Facial Video Forgery (DeepFake) Detection

Team-6

- Suyash Chintawar - 191IT109
- Naveen Shenoy - 191IT134
- Sarthak Jain - 191IT145

Introduction

Deepfake is a technique of generating synthesized video to swap the face of the person in the video with the face on the provided image in such a way that the generated video has the target person doing or saying things the source person does.

Motivation

- The popularity of smartphones and the growth of social networks in recent decades have made digital images and videos very common digital objects.
- Efficient DeepFake detection models can be used for surveillance to remove forged videos from circulating on the internet.

The aim of this research is to successfully **classify and identify facial forgery** in videos by using novel techniques in the field of deep learning and video processing.

Literature Survey

Sl. No	Paper	Methodology	Limitations	Year
1.	DeepFakeHop	Three major steps. 1) Successive subspace learning (SSL) principle to extracts features automatically. 2) Feature distillation and finally 3) Ensemble classification.	Different classifiers could be used in the classification step.	2021
2.	DeepFakeHop++	DefakeHop++ includes eight more landmarks for broader coverage in addition to 2 eyes and mouth.	Techniques like GA can improve accuracy further	2022
3.	EfficientNets for DeepFake Detection	Preprocessing involved frame rate reduction followed by augmentation and feature extraction using EfficientNets. Finally simple binary classifiers used.	Evolutionary computation techniques could have been used.	2021
4.	EfficientNet and Vision Transformers for Deepfakes	Mixed convolutional-transformer architecture used which involve face extraction using MT-CNN and vision transformers for informative global description.	Distillation/ensemble techniques could be used.	2022

Literature Survey

Sl. No	Paper	Methodology	Limitations	Year
5.	Protecting World Leaders Against Deep Fakes	Approach includes tracking of facial and head movements and then extracting the presence and strength of specific action units and applying SVM for prediction	Used videos of very few POIs (point of interest)	2019
6.	Hybrid LSTM and Encoder–Decoder Architecture for Detection of Image Forgeries	This paper proposes an architecture that utilizes long short-term memory (LSTM) cells, and an encoder–decoder network to segment out manipulated regions from non-manipulated ones	Fails to work for videos with low contrast.	2019
7.	Long-Term Recurrent Convolutional Networks for Visual Recognition and Description	Long-term Recurrent Convolutional Networks (LRCNs), based model is designed for visual recognition and description task	Fails for domain specific tasks and requires huge amount of data to train	2020

Literature Survey

Sl. No	Paper	Methodology	Limitations	Year
8.	Machine Learning approach for Deepfake detection	Methodology includes steps of frame extraction, face detection and image processing which is followed by predictor models backed by EfficientNets.	Testing on images from other sources and more varied manipulation techniques, increasing the generalization of the model for predictions.	2021
9.	Cascaded-Hop For DeepFake Videos Detection	A preprocessing method based on the image pixel matrix feature to eliminate similar images and the residual channel attention network (RCAN) to resize the scale of images.	When the image is large enough, the cropped subspace size may only cover a relatively small region	2022
10.	Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics	This paper introduces the CelebDF dataset and the techniques used to create them. Comparison of baseline models on various datasets such as DFDC.	Does not examine performance on latest CNN architectures such as EfficientNet	2020

Outcomes of Literature Review

- Many datasets are publicly available like UADFV, DFDC, CelebNet, FaceForensics++, etc. One of the most difficult of all is CelebDF [Source]
- EfficientNet features have been proved useful for this task done by many researchers. [Paper]
- Most techniques using deep learning have not using evolutionary intelligence to select frames. Frame selection have been done using naive equidistant sampling or statistical based methods.
- Other techniques used include the use of encoder-decoder architectures, use of CNNs, LSTMs, etc.

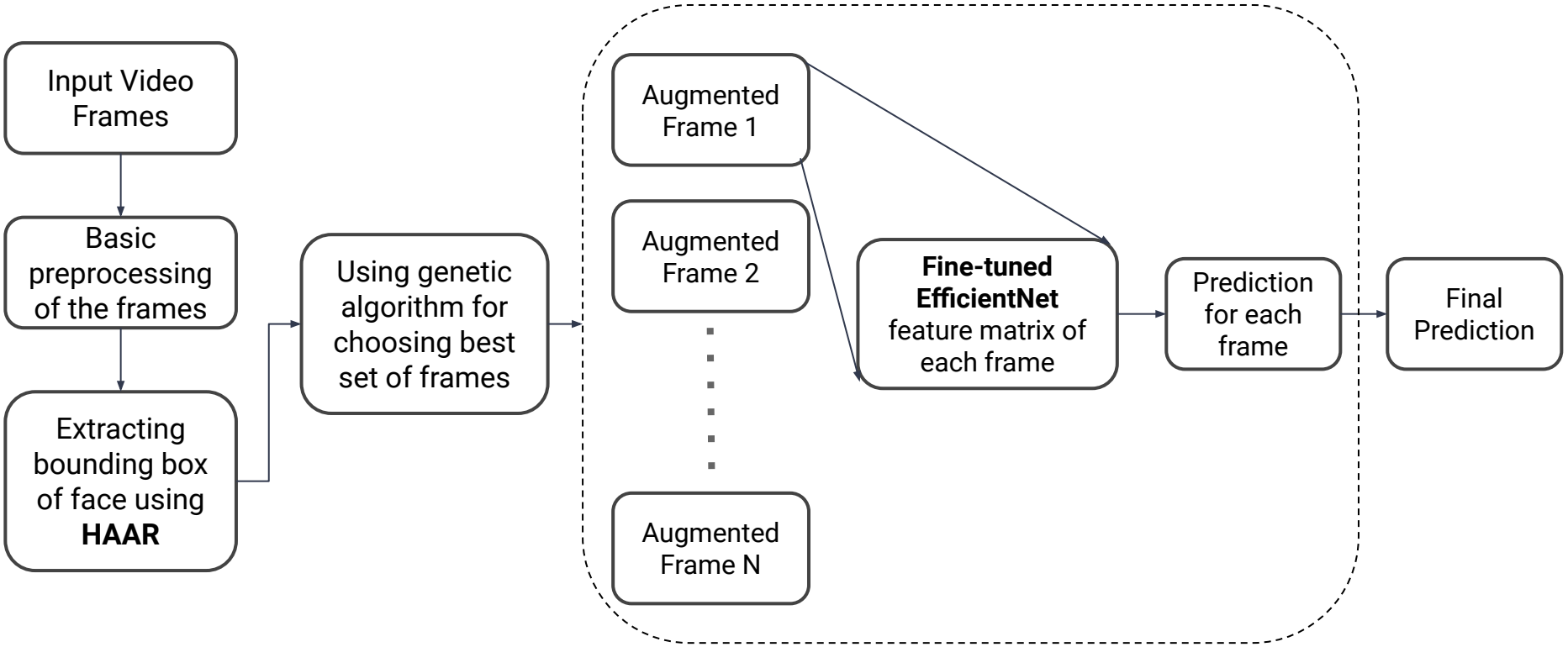
Problem Statement and Objectives

To build an efficient deep learning model to identify deepfake videos using evolutionary intelligence techniques

Objectives

- To extract facial features from video frames by using frame selection methods followed by HAAR.
- To obtain a subset of frames from the video using genetic algorithm to input into deep learning model.
- To fine-tune deep learning models such as EfficientNet, InceptionV3 to identify deepfakes.

Methodology



Methodology: Genetic Algorithm

Frame Selection Using GA: Each video can be represented as the binary string where each bit represents the frames and bit set to one implies the selection of that frame.

Fitness Function:

- We calculate the histogram of each frame and find the difference of each frame's histogram with that of other frames. If the difference is lesser than certain threshold, add it to score. Higher the score, the higher are the chances of the similarity of the current frame with the other frames.
- Factor of distance is also considered. We need dissimilar frames that are at a lesser distance to capture more data.
- Net fitness value becomes proportional to **the inverse of distance** and also logarithmic reciprocal of above mentioned score. Logarithm is used for scaling score.

Methodology: Genetic Algorithm

- Initially we start with the randomly generated strings each comprising of 25 ones (1's) as we intend to take 25 frames for the next steps of framework.
- First step is selection of parents. For this fitness values of all 4 parents are calculated then by using the roulette wheel selection two parents are selected and crossover and mutation operators are applied on them.
- Crossover: A random point is selected and interchanging is performed about that point.
- Mutation: Since the resulting string from prev step may contain more or less number of ones than required, we perform the mutation to bring a number of ones equal to the required number. If the number of ones is equal to the required number of ones than no mutation is performed.

Methodology: Genetic Algorithm

- If the number of ones exceeds the required number of ones, we perform the merging operation. It refers to the merging of adjacent segments until a number of ones become equal to 25. If the number of ones is less than the required number of ones then randomly zeroes are picked and changed to one.
- After the mutation operations both the strings are saved for the next generation. This entire process is repeated once again since we need 4 strings in any of the generation. This continues till the fixed number of iterations.

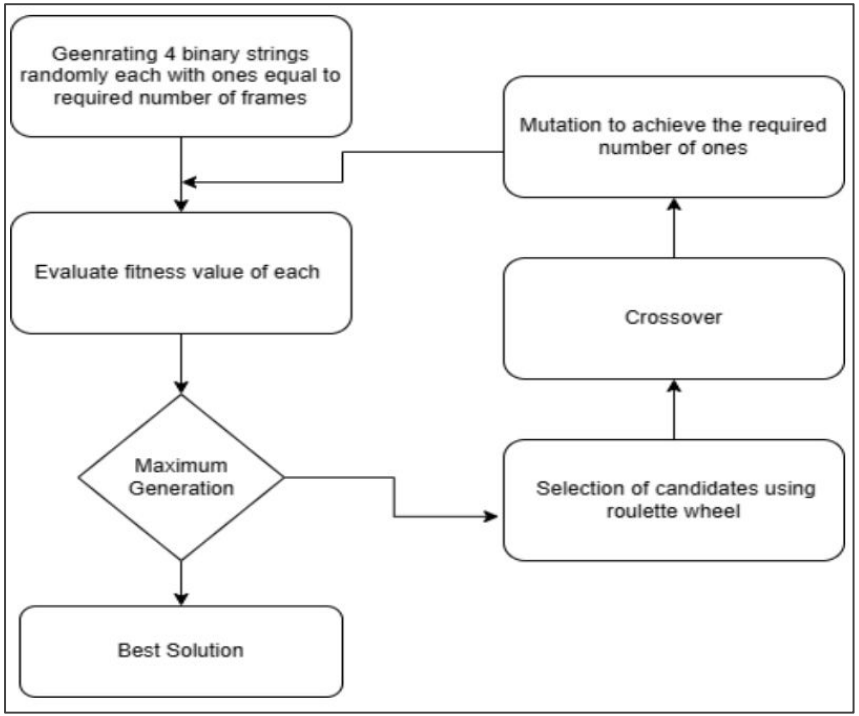
Methodology: Genetic Algorithm

Fitness Function:

- h(i, j) = histogram difference between i and j
- dh(i) = h(i −1, i)
- F ' = { j ∈ F | dh(j) > dh + σ }
- Ci = { j ∈ F ' | h(i , j) < dh + σ }

  we define a set Ci to be those elements similar to i
- Wi = |Ci| / |F'|
- I(Importance) = log(1/(Wi+1))

$$f(S_k) = \sum_{\substack{i,\, j \in S_k \\ i \neq j}} h(i,j) \frac{(I_i + I_j)}{|i-j|^2}$$



Methodology: InceptionV3

- The InceptionV3 model receives the extracted frames from the genetic algorithm as input. Initially, a pretrained InceptionV3 model is loaded.
- Additional layers are trained on top of the InceptionV3 model keeping the core Inception weights constant throughout the fine-tuning.
- The output shape of InceptionV3 is (1000,). This is converted to shape of (256,) using a dense layer followed by ReLU activation. Further dropout is applied to 20% of the neurons after which a dense layer is applied to finally receive a single probability value for a frame.

Methodology: EfficientNets

- The model architecture of EfficientNet includes inverted residual blocks which were originally proposed in MobileNetV2 architectures.
- There exists multiple variants of EfficientNet models from B0-B7. These eight models have increasing number of parameters as we move from B0 to B7. Thus, the computational complexity increases from B0 to B7.
- The image input dimension required for each variant is different. The input size for variants changes or rather increases from 224x224x3 for EfficientNet-B0 to 600x600x3 for EfficientNet-B7.
- In this research we fine tune the EfficientNet variants from B0-B4 by using multiple neural network layers like Linear layers, BatchNorm layers, etc. The input shape for each model is adjusted accordingly.

Dataset Description

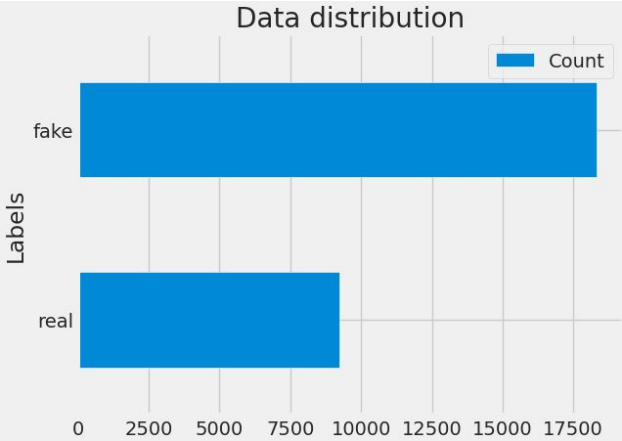


Fig. Data distribution of the final dataset

- The Celeb-DF dataset is used as the benchmark in this study.
- The Celeb-DF dataset contains 408 original YouTube clips with individuals of various ages, ethnic backgrounds, and genders, as well as 795 DeepFake videos generated from these real videos.
- It includes real and DeepFake generated videos with similar visual quality to those seen online.
- The train/test split of the dataset corresponds to 1103 training videos and 100 test videos

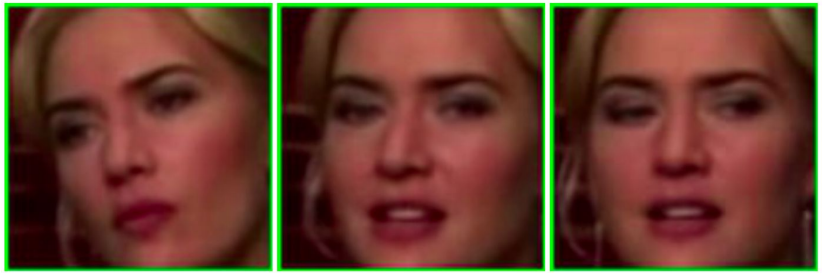
Results

Sl no.	Model	Accuracy
1.	HeadPose (Based on SVM)	54.8
2.	Meso4 (Based on CNN)	53.6
3.	Two Stream (Based on InceptionV3)	55.7
4.	DeFakeHop	95.0
5.	DeFakeHop++	97.5
6.	GA + Fine-tuned EfficientNetB4 *(proposed)*	**96.0**

Results

Sl no.	Model	Accuracy
1.	GA + Fine-tuned InceptionV3	62.0
2.	GA + Fine-tuned EfficientNetB0	89.0
3.	GA + Fine-tuned EfficientNetB1	93.0
4.	GA + Fine-tuned EfficientNetB2	94.0
5.	GA + Fine-tuned EfficientNetB3	89.0
6.	GA + Fine-tuned EfficientNetB4	**96.0**

Results (Test Data)



Ground Truth : Real
Predicted Label : Real

Ground Truth : Fake
Predicted Label : Fake

Ground Truth : Real
Predicted Label : Fake

Conclusion

- In this research, we propose a novel fine-tuned neural model for deepfake detection enhanced by genetic algorithm in the preprocessing stage.
- Our methodology involved obtaining faces from the frames of videos using HAAR cascade and applying genetic algorithm to obtain a subset of suitable frames for further fine-tuning by InceptionV3 and EfficientNet models.
- We find that the proposed models perform competitively when compared with the state-of-the-art baseline models. Improved performance of our approach when compared with similar deep learning models such as InceptionV3 without GA based preprocessing show superiority of our approach.
- Future work involves developing more task specific fine-tuned versions of the deep neural architectures.

THANK YOU