Hybrid Deepfake Video Detection Using EfficientNet and BiGRU

1st Fariz Rachman Hadi*
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Gadjah Mada
Yogyakarta, Indonesia
farizrachmanhadi@mail.ugm.ac.id

2nd Risang Panggalih*
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Gadjah Mada
Yogyakarta, Indonesia
risangpanggalih@mail.ugm.ac.id

3rd Porto Mauritio Hartley*
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Gadjah Mada
Yogyakarta, Indonesia
portomauritiohartley@mail.ugm.ac.id

4th Ichsan Setiawan*
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Gadjah Mada
Yogyakarta, Indonesia
ichsan.setiawan@mail.ugm.ac.id

***Abstract*—The rapid advancement of DeepFake technology poses a serious threat to digital information integrity, enabling the creation of highly realistic manipulated facial videos. This study proposes a hybrid deep learning model for DeepFake detection, combining EfficientNet-B0 for spatial feature extraction and Bidirectional Gated Recurrent Unit (BiGRU) for temporal dependency analysis. Experiments were conducted using the Deep Fake Detection (DFD) dataset, which presents a significant challenge due to extreme data imbalance (89.4% fake vs. 10.6% real videos). A two-phase training strategy (warmup and fine-tuning) was employed along with a weighted loss function to address class imbalance. Experimental results show that the model achieved an overall accuracy of 90.1% and a Train AUC of 95.3%. However, further evaluation using the F1-Score reveals that while the model is highly effective at detecting the Fake class (F1-Score 0.944), it struggles to correctly identify the Real class (F1-Score 0.553), indicating overfitting toward the majority class. This research provides insights into the effectiveness of the hybrid CNN-RNN architecture while highlighting the critical challenge of data imbalance in digital forensics.**

Keywords—DeepFake, EfficientNet, BiGRU, Imbalance Dataset, Computer Vision, Digital Forensics

I. Introduction

The rapid advancement of artificial intelligence, particularly in generative models, has led to the emergence of sophisticated media manipulation techniques known as "Deepfakes." These technologies allow for the creation of highly realistic synthetic videos where a person's facial appearance or expressions are swapped or altered digitally. While this technology has creative applications, its misuse poses a severe threat to the integrity of digital information. The widespread dissemination of deepfakes can facilitate the spread of disinformation, political manipulation, identity theft, and financial fraud, making it increasingly difficult for the public to distinguish between authentic and fabricated content.

Detecting deepfake videos is a complex task that requires analyzing visual artifacts at the frame level as well as temporal inconsistencies across video sequences. Single-frame analysis often fails to capture the subtle flickering or unnatural movements that occur over time. Therefore, effective detection systems must integrate both spatial and temporal feature extraction methods.

In this study, we propose a hybrid deep learning architecture that combines EfficientNet-B0 for robust spatial feature extraction and a Bidirectional Gated Recurrent Unit (BiGRU) for modeling temporal dependencies. EfficientNet is chosen for its efficiency and high performance in image classification, while BiGRU is utilized to capture context from both forward and backward temporal directions in video sequences.

II. Related Work

Deepfake detection approaches have evolved from analyzing single-frame spatial artifacts to capturing temporal inconsistencies across video sequences. Hybrid models, which integrate Convolutional Neural Networks (CNN) for feature extraction and Recurrent Neural Networks (RNN) for sequence modeling, have recently demonstrated superior performance.

Khudhur and Mohammed [2] proposed a spatio-temporal model combining EfficientNetV2-B0 with Bidirectional LSTM (Bi-LSTM). Their approach achieved 99.51% accuracy on the FaceForensics++ dataset, demonstrating high efficiency with low computational cost. Similarly, Al-Adwan et al. [1] developed a hybrid CNN-RNN model optimized using the Particle Swarm Optimization (PSO) algorithm. Their method showed robust generalization capabilities with 97.26% accuracy on the Celeb-DF dataset, although the specific CNN architecture was not explicitly detailed.

More recently, Nelson et al. [3] introduced a three-stage multi-modal framework utilizing XceptionNet for spatial features and LSTM for temporal analysis. While their model achieved 97.57% accuracy across multiple datasets (FF++, DFDC, and Celeb-DF), the use of XceptionNet typically incurs a higher computational load compared to more lightweight architectures.

Despite these advancements, most existing studies rely on balanced datasets or computationally heavier backbones like XceptionNet. Our research distinguishes itself by employing EfficientNet-B0 combined with BiGRU (Bidirectional Gated Recurrent Unit). BiGRU offers a more computationally efficient alternative to LSTM while maintaining effectiveness in capturing long-term dependencies. Furthermore, unlike previous works that often utilize balanced benchmarks, this study specifically addresses the challenge of extreme class imbalance found in the Deep Fake Detection (DFD) dataset, providing insights into model performance in realistic forensic scenarios.

III. Methodology

The proposed deepfake detection framework consists of three main stages: data preprocessing, spatial-temporal feature extraction using a hybrid architecture, and a specialized training strategy to handle class imbalance.

A. *Dataset Preparation*

This study utilizes the Deep Fake Detection (DFD) dataset, which contains 3,431 videos with a resolution of 1920x1080 at 24 fps. A significant challenge in this dataset is the extreme class imbalance, consisting of 3,068 fake videos (89.4%) and only 363 real videos (10.6%).

We employed a stratified split strategy to maintain the class distribution ratio across all subsets:

- Training Set: 2,744 videos (80%)
- Validation Set: 343 videos (10%)
- Test Set: 344 videos (10%)

B. *Preprocessing*

Fig. 1 shows the preprocessing pipeline that consists of several stages. First, a 2-second mid-clip is extracted from the temporal center of each video to ensure representative content. From this segment, 8 frames are uniformly sampled. Facial regions within the frames are localized and cropped using the Multi-task Cascaded Convolutional Networks (MTCNN). Each cropped face is then resized to 224×224 pixels to match standard input dimensions for CNN-based models. Finally, all images undergo normalization using the ImageNet mean and standard deviation to align with common pretrained model preprocessing standards.

C. *Model Architecture*

Fig. 2 shows the baseline model, a hybrid architecture combining EfficientNet-B0 as a frame-level feature extractor and a Bidirectional Gated Recurrent Unit (BiGRU) for temporal modeling. Each input consists of 8 frames of size 224×224. EfficientNet-B0, pretrained on ImageNet, encodes every frame into a 1280-dimensional feature vector. The resulting sequence is reshaped and passed to a BiGRU with 256 hidden units per direction, producing a concatenated temporal representation of size 512. This representation is processed through a dropout layer (p =

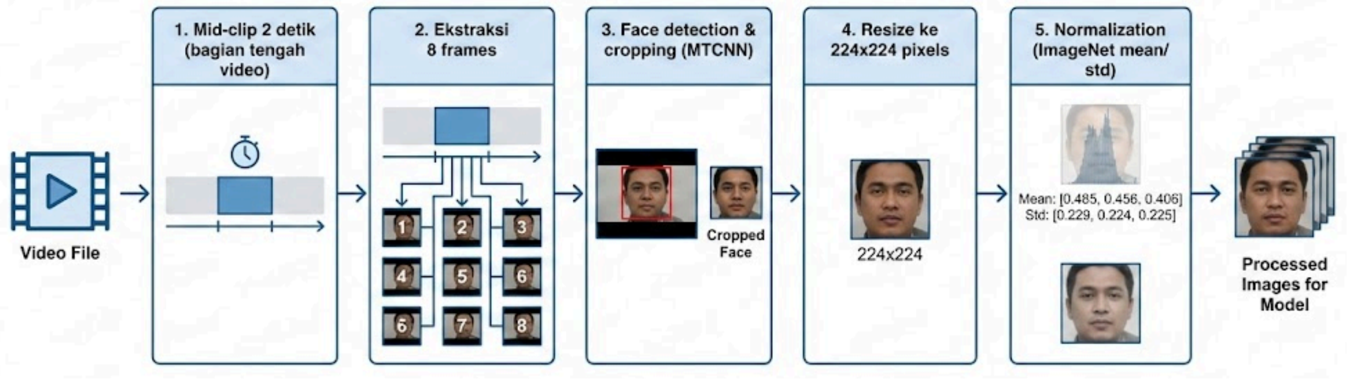
Video File
1. Mid-clip 2 detik (bagian tengah video)
2. Ekstraksi 8 frames
1
2
3
4
5
6
6
7
8
3. Face detection & cropping (MTCNN)
Cropped Face
4. Resize ke 224x224 pixels
224x224
5. Normalization (ImageNet mean/ std)
Mean: [0.485, 0.456, 0.406]
Std: [0.229, 0.224, 0.225]
Processed Images for Model

Fig. 1. Video Pre-Processing Pipeline

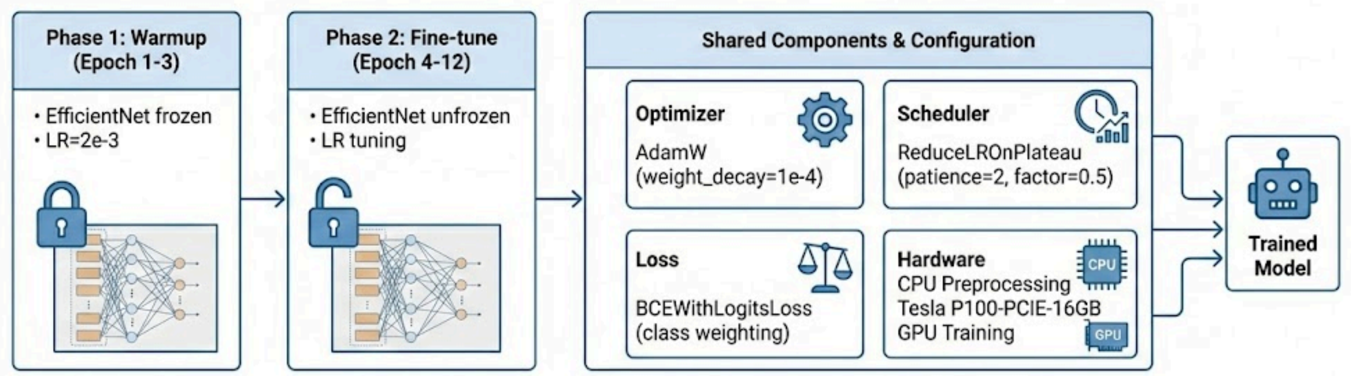
Phase 1: Warmup (Epoch 1-3)
• EfficientNet frozen
• LR=2e-3
Phase 2: Fine-tune (Epoch 4-12)
• EfficientNet unfrozen
• LR tuning
Shared Components & Configuration
Optimizer
AdamW (weight_decay=1e-4)
Scheduler
ReduceLROnPlateau (patience=2, factor=0.5)
Loss
BCEWithLogitsLoss (class weighting)
Hardware
CPU Preprocessing
Tesla P100-PCIE-16GB GPU Training
CPU
GPU
Trained Model

Fig. 2. Two Phase Fine-tuning Pipeline

0.2), followed by a fully connected layer that outputs a single sigmoid-activated probability.

As shown in Fig. 2–the training procedure is divided into two phases. In Phase 1 (Epochs 1–3), EfficientNet is frozen to stabilize the GRU training, using a learning rate of 2×10^{-3}. In Phase 2 (Epochs 4–12), EfficientNet is unfrozen for fine-tuning. Optimization is performed using AdamW with weight decay of 1×10^{-4}. A ReduceLROnPlateau scheduler (patience = 2, factor = 0.5) adjusts the learning rate adaptively. The model is trained with BCEWithLogitsLoss and class weighting to mitigate class imbalance. Preprocessing is carried out on CPU, while model training utilizes an NVIDIA Tesla P100-PCIE-16GB GPU.

IV. Experimental Results

This section presents a comprehensive evaluation of the proposed hybrid deepfake detection model on the Deep Fake Detection (DFD) dataset.

A. Training Performance

The model was trained for 12 epochs with a two-phase strategy. Table I summarizes the training metrics across all epochs. During the warmup phase (epochs 1–3), EfficientNet remained frozen while the BiGRU component was trained. In the fine-tuning phase (epochs 4–12), all parameters were optimized jointly.

TABLE I. TRAINING METRICS ACROSS 12 EPOCHS

Epoch	Train Loss	Train F1	Train AUC	Val Loss	Val F1	Val AUC
1	0.8443	0.423	0.858	1.1075	0.362	0.805
2	0.8187	0.433	0.867	0.8799	0.45	0.849
3	0.8413	0.417	0.861	0.884	0.341	0.832
4	0.6175	0.521	0.926	0.8812	0.385	0.838
5	0.5992	0.509	0.93	0.9497	0.388	0.826
6	0.534	0.578	0.947	0.9473	0.364	0.834
7	0.5513	0.537	0.942	1.0407	0.48	0.83
8	0.5169	0.591	0.949	0.9665	0.413	0.839
9	0.4917	0.59	0.954	1.1596	0.408	0.824
10	0.4803	0.587	0.956	0.992	0.424	0.845
11	0.4739	0.617	0.958	0.9634	0.422	0.85
12	0.4966	0.594	0.953	0.9812	0.4	0.848

The training loss decreased from 0.8443 to 0.4966, while validation loss stabilized around 0.98. Training AUC reached 95.3% at epoch 12, demonstrating strong discriminative capability on the training set. The best validation AUC of 85.0% was achieved at epoch 11, which was selected as the final model checkpoint. The training process required approximately 30 minutes on an NVIDIA Tesla P100 GPU.

B. Test Set Evaluation

The model trained at epoch 11 was evaluated on the held-out test set of 344 videos. Given the extreme class imbalance, we investigated multiple decision thresholds to optimize performance. The optimal threshold of 0.85 was selected to maximize the macro-averaged F1-score.

TABLE II. CLASSIFICATION REPORT (THRESHOLD = 0.85)

Class	Precision	Recall	F1-Score	Support
FAKE	0.948	0.941	0.944	307
REAL	0.538	0.568	0.553	37
Accuracy			0.901	344
Macro Avg	0.743	0.754	0.749	344
Weighted Avg	0.904	0.901	0.902	344

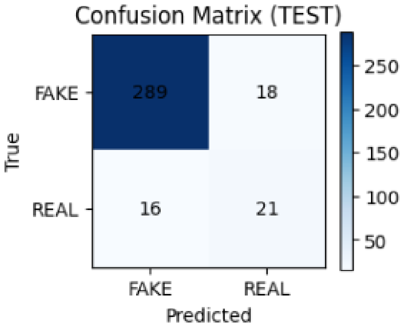

Fig. 3. Confusion Matrix

The confusion matrix in Fig. 3 provides insight into the classification behavior. The model correctly identified 289 out of 307 fake videos (True Negatives) and 21 out of 37 real videos (True Positives). However, 18 fake videos were incorrectly classified as real (False Positives), and 16 real videos were misclassified as fake (False Negatives).

C. Threshold Analysis

To understand the impact of the classification threshold on performance, we evaluated the model across a range of threshold values from 0.1 to 0.99. As shown in Table III, different thresholds optimize for different objectives.

TABLE III. PERFORMANCE AT KEY THRESHOLDS

Threshold	F1 (FAKE)	F1 (REAL)	Macro F1	Accuracy
0.5	0.944	0.553	0.749	0.901
0.85	0.959	0.563	0.761	0.906
0.95	0.974	0.486	0.73	0.897

At threshold 0.85, the model achieves the best balance between detecting both classes, with a macro-averaged F1-score of 76.1%. At higher thresholds (e.g., 0.95), the model becomes more conservative in predicting the REAL class, which increases precision for fake detection but reduces recall for real videos.

V. Discussion

A. Model Strengths

The proposed hybrid architecture demonstrates several notable strengths. First, the model achieves excellent performance on fake video detection, with an F1-score of 94.4% and precision of 94.8% at the optimal threshold. This high precision is critical for practical deployment, as it minimizes false alarms when flagging potentially manipulated content.

Second, the overall AUC-ROC of 82.1% indicates strong discriminative capability across varying decision thresholds. The particularly high PR-AUC of 97.3% for the fake class confirms the model's reliability in identifying manipulated videos.

Third, the computational efficiency of the architecture—achieved through the use of EfficientNet-B0 and BiGRU instead of heavier alternatives like XceptionNet or LSTM—makes the approach viable for real-world deployment scenarios where inference speed is critical. The entire test set of 344 videos was processed efficiently on GPU hardware.

B. Impact of Class Imbalance

Despite the overall accuracy of 90.1%, the model exhibits significantly weaker performance on the minority REAL class (F1-score of 55.3%). This disparity stems directly from the severe class imbalance in the dataset (89.4% fake vs. 10.6% real). Even with weighted loss functions, the model's learning is dominated by the majority class.

The confusion matrix reveals that 16 out of 37 real videos (43.2%) were misclassified as fake. This high false negative rate for real videos suggests that the model has learned to be overly suspicious, potentially because the training signal from authentic videos is insufficient to establish robust decision boundaries.

C. Comparison with Related Work

Table IV compares our results with recent state-of-the-art approaches from the literature.

TABLE IV. COMPARISON WITH RELATED WORK

Method	Dataset	Accuracy	Architecture
Khudhur & Mohammed [1]	FF++	99.51%	EfficientNetV2-B0 + Bi-LSTM
Al-Adwan et al. [2]	Celeb-DF	97.26%	CNN-RNN + PSO
Nelson et al. [3]	FF++/DFD C/Celeb-DF	97.57%	XceptionNet + LSTM
Our Method	DFD	90.10%	EfficientNet-B0 + BiGRU

While our accuracy is lower than previous studies, this comparison must account for dataset characteristics. The DFD dataset presents a more challenging scenario due to extreme class imbalance, whereas datasets like FaceForensics++ and Celeb-DF are typically more balanced or have been balanced through preprocessing. Our work provides realistic insights into model performance under imbalanced conditions common in real-world forensic scenarios.

Additionally, our approach offers computational advantages. BiGRU provides comparable temporal modeling to LSTM with reduced parameters and faster inference, while EfficientNet-B0 is significantly lighter than XceptionNet, making our solution more suitable for resource-constrained deployment.

D. Limitations and Challenges

Several limitations should be acknowledged. First, the class imbalance fundamentally constrains model performance on the minority class. More sophisticated sampling strategies, such as focal loss or class-balanced sampling, could potentially improve real video detection.

Second, the model was trained on a single dataset (DFD) with specific manipulation techniques. Cross-dataset generalization remains an open question. Future work should evaluate the model on other deepfake datasets to assess robustness to different generation methods.

Third, the current preprocessing pipeline extracts a fixed 2-second mid-clip from each video. This strategy may miss temporal artifacts that occur at video boundaries or throughout longer sequences. Exploring multi-segment sampling or full-video analysis could capture more comprehensive temporal patterns.

VI. *Conclusion*

This study presented a hybrid deep learning architecture combining EfficientNet-B0 and Bidirectional GRU for deepfake video detection. Experiments on the Deep Fake Detection dataset demonstrated that the model achieves 90.1% accuracy and 82.1% AUC, with particularly strong performance on fake video detection (F1-score of 94.4%). However, performance on the minority real class remains a challenge due to severe class imbalance in the dataset.

The two-phase training strategy—freezing the pretrained backbone during warmup before fine-tuning—proved effective for transfer learning in this context. The use of lightweight architectures (EfficientNet-B0 and BiGRU) provides computational efficiency advantages over heavier alternatives while maintaining competitive performance.

A. Future Work

Several directions warrant further investigation. First, advanced techniques for handling class imbalance, such as focal loss, class-balanced sampling, or synthetic minority oversampling (SMOTE), could improve detection of authentic videos. Second, ensemble methods combining multiple architectures or temporal sampling strategies may enhance robustness. Third, cross-dataset evaluation is essential to assess generalization capability across different deepfake generation methods and video characteristics.

Finally, incorporating attention mechanisms—either spatial attention to focus on facial regions most susceptible to manipulation artifacts, or temporal attention to weight frames with stronger forgery indicators—could further improve detection accuracy while maintaining interpretability for forensic analysis.

The code, trained models, and preprocessing pipeline are available for research purposes to facilitate reproducibility and future comparisons.

Acknowledgment

This work is supported by the Department of Computer Science and Electronics, Universitas Gadjah Mada. The authors would like to thank Mr. Wahyono PhD for the opportunity to conduct this research as part of the Computer Vision class at Master Program In Artificial Intelligence of Universitas Gadjah Mada. His instruction provided the necessary foundation for completing this study.

References

[1] Al-Adwan, A., Alazzam, H., Al-Anbaki, N., & Alduweib, E.. “Detection of Deepfake Media Using a Hybrid CNN–RNN Model and Particle Swarm Optimization (PSO) Algorithm”. Computers. (2024).

[2] Khudhur, R. Z., & Mohammed, M. A.. “A Spatio-Temporal Deep Learning Approach for Efficient Deepfake Video Detection”. ARO-The Scientific Journal of Koya University. (2025).

[3] Nelson, L., Batra, H., & P., R.. “Deepfake Detection in Manipulated Images/Audio/Videos: A Three-Stage Multi-Modal Deep LearningFramework”. Inteligencia Artificial. (2025)