The Visual Computer (2023) 39:6323–6344
https://doi.org/10.1007/s00371-022-02732-7

ORIGINAL ARTICLE

ResNet-Swish-Dense54: a deep learning approach for deepfakes detection

Marriam Nawaz[1] · Ali Javed[1] · Aun Irtaza[2]

Accepted: 14 November 2022 / Published online: 2 December 2022

Abstract

Development in artificial intelligence has brought a new revolution to technologies and approaches that have been employed for malicious purposes specifically after the introduction of generative adversarial networks (GANs) in 2014. GANs are empowered of generating fake visual samples with high realism. Several refined ML-based methods can produce highly realistic deepfakes videos that can be employed for harassing and blackmailing people. Moreover, deepfakes have introduced political stress by navigating disinformation which can result in societal, and political encounters. The prevailing situation has induced a severe danger to the privacy of humans and thus, urged for the introduction of automated approaches to identify deepfakes. In the presented approach, we have used deep learning (DL)-based approach namely ResNet-Swish-Dense54 for reliable and accurate detection of deepfakes. Initially, human faces are extracted from input video frames. Then, the extracted faces are passed to the ResNet-Swish-Dense54 model to perform the content classification as being real or manipulated. We have evaluated our model over the challenging datasets namely DFDC, FaceForensic++, and CelebDF datasets, and confirmed the robustness of the proposed approach through experimentation. Moreover, we have evaluated our approach for adversarial attacks and proved the explainability power of the ResNet-Swish-Dense54 model by generating heatmaps and performing cross-dataset validation. Both the quantitative and qualitative results demonstrated the effectiveness of our approach for visual manipulation detection.

Keywords Deepfakes · Deep learning · Swish · Visual manipulation · ResNet50

1 Introduction

The easier availability of low-cost digital gadgets like mobile phones, cameras, laptops, tabs, etc., has enabled people to save their memories in digital formats (audio, video, and images) [1–3]. At the same time, internet access has allowed people to share their data via using several social websites like Instagram, Twitter, Facebook, etc., [4, 5]. This information sharing among people has increased the amount of multimedia content in cyberspace. At the same time, several editing tools are available that people usually employ to improve the visual appearance of their data. However, such manipulations also introduce forensic changes in the information conveyed to the people [6, 7]. The great advancement in the field of machine learning (ML) and deep learning (DL) has introduced such tools and applications which can create fake content with extreme realism even undistinguishable by humans [8]. Due to the easy generation and spread of fake data, now it is impossible for humans to classify between the original and altered information which can ultimately cause thought-provoking concerns. Furthermore, the research community has recognized this era as “post-truth” where a chunk of misinformation or disinformation is used by malicious players to influence community belief. The manipulated information has the power to produce intense harm in society: influencing election campaigns, the foundation of warmongering conditions, affecting the reputation of individuals, etc. The word “deepfakes” is defined as the generation of synthetic audiovisual content with the help of DL-based approaches like GAN [8] to spread false narratives about a person [8]. Recently, deepfakes creation has considerably progressed and is usually employed to broadcast disinformation in society which can bring an adverse risk in the shape of

✉ Marriam Nawaz
marriam.nawaz@uettaxila.edu.pk

[1] Department of Software Engineering, UET Taxila, Taxila 47050, Pakistan

[2] Department of Computer Science, UET Taxila, Taxila 47050, Pakistan

false news. In this digital era, multimedia content is used to process legal claims; however, such data demand verification and reliability. Whereas, the extensive alterations introduced in the visual content have made the multimedia data untrusted [9, 10]. Moreover, multimedia forensic analyzers are facing a serious problem in investigating the visual content posted on social sites due to the realistic generation of altered content. Furthermore, the propagation of easy-to-use tools like Zao [11], REFACE [10–12], FaceApp [13], and GAN-based approaches[14, 15] has made the truthfulness and realism verification of images and videos even more complex job.

Visual deepfakes are broadly classified into three types namely the FaceSwap, Lip-synching, and Puppet-master. FaceSwap-based deepfakes are concerned to replace the face of the subject with the target with the aim of portraying the target to do the task that was actually performed by the source person. FaceSwap-based manipulated contents are typically produced for character assassination of well-known people by presenting them in controversial scenarios [16]. Several open-source tools and apps like ZAO [3] and REFACE [4] have the ability to create convincing FaceSwap-oriented deepfakes. With no technical expertise, these applications allow the general audience to exchange their faces with celebrities and see themselves doing those shots. Many other publically available DL-based approaches like DeepFaceLab [5] and FaceSwapGAN [6] are prominent to produce realistic synthesized visual clips. Such user-friendly tools and applications can also generate offensive content like the initial creation of the deepfakes was the generation of non-consensual pornography presenting a serious threat to women [17]. While the Lip-synching deepfakes are focused on synching the lip movements of the target person to some arbitrary audio with the aim to show him/her saying something that is not actually spoken by them. Whereas Puppet-master-oriented deepfakes are concerned to copy the expressions of the target person, like head and eye movement, or mimicking the facial expressions. The main aim of this manipulation is to capture the source's expression [18] in a visual sample. Even though there are many positive applications of deepfakes like these can be used to produce the voice for people who lost their vocal ability [19]. Moreover, drama or movie producers can use these techniques to reproduce the shots for which celebrities are no more available. However, its negative impact is more prominent with the power of claiming a serious threat to the privacy of humans.

Due to the prevailing circumstances of deepfakes, the researchers are now focusing to present such methods which are helpful to classify real and fake content. The methods which are used for deepfakes detection and classification are divided into two types namely the ML-based approaches or DL-based frameworks. In the case of the conventional ML-based feature extraction techniques, Zhang et al. [20] introduced an approach to locate real and fake content. In the first phase, the Speeded up Robust Features (SURF) algorithm was used for keypoints estimation that was later used to train the SVM to accomplish the classification task. This approach was then evaluated over the blurred samples. The method [20] works well for the manipulated images, however, lacks to generalize well to the video-based altered multimedia content. Another approach was introduced in [21] to recognize the forensic changes by approximating the 3D head orientation from 2D facial region features. The calculated variance between the head orientation was employed as a keypoints vector for the SVM training to categorize the real and fake visual content. The technique in [21] shows better deepfakes detection results; however, the performance of this method degrades for the blurred samples. Guera et al. [22] proposed a solution to recognize the synthesized faces from suspected samples. Multimedia stream descriptors [23] were employed for keypoints estimation along with the SVM and random forest to classify the real and fake images. The approach exhibits a low-cost solution to deepfakes detection; however, the detection accuracy reduces for the video re-encoding attacks. A deepfakes detection method was introduced in [24] that utilized the biological signals (e.g., heart rate) computed from the facial region of the input video sample. The computed features were used for the SVM and CNN-based classifiers to locate the original and modified data. The method works well for manipulation detection; however, it is not robust to video post-processing attacks. Jung et al. [25] presented an approach for deepfakes detection by computing the unrealistic eye-blinking pattern from the altered samples. The Fast-HyperFace [26] and EAR method (eye detection) [27] were used to locate the eye-blinking patterns. Then, an integrity authentication approach was utilized by following the variation of eye blinks based on gender, age, behavior, and time factor to locate the pristine and fake data samples. The method presented in [25] works well for visual manipulation identification, however, does not work well for visual samples of persons suffering from mental illness that caused abnormal eye-blinking movements. The conventional ML-based feature extraction techniques lack the ability to tackle the post-processing attacks like the presence of intense light variations, blurring, and compression in visual samples due to their limited feature extraction power [28, 29].

To deal with the limitations of ML-based approaches, the research community is evaluating the power of DL-based approaches for the detection of manipulated content [30, 31]. One such technique was presented in [32] where a supervised learning-based method was used for video forensic analysis. More clearly, the Xception network together with a supervised constructive loss was employed to learn the features from the input samples that were later classified as being original or modified. This approach works well for deepfakes detection; however, the evaluation power should be tested over a more challenging dataset. Another framework was proposed in [33] that employed the fusion of both landmarks and deep features to recognize the real and manipulated samples. The work [33] exhibits improved performance for deepfakes identification, however, does not generalize well to dark light visual samples. Roy et al. [34] employed three types of DL-based frameworks namely the 3D ResNet, 3D ResNeXt, and I3D to identify the visual manipulations. The framework introduced in [34] acquires better deepfakes detection results for the 3D ResNeXt network, however, lacks to generalize well for the unseen testing samples. Another work was proposed in [35] where the information from both the frame level and temporal sequence analysis was used for deepfakes detection. The work in [35] demonstrates better visual manipulation categorization results, however, does not work well for the compressed video samples. Chen et al. [36] introduced an approach for recognizing pristine and fake videos. A two-step technique named mask-guided identification and reconstruction was employed to detect the altered visual data. In the first step, the deep keypoints were calculated that were later utilized iteratively to detect the fake samples. The approach proposed in [36] is robust to deepfakes classification; however, it is unable to tackle adversarial attacks. Moreover, a technique is proposed in [37] that employed a 3D CNN approach for deepfakes detection. The approach [37] exhibits better visual manipulation results, however, with the increased computational burden. Masood et al. [38] presented a framework for deepfakes detection and classification that used numerous pre-trained frameworks for keypoints computation. Then, the extracted features were used for training the SVM classifier to categorize the real and fake videos. The approach proposed in [38] exhibits the best performance for the DenseNet-169 approach; however, it suffers from a higher computing cost.

Despite extensive work presented for the accurate detection of deepfakes, still there is a need for performance enhancement. The existing works show degraded performance for samples suffering from adversarial attacks like noise, compression, light variations, blurring, scale and position variations, etc. Moreover, the existing techniques are robust to trained data, however, showed less performance for unseen scenarios. Moreover, the creation of manipulated content with huge realism is also imposing a demand for a more accurate approach to the reliable identification of forged samples. We have tried to overcome the issues of existing works by presenting a novel DL approach namely the ResNet-Swish-Dense54 model. The usage of the Swish activation approach performs a multiplication method on input values by utilizing the sigmoid function. The employed activation approach smoothly changes the flow of negative values in place of sudden change and supports a small range of negative values to pass through the framework which enables our framework to learn the complicated patterns of input videos with a high recall rate. Moreover, the inclusion of added dense layers at the end of our ResNet-Swish-Dense54 framework allows it to better nominate an effective set of sample features for the classification task. Initially, the video samples are extracted on which the OpenFace2 toolkit is applied for human face detection. The detected faces are later passed to the ResNet-Swish-Dense54 model for deep features computation and visual manipulation classification. The main contributions of our work are as follows:

- A novel ResNet-Swish-Dense54 model is presented that improves the deepfakes detection accuracy by introducing a very little computational burden.
- Reliable and accurate identification of deepfakes due to the capability of the proposed approach to tackle the model over-fitting.
- Better model explainability power because of the reliable feature selection ability of the ResNet-Swish-Dense54 model.
- The presented work is reliable to perform well under the presence of adversarial attacks like compression, noise, blurring, translation, and rotational variations due to the capability of ResNet-Swish-Dense54 to propagate a more relevant set of visual features within neurons and perform the classification task.
- We have evaluated the proposed solution over the complex Deepfakes Detection Challenge Dataset (DFDC), FaceForensic++, and CelebDF datasets including a cross-corpus evaluation to elaborate on the effectiveness and generalization power of the ResNet-Swish-Dense54 framework for visual manipulation detection.

The remaining article follows the following structure: Sect. 2 explains the details of the proposed approach, while the evaluation results along with the performance measurement metrics and the employed dataset are given in Sect. 3. Finally, the conclusion is drawn in Sect. 4.

2 Proposed method

In this work, we have presented a novel ResNet-Swish-Dense54 framework. To effectively capture the underlying complex patterns of videos, we have introduced residual blocks with the Swish activation method. Such architecture of the proposed approach permits the flow of small negative values through the network and optimizes the model learning behavior. Moreover, we have introduced the extra dense layers at the end of the framework architecture to nominate a more reliable set of sample features for classification purposes. More clearly, initially, we have extracted the video frames from which the subject faces are extracted by using the OpenFace2 [39] toolkit. Then, the extracted faces

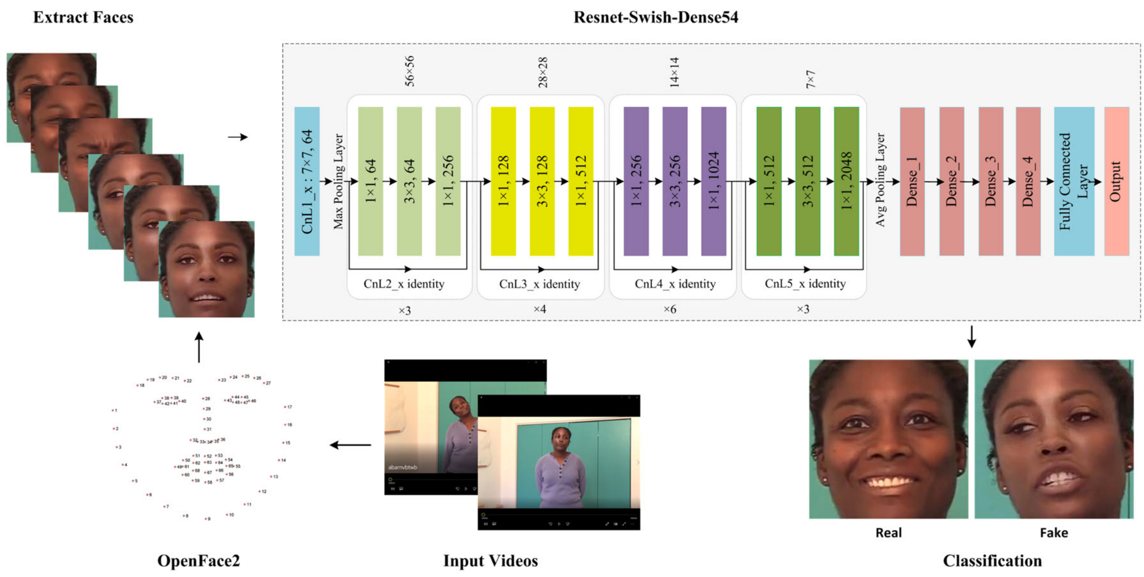


Fig. 1 Visual representation of ResNet-Swish-Dense54

are passed to the ResNet-Swish-Dense54 model to compute the deep features and classify the samples either as real or manipulated. A detailed visual representation of the proposed approach is shown in Fig. 1.

2.1 Identification and extraction of human faces from videos

After extracting the video frames, the next step is to locate the faces from them. For visual manipulation, the face part is the main area of video frames in which the alterations are made. Therefore, the presented method is focused on the facial region only. To obtain the faces, we have used a face detector namely the OpenFace2 [39] toolkit. The employed tool utilizes 2D and 3D face region landmarks to locate the facial area. Furthermore, the OpenFace2 has the ability to compute the head position, eye-gaze, and localization of mouth action units as well. The main reason to select the OpenFace2 for face detection is that this approach is effective to locate faces even under the occurrence of changes in the face orientation, intensity variations, and capturing device position. Such characteristics of the OpenFace2 toolkit enable it to accurately identify the human faces from the video samples under intense transformation changes [40]. Moreover, to maintain the computational complexity of the proposed solution, we have only taken 20 frames from all video samples.

2.2 Feature computation

After the extraction of faces from the video frames, the next task is to compute features from them. For this reason, we have used a well-known pre-trained model namely the ResNset50 and altered it by introducing the Swish activation method inside its structure. Moreover, additional dense layers are introduced at the end of the model architecture. The intuition of using the swish activation method is that it permits the model to propagate the negative values through neurons which assist to capture the complex underlying visual patterns while the additional dense layers allow nominating a representative set of features to be passed for classification. The basic purpose of selecting a pre-trained model is that the framework is already trained on an online available large dataset namely the ImageNet database and has the power to compute a more reliable set of image features. The starting layers of the model are responsible to learn the low-level image information while the later layers are focused to compute the job-specific information. So, the employment of a pre-trained model for a new task like using for deepfakes detection causes to increase in the fake recognition accuracy and reduces the execution time by fastening the training process. A visual demonstration of this task is shown in Fig. 2.

2.3 ResNet50

ResNet50 [41] is a well-known DL model that uses identity shortcut links along with the residual mapping in its entire architecture to acquire better performance results. Usually, the traditional CNN methods use the information of all previous layers to extract a dense set of image features to improve their object recognition accuracy. A pictorial representation of the original ResNet50 model is shown in Fig. 3. However, by increasing the network depth, the models with such architecture settings face a serious degradation in their

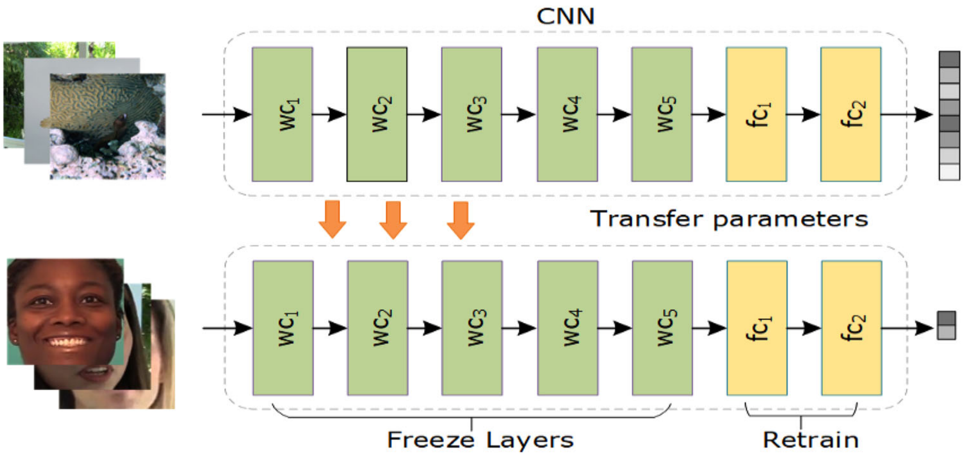


Fig. 2 A visual demonstration of transfer learning

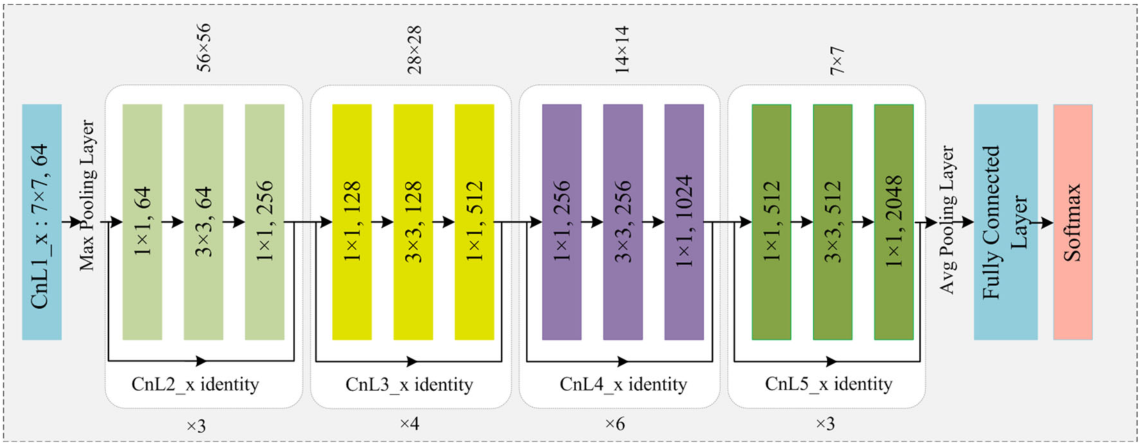


Fig. 3 The architectural description of ResNet50

performance because of the gradient vanishing issue in the training procedure [4]. To deal with the problems of existing CNN models, the ResNet approach proposes the concept of using skip connections for deep network architectures that miss one or more layers and build the foundation of residual blocks (RBs). The resultant structure permits reemploying the keypoints maps from the previous layers, which results in improved performance and easier training. The RB is the basic building block of the ResNet model, and a visual representation is shown in Fig. 4.

The RB contains numerous convolution layers, along with the ReLU activation function. Furthermore, it contains a batch normalization layer and a shortcut link. In each RB, the stacked layers are responsible for residual mapping via using shortcut links that execute identity mapping (i). The obtained result is joined with the output function of the stacked layers. The outcome from the RB can be stated as:

$$Y = F(i) + i \tag{1}$$

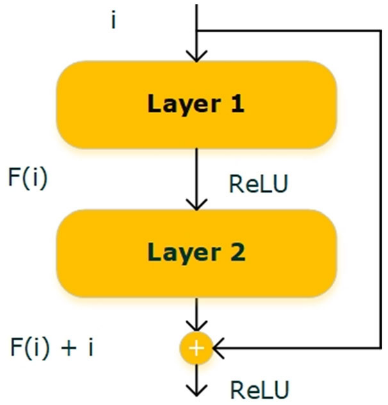


Fig. 4 The architectural description of RB

Here, I represent the input, F shows the residual function while Y is demonstrating the result obtained from the residual function.

2.4 ResNet-Swish-Dense54

The reliable computation power of the ResNet model has inspired us to employ it for recognizing real and altered visual samples. We have customized the existing ResNet50 model by introducing the swish activation method in its structure along with four additional added dense layers at the end of the architecture.

The motivation for using the swish activation method in the original model is to improve its learning behavior via exhibiting the least training loss and to enhance its empowerment to recognize complicated video patterns. Moreover, the addition of added dense layers in the ResNet50 model enhances the keypoints nomination ability of the network. The in-depth view of the novel ResNet-Swish-Dense54 model is shown in Fig. 5, while the structural description is given in Table 1. The framework comprises a total of 33 convolutions layers (CnL), which are further clustered into 5 CnL phases, where each group contains numerous RBs positioned on top of one another. We further added the global average pooling layer along with the four added dense layers. The dense layers permit the network to emphasize the manipulated regions while removing unrequired background information and enhancing deepfakes detection results under changing complex conditions, such as variations in light, color, and face orientations. The added dense layers enhance the capability of the model to learn a reliable set of image features by introducing very little overhead to model architecture. Once the features are selected by the added dense layers, these are forwarded to the softmax layer to produce the final results. A detailed description of the inner model layers is discussed in the succeeding sections.

2.4.1 Convolutional layer

In all CNN models, the convolutional layer is focused to calculate a set of deep keypoints from the input sample. The mathematical description of this layer is given in Eq. (2).

$$K_i^t = f\left(\sum_{j \in n_i}\left(Y_{ji}^t * K_j^{t-1} + \eta_i^t\right)\right) \tag{2}$$

In Eq. (2), t is denoting the total number of model layers, whereas K is representing the attained keypoints set by applying the kernel window of size Y. Moreover, the convolution operation is shown by *, while η is representing the bias value. Furthermore, n_i indicates the total number of feature maps. In the presented work, all the extracted faces attained from the OpenFace2 are set to a size of 224 × 224 to make them compatible with the framework requirements. The ResNet-Swish-Dense54 model resized the samples with the aim of increasing them in depth to extract a more representative set of sample information. The used convolution layers in the entire structure of the ResNet-Swish-Dense54 model have influenced the depth of the extracted samples. The proposed ResNet-Swish-Dense54 model has a total of 48 convolution layers which are directed to attain a nominative set of features from each input video sample.

2.4.2 Activation layer

To enhance the visual manipulation recognition ability of the proposed approach, we have presented the Swish activation method as an alternative to ReLU in our ResNet-Swish-Dense54 model after each convolution 2D layer. The swish activation approach is non-monotonic smooth unbounded above and bounded below in its behavior. Such nature of the swish activation method assists to prohibit saturation and model over-fitting problems. Smoothness assists the model to optimize its behavior by improving the recall ability which in turn enhances the generalization power of an approach. Whereas, the non-monotonic nature supports the easier flow of gradient and delivers robustness to varying learning rates. The swish activation method is simple in nature, and several studies reveal that it performs well in comparison with the most widely employed ReLU activation method in complicated research areas of image classification and object recognition [28]. The major reason for the better performance of the swish method is that the ReLU function prohibits the flow of negative values through the model which causes the loss of significant sample information. While the swish method permits the flow of small negative values through the model which are significant for computing complex patterns from the input data in deeper networks. A visual depiction of swish and ReLU activation methods is given in Fig. 6. The mathematical representation of the swish method is given as:

$$s(i) = i \times \text{sigmoid}\,(\text{£}i) \tag{3}$$

In Eq. (3), i is depicting the value of input and £ shows the trainable model parameters. Moreover, the Swish method optimizes model learning behavior and also reduces the computation complexity of the model as the employed activation approach takes less time for training as compared to other activation methods.

2.4.3 Pooling layer

This layer is employed to reduce the high dimensional keypoints vector by eliminating the unwanted information. The pooling layer assists the proposed model to avoid the overfitting problem by showing the summation of the pixel information and makes the model robust to spatial translations of the input. In the proposed approach, the computed

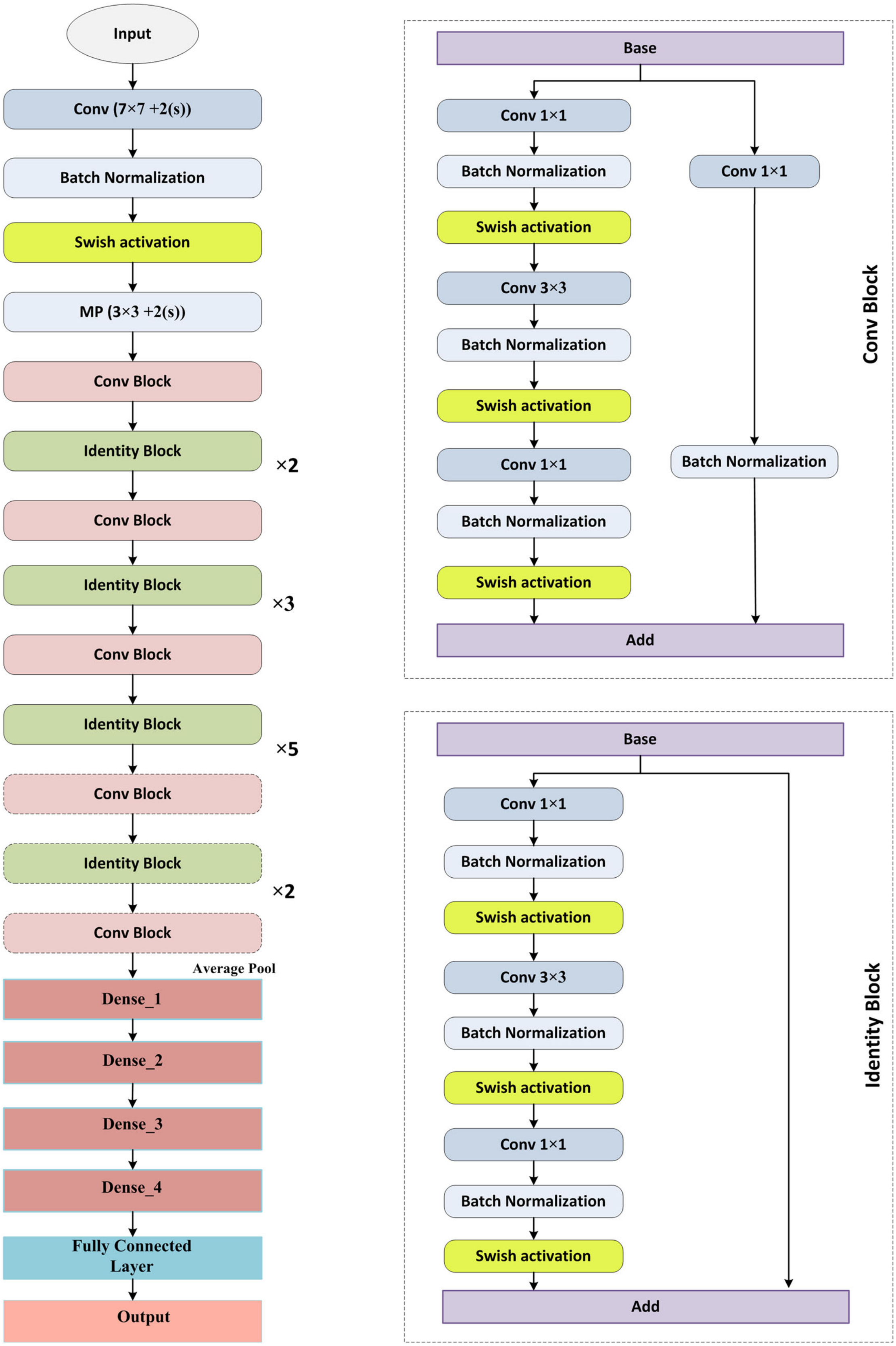


Fig. 5 The architectural description of the ResNet-Swish-Dense54

Table 1 Structural description of ResNetDense-54

Layer name	ResNet-Dense
CnL1_x	7 × 7, 64 3 × 3 max pool
CnL2_x	$\begin{bmatrix} 1\times 1, 64 \\ 3\times 3, 64 \\ 1\times 1, 256 \end{bmatrix} \times 3$
CnL3_x	$\begin{bmatrix} 1\times 1, 128 \\ 3\times 3, 128 \\ 1\times 1, 512 \end{bmatrix} \times 4$
CnL4_x	$\begin{bmatrix} 1\times 1, 256 \\ 3\times 3, 256 \\ 1\times 1, 1024 \end{bmatrix} \times 6$
CnL5_x	$\begin{bmatrix} 1\times 1, 512 \\ 3\times 3, 512 \\ 1\times 1, 2048 \end{bmatrix} \times 3$
Global Average Pooling	
Dense_1	
Dense_2	
Dense_3	
Dense_4	
Fully connected layer	
Output layer	

features are acquired from the average pool layer and then passed to the added dense layers.

2.5 Additional dense layers

After the pooling layer, we added four additional dense layers with the ReLU activation method at the end of the model architecture. The additional layers at the end of the model architecture are responsible for emphasizing the manipulated regions of the input samples by removing the least wanted information and enhancing the visual deepfakes detection results under altering transformation scenarios, like under the occurrence of position, orientation, light, and color variations. The introduced dense layers optimize the capability of the ResNet-Swish-Dense54 model to calculate a more competent set of video frame features by introducing a minute overhead to the framework architecture. Lastly, the extracted information in the form of deep keypoints is propagated to the softmax layer.

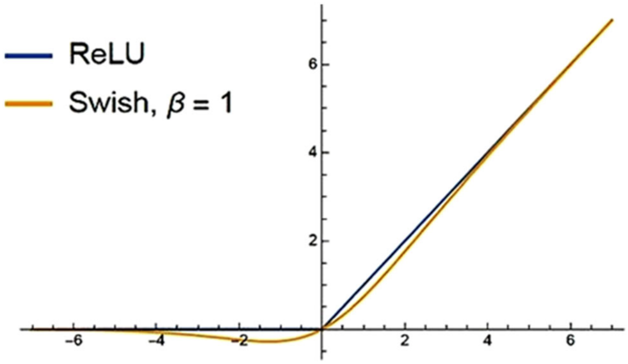


Fig. 6 ReLu vs. Swish activation method [42]

2.6 Softmax layer

The softmax layer is the last layer of the proposed work that is responsible for accomplishing the deepfakes classification task. The proposed approach utilized a softmax activation function in the last fully connected (FC) layer to estimate the proportional possibility of the output units. The mathematical description of the softmax is given in Eq. (4).

$$\delta(I_x) = \frac{\exp(I_x)}{\sum_{m=0}^{n-1} \exp(I_m)} \tag{4}$$

In Eq. (4), both (I_x) and (Z_m) are showing the input and output vectors, respectively, whereas m denotes the output classes.

3 Experimental setup and results

In this part, we have defined the performance metrics that are used to measure the deepfakes detection accuracy of the proposed model. Furthermore, we have demonstrated the details of the employed dataset. Moreover, comprehensive experimentation is illustrated to explain the effectiveness of our approach.

3.1 Evaluation metrics

To assess the deepfakes detection performance of the proposed approach, we have employed several standard metrics namely precision (Pr), recall (Re), accuracy (Ac), and F1 score. The mathematical explanation of employed metrics is explained in Eqs. (5) to (8).

$$\mathrm{Pr} = \frac{d'}{d' + \gamma} \tag{5}$$

$$\mathrm{Re} = \frac{d'}{d' + Q} \tag{6}$$

$$\mathrm{Re} = \frac{d' + \acute{r}}{d' + \acute{r} + \gamma + \mathcal{Q}} \tag{7}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{Pr} \times \mathrm{Re}}{\mathrm{Pr} \ + \mathrm{Re}} \tag{8}$$

Here, d' shows the true positives (deepfakes samples), and $\acute{r}$ denotes true negatives (original samples). While, γ determines the false positives (false real), and $\mathcal{Q}$ denotes false negatives (false deepfakes), respectively.

3.2 Dataset

For performance evaluation, we considered two challenging datasets namely the DFDC database published by Facebook [43] and FaceForensic++ (FF++) [44]. The DFDC database comprises 1131 pristine videos and 4119 altered visual samples. The manipulated samples are created by utilizing the two unknown approaches. The DFDC data are an online and publicly accessible dataset and can be downloaded from the Kaggle competition website [43]. While the FF++ dataset is also a standard and large-sized database of visual manipulations. The FF++ dataset comprises a total of one thousand real and four thousand manipulated samples of varying subjects. The samples are altered by employing numerous forensic approaches like DeepFakes [45], FaceSwap [8], Face-Reenactment [46], and Neural Textures [47]. Furthermore, the samples in this dataset are present with three different quality levels. For our work, we have tested our approach for the FaceSwap and Face-Reenactment deepfakes at all quality levels. The used databases are distributed randomly into 70–30 parts where 70% of data is employed for model training, whereas the remaining 30% of unseen samples are used for model verification. Few samples of real and fake samples from the employed datasets are shown in Fig. 7.

3.3 Implementation details

The model is implemented in Python language with TensorFlow library and executed on a system containing Nvidia GTX1070 GPU. The employed database is distributed randomly into 70/30 chunks to produce two separate sets named the training and test sets, respectively. Moreover, the same number of real and fake samples are used to maintain the class balance. For the DFDC dataset, the model is trained to detect the samples as real and fake classes, respectively, while for the FaceForensic++ dataset, the framework is tuned to detect the FaceSwap and Face-Reenactment deepfakes classes, respectively. The following settings are designed to successfully execute the model:

(i) Subtracting channel mean from each channel.
(ii) The detected face samples are resized to 224-by-224 dimensions as per model requirements.
(iii) We have trained the model for 40 epochs, and the learning rate is set to 0.0001.

To show the model behavior at training time, we have discussed both the train time deepfakes detection accuracy and training loss as these parameters help to show the model behavior during the training procedure. We have shown the visual representation of the loss graph in Fig. 8a for the presented approach. Figure 8a clearly shows that the presented work acquired an optimal value of 0.0013 at the epoch number of 40, which is demonstrating the efficient learning of the proposed technique. Moreover, the ResNet-Swish-Dense54 model attains the highest validation accuracy of 99.99% as shown in Fig. 8b.

3.4 Evaluation of the proposed method

A reliable forensic analysis model must be capable of identifying the visual manipulations with high performance. To test the deepfakes detection capability of the work, we conducted

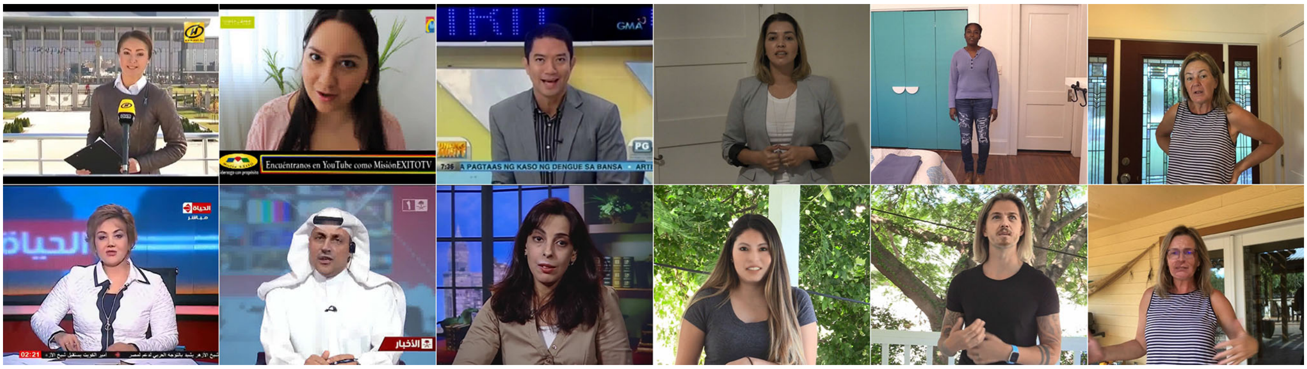


Fig. 7 Samples from the employed datasets, where the first row contains real images, while the second row shows manipulated samples

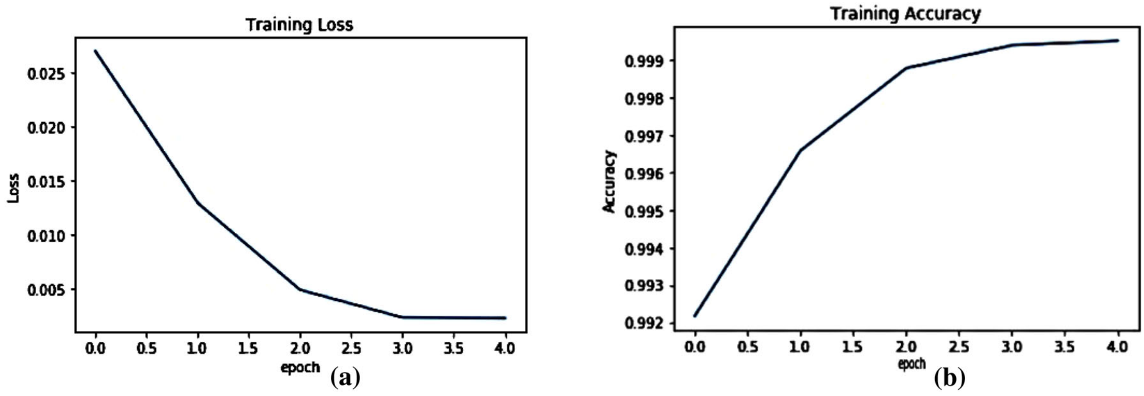


Fig. 8 Visual representation of training loss graph

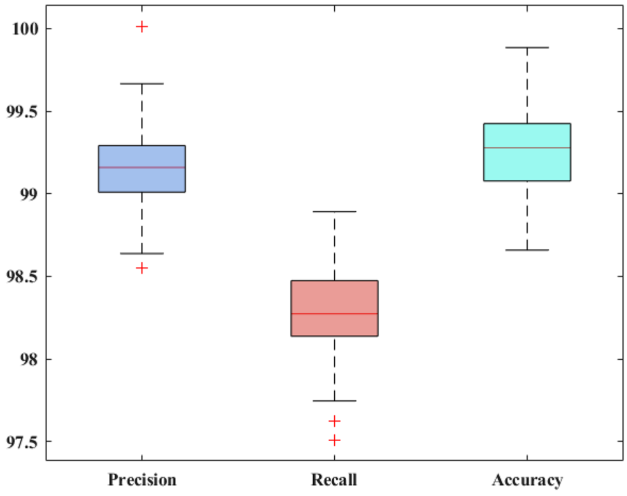


Fig. 9 Performance analysis of the proposed approach in terms of precision, recall, and accuracy over the DFDC dataset

an experiment in this section where we discussed the classification results of the proposed model. For this reason, we chose several standard metrics to demonstrate the deepfakes recognition performance.

To show the classification results of the proposed approach on the DFDC, and FaceForensic++ dataset, we have plotted the boxplots as these better show the results by showing the minimum, maximum and median values. Initially, we have shown the test performance over the DFDC dataset in terms of precision, recall, and accuracy, and obtained results are shown in Fig. 9. More specifically, we have attained the average precision, recall, and accuracy values of 99.18, 98.24, and 99.26%, respectively, which is showing the robustness of the proposed framework for deepfakes detection.

While for the FaceFornensic++ dataset, the acquired results are shown in Fig. 10. It can be seen from the values shown in Fig. 10 that the presented framework has exhibited robust performance for classifying both FaceSwap and Face-Reenactment deepfakes. More clearly, for the FaceSwap deepfakes, the proposed approach has attained precision, recall, and accuracy scores of 99.88, 98.76, and 99.13%. While for the Face-Reenactment deepfakes, we have acquired the precision, recall, and accuracy scores of 98.23, 97.78, and 98.08%, respectively, which is clearly indicating the robustness of our approach to visual manipulations. The results explained in Figs. 9 and 10 are clearly showing that our approach is effective to classify the visual manipulations. The basic reason for the robust performance of our model is due to the better keypoints extraction and selection ability of the ResNet-Swish-Dense54 model which causes the framework to present the complicated visual patterns in a viable manner.

3.5 Explainability

To assess the authenticity of the multimedia content, the role of the forensic analysis frameworks is very crucial, particularly for situations where such content can be used as proof to verify a legitimate claim. Therefore, there is a demand from such automated systems to explain their internal working by reporting the regions from the suspected visual samples that urged the system to label them as being original or manipulated. To check this, we performed an analysis to validate the explainability power of the proposed approach. For this reason, we have shown the heatmaps corresponding to the last layer of the introduced model. It is clear from Fig. 11 that the ResNet-Swish-Dense54 model focuses on those image regions where the manipulation exists. The main reason for the effective recognition power of the novel ResNet-Swish-Dense54 model is the reliable feature computation ability of our work as the introduced activation method propagates a small number of negative values in the model as well which assists it in better learning the complex visual artifacts and better capture the transformation changes of visual content.

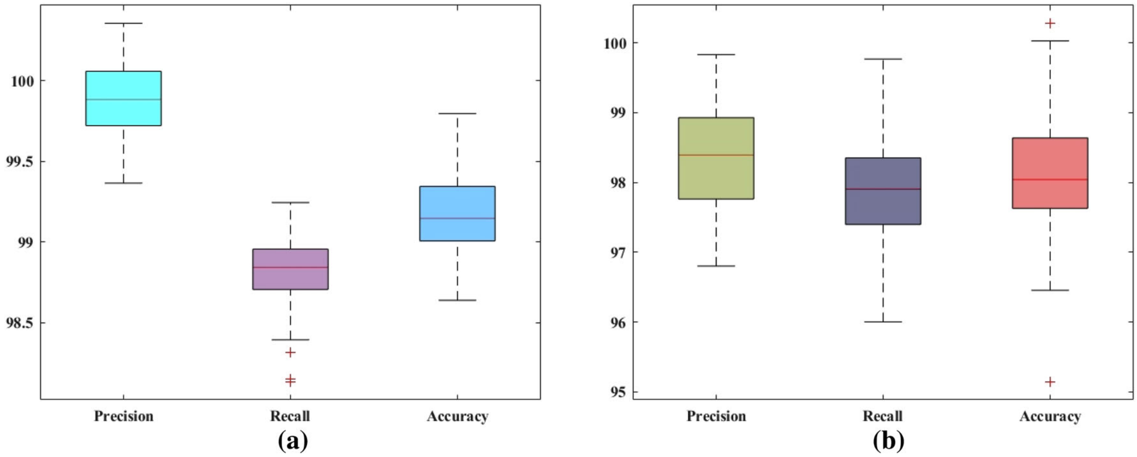


Fig. 10 Performance analysis of the proposed approach in terms of precision, recall, and accuracy over the FaceForensic++ dataset, **a** FaceSwap, and **b** Face-Reenactment deepfakes detection results

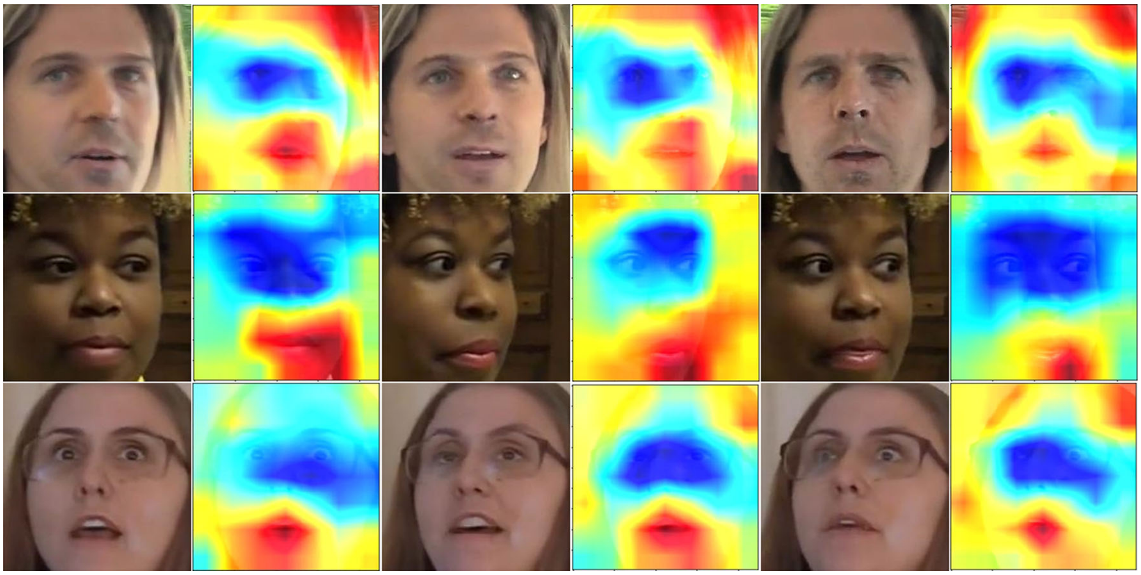


Fig. 11 Heatmaps visualization where the red color corresponds to where the manipulation exists

Table 2 Comparative analysis with the base model over the DFDC dataset

Model	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
ResNet50	95.30	95.10	95.20	97
Proposed	99.18	98.24	98.71	99.26

3.6 Comparison with the base model

In this section, we have compared the visual manipulation detection performance of the proposed approach against the original ResNet50 model. Initially, we have reported the results for the DFDC dataset and the comparison both in terms of detection accuracy and network architecture complexity is shown in Table 2. It is clear from reported results that in terms of deepfakes identification power, the proposed solution is more competent than the original network. More descriptively, for the precision evaluation metric, the original ResNet50 model shows an average value of 95.30% which is 99.18% for our model. Hence, for precision, the presented framework shows an average performance gain of 3.88%. Similarly, for the recall and F1-score measures, we have presented the average performance gains of 3.14, and 3.51%, respectively. While, for the accuracy metric, the proposed model acquires a performance gain of 2.26%.

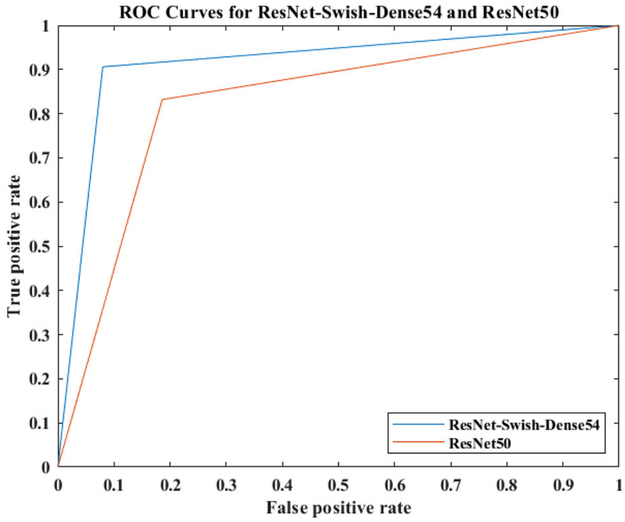


Fig. 12 AUC-ROC curves for the base and proposed models over the DFDC dataset

To further explain the visual alteration classification results of the presented method over the DFDC dataset, we have reported the AUC-ROC curves of both the original and ResNet-Swish-Dense54 models (Fig. 12). The AUC-ROC curve is an important performance estimation metric for checking the categorization power of any framework. In the AUC-ROC curve, the ROC demonstrates the likelihood curve while the AUC part exhibits the calculation of separability. In the proposed solution, the AUC-ROC curve is describing the capability of the introduced method to recognize the real and fake samples. For the AUC-ROC curve, the value proceeding toward 1 is showing the ability of the model to better recall the original and manipulated data. It is quite clear from Fig. 12 that the proposed solution is more competent than the base model.

Furthermore, for the FaceForensic++, the comparison with the base model is given in Table 3 which is clearly showing the effectiveness of our approach. We have attained the highest results for both FaceSwap and Face-Reenactment deepfakes in form of all performance measures in comparison with the ResNet-50 model. Descriptively, in the case of the FaceSwap deepfakes detection, we have reported performance gains of the 5.72, 6.57, 6.16, and 5.24% for the precision, recall, F1-score, and accuracy metrics, respectively. Similarly, for the Face-Reenactment deepfakes, we have attained performance gains of 4.82, 5.75, 5.30, and 6.32% for the precision, recall, F1-score, and accuracy metrics, respectively. Moreover, we have shown the comparative AUC-ROC curves for the FaceSwap and Face-Reenactment deepfakes in Fig. 13 which is clearly depicting the robustness of our approach in comparison with the base model.

From the conducted analysis, it is quite evident that the proposed ResNet-Swish-Dense54 model performs well for

Table 3 Comparative analysis with the base model over the FaceForensic++ dataset

Model	Precision (%)		Recall (%)		F1-score (%)		Accuracy (%)	
	FaceSwap	Face-Reenactment	FaceSwap	Face-Reenactment	FaceSwap	Face-Reenactment	FaceSwap	Face-Reenactment
ResNet-50	94.16	93.41	92.19	92.03	93.16	92.71	93.89	91.76
ResNet-Swish-Dense54	99.88	98.23	98.76	97.78	99.32	98.01	99.13%	98.08

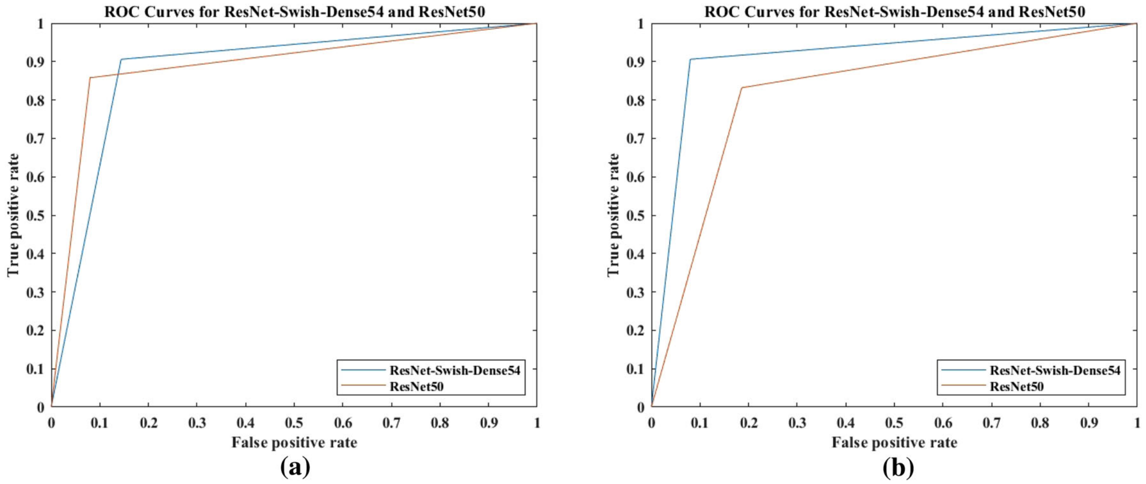


Fig. 13 AUC-ROC curves for the base and proposed models over the FaceForensic++ dataset

both datasets as compared to the base model. Even though the original ResNet50 model has fewer model parameters of 23.5 million which is 28.2 million for the presented framework, however, it is quite evident from the reported results that with a little added model complexity the proposed solution exhibits robust performance for deepfakes detection. The major reason for the better manipulation recognition performance of the proposed solution is the inclusion of the Swish activation method in the model which results in accurate feature computation. Moreover, the added dense layers assist the model in better tackling the network over-fitted data and ultimately enhances the performance.

3.7 Comparison with other activation methods

To better examine the effect of different activation functions when applied to our model, we conducted an experiment to implement our model with the ReLU, LeakyReLU, PReLU, and swish activation functions separately and evaluated the performance of our approach for both DFDC and FaceForensic++ datasets. The reason to choose the ReLU, LeakyReLU, and PReLU, activation methods for performance comparison is due to their higher employment and suitability in several image classification problems. The attained results for both DFDC and FaceForensic++ datasets are shown in Tables 4 and 5, respectively. It is quite evident from the values reported in Tables 4 and 5 that the proposed ResNet-Swish-Dense54 model performs well in comparison with all other activation approaches. The main reason for the effective deepfakes detection performance of the proposed novel ResNet-Swish-Dense54 model is because of the non-monotonic nature of the Swish activation method. Such a behavior of the employed activation approach allows the computed values to still fall even if the input rises which ultimately improves the computed values storage capacity of the proposed approach. So, the employment of the Swish activation method optimizes the model behavior by improving the feature selection power and recall ability of the proposed approach. Comparatively, the rest of the activation methods lack that ability; hence, the proposed ResNet-Swish-Dense54 model presents the highest performance results for the classification of visual manipulations.

3.8 Performance evaluation under the occurrence of adversarial attacks

A basic issue faced by the deepfakes detection approaches is the presence of adversarial attacks which are added for the purpose of fooling the detectors [48–51]. The multimedia content undergoes severe quality reduction before uploading on social sites to save bandwidth. Furthermore, several other attacks like noise, blurring, and size transformation changes are developed in the fake content to mislead the detection techniques. So, for an effective visual manipulation framework, it must be capable of dealing with such visual perturbations. For this reason, we performed an analysis to check the forensic alteration detection performance of our approach under the occurrence of such perturbations.

The employed datasets named DFDC and FaceForensic++ contain samples of varying quality levels with huge brightness changes. Besides this, we further introduced other adversarial attacks like blurring, noise, zooming, and rotational changes in the visual samples of both datasets. *It is worth stating that all the mentioned adversarial attacks are introduced only during the model's testing.* The basic reason to add such attacks only at the test time is to check the robustness of the proposed framework without training it in such scenarios. To add perturbations in the visual samples of the DFDC and FaceForensic++ datasets, we have modified the videos by translating them into dimensions with a

Table 4 Comparative analysis with the other activation methods over the DFDC dataset

Activation method	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
ReLU	99	98.10	98.50	99.10
LeakyReLU	99.01	98.12	98.56	99.11
PReLU	99.06	98.10	98.58	99.14
Swish	99.18	98.24	98.71	99.26

Table 5 Comparative analysis with the other activation methods over the FaceForensic++ dataset

Activation method	FaceSwap				Face-Reenactment			
	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
ReLU	99.23	98.15	98.69	98.65	97.83	97.58	97.70	97.41
LeakyReLU	99.31	98.29	98.80	98.70	97.94	97.61	97.77	97.54
PReLU	99.42	99.40	99.41	98.74	98.09	97.65	97.87	97.68
Swish	99.88	98.76	99.32	99.13	98.23	97.78	98.01	98.08

range of [− 2, 2] and accomplished the zoom and rotation perturbations with a range of [− 0.2, 0.2].

Furthermore, noise and blurring attacks are introduced in the visual data with varying kernel window sizes, i.e., 5,7, and 9, etc., to enhance the variety of the test data. The samples from the DFDC and FaceForensic++ datasets are tested for adversarial attacks, and attained performance results are discussed in Table 6. The values shown in Table 6 are clearly showing that the presented work is effective to tackle these attacks. More descriptively, for the DFDC dataset we attained an average accuracy of 96.89%, while for the FaceForensic++ dataset, we have acquired the average accuracy scores of 98.75, and 96.62% for the FaceSwap, and Face-Reenactment deepfakes which are clearly depicting the robustness of our approach. So, we can conclude from this evaluation that the proposed ResNet-Swish-Dense54 framework is capable of dealing with different types of adversarial attacks. The basic reason for the better performance of our approach is due to the capability of swish activation that makes the model pass neurons with more relevant features even under the incidence of noise, blurring, rotation, zooming, and light variations. Moreover, the inclusion of dense layers helps to pass a more representative set of sample features to perform the classification task.

3.9 Comparative analysis with DL methods

To further elaborate on the deepfakes detection performance of the proposed approach, we have compared it against several latest DL-based approaches which employ the same datasets.

For the DFDC dataset, we have considered state-of-the-art approaches namely VGG16 [52], VGG-19 [53], ResNet101 [54], Inception V3 [55], InceptionResV2 [56], XceptionNet [57], DenseNet-169 [58], MobileNetv2 [59], EfficientNet [60], NASNet-Mobile [61] as demonstrated in [38]. Several standard metrics for performance evaluation are used to compare the average results of the proposed approach with comparative approaches.

Initially, we compared the manipulation detection accuracy of our approach and performed a comparative analysis with selected models. The comparison results are shown in Fig. 14a. The reported results are clearly showing that the proposed approach outperforms the comparative approaches. In terms of accuracy performance metrics, the lowest results are depicted by the VGG16 model with a value of 89%. While the second lowest value is attained by the VGG19 model with a value of 89.20%. Whereas, the proposed ResNet-Swish-Dense54 model shows the highest accuracy results with a value of 99.26%. More clearly, the comparative techniques show an average accuracy value of 93.84% which is 99.26% for our approach. So, we give an average performance gain of 5.44% for the accuracy metric.

For visual manipulation detection, the charge of false classification of altered content as pristine is heavy than the false categorization of real data as fake. Therefore, the model recognition performance is checked via conducting an experiment. For this reason, we have computed the recall evaluation metric and compared our results with the peer methods. The obtained values are shown in Fig. 14b which is clearly showing the robustness of our approach. It is quite visible from Fig. 14b that the lowest recall rate is attained by the VGG19 model with a value of 85.60%. The VGG16 and MobileNetv2 show the second value of 86%. Descriptively, for the recall measure, we have achieved a performance gain of 5.29%.

Table 6 Performance analysis of the ResNet-Swish-Dense54 model in the presence of adversarial attacks on the FF++, and DFDC datasets

ResNet-Swish-Desne54 results in added adversarial attacks	Precision (%)		Recall (%)		F1-Measure (%)		Accuracy (%)	
	FS	FR	FS	FR	FS	FR	FS	FR
FaceForensic++								
C0	99.45	97.94	98.39	97.24	98.92	97.59	98.78	97.89
C23	99.04	97.71	98.05	97.11	98.54	97.41	98.75	97.56
C40	99.08	97.62	97.66	96.68	98.36	97.15	98.72	97.39
DFDC								
	98.83%		97.96%		98.39%		96.89%	

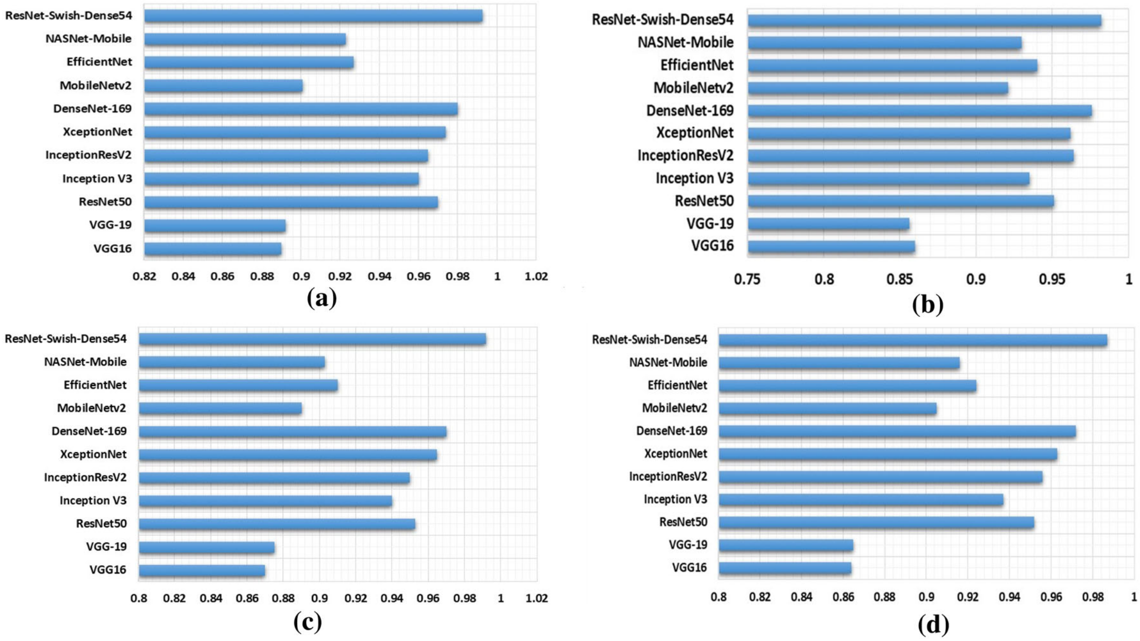


Fig. 14 Comparative analysis with the DL approaches in terms of **a** Accuracy, **b** Recall, **c** Precision, and **d** F1-score for the DFDC dataset

For multimedia forensic analysis, another main concern of models is to reduce the occurrence of false-positive rates. Hence, we have measured the precision metric to check the ability of the model to recognize the original data as well. The performance analysis with peer methods in terms of precision evaluation metric is demonstrated in Fig. 14c. The highest precision value is attained by the proposed approach with the value of 99.18%, while the lowest value is achieved by the VGG16 model with the value of 87%. More clearly, the comparative methods show the average precision value of 92.37% which is 99.18% for the introduced ResNet-Swish-Dense54 model. So, we have acquired an average performance gain of 6.92%.

We have further computed the F1-score of the proposed approach as this metric provides a better comparison of both precision and recall metrics to better describe the deepfakes recognition ability of the proposed solution. The obtained comparison results are shown in Fig. 14d. For the F1-score, the proposed approach demonstrates the highest result with a value of 98.70%. Moreover, the lowest results are acquired by the VGG16 model with a value of 86.40%. More clearly, the comparative approaches show the average F1-score of 92.63% which is 98.70% for the proposed solution. Therefore, for the F1-score measure, we have attained an average performance gain of 6.16%.

Moreover, for the FaceForensic++ dataset, we have presented the comparison of the proposed approach with other DL models like XceptionNet [44], Vgg16 [62], ResNet34 [63], InceptionV3 [64], and attained results are given in Fig. 15 It can be seen from the attained analysis that the ResNet-Swish-Dense54 model acquires robust classification results both for FaceSwap and Face-Reenactment than other DL approaches. More clearly, for the FaceSwap deepfakes, the other DL methods show an average accuracy score of 84.22% which is 99.13% for our work. So, we have

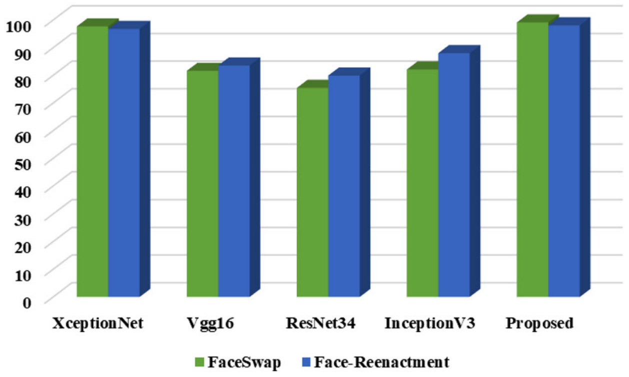


Fig. 15 Comparative analysis with the DL approaches in terms of Accuracy for the FaceForensic++ dataset

shown a performance gain of 14.91%. While for the Face-Reenactment deepfakes, the peer DL approaches show the average accuracy value of 87.07%, which is 98.08% for our work. Therefore, we have reported a performance gain of 11% for the Face-Reenactment deepfakes.

So, all the results reported in Figs. 14 and 15 clearly show that the introduced custom ResNet-Swish-Dense54 model is more robust to visual manipulation detection as compared to other DL approaches for all the nominated evaluation parameters. The main cause of the reliable performance of the introduced approach is because of the better feature computation ability of the model which empowers it to better learn the keypoints. The employment of the Swish activation method with added dense layers in the ResNet50 model allows it to better deal with the transformation changes found in the videos and hence result in improved deepfakes detection performance.

3.9.1 Comparison with state-of-the-art

To investigate the performance of our approach over other contemporary methods, we have compared the deepfakes detection ability of the proposed approach by comparing the attained results with numerous new studies for both employed datasets.

In the first part, we discussed the comparative analysis of the DFDC dataset by considering several standard performance measures with the latest methods discussed in [65–69]. The attained comparative analysis is given in Table 7. The values reported in Table 7 are clearly showing that our work is more robust to visual manipulation detection as compared to other works. The approach in [65] employed a DL framework to classify the input videos as being real or fake and exhibited an accuracy of 84.70%. Chintha et al. [66] employed the concept of using the frame, edge, and temporal level information of the video samples with a DL approach namely XceptionNet, and acquired the accuracy results of 97.94%. Another work was proposed in [67] named DPNet to locate the forensic alterations made within the visual samples and showed an AUC of 0.9244. Ganguly et al. [68] proposed an approach namely Vision Transformer along with the Xception Network for classifying the real and manipulated visual samples and showed the average accuracy and AUC values of 95.42%, and 0.9893, respectively. While the approach in [69] proposed an approach via employing the Convolutional Attention Network (CAN) to detect the visual deepfakes and attained the accuracy and AUC values of 94.40% and 0.9820, respectively. While in comparison, our work has shown the highest results in terms of both accuracy and AUC measures with the values of 99.26% and 0.9931, respectively. More clearly, for the accuracy metric, the comparative approaches have shown the average value of 95.92% which is 99.26%, so we have given a performance gain of 3.34% for the accuracy measure. While for the AUC measure the comparative techniques have attained the average value of 0.9652 which is 0.9931 for our case. Therefore, for the AUC measure, we have exhibited a performance gain of 2.78%. The conducted analysis in Table 7 is clearly depicting the robustness of our approach in comparison with the latest techniques.

Table 7 Comparative analysis with the latest methods over the DFDC dataset

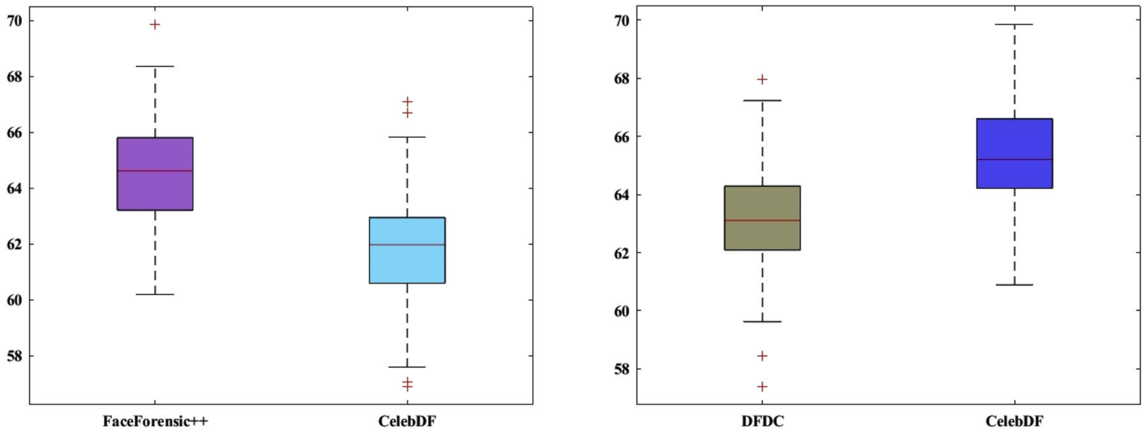
Technique	Accuracy (%)	AUC
Ranjen et al. [65]	84.70	–
Chintha et al. [66]	97.94	–
Trinh et al. [67]	–	0.9244
Ganguly et al. [68]	95.42	0.9893
Hernandezo et al. [69]	94.40	0.9820
Proposed	**99.26**	**0.9931**

Further, we have discussed the performance comparison of the proposed approach for the FaceForensic++ dataset by taking the studies given in [70–73]. The performance analysis in terms of several standard evaluation metrics is elaborated in Table 8. The technique discussed in [70] used two frameworks where the name of the first model is Mesoinception4 and later recognized as Meso4 for classifying the videos into several types of manipulations. The work has attained the accuracy of 73.91, and 66.31% for the FaceSwap and Face-Reenactment deepfakes for the Meso4 model. While the Mesoinception4 has demonstrated accuracies of 86.78 and 81.13% over the FaceSwap, and Face-Reenactment deepfakes. Li et al. [71] also used a DL approach to classify the visual samples and showed the performance results in terms of accuracy with values of 96.25 and 97.17% for the FaceSwap and Face-Reenactment deepfakes. While the method elaborated in [72] employed a Capsule network to locate the forensic changes developed inside the videos and attained the accuracy and AUC of 97.80% and 0.9979, respectively, over the FaceSwap deepfakes. While for the Face-Reenactment deepfakes the work has attained accuracy and AUC values of 97.48% and 0.9893, respectively. Pan et al. [73] introduced a methodology for the early recognition of multimedia manipulation by estimating the difference found between the human faces and the background area. For the FaceSwap deepfakes, the approach has reported the AUC and accuracy results of 0.9979 and 98.69%, which are 0.9897 and 97.57% for the Face-Reenactment deepfakes. It is quite evident from the performance analysis conducted in Table 8 that our work namely the ResNet-Swish-Dense54 has attained the highest results for both FaceSwap and Face-Reenactment deepfakes. For the FaceSwap deepfakes, we have attained the accuracy and AUC values of 99.13% and 0.9982, which are 98.08% and 0.9914 for the Face-Reenactment deepfakes, respectively. In a clearer manner, the comparative methods have shown the average accuracy results of 89.17 and 89.45%, which are 99.13 and 98.08% for our approach for the FaceSwap and Face-Reenactment deepfakes, respectively. So, for the accuracy measure, we have attained the performance gains of 9.96 and 8.63%, for the FaceSwap and Face-Reenactment deepfakes, respectively. While for the AUC measure, the competitor approaches have demonstrated the average values of 0.9455 and 0.9444 for the FaceSwap and Face-Reenactment deepfakes which are 0.9982 and 0.9914 for our case. Hence, for the AUC measure, we have acquired performance gains of 5.27 and 4.69%, respectively.

The conducted analysis for both datasets has clearly proved the proficiency of our approach to locate and recognize the visual manipulations of different types. The main reason for the effective deepfakes detection results of our model is due to the better feature selection power of the proposed model due to the inclusion of the Swish activation method which optimizes the model learning and the inclusion of dense layers which assists to nominate a more effective set of sample features to accomplish the classification task. Moreover, the light architecture of our approach assists it to avoid the occurrence of model over-fitting problems which in turn enhances the recognition power of our model. So, it can be concluded that our approach is proficient to locate the forensic changes made inside the visual samples.

Table 8 Comparative analysis with the latest methods over the FaceForensic++ dataset

Technique	FaceSwap		Face-Reenactment	
	Accuracy (%)	AUC	Accuracy (%)	AUC
Afchar et al. [70] (Mesoinception4)	66.31	0.8004	73.91	0.8104
Afchar et al. [70] (Meso4)	86.78	0.9429	81.13	0.9419
Li et al. [71]	96.25	0.9907	97.17	0.9909
Nguyen et al. [72]	97.80	0.9957	97.48	0.9893
Pan et al. [73]	98.69	0.9979	97.57	0.9897
Proposed	**99.13**	**0.9982**	**98.08**	**0.9914**



(a)

(b)

Fig. 16 Cross-dataset evaluation results of the ResNet-Swish-Dense54 model, **a** model is trained on the DFDC dataset, and tested on the FaceForensic++, and CelebDF datasets **b** Model is trained on the FaceForensic++ dataset and tested on the DFDC, and CelebDF datasets, respectively

3.9.2 Generalization ability testing

To investigate the generalization ability of our ResNet-Swish-Dense54 model, we performed an analysis to evaluate the deepfakes detection performance of the proposed approach in a cross-corpus scenario. To accomplish this task, we have chosen an additional dataset namely the CelebDF dataset. The dataset comprises a total of 408 original and 795 manipulated videos. The samples of the CelebDF dataset are complex in nature due to the small visible visual artifacts which make the manipulation detection a difficult task.

We have tested the model for scenarios where the ResNet-Swish-Dense54 model is trained on the DFDC dataset and evaluated on the FaceForensic++ and CelebDF datasets. Then, for the second task, we checked the model performance in a cross-dataset scenario where the framework is trained on the FaceForensic++ dataset and evaluated on the DFDC and CelebDF datasets. The attained results from both scenarios are shown in Fig. 16 with the help of a boxplot, as these graphs provide a better demonstration of the results by showing the largest, smallest, and average values. The values shown in Fig. 16 are clearly depicting that the proposed approach has faced performance degradation in a cross-corpus scenario as compared to the intra-database assessment case. The basic reason for this performance degradation is that the presented approach does not take into count the temporal changes that occurred within frames over time which can better assist the approach to capture the underlying manipulated artifacts. However, the ResNet-Swish-Dense54 model has enhanced the generalization results to some extent which can help the forensic analyzers for assessing the reliability of visual contents. More descriptively, for the first scenario where the model is trained on the DFDC dataset, we have attained the AUC values of 64.48 and 61.77% for the FaceForensic++ and CelebDF datasets, respectively. While for the other case, where the ResNet-Swish-Dense54 framework is trained on the FaceForensic++ dataset, we have acquired the AUC values of 63.17 and 65.22% for the FaceForensic++ and CelebDF datasets, respectively, which is depicting the robustness of our model to the unseen examples of an entirely different dataset.

Moreover, we have compared the generalization power of the presented approach with the latest techniques and attained comparison in terms of the accuracy measure is given in Table 9. We have taken state-of-the-art studies performing the same experiment to compare our results against them. It is quite clear from the values given in Table 9 that our approach performs better than the existing works in cross-corpus scenario which is exhibiting the better recall ability of the ResNet-Swish-Dense54.

Table 9 Accuracy comparison of the state-of-the-art approaches in terms of cross-dataset evaluation

Method	Test		
	DFDC (%)	CelebDF (%)	FF++ (%)
Trained on FaceForensic++			
[66]	81.29	–	97.94
[68]	–	69.30	95.42
Proposed	81.51	70.04	98.60
Trained on DFDC			
[68]	84.70	65.53	67.60
Proposed	99.26	67.14	70.12

3.9.3 Discussion

Due to the advancement of easy-to-use editing tools and DL-based platforms, an immense increase has been witnessed in the spread of fabricated data on social sites. This situation urged the scientists to develop forensic analysis tools to verify the authenticity of videos before employing them in processing any legal claims. Several methods have been represented in the past for the quick and reliable detection of real and fabricated videos; however, due to the increased realism of fake content, the effective detection of altered visual content is a complex task that degrades the recognition ability of existing works. Furthermore, existing methods are less robust to unseen cases and adversarial attacks. Moreover, most techniques lack the explainability feature which is a basic demand of video forensic analysts. In the presented work, we attempted to tackle the problems of the existing approach by introducing a DL-based approach named ResNet-Swish-Dense54.

The ResNet-Swish-Dense54 approach is capable of recognizing several types of visual manipulations like FaceSwap and Face-Reenactment and is able to differentiate the manipulated content from the pristine videos with a high recall rate. A detailed evaluation of the proposed approach is performed on two challenging datasets namely the FaceForensic++, and DFDC with the highest AUC score of 0.9982.

Furthermore, the proposed approach is proficient in tackling adversarial attacks of visual samples, i.e., compression, noise, blurring, zooming, rotation, and translation robustly. The ResNet-Swish-Dense54 is tested on visual samples of varying resolutions. Moreover, the post-processing attacks like noise, blurring, zooming, rotation, and translation are introduced only on the test samples to evaluate the effectiveness of our approach. We have confirmed through the reported results that the proposed approach is capable of detecting visual manipulations under the incidence of adversarial attacks in samples. Such a behavior of the presented work can be beneficial for the forensic analyzers as the videos uploaded on social websites have faced immense quality degradation for internet bandwidth requirements.

Another main concern for manipulation detection approaches is to be explainable to support the employment of the videos as proof in the processing of legal claims. For this reason, we have added an explainability module in the proposed approach via generating heatmaps. The visual results clearly depict that the proposed approach widely focused on regions where manipulations exist and show the accurateness of our work for visual deepfakes detection.

Additionally, to analyze the generalization capability of the ResNet-Swish-Dense54 unseen scenarios, we have conducted a cross-dataset examination where the framework is evaluated on different databases. We have observed that our technique has faced some performance reduction; however, the performance is still substantial. Therefore, after performing a huge numeric and visual examination of the proposed novel ResNet-Swish-Dense54 framework, we can conclude that our work can assist in the area of video forensic analysis. Currently, the model lacks to capture the temporal behavior of the forged content with time; therefore, our main motive is to include the temporal sequence analysis in the future to further improve the detection and generalization power of the proposed approach.

4 Conclusion

The extensive increase in manipulated content in cyberspace has urged researchers to nominate reliable methods for deepfakes detection to counter its impact on society. One such technique is presented in this work for deepfakes detection from visual samples. We have presented a novel ResNet-Swish-Dense54 model for deepfakes detection. Extensive experimentation has been conducted on the DFDC, FaceForensic++, and Celeb-DF datasets to show the effectiveness of the proposed approach for visual manipulation recognition. We have performed the cross-dataset evaluation of the proposed approach to show its generalization ability for unseen cases. Moreover, we have generated the heatmaps to show the explainability power of our framework. Furthermore, we have tested the presented work under the presence of adversarial attacks like compression, blurring, noise, rotation, and scale variations and confirmed through the reported results that our approach is effective to tackle such perturbations due to the better feature computation and selection ability of the ResNet-Swish-Dense54 model. The reported results have confirmed that the proposed work can serve as a tool in the field of multimedia forensic analysis. However, the presented framework is not evaluated on the crafted attacks like white or black box attacks; therefore, in the future, we plan to check the effectiveness of our approach on such attacks. Moreover, we plan to evaluate the proposed approach to other types of visual deepfakes like Neural Texture and lip-synching, etc. Furthermore, in the future, we also plan to include the temporal sequence analysis to further improve the detection and generalization power of the presented work.

Acknowledgements This work was supported by the grant of the Punjab Higher Education Commission (PHEC) of Pakistan via Award No. (PHEC/ARA/PIRCA/20527/21).

Declarations

Conflict of interest The authors declare that there is no conflict of interest between them.

References

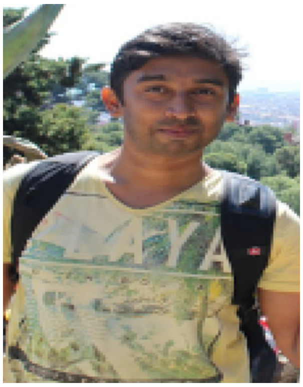
1. Khan, A.R., Doosti, F., Karimi, M., Harouni, M., Tariq, U., Fati, S.M., Ali Bahaj, S.: Authentication through gender classification from iris images using support vector machine. Micosc. Res. Tech. **84**(11), 2666–2676 (2021)
2. Palotás, Á.B., Rainey, L.C., Feldermann, C.J., Sarofim, A.F., Vander Sande, J.B.: Soot morphology: an application of image analysis in high-resolution transmission electron microscopy. Microsc. Res. Tech. **33**(3), 266–278 (1996)
3. Mahmood, M.T., Choi, W.J., Choi, T.S.: PCA-based method for 3D shape recovery of microscopic objects from image focus using discrete cosine transform. Microsc. Res. Tech. **71**(12), 897–907 (2008)
4. Nawaz, M., Mehmood, Z., Bilal, M., Munshi, A.M., Rashid, M., Yousaf, R.M., Rehman, A., Saba, T.: Single and multiple regions duplication detections in digital images with applications in image forensic. J. Intell. Fuzzy Syst. **40**(6), 10351–10371 (2021)
5. Nazir, T., Irtaza, A., Javed, A., Malik, H., Mehmood, A., Nawaz, M.: Digital image forensic analysis using hybrid features. In: 2021 International Conference on Artificial Intelligence (ICAI), 2021, pp. 33–36. IEEE
6. Vinolin, V., Sucharitha, M.: Dual adaptive deep convolutional neural network for video forgery detection in 3D lighting environment. Vis. Comput. **37**(8), 2369–2390 (2021)
7. Yang, G., Xu, K., Fang, X., Zhang, J.: Video face forgery detection via facial motion-assisted capturing dense optical flow truncation. Vis. Comput. 1–20 (2022)
8. Masood, M., Nawaz, M., Malik, K.M., Javed, A., Irtaza, A.: Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward. arXiv preprint arXiv:2103.00484 (2021)
9. He, D., He, X., Yuan, R., Li, Y., Shen, C.: Lightweight network-based multi-modal feature fusion for face anti-spoofing. Vis. Comput. 1–13 (2022)
10. Tyagi, S., Yadav, D.: A detailed analysis of image and video forgery detection techniques. Vis. Comput. 1–21 (2022)
11. Ballester, P., Araujo, R.M.: On the performance of GoogLeNet and AlexNet applied to sketches. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
12. (September 11, 2020). *Reface App*. Available: https://reface.app/
13. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
14. Setiaji, H., Paputungan, I.V.: Design of telegram bots for campus information sharing. In: IOP Conference Series: Materials Science and Engineering, vol. 325, no. 1, p. 012005. Institute of Physics Publishing (2018)

15. (January 11, 2021). *Sound Forge*. Available: https://www.magix.com/gb/music/sound-forge/
16. Boylan, J.F.: Will deep-fake technology destroy democracy? The New York Times, Oct, vol. 17, 2018.
17. Harwell, D.: Scarlett Johansson on fake AI-generated sex videos: 'nothing can stop someone from cutting and pasting my image. Washington Post (2018)
18. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5933–5942 (2019)
19. Nawaz, M., Mehmood, Z., Nazir, T., Masood, M., Tariq, U., Munshi, A.M., Mehmood, A., Rashid, M.: Image authenticity detection using DWT and circular block-based LTrP features. CMC Comput. Mater. Contin. **69**(2), 1927–1944 (2021)
20. Zhang, Y., Zheng, L., Thing, V.L: Automated face swapping and its detection. In: 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP), pp. 15–19. IEEE
21. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8261–8265. IEEE (2019)
22. Güera, D., Baireddy, S., Bestagini, P., Tubaro, S., Delp, E.J.: We need no pixels: video manipulation detection using stream descriptors. arXiv preprint arXiv:1906.08743 (2019)
23. Jack, K.: Chapter 13-MPEG-2. In: Video Demystified: A Handbook for the Digital Engineer, pp. 577–737
24. Ciftci, U.A., Demir, I.: FakeCatcher: detection of synthetic portrait videos using biological signals. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
25. Jung, T., Kim, S., Kim, K.: DeepVision: deepfakes detection using human eye blinking pattern. IEEE Access **8**, 83144–83154 (2020)
26. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans. Pattern Anal. Mach. Intell. **41**(1), 121–135 (2017)
27. Soukupova, T., Cech, J.: Eye blink detection using facial landmarks. In: 21st computer vision winter workshop, Rimske Toplice, Slovenia (2016)
28. Gupta, S., Thakur, K., Kumar, M.: 2D-human face recognition using SIFT and SURF descriptors of face's feature regions. Vis. Comput. **37**(3), 447–456 (2021)
29. Zhou, D., Liu, Y., Li, X., Zhang, C.: Single-image super-resolution based on local biquadratic spline with edge constraints and adaptive optimization in transform domain. Vis. Comput. 1–16 (2020)
30. Zhu, X., Chen, Z.: Dual-modality spatiotemporal feature learning for spontaneous facial expression recognition in e-learning using hybrid deep neural network. Vis. Comput. **36**(4), 743–755 (2020)
31. Couillaud, J., Ziou, D.: Light field variational estimation using a light field formation model. Vis. Comput. **36**(2), 237–251 (2020)
32. Xu, Y., Raja, K., Pedersen, M.: Supervised contrastive learning for generalizable and explainable DeepFakes detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 379–389 (2022)
33. Kolagati, S., Priyadharshini, T., Rajam, V.M.A.: Exposing deepfakes using a deep multilayer perceptron–convolutional neural network model. Int. J. Inf. Manag. Data Insights **2**(1), 100054 (2022)
34. Roy, R., Joshi, I., Das, A., Dantcheva, A.: 3D CNN Architectures and attention mechanisms for deepfake detection. (2022)
35. Sun, Z., Han, Y., Hua, Z., Ruan, N., Jia, W.: Improving the efficiency and robustness of deepfakes detection through precise geometric features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3609–3618 (2021)
36. Chen, Z., Xie, L., Pang, S., He, Y., Zhang, B.: MagDR: mask-guided detection and reconstruction for defending deepfakes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9014–9023 (2021)
37. Mehta, V., Gupta, P., Subramanian, R., Dhall, A.: FakeBuster: a DeepFakes detection tool for video conferencing scenarios. In: 26th International Conference on Intelligent User Interfaces, pp. 61–63 (2021)
38. Masood, M., Nawaz, M., Javed, A., Nazir, T., Mehmood, A., Mahum, R.: Classification of Deepfake videos using pre-trained convolutional neural networks. In: 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2), pp. 1–6. IEEE (2021)
39. Baltrušaitis, T., Robinson, P., Morency, L.-P.: Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10. IEEE (2016)
40. Fydanaki, A., Geradts, Z.: Evaluating OpenFace: an open-source automatic facial comparison algorithm for forensics. Forensic Sci. Res. **3**(3), 202–209 (2018)
41. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
42. Patwardhan, N., Ingalhalikar, M., Walambe, R.: ARiA: utilizing Richard's curve for controlling the non-monotonicity of the activation function in deep neural nets. arXiv preprint arXiv:1805.08878 (2018)
43. Dolhansky, B., Bitton, J.,. Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The DeepFake detection challenge dataset. arXiv preprint arXiv:2006.07397 (2020)
44. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1–11 (2019)
45. (2018, 14 March 2022). Deepfakes github. Available: http://github.com/deepfakes/faceswap
46. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2387–2395 (2016)
47. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: image synthesis using neural textures. ACM Trans. Graph. **38**(4), 1–12 (2019)
48. Gandhi, A., Jain, S.: Adversarial perturbations fool deepfake detectors. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)
49. Carlini, N., Farid, H.: Evading deepfake-image detectors with white-and black-box attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 658–659 (2020)
50. Hussain, S., Neekhara, P., Dolhansky, B., Bitton, J., Ferrer, C.C., McAuley, J., Koushanfar, F.: Exposing vulnerabilities of deepfake detection systems with robust attacks. Digit. Threats Res. Pract. (2021)
51. Hussain, S., Neekhara, P., Jere, M., Koushanfar, F., McAuley, J.: Adversarial deepfakes: evaluating vulnerability of deepfake detectors to adversarial examples. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3348–3357 (2021)
52. Nawaz, M., Masood, M., Javed, A., Iqbal, J., Nazir, T., Mehmood, A., Ashraf, R.: Melanoma localization and classification through faster region-based convolutional neural network and SVM. Multimed. Tools Appl. 1–22 (2021)
53. Carvalho, T., De Rezende, E.R., Alves, M.T., Balieiro, F.K., Sovat, R.B.: Exposing computer generated images by eye's region classification via transfer learning of VGG19 CNN. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 866–870. IEEE (2017)

54. Theckedath, D., Sedamkar, R.: Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks. SN Comput. Sci. **1**(2), 1–7 (2020)
55. Xia, X., Xu, C., Nan, B.: Inception-v3 for flower classification. In: 2017 2nd International Conference on Image, Vision and Computing (ICIVC), pp. 783–787. IEEE (2017)
56. Ferreira, C.A., Melo, T., Sousa, P., Meyer, M.I., Shakibapour, E., Costa, P., Campilho, A.: Classification of breast cancer histology images through transfer learning using a pre-trained inception resnet v2. In: International Conference Image Analysis and Recognition, pp. 763–770. Springer, Berlin (2018)
57. Kusniadi, I., Setyanto, A.: Fake video detection using modified XceptionNet. In: 2021 4th International Conference on Information and Communications Technology (ICOIACT), pp. 104–107. IEEE (2021)
58. Krešo, I., Oršić, M., Bevandić, P., Šegvić, S.: Robust semantic segmentation with ladder-densenet models. arXiv preprint arXiv:1806.03465 (2018)
59. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
60. Marques, G., Agarwal, D., de la Torre Díez, I.: Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network. Appl. Soft Comput. **96**, 106691 (2020)
61. Saxen, F. Werner, P., Handrich, S., Othman, E., Dinges, L., Al-Hamadi, A.: Face attribute detection with mobilenetv2 and nasnet-mobile. In: 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA), pp. 176–180. IEEE (2019)
62. Amerini, I., Galteri, L., Caldelli, R., Del Bimbo, A.: Deepfake video detection through optical flow based cnn. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
63. Koonce, B.: ResNet 34. In: Convolutional Neural Networks with Swift for Tensorflow, pp. 51–61. Springer, Berlin (2021)
64. Wang, C., Chen, D., Hao, L., Liu, X., Zeng, Y., Chen, J., Zhang, G.: Pulmonary image classification based on inception-v3 transfer learning model. IEEE Access **7**, 146533–146541 (2019)
65. Ranjan, P., Patil, S., Kazi, F.: Improved generalizability of deepfakes detection using transfer learning based CNN framework. In: 2020 3rd International Conference on Information and Computer Technologies (ICICT), pp. 86–90. IEEE (2020)
66. Chintha, A., Rao, A., Sohrawardi, S., Bhatt, K., Wright, M., Ptucha, R.: Leveraging edges and optical flow on faces for deepfake detection. In: 2020 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–10. IEEE (2020)
67. Trinh, L., Tsang, M., Rambhatla, S., Liu, Y.: Interpretable and trustworthy deepfake detection via dynamic prototypes. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1973–1983 (2021)
68. Ganguly, S., Ganguly, A., Mohiuddin, S., Malakar, S., Sarkar, R.: ViXNet: vision transformer with xception network for deepfakes based video and image forgery detection. Expert Syst. Appl. 118423 (2022)
69. Hernandez-Ortega, J., Tolosana, R., Fierrez, J., Morales, A.: DeepFakesON-Phys: DeepFakes detection based on heart rate estimation. arXiv preprint arXiv:2010.00400 (2020)
70. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7. IEEE (2018)
71. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656 (2018)
72. Nguyen, H.H., Yamagishi, J., Echizen, I.: 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), pp. 2307–2311. IEEE (2019)
73. Pan, Z., Ren, Y., Zhang, X.: Low-complexity fake face detection based on forensic similarity. Multimed. Syst. **27**(3), 353–361 (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Marriam Nawaz has completed BSc (Software Engineering) from University of Engineering and Technology, Taxila, and secured Gold Medal. Completed her M.Sc. degree in Software Engineering from University of Engineering and Technology, Taxila, with specialization in Software Engineering Discipline. Currently enrolled in Ph.D. Program in Software Engineering at UET, Taxila. Since 2017, she has been serving as Programmer at Computer Science Department, UET Taxila. Her research interests include Image processing, Medical Image Analysis, Digital Image Forgery detection, and Deepfakes Detection.



Ali Javed received the B.Sc. degree with honors and 3rd position in Software Engineering from UET Taxila, Pakistan in2007. He received his MS and Ph.D. degrees in Computer Engineering from UET Taxila, Pakistan in 2010 and 2016. He received Chancellor's Gold Medal in MS Computer Engineering degree. Dr. Javed is serving as an Associate Professor in Software Engineering Department at UET Taxila, Pakistan. He has served as a Postdoctoral Scholar in SMILES lab at Oakland University, MI, USA in 2019 and as a visiting PhD scholar in ISSF Lab at University of. Michigan, MI, USA in 2015. His areas of interest are Multimedia Forensics, Image Processing, Computer vision, Video Content Analysis, Medical Image Processing, and Multimedia Signal Processing. He has published more than 80 papers in leading journals and conferences including the IEEE Transactions. Dr. Javed is a recipient of various research grants from HEC Pakistan, National ICT R n D Fund, NESCOM, and UET Taxila Pakistan. He has also served as an HOD in Software Engineering Department at UET Taxila in 2014. Dr. Javed got selected as an Ambassador of Asian Council of Science Editors from Pakistan in 2016. He is also a member of Pakistan Engineering Council since 2007.

Aun Irtaza has completed his Ph.D. in 2016 from FAST-nu, Islamabad Pakistan. During his Ph.D., he remained working as a Research Scientist in the Gwangju Institute of Science and Technology (GIST), South Korea. He became an Associate Professor in 2017 and Department of Computer Science Chair in 2018 in the University of Engineering and Technology (UET) Taxila, Pakistan. He is currently working as visiting Associate Professor in the University of Michigan-Dearborn. His current research areas include computer vision, multimedia forensics, audio-signal processing, medical image processing, and big data analytics.