RESEARCH ARTICLE

IRV2-hardswish Framework: A Deep Learning Approach for Deepfakes Detection and Classification

Farooq Akhtar[1] **and Rabbia Mahum**[1,*]

[1] Department of Computer Science, University of Engineering & Technology Taxila, Taxila 47050, Pakistan

Abstract

Deep learning models are pivotal in the advancements of Artificial Intelligence (AI) due to rapid learning and decision-making across various fields such as healthcare, finance, and technology. However, a harmful utilization of deep learning models poses a threat to public welfare, national security, and confidentiality. One such example is Deepfakes, which creates and modifies audiovisual data that humans cannot tell apart from the real ones. Due to the progression of deep learning models that produce manipulated data, accurately detecting and classifying deepfake data becomes a challenge. This paper presents a groundbreaking IRV2-Hardswish Framework for deepfake detection, leveraging a hybrid deep learning architecture that synergizes residual blocks in CNNs and the Inception-Resnet-v2 model. By incorporating residual blocks to capture underlying audiovisual data layers and enhancing Inception-Resnet-v2 with Hardswish activation for robust feature extraction, our framework achieves accurate detection of deepfakes. Furthermore, additional dense layers are integrated to ensure precise classification, establishing a comprehensive and effective solution for deepfake detection. Further, a detailed comparison of our framework with the state-of-the-art CNN models reports that our framework outperforms with 98% accuracy, 96% precision, and 95% AUC using the Deep Fake Detection Challenge (DFDC) dataset. The DFDC dataset is the largest, consisting of approximately 5,000 clips, including 1,132 actual and 4,118 false ones. The results report the efficiency of the proposed framework. These results demonstrate the framework's effectiveness in deepfake detection.

Keywords: deep learning, deepfakes, hardswish framework, DFDC dataset.

1 Introduction

The proliferation of smart devices led to the creation of humongous audiovisuals, including photos and videos. Social media platforms such as Facebook, Instagram, and WhatsApp facilitate the easy sharing of the created content with the public [1, 2]. Consequently, celebrities and politicians who have a significant presence on these online platforms became the first targets of "deepfake". The term "deepfake" combines the terms "deep learning," which is a machine learning technology involving multiple layers of processing and "fake" addresses that the content is not real. Deepfakes help users create and synthesize

Submitted: 08 April 2025
Accepted: 27 April 2025
Published: 23 May 2025

Vol. 1, **No.** 1, 2025.
10.62762/JIAP.2025.421251

***Corresponding author:**
✉ Rabbia Mahum
rabbia.mahum@uettaxila.edu.pk

Citation
Akhtar, F., & Mahum, R. (2025). IRV2-hardswish Framework: A Deep Learning Approach for Deepfakes Detection and Classification. *ICCK Journal of Image Analysis and Processing*, 1(1), 45–56.

fake audiovisual data [3].

Deepfake refers to the various face and audio modification techniques using deep learning and computer vision. These face modification techniques are further categorized into four types: expression swap, full-face synthesis, attribute manipulation, and identity swap [4]. Among these, identity swap or face swap is the most used deepfake technology that enables the face of a person to be swapped with another. One such example is when an autoencoder-decoder in deep learning created fake pornographic content of a celebrity, validating the misuse of the technology. The people became aware of the identity swap back in 2017 [4].

Deepfakes also introduce forged images, videos, or audio that are difficult to tell apart from the real ones. In 2018, a minute-long video of former US President Barack Obama delivering an iconic hate speech that he never delivered became viral on social media [3]. This deepfake video of Barack Obama was produced by overdubbing his existing footage using a Generative Adversarial Network (GAN) that replicated the precise lip, head, and eye artifacts in the face using 56 hours of sample input movies [1]. Further, a transaction fooled the bank for USD 243,000 with a deepfake audio in 2019 [5–7]. Thus, deepfakes can cause political or religious tension between different countries, harm the financial market, or deceive the public by spreading false information [8].

Some ethical uses of the deepfake have also emerged in different fields, such as the ability to reshoot movie sequences in the absence of the actor, as witnessed in the Fast & Furious series, which is a prominent example of the media field [9]. Similarly, deepfakes can deliver realistic images of another person lip-syncing with the voice of another, and they can also be used to supply audio to actors who have lost their voices, or a character voice mismatch in the case of artists [10].

Editing tools can also perform different deepfake activities, such as adding, deleting, and replicating images [11]. A new object in an image can be copied using the other image, known as splicing, and an existing object can be deleted by expanding the background image to cover it, called inpainting [12]. Common picture editing software can also perform scaling, rotation, color correction, etc. [3]. However, due to the occurrence of pixel implosion, which causes unnatural-looking visual abnormalities in the skin, face, etc., and pixel density inconsistencies in pictures, a deepfake created by editing tools might be easily spotted in its early stages by human vision. However, because of recent advancements in deep learning technology and free access to vast quantities of data, deepfakes are difficult to recognize using either advanced computer techniques or hands-on human monitoring [3].

Media assets are created from scratch by using auto-encoders and GAN to synthesize the face. A segmentation map, to name a few deep learning techniques to create synthetic visual data. Any image can be synthesized with simple drawings or text descriptions using deep learning techniques. Auditory input also helps to synthesize the person's modifications. Style transfer deep learning technique can create a new image by altering the painting style [13].

In state-of-the-art, different deepfake techniques are used to solve the problem, such as deep learning methods (CNN or RNN), and machine learning algorithms (Support Vector Machine, K Nearest Neighbor, Random Forest, etc.). These techniques are implemented across a variety of different existing datasets. The widely used technique in deep learning is CNN for deepfake detection [14]. Therefore, the IRV2 Hardswish Framework based on deep learning is proposed in this research study. The contribution of our research study is as follows.

- To facilitate the detection of subject faces, we commenced by extracting individual frames from videos utilizing the Deep Fake Detection Challenge (DFDC) dataset. This preliminary step enabled us to isolate and process individual frames, setting the stage for subsequent face detection and analysis procedures.
- Hardswish activation in residual blocks is used to capture underlying visual data layers effectively. Hardswish offers improved gradient flow, enhanced feature representation, and computational efficiency. This enables our model to better detect subtle deepfake patterns and anomalies.
- Extra-dense layers are added to enhance classification accuracy. These layers amplify relevant features, reduce noise, and refine feature representations. This leads to improved deepfake detection performance and robust classification.
- We utilized the IRV2-Hardswish model to compute deep features and classify images as

real or altered. IRV2's robust architecture and Hardswish activation enable effective feature extraction and detection of subtle deepfake manipulations.

- The Deep Fake Detection Challenge (DFDC) dataset is used for experiments. Moreover, cross-validation is performed using two other datasets, i.e., Face Forensics deepfake collections (FF++) and Celeb-F. Results show the effectiveness of the proposed IRV2 Hardswish Framework.

The rest of the paper is structured as follows. Section 2 reviews the existing state-of-the-art. Section 3 presents our proposed framework, along with model training and testing parameters. Section 4 provides insight into the experimental results and covers dataset description, preprocessing steps, results, and comparison with existing state-of-the-art. Our research study is concluded in Section 5.

2 Related Work

Deepfake detection is an active research area, and several research studies have contributed to accurately performing it. Deepfake detection can be performed using two methods: Convolutional Neural Network (CNN) [15] and Region Convolutional Neural Network (RCNN). In CNN-based deepfake detection, pictures are extracted from a video and fed into the CNN model for training and prediction using spatial information only. On the other hand, RNN-based deepfake detection refers to the series of video frames for training and producing a result, considering both spatial and temporal information into account [16]. Various CNN architectures perform better in distinguishing between GAN-produced pictures and actual audiovisual data [14]. In the subsequent sections, research studies have been classified into two broad categories: 1) image-based deepfake detection and 2) video-based deepfake detection. Only the latest research studies are targeted, in which CNN-based models are used for deepfake detection.

2.1 Image-based Deepfake Detection

In this section, only those research studies are discussed in which deepfake detection using image-based content is performed, as shown in Table 1.

The model is trained on the following datasets: GDWCT, AttGAN, STARGAN, StyleGAN, and StyleGAN2, and an average of 97% accuracy, precision, and recall is reported. In [18, 19], a critical Forgery Mining (CFM) framework is proposed for forgery detection. It can be flexibly assembled with various backbones to increase the generalization. An aware loss function and Adam optimizer are used to learn global features. The model has been trained on six different datasets and achieved 83.93% AUC.

In [19], CNN is used for deep feature extraction, and the image is supplied to it after passing through the Error Level Analysis (ELA). Passing through ELA is a crucial step as it determines whether the image is modified or not. The resultant features are further classified using Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) by performing hyperparameter optimization to increase the performance. The model achieved the highest accuracy of 89.5%.

Similarly, in [20, 21] forged faces are detected using Image Quality Assessment (IQA) -based features. IQA is significant in the field of multimedia and face forensics. Images are extracted using IQA from the frequency domain and the spatial domain.

Table 1. Deepfake detection for images based on CNN.

Ref.	Classifier	Optimizer	Loss function	Dataset	Accuracy	Precision	Recall	AUC
[15]	CNN	Adam	-	GDWCT, AttGAN, STARGAN, StyleGsAN and StyleGAN2	97.72%	-	-	-
[16]	CNN	Adam	Similarity aware loss	Faceforensic++, Celeb-DFD v2, Wildfake, DFDC, DFD, DFRs	-	-	-	83.94%
[17]	CNN, SVM+, KNN	Hyper-parameter	-	Yonsei University's Computational Intelligence and Photography Lab Dataset	89.5%	-	-	-
[18]	MTCNN +Random Forest	-	IQA	VGGFace2 CASIA	99%	-	-	-
[19]	CNN	Adam	Loss function	Faceforensic++	97.52%	-	-	99.79%
[20]	DADF	AdamW	Loss function	Faceforensic++	95.94%	-	-	

Table 2. Deepfake detection for videos based on CNN.

Ref.	Classifier	Optimizer	Loss function	Dataset	Accuracy	Precision	Recall	AUC
[21]	Deep CNN	leave-one-out	ArcFace	Celeb-DF	97%	94%	98%	-
[22]	Trans-former+CNN	Stochastic gradient descent optimizer	Log loss	Celeb-DF	-	-	-	0.97
[23]	CNN+LSTM	Adam	Cross entropy	DFDC and CipLab	98.27%, 97.81%	98.24% 97.32%	-	-
[24]	CNN+LSTM	Adam	Cross entropy	Forensic, DFDC Celeb-DF	91.21%, 79.49%, 66.26%	-	-	0.91, 0.79 0.66
[25]	CNN+LSTM	Nadam	Binary Cross entropy	Celeb-DF	99.24	-	-	99.52

Subsequently, the images are transferred to the random forest classifier along with the labels of the image during the training phase. The proposed model has achieved the highest 99% accuracy with two standard datasets, VGGFace2 & CASIA.

In another research article [21], the deep learning model is trained to adapt the various face synthesis techniques using fine-grained artifact features. The proposed framework introduced the fake blender module for creating the synthetic images. Moreover, residual-based deepfake detection performs better forgery classification by detecting the residuals from the fake images. The proposed framework gained higher accuracy and AUC FF++ and WildDeepfake datasets. In [22], the Segment Anything Model (SAM) -based Detect Any Deepfake (DADF) framework is proposed to detect face forgery. This framework captures short-range and long-range forgery contexts for fine-tuning. The model was successfully tested on the Faceforensic++ dataset and achieved 95.94% accuracy.

2.2 Video-based Deepfake Detection

In this section, only those research studies are discussed in which deepfake detection using video-based content is performed. In [23], a Deep Convolutional Neural Network (DCNN) is proposed for facial recognition in videos. DCNN is based on the threshold classifier integrated with the similarity scores of facial recognition among the real videos and the videos displayed.

Consequently, the highest score is used to classify the video as real or fake. The proposed model is tested on the Celeb-DF dataset and achieves 97% accuracy with an AUC of 0.994. Similarly, local and global features are extracted by combining the vector-concatenated CNN and patch-based positioning to detect all the possible positions using the vision transformer [24]. For distillation, binary cross-entropy is used, and a comparison of the proposed model is made with the SOTA model. The results report that the proposed model outperforms SOTA by 0.006 AUC and 0.013 F1 scores using the DFDC dataset. 2500 fake videos are provided to the proposed model, and 2313 videos are accurately classified as fake. The SOTA model predicted 2276 fake videos among 2500. The model outperforms the SOTA model for the Celeb-DF (v2) dataset by achieving 0.993 AUC and 0.978 F1 score.

In another research study [25], a novel deep learning architecture combining Long Short-Term Memory (LSTM) and CNN is proposed for deepfake detection. The spatial information of CNN is integrated into the temporal information of the LSTM to combat the deepfake. The proposed architecture is implemented using Python and the Kaggle platform on two open-source datasets, DFDC and Ciplab, and achieved 98.27% & 97.81% accuracy, respectively. The error rates are 0.51% and 0.26%, as calculated by the binary cross function. The results show that the LSTM combined with CNN strongly emphasized the importance of the temporal dependencies of deepfake detection for visual data. Similarly, in [26, 27] CNN and LSTM are combined for the deepfake detection of the video frames using different datasets, and their achieved accuracy and AUC values are illustrated in Table 2.

3 Methodology

In this research study, a novel IRV2-Hardswish framework is introduced for deepfake detection in videos. The proposed framework consists of the following phases.

3.1 Face Extraction from Videos

Human faces are the prime location where modifications in the deepfake are made. In our proposed framework, human faces are identified from the video frames using the Cascade classifier of the

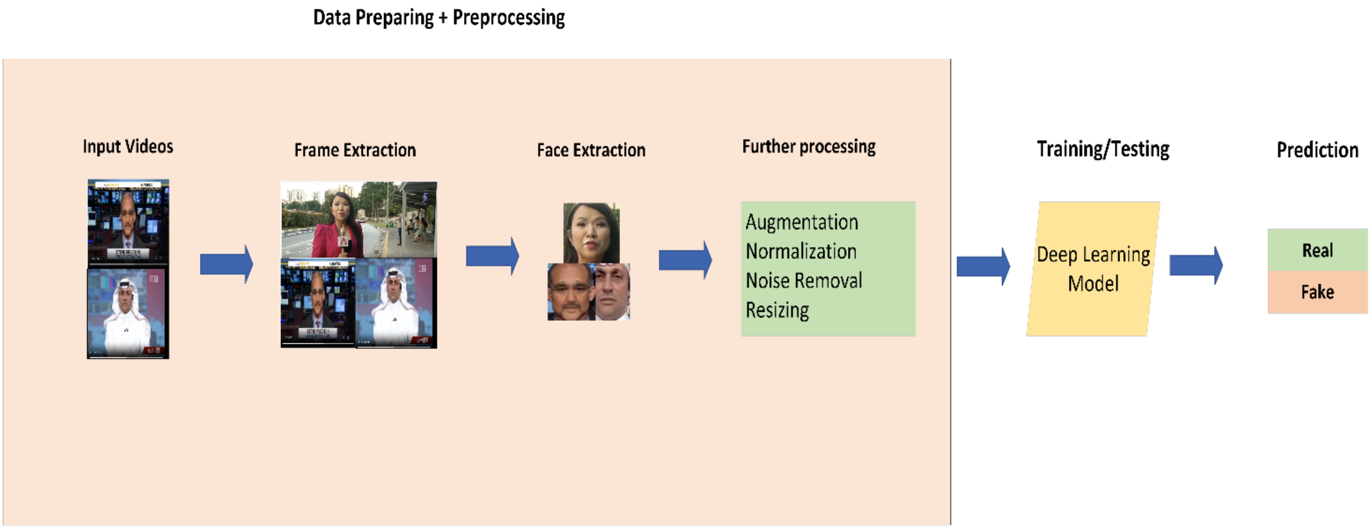


Figure 1. Illustration of the IRV2-Hardswish showcasing the efficient feature extraction and linear transformation capabilities of this architectural component.

OpenCV tool [3], as shown in Figure 1.

The cascade classifier uses an improved version of the simple classifier to examine the image area to find out whether a face exists in this area or not. Once faces have been identified in the video frames, they may be separated and utilized for additional processing, such as finding visual adjustments [28, 29]. The effectiveness of the face identification procedure can significantly affect how well the final deepfake turns out because false positives or false negatives might result in erroneous or unrealistic results.

Additionally, to maintain the computational complexity of the proposed method, we have only selected 20 frames from all video samples.

3.2 Feature Extraction

After the face extraction, the next step is to calculate the features from the video frames. Inception-Resnet-v2 [29] is modified by adding the Hardswish activation technique and used to perform feature extraction. The Hardswish activation technique is non-linear and capable of adding negative values across the neurons, while feature extraction helps in the identification of complex visual patterns.

Moreover, the proposed model design also includes additional dense layers and global max pooling in the inception-resnet architecture. The benefit of global max pooling is that it lowers the spatial dimensions of the feature maps to a single value per channel, the number of parameters in the network, and the probability of overfitting. The dense layers enable the model to suggest a representative collection of features for classification. Overall, these adjustments to the Inception-Resnet-v2 model allow it to calculate and extract critical information from the faces in the video frames with more efficiency.

The primary benefit of choosing an inception-resnet architecture is that it has previously been trained on a sizable dataset, like the ImageNet database, and can produce a more accurate collection of image features [29]. The ImageNet collection contains millions of labeled pictures that are used to train neural networks with deep connections for image recognition applications.

Figure 2 provides a graphic illustration of this activity. The ResNet design, on the other hand, makes use of residual connections to allow for the training of far deeper networks. Reusing previously learned characteristics through residual connections makes it simpler for the network to pick up new ones. For deepfake identification applications, this method has been demonstrated to be quite successful, although it can be computationally costly.

The Residual Block (RB) is the fundamental component of the ResNet model. The ReLU activation function and multiple convolution layers are both included in the RB. It consists of a convenient link, a batch normalization layer, and the stacked layers in charge of residual mapping by using shortcut linkages that carry out identity mapping. The result of the RB can be expressed as in Eq. 1.

$$Z = F(X) + X \tag{1}$$

where X stands in for the input, F for the residual

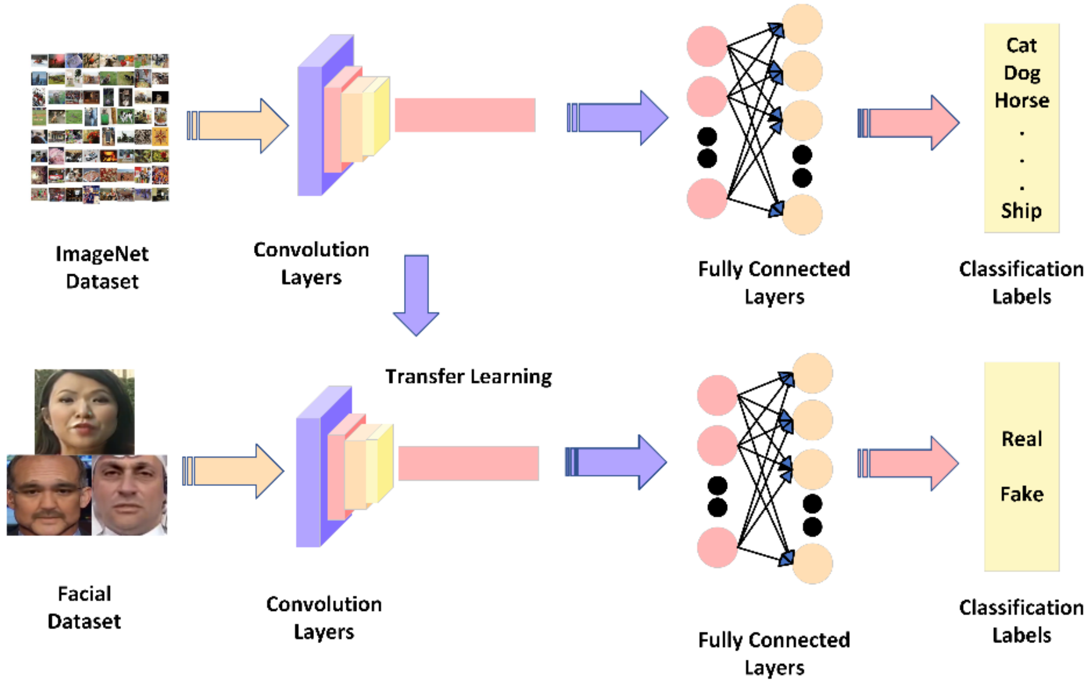


Figure 2. Schematic illustration of the feature extraction process, depicting the transformation of raw input data into meaningful representations through various techniques, including convolution and fully connected layers. The extracted features are then utilized for the classification.

function, and Z for the output of the residual function as illustrated in Figure 3.

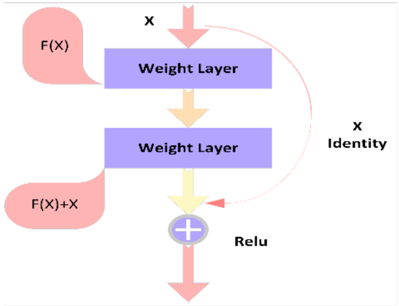


Figure 3. Schematic representation of the RELU activation function and associated weights in a ResNet model.

The network can learn features at various sizes while still being able to train successfully because of the utilization of residual connections in the Inception blocks. The network, which has 164 layers, was trained using pictures from the enormous ImageNet database as depicted in Figure 4 (a) and (b).

3.3 IRV2-Hardswish

Inception-ResNet v2 is further expanded with batch normalization, bottleneck layers, and an extra classifier to the network. Accuracy and training time both improved by these enhancements. For classification, we introduce transfer learning on the InceptionResnet-v2 network, producing output probabilities for two classes: fake and real. By adding relatively minimal cost to the model design, the additional dense layers improve the model's capacity to learn a trustworthy collection of picture attributes. Once the attributes have been chosen by the additional dense layers, they are sent to the softmax layer to yield the results. The network is trained using shape-related input pictures (128, 128, 3), which were taken from the DFDC dataset.

The model is trained on 30 epochs and produces low validation loss at the end. Hardswish only requires straightforward arithmetic operations, it is computationally efficient when compared to other non-linear activation functions like ReLU. This speeds up computation and makes implementation simpler. Since Hardswish is non-saturating, both high and low input values have no impact on its output. This may assist the model in consolidating more quickly by preventing the gradients from both vanishing or exploding throughout back propagation.

Moreover, Regularization techniques like Hardswish can assist deep learning models from overfitting. Hardswish can aid in ensuring that the model generalizes successfully to new data by limiting how quickly and aggressively it learns [30]. The hardswish activation approach is straightforward by nature, and

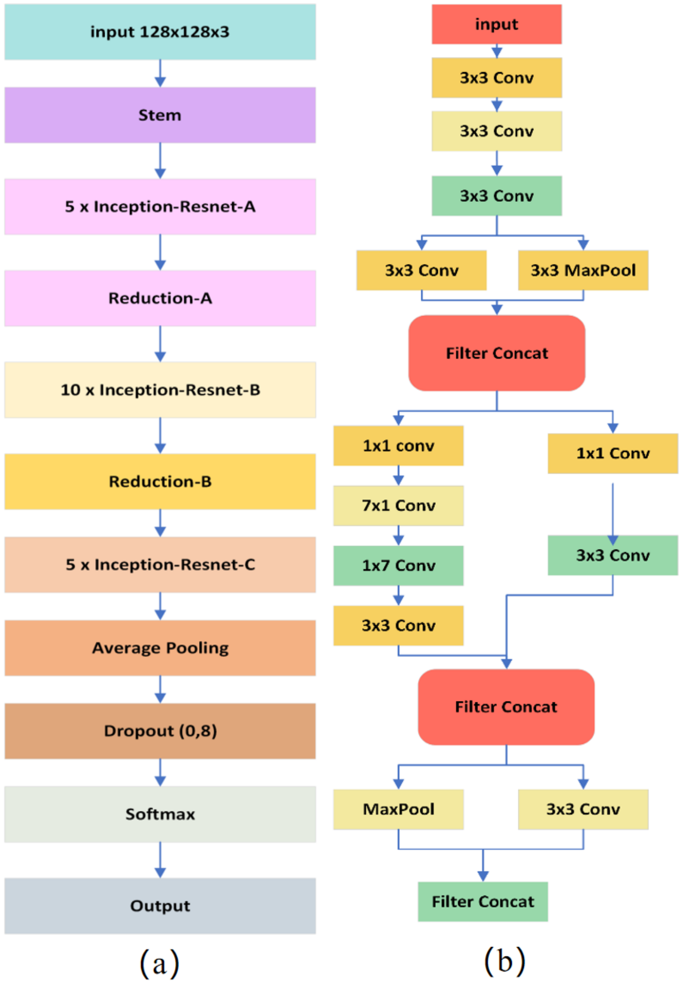


Figure 4. Feature extraction layers.

several studies show that it outperforms the most popular activation methods, such as Sigmoid, Swish, Tanh, and ReLU for images and object identification. Figure 5 provides a graphic representation of swish and other activation techniques.

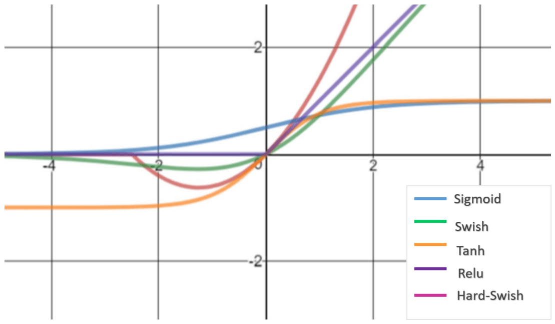


Figure 5. Hardswish vs other methods [35].

The hardswish mathematical representation is given below in Eq. (2).

$$F(i) = i \cdot \frac{\max(0, \min(6, i+3))}{6} \tag{2}$$

where i is the input to the activation function.

4 Experiments

In this section, the dataset and tools used, and implementation details are highlighted. At the end of this section, results achieved from the proposed models are reported.

4.1 Dataset

The DFDC dataset is selected to experiment with the proposed framework. There are other datasets such as Celeb-DF, Forensic, Forensic++, etc. with forgeries of videos, but we found DFDC to be fairly diversified in terms of the gender, skin tone, age, and color of the actors. Participants were free to record movies with any background of their choice, so DFDC produced visually diverse backgrounds, incorporated a variety of head angles and lighting situations, and increased visual diversity. The DFDC dataset consists of approximately 5,000 clips, including 1,132 actual and 4,118 false ones. The DFDC data are an online dataset that is available to the general public and may be downloaded from the Kaggle competition website [14].

4.2 Tools

The following tools are used to implement and validate the proposed framework in this research study.

1. OpenCV: For basic calculations, data retrieval, cleaning, processing, and visualization, the Python Imaging Library (PIL) and OpenCV are used.
2. Tensor Flow and Keras: Scikit Learn was used to import machine learning models. TensorFlow and Keras were used to construct an artificial neural network.
3. Matplotlib: Matplotlib was used to plot the different graphs for comparison using Matlab.

4.3 Method

In our research, we employ OpenCV, a powerful open-source computer vision library, to extract frames from videos and subsequently detect faces within these frames, as shown in Figure 6. OpenCV provides convenient functionalities, such as the VideoCapture class, which enables us to read video files and extract individual frames efficiently. Once we have extracted frames, we utilize OpenCV's pre-trained Haar cascade classifier for face detection. Haar cascades are a type of machine learning-based object detection algorithm that leverages a cascade of classifiers to identify objects within an image based on their features. Specifically, the Haar cascade classifier for face detection employs a series of rectangular features to detect facial features such as eyes, nose, and mouth. We use the Cascade Classifier class in OpenCV to load the pre-trained Haar cascade XML file and apply it to each frame to detect faces.

Figure 6. Sample images from the DFDC dataset.

We proceed with additional preprocessing steps to ensure uniformity and standardization of the facial images extracted from the DeepFake Detection Challenge (DFDC) dataset. We resize each detected face image to a standardized size of (128, 128) pixels, a common practice in computer vision tasks to facilitate model training and performance evaluation. This resizing step ensures that all facial images have consistent dimensions, thereby minimizing variations in size that could affect the performance of subsequent machine-learning algorithms.

By creating a structured dataset of resized facial images extracted from the DFDC videos, we establish a foundation for conducting rigorous experiments and analyses aimed at addressing the challenges posed by deepfake content. We leveraged the Inception ResNet-v2 architecture, incorporating a Hardswish activation layer, to tackle the challenge of deepfake detection.

The Inception ResNet-v2 model, renowned for its depth and efficiency in capturing complex features, was chosen for its robustness and superior performance in image classification tasks. By integrating the Hardswish activation layer, known for its non-linearity and efficiency, we aimed to enhance the model's representational power and computational efficiency.

Our experimentation on the DFDC dataset yielded promising results, showcasing the efficacy of the Inception ResNet-v2 model with the Hardswish activation layer. Specifically, we achieved a remarkable accuracy of 98.3% in discerning between genuine and deepfake videos. This high level of accuracy underscores the effectiveness of our proposed approach in accurately identifying manipulated videos, thereby contributing significantly to the ongoing efforts in combating the proliferation of deepfake content across digital platforms. A ROC plot exhibiting AUC is shown in Figure 7.

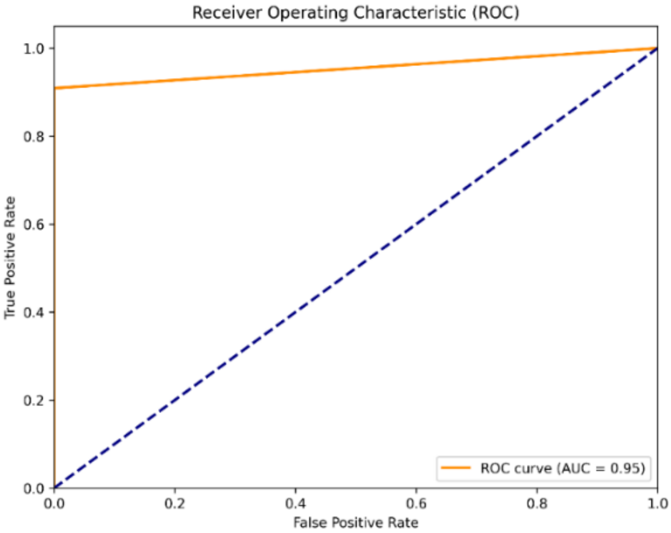


Figure 7. ROC curve with AUC value.

Therefore, a further important statistic for comparison is the classifier's AUC value. The Area Under the Curve (AUC) statistic shows how well the classifier can distinguish between distinct output classes. The stronger the model's capacity to distinguish between the positive and negative classifications, the higher its AUC score. As a result, we have displayed the Receiver Operating Characteristics (ROC) curves and determined the AUC 0.95 score for each of our neural networks as displayed in Figure 8. In Figure 9, the graph shows the accuracy of the proposed model, which improves with increasing epochs as it learns from the data. The blue line represents the accuracy of the training data on which the model learns.

The orange line shows the accuracy of the testing data, which is test data used to evaluate how well the model has learned. Both lines start at a lower accuracy and increase steadily, indicating that the model is learning and improving its accuracy over time. Figure 8 also indicates that the training accuracy (blue line) is mostly higher than the testing accuracy (orange line). This is normal because the model is repeatedly exposed to the training data, making it

Table 3. Comparison with other models on the DFDC dataset.

Ref.	Classifier	Optimizer	Loss function	Dataset	Accuracy	Precision	Recall	AUC
[23]	CNN+ LSTM	Adam	Cross entropy	DFDC	98.2 %	98%	-	-
[24]	CNN+ LSTM	Adam	Cross entropy	DFDC	79%	-	-	0.79
[31]	Efficient Net	Adam	Binary Cross entropy	Face-forensic++	99.2%	90%	95%	94
[32]	CNN+GRU	Adam	Binary Cross entropy	DFDC	92.60%	-	-	-
[33]	CNN+ LSTM	Adam	Binary Cross entropy	DFDC +Custom	92.61%	-	-	-
[34]	CNN	Adam	Binary Cross entropy	DFDC	93.7%	98%	97%	-
Our Model	**CNN**	**Adam**	**Binary Cross entropy**	**DFDC**	**98.3%**	**96%**	**96.48%**	**0.95**

better at predicting these known examples. Around the 20th epoch, both lines start to level off, meaning the model has learned most of what it can from the data, and additional training does not significantly improve accuracy.

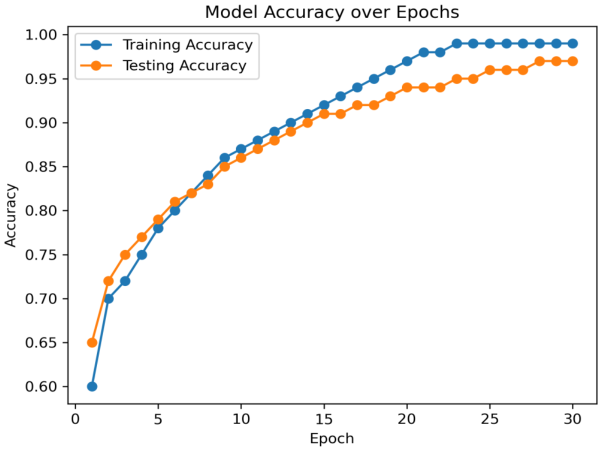


Figure 8. Model accuracy over epoch.

Figure 9 shows the error of the model during training and testing, with lower values indicating better performance. The blue line represents the error (or loss) in the training data, while the orange line shows the error in the testing data. Both lines start higher and decrease over time, meaning the proposed model is learning and provides better results at making accurate predictions.

It is analyzed from Figure 9. that the error decreases rapidly, showing that the model quickly learns the basic patterns in the data in the beginning. As training continues, the rate of improvement slows down, indicating that the model is fine-tuning its understanding and making fewer mistakes. The training error (blue line) is consistently lower than the testing error (orange line), which is expected since the proposed model is optimized to perform well on the training data.

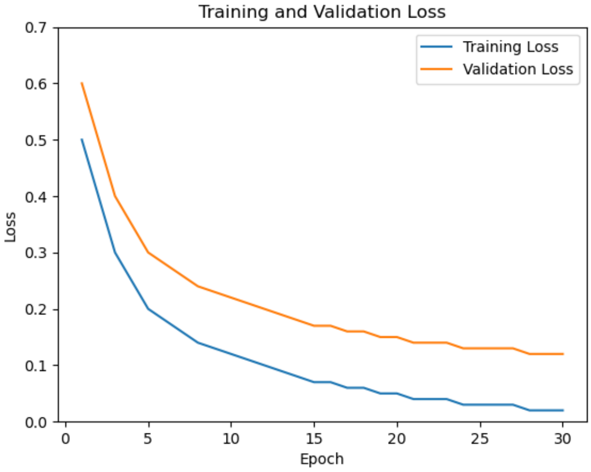


Figure 9. Training and validation loss.

4.4 Comparison

In this section, results achieved by the proposed model are compared with the state-of-the-art latest models, mostly using the DFDC dataset, as illustrated in Table 3. Since identifying deepfakes essentially involves classification, several methods are usually compared by determining the evaluation parameters of Precision, Recall, Accuracy, and Area Under Curve (AUC) metrics. Using Inception-ResNet-v2 with Hardswish activation function, our model was able to attain performance values of 98.3% Accuracy, 96.48% Recall, and 96% Precision. With an AUC of 0.95, one of the best state-of-the-art results is attained. In [31], the achieved accuracy is more than the proposed model, however, a dataset other than DFDC is used in this research article.

4.5 Cross-Validation

For cross-validation, an experiment is conducted using FF++ and Celeb-DF datasets to measure the effectiveness of the proposed. The celeb-DF dataset includes 950 edited videos and 475 original videos. The AUC values for the FF++ and Celeb-DF are 70.12%

and 65.23%, respectively. When trained on the DFDC dataset, it scored 95% AUC. The results show the effectiveness of the IRV2 hardswish framework when unseen cases of deepfakes are provided, as shown in Table 4.

Table 4. Cross-validation with other datasets.

Datasets	AUC
FF++	70.12%
Celeb-DF	65.23%
DFDC	95.00%

5 Conclusion

In this paper, a novel deep learning-based architecture is proposed for deepfake detection using a combination of Convolution Neural Network (CNN) and IRV2 Hardswish, and implemented using the DFDC dataset to evaluate its performance. The DFDC dataset comprises approximately 5,000 clips, including 1,132 actual and 4,118 false ones. Our experimental results on the DFDC dataset demonstrate the superiority of the IRV2-Hardswish Framework, achieving state-of-the-art performance in deepfake detection. The synergy of Inception ResNet-v2, CNN, and Hardswish activation enables robust feature extraction and accurate classification, outperforming existing methods. The proposed model achieved an accuracy of 98.3% by outperforming the previous state-of-the-art approaches and distinguishing between real and deepfake videos. Therefore, the fusion of IRV2 hardswish and CNN enables a focus on their respective strengths, potentially yielding an efficient and accurate model. In conclusion, this framework's effectiveness validates its potential for real-world applications, paving the way for reliable deepfake detection solutions.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Ramanaharan, R., Guruge, D. B., & Agbinya, J. I. (2025). DeepFake video detection: Insights into model generalisation—A systematic review. *Data and Information Management*, 100099. [CrossRef]

[2] Siddiqui, F., Yang, J., Xiao, S., & Fahad, M. (2025). Enhanced deepfake detection with DenseNet and Cross-ViT. *Expert Systems with Applications*, *267*, 126150. [CrossRef]

[3] Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *14*(2), e1520. [CrossRef]

[4] Gao, J., Micheletto, M., Orrù, G., Concas, S., Feng, X., Marcialis, G. L., & Roli, F. (2024). Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection. *Engineering Applications of Artificial Intelligence*, *133*, 108450. [CrossRef]

[5] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1-11). [CrossRef]

[6] Mahum, R., Irtaza, A., & Javed, A. (2023). EDL-Det: A robust TTS synthesis detector using VGG19-based YAMNet and ensemble learning block. *IEEE Access*, *11*, 134701-134716. [CrossRef]

[7] Mahum, R., Irtaza, A., Javed, A., Mahmoud, H. A., & Hassan, H. (2024). DeepDet: YAMNet with BottleNeck Attention Module (BAM) for TTS synthesis detection. *EURASIP Journal on Audio, Speech, and Music Processing*, *2024*(1), 18. [CrossRef]

[8] Sohrawardi, S. J., Wu, Y. K., Hickerson, A., & Wright, M. (2024, May). Dungeons & deepfakes: Using scenario-based role-play to study journalists' behavior towards using AI-based verification tools for video content. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-17). [CrossRef]

[9] Maniyal, V., & Kumar, V. (2024). Unveiling the deepfake dilemma: Framework, classification, and future trajectories. *IT Professional*, *26*(2), 32-38. [CrossRef]

[10] Abbas, F., & Taeihagh, A. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications*, 124260. [CrossRef]

[11] Megahed, A., Han, Q., & Fadl, S. (2024). Exposing deepfake using fusion of deep-learned and hand-crafted features. *Multimedia Tools and Applications*, *83*(9), 26797-26817. [CrossRef]

[12] Melnik, A., Miasayedzenkau, M., Makaravets, D., Pirshtuk, D., Akbulut, E., Holzmann, D., ... & Ritter, H. (2024). Face generation and editing with stylegan:

A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(5), 3557-3576. [CrossRef]

[13] Gupta, A., & Pandey, D. (2024, February). Deepfake videos generation and detection: A comprehensive survey. In *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)* (Vol. 5, pp. 1939-1944). IEEE. [CrossRef]

[14] Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access*, *10*, 25494-25513. [CrossRef]

[15] Guo, Z., Yang, G., Chen, J., & Sun, X. (2023). Exposing deepfake face forgeries with guided residuals. *IEEE Transactions on Multimedia*, *25*, 8458-8470. [CrossRef]

[16] Chauhan, R., Sethi, M., & Ahuja, S. (2024, March). Fake Faces Unveiled: A Comprehensive Study on Detecting Generated Facial Images. In *2024 International Conference on Automation and Computation (AUTOCOM)* (pp. 475-482). IEEE. [CrossRef]

[17] Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., Davidson, I. E., & Mazibuko, T. F. (2023). An improved dense CNN architecture for deepfake image detection. *IEEE Access, 11*, 22081-22095. [CrossRef]

[18] Luo, A., Kong, C., Huang, J., Hu, Y., Kang, X., & Kot, A. C. (2023). Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Transactions on Information Forensics and Security*, *19*, 1168-1182. [CrossRef]

[19] Rafique, R., Gantassi, R., Amin, R., Frnda, J., Mustapha, A., & Alshehri, A. H. (2023). Deep fake detection and classification using error-level analysis and deep learning. *Scientific Reports*, *13*(1), 7422. [CrossRef]

[20] Kiruthika, S., & Masilamani, V. (2023). Image quality assessment based fake face detection. *Multimedia Tools and Applications*, *82*(6), 8691-8708. [CrossRef]

[21] Liu, Q., Xue, Z., Liu, H., & Liu, J. (2024). Enhancing deepfake detection with diversified self-blending images and residuals. *IEEE Access, 12*, 46109-46117. [CrossRef]

[22] Lai, Y., Luo, Z., & Yu, Z. (2023, December). Detect any deepfakes: Segment anything meets face forgery detection and localization. In *Chinese Conference on Biometric Recognition* (pp. 180-190). Singapore: Springer Nature Singapore. [CrossRef]

[23] Reis, P. M. G. I., & Ribeiro, R. O. (2024). A forensic evaluation method for DeepFake detection using DCNN-based facial similarity scores. *Forensic Science International*, *358*, 111747. [CrossRef]

[24] Heo, Y. J., Yeo, W. H., & Kim, B. G. (2023). Deepfake detection algorithm based on improved vision transformer. *Applied Intelligence*, *53*(7), 7512-7527. [CrossRef]

[25] Al-Dulaimi, O. A. H. H., & Kurnaz, S. (2024). A hybrid CNN-LSTM approach for precision deepfake image detection based on transfer learning. *Electronics*, *13*(9), 1662. [CrossRef]

[26] Masud, U., Sadiq, M., Masood, S., Ahmad, M., & Abd El-Latif, A. A. (2023). LW-DeepFakeNet: A lightweight time distributed CNN-LSTM network for real-time DeepFake video detection. *Signal, Image and Video Processing*, *17*(8), 4029-4037. [CrossRef]

[27] Saikia, P., Dholaria, D., Yadav, P., Patel, V., & Roy, M. (2022, July). A hybrid CNN-LSTM model for video deepfake detection by leveraging optical flow features. In *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). IEEE. [CrossRef]

[28] Tang, L., Ye, D., Lu, Z., Zhang, Y., & Chen, C. (2024). Feature Extraction Matters More: An Effective and Efficient Universal Deepfake Disruptor. *ACM Transactions on Multimedia Computing, Communications and Applications*, *21*(2), 1-22. [CrossRef]

[29] Nawaz, M., Javed, A., & Irtaza, A. (2023). ResNet-Swish-Dense54: A deep learning approach for deepfakes detection. *The Visual Computer*, *39*(12), 6323-6344. [CrossRef]

[30] Wang, J., Qi, Y., Hu, J., & Hu, J. (2022, May). Face forgery detection with a fused attention mechanism. In *2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA)* (pp. 722-725). IEEE. [CrossRef]

[31] Suratkar, S., & Kazi, F. (2023). Deep fake video detection using transfer learning approach. *Arabian Journal for Science and Engineering*, *48*(8), 9727-9737. [CrossRef]

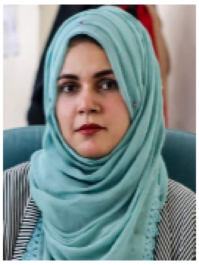
[32] Nguyen, H. H., Fang, F., Yamagishi, J., & Echizen, I. (2019, September). Multi-task learning for detecting and segmenting manipulated facial images and videos. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)* (pp. 1-8). IEEE. [CrossRef]

[33] Montserrat, D. M., Hao, H., Yarlagadda, S. K., Baireddy, S., Shao, R., Horváth, J., ... & Delp, E. J. (2020). Deepfakes detection with automatic face weighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 668-669). [CrossRef]

[34] Rahman, A., Siddique, N., Moon, M. J., Tasnim, T., Islam, M., Shahiduzzaman, M., & Ahmed, S. (2022, September). Short and low resolution deepfake video detection using CNN. In *2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC)* (pp. 259-264). IEEE. [CrossRef]

[35] Avenash, R., & Viswanath, P. (2019). Semantic Segmentation of Satellite Images using a Modified CNN with Hard-Swish Activation Function. In *VISIGRAPP (4: VISAPP)*.

Farooq Akhtar holds a bachelor's degree in computer science from HITEC University and is currently pursuing an MS in Data Science from UET Taxila. With 7 years of professional experience in artificial intelligence and big data development, their work spans deep learning, computer vision, and large-scale data processing. He have also contributed to medical AI by publishing research on COVID-19 diagnosis using datasets image and speech. He is passionate about integrating AI with big data technologies, to deliver innovative, scalable, and ethical solutions for complex real-world problems. (Email: Farooq.Akhtar@students.uettaxila.edu.pk)



Rabbia Mahum received the PhD degree in computer science from UET Taxila in 2024. She received the MS degree in computer science from UET Taxila in 2018. She achieved the B.Sc. degree in computer science from COMASTS University Wah in 2015. Besides this, she is the recipient of medal in BS degree program and several scholarships in her academic career. Her research areas include Computer Vision, Deep Fake Audio Synthesis and Detection, and Medical Imaging. Moreover, she is serving as Academic Editor in ICCK Journal of Image Analysis and Processing. (Email: rabbia.mahum@uettaxila.edu.pk)