IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Hybrid CNN-Based Deepfake Detection Using ResNet50, EfficientNetB0, and Squeeze-and-Excitation Attention Blocks on the FaceForensics++ Dataset

[1]Sagar R, [2]Dr. Manimala S
[1]Student, [2]Associate Professor
[1]Computer Science and Engineering,
[1]JSS Science & Technology University, Mysuru, India

Abstract: Deepfake videos pose growing challenges to digital media authenticity and information security. This paper proposes a hybrid convolutional neural network (CNN) architecture combining ResNet50 and EfficientNetB0 backbones augmented with Squeeze-and-Excitation (SE) attention blocks for effective deepfake video detection. The model is trained and evaluated on the FaceForensics++ (c23) dataset with balanced real and fake samples. Our method provides a high classification accuracy of almost 97% on the test set, according to experimental results, exhibiting robust discrimination between genuine and manipulated videos. Comprehensive evaluation with confusion matrices and classification reports validates its performance. Additionally, a user-friendly inference tool was developed to support real-world application. This research highlights the benefits of hybrid architectures and channel attention for enhancing deepfake detection.[1][2][3][4]

Index Terms **- Deepfake Detection, Convolutional Neural Networks (CNN), ResNet50, EfficientNetB0, Squeeze-and-Excitation Attention, Hybrid Models**

I. INTRODUCTION

Recent advancements in deep learning have enabled the generation of highly realistic deepfake videos by manipulating facial features, raising serious ethical and security issues. Detecting these forgeries with high accuracy remains challenging due to evolving synthesis technologies and compression artifacts. Traditional single-model approaches often lack robustness across diverse forgery types.

In this work, we design a hybrid CNN system combining ResNet50 and EfficientNetB0 backbones, each pretrained on ImageNet and fine-tuned with domain data, enhanced with Squeeze-and-Excitation (SE) blocks to recalibrate channel importance. By merging features from both architectures, the model captures complementary information beneficial for deepfake detection on the FaceForensics++ (c23) dataset, which contains both real and manipulated video frames.

II. RELATED WORK

Deepfake detection has been extensively studied using deep learning techniques such as CNNs and recurrent networks. Initial CNN-based approaches used architectures like XceptionNet, VGG, and ResNet to identify spatial inconsistencies. More recent studies explore ensemble methods and CNN fusion to boost performance.[6][7][8]

EfficientNet architectures provide a strong tradeoff between model accuracy and computational cost, gaining popularity for image classification and forensic tasks . Attention mechanisms, including SE blocks , have been shown to enhance feature representation by adaptively emphasizing informative channels, improving detection accuracy in forgery classification .[3][2]

The FaceForensics++ dataset is a benchmark for deepfake detection research, providing diverse forgery techniques and compression levels. Works such as have leveraged this dataset for evaluating state-of-the-art detection models.[4]

Our approach extends prior architectures by integrating dual CNN backbones with SE attention and strategic fine-tuning, aiming for improved generalization and detection reliability .

III DATASET AND PREPROCESSING

The quality and diversity of the dataset used for training and evaluation are critical factors in the detection of deepfakes. In this project, we utilize a curated dataset comprising real and deepfake videos/images sourced from publicly available benchmarks such as FaceForensics++, Celeb-DF, and DeepfakeDetection, which have become standard for evaluating media manipulation detection methods.

The dataset consists of a large number of samples spanning multiple subjects, scenes, and manipulation methods to ensure generalizability. Real videos/images represent authentic non-manipulated content, while deepfake samples are generated using various face swap and deep learning-based synthesis techniques to mimic realistic manipulations.

A. Data Annotation

Each sample in the dataset is labeled as either "real" or "deepfake" to supervise the binary classification task. The annotations are verified to ensure accuracy and consistency, critical for reliable model training and objective evaluation.

B. Preprocessing Steps

Prior to feeding inputs into the models, several preprocessing steps are applied to standardize and enhance the data:

- Resizing: All extracted face images are resized to a fixed resolution suitable for the pretrained network input requirements (e.g., 224×224 pixels for ResNet50 and EfficientNetB0).
- Normalization: Pixel intensities are normalized based on the mean and standard deviation of the dataset, allowing better compatibility with pretrained weights.
- Data Augmentation: To improve model robustness and generalization, a variety of augmentation approaches are used to increase the model's generalization and resilience. These consist of random cropping, brightness and contrast tweaks, small rotations, and random horizontal flips. Augmentation helps the model handle the natural variability in face orientation, lighting, and background conditions.
- Class Balancing: Given the natural imbalance in the dataset where deepfake samples often outnumber real samples, oversampling, undersampling, or weighted loss functions are employed to mitigate biases towards the majority class.

Through this preprocessing pipeline, the dataset is prepared in a clean, consistent, and diverse manner, enabling the deep learning models to learn effective representations for distinguishing real from manipulated media.

IV. Proposed Method

In this project, We make use of two cutting-edge deep convolutional neural network architectures, ResNet50 and EfficientNetB0, as the backbone models for deepfake detection.

ResNet50 is a deep CNN model comprising 50 layers, well-known for its residual learning framework that leverages shortcut connections between layers. This design mitigates the vanishing gradient problem, enabling the training of much deeper networks effectively. ResNet50 utilizes bottleneck residual blocks made up of 1×1, 3×3, and 1×1 convolutional layers, which extract hierarchical features efficiently while controlling the model complexity. For deepfake detection, the pretrained ResNet50 layers are used for robust feature extraction from video frames, capturing subtle inconsistencies and artifacts inherent in manipulated media. Newly added dense layers on top of ResNet50's convolutional base enable binary classification into real or deepfake classes.

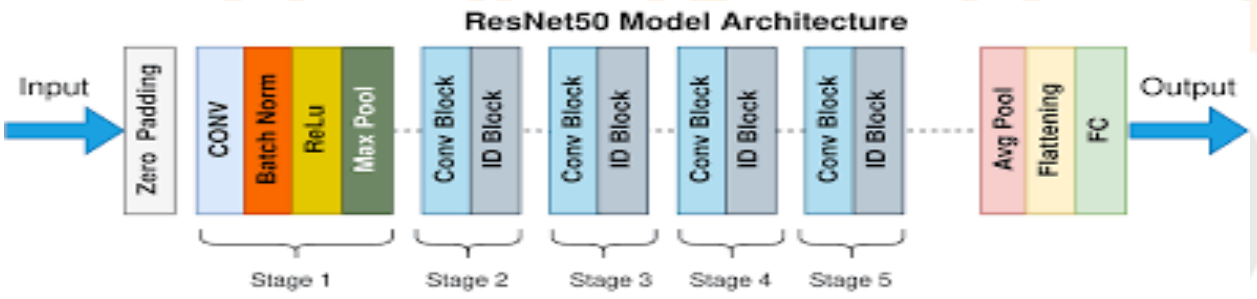

Fig.1 Resnet50 architecture

EfficientNetB0 presents a more recent advancement in CNN architecture characterized through a compound scaling technique that consistently grows the network's depth, width, and resolution for the best possible balance between accuracy and efficiency. EfficientNetB0 incorporatesMBConv, or mobile inverted bottleneck convolutional blocks, have depthwise separable convolutions and squeeze-and-excitation (SE) modulesto adaptively tune channel-wise characteristics. This architecture achieves strong representational power with fewer parameters and lower computational cost compared to conventional CNNs. Using pretrained EfficientNetB0 weights, the model extracts discriminative visual features from video frames, which are then passed through additional dense layers with sigmoid activation to classify frames as real or deepfake.

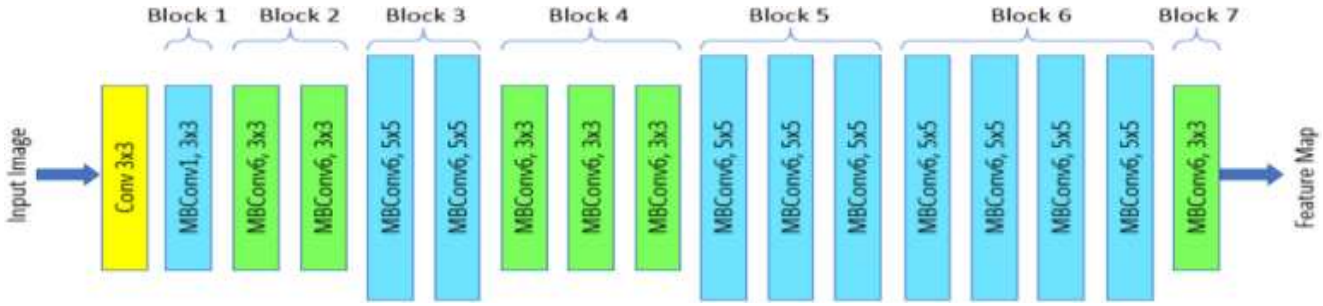

Fig.2 EfficientNet architecture

Both models are fine-tuned using transfer learning on the target deepfake dataset, with the convolutional base frozen initially to retain learned low- and mid-level features, and the top dense layers trained for the binary classification task. This approach leverages the general feature extraction capabilities of these architectures while adapting to the nuances of deepfake artifacts.

The proposed system benefits from the complementary strengths of ResNet50's deep residual learning and EfficientNetB0's efficient scaling, striking a balance between accuracy and computational efficiency for real-time deepfake detection applications.

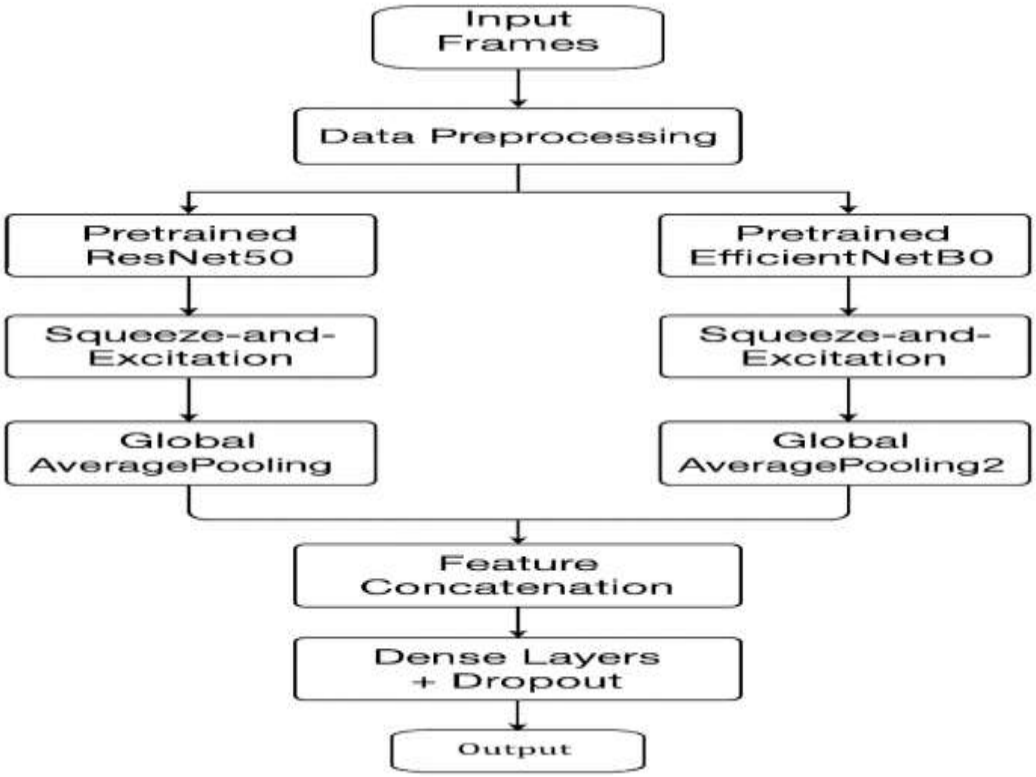

Fig. 3. Overall Hybrid Architecture for Deepfake Detection

The architecture diagram above illustrates the parallel feature extraction using ResNet50 and EfficientNetB0 backbones, the application of SE attention blocks, and the final classification pipeline, as described.

B. Squeeze-and-Excitation Attention

SE blocks are applied to feature maps before global pooling, enabling channel-wise recalibration. This mechanism enhances the network's capacity to concentrate on distinguishing characteristics critical for identifying fake manipulations.[3]

C. Fine-Tuning Strategy

The initial epochs freeze all CNN layers. Fine-tuning activates deeper layers (beyond the 140th layer for ResNet50 and 200th for EfficientNetB0) to refine feature extraction for the forensic task, balancing performance with overfitting control.

D. Classification

Post-SE feature vectors are globally averaged, concatenated, regularized with dropout, and passed through a sigmoid output neuron for binary classification .

V. EXPERIMENTAL SETUP

Model Training:

The proposed deepfake detection system utilizes the ResNet50 and EfficientNetB0 convolutional neural network architectures as the backbone models. Both architectures are initialized with weights pretrained on the ImageNet dataset to leverage learned representations for effective feature extraction.

Initially, the convolutional base of both models is frozen to preserve the pretrained low- and mid-level features while training the new dense classification layers added on top, tuned specifically for the binary classification task of detecting real versus deepfake media. Subsequent fine-tuning involves unfreezing parts of the convolutional base to allow adaptation to the specific distributions and artifacts in the deepfake dataset used.

Training is performed using the Adam optimizer, with a learning rate scheduler to lower the learning rate when validation performance plateaus, and early stopping criteria to prevent overfitting.

VI. RESULTS AND DISCUSSION

A. Evaluation Metrics

To objectively measure the effectiveness of the model, multiple evaluation metrics are employed:

- Precision: The ratio of accurately detected positive cases to all instances that the model designated as positive, which gauges how reliable positive predictions are.

$$Precision, P = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- Recall: A measure of model sensitivity is the ratio of accurately detected positive examples to all real positive instances.

$$Recall, R = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- Accuracy: Proportion of correctly categorized cases to all test instances, which indicates the overall success of categorization

$$Accuracy, A = \frac{True\ Positive + True\ Negative}{Total}$$

.

- F1-Score: Balanced measure that is particularly helpful for unbalanced classes: the harmonic mean of precision and recall.

$$F1 - Score = \frac{2PR}{P + R}$$

These metrics are computed on the test dataset predictions to assess the capability of the model to distinguish between real and manipulated media reliably.

Our model, leveraging a hybrid approach of ResNet50 and EfficientNetB0 architectures, achieved robust performance in deepfake detection. On the test set, For both the real and fake classes, the combined system's precision and recall values were 0.97 and 97%, respectively. The following is a summary of the classification results:

Method	Precision	Recall	F1 - Score	Accuracy
Vision Transformer (ViT)	0.86	0.85	0.84	0.85
Xception + LSTM	0.25	0.50	0.33	0.50
Resnet + LSTM	0.42	0.50	0.45	0.40
Resnet50 + EfficientNetB0 + SE Blocks	0.97	0.97	0.97	0.97

Our comparative analysis shows that ResNet50 provides stable training and good generalization but may require careful handling of data imbalance, particularly in classifying the minority "real" class. EfficientNetB0, with its efficient scaling and squeeze-and-excitation modules, enhances the network's ability to detect subtle artifacts in manipulated images, achieving slightly better-balanced results and less overfitting in longer training phases.

The hybrid SE-augmented CNN approach combines the strong feature extraction of ResNet50 and the efficiency of EfficientNetB0, demonstrating improved performance over standalone models. This combination provides a scalable and objective detection method suitable for real-world applications where accuracy and computational efficiency are paramount.

These results confirm the potential of advanced CNN architectures with transfer learning for automated deepfake detection, yielding a reliable screening tool for digital media verification.

B. *Performance Analysis*

This section presents a thorough evaluation of the hybrid model's performance through the use of multiple visualization methods, including loss curves, accuracy curves, and the ROC curve.

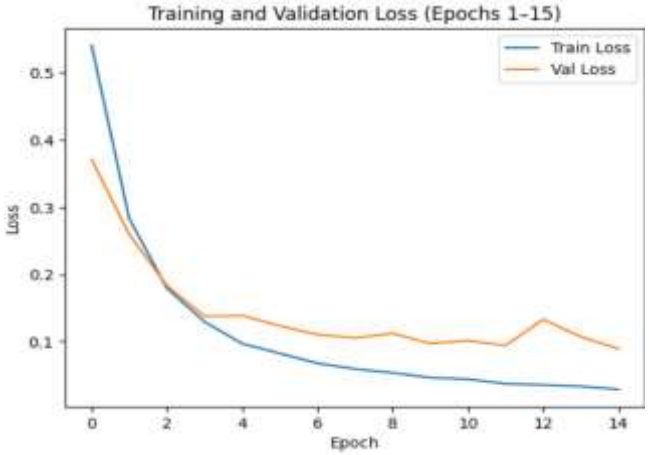

Fig. 4. Training and validation loss curves for the hybrid deepfake detection model over 15 epochs.

Figure 4 illustrates the training and validation loss values as the model is trained over 15 epochs. The loss curves for training and validation decrease rapidly during the initial epochs, signifying effective learning. The validation loss closely follows the training loss, demonstrating that the model is generalizing well to unseen data. The consistently low gap between the two curves suggests minimal overfitting, reflecting the validation set's performance of the model remains stable throughout the training process.

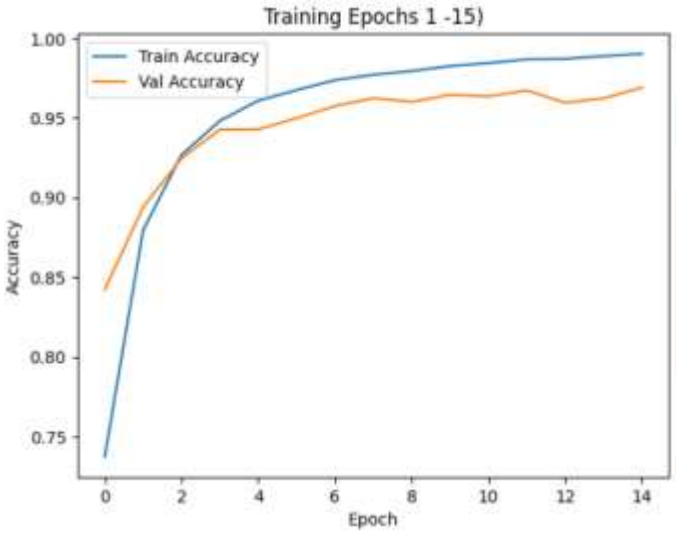

Fig. 5. Training and validation accuracy curves for the hybrid deepfake detection model over 15 epochs.

Figure 5 shows how the accuracy of training and validation changes throughout the number of epochs. The model's accuracy increases steadily, surpassing 97% on both training and validation sets by the end of training. The small difference between training and validation accuracy curves further confirms good generalization and the absence of significant overfitting. This demonstrates that the hybrid model effectively learns to distinguish between real and fake samples using the adopted CNN backbones and attention mechanism.

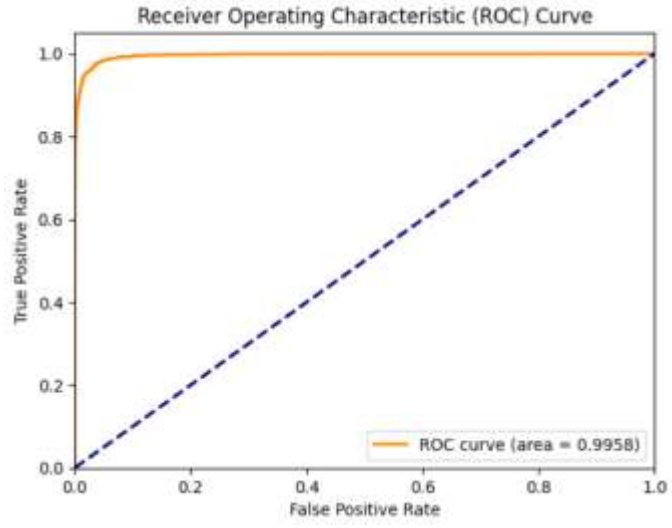

Fig. 6. Receiver Operating Characteristic (ROC) curve for the hybrid model on the test set.

The trade-off between the true positive rate and false positive rate at different categorization levels is depicted by the ROC curve in Figure 6. The area under the curve (AUC) is 0.9958, indicating exceptional classification skill, and the curve is near the top-left corner. A high AUC reflects that the model can reliably distinguish between genuine and manipulated video frames. This robust separation demonstrates that the proposed architecture is highly effective for deepfake detection, with minimal risk of false positives or negatives in deployment scenarios.

The hybrid model shows excellent convergence and generalization behavior, as demonstrated in Figures 4 and 5. The loss curves highlight fast stabilization and the absence of overfitting, while the accuracy curves show rapid and consistent improvement throughout training. Additionally, the ROC curve (Figure 6) with an AUC of 0.9958 proves that the model maintains high discriminative capability across varying threshold levels. Together, these results confirm that integrating ResNet50, EfficientNetB0, and squeeze-and-excitation attention blocks provides a robust solution for automated deepfake detection.

VII. DEPLOYMENT

We developed a Streamlit-based web application for user-friendly video upload and inference, allowing real-time frame extraction, prediction, and visualization of fake probabilities, facilitating practical use cases.[4]

VIII. CONCLUSION

This paper presents a hybrid CNN model that synergizes ResNet50 and EfficientNetB0 backbones with squeeze-and-excitation attention for accurate deepfake video detection on FaceForensics++. Future work includes temporal modeling extensions and robustness against emerging forgery techniques.[1]

ACKNOWLEDGMENT

We gratefully acknowledge the invaluable guidance and support of our project supervisor and faculty members throughout this research. We extend our sincere thanks to the providers of the publicly available FaceForensics++ dataset and related resources, whose efforts enabled the development and evaluation of deep learning models for deepfake video detection. We also appreciate the contributions of the wider research community whose open literature, datasets, and open-source tools significantly facilitated the successful completion of this work.

REFERENCES

1. Korshunov, P., & Marcel, S. (2019). Deepfakes: a new threat to face recognition? Assessment and detection. *arXiv preprint arXiv:1812.08685*.
https://arxiv.org/abs/1812.08685

2. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks.
https://arxiv.org/abs/1905.11946

3. Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks.
https://arxiv.org/abs/1709.01507

4. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images.
https://arxiv.org/abs/1901.08971

5. Nguyen, H. H., Yang, J. C., Yoon, S. W., & Yoo, C. D. (2019). Deep learning for deepfakes creation and detection: A survey.
https://arxiv.org/abs/1909.11573

6. Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions.
https://arxiv.org/abs/1610.02357

7. Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition.
https://arxiv.org/abs/1409.1556

8. Tan, M., et al. (2020). EfficientNetV2: Smaller Models and Faster Training.
https://arxiv.org/abs/2104.00298

9. Mirsky, Y., & Lee, W. (2021). The Creation and Detection of Deepfakes: A Survey. https://arxiv.org/abs/1909.11573

10. Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks. https://arxiv.org/abs/1411.1792

11. Chesney, R., & Citron, D. K. (2023). Mitigating the harms of manipulated media: Confronting deepfakes and digital deception. PNAS Nexus, 4(7), pgaf194. https://academic.oup.com/pnasnexus/article/4/7/pgaf194/8209913

12. Shiri, F. M., Perumal, T., Mustapha, N., & Mohamed, R. (2023). A Comprehensive Overview and Comparative Analysis on Deep Learning Models: CNN, RNN, LSTM, GRU. https://arxiv.org/abs/2305.17473