Deep Fake Image Detection using Xception Architecture

Paritosh Joshi
Department of Computer Science and Engineering
Amrita School Of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
paritoshj2001@gmail.com

Dr. Nivethitha V
Department of Computer Science and Engineering
Amrita School Of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
v_nivethitha@ch.amrita.edu

2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST) | 979-8-3503-5137-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRTCST61793.2024.10578398

***Abstract*–Deep Fake technology has become increasingly sophisticated, posing a significant challenge to the integrity of digital content in today's information age. This research paper introduces a novel approach in detecting deep fake images and videos utilizing the Xception model, a deep CNN known for its exceptional image classification capabilities. The study begins by collecting a diverse dataset containing real and deep fake images and videos across various domains. The Xception model is fine-tuned for deep fake detection through transfer learning. To improve model robustness, a range of data augmentation strategies and regularization techniques are incorporated during training. The results of the proposed model demonstrate an accuracy of 93.01% on the test dataset. The proposed approach exhibits promising potential in combating the spread of deceptive and malicious deep fake content across the internet and social media platforms. In conclusion, this research paper presents a reliable and efficient method for deep fake image and video detection, leveraging the Xception model's capabilities.**

Keywords— CNN, Deep Fake, Fine-tuned, Xception

I. INTRODUCTION

In an era marked by the rapid advancement of artificial intelligence and machine learning technologies, the proliferation of deep fake images and videos has emerged as a significant societal challenge. Deep fakes, manipulated digital media created with the intent to deceive, have the potential to undermine trust in visual content and pose serious threats to various aspects of society, from politics and journalism to personal privacy. Detecting these sophisticated forgeries has become a pressing need in the digital age. Deep fake technology has the power of deep learning, particularly generative adversarial networks (GANs) and deep neural networks, to produce counterfeit content that is virtually indistinguishable from genuine media. By training on extensive datasets of real faces, voices, and contexts, these algorithms can synthesize entirely fabricated content, making it increasingly challenging to discern fact from fiction. One of the most unsettling aspects of deep fake technology is its ability to manipulate public perception. Fake videos can depict individuals saying or doing things they never did, potentially leading to misinformation, character assassination, or even geopolitical turmoil. The ease with which deep fakes can be created and disseminated via the internet raises concerns about the erosion of trust in visual media. This research paper addresses the critical issue of deep fake image and video detection, presenting an approach that leverages the power of deep learning, particularly the Xception model, to combat the proliferation of deceptive visual content.

II. RELATED WORK

In the recent years, the burgeoning issue and proliferation of deep fake technology has prompted extensive research and development efforts in the realm of deep fake detection. A fundamental understanding of these efforts is crucial to contextualize the contributions of the present research. Several noteworthy research papers have emerged, each offering unique insights into the methodologies and algorithms utilized in the pursuit of effective deep fake detection.

"A Comprehensive Survey of Deep Fake Detection Techniques" (2021) provides a systematic examination of deep fake detection methods, encompassing both image and video analysis. This survey, authored by experts in the field, discusses feature-based, deep learning-based, and ensemble-based approaches. Notably, it delves into algorithms like Xception, VGG, ResNet, LSTM, and Capsule Networks (CapsNets), highlighting the latter for their ability to capture intricate spatial hierarchies within images.

Further research includes "Deep Fake Detection Using Temporal Coherence and Cyclic Consistency" (2021), [2] which introduces a unique approach emphasizing temporal information to detect deep fake videos. This research by Yangyang Su, Xudong Wang, Fengqing Qin, and Lihuo He employs optical flow-based motion analysis and Recurrent Neural Networks (RNNs) to model temporal coherence, effectively identifying deep fake videos with realistic facial movements. Additionally, "Learning to Detect Deep Fakes with Capsule Networks" (2020) by Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu explores the use of Capsule Networks (CapsNets) for spatial analysis in deep fake detection, capitalizing on their resilience against adversarial attacks.

"Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis for Robust Disentanglement of Identity and Voice" (2020) [3] by Ye Jia, Yu Zhang, Ron J. Weiss, and others, while primarily focused on voice synthesis, addresses the challenges of voice-based deep fake detection. The paper introduces techniques like speaker verification-based detection, employing Siamese networks to distinguish between real

979-8-3503-5137-8/24/$33.00 ©2024 IEEE

and synthesized voices, which is pivotal in the context of deep fake audio detection. These recent research endeavours collectively illuminate the diverse methodologies and algorithms employed to combat the ever-evolving landscape of deep fake technology.

“Robust Deepfake On Unrestricted Media: Generation And Detection” (2022) [4] by Trung-Nghia Le, Huy H Nguyen, Junichi Yamagishi, Isao Echizen is a review paper that talks about the evolution and challenges in deep fake generation as well as the deep fake detection. A full-fledged list of possible ways to improve the robustness of deep fake detection in a wide variety of media is also discussed in this review paper. Also, a future scope for fake media research is enlightened in the paper.

“The Effectiveness of Temporal Dependency in Deep Fake Video Detection" (2022) [5] by Will Rowan and Nick Pears investigates the question, whether incorporating temporal information can enhance the efficacy of deep learning models in identifying deep fake content. To address this inquiry, they introduced a structured framework that categorizes various deep fake detection methods based on their distinguishing attributes. These attributes perform feature extraction, which can be either automated or manual, as well as the temporal relationship between consecutive frames, which can be characterized as dependent or independent.

III. METHODOLOGY

Our methodology serves as a comprehensive guide to comprehensively evaluate the model's performance and contributes valuable insights to the field of multimedia forensics. We have followed the following structured approach.

A. Data Collection and Preprocessing

Our search ended at the *deepfake_faces* dataset from Kaggle which is a collection of more than 90,000 fake and real images consisting of human faces. We standardized the data by resizing, cropping, and normalizing images and videos. We also took care of consistent formatting to facilitate model training and paid special attention to handling data imbalances between genuine and deep fake samples.

B. Transfer Learning with VGG-16

Xception is chosen for its exceptional image classification capabilities and its suitability for transfer learning. We initialized the Xception model with pre-trained weights from a huge-scale classification task for images (e.g., ImageNet). Following this we fine-tuned the model's layers to adapt to the specific task of deep fake detection while retaining its ability to capture high-level features. This results in the accuracy to jump to 66.1%.

C. Data Augmentation and Regularization

We have applied data augmentation techniques in increasing the diversity of the training dataset. Techniques such as random rotation, flipping, and contrast changing can help the model generalize better. This is done to make sure that overfitting is avoided while training the model on the dataset. Also, implementing regularization methods like dropout and L2 regularization in order to prevent overfitting and enhance the model robustness.

D. Training Process

We have selected an appropriate loss function, such as binary cross-entropy to measure the dissimilarity between and actual labels (fake and real). We have employed the Stochastic Gradient Descent (SGD) to update the model weights during the training process. The learning rate is set to 0.01 along with the momentum set to 0.9. Next, we fit the model on the training set and validation set for 15 epochs. The model takes its time to train.

E. Model Architecture

The Xception model, short for "Extreme Inception," represents a groundbreaking advancement in the realm of deep learning architectures, particularly within the domain of computer vision. It embodies a fundamental shift in convolutional neural network (CNN) design by embracing the concept of depthwise separable convolutions, a key innovation that has since influenced numerous subsequent models. At the heart of the Xception architecture lies the concept of depth wise separable convolutions, which replaces traditional convolutions with a more efficient and powerful alternative. Traditional convolutional layers involve convolving each input channel with a learnable filter, resulting in a computationally intensive process. Depthwise Convolution*:* In the first stage, the input data is convolved channel-wise with separate learnable filters. This process captures spatial features within each input channel independently, reducing computational complexity. Pointwise Convolution*:* The second stage involves pointwise convolutions, where the depthwise convolved outputs are linearly combined using 1x1 convolutions. This step enables the network to learn cross-channel feature interactions. Fig 1 below shows the Depthwise and Pointwise convolution of the Xception model architecture.

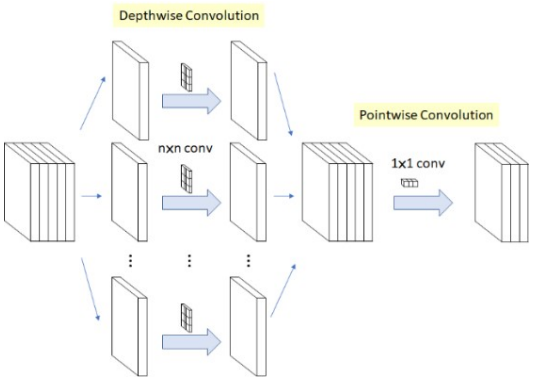

Fig. 1. Xception – Depthwise & Pointwise Separable Convolution

IV. RESULT

In this section, we present the outcomes of our deep fake detection research using the Xception model. We detail the experimental setup, provide a comprehensive analysis of the model's performance, and discuss the implications of our findings. The result includes Model Accuracy and Loss, Test Dataset Evaluation and Receiver Operating Characteristic (ROC) Curve which gives us the trade-off between the TPR and the FPR at different classification thresholds and the Confusion Matrix which illustrates how capable our model is to classify the input images into real and fake correctly, along with some false classification.

A. Training Accuracy and Training Loss

In the realm of machine learning and deep learning, the pursuit of accurate and effective models hinges on the measurement and interpretation of two fundamental metrics: model accuracy and loss. These metrics serve as crucial indicators of a model's performance. We obtain the insights of the model's learning after 15 successful epochs as shown in Fig 2 and Fig 3.

1) *Training Accuracy :* Our trained model illustrates a constant incremental growth in training accuracy during the initial epochs.
2) *Validation Accuracy :* Simultaneously, validation accuracy also experiences a rising trend and ends up validating the validation dataset correctly. Our model achieved a validation accuracy of 92.65%.
3) *Training Loss:* The training loss is observed to decrease smoothly, resulting in minimization of errors during training the model.
4) *Validation Loss:* Along with the training, validation loss also received a smooth decreasing progress, corresponding to an effective generalization.

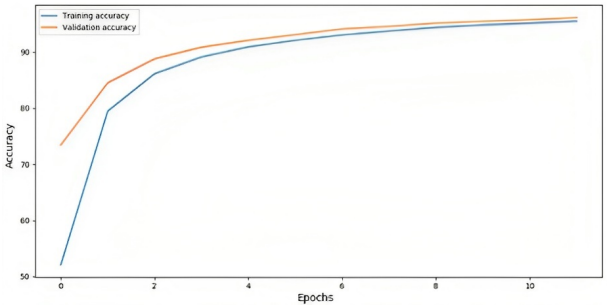

Fig. 2. Accuracy Curve

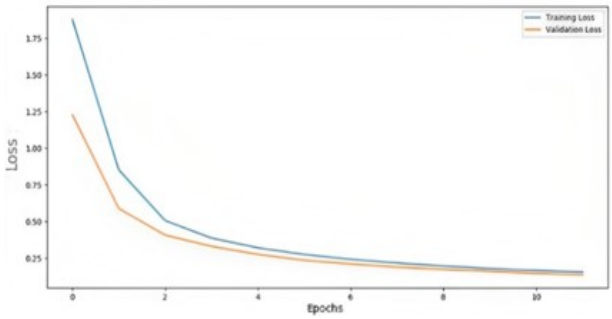

Fig. 3. Loss Curve

B. Test Dataset Evaluation

In the final phase of our deep fake detection project, we turn our attention to the evaluation of our trained model using the test dataset. This critical step aims to assess how well our model generalizes to previously unseen data and provides insights into its real-world applicability. In this section, we delve into the evaluation process, metrics, and the implications of our findings.

1) *Test Accuracy :* Our model was able to achieve an impressive accuracy of 93.01%.
2) Test Loss : Our model gets a test loss of 15.78%, which reconfirms the error minimization for detecting real and fake images on unseen dataset.

C. ROC Curve

Our trained model has achieved an AUC (Area Under the Curve) of 0.922, a metric that distinctly underscores its remarkable ability to discern deep fake images within an unseen dataset. This noteworthy AUC score serves as a beacon, illuminating the model's proficiency in detecting digital deception. The Receiver Operating Characteristic (ROC) curve, eloquently portrayed in Figure 4 below, provides a visual representation of our model's exceptional detection prowess. In practical terms, our model's capabilities extend to a plethora of applications, where the preservation of content integrity and trust in digital media is of paramount importance. Fig 4 shows the ROC Curve.

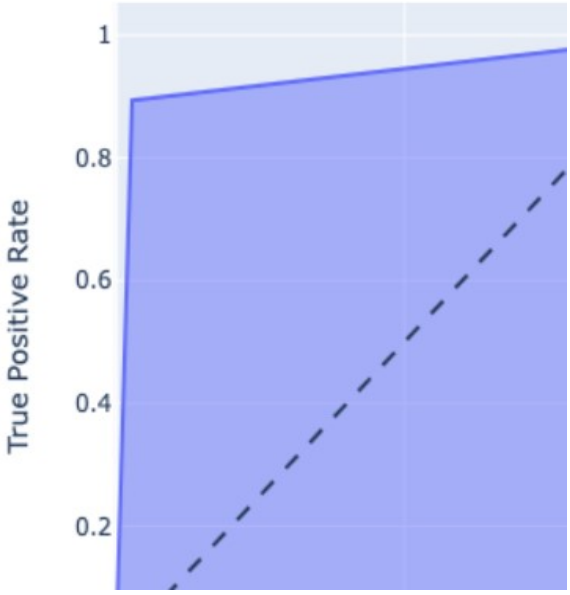

Fig. 4. ROC Curve of Xception Model

D. Confusion Matrix

The model's precision comes to the fore, with an impressive accuracy rate of 95.43% in classifying images as real. This remarkable feat stands as a testament to its unwavering commitment to the preservation of genuine media. This degree of precision positions our model as an ideal choice for applications that demand the utmost accuracy in maintaining the authenticity of real media.

Equally remarkable is the model's adeptness in categorizing 97.67% of the images as fake. This proficiency in identifying manipulated media is nothing short of invaluable, especially in scenarios where the menace of deep fake content looms large, threatening the very fabric of trust and information integrity. The model's inherent capability to unearth digital deception serves as a guardian of media veracity, acting as a bulwark against the propagation of potential misinformation. In essence, the insights gleaned from the confusion matrix, particularly the precision and proficiency exhibited in classifying real and fake media, affirm our model's pivotal role in the ongoing endeavour to safeguard the integrity of digital content. Fig 5 shows the Confusion Matrix.

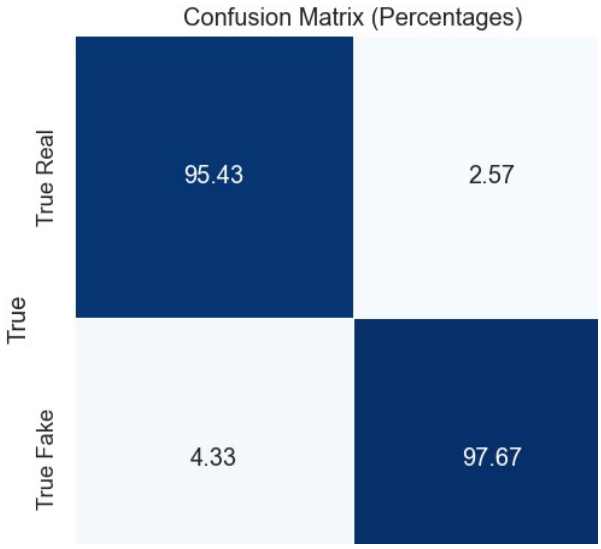

Fig. 5. Confusion Matrix of Xception Model

V. CONCLUSION

In response to this burgeoning challenge of Deep Fake technology, our research paper embarked on a journey to develop a robust and effective deep fake detection system. We have traversed the realms of data preprocessing, model architecture, training, validation, testing, and evaluation to shed light on the complexities of identifying manipulated media. Our research began with meticulous data collection and selection, curating a dataset that encapsulated the essence of deep fake media. Data preprocessing was the bedrock upon which our model was built, extracting meaningful images from video frames, stratifying for a balanced distribution, and meticulously splitting the dataset into training, validation, and test subsets. This foundational work ensured the model's training and evaluation were based on a representative and unbiased dataset. Central to our research was the adoption of the Xception model, a CNN capable of handling complex image classification tasks. Through the incorporation of transfer learning, the model was equipped with the knowledge to discern intricate features in digital media. With a carefully designed architecture and fine-tuning, our deep fake detection system was primed for performance. The results of the experiments depict that our model was able to attain an accuracy of 93.01% on the unseen dataset. The confusion matrix, a visual testament to our model's performance, illuminated the percentage of correctly classified images as real and fake. Our model demonstrated accuracy rates of 95.43% for real and 97.67% for fake media, exemplifying its prowess in preserving authenticity and uncovering digital deception. Furthermore, the Receiver Operating Characteristic (ROC) curve provided a visual representation of the model's discrimination capabilities, affirming its adeptness with an impressive 0.922 AUC score.

REFERENCES

[1] D. Afchar, V. Nozick, and O. Yamaguchi, “MesoNet: a Compact Facial Video Forgery Detection Network”, 2022.
[2] Yangyang Su, Xudong Wang, Fengqing Qin, Lihuo He, “Deep Fake Detection Using Temporal Coherence and Cyclic Consistency", 2021.
[3] Ye Jia, Yu Zhang, Ron J. Weiss, "Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis for Robust Disentanglement of Identity and Voice", 2020.
[4] Trung-Nghia Le, Huy H Nguyen, Junichi Yamagishi, Isao Echizen “Robust Deepfake On Unrestricted Media: Generation And Detection”, 2022.
[5] Will Rowan and Nick Pears, “The Effectiveness of Temporal Dependency in Deep Fake Video Detection", 2022.
[6] Vinaya Sree Katamneni, Ajita Rattani, “MIS-AVioDD: Modality Invariant and Specific Representation for Audio-Visual Deepfake Detection”, 2023.
[7] Leandro A. Passos, Danilo Jodas, Kelton A. P. da Costa, Luis A. Souza Júnior, Douglas Rodrigues, Javier Del Ser, David Camacho, João Paulo Papa, “A Review of Deep Learning-based Approaches for Deep Fake Content Detection”, 2023.
[8] Shanmin Yang, Shu Hu, Bin Zhu, Ying Fu, Siwei Lyu, Xi Wu, Xin Wang, "Improving Cross-dataset Deep Fake Detection with Deep Information Decomposition”, 2023.
[9] Felix Rosberg, Eren Erdal Aksoy, Cristofer Englund, Fernando Alonso-Fernandez, “FIVA: Facial Image and Video Anonymization and Anonymization Defense”, 2023.
[10] Qiaomu Miao, Sinhwa Kang, Stacy Marsella, Steve DiPaola, Chao Wang, Ari Shapiro “Study of detecting behavioral signatures within DeepFake videos”, 2022.
[11] Nguyen, Thanh Thi, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi, “Deep Learning for Deep Fakes Creation and Detection: A Survey”, 2023.
[12] Deeraj Nagothu, Ronghua Xu, Yu Chen, Erik Blasch, Alexander Aved, “ DeFakePro: Decentralized DeepFake Attacks Detection using ENF Authentication”, 2022.
[13] Paarth Neekhara, Shehzeen Hussain, Xinqiao Zhang, Ke Huang, Julian McAuley, Farinaz Koushanfar “FaceSigns: Semi-Fragile Neural Watermarks for Media Authentication and Countering Deepfakes”, 2022.
[14] Davide Coccomini, Nicola Messina, Claudio Gennaro, Fabrizio Falchi, “Combining EfficientNet and Vision Transformers for Video Deep Fake Detection”, 2022.
[15] Koopman, Marissa, Andrea Macarulla Rodriguez, and Zeno Geradts, “Detection of deep fake video manipulation”, 2018.
[16] Bozhi Xu, Jiarui Liu, Jifan Liang, Zhuo Wei, “Detecting Generative Deepfakes via Anomalous Facial Textures”, 2023.
[17] João C. Neves, Ruben Tolosana, Ruben Vera-Rodriguez, Vasco Lopes, Hugo Proença, Julian Fierrez, “GANprintR: Improved Fakes and Evaluation of the State of the Art in Face Manipulation Detection”, 2022
[18] Irene Amerini, Leonardo Galteri, Roberto Caldelli, Alberto Del Bimbo, “Deepfake Video Detection through Optical Flow based CNN”, 2022
[19] Guarnera et al. “DeepFake detection by analyzing physiological signals”, 2022
[20] Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, Ziwei Liu, “Detecting and Grounding Multi-Modal Media Manipulation and Beyond”, 2023
[21] Hui Miao, Yuanfang Guo, Yunhong Wang, “RFDforFin: Robust Deep Forgery Detection for GAN-generated Fingerprint Images”, 2023

[22] Yan Ju, Shan Jia, Jialing Cai, Haiying Guan, Siwei Lyu, "GLFF: Global and Local Feature Fusion for AI-synthesized Image Detection", 2023
[23] Jiazhi Guan, Tianshu Hu, Hang Zhou, Zhizhi Guo, Lirui Deng, Chengbin Quan, Errui Ding, Youjian Zhao, "Building an Invisible Shield for Your Portrait against Deepfakes", 2023
[24] Juan Hu, Xin Liao, Difei Gao, Satoshi Tsutsui, Qian Wang, Zheng Qin, Mike Zheng Shou, "Mover: Mask and Recovery based Facial Part Consistency Aware Method for Deepfake Video Detection", 2023
[25] Zhiyuan Yan, Yong Zhang, Yanbo Fan, Baoyuan Wu, "UCF: Uncovering Common Features for Generalizable Deepfake Detection", 2023
[26] Jacob Mallet, Natalie Krueger, Mounika Vanamala, Rushit Dave, "Hybrid Deepfake Detection Utilizing MLP and LSTM", 2023
[27] Yuhang Lu, Touradj Ebrahimi, "Assessment Framework for Deepfake Detection in Real-world Situations", 2023
[28] Mahsa Soleimani, Ali Nazari, Mohsen Ebrahimi Moghaddam, "Deepfake Detection of Occluded Images Using a Patch-based Approach", 2023
[29] Rui Shao, Tianxing Wu, Ziwei Liu, "Detecting and Grounding Multi-Modal Media Manipulation", 2023
[30] Duc-Tien Dang-Nguyen, Sohail Ahmed Khan, Cise Midoglu, Michael Riegler, Pål Halvorsen, Minh-Son Dao, "Grand Challenge On Detecting Cheapfakes", 2023
[31] Chuer Yu, Xuhong Zhang, Yuxuan Duan, Senbo Yan, Zonghui Wang, Yang Xiang, Shouling Ji, Wenzhi Chen, "Diff-ID: An Explainable Identity Difference Quantification Framework for DeepFake Detection", 2023
[32] Yuhang Lu, Touradj Ebrahimi, "Impact of Video Processing Operations in Deepfake Detection", 2023
[33] Tianyi Wang, Harry Cheng, Kam Pui Chow, Liqiang Nie, "Deep Convolutional Pooling Transformer for Deepfake Detection", 2023
[34] Yihao Huang, Felix Juefei-Xu, Qing Guo, Yang Liu, Geguang Pu, "Dodging DeepFake Detection via Implicit Spatial-Domain Notch Filtering", 2023
[35] Liang Shi, Jie Zhang, Shiguang Shan, "Real Face Foundation Representation Learning for Generalized Deepfake Detection", 2023
[36] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, Zheng Ge, "Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization", 2023