1. ABSTRACT:

This development has enabled deepfakes to spread very rapidly and has now challenged the entire field of online security, media integrity, and human trust. This report discusses InceptionResnet model: A convolutional neural network architecture specifically for deepfakes detection, a novel framework combining two important features, the Inception model analyzes different details of the data at different scales, while the ResNet model helps the network learn better by skipping some layers, avoiding problems during training. The paper shows the architecture of InceptionResnet model. Experimental results clearly depict a strong accuracy robustness of such an architecture towards some benchmark data, including FaceForensics++, CelebDF, Deepfake Face Mask Detection (DFFMD), and Deepfake Detection Challenge.

2. INTRODUCTION:

Deepfake technology uses advanced artificial intelligence to create highly realistic fake media, including videos, images, and audio recordings. Concern about deep fakes has spread in recent years as they have shown rapid progress and spread widely. Deep fakes will bring about opportunities for entertainment and education, for instance, but on the darker side, deep fakes come with serious risks.They have been used in misinformation, manipulation of public opinion, and identity theft—all of which have been done with serious personal and societal consequences. They are also a threat to privacy because they can be used to manipulate content through the use of someone's likeness without permission. With ongoing development and increasing accessibility of deepfake technology, it is now increasingly hard to tell the real from the fakes, making it very hard to sustain trust in digital content.With this increasing concern, it highlights the need for developing reliable methods of deepfake detection to identify and curb such misuse of synthetic media.Most important are these issues, in the realm of ensuring that AI is used ethically and safely in the creation and sharing of media.

Especially, machine learning models—more precisely, convolutional neural networks (CNNs)—achieved great success in deep fake detection through pattern analysis in images, and audio. Inception-ResNet is a powerful deep-learning tool for detecting deep fakes: those fake images and videos created to fool people. Combining two important features, the Inception model analyzes different details of the data at different scales, while the ResNet model helps the network learn better by skipping some layers, avoiding problems during training. This mixture makes Start-Reset effective in detecting small changes in deep forgeries so that it might tell the difference between real and changed media. It is a very valuable model in combating digital deception.

3. PROBLEM DESCRIPTION:

Using sophisticated artificial intelligence and machine learning techniques, deepfake is a type of modified or synthetic material that includes audio, video, and images. The name "deepfake" combines the words "deep learning" with "fake" because deep learning models, in particular Generative Adversarial Networks (GANs) are used to produce realistic-looking but entirely fake content.

Deepfakes can include many manipulations, for example: Swaps in the face, where the face of a person is replaced by another in the video or imitation of the voice , where the voice of a person is imitated. Create recordings of fake audio. Lip-syncing, where it is a relatively new practice where a human mouth inside a video is mimicking an audioclitic that appears to be speaking; they are not. These manipulations depend on AI, which has been trained on vast amounts of real-world examples, and so often permit the creation of media that is indistinguishable from real.

Deepfakes can be used creatively and artistically, but they are quite serious and worrisome issues, particularly when used as a tool to sow disinformation, create fake content, or tamper with public figures. Deepfakes bring forth many critical problems, from spreading misinformation, reputation damage, and cybersecurity issues due to impersonation and phishing, to trust in media since distinguishing between authentic and fake is quite challenging. The victims could be emotionally damaged, while business and institutional reputations and financial outcomes may be threatened. There is also a challenge to the law and ethics regarding current frameworks unable to regulate creation and misuse. These issues highlight the need for detection technologies, public awareness, and robust legal measures to curb the risks.

This is where the Inception-ResNet model comes in. It can do better analysis on media mismatches by combining the features of Inception, which can capture multi-scale features, and the skip connections used in ResNet. These capabilities make them particularly useful for detecting subtle manipulations in deep fakes, such as slight distortions in images and audio that may be missed by the human eye or less sophisticated detection methods. By combining deep learning techniques with large datasets, we are now able to train start-reset models to effectively address the complex problems presented by deepfake technology. An effective solution will be provided so that the presence of fake content will not harm society, security, and privacy.

4. RELATED WORK:

This area of research is becoming important as the advances in generative models are happening so rapidly to produce hyper-realistic fake videos. Various methods have been proposed for deepfake detection across different modalities, such as images, audio, and video. The main areas of related research in the field of deep fake detection are discussed below:

Deep fake detection has evolved a lot since its inception. Early approaches are based on conventional forensic techniques. They analyze the visual artifacts, lighting inconsistencies, and pixel-level anomalies. In most cases, these methods needed manual analysis, which was highly time-consuming. Simultaneously, researchers began researching handcrafted features, facial feature analysis and micro-movements, like eye blinking and head turning. These were fed into a deep learning model that classifies the video as real or fake. While these initial methods laid down the foundation for the detection of deep fakes, they often trailed behind the fast pace at which the developments in deepfake technology were taking place.

Deep Learning Techniques:

Deep fake detection has advanced through different deep learning models and techniques. CNNs identify hierarchical features such as simple edges and textures and more complex patterns such as facial features, lighting inconsistencies, and pixel-level anomalies by applying convolutional layers. Such skills make CNNs very effective in detecting artifacts of deepfake generation methods such as subtle inconsistencies of texture, unnatural lighting changes, and distorted facial structuresCNN’s are highly popular nowadays, mainly due to their capabilities for spatial feature extraction; a few of the most notable contributions came from FaceForensics++ benchmark when they highlighted transfer learning's efficiency.

Additionally, we used Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM) networks and Guided Recurrent Units (GRUs) to capture the temporal inconsistencies in the video sequences. A dual-stream network that combines spatial and temporal features further improves the detection capability.

Recently, attention mechanisms and transformers such as Vision Transformers (ViT) have gained attention by focusing on important regions and capturing subtle manipulation artifacts. Another promising area is frequency domain analysis, where methods leverage Fourier transform-based features to detect manipulation artifacts invisible in the spatial domain, often combined with spatial features in hybrid models to

improve accuracy. In addition, ensemble learning techniques, such as bagging, boosting, and model stacking, combine multiple models-inclusive of CNNs, RNNs, and frequency-based methods-for the exploitation of their combined strength.

GANs have been widely adopted for the construction of deep fakes, they could also be repurposed for detection using adversarial learning and deepfake identification using GAN-specific fingerprints. Baseline datasets, like DFDC, Celeb-DF, and FaceForensics++, were crucial to standardize evaluation and encourage creativity. Aggregate methods number many, chasing increased detection accuracy, robustness, and generalization across manipulations.

Researchers have used InceptionResNet with other architectures, such as CNNs, RNNs, or transformers, to tap into its potential for detection performance. One common way is to use InceptionResNet-v2 as a feature extractor and append some classifier to detect the inconsistency in video frames. The hybrid model that integrates the spatial analysis capabilities of InceptionResNet with the temporal dynamics from models like LSTM increases the accuracy. Transfer Learning with InceptionResNet.

As manipulated media is so prevalent and can be used for any nefarious purpose, deep fake detection is an area of great research. Different deep learning models and methodologies have been explored to detect deep fakes effectively.

5. CURRENT STATUS OF WORK:

InceptionResNet models have been applied with success in transfer learning to fine-tune pre-trained models on deepfake detection tasks. Researchers, using the approach of initializing the model with pre-trained weights on large datasets like ImageNet and then fine-tuning it on a specific deepfake dataset, for example, DFDC or Celeb-DF, obtain high performance in identifying processed images and videos. The model, commonly used through different deep -factor type, has an important advantage in its ability.

Some deep fake detection tasks and benchmarks use InceptionResnet as a baseline or reference model. A model built on InceptionResNet was among the best at distinguishing real videos from fake in the Facebook-sponsored DeepFake Detection Challenge (DFDC). Such a model processes high-resolution video with little or no additional overhead, which means it's great for deep fake detection in real-time systems. Detecting Manipulation Artifacts is one of the more actionable applications for InceptionResNet is how it greatly facilitates the detection of GAN artifacts as well as deep fake generation methods. According to researchers, InceptionResNet models

detect slight pixel distortions, unnatural lighting, and facial shape inconsistencies so easily missed by others.

Some works have used InceptionResNet in multimodal detection frameworks. These approaches combine both visual and audio signals from Deepfake videos. By integrating InceptionResnet with audio treatment models-for example, using LSTM networks or attention for sound analysis-researchers can more effectively identify the Faked which handle both the visual and hearing aspects of the videos.

6. PROPOSED MODEL AND METHODOLOGIES:

I have worked on various deep fake detection datasets using InceptionResnet model, which combines the forces of the Inception network and the residual network (RESNET). It uses the multi-test architecture of the initial model to capture functions at different scales, and also includes residual connections to improve the training and optimization of performance in deep networks.

6.1 Architecture:

- Entry point:
  Input image size: 299x299x3 (height x width x channels for RGB images).

- Stem:
  The Stem module down samples the input data and extracts low-level features.

- Beginning-Resnet-A:
  Repeated 5 times. Each module performs parallel convolutions (1x1 and 3x3) and combines them with residual connections. It helps extract features while preserving the input information.

- Reduction-A:
  Convolve and pool to reduce the spatial resolution of the feature maps. Spatial dimensions reduced, and depth increased to capture more abstract features.

- Inception-ResNet-B:
  10 iterations. More complex module with larger convolutional filters, like 7x1 and 1x7, to capture long-range dependencies. Residual connections are preserved for stability.

- B-Reduction:
  Further reduces sampling of feature maps with pooling and convolutions. Reduces spatial dimensions and increases depth.

- Inception-ResNet-C:
  5 iterations. It is built to extract fine-grained features by using smaller convolutions, such as 1x1 and 3x3. Residual connections are used to keep the network stable during training.

- Global Pooling Averaging:
  Reduces feature maps to one value per channel, which gives a one-dimensional vector. Efficient for summarizing spatial information.

- Dropout:
  A dropout rate of 0.8 (keeping 80% of neurons) is applied to avoid overfitting.

- Softmax:
  The final fully connected layer computes the features for class probabilities in the softmax function.

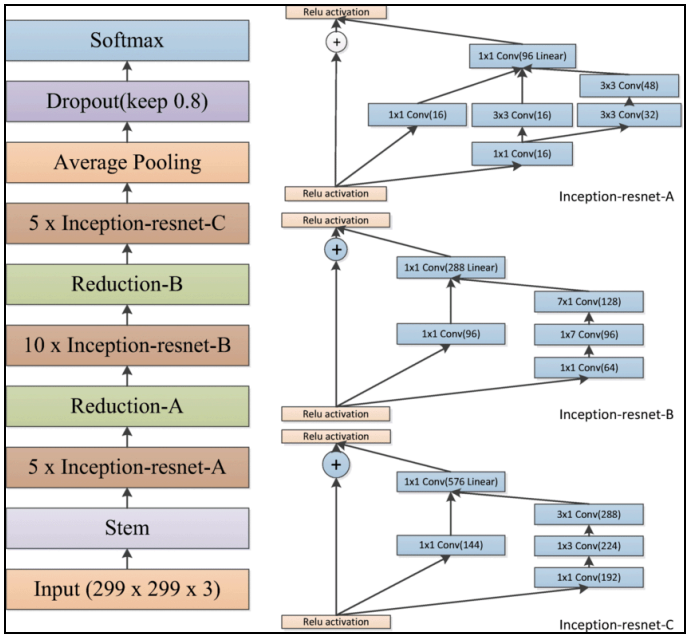


Fig1: key components of the Inception-ResNet-v2 architecture.

6.2 Datasets:

I have used four publicly available datasets to train and test our model. The four datasets that we will using for our work are:

6.2.1 Celeb DF Dataset:

This dataset has one folder of real images and one folder of fake images. CelebDF consists of 5639 fake videos and 890 real videos making a total count of 6529 images featuring 59 celebrities. The dataset includes videos of celebrities across various ages, genders, and ethnicities, ensuring diversity and helping researchers create generalizable models.

6.2.2 Face Forensics ++ Dataset:

FaceForensics++ is the dataset of short video clips which are sourced from various videos from the internet. This dataset has a total of 400 videos out of which 200 videos are real and the other 200 videos are fake, which are edited/morphed using various techniques.

6.2.3 DFDC Dataset:

DFDC dataset is also another video dataset consisting of a large number of videos. The DFDC dataset that we used is a dataset of images frame which are extracted from the original DFDC video dataset. Our dataset has 581 fake and 130 real image frames

6.2.4 DFFMD Dataset:

The DFFMD dataset (Deepfake Masks Dataset) was created as part of a study to address the challenge of detecting deep fakes due to the COVID-19 pandemic in an era of widespread mask wearing. It consists of faces of people with masks. The dataset consists of total 1000 fake videos and 834 real videos.

6.3: Methodology 1: Unmasking Deepfake Faces from Videos Using An Explainable Cost-Sensitive Deep Learning Approach

This methodology deals with the problem of detecting deep characteristics in video content using a sensitive approach to deep training. I have used pre-trained Convolutional neural networks (CNN), InceptionResnet model to be precise as well as AI (XAI) tools to detect and explain manipulations with deep content in the video.I have implemented the model on celebDF and face forensics ++ deepfake datasets available in kaggle. By addressing issues such as dataset imbalance and optimizing computational efficiency, this method prioritizes high accuracy while maintaining resource efficiency.

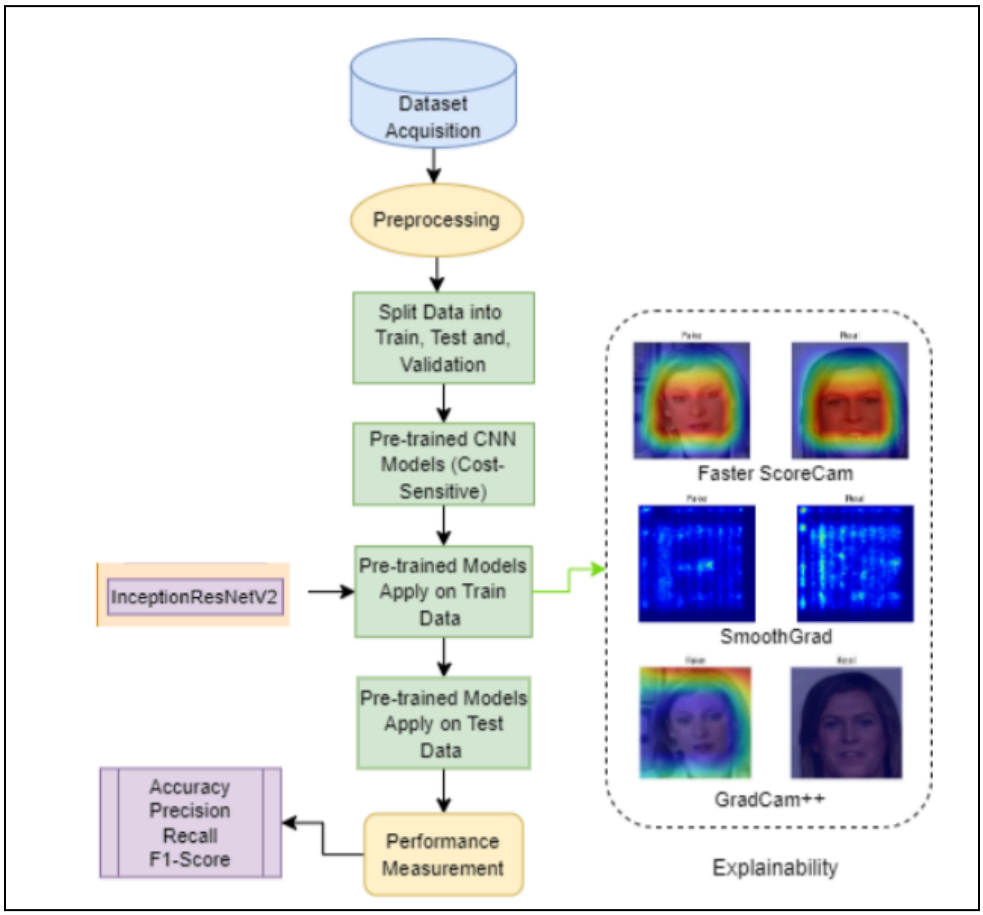


Fig2: Methodology

The methodology involves:

- Datasets:
  I have used two well known deepfake datasets - **CelebDF & Face Forensics ++**
  CelebDF consists of 5639 fake videos and 890 real videos whereas Face forensics ++ dataset consists of 200 fake videos and 200 real videos.

- Data Preprocessing:
  Face Recognition is done using the face_recognition Python library. It detects and extracts faces from video frames. Videos that are corrupted are identified and removed from the dataset. 5338 frames were extracted from celebdf dataset and 400 frames were extracted from face forensics ++ dataset which is a mix of both real and fake frames extracted from respective videos. This step also includes splitting of the data into training at 80%, validation at 10%, and a test at 10% with stratification. This will keep class proportions.

- InceptionResnet Model:
  This InceptionResNetV2 model is cost-aware and fine-tuned on datasets such as CelebDF-V2 and FaceForensics++. Class weights are computed from the distribution of fake and real samples and used during training to give more weight to minority classes.

- Training and Testing:
  The pretrained InceptionResNetV2 model is applied to both training and testing datasets after fine-tuning. The model is trained setting the hyperparameters values as follows:

  Epochs: 15
  Learning Rate: 0.001
  Optimizer: Adam
  Batch Size: 16
  Dropout: 0.5
  Activation: Softmax function

  Accuracy, precision, recall and F1-score are performance metrics.

- Explainability:
  The faster versions of ScoreCAM, SmoothGrad and GradCAM++ have been applied for visualizing parts of images that the model is trying to classify, hence making it easier to interpret and understand the results.

6.4: Methodology 2: Advances in DeepFake Detection: Leveraging InceptionResNetV2 for Reliable Video Authentication

In this methodology, I have used InceptionResnet feeding the model with some preprocessed image frames using techniques like facial identification and face cropping from the original video dataset. The process begins by gathering a video dataset from the Kaggle DFDC dataset. Preprocessing follows, including the extraction, identification, cropping, and adaptation of video frames to analyze them. These cropped frames are fed to the model while maintaining uniformity in the detection process. The InceptionResNetV2 model has been trained to classify whether the video content is original or manipulated.

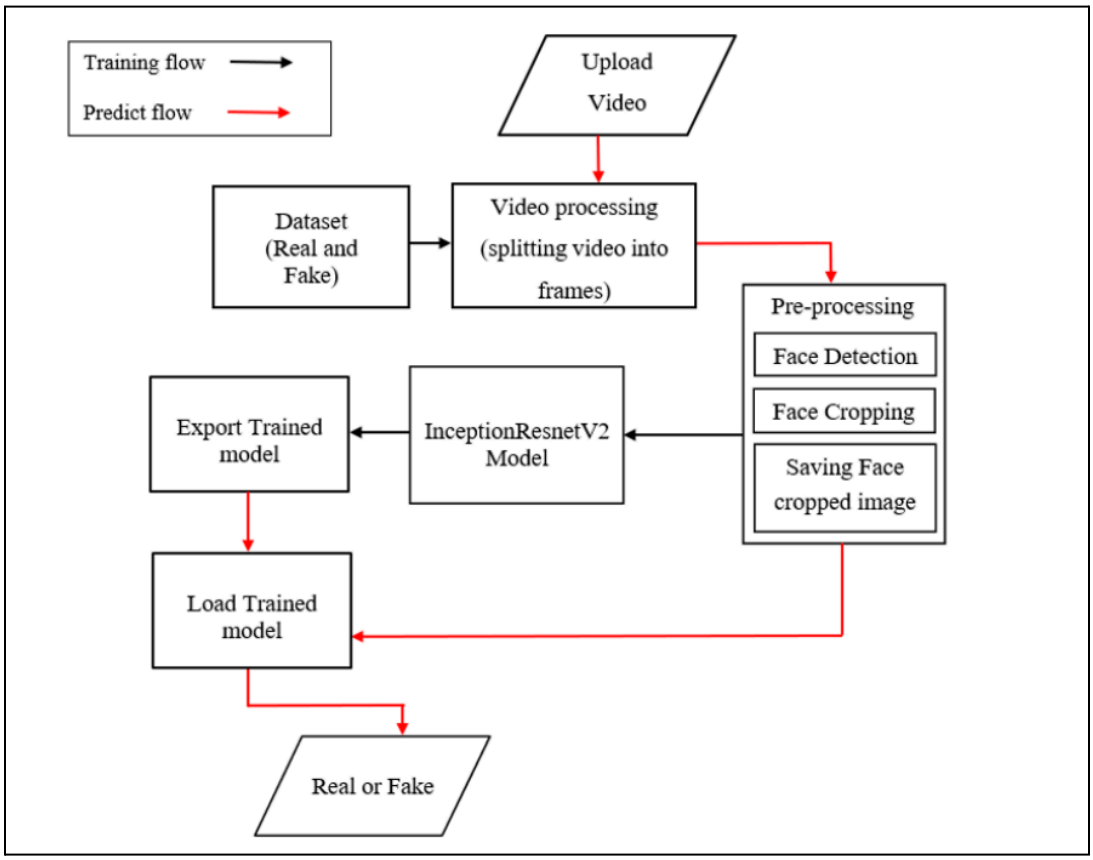


Fig3: Methodology

Initially, we have the dataset which is the extracted frames dataset from the original DFDC video dataset which is downloaded from kaggle. It consists of 581 fake image frames and 130 real image frames. These image frames were preprocessed using some techniques which ensures that the data is prepared and optimized for effective analysis by our model.

The Preprocessing techniques involves:

- Face Recognition
  Face recognition is the critical processing stage in each frame, and it is done using state-of-the-art computer vision techniques that may include the use of facial detection. I have used Haar Cascades face detection algorithm which detect and characterize facial features in each frame The algo uses **haar-Like Features** which are patterns of rectangular regions that compute the difference in pixel intensities.

- Cropping and resizing
  When a face is recognized in a frame, it is carefully cropped. It encompasses the cropping of the identified face from the rest of the frame. Isolation of the detected face enables the extraction of features. Only those areas are focused on related to facial features. The cropped faces are resized to the same size, thus uniform for further analysis.

Then, these preprocessed images are sent to InceptionResnet Model for training. The dataset goes for splitting with a ratio of 80:20 where 80% is for training and 20% is for validation. Testing data is taken from the dataset which has a total of 79 images which is a combination of both real and fake images. These image frames were trained for analysing the original and manipulated images. Hyperparameters are used to enhance training in a more efficient way. I have used the following hyperparameters:

- Epochs: 10
- Batch size: 32
- Global Average Pooling
- Dropout of 0.5
- Learning Rate: 0.001
- Adam optimizer

Performance metrics are followed after training and testing. Accuracy, precision, F1 score, recall are calculated using the confusion matrix

6.5: Methodology 3: DFFMD: A Deepfake Face Mask Dataset for Infectious Disease Era with Deepfake Detection Algorithms.

This methodology is about the InceptionResnet model performance on the DFFMD mask deep fake dataset. The InceptionResnet architecture has a slight modification here which involves addition of batch normalization which is applied just on top of the traditional layers of residual summations. The pipeline of the methodology is as follows:

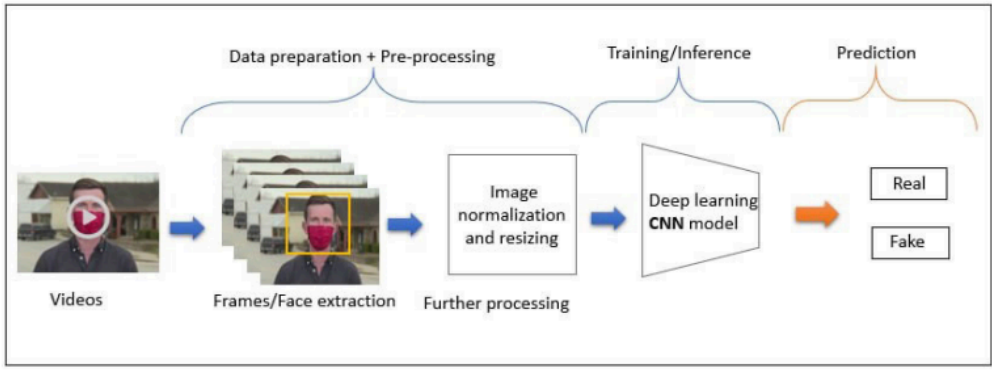


Fig3: Pipeline of the methodology

The DFFMD dataset (Deepfake Masks Dataset) was created as part of a study to address the challenge of detecting deep fakes due to the COVID-19 pandemic in an era of widespread mask wearing. It consists of faces of people with masks. The dataset consists of total 1000 fake videos and 834 real videos.

Frames are extracted from the videos using OpenCV. These extracted frames are used for model training and testing. InceptionResnet model is used as the deep learning CNN model. When building Inception-ResNet, the hyperparameters were set as follows: The fully connected layer is the last layer in the network. Additionally, we set weights as imagenet to use the weights of the pretrained model. The network was connected to convolutional layers with filters of 1024 and padding as same, followed by an activation layer with the 'ReLU' function. At the end of the network, i have included the BatchNormalization layer, GlobalAveragePooling2D, and the dense layer, a fully connected layer with two output classes (fake or real) and activation function softmax. Adam was used as an optimizer in this methodology, with a learning rate = 1e-5. THe model was trained for 20 epochs setting the value of batch size as 64. At last, we obtain performance metrics which involves accuracy, precision, recall, F1 score which tells us about the performance of our model on the dataset.

7. EXPERIMENTAL RESULTS:

7.1: Results of Methodology 1:

These results are based on Face forensics ++ dataset and CelebDF dataset. The results includes classification reports and confusion matrix of respective datasets. Training and validation accuracy graphs were also plotted to observe the model performance at every epoch. The following are the results which describes about the performance of the model.

Datasets	Accuracy	Precision	Recall	F1 Score
CelebDF	0.91	0.92	0.91	0.91
Face Forensics ++	0.62	0.79	0.62	0.56

Table 1: Performance Metrics of Weighted Average on respective datasets

7.1.1 Celeb DF:

Below are the classification report and confusion matrix of the dataset providing the insights of the InceptionResnet model:

Classification Report:

	precision	recall	f1-score	support
0	0.99	0.84	0.91	267
1	0.86	0.99	0.92	267
accuracy			0.91	534
macro avg	0.92	0.91	0.91	534
weighted avg	0.92	0.91	0.91	534

Fig4: Classification report of CelebDF dataset

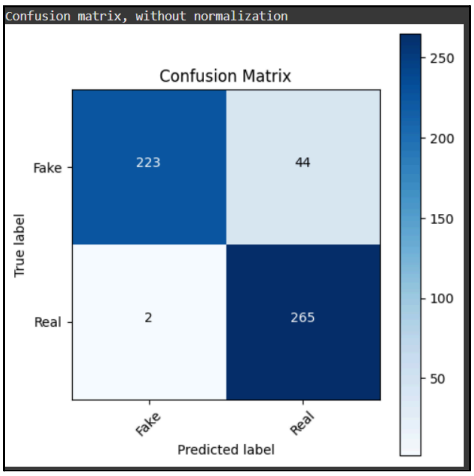


Fig5: Confusion Matrix

7.1.2 Face Forensics++:

Below are the classification report and confusion matrix of the dataset providing the insights of the InceptionResnet model. Training and validation accuracy graph is also plotted. Below is the the graph of last 5 epochs:

Classification Report:

	precision	recall	f1-score	support
0	1.00	0.25	0.40	20
1	0.57	1.00	0.73	20
accuracy			0.62	40
macro avg	0.79	0.62	0.56	40
weighted avg	0.79	0.62	0.56	40

Fig6: Classification report of Face Forensics++ dataset

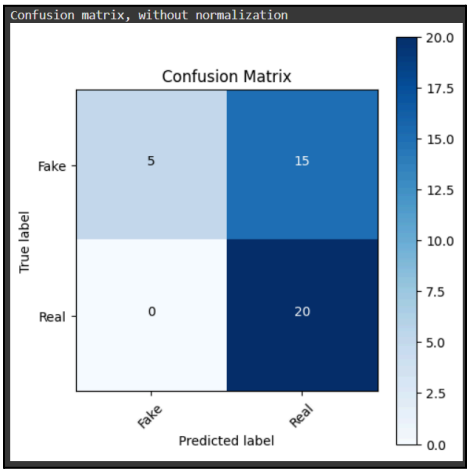


Fig7: Confusion Matrix

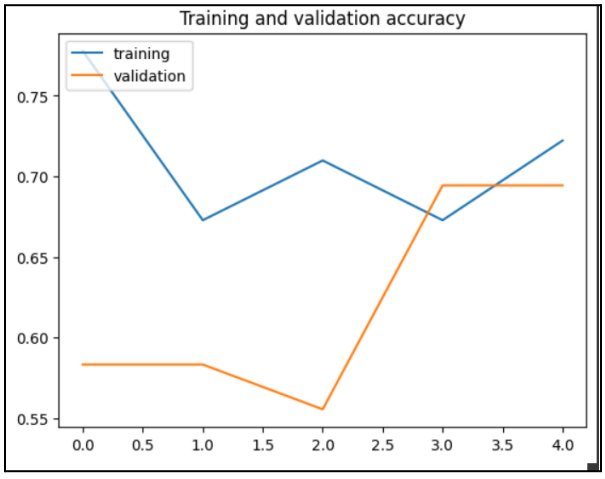
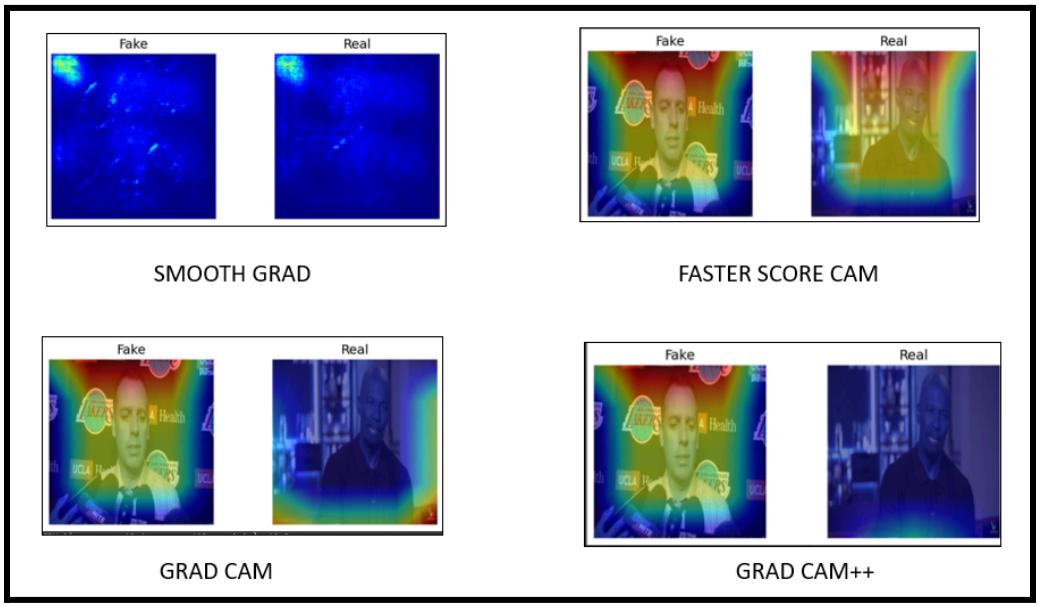


Fig8: Training & validations accuracy graph

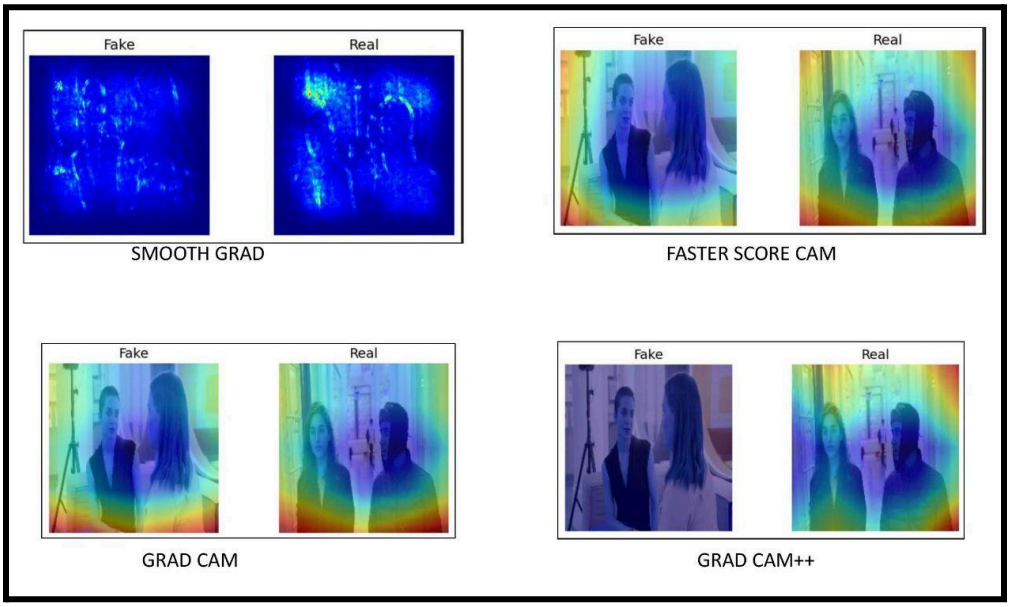
7.1.3 Explainable AI (XAI):

The methodology also integrates Explainable AI (XAI) methods, such as GradCAM, GradCAM++, and Faster Score-CAM. These tools effectively highlight the critical facial regions that distinguish real from fake frames, enhancing the system's transparency and reliability.

CelebDF:



Face Forensics ++:



7.2: Results of Methodology 2:

These results are based on DFDC dataset. The results includes classification reports and confusion matrix of the dataset. Training and validation accuracy graphs and losses were also plotted to observe the model performance at every epoch. The following are the results which describes about the performance of the model:

Dataset	Accuracy	Precision	Recall	F1 score
DFDC	0.81	0.67	0.81	0.73

Table 2: Performance Metrics of Weighted Average on DFDC dataset

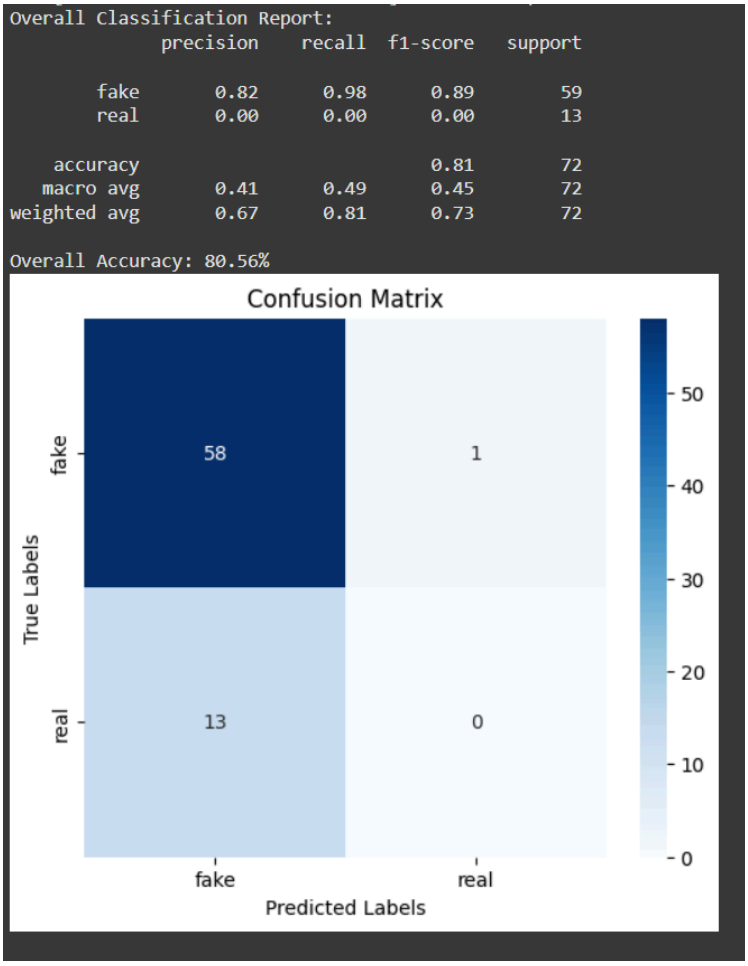


Fig9: Classification Report and Confusion matrix of DFDC dataset

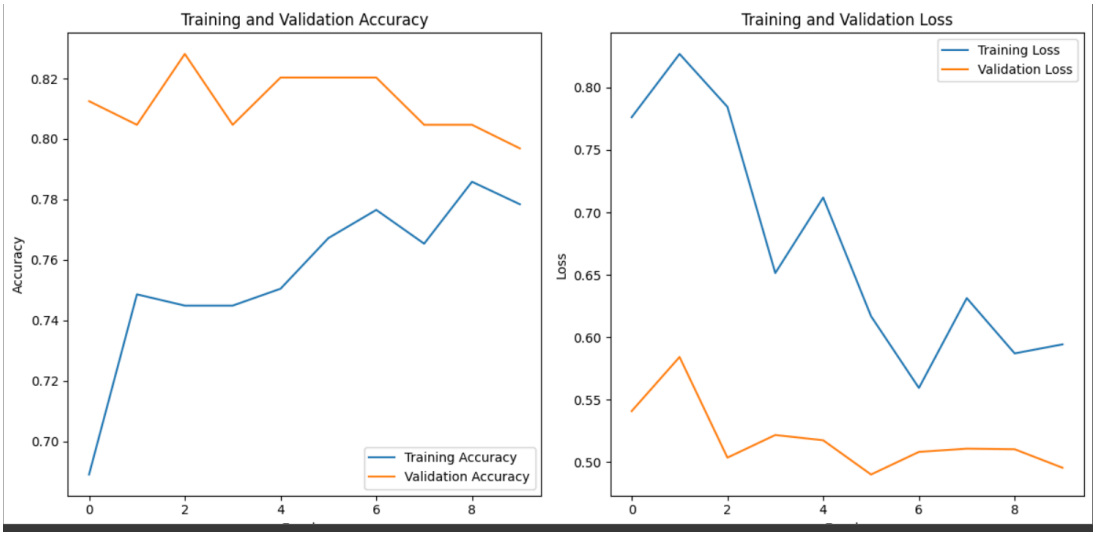


Fig10: Training & validation accuracy and loss graph

7.3: Results of Methodology 3:

These results are based on DFFMD dataset. The results includes classification reports and confusion matrix of the dataset. Training and validation accuracy graphs and losses were also plotted to observe the model performance at every epoch. The performance metrics are also represented in graphical form. The following are the results which describes about the performance of the model:

Dataset	Accuracy	Precision	Recall	F1 Score
DFFMD	0.96	0.9695	0.9673	0.9674

Table 3: Performance Metrics of Weighted Average on DFFMD dataset

```
Detailed Classification Report:
              precision    recall  f1-score   support

           0     0.9330    1.0000    0.9653       167
           1     1.0000    0.9400    0.9691       200

    accuracy                         0.9673       367
   macro avg     0.9665    0.9700    0.9672       367
weighted avg     0.9695    0.9673    0.9674       367
```

Fig10: Classification Report and Confusion matrix of DFDC dataset

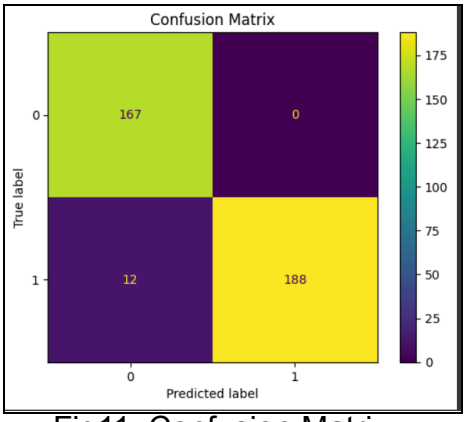


Fig11: Confusion Matrix

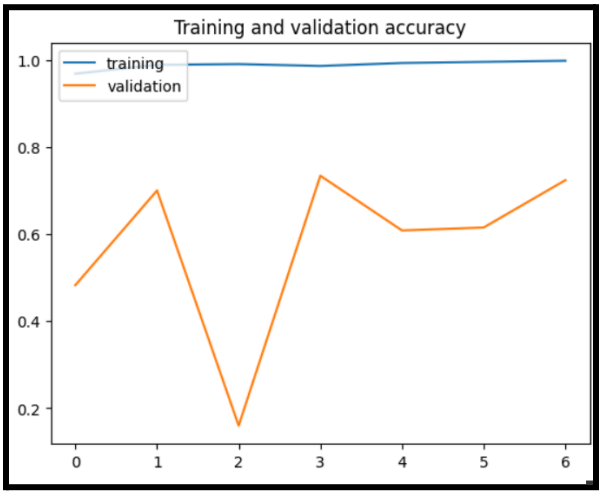


Fig12: Training & validation accuracy and loss graph

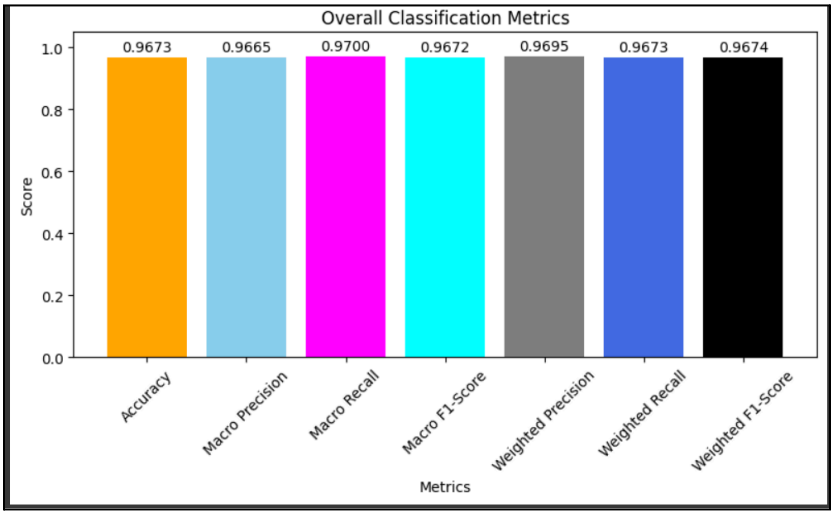


Fig13: Graph of overall classification metrics

8.CONCLUSION AND FUTURE WORK:

In my project, I have explored various issues related to deep fakes and a powerful approach for their detection using the Inception-ResNet model. Deep fakes are getting even more sophisticated, and the dangers to privacy, security, and the trustworthiness of digital media have come into sharp relief. Our model shows high accuracy in discriminating between real and manipulated content, pointing out that the combination of an advanced neural architecture with well-crafted datasets is effective.

The Inception-ResNet model is employed in a hybrid structure with Inception modules and residual connections, and it has shown especially good results in capturing spatial and temporal deep fake video features. The experimental results demonstrate that the approach is efficient in detecting manipulated content when tested on diverse datasets containing various types of deep fakes. This further increases the model's robustness to low-resolution/compressed media, making it very practical for use in scenarios such as social media and content verification systems.

Although the proposed Inception-ResNet-based deepfake detection model shows very promising results, there are several directions open for future research and development.

- One essential direction is enhancing the generalizability of the model. Deep forgery detection systems generally suffer from poor generalization performance when applied to new, unseen datasets with diverse characteristics. So, developing that model that will be able to generalize to new and emerging deep forgery techniques may be a future working scope.
- Another necessary area is to enhance the robustness of the detection systems against adversarial attacks because adversaries are still developing methods for circumventing existing defenses. The combination of adversarial training or ensemble techniques can help mitigate such vulnerabilities.
- The Inception-ResNet model's computational complexity is a challenge for real-time applications, especially on resource-constrained devices. Hence, optimization of the architecture or designing lightweight models for deployment on mobile and edge devices is a future perspective.
- Another aspect of importance is bringing explainability within the detection process. Deepfake detection models should be made interpretable to increase user trust and to give insight into how the system identifies manipulation.

Researchers will have to read carefully with regards to privacy issues and misuse of detection techniques. Collaboration with industry stakeholders and policymakers is critical in ensuring these tools are developed and deployed responsibly. Taking these directions, future research can build on the current work to create robust, scalable, and ethical solutions against the increasing threat of deep fakes.

9.SIGNIFICANT RESOURCES:

- https://www.researchgate.net/publication/378533983_Unmasking_Deepfake_Faces_from_Videos_Using_An_Explainable_Cost-Sensitive_Deep_Learning_Approach
- https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/article/view/4819
- https://www.researchgate.net/publication/368670743_DFFMD_A_Deepfake_Face_Mask_Dataset_for_Infectious_Disease_Era_with_Deepfake_Detection_Algorithms
- https://www.kaggle.com/datasets/hungle3401/faceforensics
- https://www.kaggle.com/datasets/reubensuju/celeb-df-v2
- https://www.kaggle.com/datasets/ashifurrahman34/dfdc-dataset
- https://www.kaggle.com/datasets/hhalalwi/deepfake-face-mask-dataset-dffmd