

RESEARCH ARTICLE

The Hybrid Residual–BiLSTM Model for Mitigating the Vanishing Gradient Problem in Synthetic Image Detection

Kafitra Marna Ibrahim ¹, Muhammad Munsarif ^{2,*}, Muhammad Sam'an ², and Brian Halubanza ³, Ahmad Shukri Mohd Noor ⁴, Azyan Yusra Kapi ⁵

¹ Undergraduate Program in Informatics, Universitas Muhammadiyah Semarang, Faculty of Computer Science and Digital Technology, Semarang, Indonesia

² Department of Informatics, Faculty of Computer Science and Digital Technology, Universitas Muhammadiyah Semarang, Semarang, Indonesia

³ Department of Computer Science, School of Engineering and Technology, Mulungushi University, Kabwe, Zambia

⁴ Software Engineering, Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, Kuala Nerus, Terengganu, Malaysia

⁵ Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM) Johor Branch, Pasir Gudang Campus, Johor, Malaysia

*Correspondence m.munsarif@unimus.ac.id

Received: 10 September 2026; Revised: 19 September 2026; Accepted: 28 September 2026, Published: 30 September 2026

ABSTRACT

The increasing use of generative artificial intelligence has intensified the need for automated methods that can distinguish real images from synthetically generated images. Reliable synthetic image detection requires a classification model that can represent visual patterns while maintaining stable information propagation across its computational components. This study proposes a Residual–Bidirectional Long Short-Term Memory (Residual–BiLSTM) model that integrates residual convolutional feature extraction, bidirectional sequential modeling, and attention-based representation learning for binary synthetic image detection. The model was evaluated using the CIFAKE dataset containing 120,000 images, comprising 100,000 training images and 20,000 test images. Model performance was assessed using accuracy, precision, recall, F1-score, confusion matrix analysis, training and validation behavior, and gradient-flow analysis. The proposed model achieved a test accuracy of 97.27%. The REAL and FAKE classes each obtained precision, recall, and F1-score values of 0.97. The confusion matrix recorded 9,739 correctly classified FAKE images and 9,715 correctly classified REAL images. Training and validation results showed decreasing loss and increasing accuracy during model training. Gradient-flow analysis identified non-zero gradients across the convolutional, residual, BiLSTM, and fully connected components, with a local reduction in gradient magnitude within the BiLSTM layers. These findings show that the proposed Residual–BiLSTM architecture achieved balanced classification performance across both classes while preserving gradient propagation under the evaluated experimental setting, providing an integrated framework for performance and gradient analysis in synthetic image detection.

KEYWORDS

Synthetic Image Detection; Residual Learning; Bidirectional Long Short-Term Memory; Gradient Flow; CIFAKE; Deep Learning

1. Introduction

Generative artificial intelligence has expanded the ability of computational models to create visual content from learned data representations. Latent diffusion models can generate high-resolution images by combining latent-space representation learning with diffusion-based generation [1]. Text-to-image models can also generate visual content directly from natural-language descriptions [2]. Diffusion-based text-to-image models have further improved the visual fidelity of generated images through large-scale language and image representation learning [3]. The increasing use of AI-generated visual content has also created concerns regarding authenticity and ownership in digital media. Public debate regarding AI-generated artwork increased after an AI-generated image received an award in an art competition [4]. The broader problem of unreliable digital information also affects how users evaluate the credibility of visual content distributed through online platforms [5]. Research on fake information further shows that human judgment can be influenced by the way false information is presented and interpreted [6].

The development of generative models has increased the need for computational methods that can distinguish synthetic images from authentic images. Statistical properties of image data have been examined as indicators for detecting images generated by Generative Adversarial Networks [7]. Adversarial face synthesis has also demonstrated that artificial facial images can be generated to challenge existing recognition systems [8]. Security-related studies have further shown that deep learning models can be exposed to attacks that reconstruct or manipulate visual representations [9]. Similar concerns have also been examined in verification systems that face robotic and generative adversarial attacks [10]. Deep learning methods have been widely applied to visual classification because they can learn hierarchical representations directly from image data. Residual learning introduces shortcut connections that allow information to pass across multiple convolutional layers [11]. These shortcut connections support gradient propagation during the optimization of deep neural networks [11]. This property makes residual architectures relevant to image classification tasks that require deeper feature extraction [11].

Synthetic image detection has been evaluated using the CIFAKE dataset in several previous studies. Conventional convolutional neural networks have been applied to distinguish real images from AI-generated images on this dataset [12]. CIFAKE provides balanced real and synthetic image classes for standardized evaluation of AI-generated image detection methods [13]. Residual-based architectures have also been investigated for real and AI-generated image classification [14]. An attention-enhanced ResNet architecture achieved an accuracy of 96.12% on CIFAKE by integrating a Squeeze-and-Excitation mechanism with ResNet-50 [15]. Comparative evaluation of convolutional neural networks and Vision Transformer models has also been conducted for AI-generated image detection [16]. Fine-tuned residual networks have achieved an F1-score of 0.963 on AI-generated image classification tasks [17]. Most previous CIFAKE studies have emphasized classification performance through metrics such as accuracy and F1-score [17]. These evaluations provide information about predictive performance but do not directly describe gradient propagation behavior during model training. Gradient behavior is relevant because deep neural networks can experience optimization difficulties as network depth increases [11].

Residual-based hybrid architectures can combine convolutional spatial feature extraction with bidirectional sequential modeling [18]. Bidirectional Long Short-Term Memory layers process feature sequences in forward and backward directions to capture contextual relationships between representations [18]. The integration of residual convolutional learning and bidirectional sequential modeling therefore provides a framework for examining spatial representations and contextual feature relationships within the same architecture [18]. However, the reviewed CIFAKE studies did not explicitly evaluate gradient-flow behavior in a Residual-BiLSTM architecture for synthetic image detection. This study proposes a Residual-BiLSTM model for synthetic image detection using the CIFAKE dataset. The proposed architecture combines residual convolutional feature extraction with bidirectional sequential modeling. The residual component supports gradient propagation across convolutional layers, while the BiLSTM component models contextual relationships among the extracted feature representations. The model is evaluated using accuracy, precision, recall, F1-score, confusion matrix analysis, training and validation curves, and gradient-flow analysis. The evaluation therefore examines both classification performance and optimization behavior during model training.

2. Literature Review

Synthetic image detection has been examined using conventional convolutional neural networks on the CIFAKE dataset. A convolutional architecture achieved an accuracy of approximately 91% in distinguishing real images from AI-generated images. The study demonstrated the applicability of convolutional feature extraction to binary synthetic image classification [12]. The CIFAKE dataset was developed as a balanced benchmark for classifying real and AI-generated images. The dataset contains real images and synthetic images generated using latent diffusion models, which support standardized evaluation of image detection methods. A convolutional model evaluated on this dataset achieved an accuracy of 92.98% [13].

Residual architectures have also been evaluated for real and AI-generated image classification. A ResNet-based model achieved an accuracy of 94%, while the evaluated Variational Autoencoder-based approach achieved an accuracy of 71%. The reported results indicate that the evaluated discriminative residual architecture produced higher classification accuracy under the experimental conditions [14]. Residual learning uses shortcut connections to transfer representations across network layers. These connections provide alternative paths for information and gradient propagation during backpropagation. This mechanism addresses optimization difficulties associated with increasing network depth [11].

Attention mechanisms have been combined with residual architectures for synthetic image detection. A Squeeze-and-Excitation mechanism integrated with ResNet-50 achieved an accuracy of 96.12% on CIFAKE. The same experiment reported accuracies of 93.46% for ResNet-50, 95.98% for VGGNet-16, and 96.06% for DenseNet-121 [15]. Convolutional neural networks and Vision Transformer architectures have also been compared on the CIFAKE dataset. The evaluated convolutional model achieved an accuracy of 96.31%, while the Vision Transformer achieved an accuracy of 87.09%. The results demonstrate different classification outcomes between the two architectures under the reported experimental setting [16].

Fine-tuning has been applied to deep residual networks for AI-generated image classification. A fine-tuned ResNet-152 model achieved an F1-score of 0.963. The study shows that model adaptation through fine-tuning can influence classification performance without replacing the residual backbone [17]. Recent research has also examined compact architectures for synthetic image detection. A lightweight convolutional neural network was designed to reduce model complexity while retaining classification capability for distinguishing real and synthetic images [19]. Cross-dataset evaluation has been investigated through an explainable meta-ensemble framework. The framework was evaluated using CIFAKE, FakeGPT, and PFake to examine synthetic image classification across multiple datasets [20].

Explainability has also become part of synthetic image forensic analysis. A low-resolution forensic approach examined visual artifacts associated with AI-generated images and incorporated interpretable mechanisms into the detection process [21]. Recent reviews of AI-generated visual content detection identify generalization, model efficiency, explainability, and adaptation to evolving generative techniques as continuing research issues. These issues indicate that classification performance alone does not describe all aspects required for evaluating synthetic image detection systems [22]. Hybrid architectures provide another direction for combining spatial and contextual representation learning. A Deep Residual Bidirectional Long Short-Term Memory Fusion architecture combines residual convolutional feature extraction with bidirectional sequential modeling. The residual component extracts spatial representations, while the bidirectional recurrent component processes contextual relationships among feature representations [18].

Recent studies have also examined hybrid convolutional and transformer-based architectures. A hybrid CNN–Vision Transformer model with an attention mechanism was evaluated on the CIFAKE dataset and achieved an accuracy of 99.77%. The model combines local convolutional representations with global contextual features from the Vision Transformer [23]. Extensible synthetic image detection has been investigated through adapter-based incremental learning. A content-agnostic adapter was integrated with a Vision Transformer to extract forensic features across different image generators. The framework also uses category-aware domain alignment and knowledge distillation to reduce performance degradation when new generator domains are introduced [24]. Zero-shot synthetic image detection has also been studied through forensic self-descriptions. The method learns pixel-level forensic microstructures from real images and uses these representations for zero-shot detection, open-set source attribution, and source-based clustering [25]. Transformer-based detection has also been developed through a patch-based Vision Transformer model. The PV-ISM architecture processes image patches and positional representations to distinguish authentic and AI-generated images [26]. Interpretability has also been incorporated through handcrafted forensic priors. The DC-BiNet framework combines dark-channel features and RGB representations through bilinear pooling. The method was evaluated across images produced by multiple generative models and reported an average

accuracy of 94.0% and an average precision of 98.1% [27].

The reviewed studies cover convolutional networks, residual architectures, attention mechanisms, Vision Transformers, fine-tuning, lightweight models, ensemble learning, incremental learning, zero-shot detection, explainability, and hybrid representation learning. However, the reviewed studies do not explicitly analyze gradient propagation within a Residual-BiLSTM architecture for synthetic image detection on CIFAKE. This study addresses this gap by evaluating classification performance, training behavior, and gradient-flow characteristics within the proposed Residual-BiLSTM model.

3. Methods

This section describes the dataset, image preprocessing and augmentation, proposed model architecture, and training configuration used in this study. The experimental procedure was designed to evaluate synthetic image classification performance and gradient behavior during model training.

3.1. Dataset

This study uses the CIFAKE dataset for binary classification of real and AI-generated images. CIFAKE contains 120,000 images divided into 100,000 training images and 20,000 test images. The dataset contains an equal number of real and synthetic images in each data partition. Each image has an original resolution of 32×32 pixels with three RGB channels. The real class contains camera-based images, while the synthetic class contains images generated using artificial intelligence. The dataset also includes several visual categories, including vehicles, animals, and natural scenes [13]. The balanced class distribution allows the classification models to be evaluated without applying an additional class-balancing procedure. The original image resolution provides the same input characteristics for all samples before preprocessing. Figure 1 presents examples of real and synthetic images from the CIFAKE dataset.

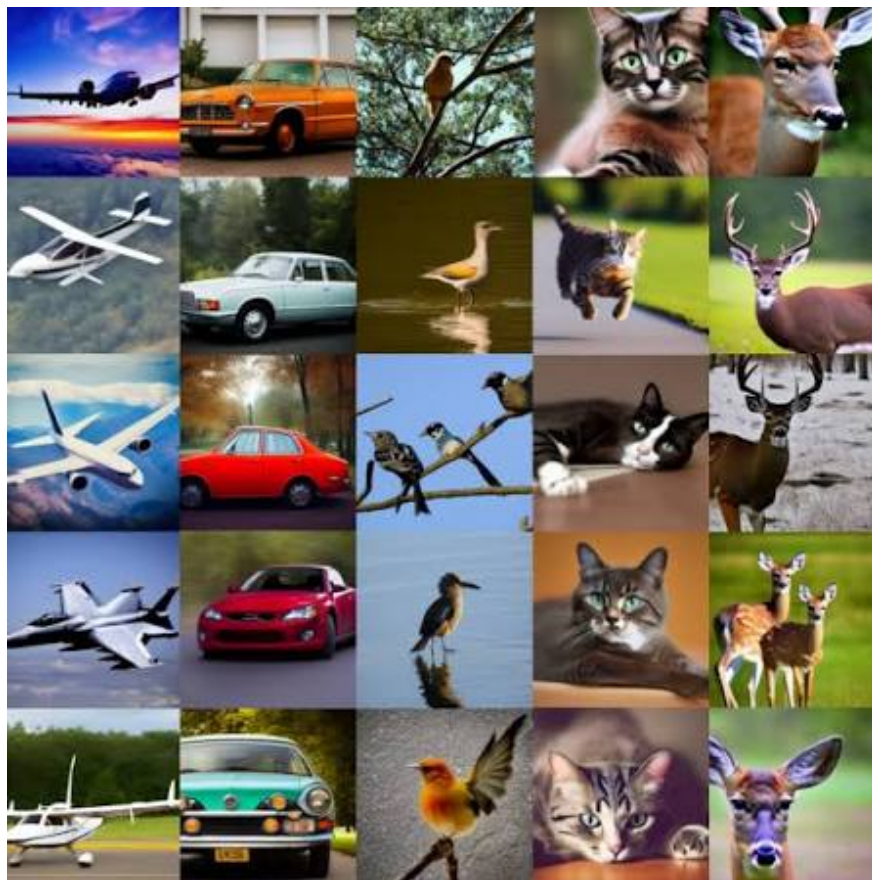


Figure 1. Examples of real and AI-generated images from the CIFAKE dataset. Source: Kaggle CIFAKE dataset [13].

3.2. Data Preprocessing and Augmentation

The CIFAKE images were preprocessed before model training. Each image was resized from its original resolution of 32×32 pixels to 224×224 pixels to match the input requirements of the residual network backbone. This resizing step provided a consistent spatial dimension for all training samples. Data augmentation was applied only to the training images. Random rotation with a maximum angle of 90° was used to expose the model to different image orientations [28]. Random sharpness adjustment was also applied to introduce variations in edge definition and texture intensity. After augmentation, each image was converted into a tensor representation. The tensor values were then normalized using the mean and standard deviation applied in the experimental pipeline. This normalization maintained a consistent numerical scale across training samples [29]. The preprocessing pipeline therefore consisted of image resizing, training-data augmentation, tensor conversion, and normalization. These steps produced standardized inputs for the proposed Residual-BiLSTM model.

3.3. Overview of the Proposed Model Architecture

The proposed model combines residual convolutional feature extraction with bidirectional sequential modeling. The architecture is adapted from a residual-based hybrid framework and is applied to synthetic image classification on the CIFAKE dataset [18]. The model consists of three main components: a Deep Spatial Network (DSN), a Deep Temporal Network (DTN), and a feature fusion module. The DSN extracts hierarchical spatial representations from the input images through convolution, activation, normalization, pooling, and residual operations. Residual connections are incorporated into the spatial network to provide shortcut paths for information and gradient propagation across deeper layers. The feature representations produced by the DSN are then processed by the DTN using a Bidirectional Long Short-Term Memory network. The BiLSTM processes the feature sequence in forward and backward directions to capture contextual relationships among the extracted representations. The outputs of the spatial and sequential components are then integrated through the feature fusion module. The fused representation is passed to the final classification layer to distinguish real images from AI-generated images. Figure 2 presents the structure of the proposed Residual-BiLSTM model.

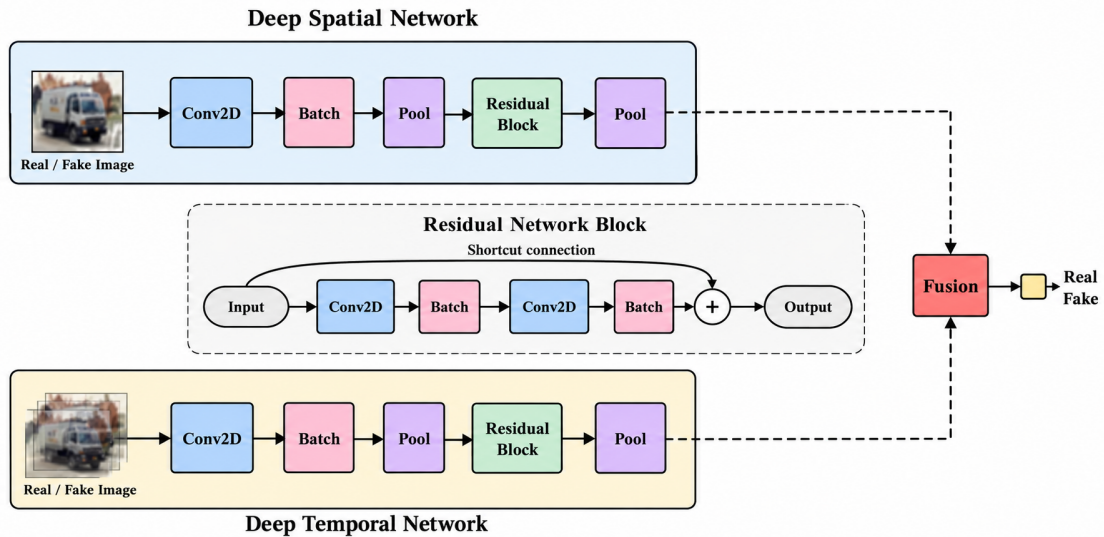


Figure 2. Architecture of the proposed Residual-BiLSTM model for synthetic image detection.

3.4. Deep Spatial Network

DSN extracts spatial features from the input image. The network applies convolution, nonlinear activation, residual learning, batch normalization, and pooling operations. The residual block contains a shortcut connection that provides an additional path for information and gradient propagation. Given an input image

X , the convolutional operation produces a feature map as follows:

$$F = W * X + b, \quad (1)$$

where W represents the convolutional kernel, $*$ represents the convolution operation, and b represents the bias term. The network then applies the Rectified Linear Unit activation function:

$$F' = \max(0, F). \quad (2)$$

The residual block combines the transformed representation with the original input through a shortcut connection:

$$Y = R(X) + X, \quad (3)$$

where $R(X)$ represents the transformation produced by the residual branch and X represents the shortcut input. Batch normalization is applied within the residual block. Pooling is then used to reduce the spatial dimensions of the feature maps. The resulting feature representation is passed to the next stage of the proposed model.

3.5. Deep Temporal Network

DTN processes the feature representations generated by the Deep Spatial Network. The extracted representations are arranged as a feature sequence before they are passed to the Bidirectional Long Short-Term Memory (BiLSTM) layer. The BiLSTM processes the sequence in forward and backward directions to capture contextual relationships among the feature representations.

For each time step t , the candidate cell state is calculated as:

$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c), \quad (4)$$

where x_t denotes the input feature at time step t , h_{t-1} denotes the previous hidden state, W_c denotes the weight matrix, and b_c denotes the bias term. The forget gate determines the information retained from the previous cell state:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \quad (5)$$

where W_f and b_f denote the weight matrix and bias term of the forget gate. The input gate determines the contribution of the current candidate state:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \quad (6)$$

where W_i and b_i denote the weight matrix and bias term of the input gate. The cell state is updated as:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \quad (7)$$

where $\odot$ denotes element-wise multiplication. The output gate is calculated as:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \quad (8)$$

where W_o and b_o denote the weight matrix and bias term of the output gate. The hidden state is then calculated as:

$$h_t = o_t \odot \tanh(c_t). \quad (9)$$

The BiLSTM generates hidden states from the forward and backward directions. These hidden states are concatenated as:

$$H_t = [\vec{h}_t; \overleftarrow{h}_t], \quad (10)$$

where $\vec{h}_t$ and $\overleftarrow{h}_t$ denote the forward and backward hidden states. An attention mechanism is applied to the BiLSTM output to assign different weights to the feature representations. The attention score is calculated as:

$$e_t = v^T \tanh(W_h H_t + b_h), \quad (11)$$

where W_h , b_h , and v denote trainable attention parameters. The attention weight is calculated using the softmax function:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^T \exp(e_k)}, \quad (12)$$

where T denotes the length of the feature sequence. The final attention-weighted representation is calculated as:

$$z = \sum_{t=1}^T \alpha_t H_t. \quad (13)$$

The resulting representation z is passed to the fusion and classification stages of the proposed model.

3.6. Training Configuration

The proposed model was trained using the hyperparameter configuration presented in Table 1. The training process was executed on a CUDA-enabled GPU. The Adam optimizer was used to update the model parameters

Table 1. Training configuration of the proposed model.

Parameter	Value
Number of epochs	50
Batch size	32
Learning rate	0.0001
Optimizer	Adam
Loss function	Cross-entropy loss

during training. The learning rate was set to 1×10^{-4} . Cross-entropy loss was used as the objective function for the binary classification task. The batch size was set to 32 samples for each training iteration. The training process was conducted for 50 epochs.

3.7. Evaluation Metrics

The classification performance was evaluated using accuracy, precision, recall, and F1-score. These metrics were calculated from the confusion matrix using true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. Accuracy measures the proportion of correctly classified samples among all evaluated samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (14)$$

Precision measures the proportion of correctly predicted positive samples among all samples predicted as positive:

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (15)$$

Recall measures the proportion of correctly identified positive samples among all actual positive samples:

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (16)$$

F1-score represents the harmonic mean of precision and recall:

$$\text{F1-score} = 2 \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right). \quad (17)$$

The confusion matrix was also used to examine the number of correct and incorrect predictions for the REAL and FAKE classes. Training and validation accuracy and loss curves were used to examine model behavior during training. Gradient-flow analysis was used to inspect the magnitude of gradients across the trainable layers.

4. Results and Discussion

4.1. Classification Performance

The proposed Residual-BiLSTM model was evaluated using 20,000 test images from the CIFAKE dataset. The test set contained 10,000 REAL images and 10,000 FAKE images. Table 2 presents the classification results for

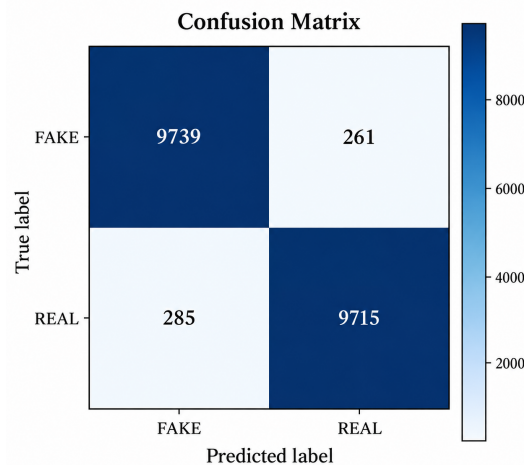
Table 2. Classification performance of the proposed Residual-BiLSTM model on the CIFAKE test set.

Class	Precision	Recall	F1-score	Support
FAKE	0.97	0.97	0.97	10,000
REAL	0.97	0.97	0.97	10,000
Macro average	0.97	0.97	0.97	20,000
Weighted average	0.97	0.97	0.97	20,000

both classes. The proposed model achieved an accuracy of 97.27% on the CIFAKE test set. The FAKE class obtained precision, recall, and F1-score values of 0.97. The REAL class also obtained precision, recall, and F1-score values of 0.97. The macro-average and weighted-average F1-scores were both 0.97. These results show that the model produced similar precision and recall values for the two evaluated classes. The balanced metric values indicate that the classification results were distributed similarly between REAL and FAKE images. However, the reported metrics alone do not describe the types of classification errors. The confusion matrix is therefore used to examine false predictions in each class.

4.2. Confusion Matrix Analysis

The confusion matrix was used to examine the distribution of correct and incorrect predictions for the REAL and FAKE classes. Figure 3 presents the classification results for the 20,000 test images. The model correctly classified 9,739 of 10,000 FAKE images. A total of 261 FAKE images were classified as REAL. The model also correctly classified 9,715 of 10,000 REAL images. A total of 285 REAL images were classified as FAKE. These results show that the numbers of misclassified samples were similar for both classes. The model produced 261 errors for the FAKE class and 285 errors for the REAL class. The difference between these two error counts was 24 samples. The confusion matrix complements the precision, recall, and F1-score results by showing the direction of the classification errors. The results indicate that errors occurred in both REAL-to-FAKE and FAKE-to-REAL predictions.

**Figure 3.** Confusion matrix of the proposed Residual-BiLSTM model on the CIFAKE test set.

4.3. Training and Validation Analysis

The training and validation curves were used to examine model behavior during the training process. Figure 4 presents the training loss, validation loss, training accuracy, and validation accuracy across the training epochs. The training loss decreased as the number of epochs increased. The validation loss also decreased during the training process. The difference between the training and validation loss remained limited during the observed epochs. The training accuracy increased during the training process. The validation accuracy followed a similar pattern and remained close to the training accuracy. The similarity between the two

accuracy curves indicates that the model produced comparable performance on the training and validation data. The loss and accuracy curves provide information about the convergence behavior of the proposed model. The observed patterns show that the optimization process progressed without a large separation between the training and validation curves.

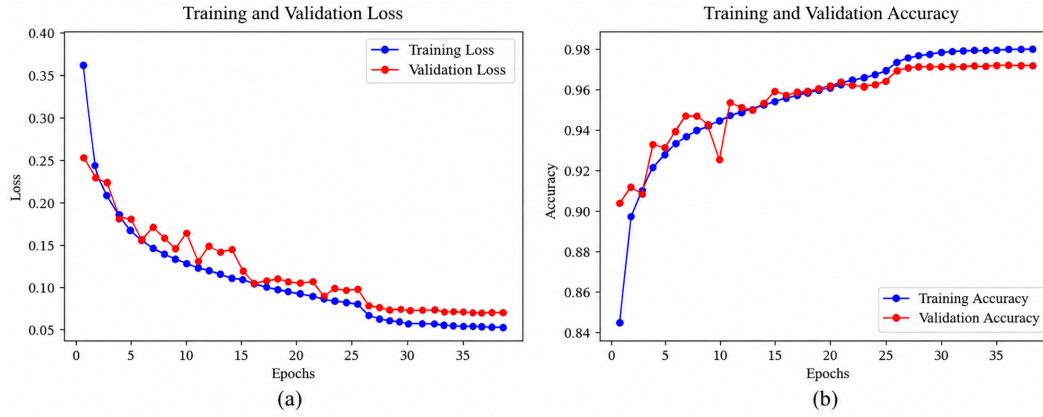


Figure 4. Training and validation performance of the proposed Residual-BiLSTM model: (a) training and validation loss, and (b) training and validation accuracy.

4.4. Comparison with Baseline and Previous Methods

The performance of the proposed Residual-BiLSTM model was compared with previously reported methods evaluated on the CIFAKE dataset. Table 3 summarizes the reported results. The proposed Residual-BiLSTM

Table 3. Comparison of synthetic image detection methods on the CIFAKE dataset.

Reference	Model	Metric	Result
[12]	CNN	Accuracy	91.00%
[13]	CNN	Accuracy	92.98%
[14]	ResNet	Accuracy	94.00%
[14]	VAE	Accuracy	71.00%
[15]	SE-ResNet50	Accuracy	96.12%
[15]	ResNet-50	Accuracy	93.46%
[15]	VGGNet-16	Accuracy	95.98%
[15]	DenseNet-121	Accuracy	96.06%
[17]	ResNet-152	F1-score	96.30%
[16]	CNN	Accuracy	96.31%
[16]	Vision Transformer	Accuracy	87.09%
Proposed method	Residual-BiLSTM	Accuracy	97.27%

model achieved an accuracy of 97.27%. The reported accuracy was higher than the accuracy values listed for the CNN, ResNet, SE-ResNet50, VGGNet-16, DenseNet-121, and Vision Transformer models in Table 3. The ResNet-152 result is reported using F1-score rather than accuracy, so it should not be interpreted as a direct accuracy comparison. The comparison shows that the proposed model produced the highest reported accuracy among the methods listed with accuracy as the evaluation metric. However, the studies used different model configurations and training procedures. Therefore, the numerical differences should be interpreted as a comparison of reported results rather than as a controlled head-to-head experiment.

4.5. Gradient Flow Analysis

Gradient-flow analysis was used to examine the behavior of gradients across the trainable layers of the proposed Residual-BiLSTM model. Figure 5 presents the gradient norm from the convolutional and residual

components to the BiLSTM and fully connected layers. The convolutional and residual layers maintain non-zero gradient values across the observed layers. The gradient magnitude changes across the residual blocks but does not decrease continuously toward zero. This pattern indicates that gradient information remains present during backpropagation through the spatial feature extraction stage. A local decrease in gradient magnitude appears in the BiLSTM layers. The gradient values increase again toward the fully connected layers. This pattern shows that gradient propagation continues through the recurrent and classification components.

The observed gradient trajectory does not show a continuous collapse of gradient values across the entire network. The result therefore provides empirical evidence that the proposed architecture maintains gradient propagation across the evaluated layers during the reported training process. The gradient-flow analysis complements the training and validation curves presented in Figure 4. The two analyses describe different aspects of the optimization process. The training curves describe changes in loss and accuracy across epochs, while the gradient-flow analysis describes gradient behavior across network layers.

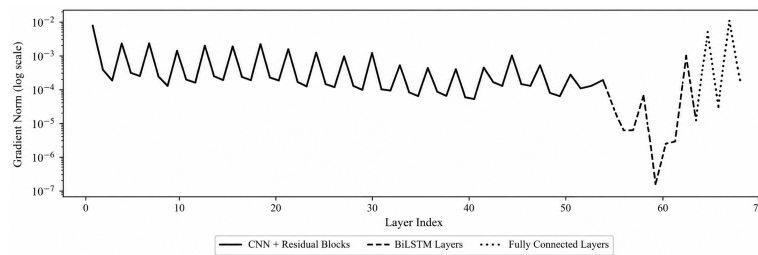


Figure 5. Gradient-flow behavior across the convolutional, residual, BiLSTM, and fully connected layers of the proposed model.

4.6. Discussion

The proposed Residual–BiLSTM model achieved an accuracy of 97.27% on the CIFAKE test set. Precision, recall, and F1-score values were similar for the REAL and FAKE classes. These results indicate that the model produced comparable classification performance for both classes. The confusion matrix shows that classification errors occurred in both directions. A total of 261 FAKE images were classified as REAL, while 285 REAL images were classified as FAKE. The similar number of errors across the two classes is consistent with the precision and recall values reported in Table 2.

The training and validation curves show that the loss values decreased during training. The accuracy values also increased across the observed epochs. The limited separation between the training and validation curves indicates similar performance trends on both data partitions. The comparison with previous CIFAKE studies shows that the proposed model achieved the highest reported accuracy among the methods listed with accuracy as the evaluation metric. However, the compared studies used different architectures, training settings, and experimental procedures. Therefore, the numerical comparison describes reported performance rather than a controlled comparison under identical conditions.

The gradient-flow analysis provides additional information about the optimization behavior of the proposed architecture. The convolutional and residual layers maintained non-zero gradient values across the observed layers. The BiLSTM layers showed a local reduction in gradient magnitude, while the gradients increased again toward the fully connected layers. This pattern indicates that gradient propagation continued across the evaluated components of the network. The combined results show that the Residual–BiLSTM architecture supports both classification and gradient propagation under the experimental setting used in this study. The residual component provides shortcut connections within the convolutional stage, while the BiLSTM component processes contextual relationships among the extracted feature representations. The gradient-flow analysis complements the classification metrics by describing the behavior of gradients across the network layers.

5. Conclusion and Future Work

This study proposed a Residual–BiLSTM model for synthetic image detection on the CIFAKE dataset. The model combines residual convolutional feature extraction with bidirectional sequential modeling and attention-

based representation learning. The proposed model achieved an accuracy of 97.27% on the CIFAKE test set. The REAL and FAKE classes each obtained precision, recall, and F1-score values of 0.97. The confusion matrix showed 261 FAKE images classified as REAL and 285 REAL images classified as FAKE. The training and validation curves showed decreasing loss values and increasing accuracy values across the training epochs. The gradient-flow analysis showed non-zero gradients across the convolutional, residual, BiLSTM, and fully connected components. A local decrease in gradient magnitude occurred in the BiLSTM layers, followed by an increase toward the fully connected layers. These results show that the proposed architecture produced consistent classification performance across the two classes and maintained gradient propagation under the experimental setting used in this study. The gradient-flow analysis also provides information about optimization behavior that is not represented by classification metrics alone. Future work can evaluate the proposed model on higher-resolution synthetic image datasets and datasets generated by different generative models. Cross-dataset evaluation can also be used to examine model behavior across different image distributions. Model compression and architectural simplification can be investigated to reduce computational requirements. Explainability methods can also be incorporated to identify visual regions that contribute to classification decisions.

Author Contributions

Conceptualization: K.M.I., B.H., and A.S.M.N.; Methodology: K.M.I.; Software: K.M.I.; Validation: B.H., A.S.M.N., and A.Y.K.K.; Formal analysis: K.M.I., B.H., and A.S.M.N.; Investigation: K.M.I. and A.Y.K.K.; Resources: B.H. and A.Y.K.K.; Data curation: M.M. and M.S.; Writing—original draft preparation: K.M.I.; Writing—review and editing: B.H., A.S.M.N., M.M., M.S., and A.Y.K.K.; Visualization: K.M.I.; Supervision: M.M. and M.S.; Project administration: A.S.M.N. and A.Y.K.K.; Funding acquisition: Not applicable. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

The CIFAKE dataset used in this study is publicly available on Kaggle [?].

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

ChatGPT, developed by OpenAI, was used to assist with language editing and manuscript formatting. The authors reviewed all generated content and remain fully responsible for the accuracy and integrity of the manuscript.

References

- [1] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-Resolution Image Synthesis with Latent Diffusion Models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2022. p. 10674-85. Available from: <https://doi.org/10.1109/CVPR52688.2022.01042>.
- [2] Ramesh A, et al. Zero-Shot Text-to-Image Generation. In: Proceedings of the 38th International Conference on Machine Learning. vol. 139 of Proceedings of Machine Learning Research. PMLR; 2021. p. 8821-31. Available from: <https://proceedings.mlr.press/v139/ramesh21a.html>.
- [3] Saharia C, et al.. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding; 2022. Available from: <https://arxiv.org/abs/2205.11487>.

- [4] Roose K. An AI-Generated Picture Won an Art Prize. Artists Aren't Happy; 2022. The New York Times. Available from: <https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html>.
- [5] Singh B, Sharma DK. Predicting Image Credibility in Fake News on Social Media Using a Multimodal Approach. *Neural Computing and Applications*. 2022;34(24):21503-17. Available from: <https://doi.org/10.1007/s00521-021-06086-4>.
- [6] Pennycook G, Rand DG. The Psychology of Fake News. *Trends in Cognitive Sciences*. 2021. Available from: <https://doi.org/10.1016/j.tics.2021.02.007>.
- [7] Bonettini N, Bestagini P, Milani S, Tubaro S. On the Use of Benford's Law to Detect GAN-Generated Images. In: 2020 25th International Conference on Pattern Recognition. IEEE; 2021. p. 5495-502. Available from: <https://doi.org/10.1109/ICPR48806.2021.9412944>.
- [8] Deb D, Zhang J, Jain AK. AdvFaces: Adversarial Face Synthesis. In: 2020 IEEE International Joint Conference on Biometrics. IEEE; 2020. p. 1-10. Available from: <https://doi.org/10.1109/IJCB48548.2020.9304898>.
- [9] Khosravy M, Nakamura K, Hirose Y, Nitta N, Babaguchi N. Model Inversion Attack: Analysis under a Gray-Box Scenario on a Deep Learning-Based Face Recognition System. *KSII Transactions on Internet and Information Systems*. 2021;15:1100-19. Available from: <https://doi.org/10.3837/tiis.2021.03.015>.
- [10] Bird JJ, Naser A, Lotfi A. Writer-Independent Signature Verification: Evaluation of Robotic and Generative Adversarial Attacks. *Information Sciences*. 2023;633:170-81. Available from: <https://doi.org/10.1016/j.ins.2023.03.029>.
- [11] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition; 2015. Available from: <http://arxiv.org/abs/1512.03385>.
- [12] Maruthiram KB, Venkataramireddy G, Klick M. Real vs. AI-Generated Image Detection and Classification. *International Journal of Innovative Research in Technology*. 2024;11(2):1076-80. Available from: <https://ijirt.org/article?manuscript=166462>.
- [13] Bird JJ, Lotfi A. CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. *IEEE Access*. 2024;12:15642-50. Available from: <https://doi.org/10.1109/ACCESS.2024.3356122>.
- [14] Bartos GE, Akyol SO, Akyol S. Deep Learning for Image Authentication: A Comparative Study on Real and AI-Generated Image Classification; 2023. Available from: <https://www.researchgate.net/publication/375952278>.
- [15] Gunukula AR, Das Gupta H, Sheng VS. Detecting AI-Generated Images Using a Hybrid ResNet-SE Attention Model. *Applied Sciences*. 2025;15(13):7421. Available from: <https://doi.org/10.3390/app15137421>.
- [16] Hossain MZ, Uzzaman F, Islam MR. Advancing AI-Generated Image Detection: Enhanced Accuracy through CNN and Vision Transformer Models with Explainable AI Insights. In: International Conference on Computer and Information Technology. IEEE; 2023. p. 1-6. Available from: <https://doi.org/10.1109/ICCIT60459.2023.10440990>.
- [17] Hakim S, et al. Classification of Artificial Intelligence-Generated Images Using the Fine-Tuning Method on a Residual Network. *Journal of Information Technology and Computer Science*. 2024;11:655-66. Available from: <https://doi.org/10.25126/jtiik.1138118>.
- [18] Munsarif M, Ku-Mahamud KR. Deep Residual Bidirectional Long Short-Term Memory Fusion: Achieving Superior Accuracy in Facial Emotion Recognition. *Bulletin of Electrical Engineering and Informatics*. 2025;14(3):2143-55. Available from: <https://doi.org/10.11591/eei.v14i3.9090>.

- [19] Lokner Ladević A, Kramberger T, Kramberger R, Vlahek D. Detection of AI-Generated Synthetic Images with a Lightweight CNN. *AI*. 2024;5(3):1575-93. Available from: <https://doi.org/10.3390/ai5030076>.
- [20] Islam MT, Lee IH, Alzahrani AI, Muhammad K. MEXFIC: A Meta Ensemble eXplainable Approach for AI-Synthesized Fake Image Classification. *Alexandria Engineering Journal*. 2025;116:351-63. Available from: <https://doi.org/10.1016/j.aej.2024.12.031>.
- [21] Sharma K. Explainable AI-Generated Image Forensics: A Low-Resolution Perspective with Novel Artifact Taxonomy. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*; 2025. p. 1587-96.
- [22] Zhang Y, Pang Z, Huang S, Wang C, Zhou X. Unmasking AI-Created Visual Content: A Review of Generated Images and Deepfake Detection Technologies. *Journal of King Saud University Computer and Information Sciences*. 2025;37:148. Available from: <https://doi.org/10.1007/s44443-025-00154-8>.
- [23] Weir S, Khan MS, Moradpoor N, Ahmad J. Enhancing AI-Generated Image Detection with a Novel Approach and Comparative Analysis. In: *2024 17th International Conference on Security of Information and Networks (SIN)*. IEEE; 2025. Available from: <https://doi.org/10.1109/SIN63213.2024.10871900>.
- [24] Tang S, He P, Li H, Wang W, Jiang X, Zhao Y. Towards Extensible Detection of AI-Generated Images via Content-Agnostic Adapter-Based Category-Aware Incremental Learning. *IEEE Transactions on Information Forensics and Security*. 2025;20:2883-98. Available from: <https://doi.org/10.1109/TIFS.2025.3546845>.
- [25] Nguyen TD, Azizpour A, Stamm MC. Forensic Self-Descriptions Are All You Need for Zero-Shot Detection, Open-Set Source Attribution, and Clustering of AI-Generated Images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2025. p. 3040-50. Available from: <https://doi.org/10.1109/CVPR52734.2025.00289>.
- [26] Çınar O, Doğan Y. Novel Deepfake Image Detection with PV-ISM: Patch-Based Vision Transformer for Identifying Synthetic Media. *Applied Sciences*. 2025;15(12):6429. Available from: <https://doi.org/10.3390/app15126429>.
- [27] Tan D, Niu C, Yang Y, Yang D, Tan B. DC-BiNet: Towards Interpretable Generated Image Detection with Dark Channel Prior. *Expert Systems with Applications*. 2025;280:127508. Available from: <https://doi.org/10.1016/j.eswa.2025.127508>.
- [28] Mumuni A, Mumuni F. Data Augmentation: A Comprehensive Survey of Modern Approaches. *Array*. 2022;16:100258. Available from: <https://doi.org/10.1016/j.array.2022.100258>.
- [29] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift; 2015. Available from: <https://arxiv.org/abs/1502.03167>.