IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 January 2023, accepted 15 February 2023, date of publication 20 February 2023, date of current version 23 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3246661

RESEARCH ARTICLE

DFFMD: A Deepfake Face Mask Dataset for Infectious Disease Era With Deepfake Detection Algorithms

NORAH M. ALNAIM[1], (Member, IEEE), ZAYNAB M. ALMUTAIRI[2], MANAL S. ALSUWAT[3], HANA H. ALALAWI[3], ALJOWHRA ALSHOBAILI[4], AND FAYADH S. ALENEZI[5]

[1]Computer Science Department, College of Sciences and Humanities in Jubail, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia
[2]Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia
[3]Computer Science Department, College of Computer Science and Information System, Umm Al-Qura University, Mecca 21955, Saudi Arabia
[4]Computer Science Department, College of Computer, Qassim University, Buraydah 52571, Saudi Arabia
[5]Department of Electrical Engineering, College of Engineering, Jouf University, Sakaka 72388, Saudi Arabia

Corresponding authors: Norah M. Alnaim (Nmalnaim@iau.edu.sa) and Fayadh S. Alenezi (Fshenezi@ju.edu.sa)

The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number 223202.

ABSTRACT Deepfake is a technology that creates fake images and videos with replaced or synthesized faces. Deepfakes are becoming a concerning social phenomenon, as they can be maliciously used to generate false political news, disseminate dangerous information, falsify electronic evidence, and commit digital harassment and fraud. The ease and accuracy of creating Deepfakes have been bolstered by the popularity of wearing face masks since the beginning of the infectious disease outbreak (2020). Because these masks obstruct defining facial features, fake videos are now even more challenging to identify, increasing the necessity for advanced Deepfake detection technology. The research also creates a real/fake video dataset with face masks because the field lacks the dataset required for detection-model training. The proposed research proposes a Deepfake Face Mask Dataset (DFFMD) based on a novel Inception-ResNet-v2 with preprocessing stages, feature-based, residual connection, and batch normalization. The combination of preprocessing stages, feature-based, residual connection, and batch normalization increases the detection accuracy of deepfake videos in the presence of facemasks, unlike the traditional methods. The study's results compared with existing state-of-the-art methods detect face-mask-Deepfakes with 99.81% accuracy compared to the traditional InceptionResNetV2 and VGG19, whose accuracy is 77.48%, and 99.25%, respectively. Future work should evaluate the accuracy of developing a subsequent experimental work for increased detection of deepfake with facemasks.

INDEX TERMS Deepfake, deep learning, CNN, generation, detection, fake videos, neural network, mask, face mask.

I. INTRODUCTION

The recent growth of technology in computer-generated editing programs has made synthesizing and modifying media content easier than ever. The potential for misinformation spread has exploded, especially with the phenomenon known as Deepfake. Deepfake is a technology that uses deep learning to create fake videos, alter existing videos, or even synthesize the speech of someone's voice. This makes it a dangerous tool in malicious applications of spreading fake news and disseminating false or dangerous information. Thus, detecting Deepfakes using machine learning techniques has been a subject of the research community since 2017. Studies have tackled the challenge from various angles, from analyzing faces to focus on specific regions like eyes and lip movements to creating new deep-learning architectures. Today, many Deepfake tools are free, open-source, and have many learning resources. The most available are Faceswap [1], Faceswap-GAN [2], DeepFaceLab [3], and DFaker [4]. These tools swap the source person's face with the target face to create a new video with

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Kashif Bashir.

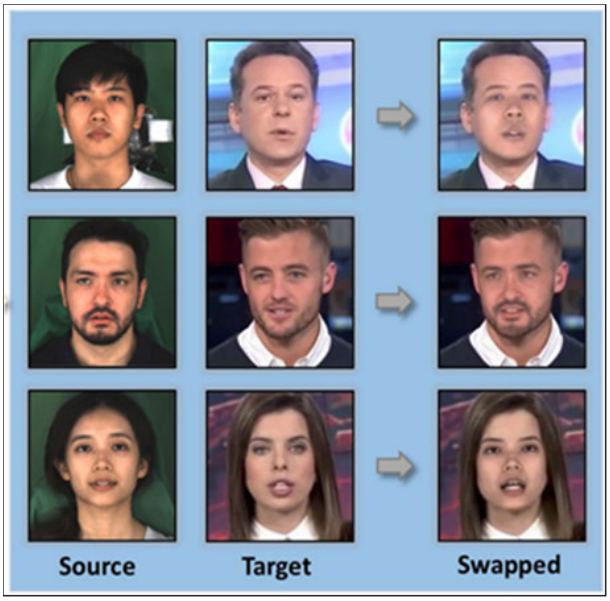

FIGURE 1. Deepfake example [6].

the same action as the source but with the target person's face, as shown in Fig. 1. The products of these convenient services are notably difficult for humans to distinguish between real and fake [5].

Therefore, developing digital forensic machine learning models and algorithms to detect fakes is more critical for digital security than ever. The two main techniques of Deepfake technology are generation and detection, both of which utilize deep learning. Many models in literature can generate fake videos.

For instance, Choi et al. [7] proposed a Star-GAN-based deepfake method using a generative adversarial network. The architecture of the method was composed of an encoder, decoder, and discriminator. The encoder consisted of $8 \times 8 \times 512$ output and $64 \times 64 \times 3$ input, and a series of output tensors followed by a self-attention block, a 2D convolution, and up-scaling blocks where activation of leaky ReLU occurs. The main shortcoming of the approach was overreliance swapping on the mouth, nose, and eyes. The look-alike targets were also obtained by averaging features map input, which generated some artefacts making the blending imperfect. This also leads to blurriness in the skin, appearing unnatural. Karras et al. [8] proposed a ProGAN-based method with a generative adversarial network. The generator and discriminator were designed to grow progressively. This speeds the training up and significantly stabilizes, allowing the production of an image with unprecedented quality. However, the poor image micro-structure compared to the results by other state-of-the-art techniques make the method seek some improvement in future research. Karras et al. [9] proposed a styleGAN based on style transfer literature. The proposed architecture yields automated learning, reducing image quality variation, using exponential in the architecture to average pixel values from logarithmic to linear for easy identification results in images with poor features. This is due to the exponential decay of the pixels resulting from the loss of some pixels. Siarohin et al. [10] proposed a first-order motion model for image animation where an object in a source image is not animated based on the motion of a driving video. However, these come with a price; that is, the transfer only allows global object geometry; thus, fine image details are lost.

The existing algorithms detecting Deepfakes are processed based on handcrafted features or CNN. However, since the Covid-19 pandemic popularized video conferencing applications, attackers have targeted them using Deepfake models to construct fake virtual identities in online video conferences [11]. For instance, Uçan et al. [11], and Graber-Mitchell [12] noted that Deepfake had been used during Covid-19 to escape and erase evidence of crime via securing videos. Graber-Mitchell [12], [13] Additionally, the popularization of face masks makes Deepfakes much easier to generate and much less detectable. This is because only the forehead and eyes are exposed, leaving most facial features. It made Deepfake much easier; as such, even armature programmers could generate Deepfake faces.

The accuracy of existing techniques in detecting deepfake videos and images is average 90% [14], [15]. However, this has further reduced to 70% with the introduction of facemasks during the Covid-19 pandemic [16], [17], [18]. This has motivated criminals to use facemasks to elude and cover their criminal activities by editing surveillance videos to escape the criminal justice system. The proposed study proposes Deepfake Face Mask Dataset (DFFMD) based on a novel Inception-ResNet-v2 with preprocessing stages, feature-based, residual connection, and batch normalization. These increase the detection accuracy of deepfake videos in the presence of facemasks, unlike the traditional methods.

II. CONTRIBUTION

We can summarize our contribution as follows:

- Numerous preprocessing enhances the accuracy and detailed information of the images, hence increasing the detection of fake images from videos.
- A rich feature-based Inception-ResNet-v2 with a variation of the Inception V3 model to increase the probability of classifying fake or real images.
- A residual connection in the multiple-sized convolutional filters in the Inception-Resnet block reduces the chances of deterioration of image feature, ensures accurate detection and reduces training time.
- An application of batch normalization in the Inception-ResNet network on top of the traditional layers instead of residual summations to compensate for a loss in dimension during training increases the model's accuracy.

This article is organized as follows: Section III is the literature review. Section IV describes the background of Deepfake generation and detection techniques. Section V presents Materials and methods. Section VI discusses the

implementation evaluation. Section VII presents the full results of the proposed models. Finally, section VIII proposes the conclusion and future work.

III. RELATED WORK

A. DEEPFAKE VIDEO GENERATION IMPLEMENTATION

Korshunov and Marcel [19] presented the first public dataset of deepfake videos generated from the VidTIMT videos dataset. The dataset has high and low-resolution videos trained via VGG and Facenet neural networks. Dolhansky et al. [14] proposed a deepfake detection challenge dataset research. Huang et al. [20] presented Deepfake MNIST+1, a novel human face animation dataset developed by a SOTA picture animation generator. Their dataset produced 10,000 face animation videos in ten distinct movements that can fool the latest aliveness detectors. Their study also included a baseline detection approach and a thorough procedure examination.

khalid et al. [21] introduced FakeAVCeleb, a unique audio-video deepfake dataset that includes not only deepfake videos (V_{only}) but also produces related lip-synced fake audios (A_{only}). They created the dataset using the most common deepfake generation techniques. The study proposed CelebDF1, a new large-scale advanced deepfake video dataset with 5, 639 high-quality Deepfake videos of celebrities made using an enhanced synthesis technique [18]. To highlight the increased difficulty of deepfake detection offered by Celeb-DF, they thoroughly reviewed deepfake detection algorithms and datasets. They used different methods and trained on their suggested dataset [18].

Hu et al. [22] proposed a two-stream strategy by evaluating compressed deepfake videos' frame and temporal levels. Because video compression adds a lot of redundant data to frames, their suggested frame-level stream pruned the network progressively to prevent the framework from trying to fit the compression noise. Table 1 summarises the related studies reviewed in this research.

B. DEEPFAKE VIDEO DETECTION IMPLEMENTATION

Pishori et al. [23] covered three deepfake detection strategies and developed algorithms for them during the deepfake detection challenge based on convolutional Long short-term memory (LSTM), eye blink detection, and grayscale histograms. The study introduced a novel deep learning-based strategy for distinguishing AI-generated fake videos from real videos [24]. The study concentrated on deepfake video facial expression detection since most algorithms were already accurate in manufacturing realistic static human faces [25]. They mixed several kinds of vision transformers with a convolutional EfficientNet employed as a feature extractor, achieving results equivalent to some relatively recent vision transformer-based approaches. Wodajo and Atnafu [26] applied a convolutional vision transformer to identify deepfake data. Their proposed convolutional vision transformer consisted of two parts: the CNN and the vision transformer (ViT) [26].

Amerini et al. [27] provided a novel forensic strategy for distinguishing between false and authentic video sequences, unlike existing state-of-the-art techniques that rely on single video frames. They suggested the use of optical flow fields to utilize probable interframe dissimilarities. They proposed a technique based on CNNs and RNNs for extracting visual and temporal characteristics from faces in videos to identify manipulations correctly. The research study proposed EffYnet as a unique architecture for identifying visual changes between altered and unaltered environments. They used an EfficientNet encoder and a U-Net with a classification component in the architecture to create a model capable of classifying and segmenting deepfake movies.

Singh et al. [28] introduced a method to identify movies distorted by deepfake efficiently and comprehensively. The authors developed an architecture that used lower-level characteristics in areas of interest and disparities over many frames. They conducted many experiments on the deep fake detection challenge dataset of 470 GB and discovered that their suggested method achieved a 97.6% test accuracy score. Xu et al. [29] also designed a novel approach for detecting deepfake movies. They built texture features using the grey level co-occurrence matrix, gradient domain, wavelet transform, and standard deviation of the face area. Table 2 summarizes studies on deepfake video detection techniques.

IV. BACKGROUND

A. TECHNICAL APPROACHES IN DEEPFAKES

Previously, generating high-quality fake videos was difficult and easy to detect based on artifacts such as eye blinking, facial expression, head position, or mouth movements. Nowadays, the situation has changed, as video editing applications and tools have improved to become more convenient with powerful editing capabilities for creating fake videos. Deepfake technology combines different methods and neural networks to produce more efficient and accurate Deepfake models. The following sections provide the technical background of Deepfake video generation and detection.

1) DEEPFAKE GENERATION TECHNIQUES

The deepfake phenomenon was invented in 2017 by a Reddit user who made manipulated videos by swapping one person's face with another's using deep learning and computer vision techniques [32]. However, deep learning methods have grown increasingly complex for generating highly-realistic synthetic content. With the rapid growth in computer vision, it is also increasingly difficult for humans to differentiate between real and fake videos. In 2018 a new tool was published, known as Deepfake, built based on GANs. The GAN was proposed by Goodfellow et al. [33]. It consists of two networks, (i) the Generator Network (G) and (ii) the Discriminator Network (D). The G network takes the image latent variable as an input. Then it converts it into a fake sample to trick the

TABLE 1. The state-of-the-art Deepfake video generation implementation.

Ref. No	Methodology	Description	Results	Pros	Cons
[19]	Created the Deep-fakes using open-source software built on Generative Adversarial Networks (GANs).	Provided the first publicly accessible collection of Deepfake videos created from VidTIMIT database videos.	VGG & Facenet neural networks were sensitive to Deepfake videos, with false acceptance rates of 85.62% & 95.00%, respectively.	Proved that training and mixing settings significantly influence the quality of the resulting videos.	Videos quality is not good.
[14]	GAN-based face swapping methods	Public dataset of the Deepfake detection challenge (DFDC) created from GAN-face swapping techniques.	-	-	-
[20]	Resnet50, Resnet152, XceptionNet, and MesoNet	Presented Deep-Fake MNIST+1, a novel human face animation dataset developed by a SOTA picture animation generator.	Resnet152 achieved the highest accuracy of 90.82% on Raw, 92.11% on LC, and 88.32% on HC quality	Their dataset produced 10,000 face animation videos in ten distinct movements that can fool the latest liveness detectors	Only animated videos are available
[21]	Introduced FakeAVCeleb, a unique Audio-Video Deepfake dataset that includes not only Deepfake videos (V_{only}) but also related lip-synced fake audios (A_{only}).	EfficientNet-B0 exhibited the most stable average accuracy of 95%	Created this dataset using the most common deep-fake generating techniques.	Data size is massive due to video an audio combination.	
[18]	Meso4, MesoInception4, HeadPose, FWA, VA-MLP, VA-LogReg, Xception-raw, Xcep-tion-c23, Xception-c40, and Capsule.	Proposed CelebDF1, a new large -scale difficult Deepfake video dataset with 5,639 high-quality Deepfake videos of celebrities made using an enhanced synthesis technique.	Claimed that Meso4 provided 54.8%, MesoInception4 expressed 53.6%, HeadPose claimed 54.6%, FWA showed 56.9%, VA-MLP achieved 55.0%, VA-LogReg expressed 55.1%, Xception-raw showed 48.2%, Xception-c23 claimed 65.3%, Xception-c40 reached 65.5%, multi-task expressed 54.3%, Capsule showed 57.5%, and DSP-FWA claimed 64.6% AUC.	Provided rigorous testing of their dataset and almost each of them achieved significant results in training and validation.	-
[22]	A two-stream DNN	Proposed a two-stream strategy in this study by evaluating compressed Deepfake videos' frame- and temporality- level. To retrieve temporal correlation characteristics, they used a temporality level stream.	Their suggested method showed an accuracy of 94.64% on Deepfakes, Their suggested method showed an accuracy of 94.64% on Deepfakes, 85.27% on FaceSwap, 86.48% on Face2Face, 80.05% on NeuralTextures, and 80.74 on Celeb-DF datasets.	As video compression adds a lot of redundant data to frames, their suggested idea prevents the framework from trying to fit the compression noise.	Method is too complicated with large training time.

discriminator by generating fake data like the real data. On the other hand, the discriminator network takes inputs and tries to differentiate between real and fake data. Since then, many video generation models have been proposed upon the fundamental idea of GAN. Motion and content decomposed GAN (MoCoGAN) is another extension of GAN - it is a recurrent network trained to produce videos based on categorized labels, images, or noises. MoCoGAN was trained to separate the content from the motion of each video, for instance, generating a new video of a person performing facial expressions with different identities. MoCoGAN is a novel generation technique in which each video frame is created from a random vector with two parts for the content and the motion [34].

MoCoGAN is comprised of four sub-networks: the recurrent neural network RM, the image generator G_1, the image discriminator D_1, and the video discriminator D_V. G_1 generates a video clip by mapping a sequence of vectors into a sequence of images. D_1 and D_V play a significant role in providing the input to G_1 and R_M. The image discriminator D_1 evaluates G_1 based on individual images. This trains the discriminator to detect whether a frame is taken from a real video clip or fake.

Vondrick et al. [35] proposed a VGAN-based method that uses the same architecture as GAN. Still, the CNN-image generator and discriminator are replaced with a spatiotemporal CNN-based video generator and discriminator. The generator network takes a low-dimensional latent random vector as an input to produce high-dimensional output in 32 frames. The fake videos are generated based on the unlabeled videos dataset. On the other hand, the video discriminator classifies the real videos from the synthesized videos. In addition, the discriminator network recognizes the real motions between the video frames and the visual behaviour by adding a five-layer Spatio-temporal convolutional network. These two networks train against each other using the min-max game, where the generator attempts to mislead the discriminator maximally. In contrast, the discriminator aims to determine which samples are fake, as,

$$\min_{wG}\max_{wD}\left(E_{x_{p_x(x)}}\Pi_A + E_{z_{p_z(z)}}\Pi_B\right) \tag{1}$$

where $\Pi_A = [logD(x : W_D)]$,
$\Pi_B = [log\ 1-D\,(G(z; w_G) : w_D)]$, z is the latent variables that is often sampled from a normal distribution, while $x_{p_x(x)}$ are samples for the data distribution. Where G is the generator and D is the discriminator networks.

Saito et al. [36] proposed a new video generation framework called the temporal generative adversarial network (TGAN). The proposed model can produce new videos by learning the representations from several unlabeled videos. TGAN is a novel model trained based on the Wasserstein GAN (WGAN) framework. It consists of two sub-networks: the temporal generator and the image generator. The temporal

TABLE 2. The state-of-the-art deepfake video detection implementation.

Ref. No	Methodology	Description	Result	Pros	Cons
[23]	CNN, RNN with combination of LSTM, eye blink and grayscale histograms.	They covered three strategies and developed algorithms while engaging in the Deepfake Detection Challenge : convolutional LSTM, eye blink detection, and grayscale histograms.	They claimed that CNN+RNN showed the highest validation accuracy of 82.81%, Eye Blink Detection gave 81.67%, and grayscale histogram ex-pressed 81.32% accuracy.	Analyzed existing understanding of Deepfake movies, a more severe type of altered media, and found the grayscale histogram methodology more relevant than others.	High false positives
[24]	Convolutional neural networks	Introduced a novel deep learning-based strategy for distinguishing AI-generated fake videos from actual ones.	Their proposed method showed an AUC of 97.4% on the UADFV dataset, 99.9% on the low-quality DeepfakeTIMIT dataset, and 93.2% on the high-quality DeepfakeTIMIT dataset.	The existing Deepfake algorithm would only create pictures with restricted resolutions, which must then be twisted to resemble the actual facial features in the source video. They were able to made pictures with any resolution.	Complex method
[25]	Transformers with EfficientNet B0	Specifically mixed several kinds of Vision Transformers with EfficientNet B0 employed as a feature extractor, getting results equivalent to some relatively recent Vision Transformer based approaches	On the Deep-Fake Detection Challenge, their proposed top model scored an AUC of 0.951 and an F1 score of 88.0%.	Focused on Deepfake facial expression detection.	Low Accuracy
[26]	Convolutional Neural Network and the Vision Transformer (ViT).	Suggested a Convolutional Vision Transformer to identify Deepfakes data.	Trained their model on the DFDC attained 91.5% accuracy, 0.91 AUC, and a 0.32 loss value.	Utilized attention mechanism to classify the learned features which have low false positives.	High false positives
[27]	CNN with Optical flow	Suggested the use of optical flow fields to utilize probable interframe dissimilarities. Then they utilized a hint-like feature for CNN classifiers to learn.	Their preliminary findings from the FaceForensics++ dataset for the Facet2Face manipulation showed an accuracy of 81.61% by using the VGG16 network and 75.46% by using the ResNet50 network.	Noncomplex method, with faster training time.	High false positives
[30]	CNN and RNN	Proposed a technique based on convolutional neural networks and recurrent neural networks for extracting visual and temporal characteristics	Trained their suggested method on the DFDC dataset. Their presented method claimed accuracy of 92.61% on validation and 91.88% in the test phase.	Fast training time with low false positives	Low accuracy
[31]	EfficientNet + ResNet + U-Net	Proposed architecture to identify visual changes between changed and unmodified environments. Used an EfficientNet encoder & a U-Net with a classification component.	Executed trials on the Deepfake Detection Challenge dataset & find that their proposed models showed an accuracy of 98.7%.	Also employed ResNet 3D to identify spatiotemporal irregularities.	Complex method
[28]	CNN	Identified distorted movies made using Deepfake efficiently and com-prehensively	Conducted many experiments on the deep Fake Detection Challenge dataset of 470 GB and discovered that the suggested method produced a 97.6% test accuracy score	The authors developed an architecture that used lower-level characteristics in areas of interest and disparities over many frames.	High False positives
[29]	Traditional Methods	Built texture features using the gray level co-occurrence matrix, gradient domain, wavelet transform, and standard deviation of the face area. Classification using SVM.	Trained their model on the FaceForensics++ dataset and claimed that their suggested method showed an accuracy of 86.3% by using C23 and 91.2% by using C40.	-	Low accuracy

generator takes two latent variables and produces a set of latent variables as an output. In contrast, the image generator can generate videos by transforming the latent variables into a series of images (frames).

$$\min_{\theta G_0, \theta G_1} \max_{\theta_D} \left(E_{([x^1, \ldots x^T]\ P_{data})} \Pi_C - E_{(z_0\ PG_0)} \Pi_D \right) \quad (2)$$

where $\Pi_C = \left[D\left(\left[x^1, \ldots x^T \right] \right) \right]$
$\Pi_D = \left[D\left(G_1\left(z_0, z_1^1, \right) \ldots G_1\left(z_0, z_1^T \right) \right) \right]$

Eq. (2) suggests that z_0 is the original latent variable, where z_1 changes with time. x^T represents the t^{th} that has been generated initially by the generator G_0 (z_0), and the p_{data} indicates the empirical data distribution. On the other hand, $N8_G$ and $N8_D$ are the generator and discriminator parameters, respectively.

2) DEEPFAKE DETECTION TECHNIQUES

Detecting synthesized and fake media content is a crucial challenge. One of the best ways to tackle it is by using machine learning methods and forensic analysis based on artifacts caused in the generation phase. The effectiveness of deep learning techniques has made remarkable results in detecting forged videos. Several methods have been recently applied, such as CNN, LSTM, and RNN [24], [32]. To detect whether the video is synthetically generated vs manipulated vs real, the detection models focus on identifying artifact categories - either spatial or temporal artifacts. Another approach to detecting manipulation is to train deep learning networks (DNN) to distinguish between real and manipulated content through classification or anomaly detection [37]. Such manipulation manifests in entire face

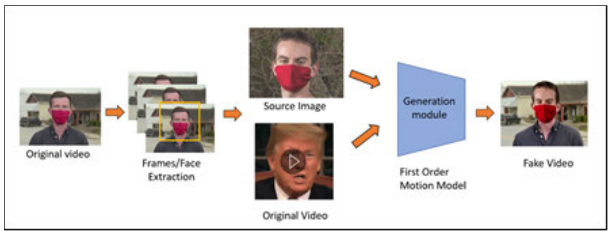

FIGURE 2. Pipeline of generating a fake video.

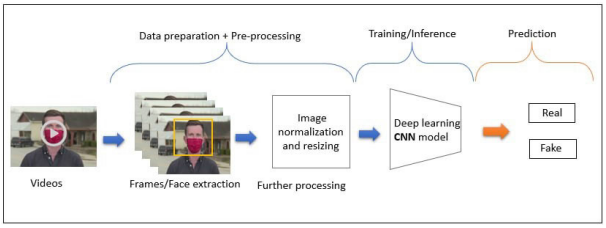

FIGURE 3. Pipeline of the fake video detection.

synthesis, attribute manipulation, identity swap, and expression swap [38].

To discriminate artificially generated faces from real faces, Mccloskey and Albright [39] proposed research based on deep learning networks and transfer learning to analyze the frames by extracting artifacts and detecting the synthesized face. This is composed of the entire face synthesis as a type of manipulation where entirely new faces are generated using a GAN. The network is trained to extract the faces from the target video to make predictions. For instance, the technique uses a linear support vector machine classifier (SVM) and the image's color feature to distinguish between real and GAN-generated images.

Transfer learning has also shown significant performance in prediction and detection tasks. This technique is often employed to solve problems in machine learning of insufficient data for training models. One develops a model for one task with massive data and then reuses and adjusts the trained model for another task [40].

V. MATERIALS AND METHODS

Previously, generating high-quality fake videos was rare and easy to detect based on artifacts such as eye blinking, facial expression, head position, and mouth movements. The situation is different today, as video editing applications and tools have improved to become both more convenient and more powerfully realistic. Deepfake creation manipulates a specific type of artificial intelligence network. Fig. 2 below shows a pipeline for generating fake videos. Due to the variety of existing neural networks, combining different networks produces more efficient and accurate Deepfake models.

Moreover, this section describes the approach for detecting Deepfake videos. Fig. 3 below shows a schematic representation of this method.

The first step is converting the raw video clip into a format that can act as input for deep learning networks. The first preprocessing step is extracting the individual frames from the video. The training process needs these numerous frames to learn accurate and detailed information. The next preprocessing step is to extract the face areas detected in every frame. The final preprocessing step involves several alterations that make the dataset compatible with the deep learning model and enhance its learning capability and performance. These transformations include resizing face images depending on the deep learning model requirements and some data normalization depending on the specific deep learning models.

After preprocessing, the data is ready for training and evaluation. Frames are fed into a convolutional neural network that is trained to predict whether the video is fake or real. In order to compare these detectors, all models were implemented on a GPU processing using the TensorFlow deep learning framework and google colab notebook.

VI. EVALUATING DEEPFAKE DETECTION METHODS

A. DEEPFAKE DETECTION MODELS

This study measures three deep learning models for their effectiveness in detecting Deepfakes with face masks: CNN, two transfer learning models, and Inception-ResNet-v2 and VGG19.

1) TRANSFER LEARNING MODELS

Transfer learning models are sometimes referred to as pre-trained models. In transfer learning, a model such as a CNN is trained on a massive and often broad volume of data, such as ImageNet, for image classification. This highly generalized pre-trained model is then applied to a more specific dataset, retaining the power from the pretraining and the specificity from the second dataset. This research will investigate two pre-trained CNN models: Inception-ResNet-v2 and VGG19.

2) INCEPTION-ResNet-v2

Inception-ResNet-v2, a variation of the Inception V3 model, is a convolutional neural network trained on the ImageNet database. The network contains 164 deep layers and can categorize images into 1000 kinds of objects, including a mouse, keypad, pencil, and numerous animals. Consequently, the network contains rich features for various images. The input size of the network is 299-by-299, and the output is a list of calculated class probabilities. This network was constructed based on a combination of the Residual connection and Inception structure and was more profound than the previous Inception V3. Multiple-sized convolutional filters merge with residual connections in the Inception-Resnet block. Using the residual connections avoids the problem of deterioration found in deep structures and reduces training time [41]. As shown in Fig. 4, Inception-ResNet merges the Inception and Residual network architectures to boost performance over Inception or ResNet alone. In Fig. 4, every

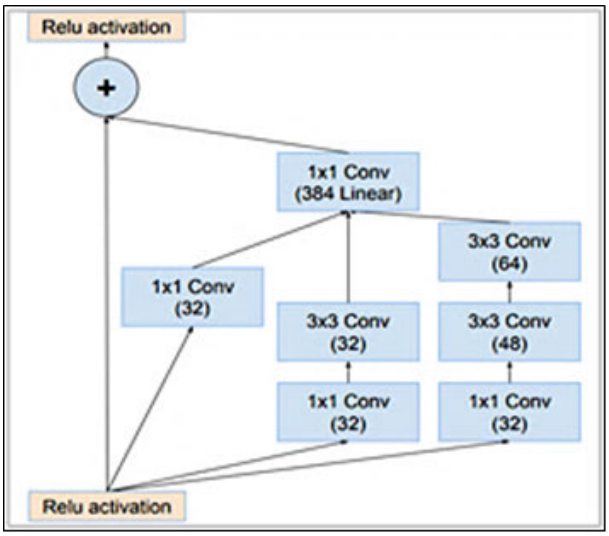

FIGURE 4. Inception-ResNet-v2 Structure [43].

Inception block is followed by layers of filter expansion using 1×1 convolution without activation. This is useful for scaling the filter bank's dimensionality before residual addition so that they are as deep as the inputs. In the Inception-ResNet network, batch normalization is applied just on top of the traditional layers rather than residual summations. For the Inception block, this is essential to compensate for a loss in dimension from the Inception block.

3) VGG19

VGG-19 is a CNN variant of the VGG model that contains 19 deep layers and was trained on the ImageNet database. VGG-19 in Fig. 5 is constructed from 16 convolution layers, 5 2×2 MaxPool layers, 3 fully connected layers, and 1 SoftMax layer (not counting SoftMax as a MaxPool layer). The kernel size is 3*3 with a stride and pad of 1, and the input size is $224 * 224 * 3$. The architecture is stacked convolution, pooling layers, and fully connected ANN. The number of filters increases as the depth of the network increases, and the spatial size of feature maps reduces due to the pooling layer. Every stack of convolutional layers is followed by a rectified linear unit (ReLU) activation function and then a max-pooling operation. In VGG, 3×3 filters were used in all convolutional layers to reduce the number of parameters and keep the structure simple [42].

4) CONVOLUTION NEURAL NETWORKS (CNN)

Convolutional neural networks (CNNs) are deep learning neural networks composed of convolution, pooling, fully connected, and nonlinear layers. This paper's proposed convolutional network was built using the Keras library containing three convolution layers deployed with the kernel size (3,3). ReLU was used as an activation function in all layers, followed by a max-pooling layer with (2,2) pooling size to reduce the size of the large images. The results were flattened before being fed to the fully connected layer with a dropout of 0.2 to avoid overfitting. Finally, a softmax activation layer was used as the output layer. The architecture of the model is shown in Fig. 6.

B. EXPERIMENTAL SETTINGS

When creating the Inception-ResNet-v2 and VGG19 models, the hyperparameters were set as follows: The value of include_top was set to "False", meaning not involving the fully connected layer in the last layer of the network. Next, input_shape: is set to (224, 224, 3) because they include_top False. In addition, we set weights='imagenet' to use the weights of the pre-trained model. The network was attached to convolutional layers with filters = 1024 and kernel_size = (2,2) and padding = 'same', followed by an activation layer with the 'ReLU' function. At the end of the network are the BatchNormalization layer, GlobalAveragePooling2D, and the dense layer, a fully connected layer with two output classes (fake or real) and activation = 'softmax'. Adam was used as an optimizer in this research, with a learning rate = 1e-5, ten epochs, and batch size = 64.

C. DATASET AND PREPROCESSING

The models require a massive dataset for the training process; therefore, approximately 1000 fake and 1000 real videos were collected for this dataset. The total dataset of roughly 2000 videos was divided 80% into training and 20% for testing. The video dataset was preprocessed before training the three selected models for fake video detection. The first step was dividing each video into multiple frames and detecting the face from each video. Then the resultant images were cropped to retain only the face and resized to 128×128 to be suitable for the models' input. Fig. 7 shows a sample from the dataset after preprocessing.

D. EVALUATION METRICS

Accuracy, precision, recall, and F1 Score were used to evaluate the performance of the detection models. These metrics are defined by the following equations:

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + FN + TN + FP} \\ Precision &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \\ F1score &= 2 \times \frac{Precision \times Recall}{Precision + Recall} \end{aligned} \tag{3}$$

where N is the number of classes.

VII. RESULTS

A. GENERATION

In this section, we shed light on our dataset generation results and the limitations faced during the generation process. Many video Deepfake generation models have been developed over the last few years. In 2019, a novel model emerged that would

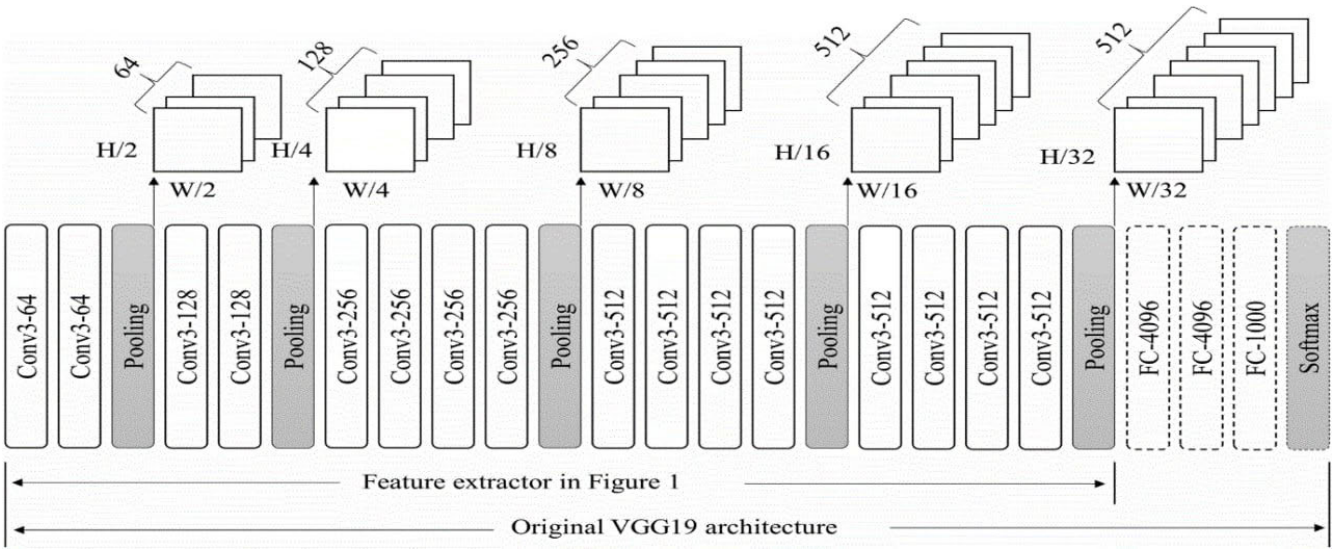

FIGURE 5. VGG-19 Structure [44].

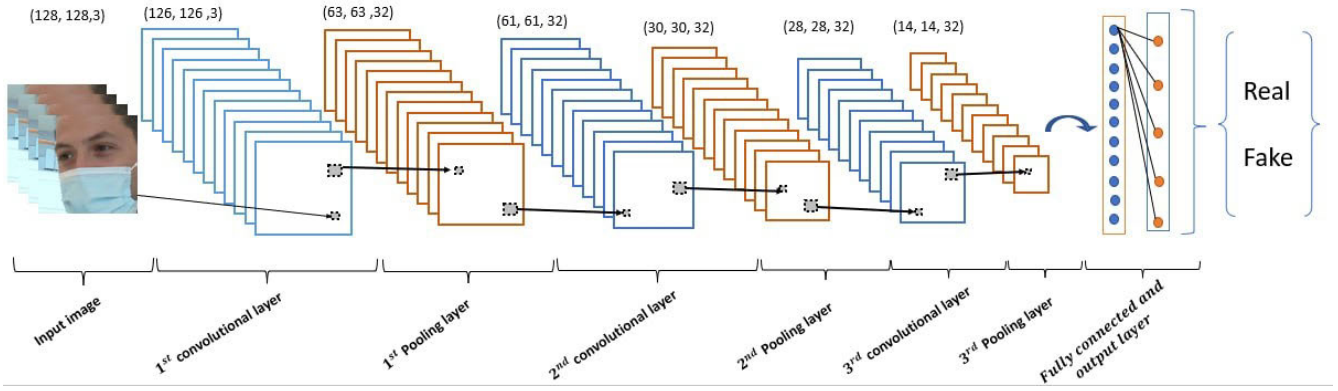

FIGURE 6. The proposed CNN Structure.

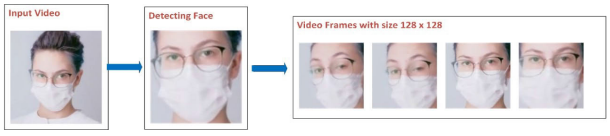

FIGURE 7. Sample of the dataset after preprocessing.

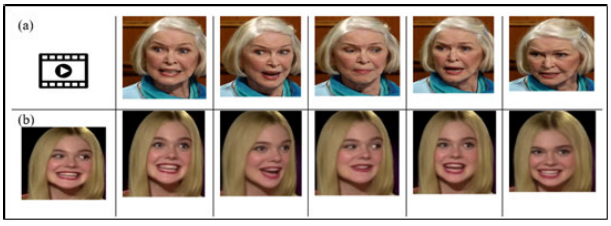

FIGURE 8. An example of images synthesized by the driving video using the FOMM model, (a) is the driving video and (b) is the source image.

FIGURE 9. Samples of generation results, (a) is the driving video and (b) is real video while (c) is the fake video generated from the source image.

change the target image motion based on the driving video, known as FOMM [45]. The main idea behind this model is to animate the source image of the person based on the facial expressions of different people to generate a fake video [10]. The DFFMD was generated using FOMM primarily because it does not rely on pre-trained models that need massive

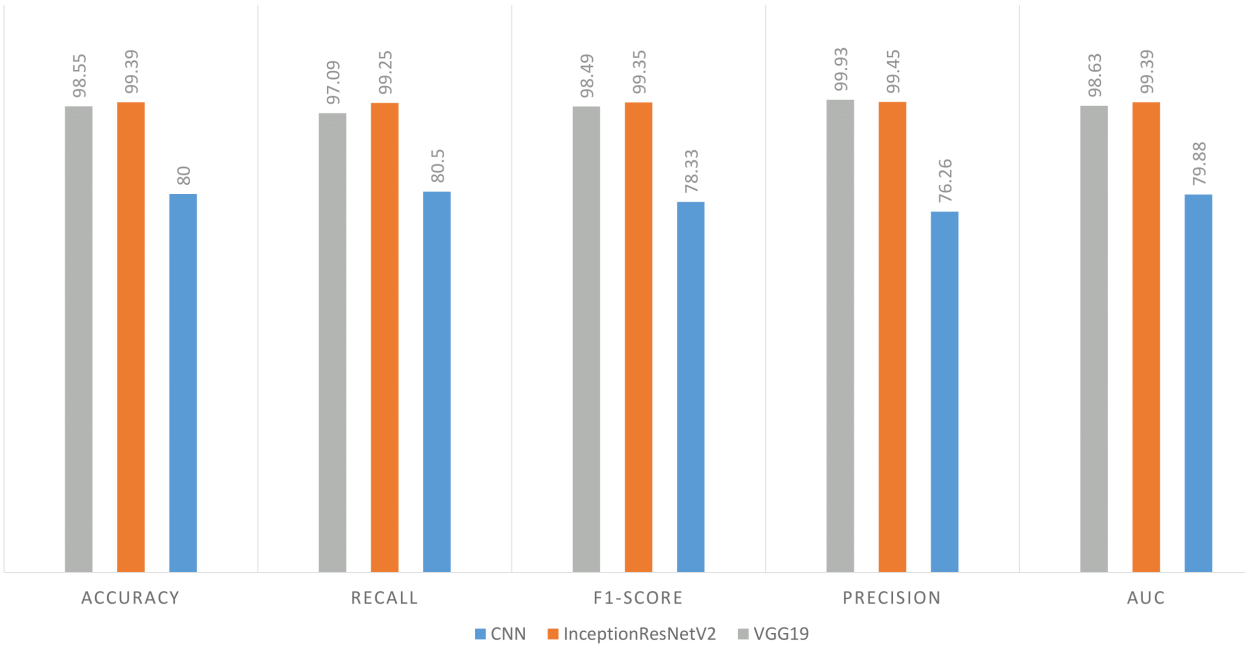

FIGURE 10. Accuracy, Precision, Recall and F1-score for the selected models. VGG19 achieved the highest values for all metrics.

ground-truth data annotations [46], [47], [48]. Fig. 8 shows an image synthesized by driving video using the FOMM model.

DFFMD contains 2000 videos of people with face masks, divided into 1000 fake and 1000 real. Each video costs approximately 5 minutes to generate. The real videos were collected from ten open-source websites, and several YouTube channels listed in Appendix A. Fig. 9 presents samples of our generation results. This dataset contains faked videos of people wearing face masks, women's hijab/Saudi niqab, and men's Ghutrah/Lithamah around their faces. The length of each real sample ranges from 2 to 10 seconds. Although FOMM does not require pre-trained models to generate Deepfake videos, it suffers from medium-quality faces in synthesized videos. Moreover, the real video voices disappeared after the end of the Deepfake generation. We concluded that FOMM might dominate image animation but needs further improvement to generate high-quality faked videos and preserve voice in natural videos.

B. DETECTION

All fake and real videos were trained and tested on VGG19, InceptionResNetV2, and proposed CNN models. As shown in Fig. 10, InceptionResNetV2 achieved the highest accuracy at 99.39% compared with the proposed CNN and VGG19 models. Furthermore, it got the highest values among all other metrics, i.e., precision, recall, and f1-score, as shown in Fig. 10. The InceptionResNetV2 model achieved very close results to the VGG19 model. On the other hand, the proposed CNN achieved lower accuracy than transfer learning models at 77.80%. Because of the dataset generated in this research, the proposed CNN model faces a challenge in classifying fake videos. Fig. 11-13 sketches the training accuracy, validation accuracy, training loss, and validation loss curves for each epoch. In all models, the train and validation losses gradually decrease, which gives the impression that the model learns well, except in the CNN model, which requires many epochs.

TABLE 3. Results of the execution time.

Task	Time
Convert video to face image (Preprocessing)	5h 30 min
CNN	50.4 s
InceptionResNetV2	9 min 44s
VGG19	9 min 24s
proposed Method	9 min 5s

Regarding execution time, Table 3 shows the execution time of the preprocessing and deep learning models used in this paper. The shortest time in the models was CNN at 50 seconds. In contrast, the InceptionResNetV2 model was the longest, but its results were the best compared to the rest of the models.

C. RESULTS COMPARISON

Based on the discussed literature, we provide a quantitative comparison to analyze the State-of-the-Art (SOTA) video Deepfake detection methods with our results based on the used measures and datasets in Table 4. As confirmed by this study [49], the most used metrics in Deepfake are accuracy,

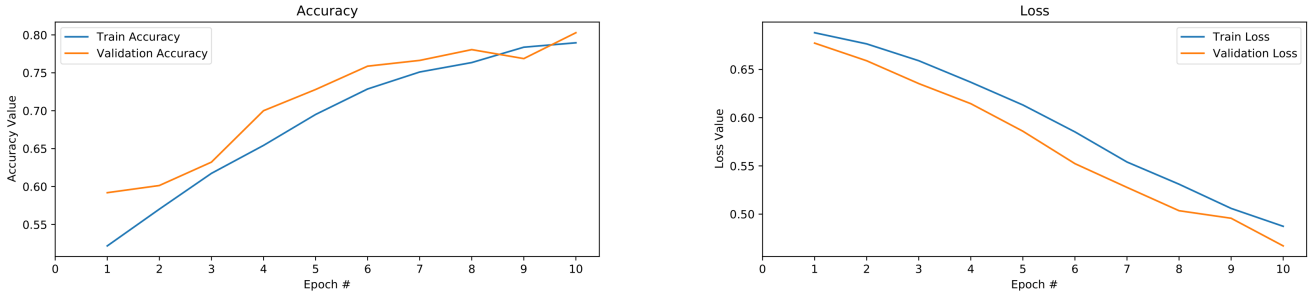

FIGURE 11. The curves of training accuracy, validation accuracy, training loss, and validation loss for CNN model.

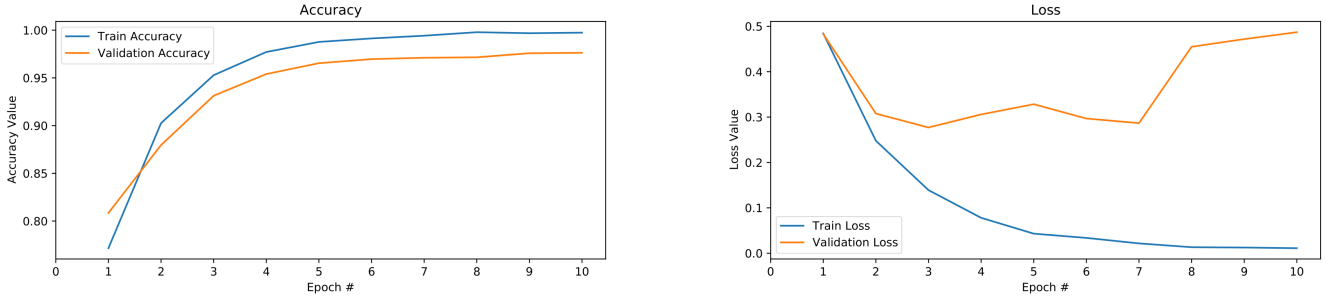

FIGURE 12. The curves of training accuracy, validation accuracy, training loss, and validation loss for inceptionResNetV2 model.

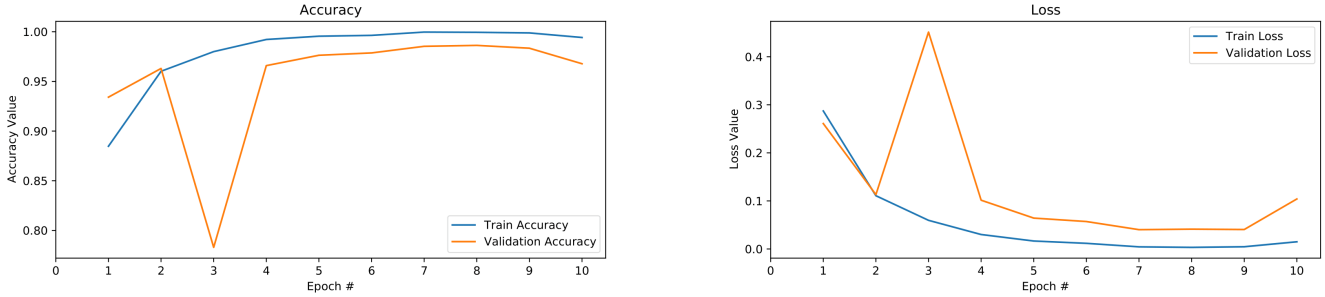

FIGURE 13. The curves of training accuracy, validation accuracy, training loss, and validation loss for VGG19 model.

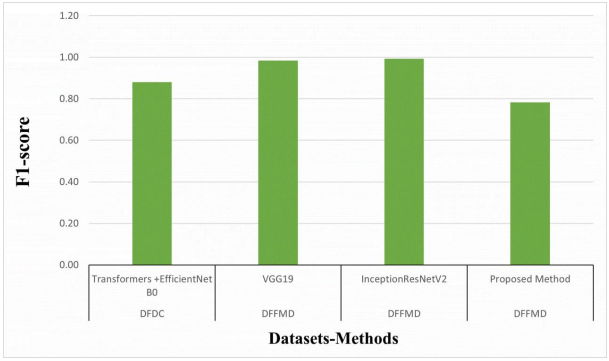

FIGURE 14. Comparison of F-score test.

an area under the ROC curve (AUC) (see Fig. 15), and F1-score (see Fig. 14). Thus, they have been used to evaluate the proposed method's performance. Starting with accuracy, we can conclude that the SVM method is not an ideal choice for Deepfake video detection due to achieved low accuracy in large datasets such as DFDC. Meanwhile, the traditional CNN model was more effective than SVM in detecting synthesized videos even though the data quality was high or low, as in the study [24]. However, it has a complex structure that extracts specific feature resolution during the pre-processing stage. Hence, this limitation limits using this method with different datasets sizes. Although EfficientNet and VGG16 detection methods do not contain a complex structure, they reported a high false positives rate and were then expected to overfit. Thus, in this paper, the proposed model was built to address these limitations. In this regard, our proposed method was CNN-based with a novel and simple structure that does not require any restriction while extracting features from the video frames. Besides, it can be used with any dataset size and quality since the proposed dataset (DFFMD) has medium quality. The model could detect faked video from it with realistic accuracy of 80.08%. Although the accuracy was not high compared to the literature, the model does not overfit since it reported the lowest training and evaluation loss. Regarding AUC and f1-score, the performance is quite similar even though the method used was traditional machine or deep learning algorithms. From Fig. 1 and 4, we observe that our proposed model can detect

TABLE 4. A quantitative comparison of SOTA video Deepfake detection methods with the proposed method results.

Measure	Dataset	Detection Method	Results
Accuracy	DFDC	SVM [29]	77%
		CNN [28]	97.6%
		EfficientNet+Resnet+U-Net [31]	98.7%
		CNN+ViT [26]	91.5%
		CNN+RNN [30]	91.88%
	Celeb-df	SVM [29]	75%
	FaceForensics++	VGG16(CNN+Optical) [27]	81.61%
		ResNet50(CNN+Optical flow) [27]	75.46%
		SVM [29]	87.3%
	DeepFake-TIMIT	SVM [29]	87.3%
		CNN (Low Quality) [24]	99.9%
		CNN (HIgh Quality) [24]	93.2%
	generated by authors	CNN+RNN+LSTM [23]	82.81%
	DFFMD	VGG19 [44]	98.55%
		InceptionResnetV2 [43]	99.39%
		Proposed Method	80.08%
AUC	UADFV	CNN [24]	97.4%
	DFDC	transformer+EfficientNet BO [25]	95.1%
		CNN+ViT [33]	91%
		SVM [29]	79.5%
	Celeb-df	SVM [29]	82.3%
	FaceForensics ++	SVM [29]	94.3%
	DeepFake-TIMIT	SVM [29]	98.2%
	DFFMS	VGG19 [44]	98.63%
		InceptionResNetV2 [43]	99.39%
		Proposed Method	79.88%
F1-score	DFDC	Transformer+EfficientNet BO [25]	88%
	DFFMD	VGG19 [44]	99.49%
		InceptionResNetV2 [43]	99.35%
		Proposed Method	78.33%

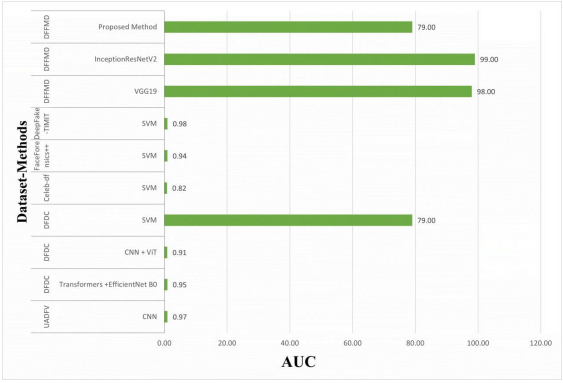

FIGURE 15. Comparison of AUC test.

fake videos in large-scale datasets generated with different fakeness types.

VIII. CONCLUSION AND FUTURE WORK

Detecting modern Deepfakes is an increasingly challenging and important problem to solve. This technology impacts security, crime, personal safety, large-scale politics, and society. With the COVID-19 virus breakout in 2020, the popularization of face masks that allow users to obstruct their faces has made the Deepfake generation easier and detecting such videos more difficult by magnitudes. This study aims to support research in Deepfake detection with its dataset of face mask Deepfakes and investigation of different deep learning models for Deepfake detection on the proposed dataset. Ultimately, the result of the study demonstrates that CNN, InceptionResNetV2, and VGG19 techniques can detect the deep-fake dataset at significant accuracies of 77.48%, 99.25%, and 99.81. The lack of video resources of humans wearing masks is a limitation we faced in this research study, which will be solved in the following research study. For future work to develop the subsequent experimental work, we will implement more deep learning techniques for detection.

ACKNOWLEDGMENT

The authors would like to thank Aljowhra Alshobaili for her contributions to this article. Aljowhra was lost to them this year to cancer but they will always remember her as the exemplary colleague and friend that she was.

REFERENCES

[1] *DeepFakes Software*. Accessed: Aug. 20, 2022. [Online]. Available: https://github.com/deepfakes/faceswap
[2] *A Denoising Autoencoder + Adversarial Losses and Attention Mechanisms for Face Swapping*. Accessed: Aug. 20, 2022. [Online]. Available: https://github.com/shaoanlu/faceswap-GAN
[3] *DeepFaceLab is the Leading Software for Creating DeepFakes*. Accessed: Feb. 24, 2022. [Online]. Available: https://github.com/iperov/DeepFaceLab
[4] *Larger Resolution Face Masked, Weirdly Warped, DeepFake*. Accessed: Feb. 24, 2022. [Online]. Available: https://github.com/dfaker/df
[5] N. J. Vickers, "Animal communication: When I'm calling you, will you answer too?" *Current Biol.*, vol. 27, no. 14, pp. R713–R715, Jul. 2017.
[6] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2889–2898.
[7] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2018, pp. 8789–8797.
[8] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.
[9] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.
[10] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.
[11] A. S. Uçan, F. M. Buçak, M. A. H. Tutuk, H. İ. Aydin, E. Semiz, and S. Bahtiyar, "Deepfake and security of video conferences," in *Proc. 6th Int. Conf. Comput. Sci. Eng. (UBMK)*, Sep. 2021, pp. 36–41.
[12] N. Graber-Mitchell, "Artificial illusions: Deepfakes as speech," Amherst College, MA, USA, Tech. Rep., 2020, vol. 14, no. 3.
[13] F. H. Almukhtar, "A robust facemask forgery detection system in video," *Periodicals Eng. Natural Sci.*, vol. 10, no. 3, pp. 212–220, 2022.
[14] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*.
[15] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A survey on deepfake video detection," *IET Biometrics*, vol. 10, no. 6, pp. 607–624, Nov. 2021.
[16] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "WildDeepfake: A challenging real-world dataset for deepfake detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2382–2390.
[17] S. R. Ahmed, E. Sonuç, M. R. Ahmed, and A. D. Duru, "Analysis survey on deepfake detection and recognition with convolutional neural networks," in *Proc. Int. Congr. Hum.-Comput. Interact., Optim. Robot. Appl. (HORA)*, Jun. 2022, pp. 1–7.
[18] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3207–3216.
[19] P. Korshunov and S. Marcel, "Vulnerability assessment and detection of deepfake videos," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2019, pp. 1–6.
[20] J. Huang, X. Wang, B. Du, P. Du, and C. Xu, "DeepFake MNIST+: A deepfake facial animation dataset," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1973–1982.
[21] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," 2021, *arXiv:2108.05080*.

[22] J. Hu, X. Liao, W. Wang, and Z. Qin, "Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1089–1102, Mar. 2022.
[23] A. Pishori, B. Rollins, N. van Houten, N. Chatwani, and O. Uraimov, "Detecting deepfake videos: An analysis of three techniques," 2020, *arXiv:2007.08517*.
[24] Y. Li and S. Lyu, "Exposing DeepFake videos by detecting face warping artifacts," 2018, *arXiv:1811.00656*.
[25] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and vision transformers for video deepfake detection," in *Proc. Int. Conf. Image Anal. Process.* Berlin, Germany: Springer, 2022, pp. 219–229.
[26] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," 2021, *arXiv:2102.11126*.
[27] I. Amerini, L. Galteri, R. Caldelli, and A. D. Bimbo, "Deepfake video detection through optical flow based CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–3.
[28] A. Singh, A. S. Saimbhi, N. Singh, and M. Mittal, "DeepFake video detection: A time-distributed approach," *Social Netw. Comput. Sci.*, vol. 1, no. 4, pp. 1–8, Jul. 2020.
[29] B. Xu, J. Liu, J. Liang, W. Lu, and Y. Zhang, "DeepFake videos detection based on texture features," *Comput., Mater. Continua*, vol. 68, no. 1, pp. 1375–1388, 2021.
[30] D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horváth, E. Bartusiak, J. Yang, D. Guera, F. Zhu, and E. J. Delp, "DeepFakes detection with automatic face weighting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 668–669.
[31] E. Tjon, M. Moh, and T.-S. Moh, "Eff-YNet: A dual task network for DeepFake detection and segmentation," in *Proc. 15th Int. Conf. Ubiquitous Inf. Manage. Commun. (IMCOM)*, Jan. 2021, pp. 1–8.
[32] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6.
[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.
[34] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1526–1535.
[35] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.
[36] M. Saito, E. Matsumoto, and S. Saito, "Temporal generative adversarial nets with singular value clipping," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2830–2839.
[37] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Comput. Surv.*, vol. 54, no. 1, pp. 1–41, 2021.
[38] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and beyond: A survey of face manipulation and fake detection," 2020, *arXiv:2001.00179*.
[39] S. McCloskey and M. Albright, "Detecting GAN-generated imagery using saturation cues," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4584–4588.
[40] S. Suratkar, E. Johnson, K. Variyambat, M. Panchal, and F. Kazi, "Employing transfer-learning based CNN architectures to enhance the generalizability of deepfake detection," in *Proc. 11th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2020, pp. 1–9.
[41] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Proc. Eur. Conf. Inf. Retr.* Berlin, Germany: Springer, 2005, pp. 345–359.
[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[43] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. 21st AAAI Conf. Artif. Intell.*, 2017, pp. 1–7.
[44] S. H. Jeong, J. P. Yun, H.-G. Yeom, H. J. Lim, J. Lee, and B. C. Kim, "Deep learning based discrimination of soft tissue profiles requiring orthognathic surgery by facial photographs," *Sci. Rep.*, vol. 10, no. 1, pp. 1–5, Oct. 2020.
[45] *This Repository Contains the Source Code for the Paper First Order Motion Model for Image Animation*. Accessed: Aug. 20, 2022. [Online]. Available: https://github.com/AliaksandrSiarohin/first-order-model
[46] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, "Synthesizing images of humans in unseen poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8340–8348.
[47] A. Shysheya, E. Zakharov, K.-A. Aliev, R. Bashirov, E. Burkov, K. Iskakov, A. Ivakhnenko, Y. Malkov, I. Pasechnik, D. Ulyanov, A. Vakhitov, and V. Lempitsky, "Textured neural avatars," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2387–2397.
[48] H. Tang, W. Wang, D. Xu, Y. Yan, and N. Sebe, "GestureGAN for hand gesture-to-gesture translation in the wild," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 774–782.
[49] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022.

NORAH M. ALNAIM (Member, IEEE) received the bachelor's degree in computer information systems from King Faisal University, in 2008, the master's degree in computer information systems from St. Mary's University, San Antonio, TX, USA, in 2012, and the Ph.D. degree in electrical engineering and electronics from Brunel London University, London, U.K., in 2020. She is currently an Assistant Professor with the Computer Science Department, Imam Abdulrahman Bin Faisal University, Saudi Arabia. She has received eight rewarding excellence regarding academic and research output in the past few years. She is also a member of different organizations. Her research interests include computer vision, machine learning, gesture recognition, image processing, and artificial intelligence fields.

ZAYNAB M. ALMUTAIRI received the bachelor's degree in information technology from Majmaah University, in 2017. She is currently pursuing the master's degree in information technology with King Saud University. She has received two rewarding excellence regarding academics and research output in the past few years. Her research interests include machine learning, big data analytics, NLP, data mining, and computer vision fields.

MANAL S. ALSUWAT received the bachelor's degree in computer science from Taif University, in 2019. She is currently pursuing the master's degree in artificial intelligence with Umm Al-Qura University. In the past few years, she has received four rewarding academic and research achievements. Her research interests include computer vision, machine learning, and speech recognition.

HANA H. ALALAWI received the bachelor's degree in computer science from Umm Al-Qura University, in 2018. She is currently pursuing the master's degree in artificial intelligence. She has received three rewarding excellence regarding academic and research output in the past few years. Her research interests include machine learning, computer vision, speech and gesture recognition, and machine translation.

ALJOWHRA ALSHOBAILI, photograph and biography not available at the time of publication.

FAYADH S. ALENEZI received the B.Sc. degree (Hons.) in electrical engineering electronics and communications track from Jouf University, Sakaka, Saudi Arabia, in 2012, the M.S. degree in electrical engineering from Southern Illinois University, Carbondale, IL, USA, in 2015, and the Ph.D. degree in electrical engineering from the University of Toledo, OH, USA, in 2019. He is currently an Assistant Professor with the Department of Electrical Engineering, Jouf University. He has authored several journals and conference papers. His research interests include artificial intelligence, image processing, signal processing, image enchantment, machine learning, neural networks, and facial recognition.