Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A deep learning model for FaceSwap and face-reenactment deepfakes detection

Marriam Nawaz [a], Ali Javed [a,*], Aun Irtaza [b]

[a] Department of Software Engineering, University of Engineering and Technology, Taxila 47050, Pakistan
[b] Department of Computer Science, University of Engineering and Technology, Taxila 47050, Pakistan

HIGHLIGHTS

- A novel AUFF-NET model is proposed for FaceSwap, and Face-Reenactment deepfakes detection.
- Accurate detection of deepfakes due to the ability of the model to tackle the model over-fitting.
- Improved model explainability power due to the reliable feature selection ability of model.
- Robust model due to its ability to detect deepfakes in the presence of various adversarial attacks.
- Perform well to unseen examples due to the better generalization power of the AUFF-NET model.

ARTICLE INFO

Keywords:
Deep learning
Deepfakes
FaceSwap
Face-reenactment
Inception-swish-Resnet-v2
Bi-LSTM

ABSTRACT

Recently, the higher availability of multimedia content on social websites, together with lightweight deep learning (DL) empowered tools like Generative Adversarial Networks (GANs) has caused the generation of realistic deepfakes. Such fabricated data has the potential to spread disinformation, revenge porn, initiate monetary scams, and can result in adverse immoral and illegal societal issues, etc. Hence, the accurate identification of deepfakes is mandatory to discriminate between real and manipulated content. In this work, we have presented a DL-based approach namely a unified network for FaceSwap (FS) and Face-Reenactment (FR) Deepfakes Detection (AUFF-Net). More clearly, both the spatial and temporal information from the video samples are used to detect two types of visual manipulations i.e., FS and FR. For this reason, a novel DL framework namely the Inception-Swish-ResNet-v2 model is introduced as a feature extractor for computing the information at the spatial level. While the Bi-LSTM model is utilized to measure the temporal information. Additionally, 3 dense layers are included at the last of the model structure to suggest a discriminative group of the feature vector We performed extensive experimentation on a challenging dataset namely the FaceForensic++, and attainede average accuracy values of 99.21 %, and 98.32 % for FS, or FR, respectively. Furthermore, we introduced an explainability module to show the reliable keypoints selection capability of our technique. Moreover, we have performed a cross-dataset evaluation to show the generalization power of our approach. Both the qualitative and quantitative results have confirmed the effectiveness of the suggested approach for visual manipulation categorization under the occurrence of various adversarial attacks.

1. Introduction

The accessibility to cost-effective digital devices i.e., cell phones, laptops, tablets, and digital cameras has resulted in an exponential escalation in digital data like audio, video, and images in the cyber world. In addition, the internet facility and social sites have connected people globally which enables them to share their memories. Meanwhile, the marvelous achievement in the area of Machine Learning (ML) proposes sophisticated algorithms that have the capability of manipulating audiovisual content to propagate fabricated information through social websites. The easier availability of several ML-based tools and apps [1–3] can help people to make their data more appealing and delightful. However, it also makes the conveyed information unreliable and untrusted, particularly for scenarios where such samples are utilized in examining a legal claim or inspecting a criminal case. Due to such reasons, the researchers have declared this era as "post-truth" where

* Corresponding author.
E-mail address: ali.javed@uettaxila.edu.pk (A. Javed).

https://doi.org/10.1016/j.asoc.2024.111854
Received 25 April 2023; Received in revised form 17 November 2023; Accepted 2 June 2024
Available online 15 June 2024

false or manipulated news (disinformation) is spread by malicious individuals to affect the public view. Disinformation is a malicious act of intentionally spreading altered information and has the power to affect political, social, and economic campaigns [4,5]. Deepfakes are heavily used these days to spread disinformation among the public. Deepfakes [6,7] represent the synthesized artificial intelligence-generated videos that contain the fabrication of audiovisual information [8]. Hence, given this ease of creating and spreading false information, it has interestingly become hard to differentiate real and fake data which may cause dangerous consequences [9–11].

Deepfakes is a combination of two terms deep and fake where DL-based methods are used to generate fake videos. In the future, deepfakes are predicted to be employed as a major disinformation weapon that may cause to loss of the trustworthiness of state institutes, information media, and others due to the incompetence of common people to distinguish the pristine and altered videos [12,13]. Deepfakes are creating extreme concern among the people because of their unrestricted access, and the capability for the scam, and cybercrimes [14,15]. Additionally, deepfakes also pose significant threats to democracy due to their rapid and unregulated progression in cyberspace [16]. Moreover, the employment of visual samples as proof in each area of proceedings and criminal hearings is appearing as a new medium [17]. Deepfakes techniques have several applications that can have a positive or negative impact on society. The positive implications of deepfakes can provide low-cost solutions to many problems such as deepfakes can generate voice speech for people without vocal sounds, entertainers can use such samples to show their creativity or drama, and movie producers can use them to update the scenes without reshooting them [18], however, the negative impact of deepfakes is creating more problems. Traditionally, manipulations were performed to show well-known people debatable in their followers, like, in 2017 an actor was plotted in a pornographic scenario by posting it on social sites. So, deepfakes can be utilized to affect the reputation of people for various objectives i.e. defamation of celebrities [19], blackmailing people for financial benefits, creating legislative or spiritual conflict by hitting government officials or religious preachers with manipulated content [20], etc. The most devastating effects of deepfakes can disturb election campaigns, causing war-type conditions among countries by viral a forged sample of missiles thrown to destroy the rival area [21]. It can mislead the armed experts by depicting false information for example presenting a fake bridge across the river to deceive troops, etc., [20]. Moreover, with fake data generation, deepfakes can show a vast influence on the estimation of the stock and affect the investors. With the progression of sophisticated DL algorithms these days, we can easily generate forged content with a small amount of data even with a single static image [22]. For example, a Chinese app named Zao [23,24] allows the layman to switch his face with actors and see himself acting in dramas and movie shoots. Therefore, these tools have put a serious privacy violation not only for renowned persons but for common people as well [25]. A visual depiction of a generic pipeline of deepfakes generation is given in Fig. 1. The discussed scenarios clearly indicate the severity of fabricated content that needs serious attention from the research community to combat the negative impact of deepfakes.

Visual deepfakes are broadly divided into three classes which are FS, FR, and Lip-synching, respectively. For FS-based deepfakes manipulations, the face of the target is placed over the source person to produce a manipulated sample of the target. The FS deepfakes are produced to show the target person doing the actions which are originally performed by the source identity. The main reason to produce the FS-oriented visual manipulations is to affect the fame of well-known persons like politicians and celebrities etc., [27] by showing them in controversial scenarios like non-consensual pornography [28]. In FR deepfakes, the gestures of the target identity i.e. eyes, facial expressions, and head alignments are copied [6] in a visual sample and animated according to the imitator's wish. While in lip-sync-based deepfakes, the lips alignments of a target identity are altered to make them consistent with an arbitrary audio sample [25]. The main motive of lip-sync-based deepfakes generation is to show the target person speaking something that in actuality he did not speak. The presented work is concerned to detect the FS and FR-based deepfakes. A pictorial representation of FS and FR deepfakes is demonstrated in Fig. 2.

Several techniques have been presented in the literature for the automated detection and classification of visual deepfakes, however, there are still several open challenges that require further

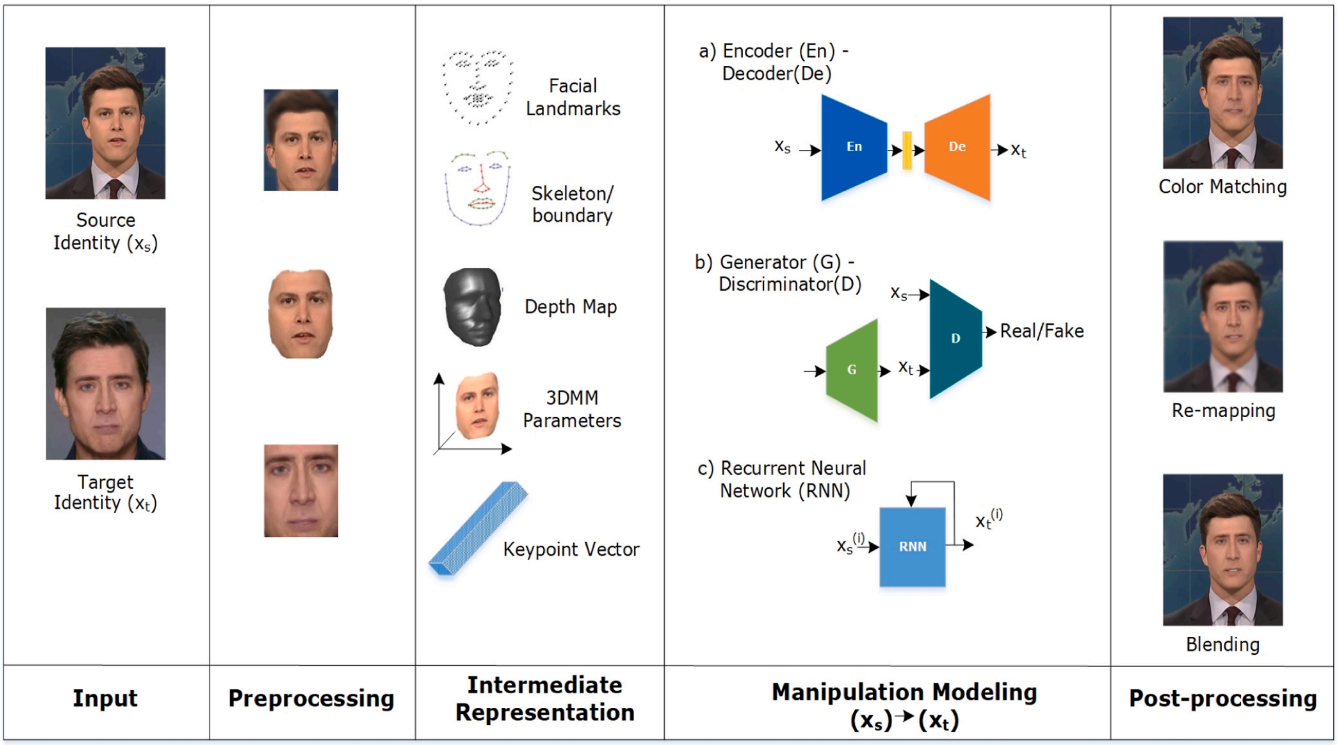


Fig. 1. A generic pipeline of deepfakes creation [26].

improvements. The approaches are not proficient enough to show better classification results well under the occurrence of several adversarial operations (i.e., compression, noise, clutter, rotation, zooming, and translation). Furthermore, the existing methods perform well for seen examples, however, unable to generalize well to untrained data. Meanwhile, the production of more truthful deepfakes data samples is deteriorating the detection performance [26]. Therefore, there is a need for such a model that can present an efficient and effective solution to visual manipulation detection. To cope with the challenges of existing works, we have proposed a DL-based approach namely the AUFF-NET. More specifically, we have extracted both the spatial and temporal information of the input samples. For this reason, a novel Inception-Swish-ResNet-v2 CNN model is introduced to capture the information at the frame level. Clearly, we have employed the Swish activation approach as an alternative to the ReLU approach in the conventional Inception-Resnet-v2 model for capturing a diverse set of sample characteristics. The employed Swish activation method uses the multiplication operation of input values by using the sigmoid function. The swish method is capable of smoothly altering the direction of negative values rather than abrupt change and allows their minimum negative range to flow through the network which assists the Inception-Swish-ResNet-v2 model to compute the complex patterns of data effectively. This activation method produces a smooth curve which eventually optimizes the model behavior by quickly converging with a small loss. While the Bi-LSTM model is utilized to compute the temporal information. Further, the addition of extra dense layers at the last of the AUFF-NET model assists it in effectively propagating a more significant group of visual key points. Lastly, the results are predicted based on both the frame level and temporal level information to make the final decision. The method is competent in classifying both FS and FR-based deepfakes with a high recall rate. The approach is competent to tackle adversarial attacks and shows effective explainability results as compared to the latest works. Further, the proposed work is evaluated in the cross-corpus scenario where it has been observed that the work has undergone some performance degradation, however, the results are still convincing in comparison to the state-of-the-art works. The following are the distinctive contributions of the presented study:

- **Unified Model:** We introduce a new technique that employs a novel spatial descriptor along with the temporal characteristics of the visual samples and is capable of detecting and classifying both FS and FR-based deepfakes.
- **Explainability:** We have visualized the computed features by using the heatmaps to present the explainability capacity of the proposed approach which helps us to show the actual manipulated portions of videos.

- **Generalizability:** The cross-dataset validation is performed where the proposed model is tested on the unseen examples to show the generalization ability of our framework and proved with the evaluations that the presented solution is effective to unseen examples due to its ability to better tackle the model over-fitting problem.
- **Robustness:** Several video adversarial operations like compression, blurring, noise, rotation, and size alterations are added at the test

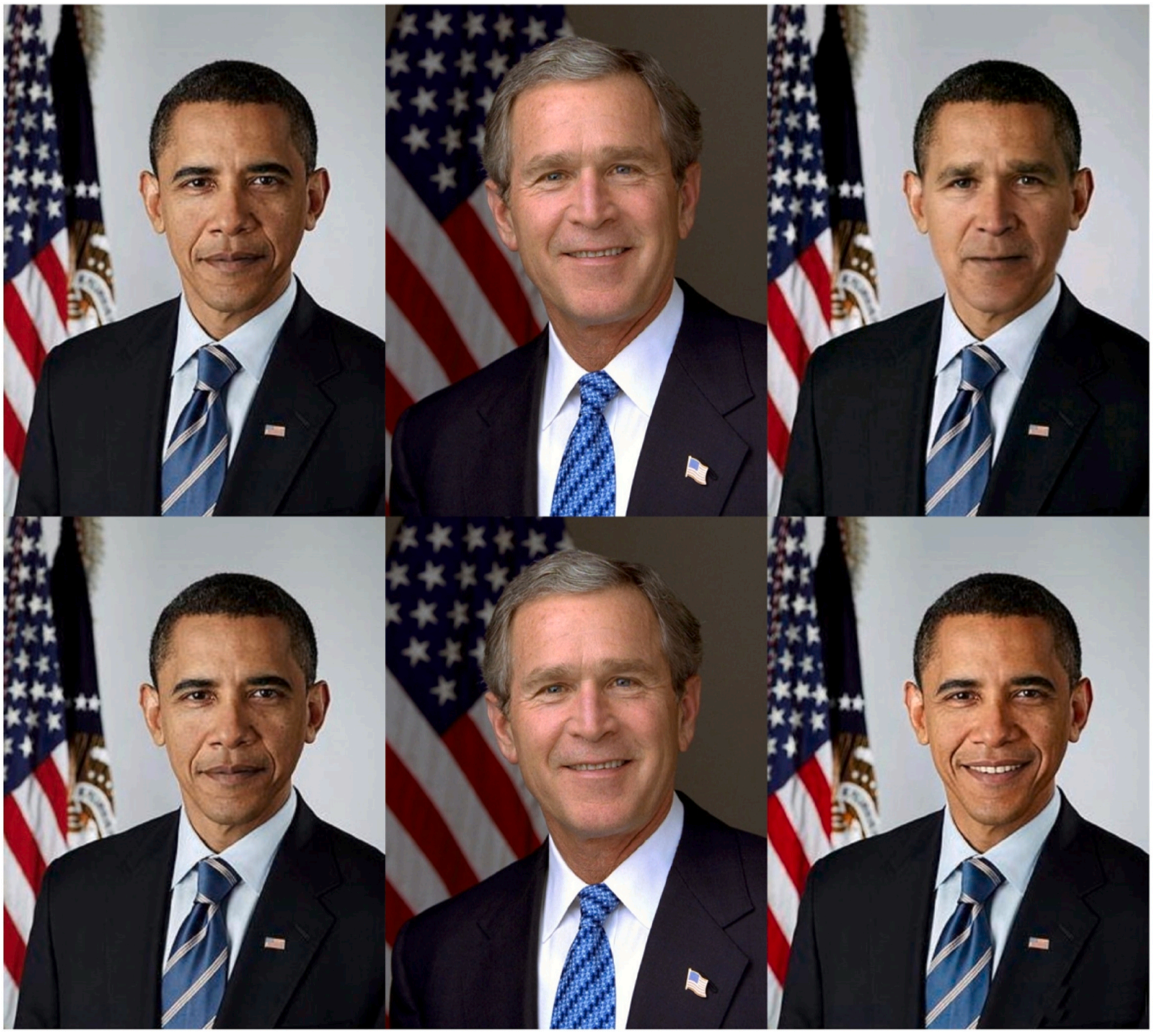


Fig. 2. Visual depiction of deepfakes, first row: FS deepfakes, second row: FR deepfakes.

time to check the robustness of the proposed approach and have proved the effectiveness of our model to such video alterations due to the distinctive features extraction empowerment of the AUFF-NET.

The later manuscript follows the given outline: Section 2 contains an analysis of work performed in history for visual manipulation classification, while an in-depth elaboration of the suggested framework is given in Section 3. Section 4 shows a comprehensive explanation of the experimental evaluation while Section 5 comprises the conclusion and future work.

2. Related work

In this part of the manuscript, we have accomplished a critical investigation of historical approaches used for the detection of Face-Swap and Face-Reenactment-based deepfakes. The approaches for Face-swap and Face-Reenactment detection techniques are generally categorized either as handcrafted-based approaches or DL-based methods.

2.1. Handcoded features

Initially, scientists tested the ML frameworks generally employed for image forensic analysis. Zhang et al. [29] discussed a conventional ML method focused on recognizing the FaceSwap-based visual manipulations. Initially, the Speeded Up Robust Features (SURF) approach was used to compute the features and later used these features for the SVM learning to perform the categorization job. The technique shows better image-based deepfakes detection, however, unable to attain effective results for video-based deepfakes samples. An approach for revealing Face-swap deepfakes was discussed in [30] by computing the alignment of the 3D head poses by calculating 2D facemask key points. The estimated dissimilarity found among the head orientations was passed as a key vector. In the next step, the extracted visual characteristics were applied to perform the SVM training to classify the pristine and manipulated content. The framework [30] performs better for deepfakes classification, however, unable to work well for blurred samples. Guera et al. [31] discussed a method to locate the forensic changes found in the visual content by applying multimedia stream descriptors [32] for feature computation. Later, the SVM method was learned over the extracted sample characteristics to discriminate the original and fabricated samples. The work [31] is effective in deepfakes recognition, however, not proficient in tackling video re-encoding operations. Another approach was proposed in [33] where the frequency of the heartbeat from the examined samples was computed to locate the alterations made within the videos. The extracted set of key features was passed as input to tuning the SVM and CNN approaches. The approach [33] enhanced visual modification identification performance, however, at the charge of an increased computing burden. Jung et al. [34] presented a model to identify Face-swap-based deepfakes by locating abnormal eye patterns from the input videos. The work employed the Fast-HyperFace [35] and EAR methods to compute eye blinking patterns from the eyes. Then, the truthfulness verification technique was used by applying the variant of eye flickers with time to locate the original and altered visual samples. The work demonstrated in [34] shows better deepfakes identification accuracy, however, this method is not generalized well for persons with mental sickness which causes irregular eye blinking patterns.

Amerini et al. [36] discussed a study to recognize Face-Reenactment-based visual fabrications. The model called the PWC-Net [37] was used to calculate the optical flow fields [38] of all suspected samples [37]. The extracted characteristics were passed for a DL-oriented predictor to classify the actual and forged data. This technique [36] shows enhanced deepfakes recognition accuracy, however, the model needs evaluation over a complex dataset. Some researchers have presented techniques that can locate more than one type of deepfakes. One such work was presented in [39] to discriminate the original and fake data by using facial landmarks i.e. eye colors, and missing reflections. The computed keypoints were employed for logistic regression and MLP classifier training, to differentiate the altered and pristine sample. The framework performs effectively for deepfakes recognition; however, not proficient in images with closed eyes or invisible teeth. Agarwal et al. [16] discussed a method to reveal the deepfakes where the facial areas are located with the help of the OpenFace2 [40] software. The estimated visual characteristics were passed for the SVM algorithm learning to differentiate the actual and altered samples. The work attains improved recognition results, however, does not perform well in cases where the subject is looking away from cameras.

2.2. Deep features

Numerous approaches have used DL methods for Face-swap-based deepfakes identification. Li et al. [41] introduced an algorithm for multimedia forensic analysis. Initially, face areas from the input videos were located by using the dlib library [42]. Then, several DL methods namely ResNet-with 50, 101, and 152 layers, along with the VGG-16 were applied for computing the visual characteristics for forged content classification. The technique [41] works well for deepfakes detection, however, lacks to perform well for highly compressed visual samples. Another framework was discussed in [43] where a CNN framework was used to compute the set of deep features at the frame level of videos. In the next step, the RNN model was used to compute the visual sequence analysis with time to identify the original and modified video samples. The approach [43] attains better deepfakes identification performance, however, workable with samples of small length. A framework was elaborated in [44] to reveal the visual fabrications by tracking the irregular eye blinking patterns. A CNN/RNN model was used to specify the missing eye blinking from the examined samples. The work [44] performs effectively in identifying the visual modifications, however, unable to perform well for visual samples having a person with closed eyes. Montserrat et al. [45] suggested a methodology to reveal Face-swap-based deepfakes detection. At the start, the human faces for the input samples were located by using a Multi-task CNN (MTCNN) [46]. In the next step, a CNN framework was used over the detected faces for deep feature computation. In the last step, an RNN method was applied to identify the altered visual content. The technique in [45] shows effective deepfakes categorization results, however, not proficient to acquire predictions from the key points in several video frames. Another DL-based approach was introduced in[47] where a VGG-16 approach was used for deep key points estimation. Then, the LSTM method was used for temporal sequence examination to detect real and fake content. This technique [47] advances visual modification identification performance, however, suffering from high processing complexity. Agarwal et al. [48] discussed a technique to identify Face-swap-based alteration by joining the facial features along with behavioral biometrics key points. The VGG-16 model along with the encoder-decoder framework was used for features computation and manipulation detection. This work presented in [48] shows better performance for unseen examples, however, the method is not robust to lip-synch-oriented visual manipulations. An approach was proposed in [49] to recognize video-based alterations by estimating the heart rate of persons. In the first step, the heart rate was measured by using several methods namely skin color alteration [50], average optical intensity [51], and Eulerian video magnification [52]. The calculated features were employed for Neural Ordinary Differential Equations (Neural-ODE) [53] training to distinguish the pristine and forged samples. The approach [49] is robust to deepfakes identification but with an increased computing burden. Kolagati et al. [54] proposed an approach for Face-swap-based deepfakes detection. Initially, the Dlib library was used to compute the landmarks of the facial region. Then, a set of deep features was computed by employing a CNN model. Both landmark and deep features were combined to perform the classification task. The work [54] performs well for visual modification recognition, however, recognition evaluation shows degraded values for videos with dark light settings.

Mazaheri et al. [55] introduced an approach for Face-reenactment detection by employing a two-stream. Initially, deep features were computed by employing the XceptionNet framework which was later passed to a decoder unit to differentiate the original and manipulated content. The work [55] attains better Face-reenactment recognition accuracy, however, suffering from high computational cost. Many researchers have employed DL-based approaches to detect both Face-swap and Face-reenactment-based deepfakes. Like, Sabir et al. [56] highlighted an observation that the manipulated content lacks temporal consistency. To investigate this, an RNN-based approach was employed to locate the alerted samples. This work [56] attains robust visual manipulation detection accuracy, however, workable only with image-based alterations. Another approach was presented in [57] that combined the handcrafted and deep key points for deepfakes detection. The work [57] shows better visual alteration detection accuracy, however, unable to perform well for compressed samples. To better assess the mesoscopic characteristics of altered content, Afchar et al. [58] presented a technique comprising two types of CNN approaches with a lightweight architecture namely Meso-4 and MesoInception-4. The work is robust from the perspective of computing burden, however, detection performance needs improvements. Nguyen et al. [59] introduced a CNN model to concurrently identify and recognize forged digital data. For classifying the deepfakes, an auto-encoder was utilized, while the manipulated area segmentation was performed via a y-shaped decoder. The approach is robust to deepfakes detection; however, does not suit well to real-life cases. To deal with the deepfakes detection degradation problem like the one that occurred in [59], Stehouwer et al. [60] introduced a CNN model for manipulation recognition. However, the strategy discussed in [60] is computationally inefficient. Another approach was presented in [61] where a framework namely the supervised contrastive (SupCon) model with Xception network was used for detecting both Face-swap and Face-reenactment-based visual manipulation. The approach [61] generalizes well to unseen cases, however, performance needs more improvements. Yu et al. [62] suggested a framework called the SegNet to reveal visual manipulations. Initially, the suspected samples were divided into four patches, then, a CNN

Table 1
Comparison of existing FS and FR detection approaches.

Reference	Method	Best Results	Database	Type	Limitations
Handcrafted keypoints-based methods					
[29]	SURF features with SVM	Precision= 97 % Recall= 88 % Accuracy= 92 %	Custom	FS	• Incapable of preserving facemask expressions. • Applicable to images only.
[30]	68-D key points with SVM	ROC=89 % ROC=84 %	UADFV Custom	FS FS	• Not robust to blurred samples.
[31]	Multimedia stream descriptor [29] with SVM and RF	AUC= 93 % (SVM) AUC= 96 % (RF)	Custom dataset.	FS	• Not generalize well to samples re-encoding attacks
[33]	Heart rate features with the CNN	• Accuracy= 96 %	Face-Forensics	FS	• Computationally complex.
[34]	Landmark key points	• Accuracy= 87.5 %	Eye Blinking Prediction database	FS	• Nor workable for people with mental health issues.
[36]	Optical flow fields with CNN	• Accuracy= 81.61 %	Face-Forensics ++	FS	• Performance needs evaluation on a more complex dataset.
[39]	16-D features with the MLP.	• AUC=.0.851(FS) • AUC=.823 (FR)	Face-Forensics ++	FS FR	• The approach is workable for cases with open eyes and visible teeth.
[16]	Landmarks feature with SVM	AUC= 93 % (FS) AUC=98 % (FR)	Custom dataset.	FS FR	• The lowest classification results are for scenarios where a subject is looking off-camera.
DL keypoints-based methods					
[41]	VGG-16, ResNet-50,101,152	AUC=84.5 (VGG16), 97.4 (ResNet50), 95.4 (ResNet101), 93.8 (ResNet152)	TIMIT	FS	• Unable to work well for compressed samples.
[43]	CNN + RNN	Accuracy=97.1 %	Custom dataset.	FS	• Workable with short-length visual samples.
[44]	CNN + RNN	TPR= 99 %	Custom dataset.	FS	• Not generalized well for subjects with closed eyes.
[45]	CNN + RNN	Accuracy=92.61 %	DFDC	FS	• Accuracy requires enhancement.
[47]	VGG11 along with LSTM	Accuracy= 98.26 %,	Celeb-DF	FS	• Economically expensive.
[48]	VGG6 + CNN	AUC= 99 % AUC= 99 % AUC= 93 % AUC= 99 %	WLDR Face-Forensics DFD Celeb-DF	FS FS FS FS	• Does not work well for unseen samples.
[49]	Biological signals with Neural-ODE model	Loss=0.0215 Loss=0.0327	Custom TIMIT	FS FS	• Computationally expensive
[54]	Landmarks and deep features with CNN	AUC=0.87	DFDC	FS	• Not robust to samples with dark light.
[55]	Two-stream network	Accuracy= 98.43 %	Face-Forensics ++	FR	• Computationally complex.
[56]	CNN + RNN	Accuracy= 96.3 % (FS) Accuracy= 94.35 % (FR)	Face-Forensics ++	FS FR	• Work for the image-based alterations only.
[58]	MesoInception-4 with deep classifier	TPR= 81.3 % (FS) TPR= 81.3 % (FR)	Face-Forensics ++	FS FR	• Not robust to compressed samples.
[59]	CNN	Accuracy=83.71 % (FS) Accuracy=92.50 % (FR)	Face-Forensics ++	FS FR	• Fewer classification results for unseen examples.
[60]	CNN	Accuracy=99.43 % (FS) Accuracy=99.4 % (FR)	DFFD	FS FR	• Computationally complex.
[57]	CNN + SVM	Accuracy= 90.29 % (FS) Accuracy= 86.86 % (FR)	Face-Forensics ++	FS FR	• Low performance on compressed videos.
[61]	SupCon model with Xception network	Accuracy= 94.74 % (FS) Accuracy= 94.36 % (FR)	Face-Forensics ++	FS FR	• Performance degrades for highly compressed samples.
[62]	Patch-based CNN framework	Accuracy= 83.8 % (FS) AUC=0.84 (FS) Accuracy= 94.6 % (FR) AUC=0.946 (FR)	Face-Forensics ++	FS FR	• Degrades detection accuracy for unseen examples.

framework was used to extract the visual characteristics. Finally, the features from all patches were combined to predict the outcome of content being real or manipulated. The approach [62] is robust to deepfakes detection, however, recognition performance degrades for unseen samples. A comparative analysis of existing methods for both FaceSwap and Face-Reenactment deepfakes detection is performed in Table 1. It is fairly visible from the investigation performed in Table 1, that despite of huge studies presented for the detection of visual manipulations, still there is a need for a more reliable and effective deepfakes detection model.

3. Proposed methodology: AUFF-NET

The introduced framework for visual manipulation detection consists of a convolutional Bi-LSTM approach to manage the video frame sequences. The proposed model namely the AUFF-NET comprises two main modules described as i) the CNN section that is concerned to extract the reliable features to capture the spatial information of the input video samples, ii) a Bi-LSTM module that is responsible for computing temporal video sequences analysis to evaluate its behavior over time. In the convolutional part, which is the first unit of the suggested framework, we have introduced a novel swish activation approach-based Inception-swish-Resnet-v2 module to obtain the sample characteristics at the frame level. Then, the computed key points from numerous consecutive frames are passed to the Bi-LSTM unit for sequence analysis. After this, three additional dense layers are added to the AUFF-Net model to designate a more representative group of key points. At last, the probability based on both the spatial and temporal features is computed to estimate whether the suspected video is original or either its FS- or FR-based deepfakes. The complete workflow of the proposed approach is given in Fig. 3.

For a given visual sample, the Inception-swish-Resnet-v2-based Bi-LSTM framework is utilized to attain the spatial and temporal information to detect and classify the forensic manipulations. Using the idea of end-to-end training, the grouping of dense layers is employed to pull the Bi-LSTM approach to an output classification score. More clearly, the Inception-Resnet-v2 technique is modified by using the swish activation approach. Moreover, it contains three extra dense layers at the end of the model structure to avoid the occurrence of the model over-fitting issue. The introduced AUFF-Net contains the CNN module and a Bi-LSTM unit to capture the local pixel information at the frame level and temporal sequences, respectively. The detailed demonstration of both modules is defined in the consequent sections.

3.1. CNN module

The first unit of the proposed approach is a CNN framework that is directed to extract the group of dense features from the examined visual sample. The computed information is later employed as input for the Bi-LSTM module to compute the final result (real, FS, and FR). In the presented approach, we have adopted a pre-trained CNN model namely the Inception-Resnet-v2 framework, and modified it by proposing the swish activation approach. The basic reason to employ a pre-trained model at the CNN unit of the AUFF-Net is that it is efficient to calculate a more robust set of key points over an extensive, publicly available database namely the ImageNet dataset. During the training procedure of such a CNN model, the initial layers are concerned to capture low-level information like the face texture and edges, etc. While the advanced layers are concentrated to extract the task-explicit key points that are far away from human intuitive understanding. More specifically for deepfakes detection, the initial layers learn the target position, while the deep layers are concerned with recognizing the manipulation. As a pre-trained framework has already learned substantial information and gained huge sample structure knowledge, hence, its employment to accomplish a new job like using it for deepfakes identification minimizes the network training time and increases the model recognition accuracy. A visual description of the 'transfer learning' procedure is elaborated in Fig. 4. The employed approach is proficient in extracting the efficient key points from the input sample i.e., face orientation, nose, and eye alignments, etc.

3.2. Custom CNN module: inception-swish-ResNet-v2

Inception-ResNet-v2 is a renowned CNN framework that is empowered to attain an effective group of image key points to perform several classification tasks. The structure of Inception-ResNet-v2 consists of a

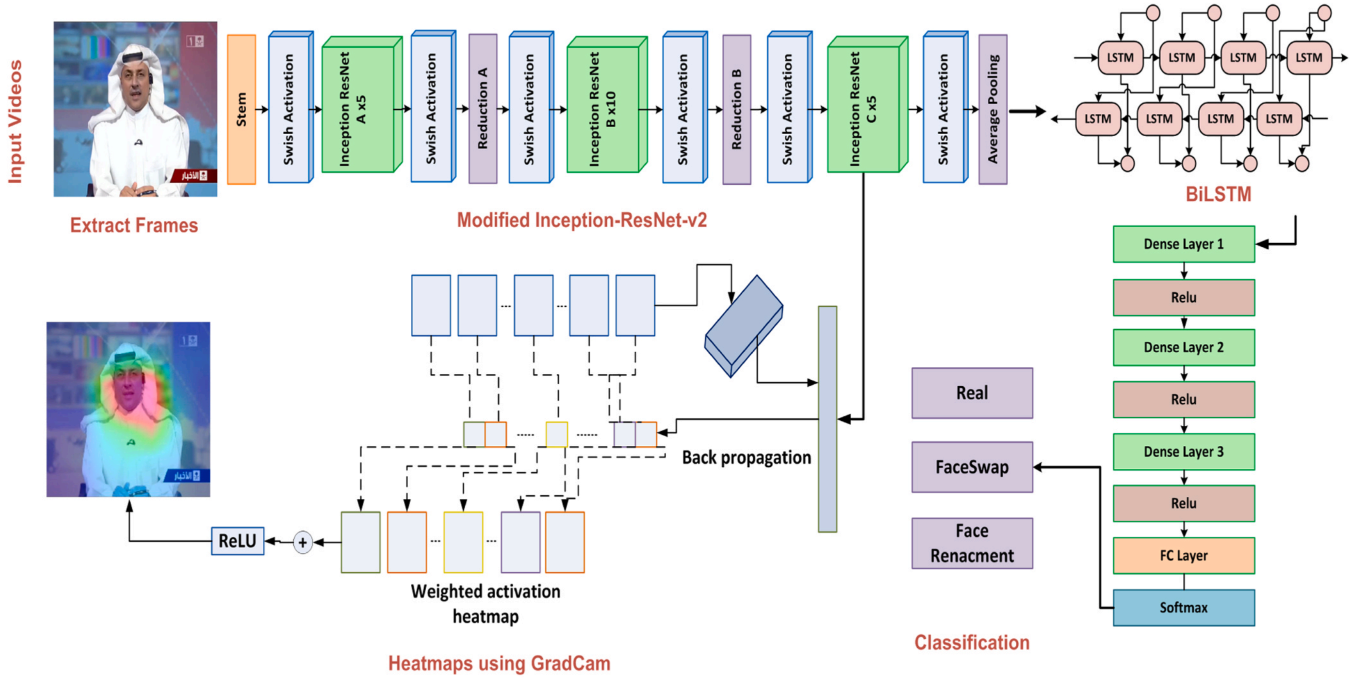


Fig. 3. A detailed view of the presented framework for the FS and FR deepfakes detection.

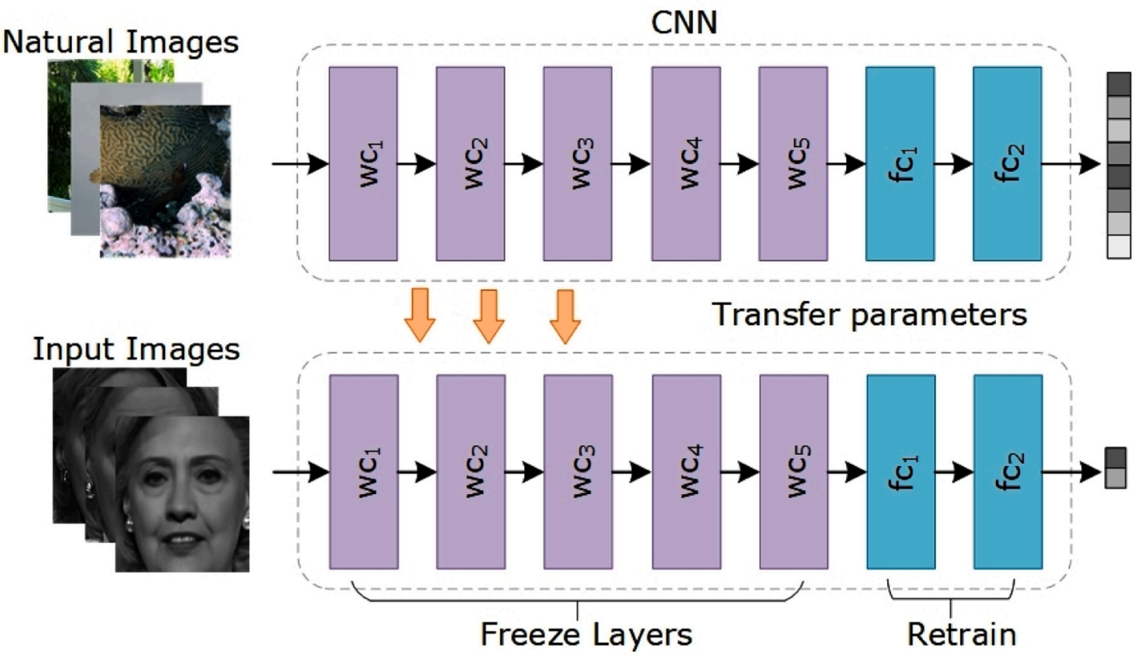


Fig. 4. A visual depiction of transfer learning.

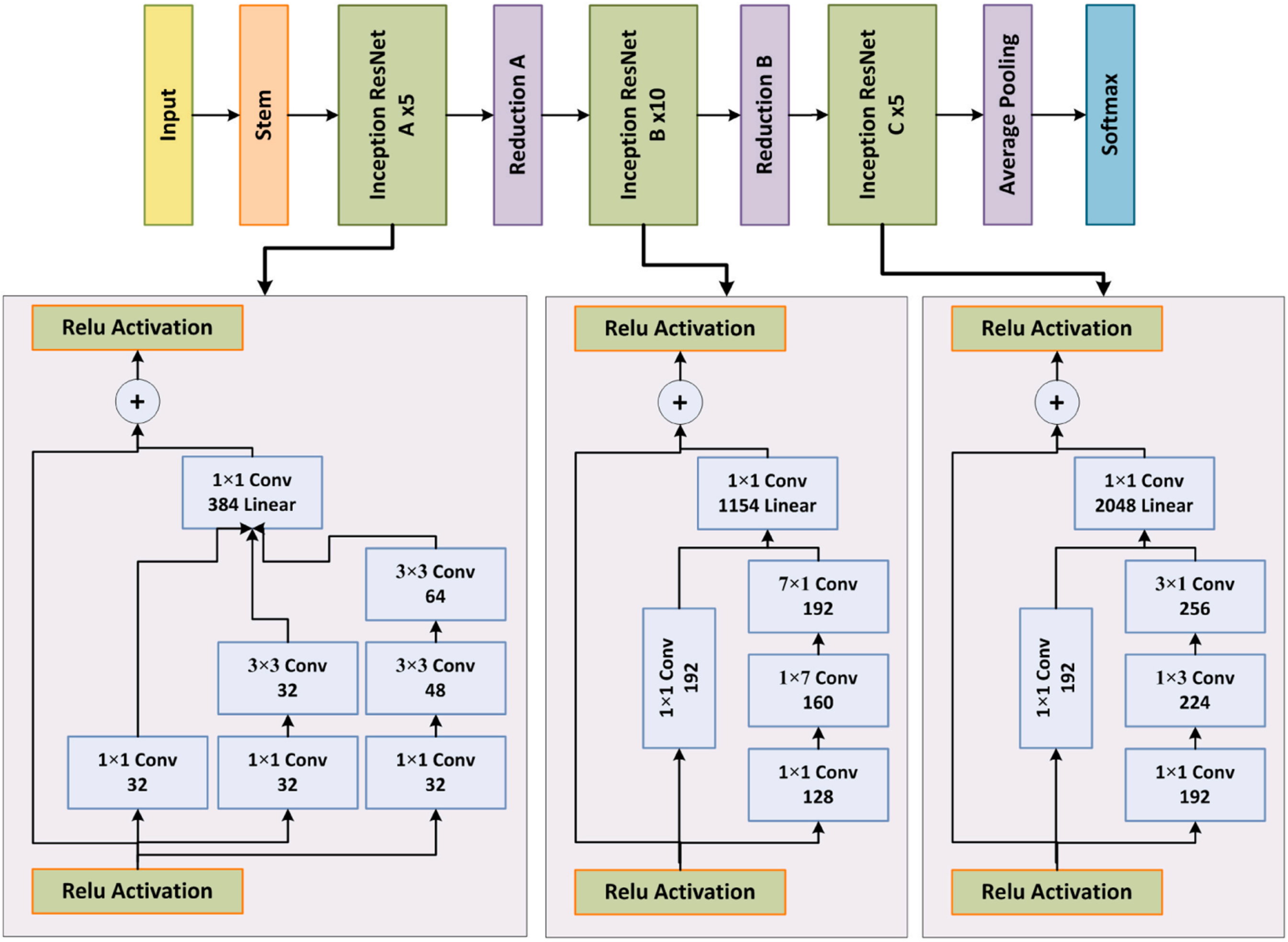


Fig. 5. Architecture view of the Inception-ResNet-v2.

mixture of Inception and residual modules (ResNet) interconnection. The main reason to employ the Inception-ResNet-v2 model for deep features computation is that this model can minimize the performance reduction issues and avoid model over-fitting problems that can occur from deep network structures and provide a concise set of image features. The inception model comprises several convolutions, and pooling layers, together with key-points maps that are combined to form a single feature. While, the ResNet framework is famous for its skip links, which efficiently join the key points from the previous to the next layer. Such structure of the ResNet model permits it to extract a robust group of sample characteristics and attain better performance in a deep network. The Inception-ResNetv2 model uses the benefits of both the Inception and ResNet models and is capable of achieving better recognition performance. The Inception-ResNet-v2 network contains Inception-Resnet-A, B, and C blocks as shown in Fig. 5. The first block of the Inception-ResNet-v2 approach is responsible for calculating the low-level sample features like edges, face, nose, and eye orientations of the targets. The second block is concerned to capture the texture, and sharpness of samples and locating the target location, while the last block is focused to capture high-level sample information like people recognition to determine the forensic alterations from the visual samples. The basic structure of the Inception-ResNet-v2 contains the convolutional, activation, and pooling layers with the ReLU activation method. We have presented a novel Inception-swish-ResNet-v2-based approach by employing the swish activation on the Inception-ResNet-v2 model. The introduced activation method permits the framework to boost its learning via showing minimum loss and a better ability to learn complex sample patterns. The comprehensive depiction of the Inception-swish-ResNet-v2 approach is illustrated in Fig. 6.

Comprehensive details of all mentioned layers are given in the below sections:

3.2.1. Convolution layer

This layer is responsible to extract the deep features from the given video sample by using Eq. (1):

$$F_i^L = f(\sum_{j\in N_i}(K_{ji}^L * F_j^{L-1} + a_i^L) \qquad (1)$$

Here, L shows the total framework layers, while F denotes the obtained feature vector with filter size K, and * is representing the convolution operation. While a is showing the bias value with N_i is depicting the feature maps. In our approach, all video frames are resized to 229×229 to meet the model requirements. The employed approach consists of a total of 825 convolution layers used for deep key points computation.

3.2.2. Activation layer

To boost the video-based fabrication detection empowerment of the suggested framework, we employed the Swish activation approach in

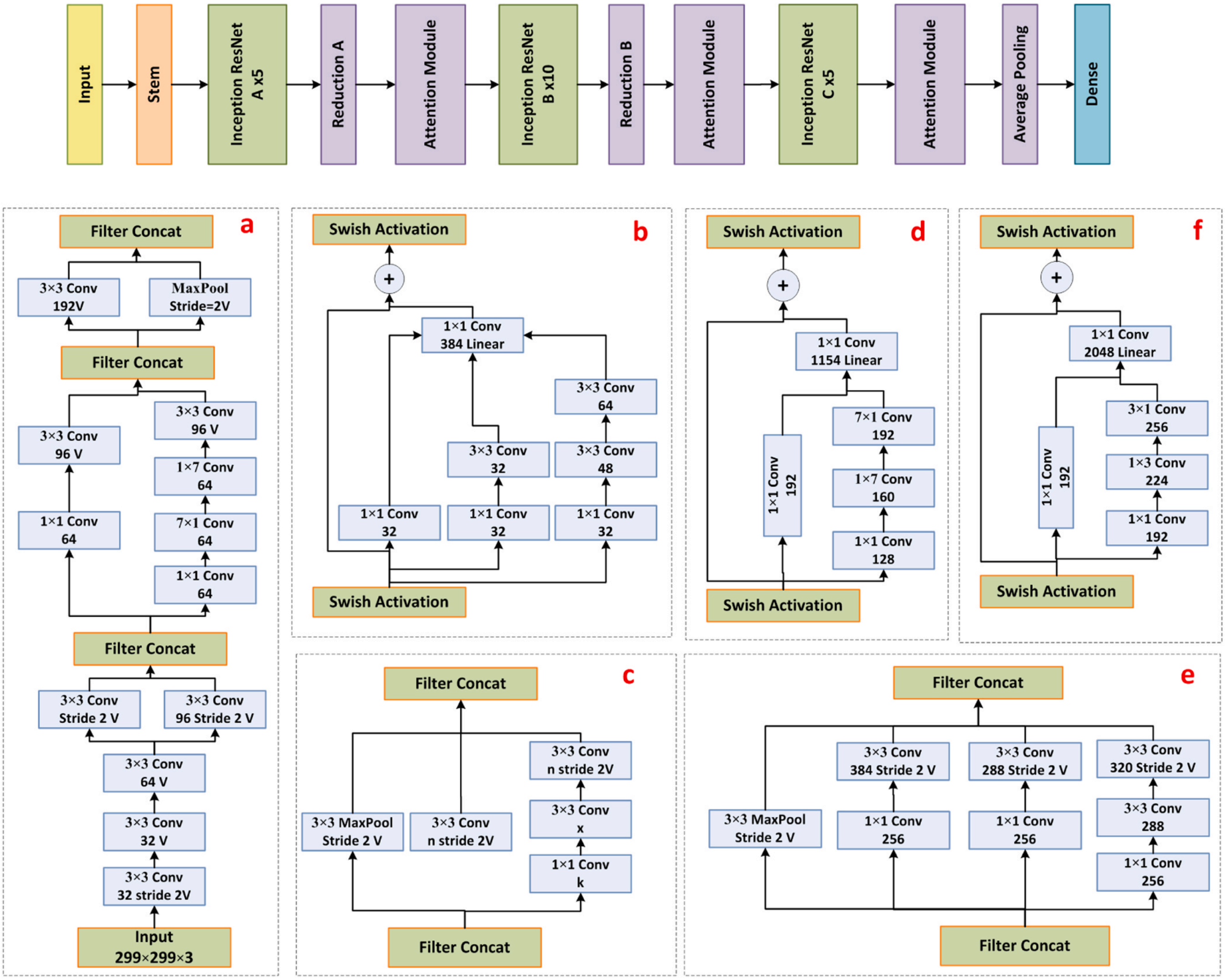


Fig. 6. A detailed depiction of the proposed Inception-swish-ResNet-v2 model: (a) stem block, (b) Inception-swish-ResNet block A, (c) reduction module, (d) Inception-swish-ResNet block B, (e) reduction B, and (f) Inception-swish-ResNet block C.

place of the ReLU function in the proposed Inception-swish-ResNet-V2 framework after all convolutional 2-D layers. The employed activation function is non-monotonic smooth unrestrained above and limited below in its learning curve. The mentioned attributes of the swish activation approach allow the CNN model to avoid saturation and overfitting issues of a framework. The swish activation approach is uncomplicated by definition, and research shows that it outperforms the most popular ReLU activation strategy in the challenging study fields of image categorization and object identification [28]. The core factor for the effective results of the swish approach is that the ReLU method blocks the propagation of negative numbers inside the framework feature engineering process which leads to the elimination of crucial sample data. Contrary the swish technique allows the slight negative numbers to circulate through the network which is crucial for calculating complicated behavior from examined samples in the dense models. Fig. 7 provides a graphic representation of the swish and ReLU activation techniques. The swish approach is mathematically represented as follows:

$$s(f) = f \quad \times sigmoid(\alpha f) \tag{2}$$

Here, f is the input, while α is showing the per-channel trainable parameter. Furthermore, the quick learning ability of the introduced activation method also makes the Inception-swish-ResNet-v2 model computationally more efficient as it requires minimum time for training in comparison to other activation functions.

3.2.3. Pooling layer

This layer is employed to minimize the feature dimension by removing the unnecessary key points. In the proposed approach, the computed key points are taken from the average pooling layer which is later propagated to the Bi-LSTM module for the temporal sequence analysis. The Inception-swish-ResNet-v2 approach calculates the feature vector with the dimension of 1536 from a single frame which is later used to accomplish the deepfakes detection task.

3.3. Bi-LSTM

The recurrent neural networks (RNNs) are proficient in analyzing the hidden progressive configurations of time-based data, however, they suffer from the vanishing gradient issue which obstructs the model parameters from being correctly updated during the backpropagation procedure. Such structure of RNN causes to degrade of the video classification performance of a model. To tackle the issues of the RNN model, the LSTM approach is presented which has the same architecture as RNN with added memory cells" as a substitute for its weight update procedure and the added memory cells can store the information for a long period. Suppose the computed feature vector from the CNN module is presented by v_t, while the final hidden state and memory cell are presented by s_{t-1} and m_{t-1}, respectively, then the Equations from 3 to 7 are used to implement the LSTM approach.

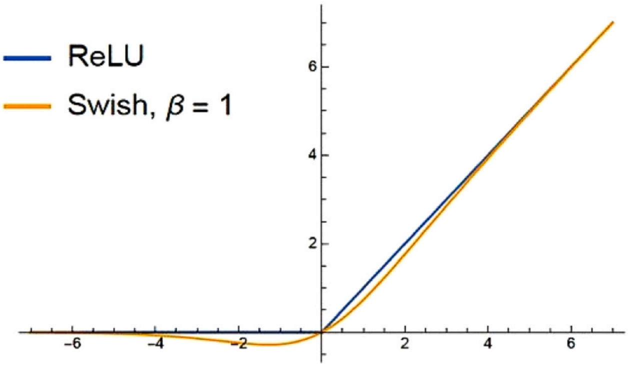


Fig. 7. ReLU vs. Swish activation method [63].

$$x_t = \varsigma(\omega_{vx} v_t + \quad \omega_{sx} s_{t-1} + \quad \omega_{mx} m_{t-1} + \beta_x \quad) \tag{3}$$

$$\mathfrak{F}_t = \varsigma(\omega_{v\mathfrak{F}} v_t + \quad \omega_{s\mathfrak{F}} s_{t-1} + \quad \omega_{m\mathfrak{F}} m_{t-1} + \beta_{\mathfrak{F}}) \tag{4}$$

$$m_t = \mathfrak{F}_t m_{t-1} + x_t \tanh(\omega_{vm} v_t + \quad \omega_{sm} s_{t-1} + \beta_m) \tag{5}$$

$$y_t = \varsigma(\omega_{vy} v_t + \quad \omega_{sy} s_{t-1} + \quad \omega_{my} m_t + \beta_y) \tag{6}$$

$$s_t = \quad y_t.\tanh(m_t) \tag{7}$$

Here, ς represents the sigmoid activation method, while for time t, the x, $\mathfrak{F}$, y, and m are showing the input, forget, output gates, and memory cell state, respectively. While the ω and β are denote the weights and biases, respectively. From the perspective of deepfakes detection and classification, the major limitation of the LSTM approach is that it computes past information only. To capture the entire context of suspected samples, it is mandatory to take both information on both ends for example past and future. Hence, for this reason, we have employed the bidirectional LSTM (Bi-LSTM) approach for deepfakes identification as it is capable of storing the information on both ends. A graphic elaboration of Bi-LSTM is given in Fig. 8.

The Bi-LSTM approach comprises two types of hidden units namely the forward (s_t^f) and backward (s_t^b) hidden layers, respectively. The s_t^f takes input in forwarding manners of time like t=1,2,3,…., T, while the s_t^b accepts input in reverse manners of time t like t= T, T-1,…1, respectively. Finally, the resultant value s_t is calculated by joining the values of both $s_t^f$and s_t^b. Equations from 8 to 10 are used for the implementation of the Bi-LSTM approach.

$$s_t^f = \tanh(\quad \omega_{vs}^f v_t + \quad \omega_{ss}^f s_{t-1}^f + \beta_s^f) \tag{8}$$

$$s_t^b = \tanh(\quad \omega_{vs}^b v_t + \quad \omega_{ss}^b s_{t+1}^b + \beta_s^b) \tag{9}$$

$$y_t = \omega_{ss}^f s_t^f + \omega_{ss}^b s_t^b + \beta_y \tag{10}$$

3.4. Added dense layers

After the Bi-LSTM layer, three dense layers along with the ReLU method and dropout layers are incorporated. The added layers allow the model to emphasize the altered visual areas while eliminating unwanted background data and improving deepfakes recognition performance under varying transformation conditions, like changes in intensity, hue, and facial area positions. The introduced layers produce huge probabilities by combining the information coming from the previous layer with the activation units of coming layers; therefore, a dropout of 0.25 is used to evade the framework over-fitting issue. After this, the calculated key points are passed to the softmax layer.

3.5. Softmax layer

The final layer of our model the softmax layer is focused on categorizing the input sample into the defined classes. In our case, this layer is used to categorize the suspected video sample into three classes namely the real, FS, and FR deepfakes, respectively. The presented technique combines a softmax activation method in the final fully connected (*FC*) layer to predict the comparative likelihood of three output neurons. Eq. (11) is used to compute the softmax activation method given as:

$$\delta(Z_x) = \frac{\exp\,(Z_x)}{\sum_{m=0}^{n-1}\exp\,(Z_m)} \tag{11}$$

Here, (Z_x) and (Z_m) are presenting the input and final vectors, while m is showing the total number of outcome categories which are three for our case.

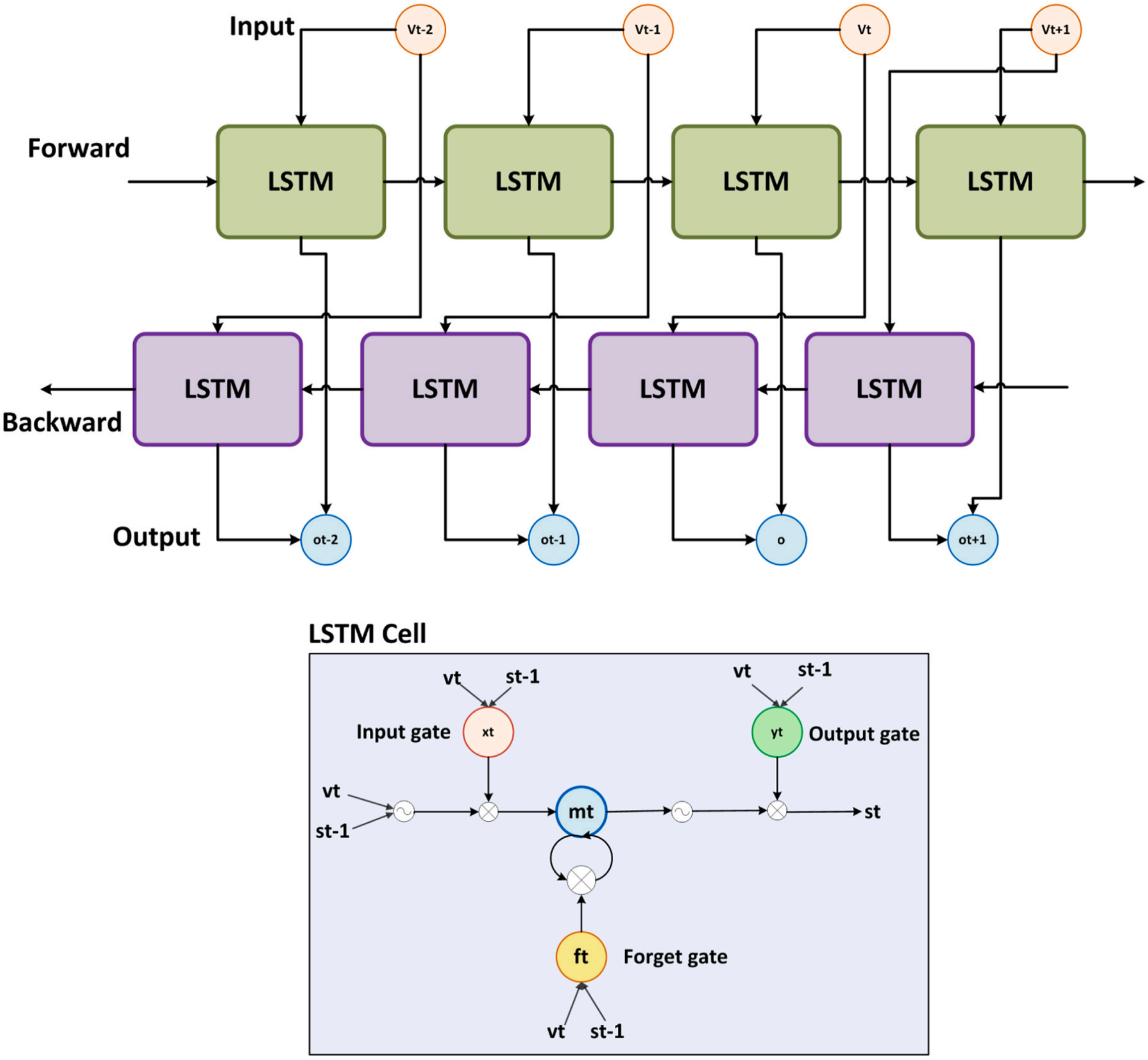


Fig. 8. A pictorial view of the structure of the Bi-LSTM model.

3.6. Loss method

The loss method assists to analyze the framework's effectiveness. For the training samples, the model employs automatic learning to detect the patterns and mark estimations. The loss method computes the degree of deviation in the real and estimated results. The method is repeatedly refined in the procedure of the model training till the optimized values are attained to reduce the prediction error. For the categorization tasks, the softmax layer uses the cross-entropy loss function [23] for estimating the variance in the predicted and original scores and is effective to handle the class imbalance problem as well. The mathematical explanation of the cross-entropy loss function L is given as:

$$L = \frac{1}{n}\sum_{j=1}^{m}\log\left(\frac{e^{s_m}}{\sum_k e^{s_j}}\right) \quad (12)$$

Here m is the total neurons in the output layer and s_m is the input vector.

3.7. Explainability module

One of the major requirements for a video manipulation approach is to make them explainable to increase their truthfulness which can assist the forensic analyzers to use them for processing legal claims. To accomplish this, we have proposed an explainability unit in our framework. More specifically, we have used the Grad-CAM approach that uses the gradient of the output classes concerning the key points map to visualize the class discrimination power of a framework. The visibility power of the Grad-CAM approach to view the inner working of a CNN model allows them to become more transparent. Such behavior assists the forensic analyzers in understanding which parts of the samples are important to determine the manipulations made within the visual content [64]. The internal working of the Grad-CAM is depicted in Fig. 9 which is clearly indicating that the backpropagation gradient of the m^{th} output class against the k^{th} feature map is elaborated as $\frac{\partial s^m}{\partial f^l}$. A weight against each l^{th} feature map is attained by employing the gradient approach over its pixels and computing an average value as given in the Eq. (13):

$$w_l^m = \frac{1}{Z}\sum_i\sum_j\frac{\partial s^m}{\partial f_{ij}^l} \quad (13)$$

Here, w_l^m is depicting the weight computed against the l^{th} feature map with respect to the m target class. While s^m is showing the confidence score of each m target class before passing the softmax method. Whereas, f_{ij}^l is denoting the element (i,j) belonging to the lth key points map containing an accumulative of Z pixels. The Grad-CAM approach uses the ReLU method to perform the weighted summation of all key points plots to produce the category activation heatmaps as given in Eq. 14:

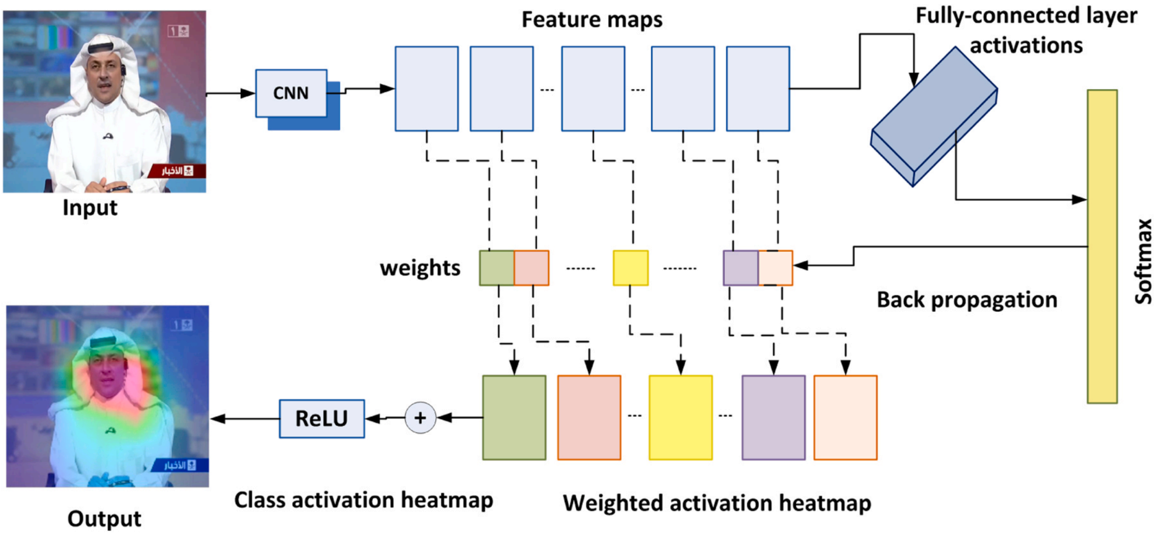


Fig. 9. A visual description of the GradCam module.

$$H_{GC}^{m} = ReLU(\sum_{l} w_{l}^{m} f^{l}) \quad (14)$$

Here, H_{GC}^{m} is representing the class activation heatmaps against each class m. The major purpose of the activation method is to lay emphasis on the region of interest and overwhelm the unnecessary information in the feature maps as we are concerned with the key points having a positive impact on the class of interest. Finally, a mask m_{Mask} for each frame from all classes is computed to show the different areas of the suspected samples with different colors by using Eq. 15:

$$m_{Mask} = \sum_{m=1}^{3} H_{GC}^{m} \quad (15)$$

The computed mask m_{Mask} contains extensive information about the sample and showed it with different colors. In our case, the red color is depicting the most significant areas where the manipulations occur.

3.8. Overview of the presented method

A thorough explanation of the entire introduced approach for FS and FR deepfakes detection is elaborated in Algorithm 1. The frames from the input videos are resized to the fixed resolution of 229×229×3 pixels which are then passed to the CNN module for deep features computation. The Inception-swish-Resnet-v2 model computed the key-points vector of the number of frames (nof)×8×8×1536 dimensions. The computed key points are restructured and passed as input to the Bi-LSTM framework. The flattening layer is introduced to convert the key points to a 1-dimensional vector. Then, a sequence of 3 FC layers along with the ReLU method is introduced. The dense layers produce extensive probabilities by joining the coming data from one layer to all activation units

Fig. 10. Samples from the FaceForensic++ dataset, first row: real videos, second row: FS deepfakes, third row: FR samples.

of coming layers; so, a dropout rate of 0.25 is applied to prohibit the model over-fitting issue. Lastly, the final layer called the softmax layer generates the resultant classification score. Further, we have discussed the big(O) notation to elaborate on the time complexity of the proposed approach. The time complexity of a spatial along with Bi-LSTM model is primarily determined by the number of time steps in the input sequence and the number of hidden units in the model. Let's assume we have a sequence of length N and H hidden units. For the forward pass of the Bi-LSTM, where the sequence is processed from the beginning to the end, the time complexity can be approximated as O (N *H). For the backward pass, where the sequence is processed in reverse order, the time complexity is also O (N *H). Since the proposed approach employs bidirectional feature sequence analysis, the overall time complexity is the sum of the forward and backward pass complexities, resulting in approximately O (2 * N * H), which can be simplified to O (N *H).

4. Results

Here, we have given the particulars of the employed data sample for model evaluation along with the used performance measures. Moreover, we have executed several experiments to show the robustness of our technique for FS and FR detection.

4.1. Dataset

To assess the visual fabrication detection accuracy of our proposed approach, we have used a large publically accessible database namely FaceForensic++ [65]. This dataset is a large-sized dataset of altered videos that contain 1000 real and 4000 forged samples of several subjects. The videos are fabricated by using several alteration approaches like DeepFakes [66], FaceSwap [65–67], Face2Face [68], and NeuralTextures [69]. Moreover, the samples are available for three quality levels which are as follows: high-quality (Raw or C0), slightly compressed (C23), and deeply compressed (C40). We have considered the samples of all quality levels for the FaceSwap, and Face2Face visual manipulations. A few examples from the employed repository are given

Algorithm 1. Steps for visual manipulation detection using the AUFF-Net approach.

```
    START
INPUT: VS
OUTPUT: IR-Bi-LSTM, Classified samples, CM, Acc, Pre, Rec, F1s,
            VS: Total video samples either real, FaceSwap, or Face-Reenactment-based deepfakes.
            IR-Bi-LSTM : AUFF-Net model.
            Classified samples: Each video sample is marked either as real, FaceSwap, or Face-Reenactment deepfakes.
            CM: Obtained confusion matrix
            Acc: Accuracy results for the entire dataset.
            Pre: Precision value acquired for the employed database
            Rec: Recall value calculated for the employed database
            F1s: F1-Score computed for the used dataset
    VideoSize ← [Total Frames]
// Resizing Frames
α← ImResize (VideoSize, [299,299])
// IR-Bi-LSTM Model
            IR-Bi-LSTM ← AUFF-Net (α)
         [ Dr, Dt] ← Dividing deepfakes dataset into train and test sets.
    // Training Unit
            For each video v in →Dr
             For each frame f in v
                    Extract Inception-swish-Resnet -v2 keypoints →sf
              End For
             Extract temporal sequence analysis tf→ LSTM (sf)
            End For
            Training IR-Bi-LSTM for all tf, and compute training time complexity t_ IR-Bi-LSTM
         η_IR-Bi-LSTM ← ClassifySample(tf)
         Ap_ IR-Bi-LSTM ← Evaluate_AP(η_IR-Bi-LSTM)
            For each sample I in → Dt
                    a) compute both spatial and temporal features through trained model €→βI
                b) [class] ←Predict (βI)
                    c) t_ IR-Bi-LSTM ←O(N * H^2) // N total samples, H hidden units of model

            End For
    // Test the Model and compute results on the trained model
            Ap_€← Test model € using η
            Output_class← IR-Bi-LSTM (Ap_€)
            CM= ConfusionMatrix (Output_class, Actual)
            Acc= $(True^{+} + True^{-})/ (True^{+} + False^{+} + False^{-} + True^{-})$
            Pre= $(True^{+} )/ (True^{+} + False^{+} )$
            Rec=$(True^{+} + True^{-})/ (True^{+} + False^{-} )$
            F1s= (2 x Pre x Rec)/(Pre + Rec)
    return trained model, Output_class, CM, Pre, Rec, F1s
FINISH.
```

in Fig. 10.

4.2. Evaluation metrics

For performance measurement, various standard evaluation measures called precision, recall, F1-measure, accuracy, and AUC/ROC curves are utilized. The technical details of utilized performance measures are given in Eq. (16) to Eq. (19), respectively.

$$\mathrm{Pr} = \frac{\text{ď}}{\text{ď} + \text{Ɣ}} \tag{16}$$

$$\mathrm{Re} = \frac{\text{ď}}{\text{ď} + \text{ɋ}} \tag{17}$$

$$\mathrm{Accuracy} = \frac{\text{ď} + \text{ŕ}}{\text{ď} + \text{ŕ} + \text{Ɣ} + \text{ɋ}} \tag{18}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{Pr} \times \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}} \tag{19}$$

Where ď indicates the true positives (correct forged detected examples), and ŕ represents true negatives (true pristine detected examples). Moreover, Ɣ determines the false positives (incorrectly classified real examples), and ɋ shows false negatives (incorrectly classified deepfakes examples).

4.3. Implementation explanation

The framework is executed in Matlab 2021 software and executes on Nvidia GTX1070 GPU-based computer system. The utilized data sample is classified randomly into 70:10:20 parts to generate three individual data parts namely the training, validation, and test sets. The given settings are used to implement the proposed approach.

i) For all channels, the mean is subtracted.
ii) The video frames are set to 299-by-299 resolution to meet the framework requirement.
iii) The network is tuned for 30 epochs with a learning rate of 0.0001 and batch size of 16.

For the introduced work, a visual demonstration of the optimal loss graph is given in Fig. 11 from which it can be visualized that our approach has reached the best loss value of 0.00018 at the epoch rate of 30. Moreover, we have shown the train time model accuracy learning graph in Fig. 12. It is visible that our framework has shown a training accuracy of 99.88 %. The values reported in both figures are clearly explaining the better learning behavior of our model.

4.4. Evaluation of the proposed framework

Here, we have discussed the deepfakes identification results of our framework via using several experiments. Initially, we exhibited the class-oriented deepfakes detection performance and then showed the performance entirely. Moreover, we have analyzed the explainability power of our approach along with the discussion of model performance for several adversarial attacks as well.

4.4.1. Class-wise performance evaluation

For an accurate forensic model, it must have the power to correctly distinguish among various types (FS, FR of visual manipulations and can separate the real data as well. To check this ability of our approach, we have performed an experiment on the FaceForensic++ dataset.

In the first phase, the box graphs are used to demonstrate the acquired category-oriented precision and recall scores as box diagrams are capable of effectively explaining the attained performance values by showing the minimum, maximum, and mean results together with the uniformity and deviation of the results (Fig. 13). The scores in Fig. 13 clearly indicates that our approach can effectively identify the Real, FS, and FR-based deepfakes content.

To further show the deepfakes detection power of our model from the visual samples, we have reported the accuracy, F1 score with the error rate in Fig. 14. One can clearly see from Fig. 14 that the presented framework is effective in categorizing the visual samples as Real, FS, and FR deepfakes by attaining an average error rate of 1.03 %. More clearly, the introduced AUFF-Net model has shown the F1 measures of 99.55 %, 99.13 %, and 98.21 % for the Real, FS, and FR deepfakes, together with the error scores of 0.45 %, 0.87 %, and 1.79 %, respectively. Furthermore, the presented AUFF-Net approach exhibits better class-wise categorization accuracy results of 99.74 %, 99.21 %, and 98.32 % for the real, FS, and FR deepfakes. The major cause for the effective categorization outcomes of our method is the addition of the Swish activation method which enhances the feature engineering capability of the presented framework at the spatial level. Further, the introduced FC layers permit the technique to effectively tackle the model over-tuned data.

Next, the confusion matrix is shown to further discuss the categorization performance of the proposed framework (Fig. 15). The employment of only an accuracy evaluation measure can be ambiguous for datasets with class imbalance problems. A model can attain results higher than 90 %, however, this performance is not good for scenarios where 90 % of images are from one class. Therefore, the confusion matrix effectively elaborates the categorization results of a model by indicating the real and estimated scores of all classes. The presented AUFF-Net model has shown the TPR rates of 99.43 %, 98.94 %, and 98.01 % for the real, FS, and FR deepfakes, respectively. Moreover, we have attained the minimum false positive rate of 0.09 % among the FS and real classes, which depicts that the model has correctly identified and differentiated the real and FS classes. However, for the real and FR classes, we have attained the false positive rate of 1.02 %, which indicates that little association has been found between the real and FR deepfakes. It can be because of the reason that for the FR deepfakes, only the expressions of identity are manipulated without making any changes to the facial attributes. From the conducted analysis it is quite obvious that our framework is proficient in the multi-class environment and can robustly recognize the FS and FR deepfakes from pristine samples.

Furthermore, the AUC-ROC graphs are elaborated (Fig. 16) for the real, FS, and FR deepfakes detection as these curves are empowered to robustly elaborate the recognition capability of the model for numerous classes. AUC-ROC graphs are the crucial model assessment metric for evaluating the behavior of an approach in categorizing the visual samples into respective classes. ROC specifies the likelihood curve while the AUC elaborates on the measurement of separability. In the presented approach, it assists in showing how much the technique is capable of differentiating the input samples and classifying them into real, FS, and FR, classes. The AUC near 1 shows that an approach is effective for classification, signifying a higher level of separability and Fig. 16 is clearly depicting that the introduced approach can correctly classify the videos into their respective classes due to its higher recall ability. Further, the proposed approach has taken a training time of 23 hours and 40 minutes, while an average time of 1 second to test a video sample of 1 minute that shows the efficacy of the suggested work as well.

The performed analysis presents that our approach namely the AUFF-Net model is robust to identify and categorize the FS and FR-based deepfakes samples and is empowered to categorize the original samples from the manipulated data. The major reason for the robustness of our technique is the inclusion of spatial and temporal information which assist in calculating a nominative group of visual characteristics and contribute to improving the recall rate of our approach.

4.4.2. Comparative analysis of the proposed AUFF (Inception-Swish-ResNet-v2 along with Bi-LSTM) model with other activation methods

The objective of this experiment is to investigate the impact of

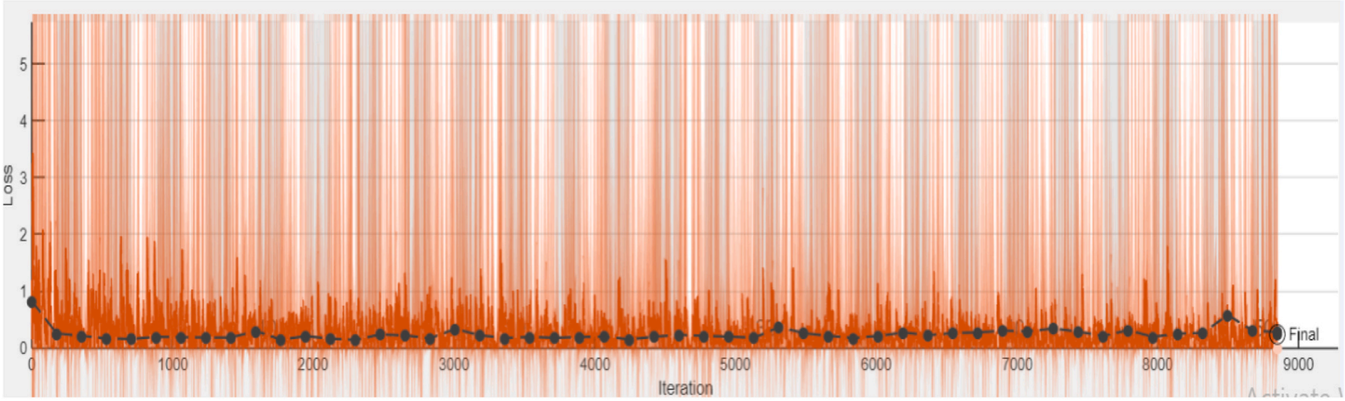


Fig. 11. Loss graph pictorial demonstration.

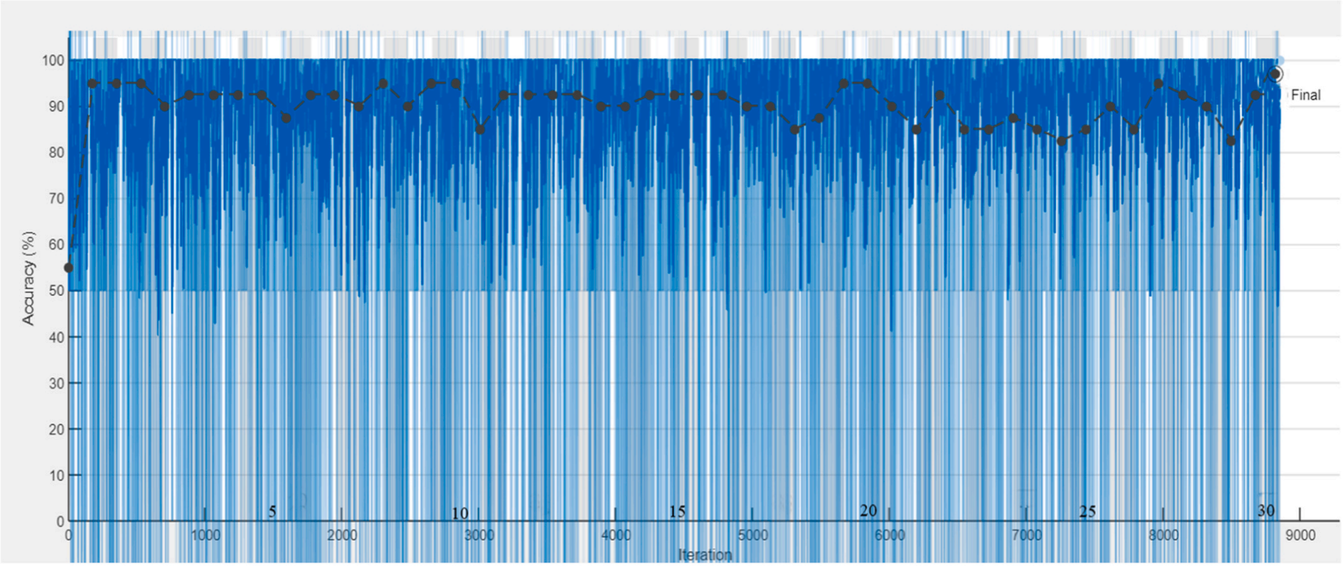


Fig. 12. Visual depiction of model training graph.

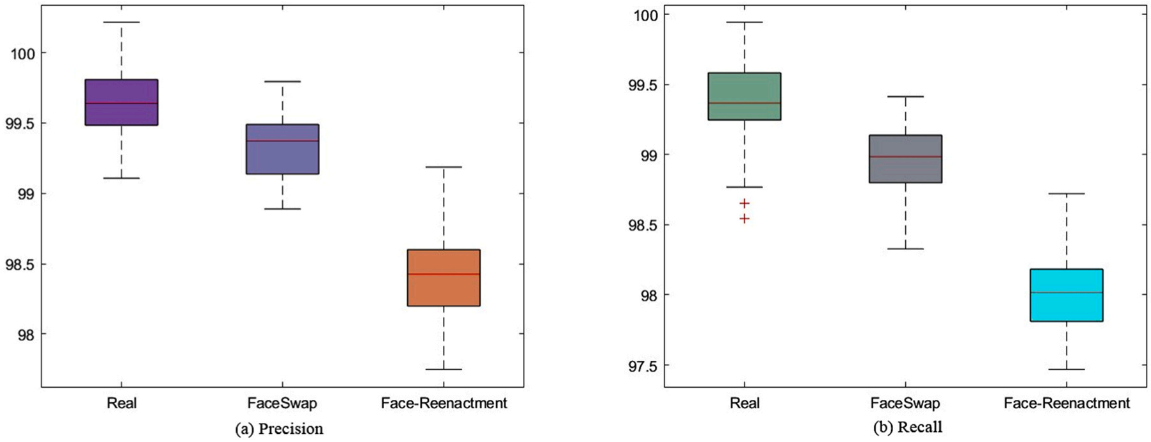


Fig. 13. Pictorial representation of obtained class-oriented Precision and Recall scores for the real, FS, and FR deepfakes detection.

various activation methods tested in the proposed framework. For this purpose, we have examined the performance of the presented network with ReLU, PReLU LeakyReLU, and swish activation methods, and the results are shown in Table 2. We have selected the mentioned activation methods for comparing the model results as these functions are heavily explored for image classification and considered standard. The scores exhibited in Table 2 are effectively indicating that the proposed Inception-Swish-ResNet-v2 along with the Bi-LSTM model outperforms the other activation methods-based model variants. The key cause for the effective results of the Inception-Swish-ResNet-v2 model is due to

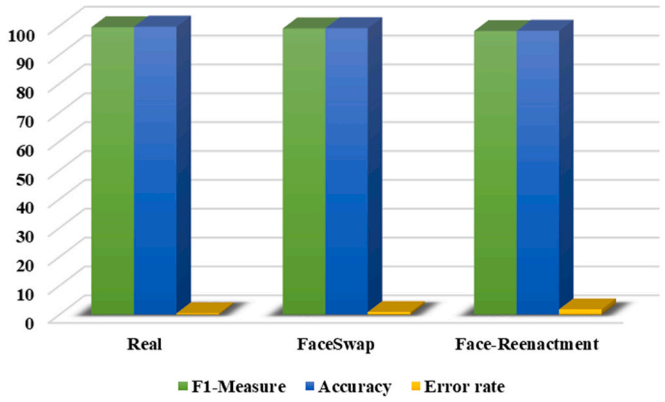
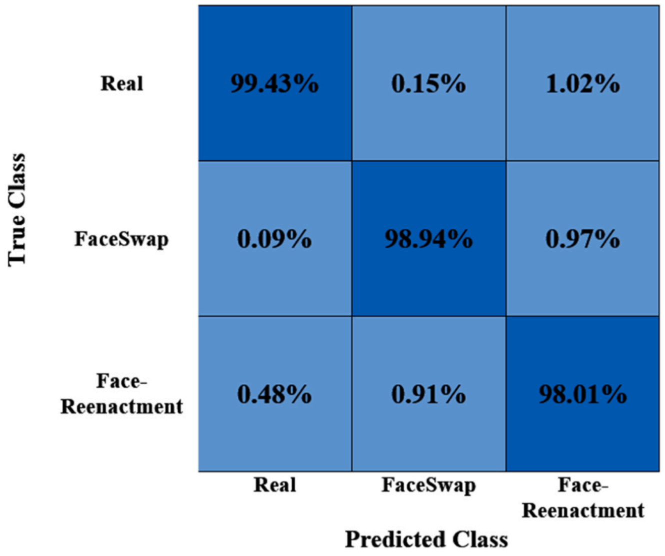


Fig. 14. A visual demonstration of the obtained F1-Measure and error scores for the Real, FS, and FR deepfakes.

the non-monotonic characteristic of the employed activation approach. This nature of the swish method permits the output to decline even for the high input scores which eventually enhances the information storage ability of the introduced framework and authorizes it to compute a robust key points set of underlying visual samples. Whereas the remaining activation techniques lack this attribute, therefore, the Inception-Swish-ResNet-v2 along with the Bi-LSTM framework shows the highest performance values in terms of all employed evaluation measures comparatively to the rest of the activation methods.

4.4.3. Explainability

The role of multimedia forensic systems is very important, especially in those scenarios where such models can be used to process legal claims. The employment of such techniques demands the explainability or reasoning of suspicious areas that have caused nominating a sample as being real or fake. To show this, we have designed an evaluation to check the explainability power of our approach. To accomplish this task, we have attained the heatmaps related to the final layer of the introduced technique with the employment of the Grad-Cam tool [68]. In Fig. 17, we have shown the heatmaps both for the FS and FR. Fig. 17 clearly shows that the presented AUFF-Net method emphasizes those areas of a suspected sample where the modification exists. The key reason for the improved explainability power of the AUFF-Net model is because of the better key points engineering capability of our approach as it utilizes both the spatial and temporal sequences which empower it to effectively tackle the forensic alterations of visual data.



True Class	Real	FaceSwap	Face-Reenactment
Real	99.43%	0.15%	1.02%
FaceSwap	0.09%	98.94%	0.97%
Face-Reenactment	0.48%	0.91%	98.01%
	Predicted Class		

Fig. 15. Confusion matrix attained by the AUFF-Net.

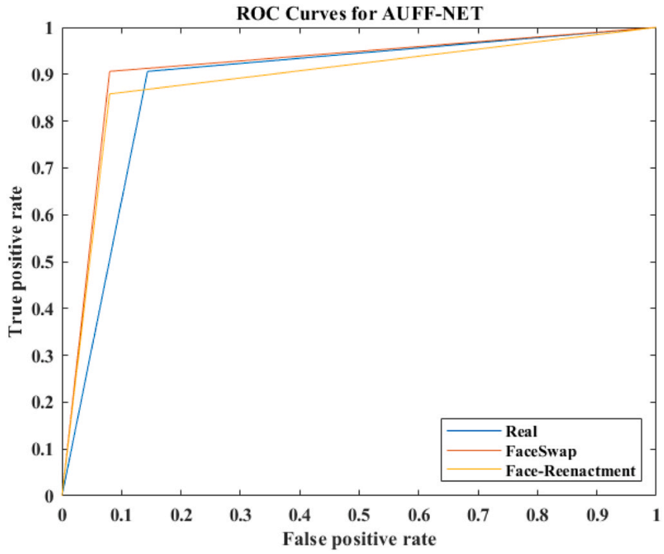


Fig. 16. The AUC curve attained by the AUFF-Net.

4.4.4. Performance analysis over adversarial attacks

One of the major hindrances to the generation of effective deepfakes recognition systems is the occurrence of several adversarial operations among which the most prominent is the compressed samples. The videos which are posted on social sites are subject to severe compression. Moreover, other post-processing attacks involve the incidence of noise, blurring, and light variations. Therefore, a robust deepfakes detection approach must be capable of performing accurately for such data samples. To validate the deepfakes identification results of our model in the presence of such attacks, we have performed an experiment.

The samples in the employed dataset namely the FaceForencsic++ are present at three quality levels and contain extensive compression and light variations attack. Moreover, we have added further perturbations like noise, blurring, zooming, and rotational variations in samples at the evaluation stage to assess the effectiveness of the presented approach over adversarial attacks. We translated the samples into a width and height span of [-2, 2] and performed the rotation and zoom operation with the span of [-0.2, 0.2]. Moreover, clutter and blur are also added to the samples with different window dimensions i.e., 5,7, and 9, etc., to increase the diversity of the evaluation samples. This experiment assists us in validating the robustness of our approach to adversarial attacks as the model is not trained on such perturbations and only faces them during the test time.

We have tested our approach for all quality stages with added adversarial operations and the obtained scores are presented in Table 3. The results demonstrated in Table 3 depict that our approach is proficient in performing well under the presence of several perturbations as well. Even for low-quality samples with a quality level of 40 and added adversarial attacks, the presented approach shows effective results for both FS and FR deepfakes with accuracy scores of 98.41 %, and 98.01 %. So, it is quite clear from the scores given in Table 3 that our model is competent in identifying the manipulated content even for unseen adversarial attacks during training which is clearly indicating its effectiveness for visual manipulation detection.

4.5. Comparative analysis with base models

We accomplished an evaluation to compare the results of our suggested strategy with several base models like VGG16 [69], GoogleNet [70], ResNet50 [71], DenseNet [72], and MobileNetv2 [73]-based Bi-LSTM models for the detection and classification of visual

Table 2
Performance assessment of the AUFF-Net with different activation approaches.

InceptionResNet-v2-based Bi-LSTM model with different Activation method	Precision (%)		Recall (%)		F1-Measure (%)		Accuracy (%)	
	FS	FR	FS	FR	FS	FR	FS	FR
ReLU	98.97	97.86	97.63	96.81	98.29	97.33	98.93	97.79
PReLU	99.38	98.11	97.99	97.23	98.68	97.66	99.29	98.09
LeakyReLU	99.64	98.19	98.31	97.52	98.97	97.85	99.67	98.13
Swish	99.93	98.41	98.94	98.01	99.13	98.21	99.21	98.32

manipulations. The attained values are given in Table 4 from where it is quite evident that the proposed framework is more effective compared to the rest of the techniques in classifying the visual manipulations from the video samples. More clearly, the GoogleNet-based Bi-LSTM approach depicts the lowest results with an accuracy of 91.53 % and 92.19 % for the FS and FR deepfakes. The second lowest result values are shown by the MobileNetv2-based Bi-LSTM with scores of 91.86 %, and 92.32 % for the FS, and FR classes, as this approach presents a lightweight architecture, however, at the compromise of performance degradation. While the DenseNet-based Bi-LSTM model shows comparable results with accuracy values of 96.30 %, and 94.35 % for the FS and FR deepfakes respectively. Comparatively, the proposed improved AUFF-Net model indicates the highest accuracy number of 99.21 % and 98.32 % for the FS and FR deepfakes. Descriptively, for the FS deepfakes, the relative base models attain an average accuracy score of 93.97 %, which is 99.21 % for our case, and the proposed model has given a performance improvement of 5.24 %. Similarly, in the aspect of the FR visual fabrications, the competitor frameworks exhibit an average accuracy score of 92.81 %, which is 98.32 % for the proposed model. So, for the FR deepfakes, we have shown a performance improvement of 5.51 % which indicates the effectiveness of the suggested work. The key reason for this effective recognition of our approach is due to the improved key points nomination capability of the AUFF-Net network which assists it in robustly identifying the altered regions and improves the deepfakes recognition results of our work. While the other approaches lack to capture the in-depth details of the manipulated samples, therefore, our model is more competent than other techniques.

4.6. Comparison with new methods

In this section, we have executed an evaluation to compare the deepfakes classification results of the introduced work with the latest techniques [56,74–76] employing the same dataset. The detection performance of our model both for the FS and FR deepfakes is compared via employing the standard evaluation metrics namely the accuracy and AUC measures, respectively. The obtained analysis shown in Table 5 elaborates on the effectiveness of our work compared to other techniques. Afchar et al. [56] proposed two approaches namely Meso4 and Mesoinception4 for visual manipulation detection and attained the best results with the second model. Similarly, the work in [74] employed a DL framework for visual manipulation detection with the best accuracy value of 97.17 % for the FR deepfakes. Nguyen et al. [75] proposed a model namely the Capsule network for the accurate identification of digital forgeries made within videos and showed the best accuracy results of 97.80 % for the FS deepfakes. The method in [76] proposed a DL approach for identifying the alterations made within the video by measuring the difference between the facial region and the background. This work showed the highest accuracy of 98.69 % for the FS deepfakes. Whereas, in comparison, the proposed approach attains the highest results for both the FS and FR deepfakes detection. Descriptively, for the FS deepfakes, the nominated approaches attained average accuracy and AUC scores of 89.17 %, and 94.55 %, which are 99.21 %, and 99.86 for the presented strategy. So, in the case of FS deepfakes, we have given the performance improvements of 10.04 %, and 5.31 % for the accuracy and AUC evaluation measures. Similarly, for the FR deepfakes, the peer models have shown the average accuracy and AUC numbers of 89.45 %, and 94.44 %, which is 98.32 %, and 99.41 for our method. Therefore,

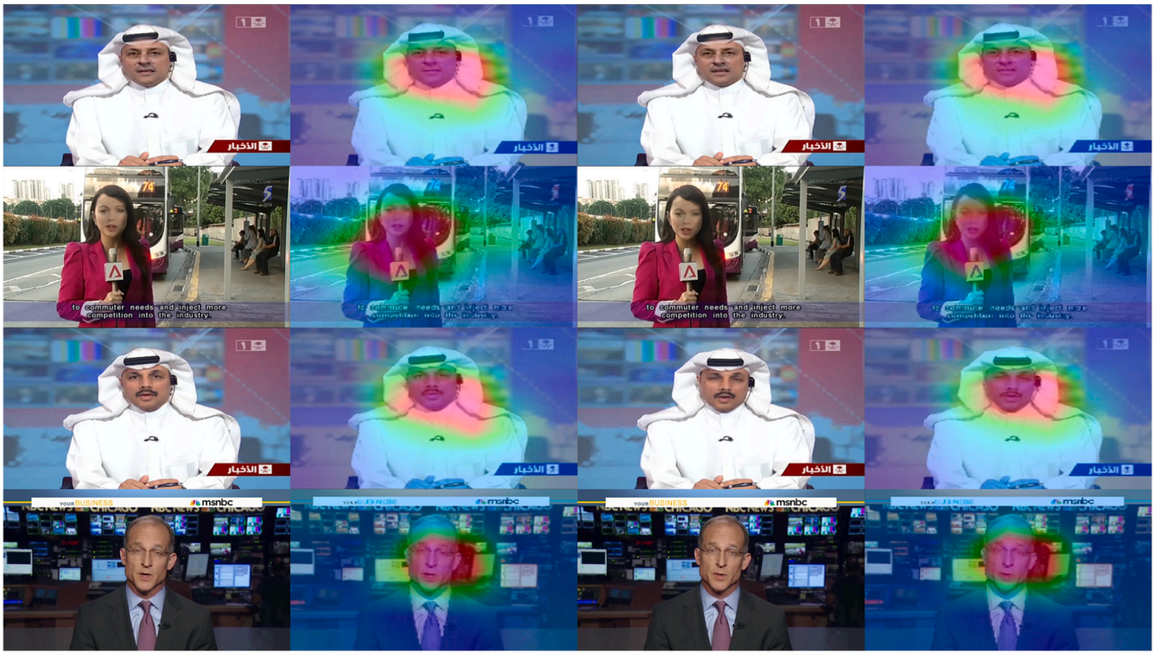


Fig. 17. Visual depiction of the heatmaps attained by the Grad-Cam module of the AUFF-Net model to show the explainability empowerment of the suggested framework. The first four rows show the results of the FS deepfakes detection. While the later four rows show the results of the FR deepfakes detection.

Table 3
Evaluation of the AUFF-Net model in the incidence of adversarial operations.

AUFF-NET results in added adversarial attacks at different compression levels	Precision (%)		Recall (%)		F1-Measure (%)		Accuracy (%)	
	FS	FR	FS	FR	FS	FR	FS	FR
C0	99.59	98.22	99.11	98.71	99.35	98.46	99.23	98.21
C23	99.28	98.11	98.63	97.69	98.95	97.89	99.12	98.13
C40	98.42	98.07	98.21	97.18	98.31	97.62	98.41	98.01

the proposed approach has shown performance gains of 8.87 %, and 4.96 % for the accuracy and AUC measures over the FR deepfakes respectively.

The key factor for the accurate deepfakes detection results of the suggested framework is that the work in [56] is unable to perform well for the low-quality visual content, while the approach in [74] is not robust to highly compressed samples, whereas the methods in [75,76] are suffering from the model over-fitting problem. The proposed approach has better tackled the limitations of existing works by proposing a more effective framework that employs both the spatial and temporal information of video frames and presents the complex content transformations more reliably. Therefore, we can say that our technique is more effective in detecting both FS and FR deepfakes.

4.7. Generalization ability testing

Here, we experimented to test the introduced model in cross-dataset evaluation. The major reason for this experiment is to check the generalization power of our technique. For this reason, we have trained the framework on the FaceForensic++ data sample and tested it on the World Leaders (WLDR) [14] repository. The WLDR [14] dataset contains visual examples of five subjects and varies in length from a range of 10 sec to 2.5 min. Also, the videos are taken at 30 fps in mp4 format. The major reason to select the WLDR [14] dataset is that it contains the samples for both FS and FR deepfakes. The obtained values are exhibited with the help of a boxplot in Fig. 18 as it better provides the elaboration of the obtained values. The reported results in Fig. 18 indicate that our model has undergone performance degradation over the cross-dataset evaluation in comparison to the intra-dataset evaluation case. The major factor for the degradation of model results is due to the fact that the FaceForensics++ dataset was created by employing computer graphics approaches while the other dataset was generated with the help of DL methods. However, we have somehow improved the generalization results which can serve in the field of multimedia forensic investigation. More clearly, we have shown the AUC value of 79.78 % for the FS deepfakes and 78.31 % for the FR deepfakes which elaborates the competence of our network to the unseen cases.

4.8. Discussion

Separating authentic material from artificial intelligence-generated bogus media is of the highest concern. Investigators have made numerous efforts to develop techniques for deepfakes sample identification. We have performed a critical investigation of historic approaches proposed for the reliable recognition of altered visual content as given in Table 1. However, from the analysis provided in Table 1, it can be seen that effective detection and classification of original and manipulated content is a challenging task as the approaches used for creating fabricated data are getting improved and ultimately resulting in the production of more complex and realistic data samples, on which the methods from history may not show effective results. Moreover, the existing approaches are not robust to real-world scenarios and post-processing attacks. Besides, the work from history lacks to provide a better aspect of model explainability which is a major requirement in the area of multimedia forensic analysis. Moreover, there is a need for a more generic model that can detect and classify several types of deepfakes. We have attempted to better tackle the challenges of historic works by proposing a DL model called the AUFF-Net framework.

Table 4
Performance comparison of the AUFF-Net with the base models.

Technique	FaceSwap Accuracy (%)	Face-Reenactment
VGG16-based Bi-LSTM	94.77	92.13
ResNet50-based Bi-LSTM	95.40	93.05
GoogleNet-based Bi-LSTM	91.53	92.19
DenseNet-based Bi-LSTM	96.30	94.35
MobileNetv2-based Bi-LSTM	91.86	92.32
Proposed	99.21	98.32

Table 5
Performance comparison of the AUFF-Net with the latest deepfakes detection techniques.

Technique	FaceSwap		Face-Reenactment	
	Accuracy	AUC	Accuracy	AUC
Afchar et al. (Meso4) [58]	66.31 %	80.04	73.91 %	81.04
Afchar et al. (Mesoinception4) [58]	86.78 %	94.29	81.13 %	94.19
Li et al. [76]	96.25 %	99.07	97.17 %	99.09
Nguyen et al. [77]	97.80 %	99.57	97.48 %	98.93
Pan et al. [78]	98.69 %	99.79	97.57 %	98.97
Proposed	99.21	99.86	98.32%	99.41

The proposed approach presents a unified model that can detect two types of deepfakes namely FS and FR visual manipulations from the original content. The model utilizes both the pixel and temporal features of videos to perform the classification task. At the spatial level, we have proposed a novel CNN framework namely the Inception-swish-Resnet-v2 for the reliable computation of deep features. Whereas, for temporal sequence analysis of videos, we have used the Bi-LSTM approach. Additionally, dense layers are added at the end of the model architecture to nominate the most significant features. Finally, the results are determined based on both the frame level and temporal level information to categorize a video into three classes i.e., real, FaceSwap, and Face-Reenactment, respectively. We have presented extensive class-wise categorization results of the proposed approach to show the generic nature of our model. The AUFF-Net has attained AUC scores of 99.86 and 99.41 for the FS and FR deepfakes.

Moreover, our work is capable of tackling adversarial operations like compression, clutter, blur, zoom, rotation, and translation attacks effectively. We have tested our model on different quality levels with the added perturbations at the test time only and have confirmed through the results that the presented work can be employed to effectively locate the forensic changes under the occurrence of adversarial attacks in visual samples as well. This nature of the proposed work can assist the forensic analyst as the videos uploaded on social media have gone to severe quality reduction to save the network bandwidth. We have further analyzed the internal working of our work to better elaborate its explainability power by generating heatmaps. The computed results show that our work majorly focused on those areas where the manipulations are made. This analysis can assist in processing legal claims where videos can be used as proof. Moreover, to check the adaptability of the introduced work to real-world cases, we have performed a cross-corpus analysis where the model is tested on another dataset. It has been observed that the work has undergone some performance degradation,

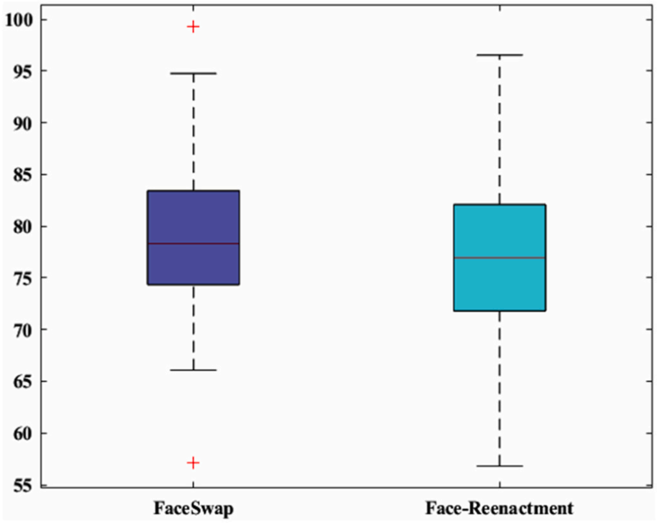


Fig. 18. Cross-corpus AUFF-Net evaluation results in the form of an AUC performance measure.

however, the results are still convincing. So, based on the extensive experimental analysis, we can conclude that the proposed work can play a vital role in the field of multimedia forensic investigation. The major reason for the improved classification performance results of the proposed approach is due to the effective feature engineering ability of the Inception-Swish-ResNet-v2 module that results in extracting the more relevant set of sample features. The key cause for the effective results of the Inception-Swish-ResNet-v2 model is due to the non-monotonic characteristic of the employed activation approach. This nature of the swish method permits the output to decline even for the high input scores which eventually enhances the information storage ability of the introduced framework and authorizes it to compute a robust key point set of underlying visual samples in comparison to historic approaches. Further, we have performed the temporal sequence analysis as well to understand the changing behavior of visual samples with time which also assists in better analyzing the manipulation of videos. Finally, the addition of dense layers before the classification module facilitates the propagation of most related visual characteristics to execute the classification task. Such overall architectural description of the presented work results in high recall, and explainability competence in comparison to other latest approaches and shows robust performance in recognizing the real and altered visual samples. One potential extension of our work is to further enhance the generalization capability of the proposed work and make it generic to other types of deepfakes as well.

As the proposed work is concerned to utilize a pre-trained model and modifying it for the computation of dense features, however, the one drawback of using a pre-trained approach is that it performs resizing of frames before model training which results in losing the morphology of the samples. So, we are also motivated to design a more effective DL strategy as future work to overcome such limitations as well as to further enhance the classification performance.

5. Conclusion

This work has presented an end-2-end DL approach that is empowered to categorize a given video sample as being original, FS, or FR deepfakes. More descriptively, our approach is based on the assumption that the manipulated content not only produces frame-level artifacts but also exhibits extensive variations within frames. Therefore, we have used both spatial and temporal information by proposing a CNN-LSTM approach. More clearly, we have proposed a novel Inception-swish-Resnet-v2 CNN model to generate the dense key points from the input videos at the frame level. While the Bi-LSTM module is used to measure the temporal information. Moreover, we added three FC layers at the last of the network configuration to nominate the most significant features. Lastly, the results are determined based on both the frame level and temporal level information to categorize a video into three classes i.e., real, FS, and FR, respectively. A huge experimental evaluation is performed on the FaceForensic++ dataset to exhibit the robustness of our approach. Besides, cross-dataset evaluation is also utilized to show the generalization capability of our approach. Moreover, the model is capable of performing effectively in the occurrence of several adversarial operations like compressed, noisy, blurred, zoomed, rotated, and translated samples. Hence, it can be said that our framework is competent for visual manipulation classification and can help forensic investigators in better recognition of altered visual content. One potential limitation of our work is to further enhance the generalization capability of the proposed work and make it generic to other types of deepfakes as well. Further, we plan to evaluate our techniques for LIME, SHAP, and ELI5 models to better analyze the explainability power of our model.

Ethical approval

Not applicable

CRediT authorship contribution statement

Ali Javed: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing. **Marriam Nawaz:** Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Aun Irtaza:** Formal analysis, Investigation, Methodology, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Data Availability

Dataset link is shared in the manuscript file

Acknowledgment

This work was supported by the grant of Punjab HEC of Pakistan via Award No. (PHEC/ARA/PIRCA/20527/21). We want to thank Prof. Hany Farid from the University of California Berkeley for providing us with their World Leaders Dataset for performance evaluation.

References

[1] *Reface App*. Available: 〈https://reface.app/〉.
[2] (September 17). *FaceApp*. Available: 〈https://www.faceapp.com/〉.
[3] M. Nawaz, et al., Image authenticity detection using DWT and circular block-based LTrP features, Comput. Mater. Contin. vol. 69 (2021) 1927–1944.
[4] M. Nawaz, et al., Single and multiple regions duplication detections in digital images with applications in image forensic, J. Intell. Fuzzy Syst. vol. 40 (6) (2021) 10351–10371.
[5] M. Masood, M. Nawaz, A. Javed, T. Nazir, A. Mehmood, R. Mahum, Classification of Deepfake videos using pre-trained convolutional neural networks. 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2), IEEE, 2021, pp. 1–6.
[6] C. Chan, S. Ginosar, T. Zhou, A.A. Efros, Everybody dance now, Proc. IEEE Int. Conf. Comput. Vis. (2019) 5933–5942.
[7] H. Kim, et al., Deep video portraits, ACM Trans. Graph. (TOG) vol. 37 (4) (2018) 163.
[8] A.R. Javed, Z. Jalil, W. Zehra, T.R. Gadekallu, D.Y. Suh, M.J. Piran, A comprehensive survey on digital video forensics: Taxonomy, challenges, and future directions, Eng. Appl. Artif. Intell. vol. 106 (2021) 104456.

[9] R.K. Kaliyar, A. Goswami, P. Narang, DeepFakE: improving fake news detection using tensor decomposition-based deep neural network, J. Supercomput. vol. 77 (2) (2021) 1015–1037.
[10] K.-K. Tseng, R. Zhang, C.-M. Chen, M.M. Hassan, DNetUnet: a semi-supervised CNN of medical image segmentation for super-computing AI service, J. Supercomput. vol. 77 (4) (2021) 3594–3615.
[11] I. Priyadarshini, C. Cotton, A novel LSTM–CNN–grid search-based deep neural network for sentiment analysis, J. Supercomput. vol. 77 (12) (2021) 13911–13932.
[12] H. Ajder, G. Patrini, F. Cavalli, and L. Cullen, "The State of Deepfakes: Landscape, Threats, and Impact," in "Amsterdam: Deeptrace," 2019.
[13] I. Goodfellow, et al., Generative adversarial nets, *Adv. Neural Inf. Process. Syst.* (2014) 2672–2680.
[14] M. Nawaz, A. Javed, A. Irtaza, Convolutional long short-term memory-based approach for deepfakes detection from videos, Multimed. Tools Appl. (2023) 1–24.
[15] N. Beuve, W. Hamidouche, O. Déforges, Waterlo: protect images from deepfakes using localized semi-fragile watermark," in *Proceedings of*, IEEE/CVF Int. Conf. Comput. Vis. (2023) 393–402.
[16] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, H. Li, Protecting world leaders against deep fake, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (2019) 38–45.
[17] R. Tolosana, S. Romero-Tapiador, R. Vera-Rodriguez, E. Gonzalez-Sosa, J. Fierrez, DeepFakes detection across generations: analysis of facial regions, fusion, and performance evaluation, Eng. Appl. Artif. Intell. vol. 110 (2022) 104673.
[18] B. Uga, "Towards Trustworthy AI: A proposed set of design guidelines for understandable, trustworthy and actionable AI," ed, 2019.
[19] T.T. Nguyen, C.M. Nguyen, D.T. Nguyen, D.T. Nguyen, and S. Nahavandi, Deep Learning for Deepfakes Creation and Detection, *arXiv preprint arXiv:1909.11573,* 2019.
[20] R. Chesney, D. Citron, Deepfakes and the new disinformation war: the coming age of post-truth geopolitics, Foreign Aff. vol. 98 (2019) 147.
[21] F. Godlee, "Why this US election matters so much," ed: British Medical Journal Publishing Group, 2020.
[22] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, Few-Shot Adversarial Learning of Realistic Neural Talking Head Models," arXiv preprint arXiv: 1905.08233, 2019.
[23] C. Su, W. Wang, Concrete cracks detection using convolutional neuralnetwork based on transfer learning, Math. Probl. Eng. vol. 2020 (2020).
[24] S.D. Olabarriaga, A.W. Smeulders, Interaction in the segmentation of medical images: a survey, Med. Image Anal. vol. 5 (2) (2001) 127–142.
[25] H. Ilyas, A. Javed, K.M. Malik, AVFakeNet: a unified end-to-end dense swin transformer deep learning model for audio–visual deepfakes detection, Appl. Soft Comput. vol. 136 (2023) 110124.
[26] M. Masood, M. Nawaz, K.M. Malik, A. Javed, A. Irtaza, H. Malik, Deepfakes Generation and Detection: state-of-the-art, open challenges, countermeasures, and way forward, Appl. Intell. 53 (2022) 1.
[27] J.F. Boylan, Will deep-fake technology destroy democracy? N. Y. Oct. vol. 17 (2018).
[28] D. Harwell, "Scarlett Johansson on fake AI-generated sex videos: 'Nothing can stop someone from cutting and pasting my image," *Washington Post,* 2018.
[29] Y. Zhang, L. Zheng, V.L. Thing, Automated face swapping and its detection. 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP), IEEE, 2017, pp. 15–19.
[30] X. Yang, Y. Li, 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.
[31] D. Güera, S. Baireddy, P. Bestagini, S. Tubaro, and E.J. Delp, "We Need No Pixels: Video Manipulation Detection Using Stream Descriptors," *arXiv preprint arXiv: 1906.08743,* 2019.
[32] K. Jack, "Chapter 13-MPEG-2," *Video Demystified: A Handbook for the Digital Engineer*, pp. 577-737.
[33] U.A. Ciftci, I. Demir, FakeCatcher: detection of synthetic portrait videos using biological signals, IEEE Trans. Pattern Anal. Mach. Intell. (2020).
[34] T. Jung, S. Kim, K. Kim, DeepVision: deepfakes detection using human eye blinking pattern, IEEE Access vol. 8 (2020) 83144–83154.
[35] R. Ranjan, V.M. Patel, R. Chellappa, Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition, IEEE Trans. Pattern Anal. Mach. Intell. vol. 41 (1) (2017) 121–135.
[36] I. Amerini, L. Galteri, R. Caldelli, A. Del Bimbo, Deepfake video detection through optical flow based CNN, *Proc. IEEE Int. Conf. Comput. Vis. Workshops* (2019), 0-0.
[37] D. Sun, X. Yang, M.-Y. Liu, J. Kautz, Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2018) 8934–8943.
[38] L. Alparone, M. Barni, F. Bartolini, R. Caldelli, Regularization of optic flow estimates by means of weighted vector median filtering, IEEE Trans. Image Process. vol. 8 (10) (1999) 1462–1467.
[39] F. Matern, C. Riess, M. Stamminger, Exploiting visual artifacts to expose deepfakes and face manipulations. 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), IEEE, 2019, pp. 83–92.
[40] T. Baltrušaitis, P. Robinson, L.-P. Morency, Openface: an open source facial behavior analysis toolkit. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2016, pp. 1–10.
[41] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," *arXiv preprint arXiv:1811.00656,* vol. 2, 2018.
[42] D.E. King, Dlib-ML: a machine learning toolkit, J. Mach. Learn. Res. vol. 10 (2009) 1755–1758.
[43] D. Güera, E.J. Delp, Deepfake video detection using recurrent neural networks. 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2018, pp. 1–6.
[44] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking," *arXiv preprint arXiv:1806.02877,* 2018.
[45] D.M. Montserrat, et al., Deepfakes detection with automatic face weighting, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops* (2020) 668–669.
[46] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. vol. 23 (10) (2016) 1499–1503.
[47] O. de Lima, S. Franklin, S. Basu, B. Karwoski, and A. George, "Deepfake Detection using Spatiotemporal Convolutional Networks," *arXiv preprint arXiv:.14749,* 2020.
[48] S. Agarwal, T. El-Gaaly, H. Farid, S.-N. Lim, Detecting Deep-Fake Videos from Appearance and Behavior. 2020 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2020, pp. 1–6.
[49] S. Fernandes, et al., Predicting Heart Rate Variations of Deepfake Videos using Neural ODE, *Proc. IEEE Int. Conf. Comput. Vis. Workshops* (2019), 0-0.
[50] D.J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *arXiv preprint arXiv:.* 2014.
[51] H. Rahman, M.U. Ahmed, S. Begum, P. Funk, Real time heart rate monitoring from facial RGB color video using webcam. in *The 29th Annual Workshop of the Swedish Artificial Intelligence Society (SAIS), 2–3 June 2016, Malmö, Sweden*, Linköping University Electronic Press, 2016.
[52] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, W. Freeman, Eulerian video magnification for revealing subtle changes in the world, ACM Trans. Graph. vol. 31 (4) (2012) 1–8.
[53] R.T. Chen, Y. Rubanova, J. Bettencourt, D.K. Duvenaud, Neural ordinary differential equations, *Adv. Neural Inf. Process. Syst.* (2018) 6571–6583.
[54] S. Kolagati, T. Priyadharshini, V.M.A. Rajam, Exposing deepfakes using a deep multilayer perceptron–convolutional neural network model, Int. J. Inf. Manag. Data Insights vol. 2 (1) (2022) 100054.
[55] G. Mazaheri, A.K. Roy-Chowdhury, Detection and localization of facial expression manipulations, *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.* (2022) 1035–1045.
[56] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, P. Natarajan, Recurrent convolutional strategies for face manipulation detection in videos, Interfaces (GUI) vol. 3 (2019) 1.
[57] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics ++: learning to detect manipulated facial images, *Proc. IEEE Int. Conf. Comput. Vis.* (2019) 1–11.
[58] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network. 2018 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2018, pp. 1–7.
[59] H.H. Nguyen, F. Fang, J. Yamagishi, I. Echizen, Multi-task learning for detecting and segmenting manipulated facial images and videos. 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2019, pp. 1–8.
[60] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, "Forensictransfer: Weakly-supervised domain adaptation for forgery detection," *arXiv preprint arXiv:1812.02510,* 2018.
[61] Y. Xu, K. Raja, M. Pedersen, Supervised contrastive learning for generalizable and explainable DeepFakes detection, *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.* (2022) 379–389.
[62] C.-M. Yu, K.-C. Chen, C.-T. Chang, Y.-W. Ti, SegNet: a network for detecting deepfake facial videos, Multimed. Syst. (2022) 1–22.
[63] N. Patwardhan, M. Ingalhalikar, and R. Walambe, "ARiA: Utilizing Richard's Curve for Controlling the Non-monotonicity of the Activation Function in Deep Neural Nets," arXiv preprint arXiv:.08878, 2018.
[64] Q. Chao, X. Wei, J. Tao, C. Liu, Y. Wang, Cavitation recognition of axial piston pumps in noisy environment based on Grad-CAM visualization technique, CAAI Trans. Intell. Technol. (2022).
[65] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics ++: Learning to detect manipulated facial images, *Proc. IEEE/CVF Int. Conf. Comput. Vis.* (2019) 1–11.
[66] (2018, 14 March 2022). *Deepfakes github.* Available: ⟨http://github.com/deepfakes/faceswap⟩.
[67] Faceswap. Available: ⟨https://github.com/MarekKowalski/FaceSwap/⟩ (2018).
[68] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Nießner, Face2face: real-time face capture and reenactment of rgb videos, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (2016) 2387–2395.
[69] J. Thies, M. Zollhöfer, M. Nießner, Deferred neural rendering: image synthesis using neural textures, ACM Trans. Graph. vol. 38 (4) (2019) 1–12.
[70] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, *Proc. IEEE Int. Conf. Comput. Vis.* (2017) 618–626.
[71] B. Liu, X. Zhang, Z. Gao, L. Chen, Weld defect images classification with vgg16-based neural network. in International Forum on Digital TV and Wireless Multimedia Communications, Springer, 2017, pp. 215–223.
[72] P. Ballester, R. Araujo, On the performance of GoogLeNet and AlexNet applied to sketches, *Proc. AAAI Conf. Artif. Intell.* vol. 30 (1) (2016).
[73] D. Theckedath, R. Sedamkar, Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks, SN Comput. Sci. vol. 1 (2020) 1–7.
[74] Y. Zhu, S. Newsam, Densenet for dense flow. 2017 IEEE international conference on image processing (ICIP), IEEE, 2017, pp. 790–794.
[75] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2018) 4510–4520.

[76] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," arXiv preprint arXiv:.00656, 2018.

[77] H.H. Nguyen, J. Yamagishi, I. Echizen, Capsule-forensics: Using capsule networks to detect forged images and videos. in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 2307–2311.

[78] Z. Pan, Y. Ren, X. Zhang, Low-complexity fake face detection based on forensic similarity, Multimed. Syst. vol. 27 (3) (2021) 353–361.