International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Securing Visual Integrity: An Efficient NetB4-Based Solution with Attention Layers and Siamese Training for Face Manipulation Detection in Videos

Nilakshi Jain[1], Shwetambari Borade[2], Bhavesh Patel[3], Vineet Kumar[4], Mustansir Godhrawala[5], Shubham Kolaskar[6], Yash Nagare[7], Pratham Shah[8], Jayan Shah[9]

Submitted: 26/01/2024 **Revised:** 04/03/2024 **Accepted:** 12/03/2024

Abstract: The public can now easily create video deepfakes because of the growth of machine learning and artificial intelligence in the current digital era. The system of society is severely affected by this technology. The phrase "Deep Fake" implies digital representations created by advanced artificial intelligence that are adapted to make erroneous sounds and sights that appear real. The identification of these motion pictures presents an important obstacle because of the occasional development of progressively realistic deepfake generating techniques. FaceSwap and deepfake are two programs that have made it easier for anyone to realistically alter faces in videos in recent years. Technological advances can be helpful, but they may also be misused, which may result in difficulties like the dissemination of misleading data or online bullying. For this reason, being able to recognize when a video has been altered is important. This research tackles the problem of face alteration detection in video sequences that aim to target modern facial manipulation methods in this research. Specifically, the research looks at a set of several Convolutional Neural Network (CNN) models that were successfully trained. The suggested method uses two separate concepts to generate multiple models starting from the fundamental network (EfficientNetB4): Layers of attention and instruction in Siamese. Thus, by such a structure this model attains an accuracy of 94% on FaceForensics and DFDC dataset.

Keywords: *Convolutional neural networks, Deepfake, Digital media forensics, Efficient netB4, Face forensics*

1 Professor, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0000-0002-6480-2796
2 Assistant Professor Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0000-0001-7547-6351
3 Professor, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0001-0363-9809
4 Founder & Global President, CyberPeace Foundation, Delhi, India
ORCID ID : 0009-0000-3806-7380
5 Student, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0005-4065-4361
6 Student, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0002-1394-7992
7 Student, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0003-1266-3709
8Student, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0006-0935-6865
9 Student, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0000-9677-9175

1. Introduction

Significant progress has been achieved with video editing techniques in recent years, especially in respect to facial adaptation. For example, now research [1] can easily and quickly change a speaker's recognition by moving their facial expressions from one film to another. Each approach maintains the same basic principle: every irreversible operation leaves an individual mark that can be recognised to figure out which editing completed the task. This forensic evidence, however, can frequently be imperceptible and subtle. This is especially true for videos that have gone through considerable sampling reduction, numerous simultaneous edits, or excessive compression. This remains true for highly realistic forgeries produced with techniques that can be difficult to formally model. This renders it extremely difficult, from a forensic viewpoint, to identify current facial modification processes.

Deepfakes are extremely realistic created media that are utilized for malicious activities, such spreading inaccurate data. These make use of customized techniques that modify specific areas of the obtained video frames from the original raw video. Certain portions that get superimposed and swapped with the target face are retained by the Deep Learning algorithms used for creating Deepfakes. As a result, the algorithm in [2] used functions in reverse. The same objectives, such as changes in lighting, lip and eye movements, are used to generate deepfake creation models.

In addition, the processed videos can show apparent evidence of tampering throughout the deepfake creation process, such as

temporal space discrepancies, lighting variations, and compression distortions.

These signs can be identified by Convolutional Neural Networks (CNN), which can be used to build strong protections against deepfake technology. The test videos are split into frame packs by the model, after which it processes them. CNNs greatly streamline the procedure of collecting faces, spatial data, and temporal features. Large-scale deepfake detection can be facilitated by CNNs because of their considerable potential and scalability [3]. Transfer learning can be used effectively to the detecting procedure. If one wants to train an improved version of the neural network on a different dataset for a specific purpose, learning takes use of pre-trained neural network weights.

The base architecture is known as EfficientNet-B4, and it's a convolutional neural network (CNN) having a simple yet powerful design. Properly tuning the model's parameters enhances the extraction of information. Adding an attention mechanism enhances its ability to identify minute irregularities related to deepfake manipulations. EfficientNetAutoAttB4's attention mechanism was developed specifically for the B4 architecture and is placed strategically across the network. This approach [4] allows the model to selectively amplify regions which are essential to identifying modified data by constantly altering the significance of spatial factors in video frames. The model's focus on significant details is enhanced by this attention-augmented technique.

The combination of a new attention mechanism integration with the mathematical ideas discovered in EfficientNet layouts is what makes EfficientNetAutoAttB4 so strong. The model's [5] attention, depth, and architectural decisions all work collectively to show visual patterns indicative of deepfake manipulations. The model's excellent navigation of altered video frames contributes to its effectiveness in deepfake recognition. The attention method makes it possible for EfficientNetAutoAttB4 to examine areas that are susceptible to digital manipulation in a specific way, increasing accuracy and consolidating its position as a reliable safeguard against emerging deepfake threats.

The 408 actual and 795 synthetic movies in the Celeb-DF dataset were produced with an altered version of the Deep-Fake generation technique. The videos feature a frame rate of thirty frames per second and average 13 seconds in length. Compared to the prior datasets, which contained high resolution videos featuring lots of visual imperfections that made it difficult to identify deepfakes, the generated videos have lesser visual imperfections and are therefore of higher quality. The challenge is made more difficult by the lower quality deepfakes in this dataset [6].

The Model has been trained with all the datasets available mainly, DFDC :[2], Celeb DF: [7] and FaceForensics++ : [8] to achieve a good amount of generalization. As mentioned in [9], EfficientNet models are very performant at deep fake detection, and many of the models used in the DFDC challenge make use of the EfficientNet models. It also shows the low correlation between the size of the model and its performance.

2. Literature Survey

Many video forensics techniques have been developed subsequently for an array of reasons. Numerous algorithms have been brought out to identify these forgeries in considering the increasing incidence of facial modification techniques and the potential dangers that they mean.Convolutional Neural Networks (CNNs) are used in specific techniques for evaluating footage frames by pixels. For example, MesoNet is a very basic CNN used for recognizing fake faces. [10]. It has been suggested, nevertheless, that XceptionNet performs better than this network when getting retrained.Other approaches track the chronological progression of video frames via the study of Long Short-Term Memory (LSTM). These techniques utilize a recurrent handle to combine frame-based features which were already obtained.Specific processing traces are the focus of some methods. [11] A certain method, for instance, takes into account the fact that deepfake donor faces have been enlarged to realistically fit onto the host video, and proposes a detector that keeps track of these indications of warping. [12]

Some techniques use semantic analysis of the frames to gain insight past the disadvantages of pixel analysis. While one technique focuses on irregular lighting effects, another learns to distinguish between actual and fake head positions. Another technique is based on the study of eye flashes, because the first set of deepfake films showed specific eye artifacts that could be detected utilizing this technique. [13] However, these semantic methods lose their effectiveness when manipulation techniques become more realistic. Lastly, certain methods offer additional information regarding localization. One multi-task learning technique provides both a segmentation mask and a detection score. An attention mechanism has been suggested by a different approach [14].

EfficientNetV2, a new family of convolutional networks, is more rapid and parameter-efficient than previous versions of it. The models were built using training-aware neural architecture search and scaling, which simultaneously improves training speed and parameter efficiency. The models were examined further using a search space that was extended with the inclusion of new operations such as Fused-MBConv. [15] EfficientNetV2 models can be up to 6.8% smaller and train far more quickly than state-of-the-art models. The authors propose a better progressive learning approach that, in order to speed up training further, adaptively modifies image size and regularization (such as data augmentation). When applied via progressive learning to the ImageNet and CIFAR/Cars/Flowers datasets, EfficientNetV2 performs significantly better than previous versions. Since EfficientNetV2 pre-trains on the same ImageNet21k, it obtains 87.3% top-1 accuracy on ImageNet ILSVRC2012, exceeding the accuracy of the most recent ViT [16].

Deepfakes, an innovative manipulation technique, enables anybody to effortlessly switch between two identities in a single video. The DeepFake Detection Challenge (DFDC) Kaggle competition began as a reaction to this new threat, and an enormous face swap video dataset has been generated to help in the training of detection models. The largest face swap video dataset that is available to the public is the DFDC dataset. It consists of over 100,000 clips that have been collected from 3,426 selected actors and produced using various Deepfake, GAN-based, and non-learned gets closer techniques. [17] Although Deepfake detection is a very difficult and unsolved topic, a Deepfake detection model trained just on the DFDC can generalize to real "in-the-wild" Deepfake films. As such, a model can be a useful analysis tool when investigating potentially Deepfaked videos [2].

A powerful face detector designed specifically for mobile GPU applications is called BlazeFace. It may display between 200 and 1,000 images per second on high-end devices. Its rapid speed makes it ideal for use in any augmented reality application where a particular facial region has to be provided as input for an individual model. These models could be used for facial segmentation, facial characteristic or recognition of emotions, or the computation of 2D/3D facial keypoints or geometry. [18] The growing number of deepfake videos has led to the development of

reliable detection systems that can warn viewers to the potential falsehood of such content on the internet and social media. The capacity to manipulate videos and swap faces is getting better every day because of to computing applications, software, and smartphone apps; nevertheless, automated systems' capacity to identify face counterfeits in videos is still somewhat unreliable and is usually biased toward the dataset that was utilized to create the specific detection system. [19] The authors of the present article examine the impact of different data enhancement and training methods on CNN-based deepfake sensors, both within and between datasets [20].

3. Methodologies

The method of study relies on the concept of ensembling, which usually increases the precision of predictions. Learning an extensive amount of CNN-based classifiers to extract different types of high-level semantic data that support one another and improve the ensemble as a whole is the objective. The EfficientNet family of models, a groundbreaking approach for automatically scaling CNNs, is the first stage in the training procedure. These models beat other innovative CNNs in terms of precision and effectiveness while still adhering to the hardware and time limitations set by DFDC. Two enhancements for an EfficientNet layout have been proposed by the researchers. Add an attention mechanism first to help analysts in finding the most informative part of the video for classification. Furthermore, it examines how incorporating siamese methods of training into the learning process may produce new data information [21].

The figure represents how the attention layer is added to the EfficientNetB4 and CNN layers. [12] the strategy relies on the concept of ensembling, which often boosts accuracy of predictions. The objective is to train many CNN-based classifiers to gather distinct kinds of high-level semantic data that improve each other and enhance the ensemble. The researchers began with the new technique for automatically scaling CNNs, called the EfficientNet family of models. These models meet the hardware and time restrictions established by DFDC and are more accurate and efficient than other state-of-the-art CNNs. The researchers suggest two improvements for an EfficientNet architecture. First, as shown in figure 1, it includes an attention mechanism to help analysts identify the most informative portion of the video for classification. Following that, it explores the incorporation of Siamese training methodologies into the educational process to extract additional data.

3.1. Attention Mechanism

The researchers picked the EfficientNetB4 model from the EfficientNet family for the purpose of this research. Utilizing the ImageNet dataset, EfficientNetB4, with 19 million parameters and 4.2 billion FLOPS, achieves 83.8% top-1 accuracy. [22] In contrast, the baseline face modification detection system, XceptionNet, requires 23 million parameters and 8.4 billion FLOPS to reach 79% top-1 accuracy. A square color image—more particularly, a face cut out of a video frame—is the input for the network. The accuracy of classification is enhanced by tracking facial data instead of using the whole frame as input. Any readily available face detector is able to simply extract faces from frames. A feature vector of 1792 elements, designated as f(I), reflects the network's output. The result of a classification layer is the face's the ultimate score.

We implement an attention mechanism similar to the self-attention mechanisms and EfficientNet's methods outlined in [23]. This is how it works:

- The feature maps from the EfficientNetB4 up to one specific layer have been selected by the researchers. This layer was chosen to provide the features sufficient input frame information without going excessive in terms of delicacy or detail. Particularly, the output features at the third MBConv block—which possess dimensions that range from 28×28×56—are selected by the researchers.
- The single convolutional layer that the researchers use to process these feature maps has a kernel size of 1. To get a single attention map, the researchers next apply a Sigmoid activation function.
- Each feature map at the selected layer has been enlarged by this attention map by the researchers. This aids the network in focusing on the most relevant parts of the input.

This simple approach not only provides the network the ability to focus solely on the most important parts of the feature maps, but it also gives us an improved awareness of the parts of the input that the network considers to be the most instructional. Furthermore, the data gathered attention map is easily mapped to the input sample, showing the elements of it that the network considered more significant. The remaining layers of EfficientNetB4 process the attention block result in the end. The authors of the study call the final network EfficientNetB4AutoAttB4, and the entire training process could be executed end-to-end.

3.2. Network Training

The researchers use two different training methods for each model:

1) End-To-End
2) Siamese

The first is a standard method of training that is utilized as well in the DFDC competition as metrics for assessment. The latter takes utilize the network's capacity for generalisation to generate a descriptor of features that highlights similarities between samples

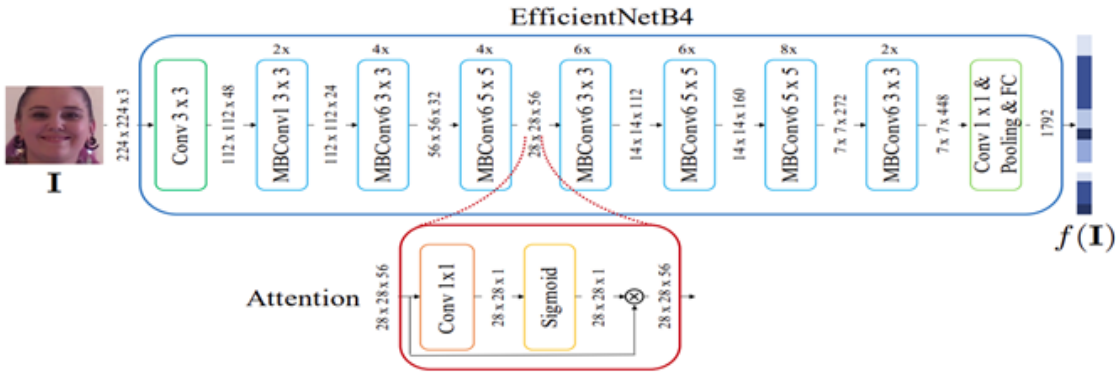

Fig 1: Architecture of EffieientNetAutoAttB4

that fall into the same class. Learning an image in the network's encoding space that successfully separates samples—that is, faces—into true and false classes is the ultimate goal.

End-To-End training: The network generates a picture-related score y after the researchers input it a sample face. Note that a Sigmoid activation isn't utilized to pass this score. The commonly utilized LogLoss function drives the weights upgrading.

$$\boldsymbol{L_L} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(\mathrm{S}(\hat{y}\mathrm{i})) + (1-y_i)\log(1-\mathrm{S}(\hat{y}\mathrm{i}))) \quad (1)$$

where yi denotes the corresponding face label and yi $\{0, 1\}$ the i-th face score. Particularly, faces from real, flawless videos have been assigned to label 0 whereas those from counterfeit videos are connected to label 1. A total of N faces was utilized in the training manage, and the Sigmoid function is denoted by S $(\cdot)$.

Siamese training: The researchers use the triplet margin loss, which was motivated by artificial intelligence works that use CNNs to construct local feature descriptors. The triplet of margins losses is defined in the following manner, keeping in consideration that f(I) is the dynamic coding that the network obtained for a source of data face I, where k·k2 is the L2 norm:

$$\boldsymbol{LT} = max(0, \mu + \delta+ - \delta-) \quad (2)$$

where,

$$\delta+ = kf(Ia) - f(Ip)k2 \quad (3)$$

$$\delta- = kf(Ia) - f(In)k2 \quad (4)$$

and

A margin of μ is merely positive.

- In this instance:
- The anchor sample $(I(a))$ is a real face;
- The positive sample $(I(p))$ is an alternate actual face that corresponds to the exact same category as $I(a)$;
- The negative sample $(I(n))$ is an unreal face that corresponds to another group than Ia.

Take into account the following networks in this experiment:

- Since XceptionNet is the most efficient model presently in use, it makes it logical to use it as the baseline for experimental campaigns. [24]
- EfficentNetB4 outperforms other current methods in terms of accuracy and efficiency. [25].
- EfficentNetB4Att, which is intended for separating out relevant facial sample segments from the irrelevant ones [26].

For each dataset, the researchers use a different data split method. The researchers use the initial 35 folders for learning, the 36–40 folders for confirmation, and the final 10 files to evaluate while adopting DFDC, according to its folder structure. The researchers use a similar split for FF++, choosing 140 films of the initial YouTube loops for testing, validation, and training, and 720 videos for training. The matching bogus videos are categorized under the same division. Each result depends on the information that the examination groups. Researchers only utilize a particular amount of pixels from every video in this experiment. This is due to the fact that overfitting might result from using too few frames, yet performance isn't actually improved through using more frames. The researchers discovered that overfitting may be avoided without raising validation loss by using 32 frames each video. Keeping into account the limitations of the DFDC challenge, the investigators use the same limit for both the training and testing stages. For FF++, this produces roughly 1.6 million pictures, and for DFDC, it yields 3.4 million frames.

Because not all pixel details are helpful for deepfake proof of identity, the researchers further focus on the area where the subject's face is placed, which helps to limit the quantity of data analyzed. The BlazeFace extractor was quicker compared to the MTCNN detector that the researchers used for obtaining features from each frame. The face with the highest rating of trust is retained by the researchers if multiple faces are found. A 224 × 224 pixel squared color image is the result that the networks use.

The researchers use a distinct split strategy for every set of data. The researchers divided DFDC based on folder structure, utilizing the initial 35 files for instruction, each of the folders spanning 36 to 40 for confirmation, and the final 10 folders for testing. Researchers divided the collection of initial sequences they obtained from YouTube using FF++ into 140 videos for validation, 140 videos for testing, and 720p videos for training. The same phony videos are in the exact same category. Each result is shown on the test sets.

In this study, researchers only account for a specific number of frames from every movie. Two primary factors affect the choice to do this during the training phase:

i. There is no rational way to increase the total number of frames with an objective to enhance efficiency; and
ii. There is a strong propensity to overfit when using an incredibly tiny number of frames per clip. By displaying both validation and training loss as an average of the training cycles with an array of frames per video, it illustrates this behaviour.

It is noteworthy that choosing 15 frames instead of 32 for each video does not result in a lower minimum validation loss; on one together, selecting 32 pixels per clip assists in avoiding overfitting. When doing testing, the researchers should consider the equipment and time limitations given by the DFDC challenge.

From this angle, researchers could further minimize the volume of data handled by the networks by considering that not every frame data is needed for the fake detection process. In fact, most study efforts are concentrated on a certain area, like the subject's face. Therefore, as a preprocessing step, extracting the human faces of the scene's subjects form each frame using the BlazeFace extractor. According to this study, the extractor is quicker than the authors' MTCNN detector. When several people are identified, researchers retain the face with the greatest confidence score. The generated color image with 224 x 224 pixels squared serves as the input for the networks.

Techniques for appending data to the source features in order to improve the model during validation and training phases. Applying hue saturation, noise, arbitrary brightness contrast, horizontally flipping it, and random downscaling in particular, and then shrinking the resultant JPEG file. Specifically, using Albumentation as a data-augmentation library and Pytorch as a system for deep learning.

As can be observed in figure 2, data augmentation approaches are used by the researchers to boost the resilience of the models during training and validation on the input faces. JPEG compression, hue saturation, random brightness contrast, downscaling, horizontal flipping, and noise addition are some of these techniques. use of Pytorch as a deep learning framework and Albumentation as a data augmentation library. The Adam optimizer is employed to train the models using specified hyperparameters.

The magnitude of the datasets prevents the researchers from retraining the networks for an entire epoch. The process is as follows:

- Train until the loss of validation reaches a plateau, which can take up to 20,000 iterations. Analyzing an entire set of 32 faces—16 real and 16 fake—randomly and evenly selected from all of the training split's films is referred to as an iteration. Every 500 training iterations, validation

will be carried out utilizing 6000 randomly selected and evenly distributed samples from every video in the validation set. The researchers cut the initial rate of learning by a factor of 0.1 if, after 10 validation processes (5000 training cycles), the loss of validation does not decrease.

- The validation process, learning rate timetable, and number of iterations used in the end-to-end training are also used to train the feature extractor. The batch's composition and the loss function that was used make a difference. The batch in this particular case consists of 12 triplet samples (6 real-fake-fake and 6 fake-fake-real), selected from every one of the set's videos.

3.3. Data Augmentation

The validation process, learning rate timetable, and number of iterations used in the end-to-end training are also used to train the feature extractor. The batch's composition and the loss function that was used make a difference. The batch in this case consists of 12 triplet samples (6 real-fake-fake and 6 fake-fake-real), selected from every one of the set's videos. The researchers emphasize on a subset of the multiple data augmentation techniques that may be utilized to represent what alterations a face could encounter in the outdoors. The following changes are taken into consideration:

- HF: Horizontal Flip
- BC: Brightness and Contrast changes
- HSV: Hue, Saturation and Value changes
- ISO: Addition of ISO noise
- GAUS: Addition of gaussian noise
- DS: Downscaling with a factor between 0.7 and 0.9
- JPEG: JPEG compression with a random quality factor between 50 and 99

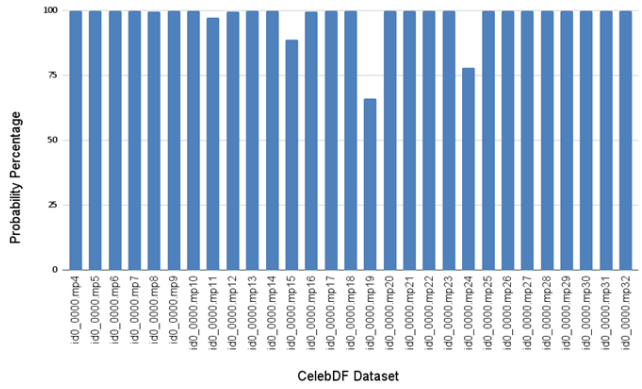

Fig 3: Probability on CelebDF dataset

By training with BCE on the CelebDF dataset, the researchers independently test the aforementioned augmentations. Every experiment that is recommended will be carried out utilizing the Albumentations framework. Figure 3 shows the results on the CelebDF test set,such that probability near 1 is representing fake videos. In light of these results, two interesting inferences might be drawn.

At first, it seems like augmentations have not much impact on enhancing intra-dataset detection. This could be because train, validation, and test groups encompass distinct video settings and scenarios. The HF enhancement is the only exception, providing an area under the curve gain of only 0.7%.

Second, in terms of cross-dataset generalization, some augmentations are beneficial (sometimes significantly). Particularly, networks trained on both CelebDF and DFD show a rise in AUC when exposed to HF, BC, HSV, and JPEG [27].

Perhaps because DFDC displays environments that are much different from those in DF, augmentations don't seem to help DF as much as they do DFDC. The latter, either inside a TV studio or in a conversation with studio-level lighting, contains almost just a single player in the center of the conduct, compared to the former, which features distant players traveling around the scene—typically two actors. The researchers create an outcomes-based data augmentation pipeline. The CNN is then rebuilt by the researchers using triplet loss in addition to BCE. The BCE loss results are available. The fusion of augmentations can lead to significant increases in cross-dataset detection AUC, with gains of up to +9% when testing on CelebDF and training on DFD, respectively. In contrast, there is little to have an effect on the intra-dataset detection performances. Table II illustrates how the small beneficial effects of triplet loss when training on the entire dataset are not as evident when augmentations are provided to the CNN trained with triplet loss. [13] As compared to BCE loss with data augmentation, triplet loss with augmentations possesses a lower AUC for almost all combinations.

A different perspective on the differences between BCE and triple losses by training EfficientNetB4 using subsampled training data in data-limited environments can be noticed. The loss of triplets assists with intra-dataset detection (DFDC, CelebDF, and DF) in this circumstance, as well as in detection among datasets and beats BCE [23].

Figure 4 illustrates how the model's efficacy is broken out

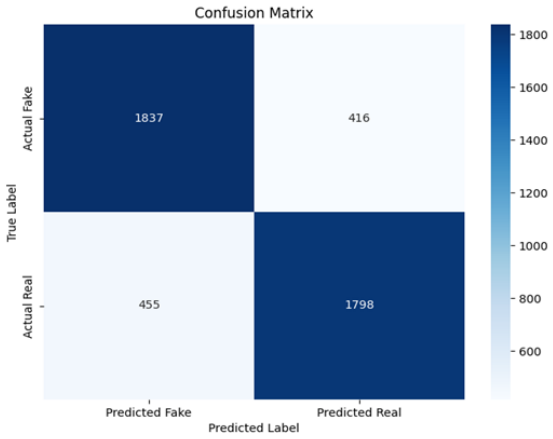

Fig 4: Confusion Matrix for EfficientNetAutoAttB4

significantly in the confusion matrix, with a high number of accurate classifications for both authentic and fake videos. The low percentage of false positives or false negatives highlights the model's dependability even further.

Figure 5 illustrates the framework's 0.805 F1 score, which

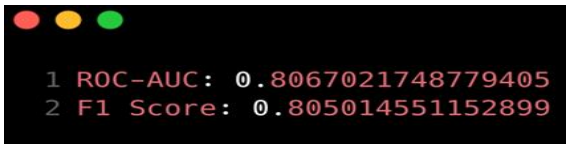

Fig 5: F1 Score and ROC Score for EfficientNetAutoAttB4

emphasizes how storage and accuracy are aligned. It shows a remarkable capacity to accurately distinguish between authentic and phony videos while minimizing false positives and erroneous negatives.

The model's 0.805 F1, as shown in figure 5, score highlights the way memory and precision are matched. It demonstrates an excellent ability of reducing false positives and false negatives

while precisely recognising real and fake videos.

As seen in figure 6, the model's excellent capacity to discriminate between genuine and fraudulent videos is exhibited by its ROC

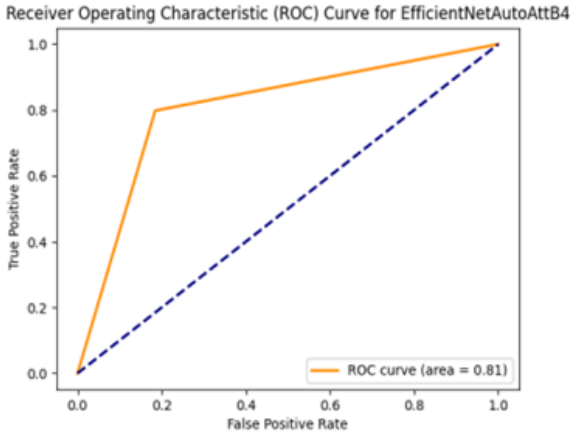

Fig 6: ROC Plot for EfficientNetAutoAttB4

value = 0.806, and its AUC of 0.81 further confirms the model's overall efficacy. These metrics show how well the model reduces false positive rates while detecting true positive rates.

4. Results

In order to show the extent to which the attention mechanism extracts the most instructive content from faces, The researchers analyze the attention map that was generated utilizing a few FF++ faces.

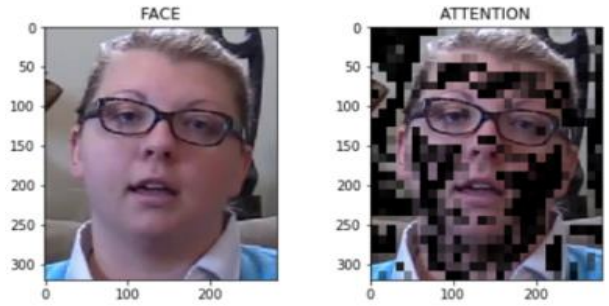

Fig 7: Image Processing

a) Real Image Input **b)** Image of Attention Layer

With the help, the researchers select the 28 x 28 2D map that is the output of the Sigmoid layer in the attention block. The researchers then overlay this on the input face after upscaling it to the input face size of 224 × 224. The results are displayed in figure 7 & in figure 8. It is notable that the most complex characteristics of faces, such as the lips, nose, ears, and eyes, may be emphasized via this basic attention process. Conversely, regions that are level or have relatively small gradients lack data related to the network.

In fact, it has been repeatedly demonstrated that facial features account for the majority of artifacts from deepfake subsequent generations methods. For example, the most distinctive features of these methods are the coarsely sketched eyes & teeth with excessively white areas. When training the network in a siamese fashion, the researchers calculated an image projection over a tiny region using the commonly used technique to see if the features generated by the net's coding are discriminating for the job. It is apparent when frames from identical videos group together into little sub-regions. Furthermore, all of the real samples are arranged in the upper portion of the chart, whereas the fake samples are shown in the lower portion. The frames of the same videos group together to form subregions which are easier to travel. The frames from the same videos group together into subregions which are simpler to traverse. This illustrates the decision to use both datasets.

The final output of two separate films created with the EfficientNet-B4 structure is shown in figures 7 and 8, where 7.a

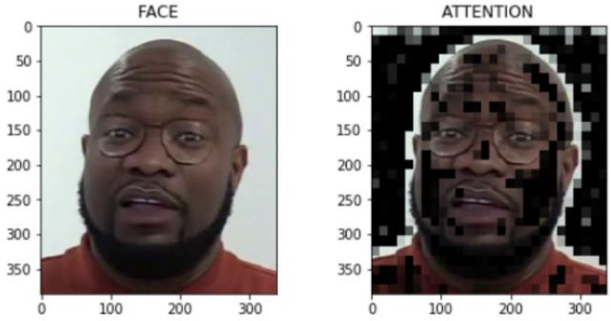

Fig 8: Image Processing

a) Fake Image **b)** Image of Attention Layer

shows what was taken from the actual video and 8.a shows the image taken from the fake video.

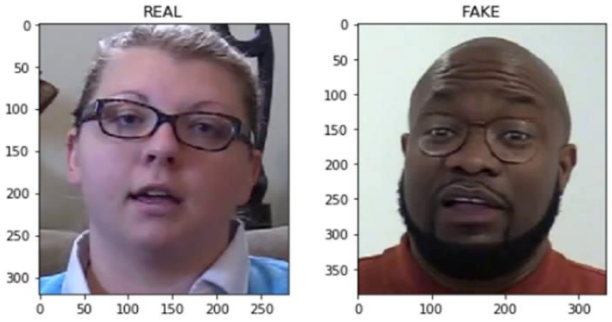

Fig 9: Classification of image as real and fake

The face extracted from the video is passed through the attention layer for getting better accuracy in the result as seen in 7.b and 8.b. As shown in figure 9, if the particular video has a score near to '0' the video will be real. Score near to '1' will show that the video is fake. The score is determined by taking into account the combination of EfficientNet-B4 architecture and the attention layer.

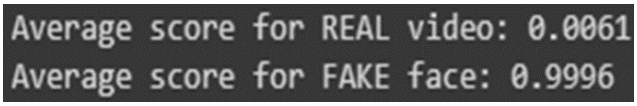

Fig 10: Final Score

In figure 10, the final result is displayed by classifying the image captured by the best frame present in the video as real and fake.

Figure 11 is the graph of the real and fake video respectively. Real video means lines will be closer to the x axis. Every line is depicting a different feature of the video frame. And if the line is closer to 1 meaning the video is fake.

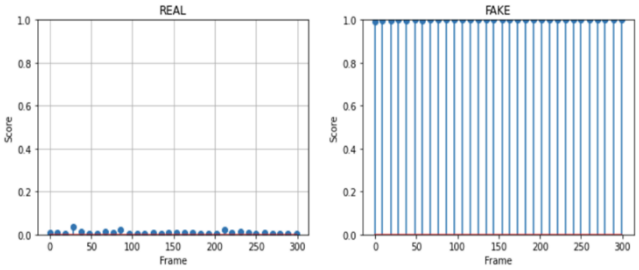

Fig 11: Graph of the final result

5. Conclusion

Detecting manipulated content in videos is crucial today due to the

widespread use of videos in daily lives and mass media. The main focus on detecting facial manipulations in videos, which could be created using traditional computer graphics or deep learning techniques. This method is inspired by EfficientNet models and improves upon a recent solution. Use of an ensemble model are trained with two main strategies:

1. An attention mechanism that helps the model focus on important features and improves its learning capability.
2. A triplet siamese training strategy that extracts deep features from data for better classification performance.

This research method has been tested on two public datasets containing nearly 120,000 videos. The results show that this ensemble strategy is effective (Accuracy of 94%) for detecting facial manipulations. In the future, the plan is to incorporate temporal information into this model. Analyzing more frames at once using intelligent voting schemes could potentially increase the accuracy of the model.

Acknowledgements

We extend our heartfelt appreciation to Cyber Peace Foundation for their generous and wholehearted funding of our research project. Their commitment to advancing knowledge and innovation in the field has been instrumental in the successful execution of our work. With their financial support, we have been able to delve deeper into our research objectives, pushing the boundaries of understanding and contributing to the broader academic and practical discourse. This collaboration exemplifies their dedication to fostering excellence in research, and we are immensely grateful for their significant contribution to our endeavors.

6. References

[1] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting World Leaders Against Deep Fakes."

[2] B. Dolhansky et al., "The DeepFake Detection Challenge (DFDC) Dataset," Jun. 2020, doi: https://doi.org/10.48550/arXiv.2006.07397.

[3] S. Verma, "Classification of Spoofing Attack Detection using Deep Learning Algorithms MSc Research Project Data Analytics."

[4] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014, doi: https://doi.org/10.48550/arXiv.1409.1556.

[5] S. Sakib, M. Tarid, and A. Abid, "Deepfake detection Using Neural Networks," 2021.

[6] S. R. Krishnan and P. Amudha, "International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING Hybrid ResNet-50 and LSTM Approach for Effective Video Anomaly Detection in Intelligent Surveillance Systems." [Online]. Available: www.ijisae.org

[7] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics," Sep. 2019, doi: https://doi.org/10.48550/arXiv.1909.12962.

[8] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in Proceedings of the IEEE International Conference on Computer Vision, 2019. doi: 10.1109/ICCV.2019.00009.

[9] A. A. Pokroy and A. D. Egorov, "EfficientNets for DeepFake Detection: Comparison of Pretrained Models," in Proceedings of the 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2021, Institute of Electrical and Electronics Engineers Inc., Jan. 2021, pp. 598–600. doi: 10.1109/ElConRus51938.2021.9396092.

[10] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2019. doi: 10.1109/ICASSP.2019.8683164.

[11] T. Wang, H. Cheng, K. P. Chow, and L. Nie, "Deep Convolutional Pooling Transformer for Deepfake Detection," Sep. 2022, doi: 10.1145/3588574.

[12] N. Bonettini, L. Bondi, E. D. Cannas, P. Bestagini, S. Mandelli, and S. Tubaro, "Video face manipulation detection through ensemble of CNNs," in Proceedings - International Conference on Pattern Recognition, Institute of Electrical and Electronics Engineers Inc., 2020, pp. 5012–5019. doi: 10.1109/ICPR48806.2021.9412711.

[13] A. Gironi, M. Fontani, T. Bianchi, A. Piva, and M. Barni, "A video forensic technique for detecting frame deletion and insertion," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2014. doi: 10.1109/ICASSP.2014.6854801.

[14] Y. Al-Dhabi and S. Zhang, "Deepfake Video Detection by Combining Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)," in 2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering, CSAIEE 2021, Institute of Electrical and Electronics Engineers Inc., Aug. 2021, pp. 236–241. doi: 10.1109/CSAIEE54046.2021.9543264.

[15] I. Ilhan, E. Bali, and M. Karakose, "An Improved DeepFake Detection Approach with NASNetLarge CNN," in 2022 International Conference on Data Analytics for Business and Industry, ICDABI 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 598–602. doi: 10.1109/ICDABI56818.2022.10041558.

[16] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," Apr. 2021, [Online]. Available: http://arxiv.org/abs/2104.00298

[17] A. Kohli and A. Gupta, "Detecting DeepFake, FaceSwap and Face2Face facial forgeries using frequency CNN," Multimed Tools Appl, vol. 80, no. 12, pp. 18461–18478, May 2021, doi: 10.1007/s11042-020-10420-8.

[18] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, "BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs," Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.05047

[19] A. Seth and A. K. Gogineni, "Detection of Deep-fakes in Videos using CNN and Transformers", doi: 10.13140/RG.2.2.23238.60480.

[20] L. Bondi, E. Daniele Cannas, P. Bestagini, and S. Tubaro, "Training Strategies and Data Augmentations in CNN-based DeepFake Video Detection," in 2020 IEEE International Workshop on Information Forensics and Security, WIFS 2020, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/WIFS49906.2020.9360901.

[21] H. T. Duong, V. T. Le, and V. T. Hoang, "Deep Learning-Based Anomaly Detection in Video Surveillance: A Survey," Sensors, vol. 23, no. 11. 2023. doi: 10.3390/s23115024.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2010. doi: 10.1109/cvpr.2009.5206848.

[23] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain, "On the Detection of Digital Face Manipulation."

[24] S. Ganguly, A. Ganguly, S. Mohiuddin, S. Malakar, and R. Sarkar, "ViXNet: Vision Transformer with Xception Network for deepfakes based video and image forgery detection," Expert Syst Appl, vol. 210, Dec. 2022, doi: 10.1016/j.eswa.2022.118423.

[25] X. Cheng, L. Yuan, Z. Liu, and F. Guo, "Comparative analysis of video anomaly detection algorithms," 2022. doi: 10.1117/12.2641049.

[26] A. Berroukham, K. Housni, M. Lahraichi, and I. Boulfrifi, “Deep learning-based methods for anomaly detection in video surveillance: a review,” Bulletin of Electrical Engineering and Informatics, vol. 12, no. 1, 2023, doi: 10.11591/eei.v12i1.3944.

[27] L. D’Amiano, D. Cozzolino, G. Poggi, and L. Verdoliva, “A PatchMatch-Based Dense-Field Algorithm for Video Copy-Move Detection and Localization,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 3, 2019, doi: 10.1109/TCSVT.2018.2804768.