iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.77910

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

Comparative Analysis of Wavelet-FFT-ANN and ResNet-Based Models for Real-Time Audio Deepfake Detection Using the In-the-Wild Dataset

Sumedha Arya

Abstract: ***Nowadays, AI-generated content is in a high trend because of advancement in technology. Deepfake audio clips looks highly realistic, posing serious concern to ethics, privacy, and security. Although, there are so many techniques has been built to classify fake from real speech, but still there exists issues. This is due to noise, compression, and speaker variations. In this study, we evaluate four different models on the "In-the-Wild" audio deepfake dataset using a balanced subset of 20,000 samples standardized to 16 kHz and 2-second clips. Two models use hand-crafted features (Wavelet + ANN and FFT + ANN), while two models apply deep transfer learning using ResNet18 and fully fine-tuned ResNet50 on log-Mel spectrograms. Experimental results show that traditional ANN models achieve moderate performance, from 58%–75% in accuracy with higher false claims. In contrast to them, deep learning models shown better generalization, with ResNet18 reaching to 97% accuracy and ResNet50 achieving the best performance at 98.9% accuracy with near-perfect F1-scores. These findings highlight that spectrogram-based representations combined with powerful pre-trained CNN architectures provide a more robust and reliable solution for real-world audio deepfake detection.***

Keywords: Audio Deepfake Detection, In-the-Wild Dataset, ResNet50, Transfer Learning, Wavelet Transform, FFT, Mel-Spectrogram, Artificial Neural Network

I. INTRODUCTION

The rapid growth of generative AI has made it possible to create similar content as that of text, images, videos, and audios [1] [3]. This is called as Deepfake technology. The speech can be created and converted in real time, which looks like real. While this technology has useful applications, it also creates serious ethical, privacy and security concerns. AI-generated speech can be used to impersonate individuals, and spread false information. Therefore, there is a strong need for real-time systems that can detect AI-generated or DeepFake voice content. Initially, misinformation was mainly limited to fake news articles, but now it includes synthetic audio and video generated using advanced algorithms. These deepfakes can negatively impact society in many ways. For example, in politics, fake audio or videos can influence voters during elections [4], [5]. Deepfakes also raise moral concerns, such as the misuse of public figures' identities without consent [2], [6]. In legal contexts, synthetic data could even be used to create false digital evidence, potentially affecting court decisions. Overall, deepfakes threaten data credibility because they can make it appear that someone said or did something that actually never happened. The commercialization of AI has further accelerated the large-scale creation of digital replicas of human behavior [7], [8].

To address these challenges, this study uses the publicly available dataset, In-the-Wild audio, which contains both real and AI-generated speech samples. In this research, four different models are applied for deepfake speech detection. The first model uses wavelet transformation to extract features from the audio signals, followed by classification using an Artificial Neural Network (ANN). The second model uses Fourier transformation to extract features from the audio signals, followed by classification using an ANN. The third model is a hybrid architecture, using Fourier transformation features, with Resnet18, followed by classification using an ANN. The fourth model is based on the deep learning architecture ResNet50. The main objective of this study is to compare the performance of these four models using evaluation metrics such as accuracy, precision, recall, and F1-score, in order to determine which approach is more effective for detecting AI-generated speech. The paper is further divided into following sections; literature review, research methodology, results analysis, conclusion and references.

II. LITERATURE REVIEW

In this section, we highlighted the latest recent done on audio deepfake detection using both traditional machine learning and deep learning techniques.

Authors [9] in their research investigated deepfake audio detection in group conversations using the UrbanSound8K dataset along with conversational data from OpenSubtitles. In data preprocessing, the included sample rate conversion, channel merging, and Mel-Frequency Cepstral Coefficient (MFCC) extraction. The pre-processed dataset was split into training and testing, and various architectures such as Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), and Deep Neural Network (DNN) were applied on it. Transfer learning was also used by applying VGG19 to improve detection performance. The CNN model achieved a highest testing accuracy of 89%, demonstrating the effectiveness of deep learning for fake speech detection. However, the study highlighted limitations such as the need for better signal denoising, improved speaker diarization, and automation in speaker segmentation. In another research, authors [10] focused on MFCC-based feature extraction combined with machine learning models for deepfake detection using the Fake-or-Real (FoR) dataset. It comprises of 195,000 real and synthetic speech samples. They created four subsets of the dataset (for-original, for-norm, for-2sec, and for-rerec) and performed preprocessing steps such as sampling rate normalization, volume adjustment, and duplicate removal. Several models were tested, including traditional machine learning models such as SVM, XGBoost, Random Forest (RF), KNN, and deep learning models as LSTM, and VGG-16. Among these, VGG-16 achieved the highest accuracy of 93% on the for-original subset. The study also highlighted the challenge of high dimensionality in certain subsets. They emphasized the importance of dimensionality reduction for robust fake speech detection.
Authors [11] addressed AI-generated fake speech detection using the DEEP-VOICE dataset. For data pre-processing, they applied exploratory data analysis (EDA) to manage outliers and missing values. Further, they also extracted features such as Chroma-STFT and Root Mean Square (RMS) energy. After preprocessing steps including resampling and normalization, a traditional machine learning model as Random Forest classifier with 5-fold cross-validation was used which achieved an accuracy of 98.5%. While the results were strong, the study highlighted limitations in terms of scalability, unseen data generalization, and reliance on manually engineered features. A hybrid framework combining machine learning and deep learning approaches, was proposed by the authors [12] in their work using the Fake-or-Real (FoR) dataset. In the feature-based approach, spectral features were extracted and classified using tradionla machine learning algorithms such as SVM, LightGBM (LGBM), XGBoost, KNN, and Random Forest. In the image-based approach, Mel-spectrograms were generated using librosa and classified using deep leraning models such as Temporal Convolutional Network (TCN) and Spatial Transformer Network (STN). Machine learning models achieved accuracies between 60% and 70%, showing poor generalization on FoR dataset. On the other hand, deep learning model such as TCN achieved an overall accuracy of 92%, outperforming STN which gave an accuracy of 80%. However, the study also identified certain limitations such as the exclusion of important spectral representations such as Short-Time Fourier Transform (STFT) and MFCC in the image-based framework. Overall, existing studies demonstrate that deep learning models, especially based on CNN and transfer learning techniques, generally outperform traditional machine learning models in audio deepfake detection. However, still challenges remain in terms of feature generalization, dimensionality reduction, noise robustness, and scalability. Therefore, the motivation of this research is to build more robust and generalized deepfake audio detection frameworks.

III.RESEARCH METHODOLOGY

In this section, we detail about the methodology used for fake audio classification on the In-the-Wild dataset, source from Kaggle. The four different algorithms with two primary paradigms: (i) hand-crafted frequency-domain feature extraction coupled with shallow artificial neural networks (ANN), and (ii) deep transfer learning applied to time-frequency image representations were used in this study. By keeping the dataset, evaluation metrics, and training hyperparameters same across all models, the study enables a controlled and fair comparison of these algorithms.

A. Dataset Description

1) In-the-Wild audio deepfake collection, approximately 31,779 audio files were present with 19,963 as real and 11,816 as fake.
2) 20,000 samples using stratified random sampling was created as 10,000 real and 10,000 fake to ensure data balancing.
3) The balanced dataset is further stratified split into train and test with a ratio of 80:20 having 16,000 training samples and 4,000 test samples.

B. Audio Standardization

All waveforms are loaded using torchaudio.load() and normalized to a canonical format before any feature extraction:

1) The target sampling rate was selected as 16,000 Hz with a fixed duration of 2 seconds with 32,000 samples. Therefore, longer signals are truncated, and shorter signals are zero-padded.
2) Multi-channel audio is converted to mono by averaging across channels.

C. Feature Extraction Pipelines

Two distinct feature representation strategies are explored, each used by two models, yielding four model–feature combinations in total.

1) Pipeline A — Wavelet-Domain Features (Models M1)

This pipeline extracts multi-resolution features from denoised speech signals using the Discrete Wavelet Transform (DWT).

a) Wavelet denoising is performed using DWT with Daubechies-4 (db4) wavelet at decomposition level 4. Noise level is estimated with the Median Absolute Deviation (MAD) of the finest detail coefficients: $\sigma = \text{median}(|cD_4|) / 0.6745$. A universal VisuShrink threshold $\lambda = \sigma\sqrt{(2 \log N)}$ is applied via soft thresholding to all detail sub-bands (levels 1–4), followed by inverse DWT reconstruction.

b) In feature extraction, the same DWT (db4, level 4) is applied to the denoised signal; all approximation and detail coefficient vectors are concatenated into a single 1-D feature vector.

c) FastICA with n_components as10, whiten as 'unit-variance', random_state as 42 was used to perform dimentionality reduction which reduces the high-dimensional wavelet coefficient vector to 10 independent components.

d) Standard scaler was applied after ICA; both transformations are fitted exclusively on the training set.

2) Pipeline B — FFT Magnitude Spectrum Features (Models M2)

This pipeline extracts compact frequency-domain features from the raw waveform using the Fast Fourier Transform.

a) Spectral feature extraction is performed by one-sided real FFT by applying it to the 32,000-sample waveform, yielding an 16,001 positive-frequency magnitude bins.

b) Dimensionality reduction is performed using FastICA with n_components as10, whiten as 'unit-variance', and random_state as 42 is applied to reduce the 16k-dimensional FFT vector to 10 independent components.

c) Standard scaler was applied after ICA; both transformations are fitted exclusively on the training set.

3) Pipeline C — Log-Mel Spectrogram Image Representations (Models M3 & M4)

Both deep learning models share the same spectrogram image representation pipeline:

a) Mel-spectrogram computation is perfomed using torchaudio.MelSpectrogram with parameters n_fft as 1024, hop_length as 512, n_mels as 128, followed by amplitude-to-decibel conversion using torchaudio.transforms.AmplitudeToDB().

b) Image preparation was done with min-max normalization to [0, 1]; conversion to uint8 grayscale; replication to 3-channel RGB (PIL.Image.convert('RGB')); bilinear resize to 224 × 224 pixels.

c) ImageNet statistics applied for M3/M4 to perform normalization using transforms.Normalize(mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]).

D. Classification Models

Four models are evaluated in total, as summarized in Table 1. Each produces a single sigmoid-activated output; binary classification is performed using a decision threshold of 0.5.

Table 1. Summary of the four evaluated model configurations.

Model	Feature Representation	Architecture	Trainable Params	Optimizer
ANN-Wavelet (M1)	10 ICA components (DWT db4 coefficients)	MLP: 10→10→7→1 (ReLU + Sigmoid)	~few hundred	Adam / BCE
ANN-FFT (M2)	10 ICA components (FFT magnitude spectrum)	MLP: 10→16→8→1 (ReLU + Sigmoid)	~few hundred	Adam / BCE
ResNet18-ANN (M3)	224×224 RGB log-Mel spectrogram (frozen ResNet18)	Frozen ResNet18 + MLP: 512→256→64→1	~200K (head only)	Adam / BCE
ResNet50-Spectrogram (M4)	224×224 RGB log-Mel spectrogram (full fine-tuning)	Full ResNet50 + fc: Linear→Sigmoid	~25M (full)	Adam / BCE

1) M1 — ANN-Wavelet (Shallow ANN on Wavelet-ICA Features)

A three-layer fully-connected network receives the 10 ICA components derived from wavelet coefficients (Pipeline A). Architecture: Linear(10→10) → ReLU → Linear(10→7) → ReLU → Linear(7→1) → Sigmoid. Total trainable parameters are minimal (~few hundred), making this the most computationally lightweight model.

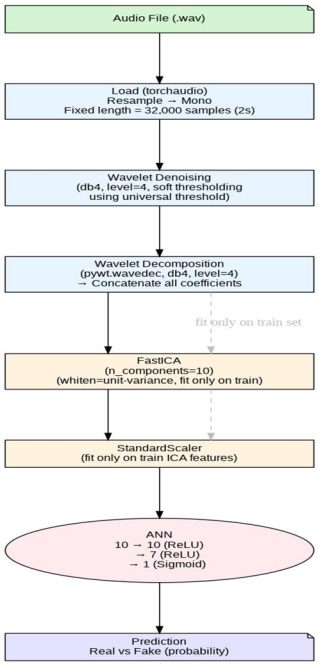


Fig. 1: ANN-Wavelet Model Architecture

2) M2 — ANN-FFT (Shallow ANN on FFT-ICA Features)

A three-layer fully-connected network operates on 10 ICA components derived from the FFT magnitude spectrum (Pipeline B). Architecture: Linear(10→16) → ReLU → Linear(16→8) → ReLU → Linear(8→1) → Sigmoid. This model mirrors M1 in scale but uses a fundamentally different frequency-domain representation.

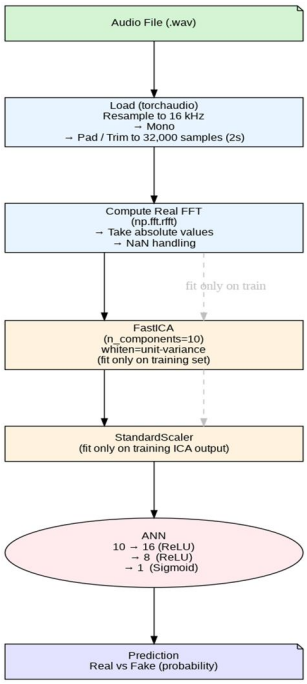


Fig. 2: ANN- Fourier Model Architecture

3) M3 — Hybrid ResNet18 + ANN (Frozen Backbone Transfer Learning)

A pre-trained ResNet18 (ImageNet weights, IMAGENET1K_V1) serves as a frozen feature extractor. The original fully-connected layer is replaced with nn.Identity(), yielding 512-dimensional feature vectors. A trainable three-layer MLP classification head is appended: Linear(512→256) → ReLU → Dropout(0.3) → Linear(256→64) → ReLU → Linear(64→1) → Sigmoid. Only the ANN head parameters (~200K) are optimized; ResNet18 weights are frozen throughout training (torch.no_grad() during the backbone forward pass), reducing computational requirements significantly relative to full fine-tuning.

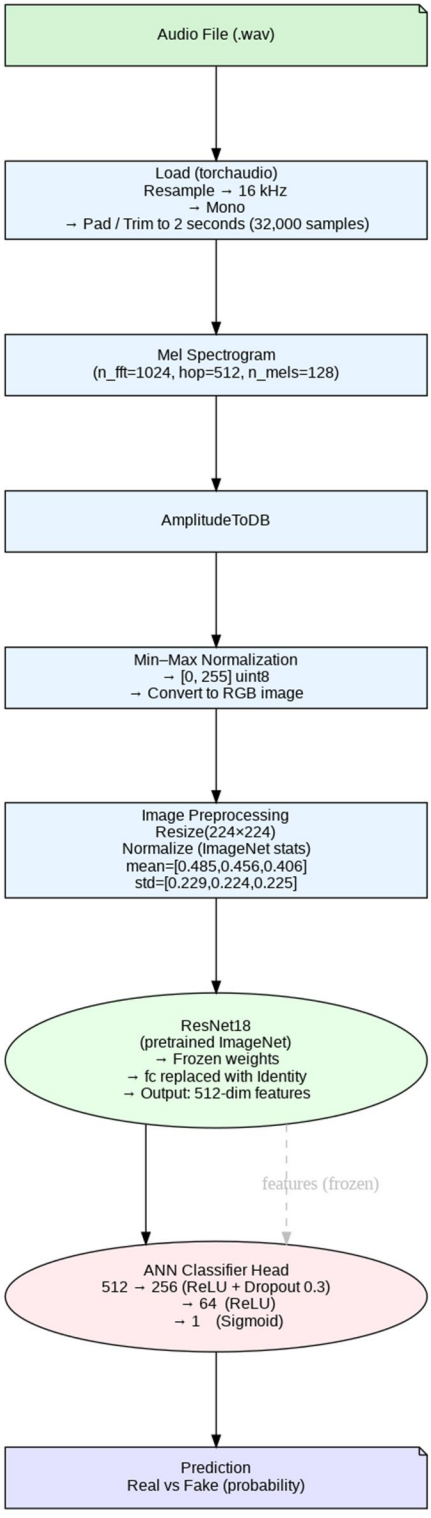


Fig. 3: Hybrid Model Architecture

4) M4 — ResNet50 Full Fine-Tuning

A full ResNet50 pre-trained on ImageNet is fine-tuned end-to-end. The original classification layer is replaced with a single linear layer followed by a sigmoid activation. All approximately 25 million parameters are trainable, making this the most expressive and computationally intensive model in the comparison.

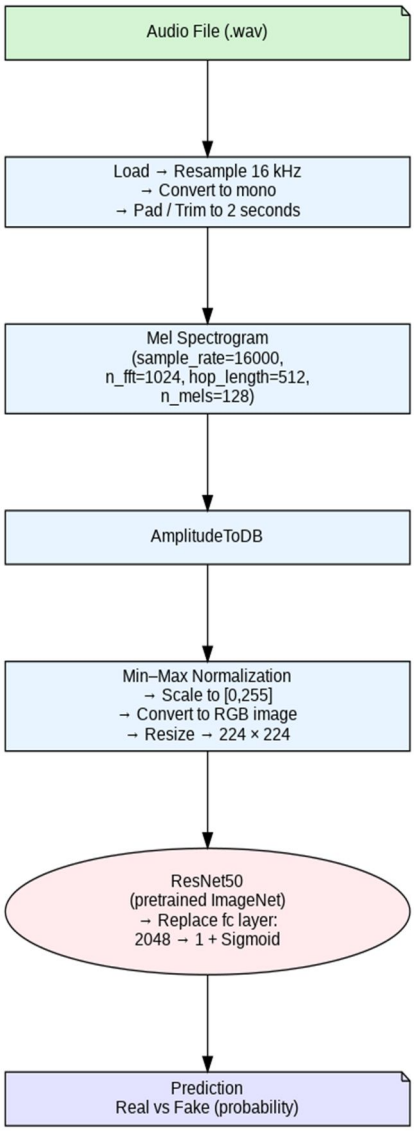


Fig. 4: ResNet-50 Model Architecture

E. Training Procedure

All four models share identical training hyperparameters to ensure that observed performance differences are attributable to architecture and feature representation rather than optimization settings.

1) Loss function: Binary Cross Entropy (nn.BCELoss).
2) Optimizer: Adam (lr = 0.001, fixed; no scheduler).
3) Batch size: 32.
4) Training epochs: 5.
5) Device: CUDA GPU if available; CPU otherwise.
6) Training loop: standard PyTorch loop with tqdm progress monitoring.

No early stopping, learning rate scheduling, weight decay, or data augmentation was applied in the reference implementation, ensuring a controlled baseline comparison. The exception is M3, which incorporates a Dropout (0.3) layer in its classification head as part of the architecture design.

F. Evaluation Protocol

All models are evaluated on the identical held-out test set of 4,000 balanced samples. The following metrics are reported for each model:

1) Confusion matrix (2×2: Real vs. Fake).
2) Precision, Recall, and F1-score per class (Real / Fake) and macro-averaged.
3) Overall accuracy.
4) Training loss curve (per epoch).

Results are generated using sklearn.metrics.confusion_matrix and classification_report. Models are compared across all metrics to determine which paradigm—hand-crafted low-dimensional features with shallow networks versus deep transfer learning on spectrogram images—is better suited for audio deepfake detection on the In-the-Wild benchmark.

IV. RESULTS ANALYSIS

This section presents the performance comparison of four models evaluated on the In-the-Wild audio deepfake dataset. All models were trained under the same conditions:

1) 20,000 balanced samples (10,000 real and 10,000 fake)
2) 80:20 stratified train-test split
3) 5 training epochs
4) Adam optimizer (learning rate = 0.001)
5) Binary cross-entropy loss
6) Batch size = 32

The results show a clear difference between:

- Shallow ANN models using hand-crafted features (M1, M2)
- Deep transfer learning models using spectrogram images (M3, M4)

Overall accuracy ranges from **58.2% (M1)** to **98.9% (M4)**, creating a performance gap of more than 40 percent. This large gap highlights the importance of feature representation in audio deepfake detection. Both deep learning models (M3 and M4) achieve more than 97% accuracy within only five epochs, while shallow ANN models show lower and less stable performance.

A. Per-Class Classification Performance

To better understand model behavior, we analyze performance separately for Real and Fake classes.

1) M1 — ANN with Wavelet Features

M1 shows strong imbalance between the two classes.

- Real class F1-score = 0.47
- Fake class F1-score = 0.65
- Real class recall = 0.38

This means 62% of real speech samples are wrongly classified as fake, which makes the model unreliable for practical use. The likely reason is that wavelet denoising and ICA compression remove important speech features needed for accurate classification.

2) M2 — ANN with FFT Features

M2 performs much more consistently:

- Real F1-score = 0.75
- Fake F1-score = 0.74

Precision and recall are balanced across both classes. The FFT magnitude spectrum preserves frequency information better than wavelet coefficients under the same compression level, leading to improved discrimination. However, its overall performance is still significantly less than the deep transfer learning models.

3) M3 — ResNet18 + ANN

M3 achieves near-perfect performance:

- F1-score = 0.97 for both classes
- Balanced precision and recall

This indicates that spectrogram images combined with pre-trained convolutional features provide highly discriminative representations.

4) M4 — ResNet50

M4 achieves the best performance:

- Accuracy = 98.9%
- F1-score = 0.99 for both classes
- Real recall = 1.00
- Fake precision = 1.00

Only 4 real samples are wrongly flagged as fake out of 2,000 real test samples. This demonstrates extremely strong separation between real and synthetic speech.

B. Confusion Matrix Analysis

The confusion matrix gives insight into absolute error counts.

1) M1 — ANN-Wavelet

- 1,249 false positives (real predicted as fake)
- 425 false negatives

This corresponds to a 62.5% false alarm rate on real speech, which is unacceptable for real-world deployment.

2) M2 — ANN-FFT

- 488 false positives
- 529 false negatives

This shows moderate discriminative ability but still too many errors for high-security applications.

3) M3 — ResNet18 + ANN

- Approximately 60 false positives
- Approximately 40 false negatives

Error rate is below 3% in both directions.

4) M4 — ResNet50 FT

- 4 false positives (0.2% false alarm rate)
- 40 false negatives (2.0% miss rate)

This indicates near-perfect classification on the test distribution.

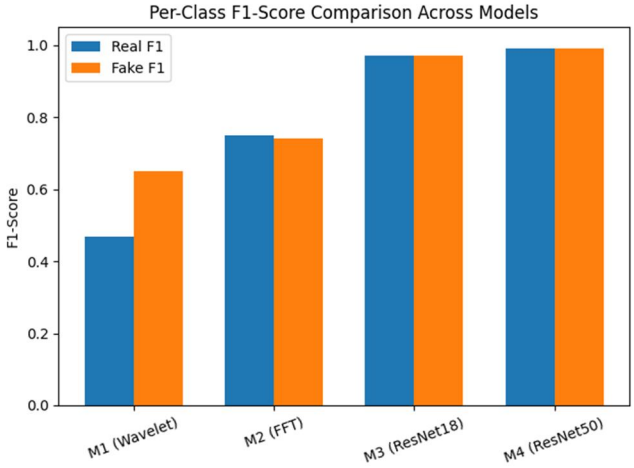


Fig. 5: Per-Class F1 Score Comparison Across Models

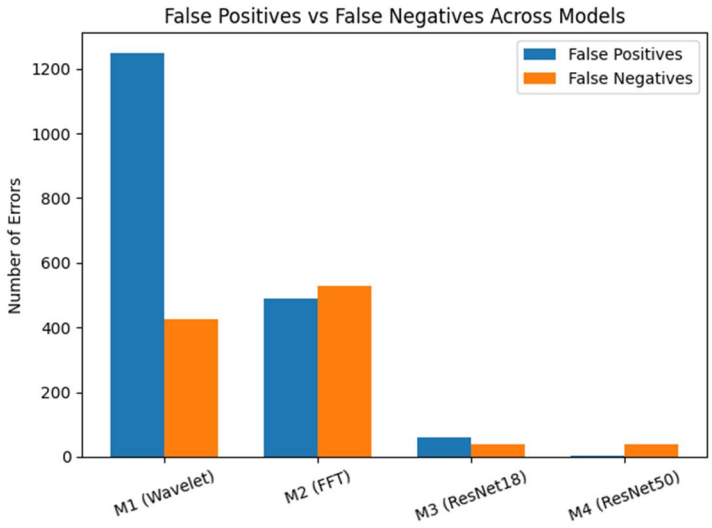


Fig. 6: False Positives vs False Negatives Across Models

C. Training Behavior and Computational Cost

Training stability and computational efficiency were also evaluated.

1) M1 — ANN-Wavelet

- Final loss = 0.742
- Loss increases in later epochs

This indicates divergence and unstable learning. The wavelet+ICA features likely remove important spectral information, leading to weak gradients.

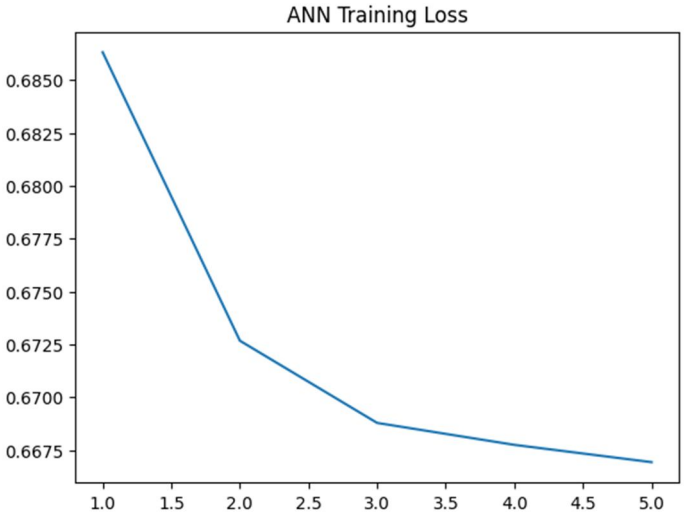


Fig. 7: ANN Training Loss

2) M2 — ANN-FFT

- Final loss = 0.583
- Non-monotonic convergence

Although somewhat unstable, the FFT features still allow partial learning.

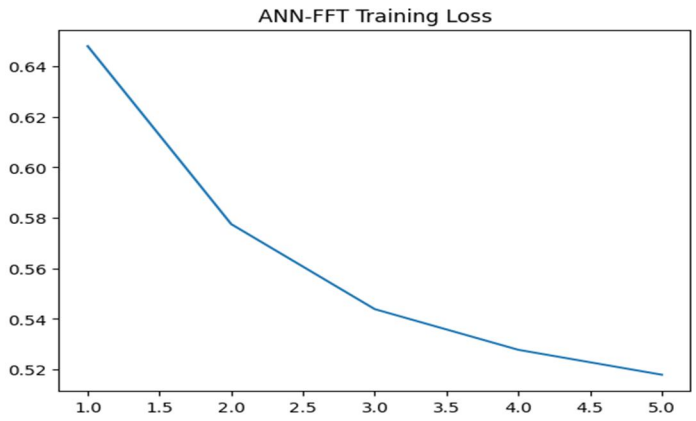


Fig. 8: ANN FFT Training Loss

3) M3 — ResNet18 + ANN

- Final loss = 0.151
- Fast and stable convergence

The frozen ResNet18 backbone provides strong 512-dimensional embeddings, requiring only minimal adaptation.

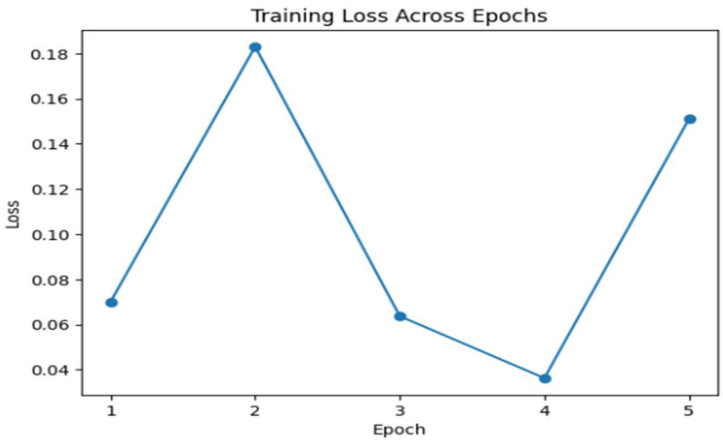


Fig. 9: Resnet18 - ANN Training Loss

4) M4 — ResNet50 FT

- Final loss = 0.013
- Most stable convergence

Although it requires longer training time, the performance gain justifies the additional computation in GPU environments.

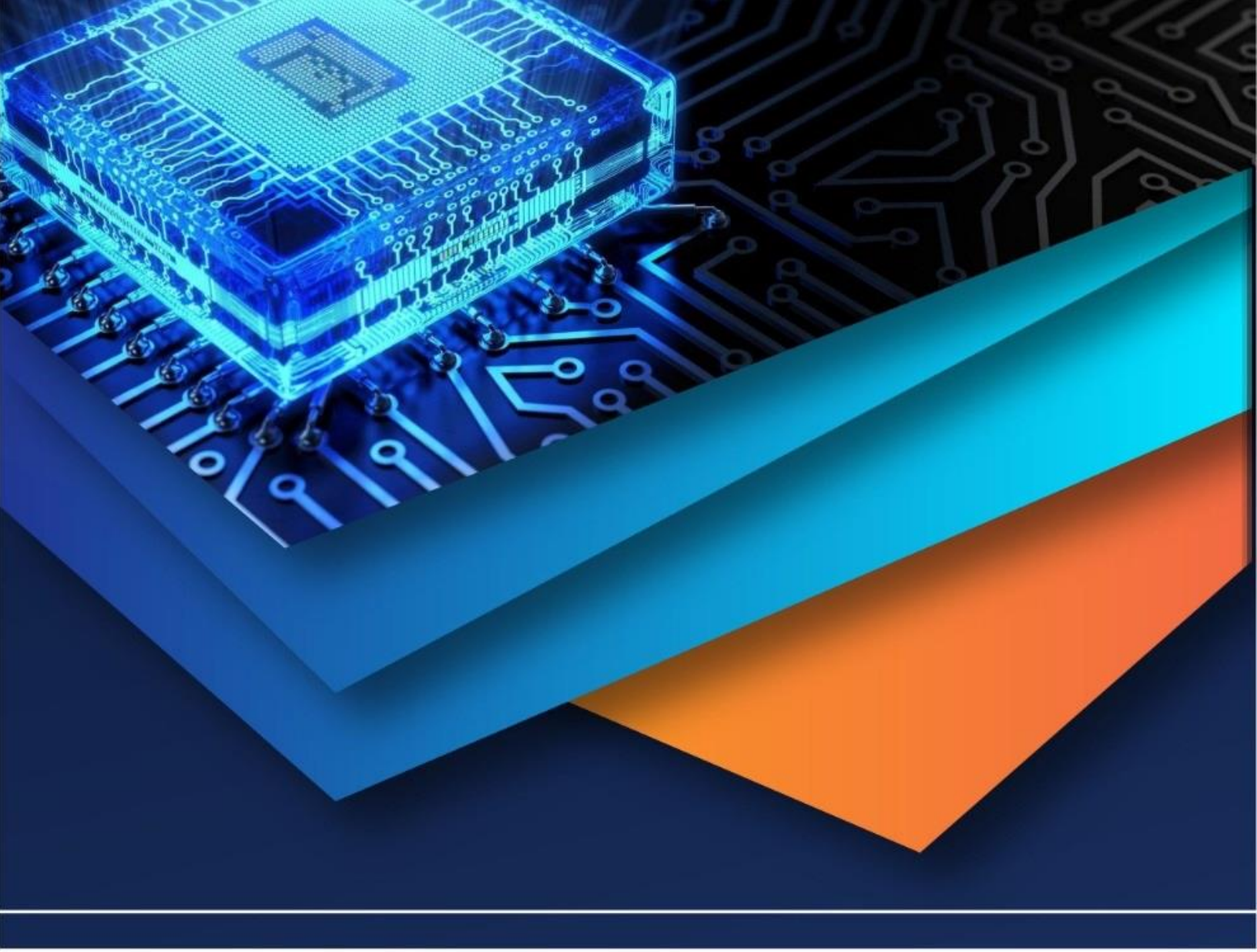
Fig. 10: Resnet-50 Training Loss

V. CONCLUSION

This study compared four different methods, wavelet, Fourier, Resnet-18 and 50 for detecting audio deepfakes using the challenging "In-the-Wild" dataset. The results clearly show that deep learning methods perform much better. The wavelet and FFT-based ANN models had lower accuracy, especially in detecting real speech correctly. In contrast, the ResNet-based models achieved very high performance within just five training epochs. These results show that spectrogram-based representations combined with powerful pre-trained CNN models are much more effective for detecting real-world audio deepfakes than traditional hand-crafted features. Future work can focus on combining models, improving noise robustness, testing newer generative speech models, and developing real-time detection systems. Overall, deep transfer learning proves to be a strong and reliable solution for tackling audio deepfake threats.

REFERENCES

[1] J. Kietzmann, L. W. Lee, I. P. McCarthy, and T. C. Kietzmann, "Deepfakes: Trick or treat?," Business Horizons, vol. 63, no. 2, pp. 135–146, 2020.
[2] B. Paris and J. Donovan, "Deepfakes and cheap fakes," Data & Society, p. 47, 2019.
[3] N. Eldien, R. Ali, and F. Moussa, "Real and fake face detection: A comprehensive evaluation of machine learning and deep learning techniques for improved performance," pp. 315–320, Jul. 2023.
[4] S. Ahmed, "Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size," Telematics and Informatics, vol. 57, p. 101508, 2021.
[5] A. Lieto, D. Moro, F. Devoti, C. Parera, V. Lipari, P. Bestagini, and S. Tubaro, "'Hello? Who am I talking to?' A shallow CNN approach for human vs. bot speech classification," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2577–2581.
[6] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A survey on deepfake video detection," IET Biometrics, vol. 10, no. 6, pp. 607–624, 2021.
[7] J. Truby and R. Brown, "Human digital thought clones: The holy grail of artificial intelligence for big data," Information & Communications Technology Law, vol. 30, no. 2, pp. 140–168, 2021.
[8] M. Waldrop, "Synthetic media: The real trouble with deepfakes," Knowable Magazine, vol. 3, 2020.
[9] R. Wijethunga, D. Matheesha, A. Al Noman, K. De Silva, M. Tissera, and L. Rupasinghe, "Deepfake audio detection: A deep learning-based solution for group conversations," in Proc. 2nd Int. Conf. Advancements in Computing (ICAC), 2020, pp. 192–197.
[10] A. Hamza, A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol, "Deepfake audio detection via MFCC features using machine learning," IEEE Access, vol. 10, pp. 134018–134028, 2022.
[11] V. Kumar, A. Kapoor, R. R. Chaudhary, L. Gupta, and D. Khokhar, "Preserving integrity: A binary classification approach to unmasking artificially generated voices in the age of deepfakes," in Proc. 11th Int. Conf. Computing for Sustainable Global Development (INDIACom), 2024, pp. 1449–1454.
[12] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, "A deep learning framework for audio deepfake detection," Arabian Journal for Science and Engineering, pp. 1–12, 2021.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY