Extracting Facial Features to Detect Deepfake Videos Using Machine Learning

Ayesha Aslam[1], Jamaluddin Mir[2], Gohar Zaman[3], Atta Rahman[4]*, Asiya Abdus Salam[5], Farhan Ali[6]*, Jamal Alhiyafi[7], Aghiad Bakry[8], Mustafa Jamal Gul[9], Mohammed Gollapalli[10], Maqsood Mahmud[11]

Department of Computer Science, Abbottabad University of Science and Technology, Havelian, Pakistan[1, 2, 3]
Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia[4, 8]
Department of Computer Information Systems-College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia[5]
College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China[6]
Department of Computer Science, Kettering University, Flint, Michigan, USA[7]
Department of Business Administration, University of York, Heslington, York YO10 5DD, United Kingdom[9]
Department of Information Technology & Engineering, Sydney Met, Sydney, NSW 2000, Australia[10]
School of Computing, Ulster University, Belfast, Northern Ireland, United Kingdom[11]

***Abstract*—Generative adversarial networks (GANs) have gained popularity for their ability to synthesize images from random inputs in deep learning models. One of the notable applications of this technology is the creation of realistic videos known as deepfakes, which have been misused on social media platforms. The difficulty lies in distinguishing these fake videos from real ones with the naked eye, leading to significant concerns. This study proposes a supervised machine learning approach to effectively differentiate between real and counterfeit videos by detecting visual artifacts. To achieve this, two facial features are extracted: eye blinking and nose position, utilizing landmark detection techniques. Both features were trained on supervised machine learning classifiers and evaluated using the publicly available UADFV and Celeb-DF deepfake datasets. The experiments successfully demonstrate that the proposed method achieves a promising and superior performance, with an area under the curve (AUC) of 97% for deepfake detection in contrast to state-of-the-art methods investigating the same datasets.**

***Keywords*—Deepfake; fake videos; facial features; GAN**

I. Introduction

The current digital age has seen an unprecedented widespread use of smartphones and other such devices, making several social networking platforms popular and part of our daily lives. Statistics show that users upload billions of pictures and videos daily on such platforms. This rise of social networking platforms has also given birth to the intent of manipulating such photos and videos for several reasons, hence the concept of Deepfake.

In recent years, deep learning algorithms such as generative adversarial networks (GANs) have been able to generate fake videos and manipulate digital media semantics. In this process, two deep learning models are created, which are pitched against each other to compete. One of these models is trained on real data and then tries to create fake images. On the other hand, the other model tries to differentiate the real images from the fake ones. The model that creates fake images keeps improving and improving to such an extent that it becomes impossible for the other model to differentiate the real images from the fake ones. Algorithms like Face2Face and Deepfake availability on the internet make the propagation of digital videos more convenient. The convenience brought ease in spreading the synthesized videos on social platforms [1].

Generative models have many applications, such as image translation tasks, generating speech with manipulated fake faces, and forging a new identity that did not exist before. The inappropriate use of deepfakes on social media is alarming for the public, whether the propagated videos are trustworthy or not. Deepfake videos commonly affect public figures (celebrities, politicians), causing security and privacy threats. Generated Deep-fake is manipulated in various ways, i.e., using specific individual attributes, swapping a complete face, manipulating facial expressions, and generating a new identity face [2] in deepfake videos.

Presently, the detection method for deepfakes relies on identifying artifacts [2], such as lip synchronization with speech [3], color inconsistencies, and the unnatural representation of eye blinks, which is less compared to natural blinks [4]. The frequency of eye blinks in humans varies according to age and gender, whereas an average adult human blinks between 2 and 10 times per second [5].

Deepfake problems are generally deemed a binary classification, where the original video is classified as real and manipulated as fake. Other methods might use classifiers such as partially fake, in which a video of multiple individuals is produced, and only one person's face is altered. Previously, the detection work dealt with hand-crafted features and extraction to explore artifacts and inconsistencies. Simultaneously, current methods utilize automatic techniques to discriminate between natural and synthesized Deepfake videos.

II. Related Work

Digital manipulation of faces in images, commonly referred to as identity swap, has become increasingly prevalent due to advancements in computer graphics and deep learning

*Corresponding Author.

techniques. Significant progress has been made since the emergence of initial deepfake databases like UADFV in 2018 and more recent ones such as Celeb-DF in 2020. As a result, detecting fake videos has become more challenging, as they appear increasingly realistic.

Researchers have developed various methods for detecting deepfake videos. Detection techniques have also advanced with the improvement of the quality of fake images and videos. Table I summarizes some of the most noteworthy research in this field. While the evaluation parameters are presented in the table, it's important to note that using different evaluation metrics complicates the comparison of these methods.

TABLE I. RELATED WORK

Reference	Detection Method	Classifiers	Best Performance	Dataset
[1]	Visual Features	Logistic Regression MLP	AUC = 85.1%	Own
			AUC = 78.0%	FF++/DFD
			AUC = 66.2%	DFDC Preview
			AUC = 55.1%	Celeb-DF
[2], [3]	Face Warping Features	CNN	AUC = 97.7%	UADFV
			AUC = 93.0%	FF++/DFD
			AUC = 75.5%	DFDC Preview
			AUC = 64.6%	Celeb-DF
[4]	Mesoscopic Features Steganalysis Features Deep learning features	CNN	Acc. ≃ 94.0% Acc. ≃ 98.0% Acc. ≃ 100.0%	FF++ (DeepFakes, LQ) FF++ (DeepFakes, HQ) FF++ (DeepFakes, RAW)
			Acc. ≃ 93.0% Acc. ≃ 97.0% Acc. ≃ 99.0%	FF++ (FaceSwap, LQ) FF++ (FaceSwap, HQ) FF++ (FaceSwap, RAW)
[5]	Deep learning features	Capsule Networks	AUC = 61.3%	UADFV
			AUC = 96.6%	FF++/DFD
			AUC = 53.3%	DFDC Preview
			AUC = 57.5%	Celeb-DF

[6]	Deep learning features	CNN + Attention mechanism	AUC = 99.4%
			EER = 3.1%
				DFFD

[9]	Deep learning features	CNN	Precision = 93.0%
			Recall = 8.4%
				DFDC Preview

[10]	Image + Temporal features	CNN + RNN	AUC = 96.9%
			AUC = 96.3%
				FF++ (DeepFakes, LQ)
				FF++ (FaceSwap, LQ)

[11]	Image + Temporal features	Dynamic Prototype Network	AUC = 99.2%
			AUC = 71.8%
				FF++ (FaceSwap, HQ)
				Celeb-DF

[12]	Eye blinking features	LRCN	AUC = 99.0%	UADFV
[13]	Eye blinking features	Distance	Acc. = 87.5%	Own

[14]	-	CapsNet	AUC = 76.8%
			AUC = 86%
				DFDC-P
				Celeb-DF

Recent advances in AI and deep learning have led to the creation and proliferation of fake digital content, including fake footage, images, audios, and videos [6]. In recent years, several machine learning-based tools have made it relatively easy to create realistic face swap videos called deepfakes [7]. These deepfakes are modern self-manipulation methods that allow users to swap identities in a single video. This negative side of machine learning is creating new challenges for the general population, as people with bad intentions alter the truth and compromise people's trust. Literature review shows that current solutions to tackle the problem lack the ability to identify the source of such fake digital media. One of the most widely used biometric authentication methods is fingerprints, which are now used in smartphones, tablets, and laptops. However, this authentication method can be easily faked [8]. A statistical feature extraction and comparative analysis method is used to determine the best features.

The study in [9] proposed a framework in which they extract features from CCTV cameras at runtime using spatial and temporal domains and build a robust and discriminative feature rendering of each sequence. In their methodology, the first phase, they are using a multi-loss function to increase inter-class variance and reduce intra-class difference. In the next phase, features are aggregated frame-wise, and temporal information is extracted from videos. In the last stage, weighted coefficients are combined, and the appearance description of the pedestrian is acquired. Interestingly, although the compression ratio used during training differs from that used for the test videos, the detector performs excellently on compressed videos.

In study [10], the authors examine methods based on GAN discriminators to detect Deepfake videos. They trained a GAN and extracted the discriminator as a standalone module to identify Deepfakes using MesoNet as a benchmark. They tested various discriminator designs on various datasets to see how the discriminator's efficacy differs depending on the setting and training approach. Using ensemble approaches, they presented a methodology to improve the efficacy of a cluster of GAN discriminators. These findings reveal that GAN discriminators do not function well enough on videos from unverified sources, even when enhanced with ensemble approaches.

Li et al. [11] presented a deep-neural network (DNN) scheme to expose Deepfake videos. Physiological signals, such as eye blinks, are not well explained in the generated Deepfake videos. Blinking refers to the eye's open-close-open movement, which varies in humans according to age and gender. The Deepfake videos usually carry fewer indications of the natural blinking pattern than the original blinks. The authors trained VGG16 with a long-short-term memory (LSTM) recurrent neural network (RNN) on the dataset having an open eye state. However, no dataset has been adequately designed to detect this feature; therefore, samples have been taken from the CEW (Closed Eye in a Wild dataset) and EBV (Eye Blink Video). The authors further investigate DNN to detect artifacts. The idea behind identifying artifacts was that the recent Deepfake algorithms produce low-resolution and limited-quality images, which leave some distinctive artifacts when mapped back to the source video. They applied Dlib models that are used to detect the facial landmarks of a person’s face. In case of multiple resolution cases, the face is aligned and smoothened by applying a Gaussian blur, and the face image is then mapped with Affine Wrap to simulate the artifacts. The CNN model was trained to detect the existence of artifacts in the face region and surroundings. The presented model was compared with the four states of the models, VGG16, ResNet50, ResNet101, and ResNet152. The model tested over UADFV and Deepfake TIMIT databases shows promising results regarding these databases' state-of-the-art features. A study in [2] surveyed the face manipulation techniques and artifacts. They proposed a methodology to identify artifacts like eye color, reflection, and missing details in teeth formation and eye area. In this regard, they have proposed a novel approach using Bi-granularity artifacts (BiG-Arts).

Yang et al. [12] presented a scheme to detect Deepfakes through head movement. The Deepfake images are created by interlacing fake face images with the actual image, and while doing so, the process leaves artifacts in the 3D head position. Analyzing 3D head estimation and inconsistency, classification can be performed to detect the modification. They proposed an SVM classifier, and the results were evaluated for each UADFV dataset. The offered method was evaluated using the frames of UADFV. Hsu et al. [13] presented a common fake feature network (CFFN) model alongside pairwise learning to detect Deepfake. A two-phase procedure was followed for feature extraction. CFFN used Siamese architecture, and classification was performed through CNN.

Another study in [14] presented an optical flow scheme based on a convolutional neural network (CNN). The proposed approaches detect the Deepfake on a single video frame, where the optical flow approach catches the inter-frame dissimilarities. An experiment is performed on VGG16 and ResNet50, and results are tested over the Face-Forensic++ dataset, showing promising performance. Likewise, a study in [15] detects artifacts' presence among real and Deepfake by examining the GAN pipeline. The proposed detection scheme chooses color feature as a detection parameter and a pre-trained machine learning SVM classifier. The method achieved 70% accuracy when evaluated over a dataset named NIST MFC2018.

The classical deepfake detection methods use a convolutional neural network (CNN) to detect real or fake images based on a dataset of still images. They are unable to perform the detection of videos. Results show that sequential features can be quite crucial for detecting deepfake videos, as some of these features can be detected in videos only, e.g., it has been observed that in deepfake videos, the eye blink rate is much lower than in real videos [14]. Fig. 1 shows the taxonomy of various methods, techniques, and classifications applied to deepfake detection methods.

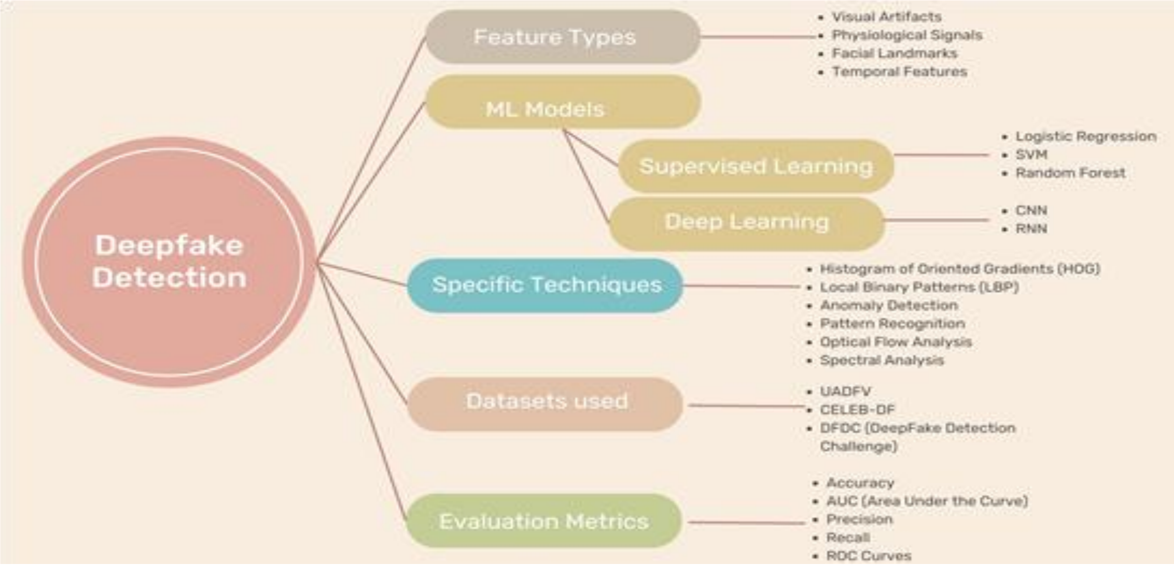

Fig. 1. Classification of deepfake detection methods.

Whereas Table II classifies and summarizes state-of-the-art methods and approaches in a hierarchical manner. Starting with main categories, that includes the feature set, followed by machine learning models and techniques, dataset used, and evaluation metrics applied. In the second column it segregates the subtypes of each category and consequently the description of each category.

TABLE II. SUMMARY OF TECHNIQUES APPLIED TO DEEPFAKE DETECTION

Main Category	Subcategory	Description
Feature Types	Visual Artifacts [1], [4], [16]	Color Inconsistencies: Abnormal color dissimilarities or mismatches in different areas of the face.
		Blurring and Boundaries: Blurred edges around the face or other areas of the image where the forged overlay blends with the real background.
		Texture and Lighting: Variations in texture and lighting that do not match the neighbouring context.
	Physiological Signals [10], [12], [17], [18]	Eye Blinking: Abnormal blinking patterns that are either too frequent or too infrequent as compared to natural human behaviour.
		Lip Sync: Reduced synchronization between lip movements and the audio, demonstrating that the speech may be dubbed or artificially generated.
		Head Movements: Unnatural head movements that do not align with the rest of the body or the environment.
	Facial Landmarks [19]–[21]	Facial Expression Analysis: Analyzing irregularities in facial expressions, which might not align logically.
		Landmark Deformation: Distortions in the placement of facial landmarks such as eyes, nose, and mouth during various expressions or movements.
	Temporal Features [21], [22]	Frame-to-Frame Consistency: Checking for irregularities across successive frames of a video that may show fiddling.
		Optical Flow: Analyzing the motion patterns in video sequences to identify irregularities.
Machine Learning Models	Supervised Learning [16], [22]	Logistic Regression: A simple, binary classification algorithm used for initial investigation.
		Support Vector Machines (SVM): Effective for high-dimensional data.
		Random Forests: An ensemble learning method that combines several decision trees for enhanced accuracy.
	Deep Learning [16], [20]	Convolutional Neural Networks (CNNs): Excellent for extracting spatial features from images and videos.
		Recurrent Neural Networks (RNNs): Appropriate for capturing sequential dependencies in video sequences.
		Capsule Networks: Capture spatial hierarchies and relationships.
		Attention Mechanisms: Focus on the most pertinent parts of the input data.
	Hybrid Models[23]	CNN + RNN: Combining spatial and temporal features for a detailed analysis.
		Ensemble Methods: Using multiple models to augment detection performance by using their combined strengths.
Specific Techniques	Handcrafted Feature Extraction [24]	Histogram of Oriented Gradients (HOG): Detecting particular facial features by examining gradients and orientations in the image.
		Local Binary Patterns (LBP): Analyzing texture by associating each pixel with its neighbors.
	Deep Learning-Based Extraction [25]	Pre-trained Networks: Utilizing networks like VGG16, ResNet, which have been pre-trained on large datasets, for feature extraction.
		GAN Discriminators: Using the discriminator component of GANs to identify fake content.
	Statistical Analysis [26]	Anomaly Detection: Identifying outliers in facial features and movements that do not follow the likely patterns.
		Pattern Recognition: Identifying and analyzing particular patterns in physiological signals and visual artifacts.
	Signal Processing [26]	Optical Flow Analysis: Identifying motion discrepancies within the video.
		Spectral Analysis: Analyzing frequency components of facial movements to distinguish anomalies.
Datasets Used	Publicly Available Datasets [27]	UADFV: Contains real and fake videos specifically created for deepfake detection research.
		Celeb-DF: A large-scale dataset with high-quality deepfake videos.
		DFDC (DeepFake Detection Challenge): A diverse dataset from the DeepFake Detection Challenge, containing numerous deepfake videos for benchmarking.
Evaluation Metrics	Accuracy [28], [29]	The overall percentage of appropriately classified instances (both real and fake).
	AUC (Area Under the Curve) [28]	Measures the capability of the model to differentiate between classes, providing insight into its performance across different thresholds.
	Precision and Recall [30]	Precision: The ratio of true positives to predicted positives, demonstrating the accuracy of the positive predictions.
		Recall: The ratio of true positives to actual positives, demonstrating the capability to identify all positive instances.
	F1-Score [29]	The harmonic mean of precision and recall, providing a stable measure of the model's performance.
	Receiver operating characteristic (ROC) Curves [31]	Receiver Operating Characteristic curves, which visualize the trade-off between true positive rate and false positive rate across different thresholds.

A. Contribution

The contribution of this research paper is as follows:

- This research work extracts facial feature eye blink using a real-time blinking method called Eye Aspect Ratio (EAR). To detect the second facial feature, the nose's position, we utilize a pre-trained machine learning Haar cascade classifier with 97% accuracy for efficient detection.
- To counter this recent threat, this research work used a supervised learning method to identify the deepfake videos from the real ones. The results show that the proposed methodology is quite efficient in distinguishing the real videos from the deepfake videos.
- The presented model is evaluated using pre-trained Eye Aspect Ratio (EAR). The proposed scheme detected a blinking ratio of 34.1/ min in real video and 3.4/ min in fake video.

This research paper is organized as follows: Related work is given in Section II. Section III discusses the proposed methodology, Section IV presents the results and discussion, and Section V discusses the conclusion.

III. Methodology

This section proposes a method to detect Deepfake videos by extracting facial features. Fig. 2 depicts the proposed methodology. Deepfake video detection differs from image detection, as manipulation is carried out frame by frame and contains temporal characteristics. We are using two features to extract the modification:

- Eyeblink.
- Nose position.

Generally, it is observed that individual Deepfake videos show abnormal eye blinking, which is less frequent compared to normal human blinking behavior. An adult human can blink in 2 to 10 seconds, and each blink consumes 0.1 and 0.4 seconds. Every individual has different blink patterns concerning the open and closed state of the eye. Typically, Deepfake methods are trained on images that possess an open eye state, so it is difficult for Deepfake methods to generate synthesized videos with normal blinking behavior. Eyeblink contains temporal dependency and is expected to appear as temporal artifacts across the frame as manipulation is performed over frame-by-frame sequence. A face landmark detector is used in the proposed study to detect the face and the eye's open /closed state using the DLIB model. Face landmarks locate the whole set of feature points of the face, like lips, eyes, nose, and contour, which can be detected from the face area. The eye blink's first feature is calculated by computing the Eye Aspect Ratio (EAR) in each video frame. EAR of the person depends on the eye's landmark locations, and it is a constant value when the person's eye state is opened and falls to 0 when the eye is closed [13].

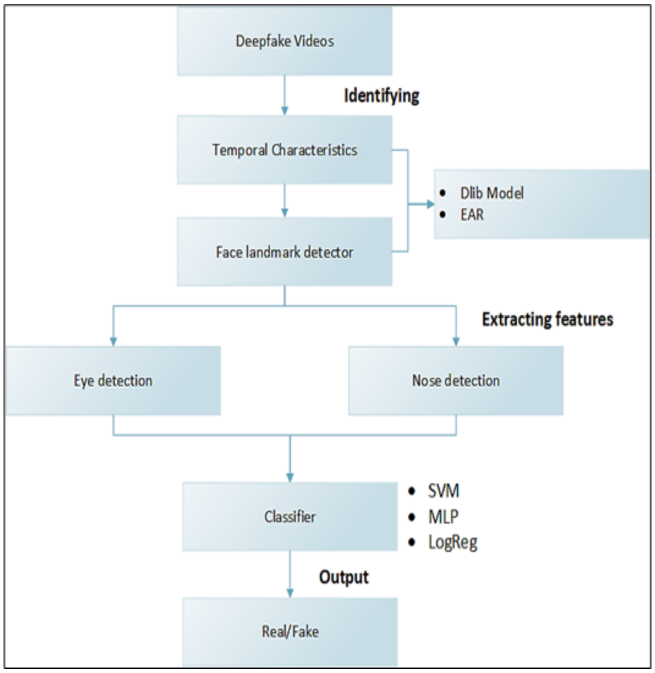

Fig. 2. Proposed methodology.

The following formula calculates EAR:

$$EAR_i = \frac{||\mathcal{P}_2-\mathcal{P}_6||+||\mathcal{P}_3-\mathcal{P}_5||}{2||\mathcal{P}_1-\mathcal{P}_4||} \tag{1}$$

Where P_i (i=1, 2…, 6) are the eye's landmark points, we do not know the deep learning algorithms for eye blink detection. Fig. 3 and Fig. 4 demonstrate the eye blink and nose positions identified in real and fake videos. Overview of the workflow for detecting eye blinks in real and fake videos. The Haar cascade classifier is used for the second feature, i.e., nose; the machine learning approach has rapid detection capability with approximately 95% accuracy. We are using a multiscale Haar cascade Classifier to detect face and nose position in the video stream. We are using the first data set to extract Features, such as UADFV, which contains both real and fake videos. For example, UADFV features the eye blink and nose position extracted from the data set with 98 videos. The dataset collected 49 real and 49 fake videos with 32,752 frames. These videos are generated from the DNN model with FakeApp and are 2 to 44 seconds long, with an average of 11.26 seconds.

Tables III and Table IV describe the UADFV and Celeb-DF datasets, respectively. Moreover, they show the number of videos in each dataset with real and fake counterparts, their average length, and frame rate (frames per second).

TABLE III. Dataset 1 UADFV Specification

Dataset	Number of videos	Average length	Frames per second
REAL	49	11.26	28FPS
FAKE	49	11.26	28FPS

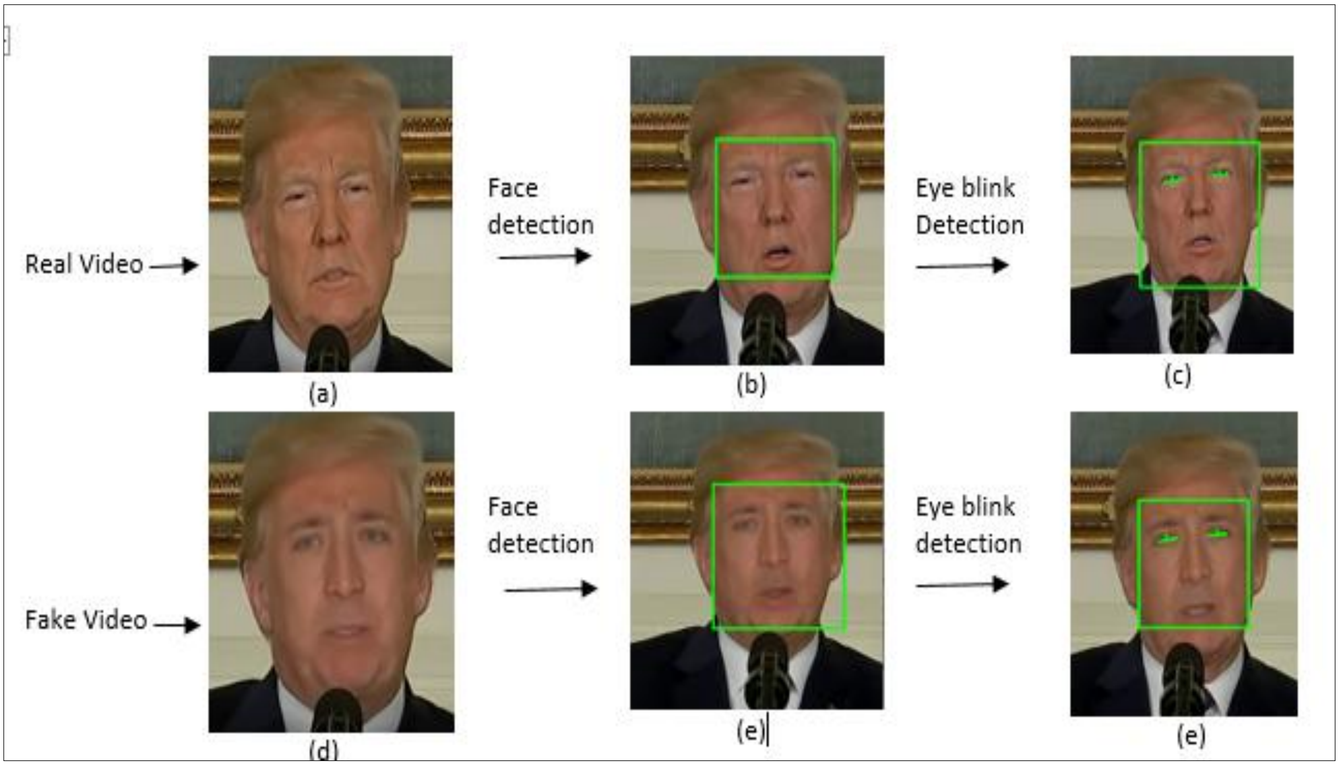

Fig. 3. Overview of workflow detecting eye blink in real and fake videos.

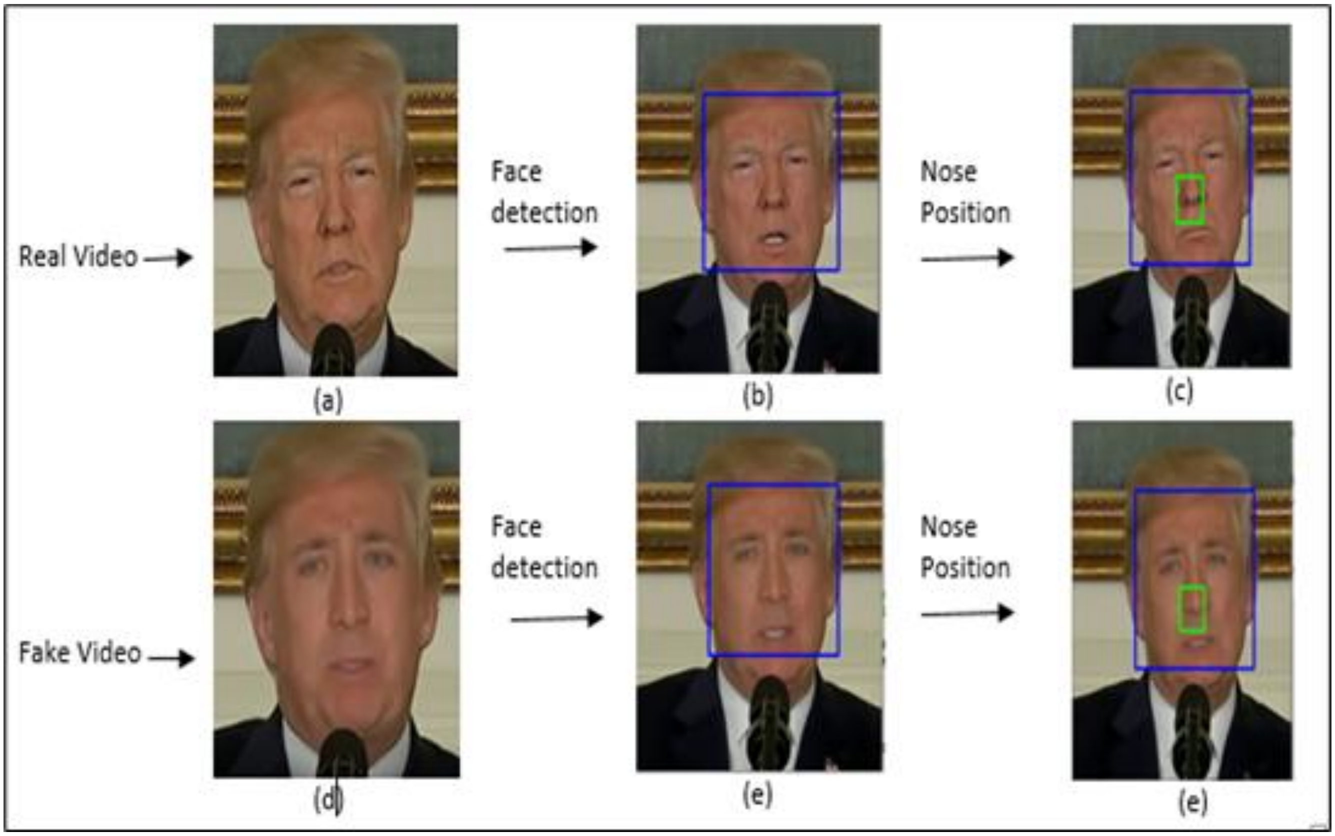

Fig. 4. Overview of workflow detecting nose positions in real and fake videos.

Celeb-DF is a new dataset proposed by Li et al. [14] using refined generating algorithms with improved quality videos and less visible artifacts. The data set used for this experiment is Celeb-DF.

TABLE IV. DATASET 2 CELEBDF SPECIFICATION

Dataset	Number of videos	Average length	Frame per second
REAL	408	13 sec	30FPS
FAKE	795	13 sec	30FPS

For a more in-depth analysis, 50 YouTube-real videos and 50 Celeb-synthesis videos datasets are also considered in addition to the two datasets mentioned above for the performance analysis of classifiers.

IV. RESULTS AND DISCUSSIONS

The extracted facial features are trained and evaluated on the datasets for three different classifiers. In the next step, a classifier was trained to distinguish between real and fake videos. Fig. 5 shows the receiver operating characteristic (ROC) curve of the proposed classifiers for the proposed feature and classifiers.

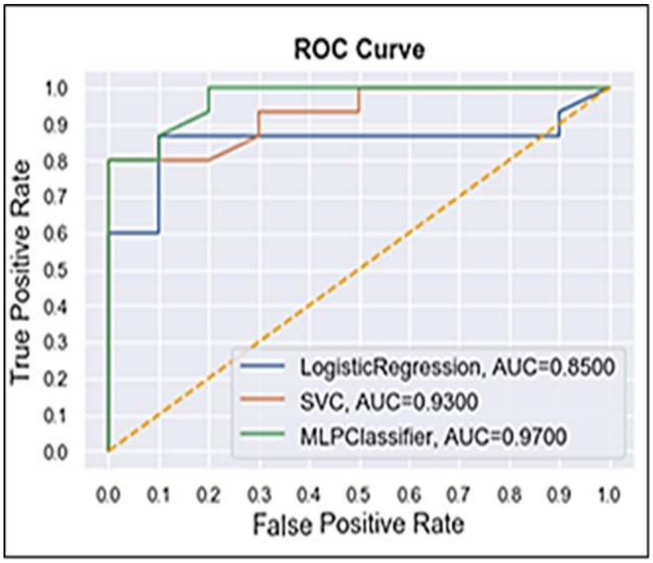

Fig. 5. ROC Comparison for UADFV dataset.

Fig. 6 shows the classification ROC curve analysis on the proposed datasets on extracted facial features for different classifiers. We refer to the supervised machine learning classifiers, SVM, Logistic Regression, and MLP.

The results show in Table V that the classifiers can adequately distinguish between the real and fake sets of videos. The Celeb-DF dataset's performance is low compared to the UADFV because it contains fewer visible artifacts that are difficult to identify. The proposed performance-based method is compared with other methods on the same datasets: UADFV and Celeb-DF.

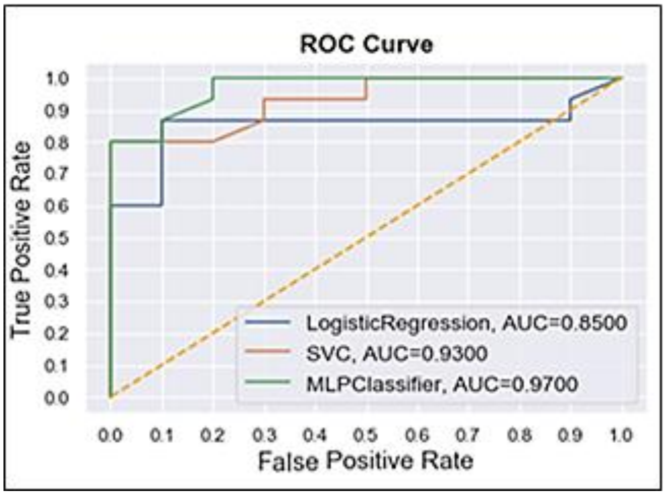

Fig. 6. ROC comparison for CELEBDF dataset.

TABLE V. PERFORMANCE COMPARISON OF MODELS FOR TWO DATASETS

Methods/Classifiers	UADFV	Celeb-DF
HeadPose-SVM [1]	89.0	54.8
VA-LogReg	70.2	48.8
VA-MLP [2]	54.0	46.9
FWA-CNN [3]	97.4	53.8
iCaps-Dfake [14]	-	86%
RNN [31]	-	73.41%
Capsule network [5]	-	57.5%
CNN[19]	84.3	54.8
SVM MLP LogReg	93.0	64.0
	97.0	75.0
	85.0	72.0

Detection is performed on the feature "Head Pose" [12] using an SVM classifier. The classifiers generated AUC (%) performance of 89.0 on UADFV and 54.8% on the CelebDF dataset. Visual artifacts like eyes, teeth, and facial texture are identified by applying Logistic Regression and MLP. The methods' AUC performances are 70.2% and 48.8% on the datasets, respectively [3]. Face Warping artifacts [11] are classified by using a CNN. The classifiers show the AUC performance of 97.4% for the UADFV dataset and 53.8% for the Celeb-DF dataset. For iCaps-Dfake method [14], performance on Celeb-DF dataset is 86%. For the RNN [31] based classifier, the performance of the Celeb-DF dataset was 73.41%. For the capsule network [5], the performance of the Celeb-DF dataset is 57.5%. For CNN [19], performance on the UADFV dataset is 84.3%, and performance on the Celeb-DF is 54.8%. This research used SVM, MLP, and Logistic Regression, which shows strong performance for deepfake detection on UADFV and Celeb-DF datasets. MLP exhibited the best accuracy among the three methods, i.e., 97% on UADFV and 75% on Celeb-DF. SVM achieved 93% on UADFV and 64% on Celeb-DF. For LogReg, 85% accuracy is achieved on the UADFV dataset and 72% on Celeb-DF.

V. Conclusion

The proposed method detects fake blinks in the UADFV dataset at a rate of 0.6 per 60 seconds and real blinks at 7.4 per 60 seconds. Similarly, in the Celeb-DF dataset, we observe a rate of 9.8 real blinks and 5.04 fake blinks per minute. Given that the average human blink rate is around 10 blinks per minute, the generated videos fall below this standard. Additionally, we note that the nose position in fake videos from the UADFV dataset deviates from its original position more than in the Celeb-DF dataset. Both features achieve a higher performance with an Area Under the Curve (AUC) of 97% on UADFV and 75% on Celeb-DF. For future work, there are several directions we plan to explore to enhance our current findings. We aim to investigate new deep learning architectures for more effective results. Furthermore, we will continue searching for facial artifacts and other physiological signals often overlooked in synthesized videos.

Authors' Contributions

Conceptualization, Jamaluddin Mir and Atta Rahman; Data curation, Dhiaa Musleh; Formal analysis, Gohar Zaman, Asiya Abdus Salam and Jamal Alhiyafi; Funding acquisition, Mustafa Jamal Gul, Jamal Alhiyafi and Aghiad Bakry; Investigation, Farhan Ali Dhiaa Musleh, Jamal Alhiyafi and Mohammed Gollapalli; Methodology, Ayesha Aslam, Jamaluddin Mir, Gohar Zaman and Atta Rahman; Resources, Mohammed Gollapalli; Software, Ayesha Aslam and Aghiad Bakry; Supervision, Jamaluddin Mir; Validation, Asiya Abdus Salam, Dhiaa Musleh and Aghiad Bakry; Visualization, Mohammed Gollapalli; Writing – original draft, Ayesha Aslam; Writing – review & editing, Gohar Zaman, Mustafa Jamal Gul, Atta Rahman, and Farhan Ali. Farhan Ali and Atta-Rahman have equal contributions.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets investigated in the study are available online in a public repository.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," Proc. - 2019 IEEE Winter Conf. Appl. Comput. Vis. Work. WACVW 2019, no. 1, pp. 83–92, 2019, doi: 10.1109/WACVW.2019.00020.

[2] H. Chen, Y. Li, D. Lin, B. Li, J. Wu, "Watching the BiG artifacts: Exposing DeepFake videos via Bi-granularity artifacts," Pattern Recognition, vol. 135, 109179, 2023. https://doi.org/10.1016/j.patcog.2022.109179.

[3] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 3204–3213, 2020, doi: 10.1109/CVPR42600.2020.00327.

[4] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manip-ulated facial images," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1–11.

[5] H. H. Nguyen, J. Yamagishi and I. Echizen, "Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 2307-2311, doi: 10.1109/ICASSP.2019.8682602.

[6] H. R. Hasan and K. Salah, "Combating Deepfake Videos Using Blockchain and Smart Contracts," in IEEE Access, vol. 7, pp. 41596-41606, 2019, doi: 10.1109/ACCESS.2019.2905689.

[7] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6, doi: 10.1109/AVSS.2018.8639163.

[8] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition, 2020, pp. 5781–5790.

[9] B. Dolhansky et al., "The deepfake detection challenge (dfdc) dataset," arXiv Prepr. arXiv2006.07397, 2020.

[10] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," Interfaces (GUI), vol. 3, no. 1, pp. 80–87, 2019.

[11] Y. Li, M. C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking," 10th IEEE Int. Work. Inf. Forensics Secur. WIFS 2018, 2019, doi: 10.1109/WIFS.2018.8630787.

[12] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2019-May, pp. 8261–8265, 2019, doi: 10.1109/ICASSP.2019.8683164.

[13] Hsu, C.-C.; Zhuang, Y.-X.; Lee, C.-Y. Deep Fake Image Detection Based on Pairwise Learning. Appl. Sci. 2020, 10, 370. https://doi.org/10.3390/app10010370.

[14] L. Trinh, M. Tsang, S. Rambhatla, and Y. Liu, "Interpretable and trustworthy deepfake detection via dynamic proto-types," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 1973–1983.

[15] T. Jung, S. Kim, and K. Kim, "Deepvision: Deepfakes detection using human eye blinking pattern," IEEE Access, vol. 8, pp. 83144–83154, 2020.

[16] S. S. Khalil, S. M. Youssef, and S. N. Saleh, "iCaps-Dfake: An integrated capsule-based model for deepfake image and video detection," Futur. Internet, vol. 13, no. 4, p. 93, 2021.

[17] M. A. Sahla Habeeba, A. Lijiya, and A. M. Chacko, "Detection of Deepfakes Using Visual Artifacts and Neural Network Classifier BT - Innovations in Electrical and Electronic Engineering," 2021, pp. 411–422.

[18] J. Cech and T. Soukupova, "Real-Time Eye Blink Detection using Facial Landmarks," Cent. Mach. Perception, Dep. Cybern. Fac. Electr. Eng. Czech Tech. Univ. Prague, pp. 1–8, 2016.

[19] 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), 2018, pp. 1–7.

[20] J. Park, H. E. Ahn, L. H. Park, and T. Kwon, "Robust Training for Deepfake Detection Models Against Disrup-tion-Induced Data Poisoning BT - Information Security Applications," 2024, pp. 175–187.

[21] S. Ganguly, S. Mohiuddin, S. Malakar, E. Cuevas, and R. Sarkar, "Visual attention-based deepfake video forgery detec-tion," Pattern Anal. Appl., vol. 25, no. 4, pp. 981–992, 2022, doi: 10.1007/s10044-022-01083-2.

[22] A. Deshmukh and S. B. Wankhade, "Deepfake Detection Approaches Using Deep Learning: A Systematic Review BT - Intelligent Computing and Networking," 2021, pp. 293–302.

[23] A. Al-Adwan, H. Alazzam, N. Al-Anbaki, and E. Alduweib, "Detection of Deepfake Media Using a Hybrid CNN–RNN Model and Particle Swarm Optimization (PSO) Algorithm," Computers, vol. 13, no. 4, pp. 1–16, 2024, doi: 10.3390/computers13040099.

[24] S. Suratkar and F. Kazi, "Deep Fake Video Detection Using Transfer Learning Approach," Arab. J. Sci. Eng., vol. 48, no. 8, pp. 9727–9737, 2023, doi: 10.1007/s13369-022-07321-3.

[25] B. Kaddar, S. A. Fezza, W. Hamidouche, Z. Akhtar, and A. Hadid, “HCiT: Deepfake Video Detection Using a Hybrid Model of CNN features and Vision Transformer,” in 2021 International Conference on Visual Communications and Image Processing (VCIP), 2021, pp. 1–5. doi: 10.1109/VCIP53242.2021.9675402.

[26] O. A. H. H. Al-Dulaimi and S. Kurnaz, “A Hybrid CNN-LSTM Approach for Precision Deepfake Image Detection Based on Transfer Learning,” Electron., vol. 13, no. 9, pp. 1–22, 2024, doi: 10.3390/electronics13091662.

[27] Y. Xu, K. Raja, L. Verdoliva, and M. Pedersen, “Learning Pairwise Interaction for Generalizable DeepFake Detection,” in 2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), 2023, pp. 1–11. doi: 10.1109/WACVW58289.2023.00074.

[28] S. Mathews, S. Trivedi, A. House, S. Povolny, and C. Fralick, “An explainable deepfake detection framework on a novel unconstrained dataset,” Complex Intell. Syst., vol. 9, no. 4, pp. 4425–4437, 2023, doi: 10.1007/s40747-022-00956-7.

[29] A. Mehra, A. Agarwal, M. Vatsa, and R. Singh, “Motion Magnified 3-D Residual-in-Dense Network for DeepFake De-tection,” IEEE Trans. Biometrics, Behav. Identity Sci., vol. 5, no. 1, pp. 39–52, 2023, doi: 10.1109/TBIOM.2022.3201887.

[30] N. M. Alnaim, Z. M. Almutairi, M. S. Alsuwat, H. H. Alalawi, A. Alshobaili, and F. S. Alenezi, “DFFMD: A Deepfake Face Mask Dataset for Infectious Disease Era with Deepfake Detection Algorithms,” IEEE Access, vol. 11, pp. 16711–16722, 2023, doi: 10.1109/ACCESS.2023.3246661.

[31] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, “Two-branch recurrent network for isolating deepfakes in videos,” in European conference on computer vision, 2020, pp. 667–684.