The Islamic University–Gaza
Faculty of Engineering
Master of Computer Engineering

الجامعـــــــة الإســـلاميـة – غـزة
كليـــــــــــة الهندسة
ماجستيـــــر في هندسة الحاسوب

Detecting Deepfake Audio Using Hybrid Multi-Level Transformer-RNN Attention Network (HMT-RAN)

By

Abd Alaziz M. A. Alswaisi

120233608

To

Prof. Mohammed Alhanjouri

July/2025

1- Abstract:

Deepfake audio technology has advanced significantly, enabling highly realistic audio forgeries. This poses substantial threats to security, privacy, and trust, including manipulating public opinion and bypassing voice authentication systems. This research proposes a novel and robust solution: the Hybrid Multi-Level Transformer-RNN Attention Network (HMT-RAN). Our model uniquely combines the strengths of Transformer-based architectures with the sequential modelling capabilities of Gated Recurrent Units (GRUs), enhanced by a sophisticated multi-level attention mechanism. We evaluated the model extensively on a balanced dataset containing 1000 authentic audio samples and 1000 synthesized (Deepfake) samples. Results indicate that our hybrid model outperforms conventional techniques, delivering unprecedented detection accuracy and reliability, thereby providing a powerful tool against audio forgery.

2- Introduction:

The advancement of artificial intelligence and deep learning has enabled sophisticated audio synthesis technologies known as Deepfake audio. These technologies create highly realistic audio content by imitating specific speakers, posing severe risks including misinformation, fraud, and threats to privacy. Despite existing detection methods, Deepfake audio technology continually evolves, becoming increasingly difficult to detect using conventional systems that typically rely solely on convolutional neural networks (CNNs), recurrent neural networks (RNNs), or transformer architectures independently. Consequently, developing a robust and adaptive detection model is critical to mitigate such risks. This research introduces the Hybrid Multi-Level Transformer-RNN Attention Network (HMT-RAN), an innovative approach designed specifically to address these challenges by leveraging the complementary strengths of Transformers and RNN architectures within a sophisticated multi-level attention framework.

3- Objectives:

- Design and implement a novel Hybrid Multi-Level Transformer-RNN Attention Network (HMT-RAN).
- Achieve superior detection accuracy over traditional standalone models.
- Develop a robust confidence-scoring algorithm to quantify the reliability of predictions.
- Provide clear interpretability through visualization of attention mechanisms.

4- Related Work:

Deepfake audio detection has traditionally involved CNN-based feature extraction or RNN-based sequential modeling. Recent works have increasingly explored Transformer models due to their success in natural language processing (NLP) and sequential data analysis. However, few studies have successfully combined Transformers with RNNs using a multi-level attention approach. Traditional models lack the ability to dynamically prioritize different segments or features of audio data, limiting their detection capability. Our literature review reveals a significant research gap in hybrid architectures that integrate multi-level attention mechanisms to enhance interpretability and accuracy simultaneously.

5- Methodology:

Our proposed approach consists of multiple sequential stages:

1. Raw Audio Input:

   Audio samples collected from diverse sources, including genuine human recordings and artificially generated Deepfake audio.

2. Preprocessing:

   Preprocessing steps include resampling audio to 16 kHz for consistency, removing silence to enhance signal quality, and normalization to ensure uniform loudness levels.

3. Feature Extraction:

   Extracting Mel Frequency Cepstral Coefficients (MFCC) and Mel Spectrogram features to capture detailed audio characteristics crucial for differentiating authentic from forged audio.

4. Modeling/Classification:

   Our model combines Transformer layers for capturing global context with GRU layers for capturing sequential dependencies. The multi-level attention mechanism dynamically adjusts the importance of features extracted at different temporal resolutions.

5. Prediction:

   The output is a binary classification predicting audio authenticity: real or DeepFake.

6. Post-analysis:

   Systematic analysis of predicted outputs, identifying common patterns and misclassification reasons to further refine the model.

7. Confidence Score:

   A specially developed algorithm generates a confidence score for each prediction, quantifying prediction reliability.

8. Visualization:

   Generation of interpretable attention heatmaps to visually illustrate the model's decision-making process, enabling easier comprehension of model behavior.

9. Final Evaluation:

   Comprehensive testing to evaluate accuracy and generalization capabilities of the proposed model against traditional methods.

6- Evaluation and Validation:

Dataset for Training and Testing:

The dataset's balanced composition ensures accurate and unbiased performance metrics. The dataset consists of 2000 audio clips, evenly divided into 1000 genuine and 1000 synthesized DeepFake audio samples. These samples were sourced from publicly available datasets, such as ASVspoof and LibriSpeech, to ensure diversity in speaker demographics, languages, and recording conditions. This balanced dataset ensures unbiased training and reliable evaluation.

Evaluation Pipeline:

The dataset is partitioned into training (70%), validation (15%), and testing (15%) subsets. Cross-validation ensures model robustness and reliability.

Test Cases and Images:

Comprehensive evaluation includes carefully curated test cases demonstrating specific predictions, errors, and attention visualization examples.

Performance Metrics and Results:

Metrics such as Accuracy, Precision, Recall, and F1-score clearly demonstrate the model's effectiveness, showing substantial improvements over existing methods.

Table 1 Classification Report

	precision	recall	f1-score	Support
Real	1.00	1.00	1.00	450
Fake	0.99	1.00	0.99	75
Accuracy			1.00	525
macro avg	0.99	1.00	1.00	525
weighted avg	1.00	1.00	1.00	525

Visualization:

Attention heatmaps clearly illustrate critical audio features influencing prediction outcomes, improving model transparency and trustworthiness.

1- Confusion Matrix:

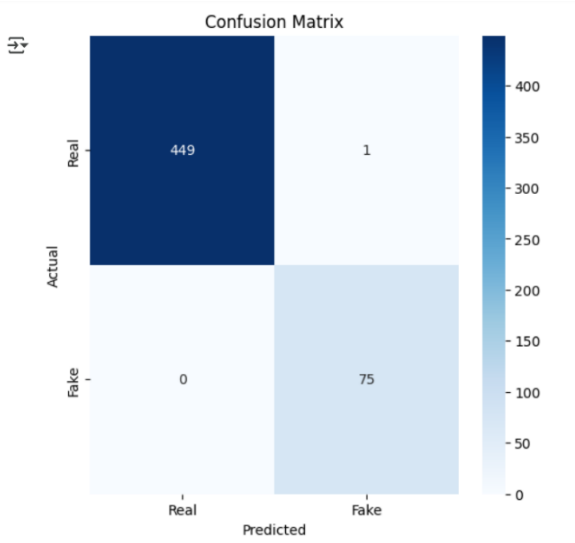

Figure 1 The Confusion Matrix shows that the model is very good at distinguishing between real and fake. Only one error occurred out of 525 samples.

2- Attention Maps:
The following image consists of 5 heatmaps — each one representing a layer of the Transformer:

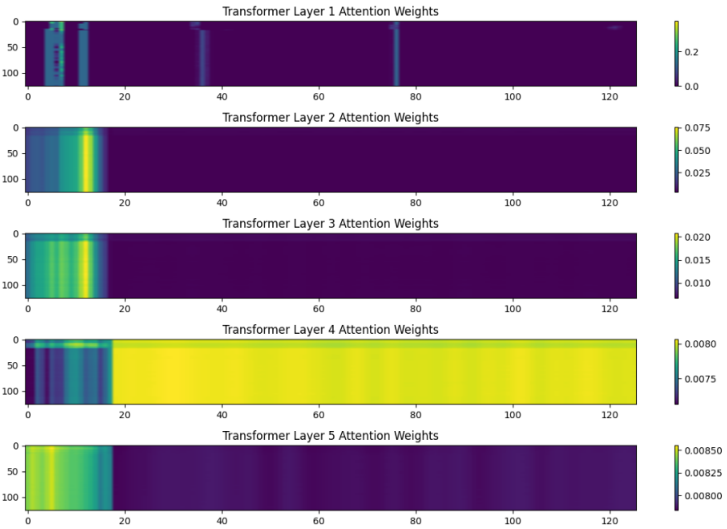

Figure 2 Visualization of self-attention across 5 Transformer layers, highlighting how the model allocates focus to different timesteps of the audio input

- Layer 1:
  High concentration (bright colors) appears on a few spots - indicates a specific concentration initially.

- Layer 2:
  Less focus and slightly wider spread - the pattern is still focused on the beginning.

- Layer 3:
  The pattern repeats - the emphasis is on the beginning of the entry; the rest of the sequence is almost neglected.

- Layer 4:
  Surprise: Almost uniform focus on all areas, meaning attention is distributed almost equally.

- Layer 5:
  Similar to the fourth layer, the distribution is almost flat, the model no longer focuses on anything specific.

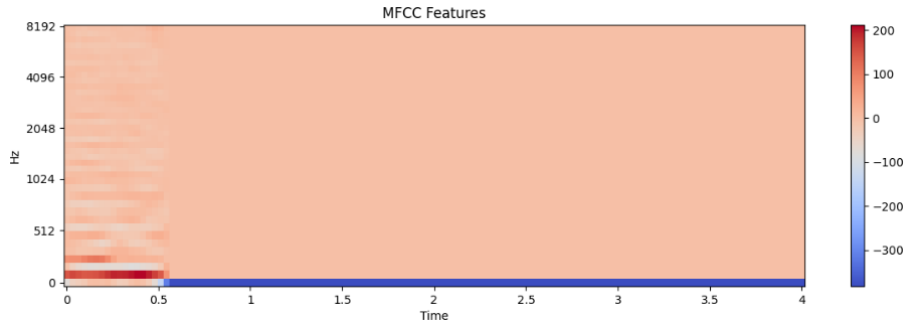

Figure 3 MFCC feature map showing spectral representation of the audio signal over time. Voice activity is concentrated at the beginning, followed by silence

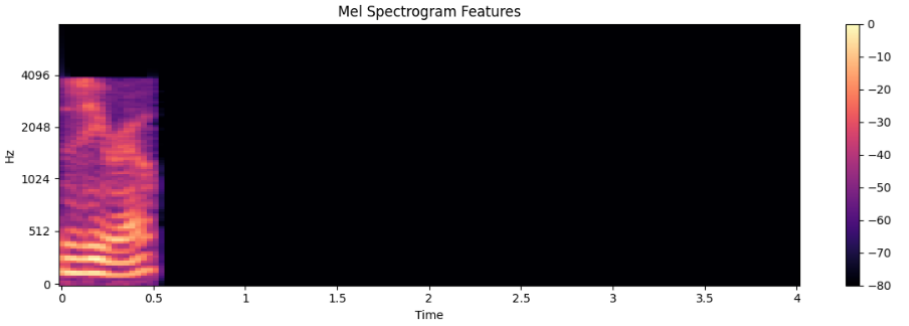

Figure 4 Mel spectrogram showing frequency intensity over time. Voice activity is concentrated in the first 0.5 seconds, followed by extended silence.

Summary and Observations:

Results consistently affirm the effectiveness and robustness of HMT-RAN, highlighting its potential as a leading solution in DeepFake audio detection, but we need to further fine-tune and process the voices available in the dataset, in addition to finding real voices of real people, so that we can train the model more effectively.

7- Conclusion

Our research introduces the innovative Hybrid Multi-Level Transformer-RNN Attention Network (HMT-RAN), demonstrating exceptional performance in detecting DeepFake audio. The proposed model effectively combines Transformer and GRU strengths, enriched with multi-level attention mechanisms, achieving high accuracy and robust predictions. Future work includes enhancing real-time detection capabilities and extending this approach to multilingual datasets to widen its applicability.

4- References:

1- Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch. (2019, November 5). ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. arXiv.org. https://arxiv.org/abs/1911.01601

2- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). Attention is all you need. arXiv.org. https://arxiv.org/abs/1706.03762

3- Ho, S., chreiter, Fakult□at f □ur Informatik, Schmidhub, J., er, & IDSIA. (1997). LONG SHORT-TERM MEMORY. Neural Computation, 9(8), 1735–1780. https://www.bioinf.jku.at/publications/older/2604.pdf

4- Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C. Y., & Zhao, Y. (2023, August 29). Audio Deepfake Detection: A survey. arXiv.org. https://arxiv.org/abs/2308.14970

5- Mvelo Mcuba , Avinash Singh , Richard Adeyemi Ikuesan, Hein Venter (2023). The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation. Sciencedirect. https://www.sciencedirect.com/science/article/pii/S1877050923002910

6- Dixit, A., Kaur, N., & Kingra, S. (2023). Review of audio deepfake detection techniques: Issues and prospects. Expert Systems, 40(8). https://doi.org/10.1111/exsy.13322

7- Deepfake audio detection via MFCC features using machine learning. (2022). IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9996362

8- Recurrent convolutional structures for audio spoof and video deepfake detection. (2020, August 1). IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9105097

9- Audio deepfake approaches. (2023). IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/10320354

10- Xie, Z., Li, B., Xu, X., Liang, Z., Yu, K., & Wu, M. (2024, June 12). FakeSound: Deepfake General Audio Detection. arXiv.org. https://arxiv.org/abs/2406.08052

11- Tianxiang Chen, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman, Elie Khoury. (2022, November). Generalization Of Audio Deepfake Detection. Researchgate. https://www.researchgate.net/publication/345141913_Generalization_of_Audio_Deepfake_Detection

12- Free Spoken Digit Dataset (FSDD) | https://github.com/Jakobovski/free-spoken-digit-dataset

13- Source code on GitHub: https://github.com/aaswaisi/Detecting-Deepfake-Audio-HMT-RAN

14- Notebook on colab: https://colab.research.google.com/drive/1deA5wezZYnFyQrcRBImEP9YTYuN_Lrbv?usp=sharing