Enhancement of Deepfake Detection Framework Integrating Efficient NetB0, Graph Attention Networks, and Gated Recurrent Units

Rafael Alden F. Agoncillo[1], James Kenneth M. Kiunisala[1], Raymund M. Dioses[2], Dan Michael A. Cortez[2]

[1]*Student, College of Information System and Technology Management, Pamantasan ng Lungsod ng Maynila, Philippines*

[2]*Professor, College of Information System and Technology Management, Pamantasan ng Lungsod ng Maynila, Philippines*
Corresponding Author: rafaelagoncillo@gmail.com

***Abstract*: Deepfake technology has emerged as a significant threat to media integrity, cybersecurity, and public trust, enabling the creation of highly realistic manipulated content. Existing detection methods often fail to effectively capture temporal inconsistencies, model complex spatial relationships, and handle noisy labels in training datasets. To address these challenges, this study proposes an enhanced deepfake detection framework that integrates EfficientNetB0 for feature extraction, Graph Attention Networks (GAT) for spatial relationship modeling, and Gated Recurrent Units (GRU) for temporal pattern analysis. To further improve reliability, Jensen-Shannon divergence is employed during training to mitigate the impact of noisy labels. The proposed framework achieved a remarkable accuracy of 80.60% and a low loss of 0.28 on the Celeb-DF (v2) dataset, outperforming widely-used models such as Mesonet, ResNet-50, VGG-19, and Xception. These results highlight the framework's ability to effectively identify manipulated content and address critical limitations in current detection systems. Its scalability and reliability make it suitable for real-world applications, reinforcing public trust in digital media and ensuring content authenticity.**

***Keywords*: Deepfake Detection, EfficientNetB0, Gated Recurrent Unit, Graph Attention Network, Manipulated Content Detection, Spatiotemporal Analysis.**

1. Introduction

The proliferation of deepfakes underscores a significant threat to societal integrity, political stability, and commercial enterprises, which results in journalistic ambiguity, eroded confidence in authoritative sources, and the widespread of misinformation [1]. Deepfakes, convincingly manipulated videos or images, can be easily created using various online software accessible to users of all technical levels. These tools utilize artificial intelligence and deep learning techniques, including auto-encoders and Generative Adversarial Networks (GANs) [2].

Traditional deepfake detection frameworks utilizing Convolutional Neural Networks (CNNs) have demonstrated effectiveness in deepfake detection but exhibit inherent limitations, particularly in their ability to model spatiotemporal and relational properties. Studies indicate that many deepfake video creation methods process or swap faces frame by frame independently, leading to a lack of coherence in temporal information [3]. Convolutional Neural Networks (CNNs) are primarily designed for extracting spatial features from individual frames and they are not equipped to capture or utilize the temporal inconsistencies that arise across sequential frames. This emphasizes the need for more advanced models capable of capturing both spatial and temporal dynamics to enhance deepfake detection accuracy and reliability.

This study advances the development of deepfake detection frameworks by presenting a refined approach that addresses key challenges in their real-world application. The enhanced framework integrates sophisticated temporal analysis techniques to more accurately identify inter-frame anomalies in video deepfakes, employs optimized spatial modeling strategies to discern the complex interactions within facial regions, and incorporates structural modifications to increase the network's resilience against various manipulation methods. These improvements are designed to counter the escalating threats posed by deepfake technology, particularly within critical

Manuscript revised December 16, 2024; accepted December 18, 2024. Date of publication December 21, 2024.
This paper available online at www.ijprse.com
ISSN (Online): 2582-7898; SJIF: 5.59

sectors such as media, political communications, and commercial enterprises, where preserving the integrity and authenticity of digital content is essential for maintaining trust and transparency.

2. Review of Related Literature

Deepfakes are hyper-realistic media created using artificial intelligence (AI) and deep learning algorithms to manipulate facial features and voices, resulting in convincing fabricated content that is often indistinguishable from authentic material [4]. Advances in multimodal generative models such as Generative Adversarial Networks (GANs) and diffusion models have significantly increased the realism of synthetic videos and audio, making deepfakes harder to detect and heightening risks of misuse in misinformation, fraud, and reputation damage [5]. These advancements have fueled a growing arms race between deepfake generation and detection technologies, necessitating innovative approaches to counter increasingly sophisticated fakes, which evolve at an alarming pace with each technological breakthrough. Detection is further complicated by noisy labels and dataset inconsistencies, which can lead to overfitting and poor generalization across diverse manipulations, limiting the effectiveness of existing detection models [6]. To address these challenges, robust loss functions like Generalized Jensen-Shannon Divergence have been developed to enhance resilience to label noise and improve detection performance by ensuring better optimization and adaptability [7]. Additionally, integrating attention mechanisms and Graph Attention Networks (GATs) improves models' ability to focus on critical features and manage complex relationships, enabling the identification of subtle and intricate patterns that are crucial for accurate detection [8]. Convolutional Neural Networks (CNNs) are effective in deepfake detection due to their capacity for automatic feature learning and ability to capture spatiotemporal dependencies within the data, making them invaluable for tasks requiring high-dimensional analysis [9]. However, CNNs struggle with handling temporal information across multiple video frames and are vulnerable to adversarial perturbations, which exploit weaknesses in model robustness and can significantly degrade detection accuracy, highlighting the need for more robust detection methods [10]. Recurrent Neural Networks (RNNs), particularly Gated Recurrent Units (GRUs), effectively model temporal relationships, making them suitable for deepfake detection tasks that require understanding frame sequences and preserving temporal coherence, which is often a key indicator of manipulated content [11].

3. Research Methodology

A. *Existing EfficientNet Convolutional Neural Network*

EfficientNet is a type of convolutional neural network that employs a scaling method to enhance its performance. This method involves uniformly adjusting the network's depth, width, and resolution using a single, combined coefficient [12].

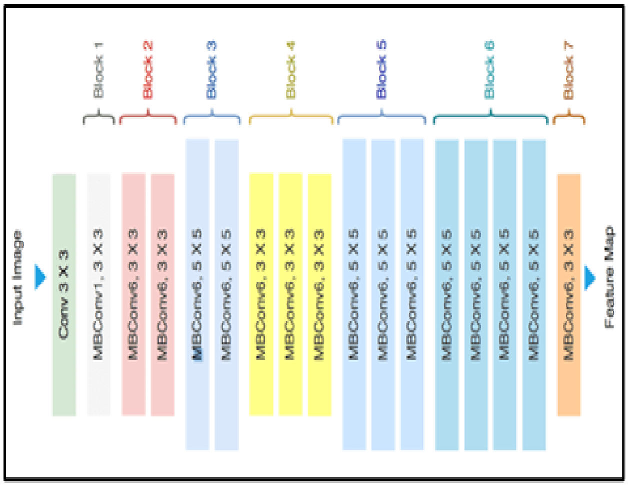

Fig. 1. The efficient net architecture

The architecture of the base model of EfficientNet (EfficientNet-B0) consists of three parts: the stem, body, and the head. The Stem is the initial stage that includes a 3x3 convolution with a stride of 2, followed by Batch Normalization and the Swish activation function. It is responsible for extracting low-level features from the input image. The body is the core of the network consisting of a series of MBConv blocks, which are optimized inverted residual blocks. Each MBConv block integrates depthwise separable convolutions and squeeze-and-excitation (SE) layers to enhance channel-wise feature recalibration. These blocks are configured with varying kernel sizes, expansion ratios, and strides to effectively capture features at multiple scales. The head is the final stage containing a 1x1 convolutional layer, followed by a global average pooling layer, a fully connected layer, and a softmax activation function for classification tasks. This stage consolidates the learned features and maps them to the output classes.

B. *Dataset*

The study used the Celeb-DF V2 dataset, a large-scale collection of videos designed to facilitate deepfake detection studies. The dataset comprises 890 real videos (590 from celebrities and 300 from YouTube) and 5,639 manipulated deepfake videos generated using mainstream techniques such as FaceSwap and DFaker. These videos feature a diverse range of individuals across multiple age groups, genders, and ethnicities. The deepfake videos are exceptionally realistic, incorporating significant visual improvements that make them nearly indistinguishable to the human eye. Celeb-DF V2 also exhibits extensive variations in facial expressions, head poses, lighting conditions, face sizes, orientations, and background environments [13].

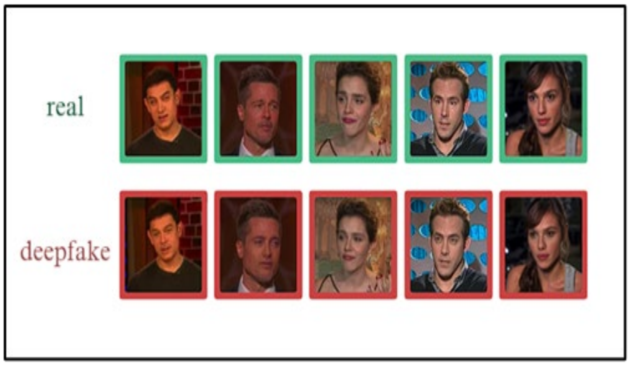

Fig. 2. Celeb-DF V2 deepfake dataset

To ensure a balanced dataset, the researchers used all 890 real videos from the original Celeb-DF V2 and randomly selected a corresponding 890 deepfake videos, resulting in a total of 1,780 samples. This design choice was made to create an equal representation of real and deepfake samples, enhancing the model's ability to effectively distinguish between the two classes. The balanced subset retains the dataset's diversity in terms of age, gender, ethnicity, facial expressions, head poses, lighting conditions, and background environments, providing a robust foundation for deepfake detection.

Table 1
Celeb-DF data set table

Dataset	Total Data	Real	Deepfake	Training Ratio
CelebDF-v2	1780	890	890	70:30

The researchers adopted a 70:30 train-test split for the dataset. This was based on widely recognized practices in machine learning, where a 70:30 split between training and testing data is commonly used to ensure both adequate training and reliable evaluation. It is suggested to use a 70/30 ratio for datasets between 100 and 1,000,000 instances. Studies have demonstrated that reserving 20-30% of the dataset for testing, with the rest allocated for training, strikes a good balance between the quality of training and the accuracy of evaluation (Muraina, 2022).

C. *Pseudocode of Enhanced Deepfake Detection Framework*

```
Load Video Frames
    1)  Load video file video using a custom video loader.
    2)  Extract N frames evenly across the video timeline.
    3)  Ensure frames have the same dimension (224x224)
    Preprocess Frames
    4)  Use YOLO for person detection and MTCNN for precise face cropping.
    5)  Apply necessary data augmentation: Gaussian Noise, Random Brightness / Contrast, Rotation, Random Gamma, CLAHE, Elastic Transform, Median Blur, Sharpen, Horizontal and Vertical Flip
    Define CNN Architecture
    6)  Use EfficientNet-B0 as the base architecture:
        Stem: A 3x3 convolutional layer with BatchNorm and ReLU activation.
        Body: Stacked MBConv blocks with depthwise convolutions and squeeze-and-excitation layers.
        Head: Fully connected layer with softmax activation for binary classification ('real' or 'fake').
    Build EfficientB0 + GAT + GRU
    7)  Extract spatial features with EfficientNet-B0.
    8)  Process temporal dependencies with GRU layers.
    9)  Refine spatial relationships with GAT layers.
    10) Replace the final layer with a binary classification head for 'real' or 'fake'.
    Define Loss Function
    11) Use BCE + Jensen-Shannon Divergence as the loss function.
    Train the EfficientB0 + GAT + GRU Architecture
    12) Use the preprocessed dataset (from Step 3.5) for training
    13) Train with the Adam optimizer and a learning rate of 1e-4.
    14) Train with a batch size of 32 over 20 epochs, using class balancing Verify Prediction
    15) Input new video sequences and their corresponding labels.
    16) Compute the model output probabilities and apply a sigmoid function to convert them to probabilities.
    17) Compare probabilities against a threshold (e.g., 0.5) to classify videos as 'real' or 'fake'.
    Evaluate Performance
    18) Compute metrics: Accuracy and Loss
    19) Maximize Accuracy
    Display Results
    20) Display performance metrics for evaluation.
```

D. *Framework of Enhanced Deepfake Detection*

This study focuses on developing a Deepfake Detection Framework that utilizes EfficientNet as a CNN base model, combined with Graph Attention Networks (GAT) for Spatial Relation Enhancement and Gated Recurrent Units (GRUs) for Temporal Analysis. This framework is designed to combat the proliferation of deepfake videos by providing an effective and robust deepfake detection solution.

The input stage utilizes the Celeb-DF (v2) dataset comprising real and fake video samples, processed via frame extraction, face detection with YOLOv5, and data augmentation (e.g., rotation, blurring) to enhance generalization. In the process stage, EfficientNet-B0 extracts features from facial regions, GRUs analyze temporal inconsistencies, and GATs refine spatial relationships among facial features. The output stage employs Binary Cross-Entropy for prediction accuracy and Jensen-Shannon Divergence to ensure consistent and reliable classification of real or fake videos.

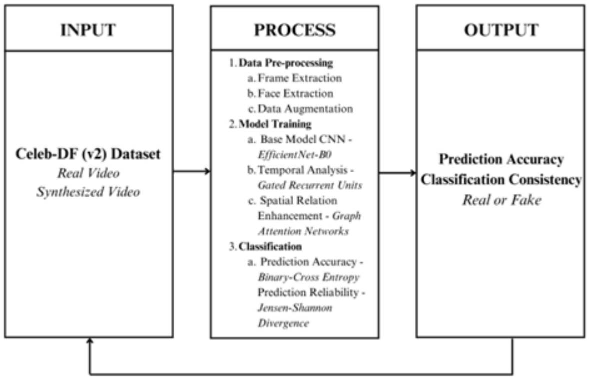
INPUT
Celeb-DF (v2) Dataset
Real Video
Synthesized Video
PROCESS
1. Data Pre-processing
a. Frame Extraction
b. Face Extraction
c. Data Augmentation
2. Model Training
a. Base Model CNN - EfficientNet-B0
b. Temporal Analysis - Gated Recurrent Units
c. Spatial Relation Enhancement - Graph Attention Networks
3. Classification
a. Prediction Accuracy - Binary-Cross Entropy
Prediction Reliability - Jensen-Shannon Divergence
OUTPUT
Prediction Accuracy
Classification Consistency
Real or Fake

Fig. 3. IPO Model of enhanced deepfake detection framework

4. Metrics for Evaluation

To evaluate the performance of the enhanced Deepfake Detection Framework, the following metrics were used:

1) Accuracy - Represents the percentage of predictions where the model correctly identifies real and fake videos, providing an overall assessment of its performance in accurately classifying the input data. It is a key metric for evaluating the model's effectiveness in distinguishing between authentic and manipulated content.
2) Loss - Reflects the level of error in the model's predictions during the training process, serving as an indicator of how well the model is learning to differentiate real videos from deepfakes. By minimizing this value, the model can improve its ability to make accurate classifications and generalize better to unseen data.

5. Result and Discussion

Table 2
Enhanced deepfake detection framework vs. baseline models

	Accuracy	Loss
Mesonet	73.19%	25.83
ResNet-50	75.26%	6.55
VGG-19	74.92%	1.06
Xception	77.83%	11.69
EfficientB0 + GAT + GRU	80.60%	0.28

The integration of EfficientB0 with Graph Attention Network (GAT) and Gated Recurrent Unit (GRU) in the proposed framework demonstrates a significant enhancement in both accuracy and loss metrics for deepfake detection compared to other models (Table 4.1). Achieving an accuracy of 80.60% and a remarkably low loss of 0.28, this model outperforms widely-used architectures such as Mesonet, ResNet-50, VGG-19, and Xception. For instance, while Xception shows competitive accuracy at 77.83%, its loss remains substantially higher at 11.69. Similarly, despite ResNet-50 achieving 75.26% accuracy, its loss of 6.55 highlights the superior optimization and robustness of the EfficientB0 + GAT + GRU approach. These results underline the framework's capability to better generalize and effectively minimize error, marking it as a leading solution for reliable deepfake detection.

6. Conclusion

This study addressed the challenges of deepfake detection by developing an enhanced framework combining EfficientB0, GAT, and GRU. Traditional CNNs struggled to capture temporal inconsistencies and relational features in deepfake videos, limiting their reliability. The proposed model achieved a remarkable accuracy of 80.60% and a low loss of 0.28, outperforming Mesonet, ResNet-50, VGG-19, and Xception. By leveraging advanced spatiotemporal analysis and attention mechanisms, the framework effectively detects inter-frame anomalies and manipulation inconsistencies. These results highlight the model's potential to enhance deepfake detection in critical sectors like media, politics, and commerce, ensuring content authenticity and public trust. While demonstrating robust performance, the findings also emphasize the need for further research to optimize scalability and computational efficiency.

References

[1] Mahmud, B. U., & Sharmin, A. (2023, January 8). Deep Insights of Deepfake Technology: A Review.

[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., Profile, I. G. B., Profile, J. P.-A. de M., Profile, M. M. de M., Profile, B. X. de M., Profile, D. W.-F. de M., Profile, S. O. de M., Profile, A. C. de M., & Profile, Y. B. de M. (2020, October 22). Generative Adversarial Networks. Communications of the ACM.

[3] Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep Learning for Time Series classification: A Review. Data Mining and Knowledge Discovery, 33(4), 917–963.

[4] Alanazi, S., & Asif, S. (2024, September 18). Exploring deepfake technology: Creation, consequences and countermeasures - human-intelligent systems integration. SpringerLink.

[5] Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? Business Horizons, 63(2), 135–146.

[6] Gong, L. Y., & Li, X. J. (2024). A contemporary survey on Deepfake Detection: Datasets, algorithms, and challenges. Electronics, 13(3), 585

[7] Englesson, E., & Azizpour, H. (2021, October 29). Generalized jensen-shannon divergence loss for learning with noisy labels.

[8] Das, A., Das, S., & Dantcheva, A. (2021). Demystifying attention mechanisms for deepfake detection. 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), 1–7.

[9] Thing, V. L. L. (2023). Deepfake detection with deep learning: Convolutional Neural Networks versus transformers. 2023 IEEE International Conference on Cyber Security and Resilience (CSR), 246–253.

[10] Tolosana, R., Vera-Rodriguez, R., & Busch, C. (2022). Handbook of digital face manipulation and detection: From deepfakes to morphing attacks. Springer.

[11] Dey, R., & Salem, F. M. (2017). Gate-variants of gated recurrent unit (GRU) neural networks. 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), 1597–1600.

[12] Tan, M., & Le, Q. V. (2020, September 11). Efficient Net: Rethinking model scaling for Convolutional Neural Networks.

[13] Gong, L. Y., & Li, X. J. (2024). A contemporary survey on deepfake detection: Datasets, algorithms, and challenges. Electronics, 13(585).