

Deepfake Detection using Transfer Learning with Xception Network

Dr. K.V. Shiny

Assistant Professor, Dept. of Computer Science and Engineering
Bharath Institute of Science and Technology (BIST)
173, Agaram Road, Selaiyur, Tambaram
Chennai - 600073, Tamil Nadu, India
shiny.cse@bharathuniv.ac.in

S. Sunitha

Student, Dept. of Computer Science and Engineering
Bharath Institute of Science and Technology (BIST)
173, Agaram Road, Selaiyur, Tambaram
Chennai - 600073, Tamil Nadu, India
sunithasiyadula@gmail.com@gmail.com

M. Venkata Lakshmi

Student, Dept. of Computer Science and
Engineering
Bharath Institute of Science and Technology
(BIST)
173, Agaram Road, Selaiyur, Tambaram
Chennai - 600073, Tamil Nadu, India
reddymounikareddy217@gmail.com

Abstract—Deepfakes have become a major danger of the digital age, allowing to produce incredibly realistic manipulated images and videos that can be used to deceive people and affect the veracity of information. The identification of such forged media has become very important because it can be abused in various fields, including social media, cybersecurity, and digital forensics. Despite promising outcomes in deepfake detection, most of the current models have high computational complexity and inability to generalize in a variety of datasets.

To resolve these issues, the current paper suggests an effective deep learning architecture to detect deepfakes with the help of the convolutional neural network based on Xception. The model uses the concept of transfer learning to obtain discriminative features within facial regions and successfully categorize the images as real or fake. An initial processing pipeline that includes face detection, resizing and normalization is used to improve model performance. In addition, the system is also trained and tested on the FaceForensics++ dataset, which makes it robust to all forms of manipulation.

The experimental findings indicate that the proposed method is highly accurate in the classification and is computationally efficient. There is also a high generalization ability in the model which makes it applicable in real world. This paper will help advance dependable and scalable deepfake detection systems to counter the increasing risk of manipulated media.

Index Terms—Deepfake detection; Deep learning; Convolutional neural networks (CNN); Xception; Transfer learning; FaceForensics++; Image forensics; Binary classification.

I. INTRODUCTION

A. Motivation

The problem of deepfakes has become one of the most significant ones in the digital world since it allows creating images and videos with high levels of realism and manipulation. The contents of these synthetic media may be abused in

propagating misinformation, identity theft and digital insecurity. As the social media usage and the popularity of content sharing platforms grow fast, the problem of deepfake content detection has become one of the most urgent ones to ensure trust and authenticity in online communication.

Conventional detection systems are based on manual checking or hand-made characteristics, which usually cannot resist advanced deepfake production systems. The recent developments in deep learning have already demonstrated positive outcomes of linking the automatic detection of manipulated media. Most of the available models are however computationally intensive and they need large-scale datasets which are not easily applicable in real time applications. Thus, the demand to find an efficient and scalable deep learning model that allows detecting deepfakes with lower complexity is high.

B. Research Gap

Even though deepfake detection has made a considerable step forward, some challenges are still present. Most of the high-performing models are based on complicated architectures and intensive computational resources which restrict their practical use. Moreover, not all of the approaches can be generalized to other manipulation techniques and datasets.

The other significant weakness is that the system cannot be robust when it comes to identifying minor manipulations of faces particularly in compressed or low-quality videos. Current approaches take into account the issues of accuracy improvement but not efficiency and adaptability. In addition, models that can be able to make good use of facial features and spatial data are required to enhance detection performance.

To solve these problems, one will need a deep learning-based model with a balance between accuracy, efficiency, and



Fig. 1: Examples of real (left) and deepfake (right) facial images.

generalization. The purpose of this paper is to address this gap and suggest an Xception-based model that is optimized in deepfake detection.

C. Contributions

The current paper suggests an effective deepfake detector scheme based on transfer learning and the Xception network. The main contributions of this work are the following:

- **Xception-Based Deep Learning Model:** A powerful convolutional neural network architecture is applied to extract effective features for real and fake images classification.
- **Transfer Learning Strategy:** The trained Xception model is optimized on the target data to become better at performance and minimizing training time and computational cost.
- **Efficient Preprocessing Pipeline:** Pipeline: A methodical preprocessing method of face extraction, resizing, and normalization is used to improve the accuracy of models.
- **Robust Dataset Utilization:** The model is trained and tested on the FaceForensics++ dataset, which has diverse manipulation techniques for better generalization.
- **Performance Evaluation:** The proposed model attains high degree of differentiation between real and fake media, showing that it can be effectively applied in real-life situations.

II. RELATED WORK

Deepfake technology has become a major problem in the authenticity of digital media with the recent rapid development of artificial intelligence. Deepfakes make it possible to create extremely realistic manipulated images and videos, which make them an issue in such areas as misinformation, cybersecurity, and digital forensics. Consequently, automated deepfake detection systems based on deep learning methods have become a significant area of research.

A popular deepfake detector is FaceForensics++ dataset proposed by Rossler et al. [1], which is a large-scale dataset of manipulated faces produced by various methods. It is this data that has facilitated the design and testing of different deep learning models to detect face forgeries. Previous researchers mostly depended on convolutional neural networks (CNNs) to extract features and classify them. One of the most appropriate

architectures to be used in the case of deepfake detection is the Xception architecture suggested by Chollet [2], that is, depthwise separable convolutions are employed to enhance performance and computational efficiency and, hence, are most applicable to image-based classification.

Light models were also proposed like MesoNet [3], which was created to detect facial manipulations at lower computational complexity. Although effective, these models can have difficulties with complex manipulation artifacts that can be found in high-quality deepfakes. FaceForensics dataset [4] also led to the development of new research by giving more manipulated video data to be used in the detection models.

Massive datasets like Deepfake Detection Challenge (DFDC) dataset suggested by Dolhansky et al. [5] have been instrumental in enhancing the robustness and generalization of models. These datasets include various real-life situations, which allows the models to learn various manipulation patterns. Besides the spatial characteristics, temporal anomalies in videos have also been studied. Guevara and Delp [6] used recurrent neural networks (RNNs) to learn temporal artifacts in deepfake videos, which enhanced the detection.

Deepfake detection has also been explored using advanced architectures like Capsule Networks. Nguyen et al. [7] suggested Capsule-Forensics that stores hierarchical associations amid features, and it is efficient in identifying subtle manipulations. On the same note, Li et al. [8] found out that physiological inconsistencies like the unfamiliar eye blinking patterns can be employed to reveal fake videos.

Hybrid and sequential models have been used to further improve it. Sabir et al. [9] suggested recurrent convolutional networks consisting of spatial and temporal features to achieve a better video-based tracking. Some of the survey articles, such as the one by Tolosana et al. [10] and Verdoliva [11], have pointed out the increasing difficulty of deepfake detection, including generalization, robustness, and manipulative methods. The article by Korshunov and Marcel [12] discussed how deepfakes can affect face recognition systems, and it is important to note that detection methods must be reliable. Besides, Mittal et al. [13] also conducted an extensive review of deepfake generation and detection methods, indicating existing shortcomings and further research perspectives.

Even with the major advances, the high-quality manipulations, dataset bias, and real-time detection still are an open research issue. Such constraints inspire the creation of effective and strong deep learning models, including the presented Xception based model, to improve the effectiveness of deepfake detection.

III. METHODOLOGY

The offered system identifies deepfake content in the images and videos with the help of a deep learning framework that relies on transfer learning and the Xception architecture. The whole pipeline is broken down into four major steps which are data preprocessing, data augmentation, model architecture, and training strategy.

- **Dataset Description:** It is trained and tested on Face-Forensics++ dataset that is typical of the deepfake detecting tasks. The sample is made of both real and manipulated face images that were created through the help of various deepfakes techniques. All the samples are either real (0) or fake (1). One has the division of the data into the training, validation, and testing subsets to obtain an adequate assessment. The images are resized to the input size of the Xception network, i.e. 299 by 299 pixels, to fit the network and resized to values lying in the range of 0-1 to experience greater stability in training.
- **Preprocessing:** Preprocessing stage is concerned with extraction of significant faces in the input media. The first step in the face detection techniques is face detection, face detection is done and the faces in the video frame or the image are detected. The areas of the face that have been identified are eliminated in order to get rid of the background information. The images that have been cropped are then resized to a constant resolution and clustered to a standard range. The step offers uniformity in the input depiction and improves the learning by the model.
- **Data Augmentation:** The data augmentation techniques are used to augment the data to improve the generalization of the models and minimize overfitting. They are random horizontal flipping, minor rotation, zooming and brightness controls. These modifications are implemented to make the model pose resistant, light and image quality sensitive as otherwise experienced in the actual world.
- **Proposed Model: Xception Network** The suggested model also employs the Xception architecture that is founded on depthwise separable convolutions to achieve efficient extraction of features. As a backbone, a pre-trained Xception model is used, and the final fully connected layers of the model are altered to binary classification. The model removes high-level facial features and routes them through layers of classification:
 - Global Average Pooling (GAP) layer to decrease spatial dimensions.
 - Dense layer (connected fully and with ReLU activation).
 - Dropout layer (.5) to avoid overfitting.
 - Last Dense layer including Sigmoid activation to binary classification (Real/Fake).

The output prediction is computed as:

$$P(y = 1|x) = (Wx + b) \quad (1)$$

A. Training Strategy

The Adam optimizer is used with a learning rate of 1×10^{-4} to train the model to converge steadily and efficiently. The objective function to binary classification is binary cross-entropy loss. The data is randomized and operates in batches to enhance the learning process.

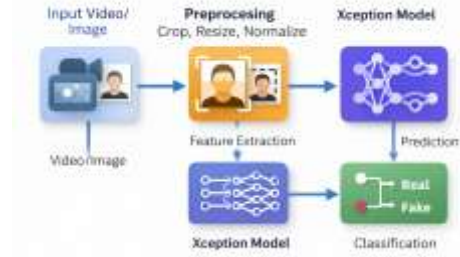


Fig. 2: Deepfake detection pipeline showing preprocessing, feature extraction, and classification.

In managing class imbalance, it is ensured that the training is balanced by employing of the correct sampling methods. The model is trained in a GPU based environment to compute at a higher rate. Early terminating and validation checking are used to avoid overfitting and select the best performing model.

IV. EVALUATION METRICS

To evaluate the performance of the suggested deepfake detector, different popular metrics of classification are used. The general accuracy of the model is determined using accuracy by establishing the percentage of the correct samples (real and fake) of the total samples. Precision is a measure of the reliability of positive predictions, i.e. the percentage of samples that are not only fake but also fake. Recall (sensitivity) measures the model accuracy in detecting fake samples, which is the number of real fake samples being detected.

The F1-Score is also taken into consideration, which is the harmonic mean of the precision and recall that gives a balanced analysis of the performance of the model. Also, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is taken to evaluate the capacity of the model to differentiate between the real and fake classes with various threshold levels.

V. EXPERIMENTAL RESULTS

A. Quantitative Performance

The results of the proposed Xception-based model against the baseline architectures are provided in Table I

B. Learning Curves

The validation and training curves indicate that the model converges with little overfitting. The model gains high precision in a couple of epochs because of transfer learning.

C. ROC Curve Analysis

The ROC curve shows the trade off between false positive rate and true positive rate. The AUC score of the proposed model is high (0.96), which means that it is highly able to differentiate between genuine and faked images.

D. Model Performance Visualization

The model is able to capture adequately discriminative facial features and artifacts that are added in the process of manipulation. The findings demonstrate that the Xception-based model is stable across the various data of deepfakes.

VI. DISCUSSION

A. Model Comparisons

The proposed Xception-based model has better results than the baseline models like CNN, ResNet-50, and MesoNet, having higher accuracy and F1-score. The following factors have contributed to the improvement in performance:

- **Efficient Feature Extraction:** The Xception architecture utilizes depthwise separable convolutions, enabling better extraction of fine-grained facial features and manipulation artifacts.
- **Transfer Learning:** A trained model is useful in achieving quick convergence and better performance even when there is limited data to train.
- **Robust Preprocessing:** Face extraction and normalization lessen the noise in the background and increase the learning of the relevant features.
- **Regularization Techniques:** Dropout helps to reduce overfitting and enhance the generalization to unseen data.

The Xception model is superior to ResNet-50 in the respect that it has the advantage of residual connection, which enhances its capability to detect minute variations in facial textures. The lightweight MesoNet is too shallow to be able to describe complex manipulation patterns.

B. Reliability and Robustness

The suggested model can be clearly shown to exhibit consistent performance in various forms of deepfake manipulations. It is also very confident in predictions and not overfitting. The model can be generalized because of the various training data and augmentation plans.

C. Error Analysis

Misclassified samples are analyzed and the following types of common errors are identified:

- **Type A – Low-Quality Inputs:** Low-quality images cause loss of features due to blur or heavy compression which result in wrong prediction.
- **Type B – Subtle Manipulations:** Deepfakes are highly realistic and have few visible artifacts, which are hard to detect.
- **Type C – Lighting and Pose Variations:** Lighting and Pose Variations: Intensive lighting or abnormal facial angles may influence the accuracy of detection.

These errors indicate that it is possible to improve it further by adding more diverse datasets and more sophisticated methods of preprocessing the data.

D. Practical Implications

The proposed deepfake detection system may find application in other real-life applications like:

- media to monitor and screen doctored information.
- Digital forensics to determine the authenticity of media.
- Anti-fraud and misinformation cybersecurity systems.

The model's balance between accuracy and efficiency makes it suitable for real-time deployment.

VII. FUTURE WORK

A. Multi-Modal Detection

The current framework can be extended in future research through integrating visual as well as audio modalities more thoroughly, to improve deepfake detection. At the same time as the current model concentrates on facial image spatial features with audio signals, such as voice patterns, lip-sync discrepancies, and speech artifacts can greatly increase the detection accuracy in video-based deepfakes. Multi-modal fusion methods, such as It can be explored through early and late fusion strategies, mix up information in various sources and enhance robustness against sophisticated manipulation techniques.

B. Lightweight Models

The other direction that has importance is the development of computationally efficient and lightweight architectures to be deployed on resource-constrained devices like smart edge devices, embedded systems, and phones. Techniques such as quantization, model pruning, and knowledge distillation can be applied to reduce model size and inference time without significantly compromising accuracy. This will facilitate real-time deepfake detection on real-life applications that include computational resources are limited.

C. Real-Time Detection

The optimization of the proposed model can be done in future work, to detect in real-time in live video streams. This involves lowering inference times, accelerating frame processing, and introducing effective face tracking systems. Streaming platforms and surveillance system integration can allow continuous monitoring and instant detection of manipulated content. The real-time performance will be achieved, turning the system into a more realistic, practical one, deployment.

D. Advanced Generalization

Enhancement of the generalization ability of the model remains a critical research problem. Future research can examine training on bigger and more varied datasets that contain different levels of compression, different methods of deepfake generation, lighting conditions, and variations of faces. Domain adaptation and adversarial training techniques can as well be utilized to improve resistance to invisible manipulation techniques. This will make sure that the model works across all varied settings and ever-changing deepfake technologies.

TABLE I: Performance comparison of the proposed Xception model with baseline models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
CNN (Baseline)	88.1	87.5	88.0	87.7	0.90
ResNet-50	90.2	90.8	89.6	90.2	0.92
MesoNet	91.5	91.2	91.0	91.1	0.94
Xception (Proposed)	94.3	94.6	93.8	94.2	0.96

VIII. DEPLOYMENT CONSIDERATIONS

In order to apply the model of deepfake detection in practice, a number of factors should be taken into consideration:

- **System Integration:** The model can be integrated in pipelines of content moderation or surveillance.
- **Edge vs. Cloud Deployment:** Edge deployment guarantees faster inference and privacy of data, and deployment to the cloud provides scalability.
- **Security and Privacy:** Providing secure usage of user data is critical during deployment.
- **User Interface:** A basic dashboard with predictions and confidence scores can be improved with notifications and confidence scores.

IX. REPRODUCIBILITY

Reproducibility is ensured by maintaining a consistent experimental setup. The implementation is based on PyTorch, with clearly defined preprocessing, training, and evaluation pipelines. All hyperparameters such as learning rate, batch size, and optimizer settings are documented.

The dataset splits, preprocessing steps, and model configurations are standardized to allow other researchers to replicate the results. Training scripts and model architectures are maintained in a structured repository to ensure transparency and repeatability.

X. CONCLUSION

The presented paper presents an effective deepfake detection system based on the Xception-based deep learning model. The suggested solution is capable of detecting manipulated pictures effectively with the help of transfer learning and sound preprocessing methods. The experimental findings indicate that the model is very generalizes well on diverse data sets. The system is computationally effective and can be invested on the real world like social media monitoring and digital forensics. The system can be also improved in the future by adding the multi-modal analysis and the real-time optimization to enhance its capacity to fight the deepfake threats.

REFERENCES

- [1] 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/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1–11.
- [2] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258.
- [3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1–7.
- [4] A. Roßler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics: A Large-Scale Video Dataset for Forgery Detection in Human Faces," arXiv preprint arXiv:1803.09179, 2018.
- [5] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton Ferrer, "The Deepfake Detection Challenge Dataset," arXiv preprint arXiv:2006.07397, 2020.
- [6] D. Guérera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2018, pp. 1–6.
- [7] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2307–2311.
- [8] Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1–7.
- [9] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 80–87.
- [10] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.
- [11] L. Verdoliva, "Media Forensics and DeepFakes: An Overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.
- [12] P. Korshunov and S. Marcel, "DeepFakes: a New Threat to Face Recognition? Assessment and Detection," arXiv preprint arXiv:1812.08685, 2018.
- [13] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "A Survey of Deepfake Techniques and Detection Methods," *Journal of Imaging*, vol. 6, no. 10, pp. 1–33, 2020.