TruthLens: Deepfake detection using Xception Model

1st Samrudhi Bhosale
EECS
Queen Mary University of London
London, UK
bhosalesamruddhi3535@gmail.com

2nd Ammar Yasir Naich
EECS
Queen Mary University of London
London, UK
a.y.naich@qmul.ac.uk

***Abstract*—In recent years, the proliferation of deepfakes—synthetic media that convincingly mimics real human features—has posed a significant threat to the integrity of digital information. This research presents a novel approach for detecting deepfakes using a fine-tuned Xception model, a convolutional neural network known for its efficiency and accuracy in image classification tasks. The model's effectiveness is further enhanced by preprocessing steps that involve precise face detection, alignment, and cropping, ensuring that the network focuses on relevant features. Our model is trained on the challenging Celeb-DF dataset, known for its high-quality and realistic deepfake manipulations. Through rigorous evaluation, the advanced model demonstrates significant improvements in accuracy, precision, and recall, particularly when compared to a basic model. This is evidenced by metrics such as an Area Under the Curve (AUC) of 0.95 and an average precision (AP) of 0.89, highlighting the model's ability to accurately distinguish between real and manipulated images. The findings suggest that our approach is not only effective but also generalizes well across different types of deepfakes, offering a robust solution for real-world applications. Future work will explore expanding the model's capabilities to detect multimodal deepfakes and improving real-time detection performance to keep pace with the evolving threats posed by deepfake technology**

***Index Terms*—DeepFake Detection, Celeb DF, Convolutional Neural Network, XceptionNet, Dlib, GradCAM, Image forgery detection.**

I. Introduction

Recent widespread popularization of one advanced smart device after another has taken global connectivity to new heights with all these devices coming with high-quality cameras. That said laptops, desktops can also be found in almost every family now. Such units wherever they may be in the world are connected through Internet to remote server. This means that people have more access than before to visual digital information of immeasurable amount and density.

Images and videos are now essential sources of information. Professionals and everyday people alike use smartphones to capture every event, putting powerful tools right at our fingertips. Thanks to advanced digital technologies like sophisticated compression methods, fast network services, and user-friendly apps, we can share visual content incredibly quickly. Beyond that platforms such as Instagram , Facebook , Titok let people put out user generated material does in fact help to accelerate the spread of visual data. Moreover, both premium and free image editing software, accessible on PCs and mobile apps, allow users to modify images easily.

These advancements have led to the widespread distribution of altered or fake visual content, sometimes altered with harmful intentions for political or commercial gain. Major social media networks continue to struggle to filter the flow of manipulated content down to its roots of distribution, especially among the viral segments of vulnerable groups. This problem also spawned a host of legal discussions concerning responsibility for the after-effects of such content.

Humans are often unable to spot visual forgeries because of a cognitive limitation. This highlights the need for advanced digital methods to detect such manipulations. Although various media can be altered, this paper specifically concentrates on identifying forgeries in still images, as they are the most common type of manipulated media.

II. Motivation

The rapid advancement of deepfake technology is making it increasingly difficult to tell real visual content from fake, which is becoming a serious concern. As these manipulated videos spread widely across online platforms, especially social media, the risk of spreading misinformation and causing social disruption is growing. This widespread use of deepfakes is eroding trust in digital media, making it essential to address this problem quickly. My research is motivated by the need to develop an effective method for detecting deepfakes by using advanced deep learning techniques like convolutional neural networks (CNNs) and convolutional vision transformers (CVTs). By integrating these powerful technologies, I aim to help organizations improve their ability to identify and reduce the risks posed by deepfake content. Investing in such detection systems is crucial to keeping up with the fast-changing digital world, particularly in areas like media, politics, and finance, where the impact of deepfakes can be most harmful.

III. Literature Review

As part of this research, an in-depth review of contemporary literature on image forgery detection was conducted to understand the evolution of detection techniques and to know the current challenges and future directions in this domain.

Among the notable contributions to the field of deepfake detection are the works of Kaddar et al. [1] and Das et al. [2], which offer comprehensive overviews of the methodologies used in detecting deepfake videos. These studies are particularly significant for their detailed examination of deep learning techniques, which have become pivotal in the development of advanced deepfake detection strategies.

The architecture introduced in this study, known as HCiT, embodies a hybrid approach that combines the strengths of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for detecting deepfakes. Kaddar et al. [1] highlight how HCiT leverages CNNs' ability to extract detailed local features from video frames, integrating this with ViTs' self-attention mechanism, which excels in capturing global patterns across the entire input sequence. The architecture comprises three main components: face cropping, feature extraction using a CNN backbone, and binary classification through a ViT model. Specifically, the CNN, which uses the Xception network, processes cropped face regions to produce feature maps that are subsequently analyzed by the ViT. The self-attention layers in the ViT concentrate on the most relevant features, enhancing the model's accuracy in detecting deepfakes. According to Kaddar et al. [1], this architecture shows significant improvements in both generalization and accuracy across various deepfake generation techniques compared to existing models.

In conclusion, Kaddar et al. [1] state that the HCiT model not only surpasses current state-of-the-art deepfake detection methods on essential benchmark datasets, such as Faceforensics++ and the DeepFake Detection Challenge preview, but also exhibits superior generalization capabilities. Particularly in cross-dataset evaluations, HCiT efficiently manages previously unseen types of deepfake manipulations. This research emphasizes that the hybrid CNN-ViT architecture offers a robust and promising solution to the escalating challenge of deepfake detection by merging the detailed local feature extraction of CNNs with the extensive global attention provided by transformers.

Das et al. [2] offers an in-depth exploration of the current strategies used to detect deepfake videos, with a particular emphasis on deep learning methods such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), alongside hybrid models that combine these approaches. The paper organizes detection techniques according to the technology they utilize, including visual, temporal, and deep feature-based methods, and provides a thorough comparison of their effectiveness.

The survey highlights that CNN-based models like ResNet, VGG16, and EfficientNet are commonly used to extract spatial features from video frames, while RNN models, such as Long Short-Term Memory (LSTM) networks, are employed to capture the temporal relationships between frames. Additionally, Das et al. [2] examine hybrid models that combine CNNs and RNNs to tackle issues such as temporal inconsistency and resolution variation. Furthermore, the paper discusses how Support Vector Machines (SVM) can boost classification accuracy when paired with CNNs, noting that SVM-based methods often surpass purely neural network-based models in terms of both speed and precision.

In summary, Das et al. [2] emphasize the critical need for effective and reliable deepfake detection systems, especially given the risks associated with the misuse of deepfake technology. The paper identifies CNN-SVM hybrid models as particularly promising, offering a good balance between computational efficiency and detection accuracy. It is also noted that as deep learning techniques continue to advance, there remains a significant need for models capable of generalising well across different datasets and types of deepfake alterations.

IV. Dataset

One of the significant challenges in the field of deepfake detection is the reliance on datasets that do not adequately represent the complexity of real-world deepfake videos. Many existing datasets, such as UADFV dataset [3] and DeepFake -TIMIT dataset(DF-TIMIT) [4], suffer from issues like low visual quality, evident splicing boundaries, color mismatches, and visible parts of the original face, which make the forged content relatively easy to detect, See Fig.1 [5]. These visual artifacts do not accurately reflect the more sophisticated manipulations found in deepfake videos that are widely circulated on the internet. Consequently, models trained on these datasets may perform well in controlled environments but struggle to generalize when applied to more realistic, high-quality forgeries [5] .

Despite these limitations, much of the existing research has relied on these older datasets, as highlighted in the survey by Das et al. [2]. Datasets like FaceForensics++ [6] and FaceBook Deep- Fake detection challenge (DFDC) dataset [7] have been widely used due to their early availability and the scale of data they provide. However, with the rapid advancement in deepfake generation techniques, there is an increasing need for more challenging datasets to develop detection methods that are robust and reliable in real-world applications.

In light of these shortcomings, the CelebDF [5] dataset was chosen for this research because it offers a more realistic and challenging set of deepfake videos. Unlike previous datasets, CelebDF [5] addresses many common issues by providing high-resolution videos with significantly fewer visual artifacts. The dataset's focus on subtle manipulations that closely mimic real human expressions and actions makes it particularly suitable for developing and testing advanced deepfake detection techniques .

CelebDF [5] is a large-scale dataset designed to overcome the limitations of earlier deepfake datasets, See Fig.2 [5]. It includes 590 real videos and 5,639 deepfake videos, with an average length of about 13 seconds per video at a standard frame rate of 30 frames per second. These videos are sourced from publicly available YouTube clips featuring 59 celebrities, ensuring a diverse representation in terms of gender, age, and ethnicity. The deepfake videos in CelebDF are generated using an improved synthesis method that enhances visual quality and creates more realistic facial expressions. For this research,

Fig. 1. Visual artifacts of DeepFake video frames in existing datasets that includes low-quality synthesized faces

due to constraints in time and computational resources, I have utilized 10% of the total CelebDF dataset. This subset was further divided into training, testing, and validation sets, with 70% allocated for training, 15% for validation, and 15% for testing. The selection includes both real and fake videos to develop models that can effectively distinguish between authentic and manipulated content. The realistic and challenging nature of the forgeries in CelebDF ensures a robust testing ground, allowing for a thorough evaluation of the deepfake detection models in conditions that closely mirror real-world scenarios.

V. Methodology

The architecture implemented in the code utilizes the Xception model, a deep convolutional neural network known for its efficiency and accuracy in image classification tasks, particularly when fine-tuned for specific datasets. This architecture is adapted for binary classification targeting the detection of deepfake images.

The first step in the architecture involves data preprocessing and augmentation, which is critical for improving the generalization ability of the model. Images from the dataset are loaded and processed using the PyTorch ImageFolderclass, which organizes the images based on their respective labels.

Fig. 2. Frames from the Celeb-DF dataset. Left column is the frame of real videos and right five columns are corresponding DeepFake frames generated using different donor subject

The images are resized to 299x299 pixels to match the input size required by the Xception model. Various transformations are then applied to the images, including random horizontal flips, random rotations, and color jittering (altering brightness, contrast, saturation, and hue). These transformations are designed to introduce variability in the training data, helping the model become more robust to different image conditions. Finally, the images are converted to tensors and normalized to ensure that the pixel values are centered around zero, which facilitates faster convergence during training.

The core of the architecture is the Xception model, which is trained on ImageNet dataset. The Xception model is chosen for its depth and the efficient use of depthwise separable convolutions, which allow it to capture complex patterns in images while maintaining computational efficiency. The pretrained version of the Xception model is loaded, leveraging the knowledge it has already acquired from extensive training on ImageNet dataset. The final fully connected layer of the Xception model is modified to suit the binary classification task. Specifically, the last layer is replaced with a fully connected layer that outputs a single value, followed by a sigmoid activation function. This setup ensures that the model outputs a probability value between 0 and 1, representing the likelihood of an image being a deepfake.

The training process is carried out using the Adam optimizer, known for its ability to handle sparse gradients and maintain a stable learning rate throughout the training. The loss function used is Binary Cross-Entropy Loss (BCELoss), which is appropriate for binary classification tasks as it measures the distance between the predicted probabilities and the actual labels. The model is trained over a large number of epochs, with each epoch consisting of several iterations where the model processes a batch of images, computes the loss, and updates its weights accordingly. During each iteration, the model's predictions are compared against the actual labels

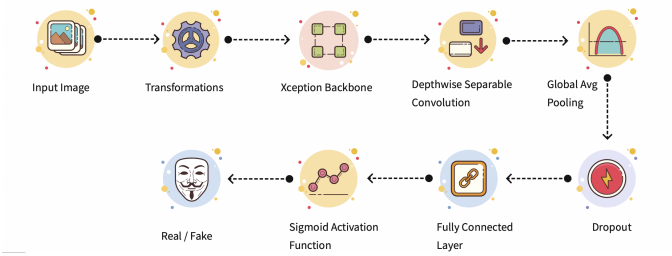

Fig. 3. Architecture diagram of the basic model

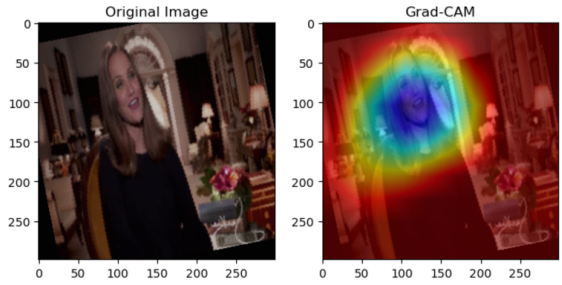

Fig. 4. Heatmap generated using Grad-CAM, illustrating the regions of the image that the model focuses on during prediction

to calculate the loss. The optimizer then adjusts the model's parameters to minimize this loss.

The model's performance is evaluated on a separate test dataset, which was not used during training to ensure an unbiased assessment of the model's generalization capability. During evaluation, the model is set to evaluation mode, and no gradients are calculated to reduce memory usage and computational overhead. The model processes the test images and outputs predicted probabilities, which are then rounded to produce binary class predictions. These predictions are compared against the true labels to calculate the overall accuracy of the model.

In this study, we used a technique called Grad-CAM Gradient-weighted Class Activation Mapping [8]. to visually interpret the focus areas of our convolutional neural network (CNN) on the test dataset. Grad-CAM allows for the generation of heatmaps that highlight the regions within an image that the model considers most influential for its predictions. Upon applying Grad-CAM to our trained model, it was observed that the model was not exclusively concentrating on facial features but was also influenced by background elements present in the images, See Fig. 4. This was likely due to the fact that the training images were randomly cropped, so they sometimes included parts of the background along with the face. Such distractions potentially diluted the model's ability to accurately capture and learn discriminative facial features . To address this issue and enhance the model's performance, particularly its accuracy, I have proposed the development of a refined model architecture. This new model is designed to more effectively isolate and focus on facial regions, thereby improving the accuracy and reliability of the classification process.

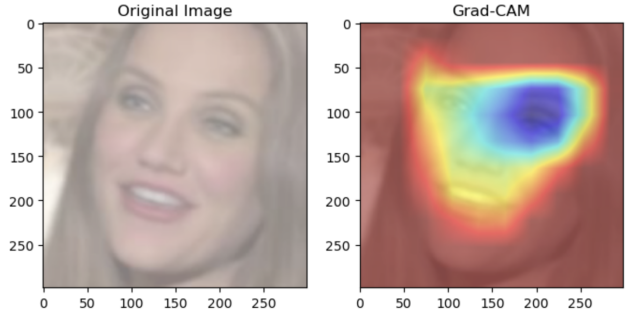

Fig. 5. The original image and Grad-CAM heatmap with a tilted face before applying corrective rotation, highlighting the misalignment of facial features.

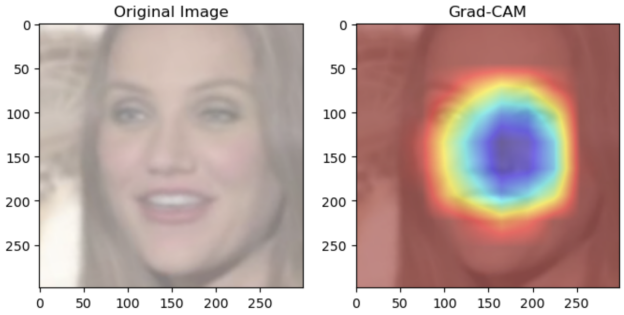

Fig. 6. The image and Grad-CAM after applying corrective rotation, ensuring consistent orientation for accurate feature extraction.

The first step involved the creation of data loaders for the training, validation, and testing datasets. A crucial part of the pre-processing pipeline involves accurately detecting and cropping faces in the images to make sure the model can focus on the relevant facial features. This step begins with turning each image into grayscale, which simplifies the data and reduces the computational complexity, while keeping the key details needed to recognise facial features.

We utilize dlib, a powerful machine learning library, to detect faces in grayscale images. Dlib's face detection algorithm , which combines a histogram of oriented gradients (HOG) with a linear classifier is highly effective in identifying faces under different lighting and angles. To enhance accuracy, we incorporate dlib's shape predictor, which relies on a pre-trained model. This file contains a trained ensemble of regression trees designed to predict the locations of 68 key facial landmarks, such as the eyes, nose, and jawline.

These facial landmarks are essential for understanding the orientation of the face. By identifying the angle between the eyes, allowing us to apply corrective rotations to align the face horizontally, See Fig. 4. This step is important because it ensures that all faces in our dataset are aligned in the same way, which makes it easier for our model to learn and recognize the essential features of the face, See Fig. 5. The detected face is cropped with a margin to include some background, and the image is resized to the required input size of 299x299 pixels. This pre-processing step is crucial for ensuring that the model receives input images with

standardized dimensions and properly aligned facial features, thereby improving its ability to learn meaningful patterns.

The backbone of our model is built on the Xception network, a highly efficient and accurate deep convolutional neural network (CNN) that is both highly efficient and accurate for classifying images. Xception, which stands for ”Extreme Inception,” leverages depthwise separable convolutions to optimize computational efficiency without sacrificing performance. This model was pre-trained on a large-scale ImageNet dataset to capture a wide range of visual features. In our project, we fine-tuned it to adapt to the specific requirements of our binary classification task, distinguishing between real images and manipulated ones.

The Xception model is composed of multiple convolutional layers, which are the building blocks that extract hierarchical features from the input images. Each convolutional layer applies filters to the image, detecting patterns like edges, textures, and eventually more complex structures as the data passes through deeper layers. To improve the stability and speed of the training process, batch normalization is applied after certain layers in the custom fully connected layer of the model. This step helps the model learn more consistently by keeping the data at each layer balanced and preventing extreme values from disrupting the learning process.The model also includes a Dropout layer with a probability of 0.5 in the custom fully connected layer. This dropout layer helps to prevent overfitting by randomly setting a portion of the neurons to zero during training, which forces the model to learn more robust features.

Additionally, the model incorporates the BatchNorm1d layer after the final linear layer in the custom fully connected layer, followed by the Sigmoid activation function. The BatchNorm1d layer helps stabilize and speed up the training process by normalizing the output of the linear layer, while the Sigmoid activation function introduces non-linearity, enabling the model to output a probability score between 0 and 1. This probability score represents the likelihood that the input image is a deepfake.

The final model architecture involves replacing the original classifier in the Xception network with a custom fully connected layer, specifically designed for our binary classification task. This new layer outputs a single value, which is then passed through a Sigmoid activation function to transform it into a probability score, indicating whether the input image is real or fake.

VI. Key Findings

The analysis of the training and validation loss curves reveals that the model is learning rapidly, as evidenced by the sharp decrease in training loss during the initial epochs. However, this rapid learning appears to come at a cost. While the training loss continues to decline steadily, the validation loss initially decreases but then plateaus and even begins to rise slightly. This divergence between training and validation performance suggests that the model is starting to overfit to the training data, struggling to generalize effectively to new, unseen data, See Fig. 8.

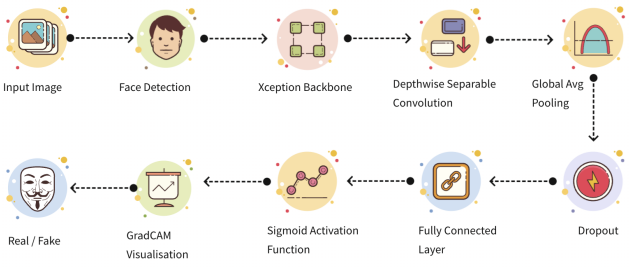

Fig. 7. Architecture diagram of advance model with face detection and cropping

Similarly, the training and validation accuracy curves show a marked difference in performance. The model's accuracy on the training set increases dramatically after the first epoch, quickly surpassing 90%. In contrast, the validation accuracy improves at a much slower rate, indicating that while the model performs well on the data it has seen before, it does not perform as reliably on new data. This growing gap between training and validation accuracy further underscores the overfitting issue, See Fig. 8.

The confusion matrix also gives more detailed information about the performance of the model in classification. The model demonstrates a strong ability to correctly identify real images, with a high number of true negatives. However, it is not as effective in detecting fake images. The model has moderate classification accuracy in identifying fake images and there is also high rate of false negatives whereby fake images are classified as real images. Additionally, the model shows a tendency to be overly cautious, resulting in a higher number of false negatives - real images misclassified as fake, See Fig. 9.

The ROC curve presents a more extensive vision of the model's discriminative ability, with an AUC of 0. 74. While this suggests that the model has a reasonable ability to distinguish between real and fake images, it is clear that there is room for improvement. This is observed because the curve shows that the model is not very sure of the distinctions between the classes, as seen in the relatively gradual slope of the curve, See Fig. 10.

Finally, the precision-recall curve highlights the model's struggle to maintain a balance between precision and recall. The average precision score of 0.53 suggests that the model's precision particularly in identifying fake images is not particularly high. As recall increases, precision drops significantly, indicating that while the model attempts to identify all fake images, it does so at the expense of increasing the number of false positives, See Fig.11 .

In summary, the results indicate that while the model performs well on the training data, it struggles to generalize to new, unseen data, as evidenced by the growing gap between training and validation performance. The model is particularly challenged in accurately identifying fake images, as shown by

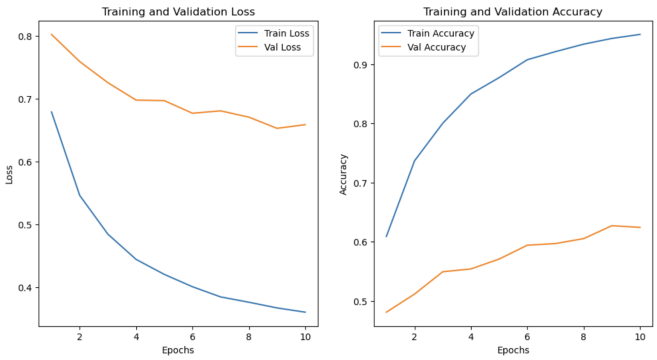

Fig. 8. Accuracy and Loss of the Basic model

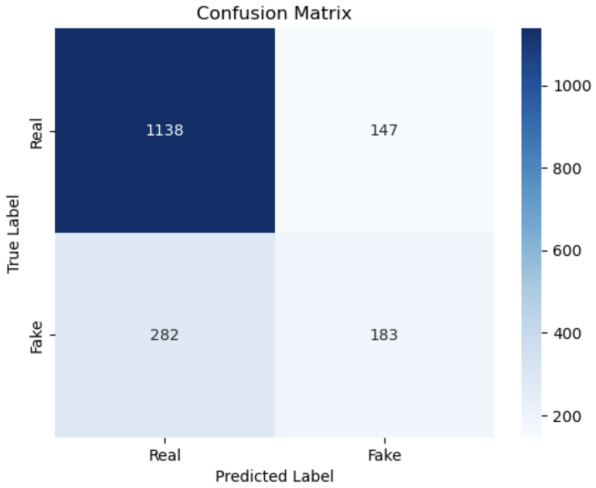

Fig. 9. Confusion Matrix

the confusion matrix and precision recall metrics.

The accuracy and loss graphs provide valuable insights into the behavior and performance of the advance model. The sharp decrease in training loss and the corresponding rapid increase in training accuracy suggest that the model is quickly

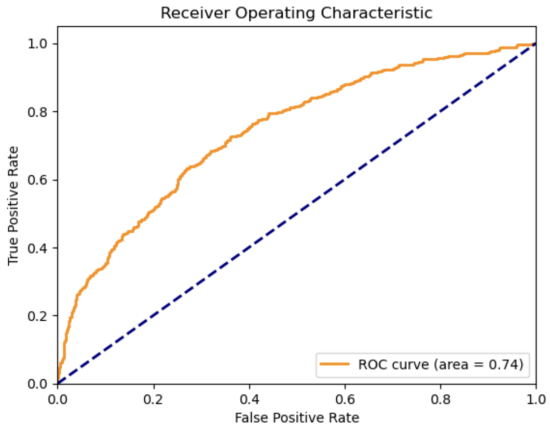

Fig. 10. Basic Architecture ROC Curve

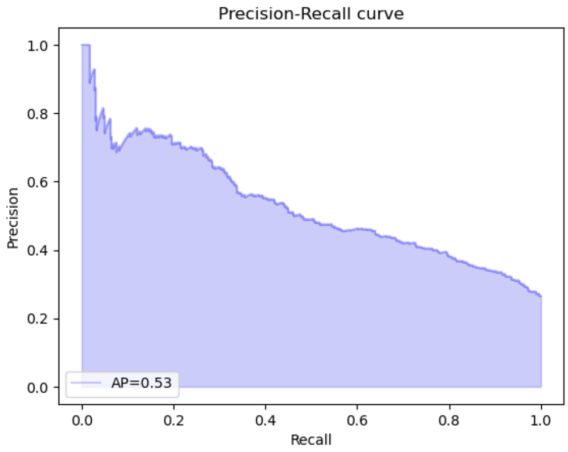

Fig. 11. Basic Architecture Precision-Recall Curve

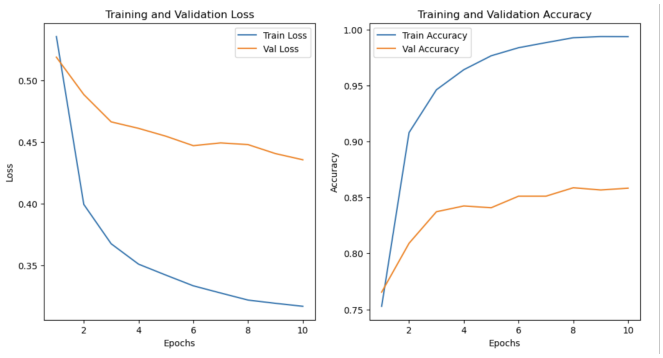

Fig. 12. Accuracy and Loss of the Advance Model

learning to classify the training data correctly. However, the relatively flat validation loss and accuracy curves indicate that the model's ability to generalize to unseen data is limited, See Fig. 12.

In the basic model, as shown in the first set of graphs, the training accuracy rapidly escalates, reaching nearly 100%, while the validation accuracy lags significantly, demonstrating a wide gap that highlights severe overfitting, See Fig. 8. On the other hand, the advanced model shows a more controlled improvement. In the second set of graphs, while the training accuracy still increases, it does so at a slightly more gradual pace, and the gap between training and validation accuracy has been reduced, See Fig. 12. The progress from the basic to the advanced model is evident in how the advanced model handles training and validation. While the basic model learned too quickly and became overly confident on the training data, the advanced model, with its more sophisticated techniques, has started to balance learning and generalization. The improvements are subtle but important, showing that the model is becoming better at understanding the nuances of the data rather than just memorizing it. This progress is a promising step towards building a more reliable and effective model for detecting deepfakes.

The confusion matrix of the advanced model shows the

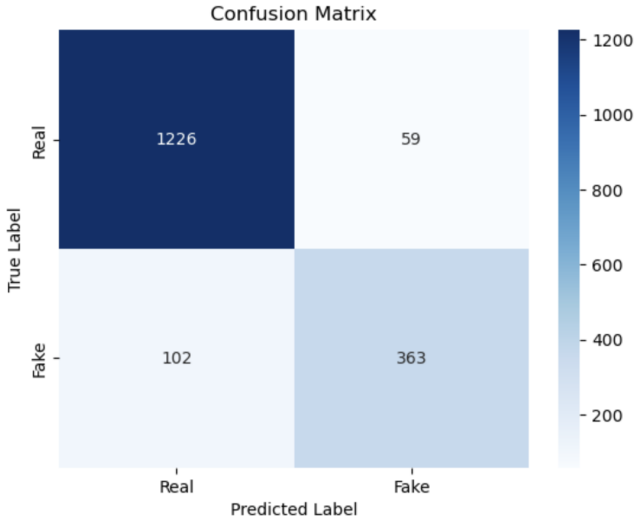

Fig. 13. Confusion Matrix of Advance Model

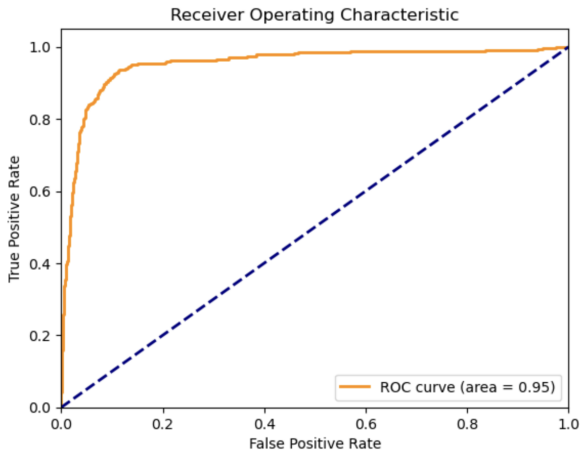

Fig. 14. ROC Curve of advanced model

performance measure of the model distinguishing the real images from fake ones. The model was able to detect 1226 real images and 363 fake images which proves that most images can easily be classified by the model. Nonetheless, there were still some errors; the classifier misclassified 59 real images as fake images, and 102 fake images as real images. This pattern indicates that overall the model is quite successful, yet it is not immune to difficulties in detecting peculiarities, especially in the case of fake images. The fact that the model has fewer false positives, actual images being classified as fake and false negatives , fake images being classified as real shows that it is a little on the cautious side, and it is right to be that way since it is more important to ensure that the model does not classify real images as fake. In general, this advanced model suggests a certain improvement in fake image detection, however, there is a number of cases of fake images being classified as real that needs further improvement, See Fig. 13.

The comparison between the confusion matrices of the basic and advanced models reveals a clear enhancement in the model's performance. The advanced model demonstrates a more refined ability to distinguish between real and fake images, significantly reducing the number of misclassifications compared to the basic model. This improvement reflects the advanced model's enhanced understanding and generalization, leading to better detection of fake images and fewer false positives. The advanced model's ability to maintain a cautious approach while still improving accuracy shows that it is more reliable and effective in detecting deepfakes, which is crucial for the robustness of the overall system.

The ROC curve for the advanced model clearly shows that the model is very efficient in distinguishing between the real images and the fake ones. The curve comes very close to the top left corner, which is a good sign since it means that the model is very accurate in its predictions. It is scoring an AUC of 0. 95, this model proves its efficiency in identifying images and therefore can effectively work in fake content detection. The obtained AUC value is very high, which confirms the model's fine-tuning and, thus, the ability to exclude errors in classification.

Comparing ROC curves of the basic and advanced models shows that there is a good increase in the performance of the model in terms of discriminating between the real and fake images. In the basic model, the ROC curve with the AUC of 0. 74, therefore, indicate moderate level of discrimination. The fact that the curve is close to the diagonal line means that the model's predictions are barely better than the toss of a coin. On the other hand, the advanced model shown a much better performance with AUC of 0. 95. This near to being a perfect curve indicate that the advanced model is very effective in detecting the fake images and real images with very little error of wrongly identifying a fake image as a real image and vice versa. The increase in the value of the ROC curve when going from the basic to the advanced model means that the changes introduced in the latter, including face detection , fine-tuning and batch normalization, have contributed to significant improvements in the accuracy of predictions. Such a difference between two models supports the further development of the models and the usage of more complex approaches to improve accuracy and efficiency of deepfake detection.

The Precision-Recall curve for the advanced model paints a promising picture of its capabilities. With an impressive area under the curve (AP) of 0.89, the model demonstrates that it can accurately pinpoint fake images while keeping errors to a minimum, even as it strives to identify as many fake images as possible. The curve's steady performance across various recall levels indicates that the model is consistently good at finding fake images without falsely labeling real ones as fake. This balance between precision and recall is vital, especially in the realm of deepfake detection, where the stakes are high. The advanced model's ability to maintain both precision and recall at such high levels speaks volumes about its reliability and potential in real-world applications.

When comparing the Precision-Recall curves between the basic and advanced models, a clear distinction in performance

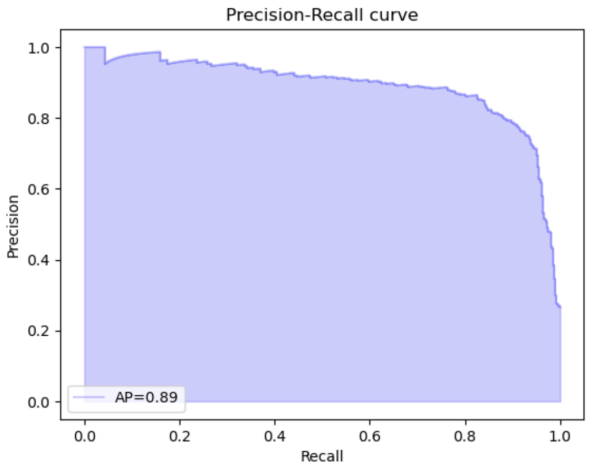

Fig. 15. Precision Recall Curve of advanced model

becomes evident. The advanced model exhibits a much higher precision across a wide range of recall values, as shown by the area under the Precision-Recall curve (AP=0.89), which is significantly better than the basic model's performance (AP=0.53). This indicates that the advanced model is far more effective in distinguishing between real and fake images, consistently maintaining high precision even as it captures more true positives. The advanced model's ability to sustain a high level of precision while increasing recall suggests a robust capacity to correctly identify both real and fake images, minimizing false positives and false negatives. In contrast, the basic model struggles to balance precision and recall, leading to a much lower overall performance. The improvements in the advanced model highlight the success of the enhancements made in its architecture and training process, leading to a more reliable and accurate deepfake detection system.

VII. Future Enhancements

To enhance the generalisation capabilities of the detection model, future work could focus on integrating multiple datasets that contain high-quality videos. As they can offer diverse and challenging deepfake videos that can help in training models that generalise better across different types of deepfakes. It would also reduce models overfitting to specific artifacts present in a single dataset, making it more robust to variations found in real-world scenarios.

Beyond detecting deepfakes in images and videos, there is a growing need to address multimodal deepfakes that involve not just visual content but also audio, text, and even synthesized behavioral patterns. Future research could explore integrating multiple models that specialise in detecting audio deepfakes, text-based manipulations, and other forms of synthetic media. Integration of these models could result in a comprehensive system that is capable of identifying a wider range of deepfakes across different media formats. Another is to identify deepfakes in real-time, which would be highly useful in daily use, for example, when working with live broadcasts or moderating content in social networks. Future work can be focused on the further enhancement of the Xception model for real-time application where the main objectives will be to minimize the latency and computational complexity. Other consideration is scalability whereby the model should be able to handle massive amount of data flow without a significant reduction in accuracy.

The rapid evolution of deepfake generation techniques necessitates the development of models that can continuously update and adapt. Future models should incorporate mechanisms for ongoing learning, potentially through techniques like online learning or the integration of active learning strategies. By continuously updating the model with new data, it can remain effective against emerging deepfake techniques that are not covered by the initial training data.

Since deepfake detection systems are being integrated in key fraud detection solutions, it is crucial to build models that are not only effective but also transparent. Future work could focus on incorporating such methods into the deepfake detection pipeline. This would entail developing mechanisms or interfaces to provide users with reasons why a given video or image was classified as deepfake which becomes essential in the critical applications or domains including legal systems and media verification.

Deepfake detection models should be invariant to cultural differences and domains, as a result of differences in facial structure, pronunciation, and ethnicities. The following improvements could be made; training of the model on new datasets containing information on different subjects from different culture. Further, other methods such as transfer learning approaches may be used to fine-tune the model to other domains without having to train it from scratch.

VIII. Conclusion

On the same note, the study showed the limitations, and suggestions for future improvements of the research. The evaluation of the model shows that it is possible to achieve a better performance when using multiple high quality datasets that would make the model more generalised for different types of deepfakes. Additionally, expanding the detection capabilities to include multimodal deepfakes those involving audio, text, and behavioral patterns would create a more comprehensive system.

In the future, it is crucial to have a model that is updated in real-time, as deepfake generation technologies are rapidly evolving. Incorporating online learning and active learning strategies can ensure that the model remains effective against emerging threats. Furthermore, the development of explainable AI techniques within the detection framework will be crucial for critical applications, such as legal and media verification, where transparency is essential.

In conclusion, while the current system shows promise, there is a clear path forward that involves integrating more diverse datasets, expanding to multimodal detection, and ensuring the model can adapt and remain transparent as deepfake technology continues to advance. These steps will be vital

in maintaining the effectiveness of deepfake detection in the increasingly complex digital landscape.

REFERENCES

[1] B. Kaddar, S. A. Fezza, W. Hamidouche, Z. Akhtar, and A. Hadid, "HCiT: Deepfake Video Detection Using a Hybrid Model of CNN Features and Vision Transformer," International Conference on Visual Communications and Image Processing (VCIP), 2021.

[2] A. Das, K. S. A. Viji, and L. Sebastian, "A Survey on Deepfake Video Detection Techniques Using Deep Learning," Second International Conference on Next Generation Intelligent Systems (ICNGIS), 2022.

[3] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[4] Pavel Korshunov and Se' bastien Marcel. "Deepfakes: a new threat to face recognition assessment and detection," arXiv preprint arXiv:1812.08685, 2018.

[5] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 3207-3216, doi: 10.1109/CVPR42600.2020.00327.

[6] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Chris- tian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In ICCV, 2019.

[7] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) preview dataset. arXiv preprint arXiv:1910.08854, 2019.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.