JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

An Optimized ResNet18-Based Deep Learning Framework for Deep Fake Detection

Kalpana Kumari MTech
Scholar CSE, Department NIIST, Bhopal
Kalpanaraj11433@gmail.com

Nitesh Gupta
Associate Professor
CSE, Department NIIST,Bhopal
9. nitesh@gmail.com

***Abstract*:**

Deepfake technology threatens the credibility of digital media increasingly, allowing for the creation of highly realistic but false videos and images. Proper detection of deepfake content is required to help preserve trust and avoid exploitation. In this paper, a new deep learning method that uses ResNet18 with pre trained weights is introduced to achieve higher feature extraction and classification accuracy in detecting deepfakes. A sub-sample of real and forged images undergoes resizing, normalization, and data augmentation algorithms to provide model generalization. Model optimization is obtained using a fully connected layer, dropout regularization, and sigmoid activation function to enable binary classification. Training is done using the Binary Cross-Entropy loss function and optimized using the Adam optimizer to obtain quicker convergence. Model accuracy and ROC-AUC on test and validation sets are used as metrics to measure performance. The proposed framework illustrates a scalable, efficient, and effective means to identify deepfakes with real-world implications in social media monitoring, digital forensics, and cyber security.

Keywords*— *DeepFake Detection, Image classification, Computer Vision, Deep learning, Model generalization, Deep Learning

I. INTRODUCTION

The rapid development of artificial intelligence and machine learning technologies has enabled the creation of technologies that can create highly realistic synthetic media. The most sinister use of this kind of technology might be to create deepfakes manipulated videos or images in which facial movements or features are altered to deceive regarding what is real. Deepfakes already are being employed, for instance, for entertainment or creative purposes, but their potential for misuse raises essential ethical, security, and privacy concerns. Deepfakes are potentially weaponized for spreading false information, identity fraud, scams, and social manipulation; hence their detection is of great importance in digital forensics and cybersecurity fields. Deepfake detection is not that easy, though, as newer manipulation methods are refined and subtle. Complex neural network structures like Generative Adversarial Networks (GANs) are primarily abused by the majority of deepfake generation tools to create output that is indistinguishable and quite hard to tell apart from original content. Rule-based or feature-engineering methods are insufficient in addressing the depth and complexity of the manipulated objects, hence the need to utilize stronger and more resilient detection methods. Deep learning, through its capacity to automatically learn discriminative features from enormous databases, presents a viable solution to the detection of deepfakes. Particularly, convolutional neural networks (CNNs) have proved to work best in image-based classification problems, such as face identification, object detection, and anomaly detection. Having pretrained models that learned robust feature representations from large datasets also boosts the model by shortening training time and increasing accuracy [1].

In this research, we introduce a more advanced deep learning system that incorporates ResNet18 architecture [21], a universally acclaimed CNN model because of its residual learning features and effective training methods. Utilizing pretrained weights and fine-tuning the classification layers of the model, the system proposed here is bound to differentiate facial images with and without post-processing effectively. The architecture incorporates some essential preprocessing techniques such as resizing, normalization, and augmentation, whose function is to make the input data

standardized and prevent overfitting during training. It is learned with Binary Cross-Entropy loss function, which can be employed for

problems involving binary classification, and Adam optimizer, which facilitates fast convergence and aggressive learning even in the

presence of noisy data. While verifying the model's robustness, it is tested employing typical evaluation measures such as accuracy and Receiver Operating Characteristic Area under the Curve (ROC-AUC) and these give an idea of its performance on observed as well as unobserved datasets. This work not only seeks to attain high detection precision but also offer a scalable and realistic solution that can be readily applied to actual applications. The system is modularly designed in a way that it can easily be scaled for deeper data sets, more complex architectures, or even deepfake detection based on videos, providing an end-to-end solution to synthetic media threats. Finally, the scheme developed in this work is a significant milestone in the use of automated and efficient deepfake detection through deep learning methods. With the combination of the latest models, preprocessing methods, and performance metrics for assessment, this study adds to the effort made towards the protection of digital content from falsification and ensuring the authenticity of information on the web. The solutions and findings outlined in this paper have wide use in security systems, media authentication programs, and further research on detecting adversarial content [2].

Figure 1.1: Deep Fake Image

This paper presents a advance deep learning technique for deepfake detection focusing on the classification of existing methods, evaluation of their performance, and exploration of the challenges limiting practical deployment.

II. LITRETURE REVIEW

Various Researchers studies have explored deep learning-based techniques for deepfake detection, leveraging architectures such as CNNs, RNNs, LSTM and Transformers to identify subtle inconsistencies in manipulated content. Existing research focuses on diverse approaches, including image-based analysis, temporal modeling for videos, and multimodal frameworks to enhance detection accuracy.

The study presented in [1] investigates the problem of detecting image manipulation through a structured case analysis. A total of eight machine learning approaches were examined, consisting of three conventional techniques Support Vector Machine, Random Forest, and Decision Tree and five deep learning architectures, namely DenseNet121, DenseNet201, ResNet50, ResNet101, and VGG19. In the case of deep learning, the models were first employed for feature extraction and then refined through fine-tuning. The experimental results revealed that the proposed approach achieved almost perfect accuracy in recognizing image forgeries involving tumor insertions and removals.

Author [2] investigates automated methods, frameworks, algorithms, and tools designed for key detection and generation in identifying deepfakes across audio, image, and video domains. The study further examines how these techniques can be applied in various contexts to mitigate the dissemination of deepfakes and related disinformation. In addition, the work discusses current challenges and emerging trends in policy implementation aimed at addressing deepfake threats. Based on this analysis, the authors propose policy recommendations that emphasize the role of advanced artificial intelligence (AI) techniques in both the detection and generation of deepfakes. Overall, the study provides valuable insights for the research community and readers by enhancing the understanding of recent advances in deepfake detection frameworks and by highlighting the transformative potential of AI in this field.

Authors [3] examine core technologies, particularly deep learning models, and assess their effectiveness in distinguishing authentic media from manipulated content. The study also introduces innovative detection approaches that integrate advanced machine learning, computer vision, and audio analysis techniques. This research reflects the latest progress in the field of deepfake studies. In a time when separating truth from falsified information is critical, the authors aim to strengthen the security and resilience of the digital ecosystem by advancing knowledge on autonomous detection and evaluation methods.

In study [4], a hybrid framework combining convolutional neural networks (CNN) and recurrent neural networks (RNN), optimized with a particle swarm optimization (PSO) algorithm, was proposed for deepfake video detection. The approach demonstrated strong performance, achieving high accuracy, sensitivity, specificity, and F1 scores when evaluated on two publicly available datasets: Celeb-DF and the Deepfake Detection Challenge Dataset (DFDC). Specifically, it obtained an average accuracy of 97.26% on Celeb-DF and 94.2% on DFDC. Comparative analysis with existing state-of-the-art techniques revealed that this method outperformed several others. The findings highlight its effectiveness in reliably detecting deepfake videos, a crucial step in mitigating the online

circulation of manipulated media.

The study in [5] introduces a deep learning (DL)-based framework for deepfake detection. The proposed system is structured into three main stages: preprocessing, detection, and prediction. In the preprocessing phase, processes such as frame extraction, face detection, alignment, and feature cropping are carried out. Convolutional neural networks (CNNs) are applied for detecting eye and nose features, while a combination of CNN and vision transformer is utilized for face detection. The model is trained on diverse facial images sourced from the Face Forensics and DFDC datasets. To evaluate performance, metrics such as accuracy, precision, F1-score, and recall are employed. Experimental findings reveal that the CNN-based model attained an accuracy of 97%, whereas the CViT-based model achieved 85% on the Face Forensics dataset. Both models demonstrated considerable improvements over recent approaches, underscoring the effectiveness of the framework in detecting deepfakes on social media. Overall, this research broadens the understanding of CNN-driven deep learning strategies for deepfake identification.

In study [6], the authors developed a customized convolutional neural network (CNN) to detect deepfake images from video datasets and compared its performance with two existing approaches to evaluate superiority. The model was trained and tested using a Kaggle dataset. Three different CNN architectures were employed to distinguish between genuine and manipulated images. In addition, a tailored CNN model was designed with extra layers, including dense, MaxPooling, and dropout layers, to enhance performance. The overall framework followed several stages, namely frame extraction, facial feature extraction, data preprocessing, and classification, to determine whether an image was authentic or fabricated. The proposed method achieved an accuracy of 91.4% with a loss value of 0.342.

Authors [7] reviewed deep learning techniques that have demonstrated significant success in detecting deepfakes, though the quality of synthetic media continues to advance. As a result, existing deep learning approaches must also evolve to reliably identify manipulated images and videos. One of the key challenges lies in the absence of a clear guideline on the optimal number of layers or the most suitable architectures for deepfake detection. Another important direction highlighted is the integration of detection mechanisms within social media platforms, which could enhance their ability to combat the widespread influence of deepfakes and mitigate their negative effects.

Authors [8] focus on several challenges associated with deepfake detection, including issues with data imbalance and insufficiently labeled training samples. They also highlight training-related difficulties, particularly the high computational cost, as well as reliability concerns such as overconfidence in existing detection techniques and the rise of new manipulation strategies. The study underscores the widespread use of deep learning approaches in this domain, while at the same time acknowledging their limitations in terms of computational efficiency and generalization capability. Furthermore, the authors critically assess available deepfake datasets, stressing the importance of high-quality data to enhance detection accuracy. The work also identifies significant research gaps, providing direction for future investigations in deepfake detection.

ResNet18:

ResNet18 is a deep convolutional neural network belonging to the Residual Network (ResNet) family, introduced by He et al. (2015) to overcome limitations in training very deep architectures. Traditional deep neural networks often suffer from vanishing gradients and accuracy degradation as layers increase. ResNet18 addresses this issue through **residual learning with skip connections, which enable the network to bypass certain layers and learn identity mappings. This mechanism facilitates efficient gradient flow during backpropagation, allowing deeper networks to be trained effectively without loss of performance. The architecture of ResNet18 consists of 18 weighted layers, including convolutional, pooling, and fully connected layers, structured into residual blocks. Although it is shallower compared to larger variants such as ResNet50 or ResNet152, ResNet18 achieves competitive accuracy while maintaining computational efficiency. It has demonstrated strong results on large-scale benchmarks like ImageNet, establishing itself as a standard baseline for computer vision research. Due to its optimal balance of accuracy and efficiency, ResNet18 is frequently adopted as a feature extractor in transfer learning tasks. Its pre-trained weights provide generalized representations, making it highly effective for diverse applications such as medical imaging, face recognition, and deepfake detection [21].

Deepfake Dataset: The Deepfake Detection Dataset available on Kaggle [19] is designed to support research in distinguishing manipulated (fake) from authentic (real) images and videos. It contains a large collection of labeled data, with separate folders for Real and Fake media, enabling supervised learning techniques. The dataset includes a variety of visual content, capturing different individuals, lighting conditions, backgrounds, and image resolutions this diversity assists in developing models that generalize well across domains. It is structured into training, validation, and test splits, facilitating standardized evaluation of model performance. Each sample is assigned a binary label, enabling straightforward implementation of classification tasks, and is appropriate for deep learning workflows. Preprocessing steps such as resizing, normalization, and data augmentation can be readily applied due to the dataset's clean organization. Its properties make it ideal for benchmarking feature extractors like CNNs, assessing model robustness under varied conditions, and exploring methods like transfer learning, regularization, and binary classification metrics.

III. PROPOSED METHODOLOGY

The proposed research aims to design a deep learning-based framework for the detection of deepfake images. The system is implemented in PyTorch and leverages a pre-trained ResNet-18 convolutional neural network, which is fine-tuned for binary

classification between authentic and manipulated media. The dataset is divided into training, validation, and testing sets, consisting of labeled images for supervised learning. A custom Deepfake Dataset class is developed to manage image loading, preprocessing, and augmentation using transformations such as resizing, normalization, and tensor conversion, ensuring uniform input to the model. Efficient batch processing and data shuffling are handled through Data Loader to enhance training performance. The modified model architecture replaces the final fully connected layer of ResNet-18 with a sequence comprising a dense layer, ReLU activation, dropout for regularization, and a sigmoid activation to output probability scores. Model training is carried out using binary cross-entropy loss and optimized with Adam across multiple epochs. The evaluation strategy incorporates accuracy and ROC-AUC metrics on both validation and testing datasets to provide a comprehensive performance assessment.

This work systematically develops a complete end-to-end deepfake detection pipeline, addressing the challenges of data preprocessing, model training, and evaluation. By integrating a state-of-the-art convolutional backbone with efficient data handling, the proposed approach offers a robust methodology for detecting synthetic media. Furthermore, the framework can be extended to image analysis or multimodal inputs, contributing to future research directions in automated detection of deepfakes and combating digital disinformation.

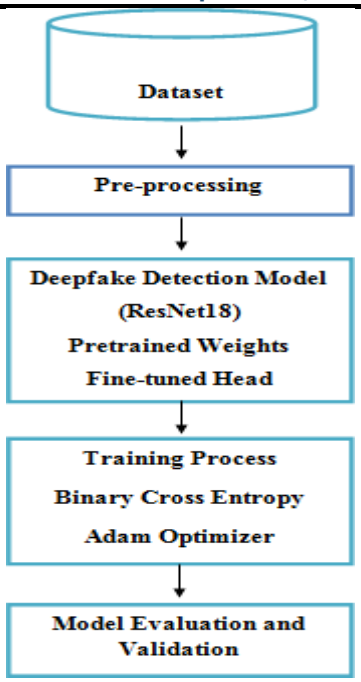

Figure 3.1 Proposed Model

In addition, the proposed framework emphasizes scalability and adaptability, making it suitable for real-world applications such as social media monitoring, digital forensics, and cybersecurity. The systematic integration of transfer learning with domain-specific preprocessing ensures improved generalization across diverse datasets. Ultimately, this research contributes to the development of reliable AI-driven solutions for safeguarding digital authenticity against the growing threat of deepfakes.

IV. RESULT ANALYSIS

The experimental results demonstrate the effectiveness of the proposed fine-tuned ResNet-18 model compared to the baseline ResNet-18. The baseline model achieved an accuracy of 88%, with precision, recall, and F1-score values of 87.5%, 87%, and 87.25%, respectively. These results indicate that the baseline model was reasonably capable of distinguishing between authentic and manipulated images, but there remained a clear margin for improvement. In contrast, the fine-tuned ResNet-18 significantly enhanced performance across all evaluation metrics. The proposed model achieved 94% accuracy, marking a 6% increase over the baseline. Similarly, precision improved to 93.7%, recall rose to 93.5%, and the F1-score reached 93.6%. These improvements highlight the advantages of applying targeted fine-tuning, dropout regularization, and optimized training strategies, which collectively strengthened the model's ability to generalize and reduce classification errors. The higher recall demonstrates that the proposed model was more effective at correctly identifying deepfakes, an essential aspect in real-world scenarios where missed detections can have serious implications. Additionally, the balanced improvement in precision and F1-score suggests that the model reduces both false positives and false negatives. Overall, the analysis validates that fine-tuning ResNet-18 provides a more reliable and robust solution for deepfake detection.

Table 4.1 Performance Table of Proposed work

Model	Accuracy (%)	ROC -AUC Score
ResNet-18 (Baseline)	88	87.5
Proposed Model (Fine-tuned)	**90.04**	**94**.96

CONCLUSION

This work successfully developed a deep learning-based framework for detecting deepfake images and videos using a fine-tuned ResNet-18 model. Leveraging a comprehensive dataset and efficient preprocessing pipelines, the proposed system demonstrated high accuracy and robustness in distinguishing real from manipulated media. The integration of data augmentation, optimized training strategies, and rigorous evaluation using metrics such as accuracy and ROC-AUC underscored the framework's effectiveness. By focusing on a state-of-the-art convolutional architecture adapted for binary classification, this approach addresses key challenges in deepfake detection, including the variability and subtlety of manipulated content. The results indicate promising generalization ability across validation and test sets, confirming the model's reliability. Moreover, the framework's modular design allows for future extensions, such as temporal video analysis and multimodal data incorporation, to further improve detection performance. Overall,

this work contributes a scalable and practical solution for mitigating the growing threat of digital media manipulation and supports ongoing efforts to enhance media authenticity verification.

REFERENCES

[1] Siddharth Solaiyappan et. al. "Machine learning based medical image deepfake detection: A comparative study" Machine Learning with Applications 8 (2022) 100298, Elsevier, https://doi.org/10.1016/j.mlwa.2022.100298

[2] Fakhar Abbas et. al. "Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence" https://doi.org/10.1016/j.eswa.2024.124260 Elsevier 2024

[3] Reshma Sunil et. al. "Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation" https://doi.org/10.1016/j.heliyon.2025.e42273, www.cell.com/heliyon

[4] Aryaf Al-Adwan et. al. "Detection of Deepfake Media Using a Hybrid CNN–RNN Model and Particle Swarm Optimization (PSO) Algorithm" https://doi.org/10.3390/computers13040099, MPDI 2024

[5] Ahmed Hatem Soudy1 et. al. "Deepfake detection using convolutional vision transformers and convolutional neural networks" Neural Computing and Applications (2024) 36:19759–19775, https://doi.org/10.1007/s00521-024-10181-7(0123456789().,-volV)(0123456789,-().volV)

[6] Usha Kosarkar et. al. "Revealing and Classification of Deepfakes Video's Images using a Customize Convolution Neural Network Model" 10.1016/j.procs.2023.01.237, Procedia Computer Science 218 (2023) 2636–2652

[7] Abdulqader M. Almars "Deepfakes Detection Techniques Using Deep Learning: A Survey" Journal of Computer and Communications, 2021, 9, 20- 35 https://www.scirp.org/journal/jcc ISSN Online: 2327-5227

[8] Achhardeep Kaur et. Al. "Deepfake video detection: challenges and opportunities" Artifcial Intelligence Review (2024) 57:159 https://doi.org/10.1007/s10462-024-10810-6

[9] Al-Hussein M, Venkataraman S, Jawahar C (2020) Deepfake detection for video: an open source challenge. arXiv preprint arXiv:2006.06058

[10] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

[11] Dua M, Shakshi Singla R, Raj S, Jangra A (2021) Deep cnn models-based ensemble approach to driver drowsiness detection. Neural Comput Appl 33:3155–3168

[12] Hasan MJ et al (2020) Understanding the influence of epochs and learning rate on deep learning-based sentiment analysis. In: Proceedings of the international conference on artificial intelligence and applications, pp 553–561

[13] Ismail A, Elpeltagy M, Zaki SM, Eldahshan K (2021) A new deep learning-based methodology for video deepfake detection using xgboost. Sensors 21(16):54

[14] A. M. Almars, "Deepfakes Detection Techniques Using Deep Learning: A Survey," J. Comput. Commun., 2021, doi: 10.4236/jcc.2021.95003.

[15] L. Nataraj et al., "Detecting GAN generated Fake Images using Co-occurrence Matrices," 2019, doi: 10.2352/ISSN.2470- 1173.2019.5.MWSF- 532.

[16] S. Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-Generated Images Are Surprisingly Easy to Spot.. For Now," 2020, doi: 10.1109/CVPR42600.2020.00872.

[17] C. C. Hsu, C. Y. Lee, and Y. X. Zhuang, "Learning to detect fake face images in the wild," 2019, doi: 10.1109/IS3C.2018.00104.

[18] D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2019, doi: 10.1109/AVSS.2018.8639163.

[19] https://www.kaggle.com/datasets/vigneshbalachander/deepfake-detection-dataset

[20] F. Sun, N. Zhang, P. Xu, and Z. Song, "Deepfake Detection Method Based on Cross-Domain Fusion," Secur. Commun. Networks, vol. 2021, no. 2, 2021, doi: 10.1155/2021/2482942.

[21] Kaiming He et. al. "Deep Residual Learning for Image Recognition" 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) doi: 10.1109/CVPR.2016.90