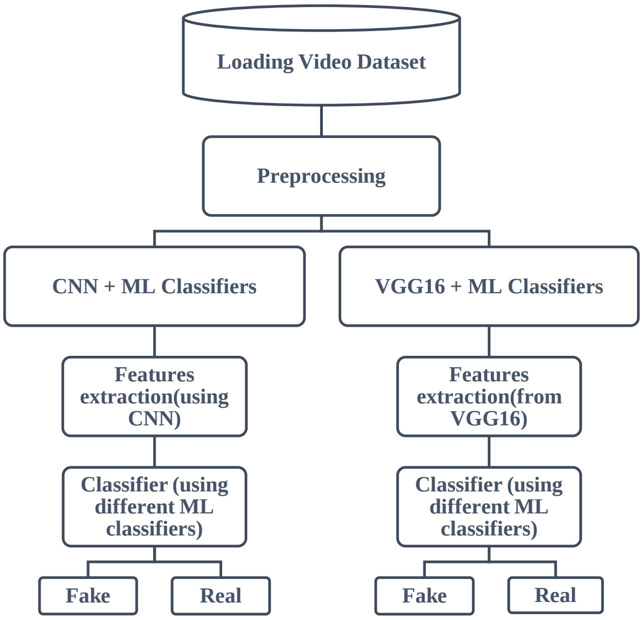

Loading Video Dataset
Preprocessing
CNN + ML Classifiers
VGG16 + ML Classifiers
Features extraction(using CNN)
Features extraction(from VGG16)
Classifier (using different ML classifiers)
Classifier (using different ML classifiers)
Fake
Real
Fake
Real

Figure 2: The proposed models

3.2 Pre-processing dataset

A video dataset must be converted to an image dataset for use in any model. As a first step, the FaceForensics++ video dataset was divided into four sub-datasets. Each sub-dataset contains 1,000 real videos and 1,000 videos from each deepfake tool. The four sub-datasets are FaceForensics++_Face2face, FaceForensics++_Deepfake, FaceForensics++_ Neuraltexture , and FaceForensics++_Faceswap. Secondly, all the videos should be converted to frames using the OpenCV library, and just one frame is chosen from each of the 25 extracted frames, which helps reduce the time and cost needed for computation. In the third step, cropping and resizing operations were applied to the extracted frames before saving them in a new folder. Figure 3 shows the pre-processing steps. The pre-processing steps produce from each 1,000 videos around 16k–20k frames. Each sub-dataset is split into a train set (85%) and a test set (15%).

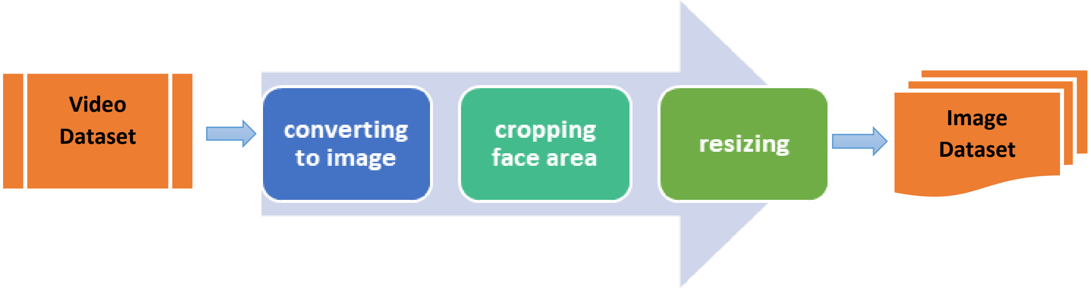


Figure 3: The preprocessing steps of the Dataset

3.3 Proposed Detection Models

3.3.1 CNN and Machine Learning Classifiers based Deepfake Detection Model

The CNN-ML model is suggested based on a hybrid of convolutional neural networks (CNN) [19] and machine learning classifiers [20] for deepfake video detection. CNN is the most commonly used deep learning method in image recognition and classification. It is one of the most effective and efficient methods for detecting deepfakes.

In the proposed method, the CNN is used as a feature extractor that works on extracting important features automatically from the pre-processing dataset to transfer them to the different classifiers. The first layers of the CNN architectures (convolution layer, max-pooling layer, dropout layer) worked on extraction features. The input size of the CNN is 32*32*3. The CNN model used for extracting features consists of three convolution layers with a 3*3 filter size, filter numbers of 32, 64, and 132 sequentially, and a ReLU activation function. These convolution layers are followed by three stacks of max-pooling layers with pool size = 2 and stride = 2. Additionally, two dropout layers with a dropout of 0.25 followed the final two max-pooling layers. The output (extraction features) was converted to one vector by a flattening layer. Finally, this vector of the extraction features will be fed to the fully connected layer (the dense layer) to transfer them to the different machine-learning classifiers to get results.

This model uses the six most common machine learning classifiers: k nearest neighbors (KNN), support vector machine (SVM), decision tree (DT), logistic regression (LR), naive bayes (NB), and random forest (RF) [21, 22]. These classifiers are implemented using the sklearn library. The extraction features are fed into each classifier to train it first. Then followed a classification process.

3.3.2 VGG16 and Machine Learning Classifiers based Deepfake Detection Model

Model VGG16-ML is proposed based on a hybrid of transfer learning [23], which uses the VGG16 model and different machine learning classifiers for deepfake video detection. VGG16 is a pre-trained model based on CNN architecture that is used on different image recognition tasks [24]. It trained on the ImageNet dataset, which has more than 1.2 million images and 1,000 classes. The input size of the VGG16 is 224*244*3.

In this approach, the richly learned features will transfer from the VGG16 to the new dataset. The last two layers of the VGG16 architecture were cut off to make this change. The fully connected layer's output (a learnable feature) was then sent to the dataset that was used. This dataset will then be fed to the different machine learning classifiers for training and sorting. The six most commonly used machine learning classifiers have been used in this proposed model (k nearest neighbor (KNN), support vector machine (SVM), decision tree (DT), logistic regression (LR), naive bayes (NB), and random forest (RF)) [21, 22]. These classifiers are implemented using the ‘sklearn’ library. The learnable features were fed into each classifier to train it first. Then followed a classification process.

4. RESULTS AND DISCUSSION

4.1 Implementation Details

The pre-processing operation is implemented in Google Colab with a GPU, while the proposed models are done on a Kaggle notebook. The Keras and TensorFlow libraries are used to implement the CNN and the VGG16. The dataset was split using the train-test-split function found in the Sklearn library. In the CNN model, the batch size was 32, and the Adam optimizer was used with a learning rate of 0.001 and 50 epochs for training.

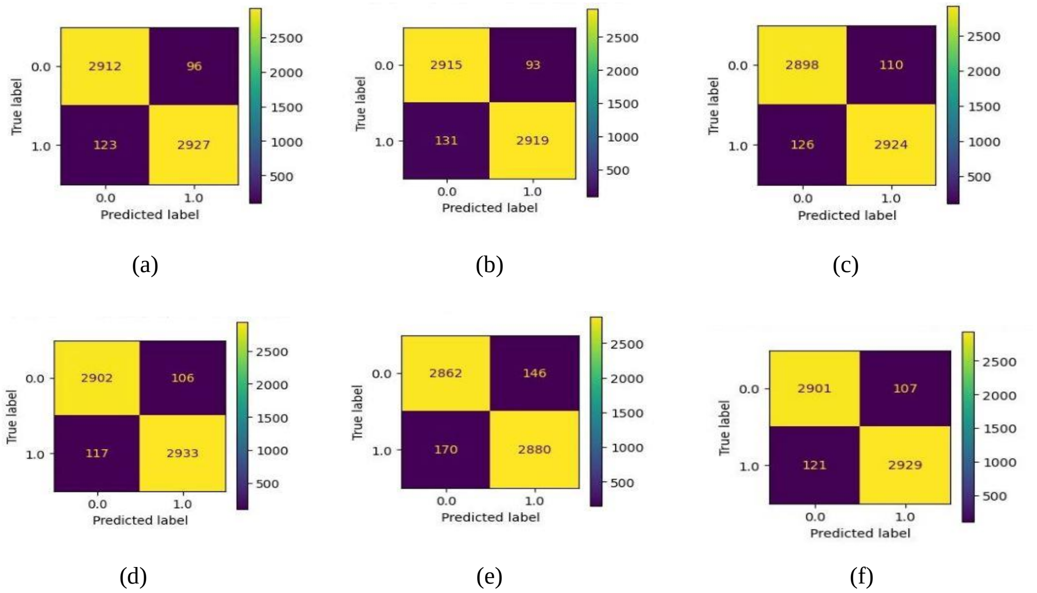
4.2 Models Results

The accuracy results of the CNN-ML classifier model are shown in Table 1. Different metrics are used to evaluate the proposed model (accuracy, F1-score, recall, and precision) [25]. This model has good performance with the FaceForensics++_Deepfake dataset, the FaceForensics++_Face2Face dataset, and the FaceForensics++_Faceswap dataset, which obtained the highest accuracy with the SVM classifier at 0.96, 0.87, and 0.90, respectively. The accuracy has a top value on the FaceForensics++_Deepfake dataset, while the lowest accuracy is on the FaceForensics++_Neuraltexture dataset. This difference in accuracy values belongs to the deepfake creation techniques. In comparison, the deepfake method depends on the whole face to generate the deepfake video, unlike the neural-texture method, which depends on the mouth area only. The neural-texture method makes the deepfake videos hard to detect and produces videos closest to the real videos, which makes the models confusing. In contrast, the results of the other metrics are relatively close to the accuracy results and show good performance. The precision results illustrate how the model can correctly predict the deepfake images, and the recall results show how the model can correctly identify the true deepfake images from all the deepfake images. Furthermore, the F1-score results proved that the misclassification of deepfake and real images is low.

Table 1: Results of the CNN-ML classifier model

Classifiers	Metrics	FaceForensics++_ Deepfake	FaceForensics++_ Face2face	FaceForensics++ _Faceswap	FaceForensics++_ Neuraltexture
SVM	Accuracy	0.96	0.87	0.90	0.64
	Recall	0.95	0.86	0.90	0.65
	F1-score	0.95	0.86	0.89	0.64
	Precision	0.96	0.87	0.89	0.64
KNN	Accuracy	0.96	0.86	0.90	0.61
	Recall	0.95	0.86	0.90	0.62
	F1-score	0.95	0.86	0.89	0.62
	Precision	0.96	0.87	0.89	0.64
RF	Accuracy	0.96	0.87	0.90	0.64
	Recall	0.95	0.86	0.90	0.65
	F1-score	0.95	0.86	0.89	0.65
	Precision	0.96	0.87	0.89	0.65
LR	Accuracy	0.96	0.87	0.90	0.64
	Recall	0.96	0.86	0.90	0.65
	F1-score	0.96	0.86	0.90	0.64
	Precision	0.96	0.87	0.90	0.64
DT	Accuracy	0.94	0.85	0.87	0.59
	Recall	0.94	0.84	0.88	0.60
	F1-score	0.94	0.84	0.86	0.60
	Precision	0.95	0.85	0.86	0.60
NB	Accuracy	0.96	0.86	0.89	0.63
	Recall	0.95	0.85	0.90	0.61
	F1-score	0.95	0.85	0.88	0.67
	Precision	0.96	0.87	0.88	0.75

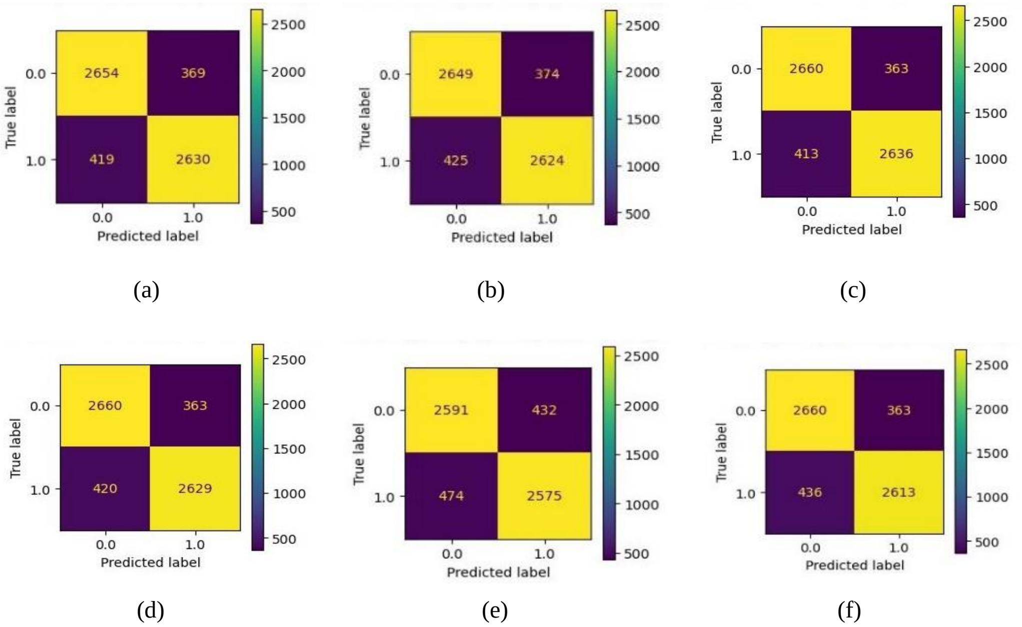
Figure 4 (a), (b), (c), (d), (e), and (f) show the confusion matrix (CM) [26] of the CNN_ML classifiers model on the FaceForensics++_Deepfake dataset. Figures 5 (a), (b), (c), (d), (e), and (f) show the confusion matrix of the CNN_ML classifier model on the FaceForensics++_Face2face dataset. Figures 6 (a), (b), (c), (d), (e), and (f) show the confusion matrix of the CNN_ML classifier model on the FaceForensics++_Faceswap dataset. The confusion matrix illustrates the performance of each classifier in the proposed model and how effectively the different deepfake methods are classified. As noticed, Figure 4 (e), Figure 5 (e), and Figure 6 (e) have more misclassification by the decision tree classifier than other classifiers. The SVM classifier in the FaceForensics++_Deepfake dataset has a best classification at Figure 4(a), while the random forest classifier in the FaceForensics++_Face2face dataset and the FaceForensics++_Faceswap dataset have a best classification at Figures 5(c) and 6(c). Figure 7(a), (b), (c), (d), (e), and (f) show the CNN_ML classifiers model on the FaceForensics++_Neuraltexture dataset. In Figures 7 (e) and (f), the decision tree and naïve bayes classifiers have more misclassification than other classifiers. while the SVM and random forest classifiers in Figures 7(a) and (c) have the best classification. In general, the FaceForensics++_Neuraltexture dataset has misclassification in all classifiers compared with the remaining datasets.



(a) (b) (c)

(d) (e) (f)

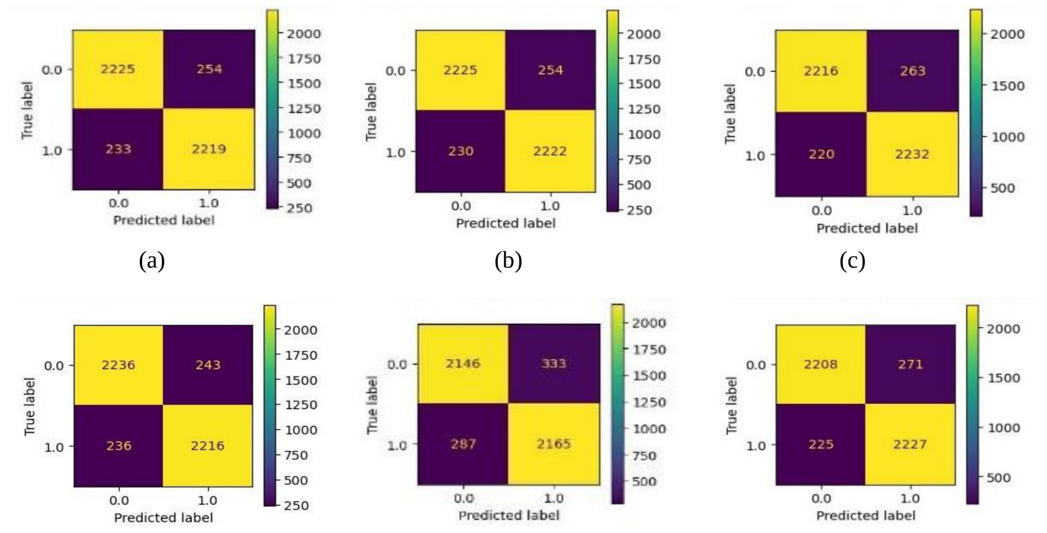
Figure 4: The confusion matrix of the CNN_ML classifier model on the FaceForensics++_Deepfake dataset, (a) CM of SVM classifier, (b) CM of KNN classifier, (c) CM of RF classifier, (d) CM of LR classifier, (e) CM of DT classifier, and (f) CM of NB classifier



(a) (b) (c)

(d) (e) (f)

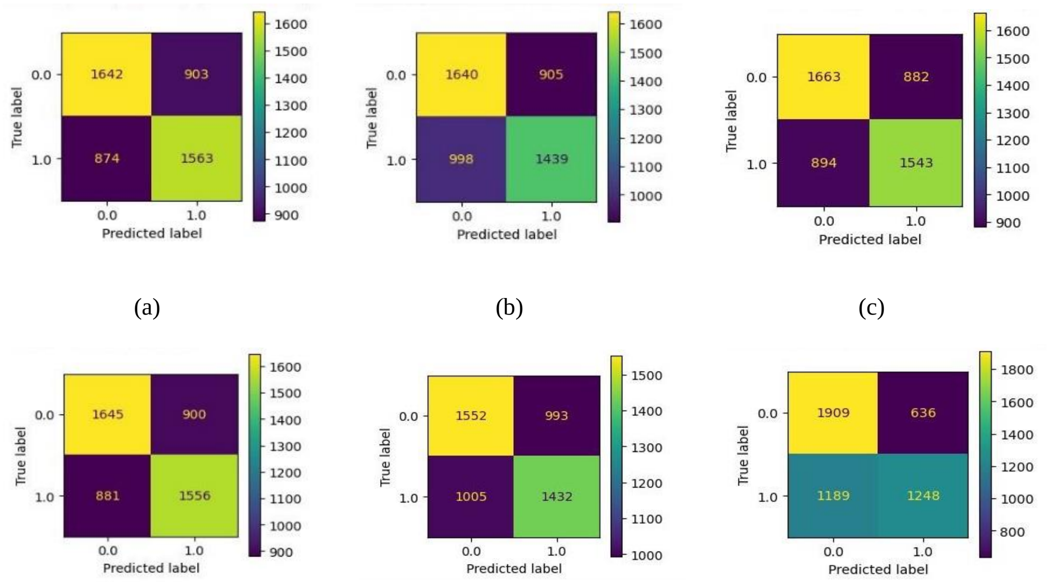
Figure 5: The confusion matrix of the CNN_ML classifier model on the FaceForensics++_Face2face dataset, (a) CM of SVM classifier, (b) CM of KNN classifier, (c) CM of RF classifier, (d) CM of LR classifier, (e) CM of DT classifier, and (f) CM of NB classifier



(a) (b) (c)

(d) (e) (f)

Figure 6: The confusion matrix of the CNN_ML classifier on FaceForensics++_Faceswap dataset, (a) CM of SVM classifier, (b) CM of KNN classifier, (c) CM of RF classifier, (d) CM of LR classifier, (e) CM of DT classifier, and (f) CM of NB classifier



(a) (b) (c)

(d) (e) (f)

Figure 7: The confusion matrix of the CNN_ML classifier on FaceForensics++_Neuraltexture dataset, (a) CM of SVM classifier, (b) CM of KNN classifier, (c) CM of RF classifier, (d) CM of LR classifier, (e) CM of DT classifier, and (f) CM of NB classifier

The accuracy results of the VGG16-ML classifier model are shown in Table 2. In this table, we notice that the classifiers' results are not approximately like the first approach. The SVM and KNN get better results than the other four classifiers in the FaceForensics++_Deepfake and FaceForensics++_Faceswap datasets. The FaceForensics++_Face2face and FaceForensics++_Neuraltexture datasets have the best results with SVM and LR classifiers.

In contrast, the results of the other metrics are relatively close to the accuracy results and show good performance. The precision results illustrate how the model can correctly predict the deepfake images, and the recall results show how the model can correctly identify the true deepfake images from all the deepfake images. Furthermore, the F1-score results proved that the misclassification of deepfake and real images is low. Except for the precision and F1-score of the NB classifier on the FaceForensics++_Deepfake and FaceForensics++_Neuraltexture datasets, their low results illustrate how the misclassification of deepfake and real images is high.

Table 2: Results of the VGG16-ML classifier model

Classifiers	Metrics	FaceForensics++_ Deepfake	FaceForensics++_ Face2face	FaceForensics++ _Faceswap	FaceForensics++_ Neuraltexture
SVM	Accuracy	0.92	0.86	0.88	0.77
	Recall	0.92	0.86	0.89	0.76
	F1-score	0.92	0.86	0.87	0.76
	Precision	0.93	0.87	0.87	0.77
KNN	Accuracy	0.95	0.74	0.91	0.56
	Recall	0.95	0.70	0.90	0.53
	F1-score	0.95	0.77	0.90	0.60
	Precision	0.96	0.86	0.91	0.71
RF	Accuracy	0.85	0.76	0.81	0.65
	Recall	0.84	0.75	0.80	0.64
	F1-score	0.84	0.76	0.81	0.64
	Precision	0.86	0.78	0.83	0.66
LR	Accuracy	0.87	0.82	0.85	0.75
	Recall	0.88	0.81	0.85	0.74
	F1-score	0.87	0.81	0.85	0.74
	Precision	0.87	0.83	0.85	0.75
DT	Accuracy	0.68	0.60	0.65	0.56
	Recall	0.69	0.60	0.64	0.55
	F1-score	0.68	0.60	0.64	0.55
	Precision	0.68	0.60	0.64	0.55
NB	Accuracy	0.54	0.58	0.66	0.53
	Recall	0.72	0.56	0.73	0.61
	F1-score	0.23	0.66	0.59	0.25
	Precision	0.14	0.81	0.50	0.16

Figure 8 (a), (b), (c), (d), (e), and (f) show the confusion matrix [26] of the VGG16_ML classifiers model on the FaceForensics++_Deepfake dataset. As noticed, the misclassifications in different classifiers are disparate. The KNN classifier in Figure 8 (b) has the lowest misclassification, while the Gaussian classifier in Figure 8 (f) has the highest misclassification. Figure 9 (a), (b), (c), (d), (e), and (f) show the confusion matrix of the VGG16_ML classifiers model on the FaceForensics++_Face2face dataset. The efficiency of the classifiers on this dataset was lower than the previous one. In Figure 9(a), the SVM classifier shows it has the best results, while the NB classifier in Figure 9(f) has the highest misclassification.

Figure 10(a), (b), (c), (d), (e), and (f) show the confusion matrix of the VGG16_ML classifiers model on the FaceForensics++_Faceswap dataset. As noticed, the KNN classifier in Figure 10 (b) has the lowest misclassification, while the NB classifier in Figure 10 (f) has the highest misclassification. Figure 11(a), (b), (c), (d), (e), and (f) show the confusion matrix of the VGG16_ML classifiers model on the FaceForensics++_Neuraltexture dataset. The SVM classifier in Figure 11(a) has the best result, while the misclassifications are highest for most classifiers in this type of dataset.

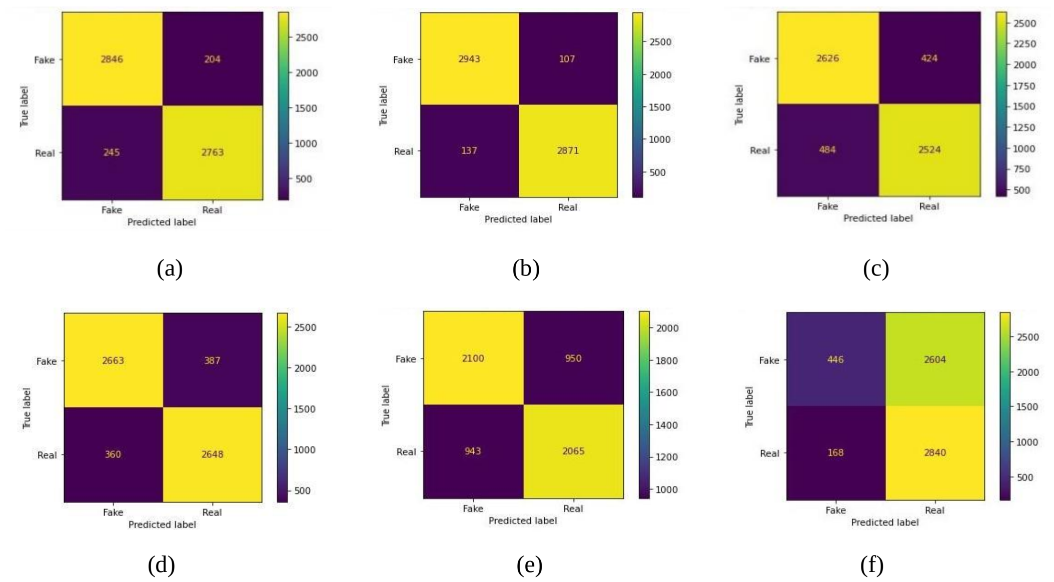


Figure 8: The confusion matrix of VGG16_ML classifiers on the FaceForensics++_Deepfake dataset, (a) CM of SVM classifier, (b) CM of KNN classifier, (c) CM of RF classifier, (d) CM of LR classifier, (e) CM of DT classifier, and (f) CM of NB classifier

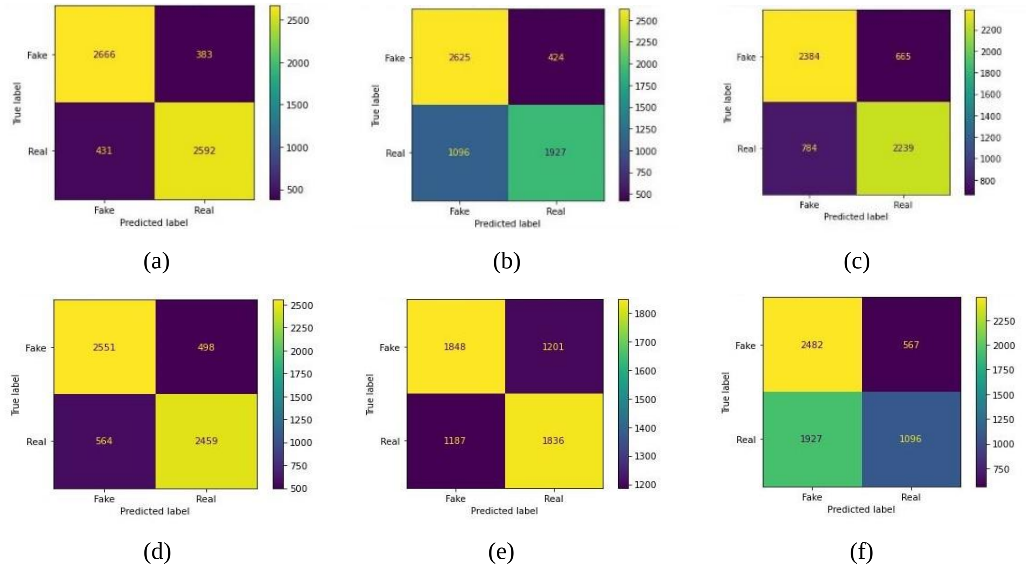
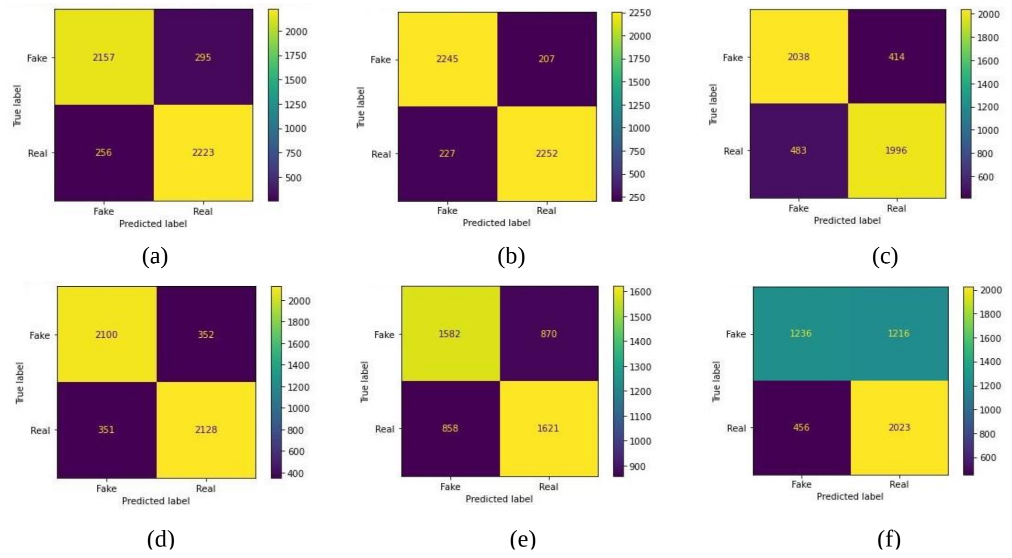


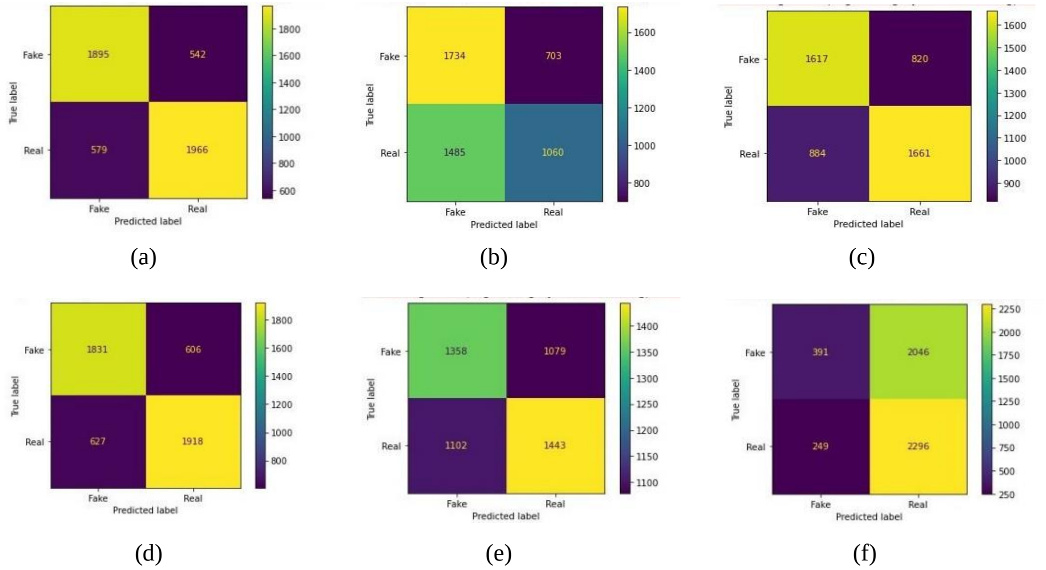
Figure 9: The confusion matrix of VGG16_ML classifiers on FaceForensics++_Face2face dataset, (a) CM of SVM classifier, (b) CM of KNN classifier, (c) CM of RF classifier, (d) CM of LR classifier, (e) CM of DT classifier, and (f) CM of NB classifier



(a) (b) (c)

(d) (e) (f)

Figure 10: The confusion matrix of VGG16_ML classifiers on the FaceForensics++_Faceswap dataset, (a) CM of SVM classifier, (b) CM of KNN classifier, (c) CM of RF classifier, (d) CM of LR classifier, (e) CM of DT classifier, and (f) CM of NB classifier



(a) (b) (c)

(d) (e) (f)

Figure 11: The confusion matrix of VGG16_ML classifiers on the FaceForensics++_Neuraltexture dataset, (a) CM of SVM classifier, (b) CM of KNN classifier, (c) CM of RF classifier, (d) CM of LR classifier, (e) CM of DT classifier, and (f) CM of NB classifier

Table 3 shows the best results in both proposed models. The results, in general, proved that this CNN model has better results, which means it extracted richer features than the features from the VGG16. But on the other side, it is important to notice that the VGG16-ML model can improve the performance of the FaceForensics++_Neuraltexture dataset with an SVM classifier. Both of the proposed models proved that ML classifiers have good performance in detecting problems created by deep learning.

Table 1: The highest accuracy in both proposed models

	CNN-ML classifiers proposed model						**VGG16-ML classifiers proposed model**					
	SVM	KNN	RF	LR	DT	NB	SVM	KNN	RF	LR	DT	NB
FaceForensics++_ Deepfake	**0.96**	**0.96**	**0.96**	**0.96**	0.94	**0.96**	0.92	**0.95**	0.85	0.87	0.68	0.54
FaceForensics++_ Face2face	**0.87**	0.86	**0.87**	**0.87**	0.85	0.86	**0.86**	0.74	0.76	0.82	0.60	0.58
FaceForensics++_ Faceswap	**0.90**	**0.90**	**0.90**	**0.90**	0.87	0.89	0.88	**0.91**	0.81	0.85	0.65	0.66
FaceForensics++_ Neuraltexture	**0.64**	0.61	**0.64**	**0.64**	0.59	0.63	**0.77**	0.56	0.65	0.75	0.56	0.53

5. CONCLUSION

This study has presented two models for detecting deepfake videos: 1) CNN and machine learning classifiers; and 2) transfer learning (VGG16) and machine learning classifiers. The FaceForensics++ dataset is used in both proposed models. Both proposed models achieved good accuracy on the three deepfake techniques (faceswap, face2face, and deepfake), while the neural texture technique was hard to detect. The obtained results proved the efficiency of both proposed models in detecting deepfake videos and how the combination of deep learning and machine learning can build an efficient model. In addition, the use of different ML classifiers shows the power of the ML algorithm in resolving problems created by deep learning techniques. It is also important to notice that the CNN in the first model, for which its parameters were chosen, had better extraction features than VGG16, which is a pre-trained model, in the second model.

Furthermore, we concluded that the proposed models in this paper can assist in detecting deepfake videos by using them on online websites or applications dedicated to recognizing and detecting deepfake videos. On the other side, there are some limitations to the proposed models. First, the limitation of memory in the Kaggle environment, which specified the input size of the CNN in the first model, means that it should expect an increase in results with an increase in the input size of the CNN because the number of trainable parameters and the number of extractor features depend on it. And secondly, in the pre-processing stage, it is hard to detect blurred, incomplete, and far-fetched faces.

For future work, it is intended to improve the detection accuracy of the neural texture technique by using different classification techniques, such as ANN, or by using a different way of extracting and selecting features. We also like to evaluate the proposed models on different video datasets, like DFDC (Deepfake Detection Challenge), and make comparisons between the results

6. Disclosure and conflict of interest

The authors declare that they have no conflict of interest.

References

[1] L. Verdoliva, "Media forensics and deepfakes: an overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910-932, 2020. Doi: 10.1109/JSTSP.2020.3002101

[2] L. Deng, H. Suo and D. Li, "Deepfake video detection based on EfficientNet-V2 network," *Computational Intelligence and Neuroscience*, vol. 2022, p. 3441549, 2022. Doi: https://doi.org/10.1155/2022/3441549

[3] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini and S. Tubaro, "Video face manipulation detection through ensemble of CNNs," in *2020 25th international conference on* *pattern recognition (ICPR)*, pp.5012-5019, Milan, Italy, Jan. 2021. Doi: 10.1109/ICPR48806.2021.9412711

[4] P. Yu, Z. Xia, J. Fei and Y. Lu, "A survey on deepfake video detection," *Iet Biometrics*, vol. 10, no. 6, pp. 607-624, 2021. Doi: https://doi.org/10.1049/bme2.12031

[5] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, *et al.* , "Deep learning for deepfakes creation and detection: A survey," *Computer Vision and Image Understanding*, vol. 223, p. 103525, 2022. Doi: https://doi.org/10.1016/j.cviu.2022.103525

[6] M. S. Rana, B. Murali and A. H. Sung, "Deepfake Detection Using Machine Learning Algorithms," in *2021 10th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 458-463, Niigata, Japan, 2021. Doi: 10.1109/IIAI-AAI53430.2021.00079

[7] B. U. Mahmud and A. Sharmin, "Deep insights of deepfake technology: A review," *Dhaka University Journal of Applied Science and Engineering*, vol. 5, no. 1&2, pp. 13-23, 2020. Doi: https://doi.org/10.48550/arXiv.2105.00192

[8] D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horváth, *et al.* , "Deepfakes detection with automatic face weighting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops(CVPRW)*, pp. 668-669, Seattle, USA, Jun. 2020. Doi: 10.1109/CVPRW50498.2020.00342

[9] A. A. Abu-Ein, O. M. Al-Hazaimeh, A. M. Dawood and A. I. Swidan, "Analysis of the current state of deepfake techniques-creation and detection methods," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 3, pp. 1659-1667, 2022. Doi: 10.11591/ijeecs.v28.i3.pp1659-1667

[10] A. Mitra, S. P. Mohanty, P. Corcoran and E. Kougianos, "A novel machine learning based method for deepfake video detection in social media," in *2020 IEEE International Symposium on Smart Electronic Systems (iSES)(Formerly iNiS)*, pp. 91-96, Chennai, India, 2020. Doi: 10.1109/iSES50453.2020.00031

[11] G. Lacerda and R. Vasconcelos. "A Machine Learning Approach for DeepFake Detection", in *Anais Estendidos do XXXV Conference on Graphics, Patterns and Images*, pp. 110-113, Natal/RN, Brasil, 2022. Doi: https://doi.org/10.5753/sibgrapi.est.2022.23272

[12] M. Masood, M. Nawaz, A. Javed, T. Nazir, A. Mehmood and R. Mahum, "Classification of Deepfake videos using pre-trained convolutional neural networks," in *2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, pp. 1-6, Islamabad, Pakistan, May. 2021. Doi: 10.1109/ICoDT252288.2021.9441519

[13] M. Nawaz, M. Masood, A. Javed and T. Nazir, "FaceSwap based DeepFakes Detection," *International Arab Journal of Information Technology*, vol. 19, no. 6, pp. 891-896, 2022. Doi: https://doi.org/10.34028/iajit/19/6/6

[14] L. Pryor, R. Dave, J. Mallet and M. Vanamala, "Deepfake Detection Analyzing Hybrid Dataset Utilizing CNN and SVM," *In Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, pp. 7-11,Malaysia, Apr. 2023. Doi: https://doi.org/10.1145/3596947.3596954.

[15] A. Raza, K. Munir and M. Almutairi, "A Novel Deep Learning Approach for Deepfake Image Detection," *Applied Sciences*, vol. 12, no. 19, p. 9820, 2022. Doi: https://doi.org/10.3390/app12199820

[16] Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, “FaceForensics++: Learning to Detect Manipulated Facial Images,” in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea(south), Oct. 2019. Doi: 10.1109/ICCV.2019.00009

[17] E. Johansson, "Detecting deepfakes and forged videos using deep learning," *Master's Theses in Mathematical Sciences*, LUND University, Lund, Sweden, 2020. [Online]. Available: https://lup.lub.lu.se/student-papers/record/9019746/file/9019761.pdf

[18] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt and M. Nießner, "Face2face: Real-time face capture and reenactment of RGB videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2387-2395, Las Vegas, USA, Jun. 2016. Doi: 10.1109/CVPR.2016.262

[19] A. M. Almars, "Deepfakes detection techniques using deep learning: a survey," *Journal of Computer and Communications*, vol. 9, no. 5, pp. 20-35, 2021. Doi: 10.4236/jcc.2021.95003

[20] J. Kang, Z. Ullah and J. Gwak, "MRI-based brain tumor classification using ensemble of deep features and machine learning classifiers," *Sensors*, vol. 21, no. 6, p. 2222, 2021. Doi: https://doi.org/10.3390/s21062222

[21] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381-386, 2020. Doi: 10.21275/ART20203995

[22] P. Jain, S. C. Coogan, S. G. Subramanian, M. Crowley, S. Taylor and M. D. Flannigan, "A review of machine learning applications in wildfire science and management," *Environmental Reviews*, vol. 28, no. 4, pp. 478-505, 2020. Doi: https://doi.org/10.1139/er-2020-0019

[23] N. Akhtar, M. Saddique, K. Asghar, U. I. Bajwa, M. Hussain and Z. Habib, "Digital video tampering detection and localization: review, representations, challenges and algorithm," *Mathematics*, vol. 10, no. 2, p. 168, 2022. Doi: https://doi.org/10.3390/math10020168

[24] M. M. Islam, M. B. Hossain, M. N. Akhtar, M. A. Moni and K. F. Hasan, "CNN based on transfer learning models using data augmentation and transformation for detection of concrete crack," *Algorithms*, vol. 15, no. 8, p. 287, 2022. Doi: https://doi.org/10.3390/a15080287

[25] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, et al. , "On evaluation metrics for medical applications of artificial intelligence," Scientific Reports, vol. 12, no. 1, p. 5979, 2022. Doi: 10.1038/s41598-022-09954-8

[26] M. Heydarian, T. E. Doyle and R. Samavi, "MLCM: Multi-label confusion matrix," *IEEE Access*, vol. 10, pp. 19083-19095, 2022. Doi: 10.1109/ACCESS.2022.3151048