

A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow Features

Pallabi Saikia^{1,*}, Dhvani Dholaria¹, Priyanka Yadav¹, Vaidehi Patel¹, Mohendra Roy^{2,*}

¹Computer Science and Engineering Department, School of Technology

²Information & Communication Technology, School of Technology

Pandit Deendayal Energy University, Gandhinagar -382007, India

*Corresponding Authors: mohendra.roy@ieee.org; pallabi.iitg@gmail.com

Abstract—Deepfakes are the synthesized digital media in order to create ultra-realistic fake videos to trick the spectator. Deep generative algorithms, such as, Generative Adversarial Networks (GAN) are widely used to accomplish such tasks. This approach synthesizes pseudo-realistic contents that are very difficult to distinguish by traditional detection methods. In most cases, Convolutional Neural Network (CNN) based discriminators are being used for detecting such synthesized media. However, it emphasise primarily on the spatial attributes of individual video frames, thereby fail to learn the temporal information from their inter-frame relations. In this paper, we leveraged an optical flow based feature extraction approach to extract the temporal features, which are then fed to a hybrid model for classification. This hybrid model is based on the combination of CNN and recurrent neural network (RNN) architectures. The hybrid model provides effective performance on open source data-sets such as, DFDC, FF++ and Celeb-DF. This proposed method shows an accuracy of 66.26%, 91.21% and 79.49% in DFDC, FF++ , and Celeb-DF respectively with a very reduced No of sample size of ≤ 100 samples(frames). This promises early detection of fake contents compared to existing modalities.

1

I. INTRODUCTION

Deepfakes are images and videos, usually created by deep neural networks to superimpose target subject's face features over another in order to produce fake media. According to a recent report from the Deeptrace Lab [1], there are almost 15,000 deepfake media over the internet consisting of non-obscene videos targeting the politicians and the functioning of democratic societies. More than 13,000 deepfake videos were found on different deepfake-specific porn sites [2], and about 96% of the deepfake content on web are related to pornography and mostly related to famous celebrities [2] to defame the individuals. Day by day, deepfake's findings are becoming so realistic that they are almost indistinguishable, and the substituted subjects are rigged to say things they never spoke [3]. Deepfake methods have been extensively applied nowadays to produce enormous fake news, posing a serious threat to communities worldwide, and have the potential to influence the masses as well as democratic and geopolitical structure of a region [4].

Many organizations as well as private companies are investing heavily to counter the challenges of deepfake. Various

techniques have already been studied in the field of deepfake detection, including Machine Learning (ML) techniques such as Support Vector Machines (SVMs) [5], and Deep learning techniques such as Convolution Neural Network (CNN) [6], CNN with SVM [7], Recurrent Neural Network (RNN) [8], CNN with Long Short Term Memory (LSTM) [8], etc. Moreover, many traditional ways have also been explored to detect manipulated media, such as exposing inconsistent in head-poses, consideration of the background color manipulation [9]. However, most of these techniques are focuses primarily on the spatial feature analysis and does not include any temporal information. Since, most of the deepfake media are in the form of video, therefore, identifying inconsistencies in temporal information (along with spatial inconsistencies) may enhance the classification accuracy of deepfakes.

In this work, we investigated a hybrid deep learning approach on modelling the intra-frame as well as inter-frame features of videos to accurately identify its authenticity. Further, we incorporated a traditional temporal feature analysis method, optical flow to help in extraction of the temporal features. The optical flow implementation is based on to characterization of the motion of the subject's face and the technique exploits the possible inter-frame dissimilarities. A detail analysis has been carried out on to characterize the proposed model on various sets of video data. The proposed model is evaluated based on its various performance parameters such as Accuracy, Recall, Precision, F1-score, and AUC.

II. BACKGROUND AND RELATED WORKS

A. Deepfake Generation

There are several machine learning algorithms that can produce credible deceptive videos. Moreover, the recent advancement in adversarial techniques such as generative adversarial network (GAN) has fuel the rapid development of digital forgery. The algorithms have been widely used in many of the modern deepfake generation approaches [10], [11]. The use of Generative Adversarial Nets (GANs) in deepfake generation has been a prominent method based on neural networks [12]. It works on the idea of setting dual neural networks in conflict with one another, i.e., the generator G that generates the output image and the discriminator D that determines whether its fake or real [11]. GAN was first introduced in 2014 by Goodfellow et al. [12]. The generator G generates fake data

¹©IEEE, Paper No: 832, IJCNN, 2022 IEEE World Congress on Computational Intelligence

x_g to mislead the discriminator D . D also learns how to differentiate between the fake media ($x_g = G(x)$ where $z \sim N$) and real media ($x \in X$). G and D are trained on an adversarial loss respectively as follows,

$$L_{adv}(D) = \max \log D(x) + \log(1 - D(G(z))) \quad (1)$$

$$L_{adv}(G) = \min \log(1 - D(G(z))) \quad (2)$$

There are two major approaches in deepfake generation, these are FaceSwap and Face Synthesis. In the FaceSwap, a target face is swapped onto a source face and in Face synthesis the facial features are being synthesised. Recent High-Resolution Face Swapping method from Disney Research is one of the very successful face swapping method [13]. Similarly the LandmarkGAN is a Face Synthesis method based on facial landmark as input [14]. All these state-of-art techniques are capable of generating ultra-realistic synthesized media, which are almost indistinguishable by traditional means. This demands a sophisticated detection technique. Recently, the adversarial detectors are being used to counter this issue. However, in many cases the adversarial detectors can be tricked [15]. Therefore, a more generalized multi-model approach is focused in nowadays research.

B. Deepfake Detection

Traditionally, inconsistencies or unrealistic elements in the forgeries are targeted by several detecting approaches [16]. Most of the detection techniques nowadays, mainly rely on machine learning techniques to generalise the detection process [17]. Peng Chen, et. al. [18] has developed FSSPOTTER, a unified framework for detection of deepfakes. It investigates the rich spatial features within a single frame with the help of a Spatial Feature Extractor (SFE) along with a Temporal Feature Aggregator (TFA) which extracts the inconsistencies between the frames. Digvijay Yadav, et al. [19] considered blinking of eyes as one of the important features to detect the deepfake, and for detecting the DeepFakes, CNN architecture is combined with LSTM to detect the temporal inconsistencies in changes in the frames. Irene Amerini, et. al. [20] introduced a method to exploit the temporal inconsistencies of videos with optical flow fields between two consecutive frames so as to differentiate between original and fake videos. Shivangi et al. [21] proposed a transfer learning based approach, named as Deep Distribution Transfer, to overcome the problem of zero-shot and few-shot transfer for facial forgery detection. Distribution-based loss formulation is used to bridge the gaps between different facial forgery techniques involved in the creations of the datasets. XTao, et. al. [22] proposed another framework that accentuates the way to accomplish better outcomes, appropriate edge arrangement and motion compensation. The work consists of introducing a “sub-pixel motion compensation” (SPMC) layer in a CNN framework. The versatile CNN structure joins the SPMC layer and fuses numerous frames to reveal image details. The paper furnish an investigation into how to coordinate numerous casing

inputs for improving outcomes. David Guera, et. al. [8] demonstrated the effectiveness of Long Short-Term Memory (LSTM) networks along with CNN model for the detection of the deepfakes. InceptionV3 with fully connected layer at the top of the network outputs a deep representation of each frame and further LSTM model takes the feature sequences to model the temporal dependencies. This method thus elaborates a temporal-aware system to automatically detect deepfake videos. Typically, well-known and pre-trained CNN algorithms are widely explored in literature to learn discrete aspects from each frame of the video sequence. Most of the state-of-the-art algorithms focus primarily on extraction of the intra-frame features for deepfake detection. However, effective extraction of inter-frame feature to exploit temporal inconsistencies is also a promising direction in the research of deepfake detection.

III. PROPOSED METHOD

In this work, we have focused on the facial features of a video to find characteristics of a deepfake content. The most common characteristics in the manipulated media is that an individual in a video is supplanted with another person’s profile. On account of that we are contemplating the facial characteristics, as the warping leaves a few distinguishable artifacts in the deepfake videos. We exploit this information and incorporated various pre-processing steps, followed by model building as illustrated below:

A. Frame Extraction

To grab the frames from the video, uniform sampling is applied on the video duration. During preprocessing, we disregarded the frames that doesn’t contain a face. Handling highly sampled frames will require a great deal of computational power. So, for probing reason on an average we have extracted 148 frames per video. Further, we have extracted the face of a subject from these videos as mentioned below.

B. Face Extraction

The proposed methodology considers the face as a region of interest (ROI). We extracted the ROI using a `batch_face_locations` algorithm within `face_recognition` library [23]. `Face_Recognition` library wraps `dlib`’s face recognition functions [24] into a simple, easy to use API. It captures 128 data points per face, resulting in unique parameters for the hash. A re-scaling was performed to remove extra-remaining background to reduce the memory complexity, which eventually help in reducing the computational complexity of the model. Thus, we obtain a modified video data-set with frames having size of 112×112 .

C. Optical flow field Feature Extraction

Optical flow fields for consecutive pairs of frames are generated to detect the pattern of apparent motion in the individual pixels on the image plane [25]. It is used to extricate the movement of patterns in an image based on apparent velocities distribution. The image intensity between consecutive frames can be expressed as a function of time (t) and space (x,y).

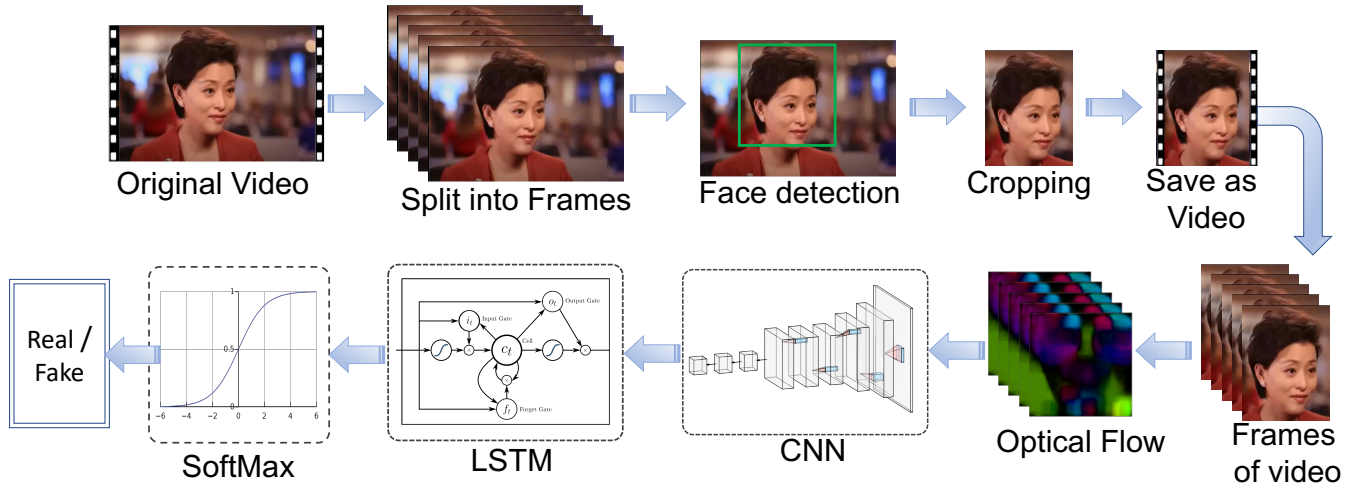


Fig. 1: Proposed Workflow of Deepfake Detection in video. Here, the original video was splitted into frames and then the face ROI was segmented out by cropping the ROI. The cropped frames are again saved as video. This reduced sizes videos are then used for generating optical flow. The optical flow features are then fed as input feature to the hybrid classification model consisting of combination of CNN and LSTM. The output of the LSTM were finally activated through SoftMax function to find the probability of classes. Thus obtain the classification.

Thus, an frame or image can be represented as $I(x, y, t)$. If the image makes a displacement of (dx, dy) in time dt , then the new image will be $I(x + dx, y + dy, t + dt)$.

Using the Taylor series expansion, the change can be expanded as:

$$I(x+dx, y+dy, t+dt) = I(x, y, t) + \frac{\partial I}{\partial x} dx + \frac{\partial I}{\partial y} dy + \frac{\partial I}{\partial t} dt + \dots \quad (3)$$

If the change in intensities remain constant in both the frames, then $I(x, y, t) = I(x + dx, y + dy, t + dt)$

$$\Rightarrow \frac{\partial I}{\partial x} dx + \frac{\partial I}{\partial y} dy + \frac{\partial I}{\partial t} dt = 0 \quad (4)$$

Now, dividing the equation (4) by dt , we get:

$$\Rightarrow \frac{\partial I}{\partial x} u + \frac{\partial I}{\partial y} v + \frac{\partial I}{\partial t} = 0 \quad (5)$$

where, $u = \frac{dx}{dt}$ and $v = \frac{dy}{dt}$

$\frac{\partial I}{\partial x}$ = image gradient along the x-axis

$\frac{\partial I}{\partial y}$ = image gradient along the y-axis

$\frac{\partial I}{\partial t}$ = image gradient along time

The cropped face videos, that obtained in the previous step of face extraction are splitted into frames and optical flow between two consecutive frames is calculated for identifying temporal inconsistencies for deepfake detection. Through calculation of optical flow, we detected the change in motion of every pixel of an image. The HSV color representation of the optical flow vector using the magnitude and the hue component

envisioned by the direction vector is shown in Figure 2. From our study, we find that the fake videos have distorted motion vectors as compared to the real ones. This motion vector plot of each axis was then converted to 3 channel images using predefined color code for feeding into the hybrid model discussed in the next step.

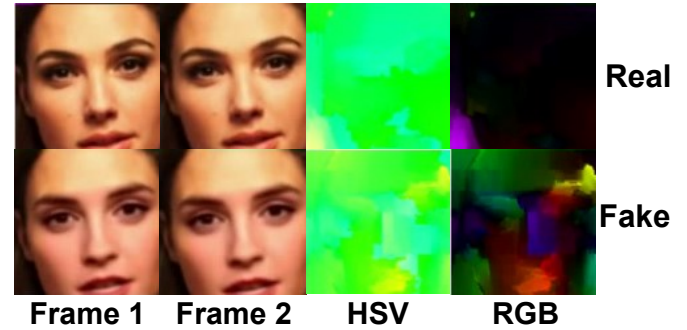


Fig. 2: Optical flow features of real and fake video frames. From this figure a distinct variation in the optical flow features can be visible between real and fake frames.

D. Hybrid CNN-RNN Architecture for modelling

The combined arrays of the color coded frames obtained as a result of motion based feature extraction process as mentioned above, provide us the dataset having explicit temporal information. This data are then fed to a pre-trained CNN model. We have explored several state-of-the-art pre-trained models, such as, VGG16 [26], InceptionV3 [27], ResNet50 [28], Xception [29], MobileNetV2 [30], EfficientNetB7 [31]. A pre-trained model, piled up with several layers of various architectural blocks to frame an exceptionally profound network, generally perform very well and require less time to re-train. Since, it

TABLE I: Comparison of the performance of the various base models

Batch Size:128	DFDC (Frames:20)			Celeb-DF (Frames:50)			FF++ (Frames:30)		
	Test Accuracy	F1	AUC	Test Accuracy	F1	AUC	Test Accuracy	F1	AUC
VGG16	64.73%	64.28%	0.64	67.09%	68.14%	0.68	78.39%	78.45%	0.78
InceptionV3	45.19%	50.00%	0.5	55.12%	52.27%	0.52	51.00%	50.51%	0.5
ResNet50	64.58%	59.87%	0.59	67.09%	68.44%	0.68	89.67%	89.65%	0.89
Xception	63.20%	64.19%	0.64	59.22%	63.49%	0.63	73.86%	73.04%	0.73
MobileNetV2	57.09%	55.68%	0.55	63.24%	65.21%	0.65	76.63%	76.76%	0.76
EfficientNetB7	59.84%	56.16%	0.56	70.08%	69.13%	0.69	83.66%	84.04%	0.84

TABLE II: Comparison of the performance of the base models with optical flow as input features on various datasets

	Celeb DF				DFDC				FF++			
	Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall	AUC
OF+RNN	52.13%	26.06%	34.21%	0.5	54.08%	24.96%	35.08%	0.5	47.73%	23.68%	50%	0.5
OF+CNN	83.33%	83.78%	83.71%	0.83	69.77%	69.36%	68.64%	0.68	89.19%	89.51%	88.92%	0.88
OF+RNN+CNN	79.49%	82.49%	79.08%	0.79	66.26%	67.11%	65.73%	0.66	91.21%	91.20%	91.21%	0.91

explicitly trained on a million images, e.g. "ImageNet" [32], hence effective for modelling vision related problems [33].

Image classification by the above pretrained models is carried out in mainly two main phases: feature extractor with the convolution layers, and discrimination with fully connected layers of CNN. The last layers of the pretrained model are fully connected, are also called as dense layers, are specific to the task, and hence was excluded for fine tuning the models on the deepfake datasets. Two LSTM layers have been added after the conv-layers and before the fully connected layers for classification. The LSTM layers also examines the interframe inconsistencies on the extracted abstract features with a dropout of 0.5. Dropout layers are significant in training complex models because they prevent the training data from being overfit. As a result, it may be possible to avoid learning of features that only appear in later samples or batches. Finally, the architecture completes with softmax layer added at the end, that compute the probabilities of the frame sequence being either fake or real. Categorical cross-entropy loss function applied to calculate the loss of the deepfake classification model, is provided in equation 6, where \hat{y}_i is the predicted score of class i at the softmax layer:

$$Loss = - \sum_{i=1}^{outputsize} y_i \times \log \hat{y}_i \quad (6)$$

IV. DATASET DESCRIPTION

We have applied our proposed methodology on three sets of data, namely FaceForensics++ [34], Celeb-DF [35], and the Deep Fake Detection Challenge (DFDC) dataset [36]. We splitted the datasets into 80:20 ratio for training and testing respectively.

A. Celeb-DF

This dataset consists of 408 real videos sourced from YouTube and synthesized 795 videos with improvements in the usual Deep-Fake generation model. Although the visual quality of the videos is low but the quality of face-swaps seems quite realistic [35].

B. DeepFakeDetectionChallenge (DFDC)

The dataset is one of the most recent participant within the category of Deep-Fake datasets, has been compiled by Facebook AI. Sixty Six paid actors were involved in the train and test sets, and their filmed sequences were considered to generate manipulated videos internally to avoid cross-set face-swaps. Dataset comprises a total of 5214 videos out of which 78.125% are manipulated. They achieved high quality of manipulations by choosing pairs of similar appearances, and visual quality is high as well [36].

C. FaceForensics++

The forensic dataset consisting of thousand original video sequences. Four automated face manipulation methods such as, Face2Face, Deepfakes, FaceSwap, and NeuralTextures have been applied to manipulate and generate the videos sequences. The data has been sourced from 977 youtube videos and all videos contain a trackable, mostly frontal face and without occlusions, which enables automated tampering methods to generate realistic forgeries [34].

V. EXPERIMENTAL RESULTS AND ANALYSIS

The experiments were performed on Google Colab Pro with 25 GB RAM and the codes were developed using python 3. We have used VGG16, InceptionV3, ResNet50, Xception, MobileNetV2, and EfficientNetB7 baseline unimodals to perform experiments. Adam optimizer with a learning rate of 1×10^{-5} is employed to train the neural network models. We used F1-score, Precision, Recall, AUC, Accuracy metric for model evaluation. Different other libraries are used for the experimentation, such as OpenCV, Keras, sklearn, Scipy, Pandas, and face recognition.

We first experimented by varying the pre-trained CNN models as shown in table I. As we can see, InceptionV3 is not able to differentiate well as it gives out almost 0.5 value for the AUC curve. For the DFDC dataset, VGG16 is giving better performance with 20 Frames. EfficientNetB7 is giving better accuracy for Celeb-DF dataset with 50 frames and VGG16 is

TABLE III: Comparison of the performance of the hybrid model on various datasets w.r.t the number of frames

Frames	DFDC Dataset			FF++ Dataset			CelebDF Dataset		
	Accuracy	Precision	AUC Score	Accuracy	Precision	AUC Score	Accuracy	Precision	AUC Score
10	64.58%	64.39%	0.63	74.87%	75.71%	0.75	63.25%	66.67%	0.65
20	64.27%	65.99%	0.64	78.39%	78.80%	0.78	67.09%	68.64%	0.68
30	66.26%	67.11%	0.66	83.17%	83.26%	0.83	73.07%	72.94%	0.73
40	64.12%	69.04%	0.63	77.89%	82.31%	0.79	73.08%	72.84%	0.73
50	-	-	-	86.68%	86.95%	0.87	78.21%	78.19%	0.78
60	-	-	-	83.17%	86.32%	0.84	72.65%	76.33%	0.73
70	-	-	-	91.21%	91.20%	0.91	74.36%	76.34%	0.74
80	-	-	-	-	-	-	76.07%	81.83%	0.76
90	-	-	-	-	-	-	79.49%	82.69%	0.79
100	-	-	-	-	-	-	76.07%	79.92%	0.76

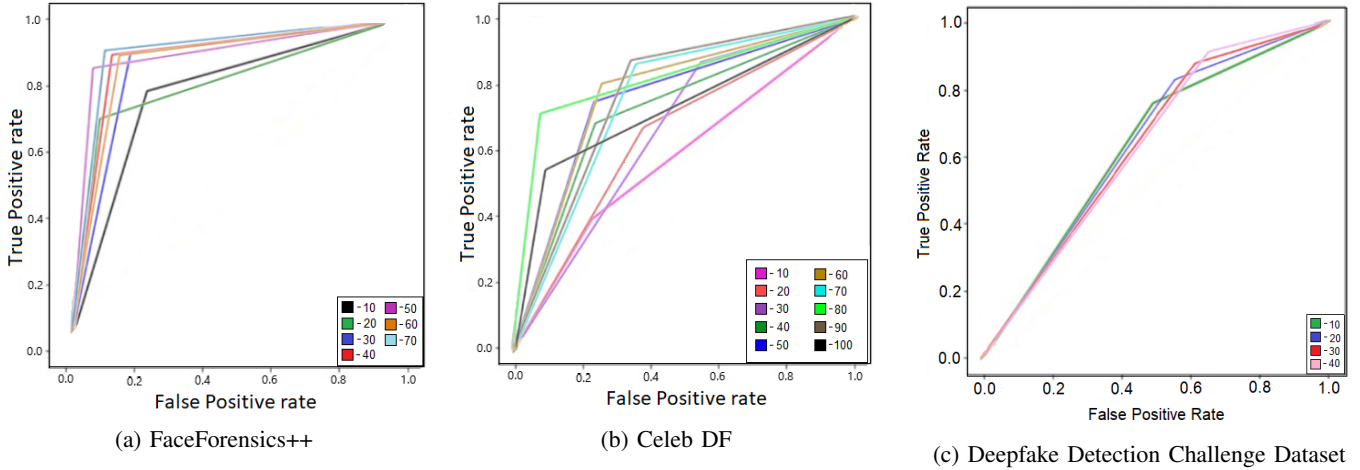


Fig. 3: AUC curves for the three datasets. Here the color indicates the corresponding number of frames as indicated in labels in respective plots.

the second best for the same. For FF++, Resnet50 is giving higher accuracy with 30 frames and VGG16 is the third best performing for FF++. Taking all metrics into consideration ResNet 50 performs best for all the three datasets for the reason that it is faster to train and easier to use and deploy. Despite the fact that ResNet 50 is much deeper than VGG16 the model size is significantly smaller because of the utilization of global average pooling rather than fully-connected layers - this diminishes the model size down perhaps that might be the reason why it obtains better result than other models. However, all the models don't have much differences in performance irrespective of the pre-trained architecture chosen.

Table II provides the comparison result of the proposed model with various compositions of the models components: optical flow field (OF), CNN and RNN. From the results on experimentation on 40 frames, 70 frames and 100 frames of DFDC, FF++ and Celeb-DF datasets, it can be deduced that, the (OF+RNN) model is not able to differentiate well as all the datasets gives out 0.5 value for the AUC curve. This means the classifier is ineffective to properly distinguish between positive and negative class points, and hence for all available data points, it predicts a constant or random a class. The (OF+CNN) model performs better results than previous

model and can distinguish between Real and Fake. For the Celeb DF and DFDC, (OF+CNN) model yields finer results than the hybrid (OF+CNN+RNN) model. On FF++ Dataset, the hybrid model outperforms both the other models.

Figures 3a, 3b, and 3c illustrate the change in the AUC (Area Under The Curve) curve on the considered dataset by our proposed model (OF+RNN+CNN). AUC score of 0.5 means model cannot distinguish between real or fake. AUC score of 1 depicts that the model is able to perfectly classify the videos into real or fake. Thus an excellent model has an AUC score closer to 1. The AUC curves obtained for our hybrid model at varying number of frames have been merged into a single graph for each dataset, to showcase how the model reaches better performance with increasing number of frames. The different coloured lines represent curves at different number of frames. For FF++ as shown in Figure 3a, the curves up to 70 frames have been considered. The light blue curve corresponds to an AUC score of 0.91, which is the highest the model has achieved. For CelebDF dataset as shown in Figure 3b the curves up to 100 frames have been plotted. It contains a total of 1168 videos, the least as compared to DFDC and FaceForensics++. For the DFDC dataset as shown in Figure 3c, plotted curves up to 40 frames

have been considered, as DFDC contains 3293 videos, the highest as compared to other two datasets used in this paper.

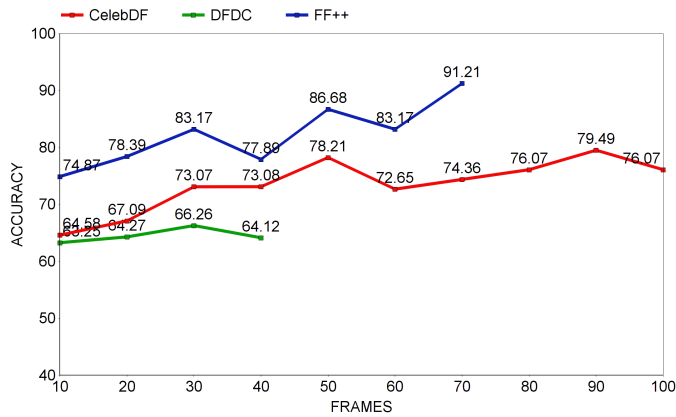


Fig. 4: Performance of the 3 datasets w.r.t the number of frame. Here the color indicates the corresponding dataset.

Table III provides the detailed comparison of the proposed model on varying the number of frames used per video for training and testing. Figure 4 illustrates the observations from the table III in a graphical manner. From the table it is observed that there is a general trend that the performance of the model improves as the number of frames increases. This is because with more frames per video the model can better detect the temporal variances for classification as real or fake. There have been certain deviations from this trend but overall performance has increased. FaceForensics++ was the best performing dataset and DFDC being the dataset with the least performance metrics. In this table only frames upto 40 have been considered for DFDC dataset and upto 70 frames for FaceForensics++ to reduce computational complexity. Most of the work in the area of Deepfake detection has used more frames than used in our work. There is no doubt that with the increase in the number of frames, the accuracy of our model would also increase to a great extent.

Table IV provides the results of our proposed model with the baseline models in the literature. One study on the effect of optical flow and CNN training was also discussed in the work [20]. Work by Peng Chen et. al. [18] investigates the rich spatial features with the help of spatial and temporal clues. They employ a two stage training strategy by learning temporal features and spatial inconsistencies separately. They have only used the AUC score percentage to present their results. Another work [21] have adopted the approach of Deep Distribution Transfer learning. Results are obtained with 4 datasets individually as well by combining all the datasets. Comparing their results on Celeb-DF dataset with the results we obtained, our model has achieved higher accuracy. Other works like [6], [37], [38], [39], [40] have presented better performance only based on accuracy. The reason for this can be accounted to the fact that our work was limited to a certain number of frames of the videos. In this paper we are using CNN along with the LSTM and then we are further improving it by combining optical flow features. Our model has been

TABLE IV: Comparing proposed model with existing works.

Paper	Method	Model	Datasets	Accuracy	AUC	
Irene Amerini, et. al. [20]	Optical flow + CNN	VGG16	FF++	81.61%	-	
		ResNet 50	FF++	75.46%	-	
Peng Chen, et. al. [18]	CNN & LSTM	VGG16	FF++	-	100	
			Celeb-DF	-	77.6	
			UADFV	-	91.1	
			Deepfake	HQ	-	98.5
			TIMIT	LQ	-	99.5
D Afchar, et al. [6]	CNN	Meso-4,	FF++ (Face2face)	95%	-	
			Deepfake	-	-	
		Meso-Inception-4	FF++ (Face2face)	98%	-	
			Deepfake	-	-	
Shivangi Aneja, et al. [21]	CNN & LSTM	Resnet18	FF++	92.23%	-	
			Google DFD	81.21%	-	
			AIF	60.79%	-	
			Dessa	74.28%	-	
			Celeb-DF	68.83%	-	
			Combined	75.47%	-	
			Celeb-DF	83.49%	-	
Pranjal Ranjan, et. al [37]	CNN + LSTM	Xception Net	DFDC	78.13%	-	
			DFD	94.33%	-	
			Combined	79.62%	-	
			Celeb-DF	85.11%	-	
			DFDC	98.84%	-	
X Li, et. al [38]	Multiple Instance Learning	Xception Net	FFPMS	90.71%	-	
			Celeb-DF	85.11%	-	
			DFDC	98.84%	-	
De Lima, et. al [39]	Spatio-temporal Convolutional Networks	RCN	Celeb-DF	76.25%	74.87	
		R2Plus1D	Celeb-DF	98.07%	99.43	
		I3D	Celeb-DF	92.28%	97.59	
		R3D	Celeb-DF	98.26%	99.73	
		MC3	Celeb-DF	97.49%	99.30	
SA Khan, et. al [40]	CNN	VGG16	DFDC	96.75%	-	
Our Work	Optical Flow + CNN + LSTM	VGG16	FF++	91.21%	0.91	
			Celeb-DF	79.49%	0.79	
			DFDC	66.26%	0.66	

evaluated on 5 metrics namely accuracy, F1-score, Precision, Recall, and AUC, for Celeb-DF they are 79.80%, 78.80%, 82.49%, 79.08%, and 0.79, respectively, for FF++, they are 91.21%, 91.20%, 91.20%, 91.21%, and 0.91, respectively, and for DFDC, they are 66.26%, 65.35%, 67.11%, 65.73%, and 0.66, respectively. , whereas most of the work in this field evaluate their performance only on accuracy and AUC score. From this comparison we can find that our model is showing a comparable result even with a reduced frame number.

VI. CONCLUSION AND FUTURE SCOPE

This work is based on the use of Optical Flow vectors with pre-trained CNN model, appended with LSTM layers to model the inconsistent motion of each pixel of the frames of videos, which can be evaluated to classify a video into fake or real. To reduce the computational constraints, the experiment was performed on a subset of frames as considering all the frames of the videos require higher computational power. However, from the experimentation it is observed that the model performed better with an increasing number of frames per video. Our work paves the way for many possible future works: firstly the model can be improved by training on huge set of the frames of the videos. Secondly, more datasets can be incorporated for better performance so that the model can be trained to detect videos of all kinds of deepfake manipulation techniques. Further the comparable score of our proposed model with the reduced number of frames indicate the possible

realization of early detection of the fake content. Thus, the application of optical flow field seems to be promising in this domain and can be further investigated on explainability of ultra-realistic deepfakes.

REFERENCES

- [1] K. Kikerpill, "Choose your stars and studs: the rise of deepfake designer porn," *Porn Studies*, vol. 7, no. 4, pp. 352–356, 2020.
- [2] H. Ajder, G. Patrini, F. Cavalli, and L. Cullen, "The state of deepfakes: Landscape, threats, and impact," *Amsterdam: Deeptrace*, vol. 27, 2019.
- [3] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.
- [4] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *Calif. L. Rev.*, vol. 107, p. 1753, 2019.
- [5] F. F. Kharbat, T. Elamsy, A. Mahmoud, and R. Abdullah, "Image feature detectors for deepfake video detection," in *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*. IEEE, 2019, pp. 1–4.
- [6] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2018, pp. 1–7.
- [7] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8261–8265.
- [8] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE, 2018, pp. 1–6.
- [9] S. Agarwal, H. Farid, T. El-Gaaly, and S.-N. Lim, "Detecting deepfake videos from appearance and behavior," in *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2020, pp. 1–6.
- [10] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.
- [11] S. Singh, R. Sharma, and A. F. Smeaton, "Using gans to synthesize minimum training data for deepfake generation," *arXiv preprint arXiv:2011.05421*, 2020.
- [12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.
- [13] J. Naruniec, L. Helming, C. Schroers, and R. M. Weber, "High-resolution neural face swapping for visual effects," in *Computer Graphics Forum*, vol. 39, no. 4. Wiley Online Library, 2020, pp. 173–184.
- [14] P. Sun, Y. Li, H. Qi, and S. Lyu, "Landmarkgan: Synthesizing faces from landmarks," *arXiv preprint arXiv:2011.00269*, 2020.
- [15] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Advances in adversarial attacks and defenses in computer vision: A survey," *IEEE Access*, vol. 9, pp. 155 161–155 196, 2021.
- [16] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep learning for deepfakes creation and detection: A survey," *arXiv preprint arXiv:1909.11573*, 2019.
- [17] A. A. Maksutov, V. O. Morozov, A. A. Lavrenov, and A. S. Smirnov, "Methods of deepfake detection based on machine learning," in *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIconRus)*. IEEE, 2020, pp. 408–411.
- [18] P. Chen, J. Liu, T. Liang, G. Zhou, H. Gao, J. Dai, and J. Han, "Fsspotter: Spotting face-swapped video by spatial and temporal clues," in *2020 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2020, pp. 1–6.
- [19] D. Yadav and S. Salmani, "Deepfake: A survey on facial forgery technique using generative adversarial network," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. IEEE, 2019, pp. 852–857.
- [20] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based cnn," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.
- [21] S. Aneja and M. Nießner, "Generalized zero and few-shot transfer for facial forgery detection," *arXiv preprint arXiv:2006.11863*, 2020.
- [22] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4472–4480.
- [23] A. Geitgey, "Face recognition documentation," *Release 1.2*, vol. 3, pp. 3–37, 2019.
- [24] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.
- [25] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.
- [26] H. Qassim, A. Verma, and D. Feinzimer, "Compressed residual-vgg16 cnn model for big data places image recognition," in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2018, pp. 169–175.
- [27] X. Xia, C. Xu, and B. Nan, "Inception-v3 for flower classification," in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 2017, pp. 783–787.
- [28] L. Wen, X. Li, and L. Gao, "A transfer convolutional neural network for fault diagnosis based on resnet-50," *Neural Computing and Applications*, vol. 32, no. 10, pp. 6111–6124, 2020.
- [29] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.
- [30] S. L. Rabano, M. K. Cabatuan, E. Sybingco, E. P. Dadios, and E. J. Calilung, "Common garbage classification using mobilenet," in *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*. IEEE, 2018, pp. 1–4.
- [31] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.
- [32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [33] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.
- [34] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1–11.
- [35] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3207–3216.
- [36] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.
- [37] P. Ranjan, S. Patil, and F. Kazi, "Improved generalizability of deepfakes detection using transfer learning based cnn framework," in *2020 3rd international conference on information and computer technologies (ICICT)*. IEEE, 2020, pp. 86–90.
- [38] X. Li, Y. Lang, Y. Chen, X. Mao, Y. He, S. Wang, H. Xue, and Q. Lu, "Sharp multiple instance learning for deepfake video detection," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1864–1872.
- [39] O. de Lima, S. Franklin, S. Basu, B. Karwoski, and A. George, "Deepfake detection using spatiotemporal convolutional networks," *arXiv preprint arXiv:2006.14749*, 2020.
- [40] S. A. Khan, A. Artusi, and H. Dai, "Adversarially robust deepfake media detection using fused convolutional neural network predictions," *arXiv preprint arXiv:2102.05950*, 2021.