RESEARCH ARTICLE

# AUDD: Audio Deepfake Detection Using Paralinguistic Feature Extraction Techniques

**Zahoor Ahmed**[1], **Gul Sher Ali Khan**[1] **and Raja Vavekanand**[2,*]

[1] Balochistan University of Information Technology, Engineering and Management Sciences, Baleli, Quetta 87300, Pakistan
[2] Benazir Bhutto Shaheed University Lyari, Karachi 75660, Sindh, Pakistan

## Abstract

**This work investigates the effectiveness of incorporating paralinguistic feature extraction in audio deepfake detection models. The proposed model extracts paralinguistic features from audio clips and represents them as 1024-dimensional vector embeddings. These embeddings are then used as input for a logistic regression model, which performs binary classification to distinguish between real and deepfake audio samples. The ASVspoof2019 dataset, comprising both genuine and spoofed audio clips, is used to evaluate the model's performance. The results are assessed using evaluation metrics such as Equal Error Rate (EER) and accuracy, which provide insight into the model's effectiveness compared to state-of-the-art methods. The proposed model achieves an EER of 3.04% and an accuracy of 97.9%, indicating that paralinguistic feature extraction is a promising approach for audio deepfake detection. These results suggest that incorporating paralinguistic features can improve the performance of audio deepfake detection systems, making it a valuable tool for future research in this area. Overall, the study demonstrates the potential of paralinguistic feature extraction in enhancing the accuracy and reliability of audio deepfake detection methods.**

**Keywords**: audio, deepfake, paralinguistic, deep learning.

## 1 Introduction

The rapid advancement of artificial intelligence (AI) has brought numerous benefits, including the revolutionization of environmental monitoring and enhanced data protection in cybersecurity [3, 4]. However, the increasing availability of AI-driven technologies also raises concerns about their potential misuse, particularly in the fields of privacy, security, ethics, and trust [5–7]. The growing problem of deepfakes, which can be created using deep learning techniques and publicly available data, including voice samples, has made it essential to develop effective detection methods [8, 9].

Most state-of-the-art audio deepfake detection methods focus on extracting verbal aspects of speech, such as words and syntax, using techniques like automatic speech recognition (ASR) [10–15]. However, these methods may overlook subtle hints that distinguish human voice from synthesized audio, such as intonation, emotional tone, rhythm,

**Submitted:** 30 November 2024
**Accepted:** 27 June 2025
**Published:** 03 August 2025

**Vol.** 1, **No.** 1, 2025.
10.62762/JCI.2024.667518

***Corresponding author:**
✉ Raja Vavekanand
zahoorahmed54217@gmail.com

**Citation**
Ahmed, Z., Khan, G. S. A., & Vavekanand, R. (2025). AUDD: Audio Deepfake Detection Using Paralinguistic Feature Extraction Techniques. *Journal of Computing Intelligence*, 1(1), 3–8.

and stress. These non-verbal cues, referred to as paralinguistic features, can capture inconsistencies in AI-generated speech that are missed by current methods [14, 15]. Paralinguistic features, including stress, intonation, rhythm, and emotional tone, have been shown to capture inconsistencies in AI-generated speech that are often missed by traditional methods focusing on verbal content [21]. Such features play a significant role in distinguishing human speech from synthetic audio, making them valuable for deepfake detection [16].

This thesis explores the effectiveness of paralinguistic feature extraction in audio deepfake detection methods. By utilizing the TRILLsson method, developed by Google Research [2], paralinguistic features are extracted from audio samples and represented as 1024-dimensional embedding vectors. These vectors serve as input for a logistic regression model, which performs binary classification to distinguish between real and deepfake audio samples. The ASVspoof2019 dataset, containing over 125,000 audio samples, is used to train and evaluate the model [15]. The performance of this method is compared to six other recent audio deepfake detection models that have used the same dataset [10–12, 17–19]. The equal error rate (EER) is used as the primary evaluation metric to assess the effectiveness of the proposed method [19].

## 2 Related Work

The rapid evolution of deepfake technologies has drawn significant attention in recent years. Deepfakes, created using advanced deep learning methods, can convincingly impersonate individuals if the necessary data is available [9]. The prevalent technique used to create deepfakes is Generative Adversarial Networks (GANs), which has made it possible to easily manipulate multimedia content [9]. Recent advancements in deepfake detection draw on techniques like data augmentation, seen in ultrasound imaging for medical diagnostics, molecular structure prediction via NMR spectra, and load-aware routing for UAV networks. Additionally, data augmentation in text classification for languages like Sindhi and multimodal image classification in ECG diagnosis further demonstrate these approaches' effectiveness.

### 2.1 State-of-the-Art Approaches in Audio Deepfake Detection

Several approaches have been proposed for audio deepfake detection in recent years. These approaches have utilized the ASVspoof2019 dataset, which contains real and spoofed audio clips from 107 different speakers [1]. Synthetic Speech Attribution Transformer (SSAT), which achieved an accuracy of 90.2% on the ASVspoof2019 dataset [11]. Online Hard Example Mining (OHEM) algorithm, which achieved an EER of 0.77% on the ASVspoof2019 (LA) dataset [12]. A fully automated end-to-end deepfake audio detection method, which achieved an EER of 1.08% on the ASVspoof2019 dataset [17]. RawNet, a deep neural network architecture that operates directly on raw audio signals, which achieved a 66.09% decrease in EER compared to RawNet [10]. AASIST2 architecture, which improved performance in short utterance evaluation in speech anti-spoofing [13]. A method that incorporates high-level features from Speech Emotion Recognition (SER) systems into audio deepfake detection methods, which achieved an accuracy of 94% [18].

### 2.2 Paralinguistic Feature Extraction for Detection Methods

Paralinguistic features, such as stress, intonation, emotional tone, and rhythm, have been demonstrated to be crucial in distinguishing real and synthetic speech, enhancing deepfake detection performance [13]. Research by Conti et al. [18] also supports that emotion recognition, a key paralinguistic feature, improves deepfake detection in speech.

### 2.3 Implementation of Techniques for Audio Deepfake Detection

Logistic regression can be used to effectively handle the paralinguistic embedding vectors generated by TRILLsson. Logistic regression is a suitable approach for this task because it is simple, effective, and interpretable. The sigmoid function is used to calculate the probability that a given input belongs to a particular class [20, 22]. The Equal Error Rate (EER) is a commonly used metric to evaluate the performance of deepfake detection methods [19]. In addition to EER, standard performance metrics such as accuracy, precision, recall, and F1-score can also be used to evaluate the effectiveness of the model [23].

## 3 Methodology

This thesis employs a logistic regression model for binary classification of genuine and spoofed audio samples. The ASVspoof2019 Logical Access database is used, comprising 80,000 audio samples. TRILLsson, a TensorFlow2 framework, extracts 1024-dimensional paralinguistic feature embeddings from the audio clips. The dataset is preprocessed by pairing embeddings with labels and divided into training (70%), validation (20%), and testing (10%) sets. The model is trained and evaluated using accuracy, precision, recall, F1-score, Equal Error Rate (EER), ROC curve, and confusion matrix. Python libraries, including Librosa, Pandas, and Sklearn, are utilized for data processing and analysis. In this study, logistic regression was selected for its simplicity, interpretability, and its suitability for a relatively straightforward binary classification task. The paralinguistic features extracted from the audio samples are sufficiently structured, making logistic regression an effective baseline model. While more sophisticated models like Support Vector Machines (SVMs) or Convolutional Neural Networks (CNNs) could potentially offer higher performance, logistic regression was chosen due to its computational efficiency and ease of implementation.

However, it is important to note that SVMs with appropriate kernel functions can capture non-linear decision boundaries and might be particularly useful when feature embeddings exhibit complex relationships. Similarly, neural networks, such as CNNs or recurrent models (LSTMs), can better capture hierarchical patterns in the data, which might be advantageous for deepfake detection in more challenging datasets.

## 4 Experimental Setup

This thesis utilizes the ASVspoof2019 Logical Access (LA) database as the sole dataset, comprising 80,000 audio samples in flac format [1]. The dataset is designed for spoofing detection architectures and provides labels indicating genuine or spoofed samples. The audio samples are processed using TRILLsson paralinguistic feature extraction, resulting in 1024-dimensional feature embedding vectors. These vectors serve as input for the logistic regression model, used for the binary classification task.

### 4.1 Dataset

Each audio clip within the dataset is standardized to a length of one second at a 16.0 kHz sample rate, aligning with state-of-the-art methods [10–13, 17, 18]. This standardization ensures consistent input data, crucial for effective feature extraction and classification.

### 4.2 TRILLsson

TRILLsson, a TensorFlow2 framework, is employed to extract paralinguistic features from the audio clips [2]. The model extracts 1024-dimensional feature embeddings containing paralinguistic information about individual audio samples.

### 4.3 Preprocessing

The preprocessing stage involves pairing TRILLsson embeddings with their labels for supervised learning. Label-embedding vector pairs are placed in a data frame to form a coherent input item. Due to dataset size and pairing issues, the input size is 80,000 embedding vectors, with pre-defined train, validation, and test sets merged into one set. The dataset is divided into 70% for training, 20% for validation, and 10% for testing.

### 4.4 Logistic Regression

A logistic regression model is selected for its efficiency in binary classification tasks. The model architecture is built to suit the input data, optimizing performance, simplicity, and interpretability. The input data is scaled using 'StandardScaler' to improve algorithm performance [24].

### 4.5 Evaluation

The performance of the logistic regression model is evaluated using accuracy, precision, recall, F1-score, Equal Error Rate (EER), ROC curve, and confusion matrix. The EER is a common evaluation metric in deepfake detection studies, reflecting the balance between false acceptance and rejection rates [19]. Accuracy is measured separately for validation and test sets to assess performance and potential overfitting. Precision, recall, and F1-score are calculated for both sets to illustrate overall performance and provide insights into prediction balance.

## 5 Results

The logistic regression model's performance on the ASVspoof2019 dataset is presented in this section. The model achieved high accuracy rates, with 97.6% on the validation set and 97.9% on the test set (Table 1). The precision, recall, and F1-score also demonstrated strong performance, with scores of 88.6%, 88.6%, and 88.6% on the validation set, and 90.5%, 89.4%, and 89.9% on the test set, respectively (Figure 3).

The Equal Error Rate (EER) of 3.04% indicates that the model is reliable and effective in detecting deepfakes. The Receiver Operating Characteristics

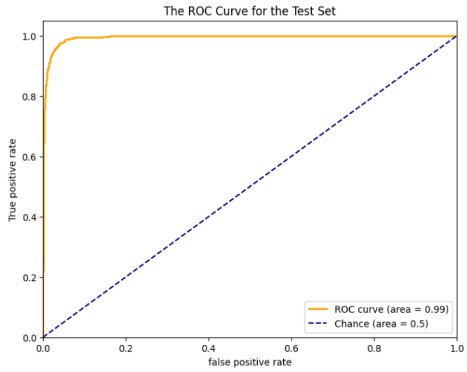

**Figure 1.** ROC curve.

**Table 1.** Performance metrics.

| Metric | Validation Set | Test Set |
|---|---|---|
| **Accuracy** | 97.6% | 97.9% |
| **Precision** | 88.6% | 90.5% |
| **Recall** | 88.6% | 89.4% |
| **F1-Score** | 88.6% | 89.9% |
| **EER** | 3.04% | - |

(ROC) curve and confusion matrix further illustrate the model's performance (Figure 1), with an area under the ROC curve (AUC) of 0.99 and correct identification of 756 true positives and 7104 true negatives. Figure 1 provides a visual representation of the model's trade-off between true positive rate (recall) and false positive rate (FPR). AUC is a key metric derived from the ROC curve, which indicates the model's ability to discriminate between real and deepfake audio. In our case, the AUC value of 0.99 demonstrates that the model performs exceptionally well, as values closer to 1 indicate a strong ability to distinguish between the two classes. A higher AUC value reflects that the model is more confident and accurate in classifying both real and deepfake samples, without favoring one class over the other.

The proposed model's performance is compared to state-of-the-art models, including SSAT, OHEM, Light-DARTS, TO-RawNet, AASIST2, and the SER network. While some of these models achieved lower EERs, the proposed model's performance is notable, especially considering the modified dataset size and division. Figure 2 presents the confusion matrix, which further clarifies how the model's predictions are distributed between the four possible outcomes: true positives (real audio correctly classified as real), false positives (deepfake classified as real), false negatives (real audio classified as deepfake), and true negatives (deepfake correctly classified as deepfake). From the confusion matrix, we observe that the model correctly identified 756 true positives and 7104 true negatives, resulting in relatively few misclassifications.

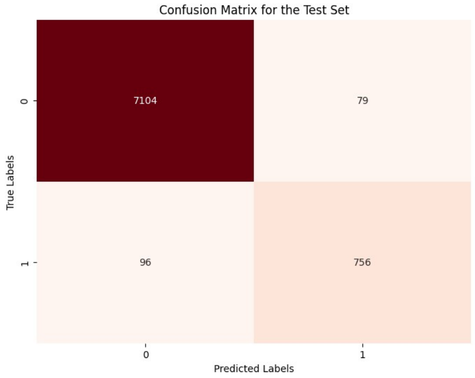

**Figure 2.** Confusion matrix.

The logistic regression model's performance on the ASVspoof2019 dataset demonstrates strong and promising results, with high accuracy, precision, and recall rates, as well as a low EER. The model's performance is comparable to state-of-the-art models, suggesting that paralinguistic feature extraction in combination with a logistic regression model is an effective and reliable method for deepfake detection (Table 2).

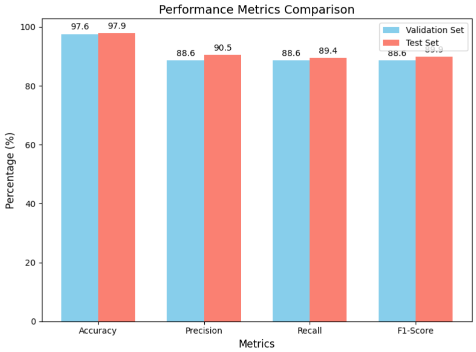

**Figure 3.** Model performance metrics graph.

We compare our logistic regression-based deepfake detection model with SSAT, AASIST2, and OHEM. SSAT uses a transformer model on spectrogram features and achieved 90.2% accuracy on ASVspoof2019. Our model, however, reached 97.9% accuracy, demonstrating better performance with simpler features and lower computational cost. AASIST2 works well for short utterances but has a higher EER of 8.36% compared to our model's 3.04%. Our model provides a more balanced performance across various audio types. OHEM focuses on hard-to-classify samples and achieves a low EER of 0.77%, but at a higher computational cost. Our model achieves a comparable EER of 3.04% with faster

**Table 2.** State-of-the-Art model comparison.

| Model | EER | Accuracy |
|---|---|---|
| **SSAT** | - | 90.2% |
| **OHEM** | 0.77% | - |
| **Light-DARTS** | 1.08% | - |
| **TO-RawNet** | 1.58% | - |
| **AASIST2** | 8.36% | - |
| **SER network** | - | 94% |
| **Proposed Model** | 3.04% | 97.9% |

training and lower complexity. Overall, our model offers a more efficient and practical solution with higher accuracy and lower computational demands.

## 6 Discussion

This work explores the effectiveness of paralinguistic feature extraction in audio deepfake detection using the ASVspoof2019 dataset. Despite minor issues with data imbalance, the logistic regression model achieved impressive results, with an accuracy of 97.9% and an Equal Error Rate (EER) of 3.04%. This suggests that paralinguistic feature extraction is a promising approach for deepfake detection. Despite minor issues with data imbalance, the logistic regression model achieved impressive results, with an accuracy of 97.9% and an Equal Error Rate (EER) of 3.04%. This validates the hypothesis that paralinguistic feature extraction can enhance deepfake detection. Empirical studies have shown that these features are often overlooked by traditional speech recognition methods but significantly contribute to identifying subtle discrepancies in AI-generated audio [20].

However, limitations of the study include the use of a 2019 dataset, which may not be representative of current deepfake generation techniques. Future research should incorporate newer datasets and consider multiple dataset options to improve generalizability. Additionally, exploring other machine and deep learning models, such as linear SVMs and Convolutional Neural Networks, may lead to further improvements.

## 7 Conclusion

This paper demonstrates the potential of paralinguistic feature extraction in audio deepfake detection. The use of TRILLsson-generated embeddings and logistic regression achieved impressive results, highlighting the effectiveness of this approach. Future research should build upon this study by incorporating diverse datasets, exploring alternative models, and examining combinations of approaches. By advancing deepfake detection methods, we can stay ahead of the rapidly evolving deepfake generation techniques and mitigate potential threats.

## Data Availability Statement

Data will be made available on request.

## Funding

This work was supported without any funding.

## Conflicts of Interest

The authors declare no conflicts of interest.

## Ethical Approval and Consent to Participate

Not applicable.

## References

[1] Yamagishi, J., Todisco, M., Sahidullah, M., Delgado, H., Wang, X., Evans, N., ... & Nautsch, A. (2019). Asvspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *ASV Spoof, 13.*

[2] Shor, J., & Venugopalan, S. (2022). TRILLsson: Distilled Universal Paralinguistic Speech Representations. *Interspeech 2022*. [CrossRef]

[3] Kaur, R., Gabrijelcic, D., & Klobucar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion, 97*, 101804. [CrossRef]

[4] Chisom, O. N., Biu, P. W., Umoh, A. A., Obaedo, B. O., Adegbite, A. O., & Abatan, A. (2024). Reviewing the role of AI in environmental monitoring and conservation: A data-driven revolution for our planet. *World Journal of Advanced Research and Reviews, 21*(1), 161-171. [CrossRef]

[5] Oladoyinbo, T. O., Olabanji, S. O., Olaniyi, O. O., Adebiyi, O. O., Okunleye, O. J., & Alao, A. I. (2024). Exploring the challenges of artificial intelligence in data integrity and its influence on social dynamics. *Asian Journal of Advanced Research and Reports, 18*(2), 1-23. [CrossRef]

[6] Sontan, A. D., & Samuel, S. V. (2024). The intersection of artificial intelligence and cybersecurity: Challenges and opportunities. *World Journal of Advanced Research and Reviews, 21*(2), 1720-1736. [CrossRef]

[7] Familoni, B. T. (2024). Cybersecurity challenges in the age of AI: Theoretical approaches and practical solutions. *Computer Science & IT Research Journal, 5*(3), 703-724. [CrossRef]

[8] Khan, A., & Malik, K. M. (2023). Securing voice biometrics: One-shot learning approach for audio

deepfake detection. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1-6). IEEE. [CrossRef]

[9] Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence, 53*(4), 3974-4026. [CrossRef]

[10] Wang, C., Yi, J., Tao, J., Zhang, C., Zhang, S., Fu, R., & Chen, X. (2023).TO-Rawnet: Improving RawNet with TCN and Orthogonal Regularization for Fake Audio Detection. *INTERSPEECH 2023*, 3137-3141. [CrossRef]

[11] Yadav, A. K. S., Bartusiak, E. R., Bhagtani, K., & Delp, E. J. (2023). Synthetic speech attribution using self supervised audio spectrogram transformer. *Electronic Imaging, 35*(4), 372-1-372-11. [CrossRef]

[12] Hu, C., & Zhou, R. (2022). Synthetic voice spoofing detection based on online hard example mining. *arXiv preprint arXiv:2209.11585*.

[13] Zhang, Y., Lu, J., Shang, Z., Wang, W., & Zhang, P. (2024). Improving short utterance anti-spoofing with AASIST2. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 11636-11640). IEEE. [CrossRef]

[14] Pastor, E., Koudounas, A., Attanasio, G., Hovy, D., & Baralis, E. (2023). Explaining Speech Classification Models via Word-Level Audio Segments and Paralinguistic Features. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics* (Volume 1: Long Papers), 2221-2238. [CrossRef]

[15] Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., ... & Lee, K. A. (2020). ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language, 64*, 101114. [CrossRef]

[16] Liu, T., & Yuan, X. (2023). Paralinguistic and spectral feature extraction for speech emotion classification using machine learning techniques. *EURASIP Journal on Audio, Speech, and Music Processing, 2023*(1), 23. [CrossRef]

[17] Wang, C., Yi, J., Tao, J., Sun, H., Chen, X., Tian, Z., ... & Fu, R. (2022). Fully automated end-to-end fake audio detection. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia* (pp. 27-33). [CrossRef]

[18] Conti, E., Salvi, D., Borrelli, C., Hosler, B., Bestagini, P., Antonacci, F., ... & Tubaro, S. (2022). Deepfake speech detection through emotion recognition: A semantic approach. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8962-8966). IEEE. [CrossRef]

[19] Zhang, L., Wang, X., Cooper, E., Evans, N., & Yamagishi, J. (2023). Range-based equal error rate for spoof localization. *Interspeech 2023*, 3212-3216. [CrossRef]

[20] Saha, S., Sahidullah, M., & Das, S. (2024). Exploring green AI for audio deepfake detection.In *2024 32nd European Signal Processing Conference (EUSIPCO)*, 186-190. [CrossRef]

[21] Crystal, D., & Quirk, R. (2021). Systems of prosodic and paralinguistic features in English. *Walter de Gruyter GmbH & Co KG*. [CrossRef]

[22] Bhavitha, B., Rodrigues, A. P., & Chiplunkar, N. N. (2017). Comparative study of machine learning techniques in sentimental analysis. In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 216-221). IEEE. [CrossRef]

[23] Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access, 10*, 25494-25513. [CrossRef]

[24] Ahsan, M. M., Mahmud, M. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies, 9*(3), 52. [CrossRef]

**Zahoor Ahmed** graduated with as gold medalist in Bachelor's degree in Computer Engineering from Balochistan University of Information Technology Engineering & Management Sciences, Pakistan in 2024. He completed various projects focusing on IoT, deep learning, and web technologies. His research interests encompass Internet of Things, machine learning, and technology, big data Analysis. (Email: engr.zahoorahmed54217@gmail.com)

**Gul Sher Ali Khan** received a Bachelor's degree in Computer Engineering from Balochistan University of Information Technology and Management Sciences (BUITEMS), Quetta, Pakistan, in 2024. He has completed various academic and industry-oriented projects in machine learning, computer vision, and IoT, including a smart thermoelectric air conditioner and Safe Drive Insights for road safety analytics. His research interests include artificial intelligence, natural language processing, Machine Learning and Deep Learning. (Email: engr.gulsheralikhan@gmail.com)

**Raja Vavekanand** received a Bachelor's degree in Information Technology from Benazir Bhutto Shaheed University, Karachi, Pakistan in 2024. He has completed different research projects based on IoT, deep learning, and image processing. His research interests include generative AI, machine learning, medical imaging and computer vision. (Email: bharwanivk@outlook.com)