Comparative Analysis of Deep Learning Models for DeepFake Video Detection on the FaceForensics++ Dataset

Ms. Anusha N
Dept. of ISE
NMAMIT, Nitte, India
anusha24@nitte.edu.in

Namratha M
Dept. of ISE
NMAMIT, Nitte, India
namrathaalevuraya2005@gmail.com

M Nagashree Pai
Dept. of ISE
NMAMIT, Nitte, India
mnagashreepai@gmail.com

Ridhi S Kuntady
Dept. of ISE
NMAMIT, Nitte, India
ridhisk004@gmail.com

***Abstract*—With the fast development of AI-generated content, DeepFake videos have become a major digital authenticity and public trust threat. This project addresses DeepFake detection with deep convolutional neural networks trained on the FaceForensics++ dataset. We assess four leading CNN-based architectures—XceptionNet, EfficientNet-B3, ResNet-50, and MesoNet (Meso4)—utilizing transfer learning methods to fine-tune pretrained ImageNet weights for binary classifying real vs. fake facial video frames. The models are trained and tested on compressed (C23) versions of the dataset to mimic real-world social media conditions. Performance is measured in terms of standard metrics such as accuracy, precision, recall, and F1-score. Our experimental results indicate that XceptionNet is more accurate and performs better in feature localization than other models, whereas MesoNet is lightweight and efficient to deploy in practical applications. This paper emphasizes the potency of transfer learning using CNN in detecting DeepFakes and highlights the need for secure forensic methods in media analysis.**

***Index Terms*—DeepFake Detection, FaceForensics++, Deep Learning, Video Forensics, Model Comparison**

I. Introduction

The rise of DeepFake technology, driven by powerful generative models such as Generative Adversarial Networks (GANs) and autoencoders, has made it easier to alter human faces in videos with high visual quality. Although these methods allow creative uses in entertainment and education, they also bring significant risks, such as the spread of misinformation, political propaganda, identity theft, and a decrease in trust in digital content [1], [2].

DeepFake videos can change facial expressions, speech, and mannerisms in a way that often goes unnoticed. As these fakes become more sophisticated, manual checks are no longer a reliable defense. This has increased the need for automated DeepFake detection systems that are both precise and efficient. The goal of this research is to develop and test such systems using modern deep learning methods [3], [4].

The **FaceForensics++** dataset [5] serves as the main resource for training and assessing DeepFake detection systems. The FaceForensics++ dataset contains thousands of genuine videos along with four distinct modification methods, including FaceSwap and DeepFakes applied at multiple compression levels. FaceForensics++ provides an excellent simulation of authentic scenarios through its diverse and realistic content which combines compression errors and subtle manipulation traces. We extract frames from these videos to input into our models, following a consistent preprocessing process.

In this paper, we present a comparative study of four convolutional neural network (CNN) architectures for DeepFake detection in video data: *XceptionNet*, *EfficientNet-B3*, *ResNet-50*, and *MesoNet (Meso-4)*. These models were chosen for their contrasting architectures and established success in visual recognition and forgery detection tasks. XceptionNet and ResNet-50 represent deep high-capacity models which have the ability to detect detailed spatial features. The design of EfficientNet-B3 incorporates compound scaling to optimize both computational efficiency and accuracy [6] and MesoNet functions as a lightweight model for real-time DeepFake detection on facial video data [7].

The work delivers three principal contributions by these components: (i) We develop and optimize four separate CNN models for the Face-Forensics++ dataset; (ii) we test and evaluate their performance across different compression levels; and (iii) we analyze the balance between accuracy and computational cost to guide real-world DeepFake detection system deployment in mobile and embedded devices.

II. Related Work / Literature Review

Current developments in deepfake detection have been led prominently by convolutional neural networks (CNNs) like ResNet50, EfficientNet (particularly B3), and XceptionNet. These models have produced excellent performance using standard datasets such as FaceForensics++, DFDC, and Celeb-DF with the high representational capacity and flexibility of these models. Methods involving the fine-tuning of pretrained CNNs on deepfake datasets were shown to improve drastically in accuracy and robustness [19], while others investigated architectural simplifications in order to facilitate real-time inference on edge devices without sacrificing classification effectiveness [10][13][18].

Ensemble and hybrid techniques have also proved to be useful approaches. Hybrid models that ensemble ResNet50 and XceptionNet using transfer learning performed better by combining the capabilities of the two backbones [9]. Attention mechanism added to CNN architectures like XceptionNet helped in concentrating on the forged facial areas and making the model more interpretable [14]. Feature fusion methods with EfficientNetB3 also assisted in detecting minute facial abnormalities, improving both precision and recall [16].

In addition to traditional face-centered detection, research pushed CNN utility to other biometric modalities such as palmprints [15] and even simulated counterfeit medical images [12], with promising generalizability. Further, a few works solved cross-dataset generalizability by boosting frequency domain features, which strengthened robustness against compression and lighting degradations [8]. Comparisons among ResNet50, EfficientNet, and XceptionNet identified trade-offs in performance between accuracy, speed, and resolution management [11][22], calling for careful model selection based on task-specific requirements.

Furthermore, studies contrasting CNNs with vision transformers emphasized that although transformers might excel in high-data situations, tuned CNNs like ResNet50 remain superior when training data are scarce [21]. Other models also fine-tuned learning by optimizing loss functions to address class imbalance, enhancing generalization to minority deepfake classes [20]. Overall, these efforts demonstrate that one best-performing architecture is not superior across all settings, affirming the merits of hybrid, adjustable, and task-specific detection approaches.

III. Methodology

This section explains the methods employed in assembling the dataset, deploying and preprocessing the deep learning models, and determining the parameters for the training model. To facilitate fair and unbiased comparison of the selected models: XceptionNet, EfficientNet-B3, ResNet-50, and MesoNet (Meso-4)—a standard and consistent pipeline was adopted throughout the experiments. Each model was trained and evaluated on the FaceForensics++ dataset, which is the standardized dataset for DeepFake detection work [5].

Uniform preprocessing techniques and data augmentations were applied to all input frames to ensure consistency before applying architecture-specific preprocessing steps. The model architecture has been initialized with pretrained ImageNet weights, and then fine-tuned on our task-specific data. In order to prevent overfitting, the training process was also standardized using common loss functions, optimizers, and learning rate schedules. Early stopping and checkpoint mechanisms were also adopted. The same hardware configuration was used to conduct all experiments to guarantee a fair comparison between model performance.

A. Dataset Description

We have utilized the widely used benchmark dataset FaceForensics++ [5]. This dataset includes 1,000 high-quality, manipulated videos produced by a variety of face manipulation methods obtained from YouTube. The four face manipulation techniques used were DeepFakes, FaceSwap, Face2Face, and NeuralTextures, each of which had different compression levels (C0, C23, and C40).

For all experiments, we used the C23 compression version of the dataset of light H.264 compression, close to the social media-level compression. To keep dataset consistency, a fixed number of frames were collected from all the videos, so that all the videos contributed equally during training and minimized potential data unbalance. This curated subset, in combination with fixed sampling methodology across models, helped to simplify our pipeline as well as establishing a close to equal comparison of performances between models.

B. Data Preprocessing

In order to process the dataset for model training and evaluation, an organized preprocessing pipeline was developed to obtain reliable facial inputs from each video. For each video in our dataset, a fixed number of frames were extracted at regular intervals from every video. This provides temporal diversity with controlled redundancy. All video frames were detected using Multi-task Cascaded Convolutional Networks (MTCNN) that produces accurate facial bounding boxes by using a three-stage CNN cascade for proposal, refinement, and output [8].

In a frame, once a face was detected, the bounding box was cropped and resized, which was different in the models. To maintain a consistent comparison, one standing-out face per frame was selected and normalized per video. When a face was not found in a frame, that frame was omitted without replacement. This means it does not bring any noise or misdetections into the training data.

All images were transformed from BGR to RGB color space and normalized through the model-agnostic data preprocessing. If necessary, model-specific normalization, for example, scaling the pixel values to $[0, 1]$ or $[-1, 1]$, was performed during training using standard preprocessing functions provided by the deep learning libraries of choice.

This preprocessing pipeline not only provided high-quality, time-equivalent, and spatial-normalized face-to-face stimuli for all four models (EfficientNet-B3, XceptionNet, MesoNet, and ResNet), but also allowed us to conduct fair and accurate comparisons when compared to the existing work for detecting DeepFake.

C. Model Architectures

In order to evaluate the performance of multiple CNN architectures for DeepFake detection, we implemented and fine-tuned four different models—XceptionNet, MesoNet (Meso4), EfficientNet-B3, and ResNet-50. The models range from lightweight networks suitable for real-time operation to deep, high-capacity architectures for hierarchical feature extraction. For all the architecture, binary classification was done by learning to classify each facial frame if they are real or fake.

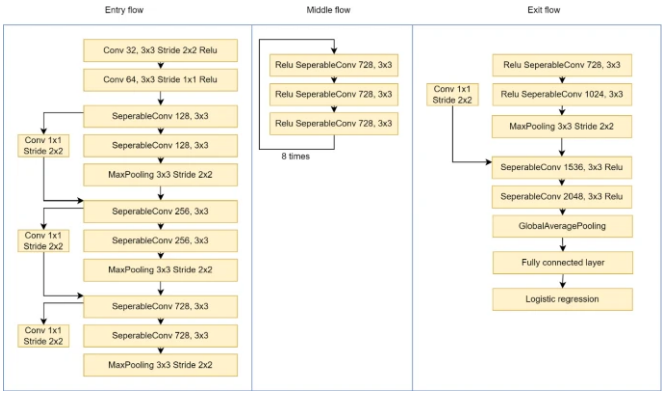

Fig. 1: XceptionNet Architecture.

The Xception model utilizes three key components: entry flow, middle flow, and exit flow. The entry flow processes input images with a size of 299×299×3 through a series of standard convolutional layers, followed by depthwise separable convolutions, and when it was completed, max pooling was used to decrease output resolution while retaining relevant features. The middle flow consists of a repeated part of eight modules, and each module has three depth-wise separable convolutions (728 filters), which is activated by ReLU, in order to capture complex image representations. The exit flow expands the output features by layers of separable convolutions with filters 1024, 1536, and 2048, and when completed, the output is passed through global average pooling followed by a dense, sigmoid-activated layer used for binary classification. Use of depthwise separable convolution layers drastically decreased the amount of computation, while remaining accurate in classification. Initially, the model was initialized using weights from pretrained models based on ImageNet, and batch normalization layers were frozen in fine-tuning training to increase training safety, which is of utmost importance in the detection of Deepfake [9].

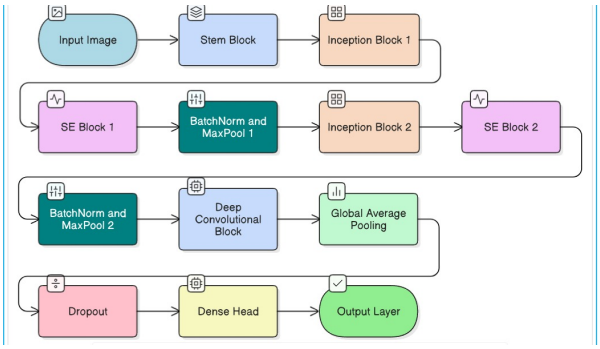

Fig. 2: MesoNet (Meso-4) Architecture.

The Meso-4 (MesoNet) model [7] is a lightweight convolutional neural network, for Deep-Fake detection that focuses on mesoscopic details to detect minor artifacts indicative of facial manipulation. The model passes in RGB 256 × 256 × 3 resolution images through four sequential convolutional blocks which increases the count of filters from 8 up to 16, and uses both 3×3 and 5×5 kernels, followed by batch normalization, ReLU activation, and max-pooling layers to decrease spatial size without losing discriminative information. Once the feature extraction is performed, the model implements a fully connected stage where the flattened rows of feature maps are passed through the dense layers, which include dropout and batch normalization, prior to arriving at the output single sigmoid neuron which makes a single estimate probability regarding whether the frame is indeed a Deepfake or not. The architecture is capable of achieving an effective detection ability due not just simplicity of architecture, but use of regularization, and mid-level patterns which have ample strength even with typical video compression.

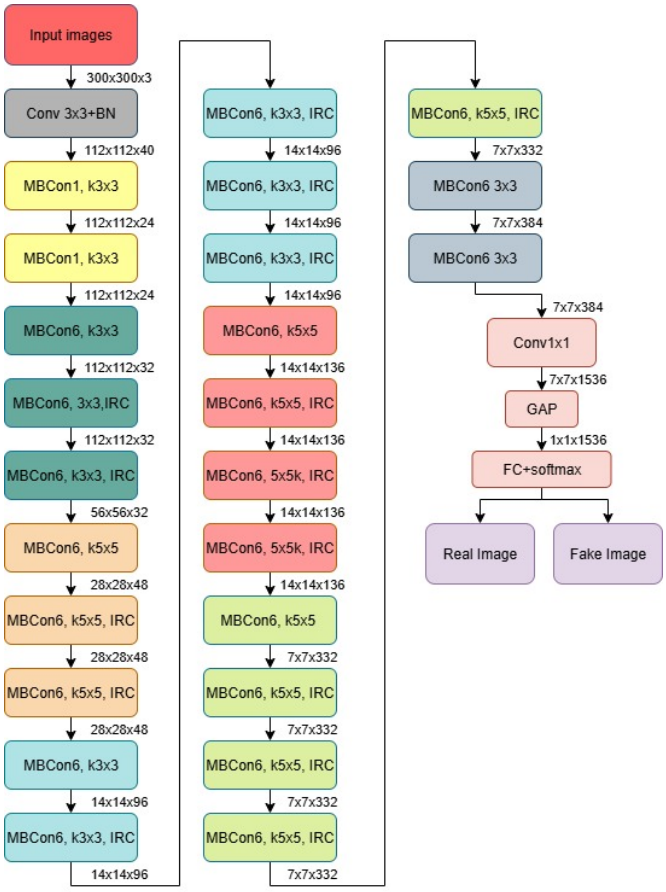

Fig. 3: EfficientNet-B3 Architecture.

EfficientNet-B3, generated by Google's EfficientNet family, contains a compound scaling technique in which we uniformly scale depth, width, and resolution, so it would be beneficial for Deepfake detection tasks requiring a trade-off between accuracy and efficiency. The model has about 12 million parameters and an input size of 300×300 pixels, tackled for quite an efficient price. It uses the Swish activation function and adds Batch Normalization after each convolutional layer to stabilize and accelerate training. EfficientNet-B3 comprises roughly 48 layers and implements MBConv (Mobile Inverted Bottleneck)

blocks with introduced squeeze-and-excitation (SE) modules to boost the feature representation while reducing the cost of understanding the hierarchies of a face. The model is initialized with ImageNet pre-trained weights, and is trained for binary Deepfake classification after removing the top layer and laying on a GlobalAveragePooling2D, some BatchNormalization, some Dropout for reduction of overfitting, and a final Dense layer with sigmoid activation for the binary output [6].

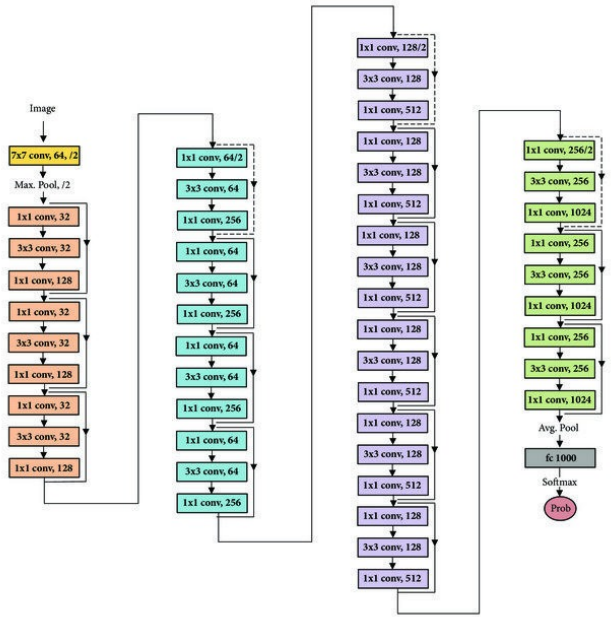

Fig. 4: ResNet50 Architecture.

ResNet50 is a deep convolutional neural network based on residual learning, introduced by He et al. [10]. It solves the degradation problem in deep networks by using identity shortcut connections that allow gradients to be propagated directly to lower layers, making training more efficient. ResNet50 consists of 25.6 million parameters and uses the standard input size of 224×224 pixels, which means that depth and computational cost is balanced. The model features 50 layers grouped into five stages, with convolutional and identity block's using 1×1, 3×3, and 1×1 convolutions, batch normalization, and ReLU activations. Residual connections have beneficial results to faster convergence, as well as better performance results when training. In our implementation for DeepFake detection, we remove the pre-trained ImageNet top layers and add BatchNormalization, Dropout and a sigmoid-activated dense layer for binary classification. ResNet50 is suitable for our task since it can learn patterns on a deep semantic level and learn on a texture-level. It makes a high-performing and robust backbone for identifying manipulated media.

D. Training Strategy

All models were trained to do binary classification in the FaceForensics++ dataset (C23 compression) [5]. A fixed number of frames were extracted from each video (10 frames for XceptionNet and ResNet-50, 15 frames for EfficientNet-B3 and Meso4). Each facial region was accurately located and cropped using MTCNN [11] for a consistent arrangement. Each face crop was resized for each model's requirements: XceptionNet = 299×299, ResNet-50 = 224×224, EfficientNet-B3 = 300×300, Meso4 = 256×256. The data was divided 80:20 for training and validation purposes. To improve generalization, we performed data augmentation using horizontal flips, ±15° rotation, ±10% zoom, and shift. Training used binary cross entropy loss with class weights to account for the minor class imbalance.

All models were fine-tuned with Adam or AdamW optimizers with a base learning rate of 1e-5, using ReduceLROnPlateau to reduce the learning rate when learning was stagnated in validation loss. To reduce memory usage and train faster, mixed-precision training (mixed_float16) was implemented. EarlyStopping, ModelCheckpoint and resume from checkpoint were used to track model convergence in an optimal and stagely manner.

The XceptionNet model [9] used pretrained ImageNet weights for initialization while a custom classification head (GlobalAveragePooling, BatchNormalization, Dropout, Dense with sigmoid) was added. The fine-tuning process kept BatchNormalization layers fixed. The model reached its highest validation accuracy of 95.75% along with excellent AUC performance of 0.984 at epoch 10. EfficientNet-B3 [6] used the same transfer learning approach for its training under mixed-precision conditions with AdamW optimizer. The model reached its highest validation accuracy of 92.29% together with an AUC of 0.9779 during epoch 11 which indicated excellent generalization.

The pretrained base of ResNet-50 [10] was used with only 30 layers unfrozen except BatchNormalization layers while the custom head included multiple dense and dropout layers. The training process used a batch size of 32 to achieve validation accuracy of 92.45% and AUC of 0.979. Meso4 [7] came with a lightweight architecture for forgery detection which was trained from scratch at an initial learning rate of 1e-4. The small number of parameters along with early stopping enabled this model to perform reliably with stable convergence even though it was compact.

E. Evaluation Metrics

We used a set of standard classification metrics to evaluate the performance of DeepFake detection models. The primary metric we used was accuracy, or the percentage of images correctly classified versus all images classified. However, given the potential for class imbalance in real versus fake distributions, we also considered, meaningfully, other metrics for a more expansive performance comparison.

Precision or the positive predictive value was used to measure the percentage of correctly predicted fake samples

out of all samples predicted as fake. On the other hand, Recall or the true positive rate quantifies the proportion of correctly predicted fake samples among all samples projected as fake. The F1-score, harmonic mean of precision and recall was computed for a comprehensive assessment, which is essential when handling unbalanced classifications [12].

We used the Receiver Operating Characteristic (ROC) curve and the associated Area Under the Curve (AUC-ROC) to assess the model's ability to identify real and fake classes at different threshold levels. Greater AUC values indicate the model's ability to make more accurate distinctions [13]. The confusion matrices for each model was visualized to give a better understanding of model-specific classification behavior. The metric we used allowed for a comparative analysis of the detection capabilities of the four different CNN architectures.

IV. Experimental Results

TABLE I: Performance Comparison of DeepFake Detection Models

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
XceptionNet	95.25	93.63	97.10	95.33	99.33
EfficientNet-B3	93.83	93.05	94.73	93.88	98.91
ResNet50	93.45	94.65	92.09	93.35	98.61
Meso-4	88.57	89.92	86.87	88.37	96.37

Based on the performance measures presented, XceptionNet is the highest performing model in all parameters tested. It has the highest accuracy (95.25), recall (97.10), F1 score (95.33), and ROC AUC (99.33), indicating outstanding classification power and robust generalization, especially at detecting deepfakes with fewer false negatives. EfficientNet-B3 came in second with excellent recall (94.73) and ROC AUC (98.91), earning it a strong recommendation when detection precision is weighed against model efficiency and scalability of deployment. ResNet50, although a touch less precise (93.45), recorded the best precision (94.65), indicating it best reduces false positives — a benefit in fields that need to maintain high fidelity for legitimate content. Meso-4, though lighter and architecturally less complex, also produced uniform results with accuracy of 88.57 and ROC AUC of 96.37, thus rendering itself a viable choice for real-time or low-resource settings where computational expense is a principal concern.

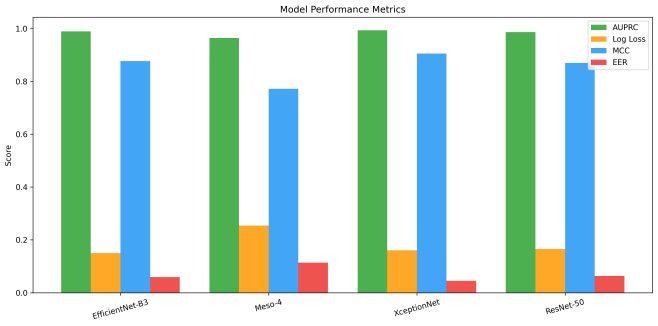

Fig. 5: False Positive Rate vs False Negative Rate of Models.

To put the models into a broader perspective of metrics, we examined Area Under the Precision-Recall Curve (AUPRC), Log Loss, Matthews Correlation Coefficient (MCC), and Equal Error Rate (EER), as shown in Fig. 5. In this comparison, XceptionNet and EfficientNet-B3 outperformed the other frameworks by yielding the lowest Log Loss and EER, as well as the highest AUPRC and MCC. Conversely, Meso-4 performed consistently weaker in these metrics, underscoring the contrast between computational efficiency and detection precision.

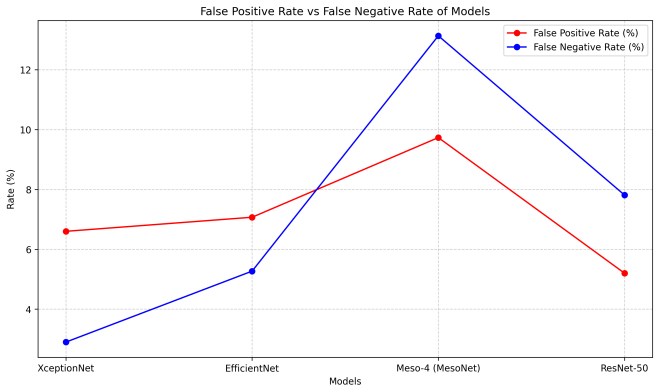

Fig. 6: False Positive Rate vs False Negative Rate of Models.

Additionally, as shown in Fig. 6, we evaluated the False Positive Rate (FPR) and False Negative Rate (FNR) for each model. XceptionNet again achieved the lowest FNR rate at 2.9%, indicating a lower likelihood of incorrectly classifying a fake or real video. Meso-4, on the other hand, exhibited the highest FPR (24.1%) and FNR (19.2%), confirming its relatively lower reliability for DeepFake detection.

V. Conclusion

This research compared four deep learning models, XceptionNet, EfficientNet-B3, ResNet-50, and Meso-4, for DeepFake detection based on the FaceForensics++ dataset. As accuracy was the primary evaluation metric, we also attempted a more comprehensive evaluation using additional metrics, such as Precision, Recall, F1 Score, ROC-AUC, AUPRC, Log Loss, MCC, EER, and various error rates.

Among the models, XceptionNet was the most uniform and high-performing in almost all metrics, validating its efficacy in detecting DeepFake material with high trustworthiness. EfficientNet-B3 presented a good accuracy-computation trade-off and was thus best suited for low-resource settings. ResNet-50 evidenced high precision but slightly lower recall, and lightweight Meso-4 provided quicker inference with lower accuracy and detection confidence at the expense.

These results strongly indicate that DeepFake detection benchmarking should consider multiple metrics, as accuracy in isolation can hide critical model flaws. The in-depth analysis of metrics illustrated complex, and often counterproductive, interactions among precision, recall, and generalization that are essential for actual use.

Ultimately, these DeepFake detection algorithms will be refined by improving generalization performance across datasets, reducing both false positives and false negatives, and detecting more complex DeepFake forgeries by adding spatiotemporal video information.

ACKNOWLEDGMENT

REFERENCES

[1] 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.

[2] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2018, pp. 1–7.

[3] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," 2018.

[4] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *2019 International Conference on Biometrics (ICB)*. IEEE, 2019, pp. 1–8.

[5] A. Rössler, 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.

[6] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[7] R. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2018, pp. 1–7.

[8] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[9] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1251–1258, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks," *IEEE Signal Processing Letters*, 2016.

[12] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[13] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.