AudioIntegrityNet: A Convolutional Neural Network Framework for Audio Deepfake Detection Using RMS-ZCR Threshold Intersections

Prasanth K V
Dept of Statistics
Cochin University of Science & Technology,
Kochi, India
2000prasanth@protonmail.com

Dr KC James
Professor, Dept of Statistics,
Cochin University of Science & Technology,
Kochi, India

***Abstract*— The rise of spoofed audio poses significant risks, including the spread of misinformation, fake news, and substantial financial fraud. Recent developments in audio generation, such as GANs, diffusion models, and autoencoders, are rapidly evolving, and only three seconds of a person's audio are required to clone their entire sound and generate new audio that feels like it is being said by them. While recent studies have predominantly focused on spectral features and their derivatives for detecting spoofed audio, this research explores the role of Root Mean Square (RMS) and Zero-Crossing Rate (ZCR) values that cross a specific threshold, interpreted as "breath," as potent discriminators. We use statistical methods to explore how the feature differ in real and fake audio data. Additionally, we develop AudioIntegrityNet, a Convolutional Neural Network (CNN) to classify audio as real or fake.**

Keywords—Deepfakes, RMS, ZCR, Mann Whitney U Test, MFCC, CNN, EER

I. INTRODUCTION

In recent years, voice synthesis and voice-to-voice cloning technologies have become more prominent. They find many creative applications, such as generating podcasts and audiobooks with a particular author's voice for a given script, as well as text-to-speech synthesis systems. Like any technology, the rampant misuse of these models has raised significant concerns. These models have been misused to spread misinformation and degrade trust in public authorities. According to World Economic Forum report, 2024 [2] , during the recently held 2024 Indian General Elections, many social media platforms like X were flooded with deepfakes of politicians making controversial statements.

A report by Signicat, 2024 [3] indicates that deepfakes now represent 6.5% of total fraud attempts, marking a 2137% increase over the past three years. Another threat is the increased chances of spoofing attacks on voice-based authentication systems.

The detection of audio deepfakes is critical in today's digital landscape, where advanced voice synthesis technologies pose significant risks to privacy, public trust, and societal stability. As these deepfakes become more sophisticated, they can be exploited to spread misinformation, manipulate public opinion, and commit fraud, particularly in sensitive areas such as politics and finance. The ability to convincingly mimic an individual's voice raises concerns about identity theft and the creation of fabricated audio recordings that can harm reputations. As deepfake technology evolves, distinguishing between genuine and manipulated audio becomes increasingly complex, highlighting the need for robust detection methods. Effective audio deepfake detection is essential for protecting individuals and organizations and for maintaining the integrity of information in an era of declining trust in media and communication

While traditional approaches to audio deepfake detection have relied heavily on spectral features since these are the same features used by Generative Adversarial Networks (GANs) and other synthesizers to generate audio in the first place the adversarial game raises concerns about their ability to detect future deepfakes and recently there is a growing interest in understanding the unique attributes of genuine human voices. In this work, we aim to explore how breath as a feature fare with respect to different types of Fake audio generated from Text to Speech(TTS) synthesizers, voice-to-voice conversion models, and Generative Adversarial Networks.This research endeavors to develop a robust classifier incorporating breath as a feature that effectively distinguishes between genuine and manipulated audio, thereby contributing to the field of audio forensics and addressing the urgent need for reliable detection mechanisms in the face of evolving audio synthesis technologies.

II. LITERATURE REVIEW

Voice synthesis technology has evolved significantly since its inception in the 1970s. Early models relied on a dictionary of spoken words, which were concatenated to create audio streams for applications such as railway station announcements. The development of concatenative systems marked a pivotal advancement, as these systems stitched together small, pre-recorded phonemes to form coherent sentences. By the mid-2000s, the introduction of Hidden Markov Models (HMMs) [8] further improved the naturalness of synthesized speech,

although these models still produced audio that could be distinguished from authentic human voices.

The landscape of voice synthesis underwent a transformative shift with the advent of neural network-based approaches, particularly those utilizing Recurrent Neural Networks (RNNs). This was followed by the emergence of Generative Adversarial Networks (GANs) [4] which introduced an adversarial learning paradigm. In this framework, a generator and discriminator are trained in opposition to one another, resulting in the generation of human-like voice. GANs synthesize audio by producing spectrograms from continuous Short-Time Fourier Transform (STFT) values, thereby enhancing the realism of the output.

Autoencoders [6] also play a significant role in voice synthesis by compressing data from a higher-dimensional space to a lower-dimensional representation and subsequently reconstructing it back to a higher dimension. Additionally, diffusion models [7] have gained traction, operating by gradually adding noise to an audio signal and then regenerating the original audio file through a process of reverse diffusion.

As voice synthesis technologies have advanced, distinguishing between real and spoofed audio has become increasingly challenging. Early approaches to audio deepfake detection used hand crafted features short-term spectral features, long term spectral features and prosodic features. Classifiers such as Gaussian Mixture models and Neural Networks were mostly employed [1].

[12] propose a system known as SpecRNet, inspired by RawNet2, which utilizes LFCC features and achieves an average Equal Error Rate (EER) of 0.1549 on the WaveFake dataset. Meanwhile [13] address the predominance of single-channel audio in existing research, exploring the implications of mono-to-stereo (dual-channel) conversion for audio deepfake detection. They introduce a M2S converter that employs neural time warping and Temporal CovNet to classify audio as real or fake by analyzing signals split into right and left branches.

The work of [17] takes a novel approach to deepfake detection by investigating the fundamental characteristics that define human authenticity in video. Their research focuses on biological feature extraction, specifically examining changes in skin color due to blood flow through arteries and veins.

Additionally, [15] developed a Breath-Silence-Talking Encoder (BTSE) Network that utilizes neural networks to detect and classify three distinct activities within an audio stream. Their Raw2Net model, which employs the [BTSE] encoder, achieves an EER of 9.79 on the ASVspoof 2021 dataset. Building on this approach, we analyze existing literature on the use of breath patterns as a feature for audio deepfake detection. The paper "Every Breath You Do Not Take" [14] presents a method for extracting breath patterns by examining the intersection of ZCR (Zero Crossing Rate) and RMS (Root Mean Square) energy values. The authors assert a significant correlation between the timestamps where RMS and ZCR values intersect and the presence of breath. The challenge of generalization and evolving deepfake techniques necessitates exploring diverse features beyond traditional spectral analysis. Multiple sources emphasize the poor generalization ability of current audio deepfake detection systems when faced with out-of-domain data, unseen attacks, or different acoustic conditions [20].

As deepfake generation technologies rapidly advance, relying solely on low-level spectral imperfections might become increasingly ineffective.

While "Every Breath You Don't Take" hypothesizes and shows that current deepfake generation techniques often fail to incorporate breaths adequately, there is a lack of detailed analysis in these papers on how Text-to-Speech (TTS) and Voice Conversion (VC) models specifically affect the RMS and ZCR patterns that are characteristic of natural human breaths.

There is a clear gap in research that compares the effectiveness and robustness of features derived from RMS-ZCR intersections for breath detection against these more traditional or deep learning-based features in the specific task of audio deepfake detection.

RMS and ZCR capture essential temporal and energy-related aspects of speech that might reveal subtle inconsistencies in deepfakes. RMS reflects the signal's power or amplitude over time, while ZCR provides information about its frequency content and rapid changes. The dynamic interplay between these fundamental properties, particularly their threshold crossings and intersections, could expose artifacts or unnatural patterns introduced by deepfake generation processes that are not adequately captured by spectral features alone. Hence, studying these features becomes essential, as it allows us to understand the nuances and variations in synthetic audio generated by different techniques, such as Generative Adversarial Networks (GANs), autoencoders, and Text-to-Speech (TTS) systems. By examining how breath features manifest across these diverse generation methods, we can identify specific characteristics that differentiate genuine audio from manipulated content. This extensive analysis can reveal critical insights into the acoustic signatures associated with various generation techniques, paving the way for the integration of breath patterns as a feature in audio classification models. Such advancements hold the potential to significantly enhance the effectiveness of deepfake audio detection systems, contributing to the development of more robust and reliable tools for identifying manipulated audio content

II. METHODOLOGY

In this study, we seek to investigate the statistical properties of breath features in authentic and

manipulated audio recordings across diverse datasets, with a primary focus on exploring their potential utility in developing a robust classifier for detecting audio deepfakes. By analyzing the acoustic characteristics of breathing patterns in real and fake audio samples, we aim to identify distinct statistical patterns and anomalies that can be leveraged to distinguish between genuine and synthesized speech.

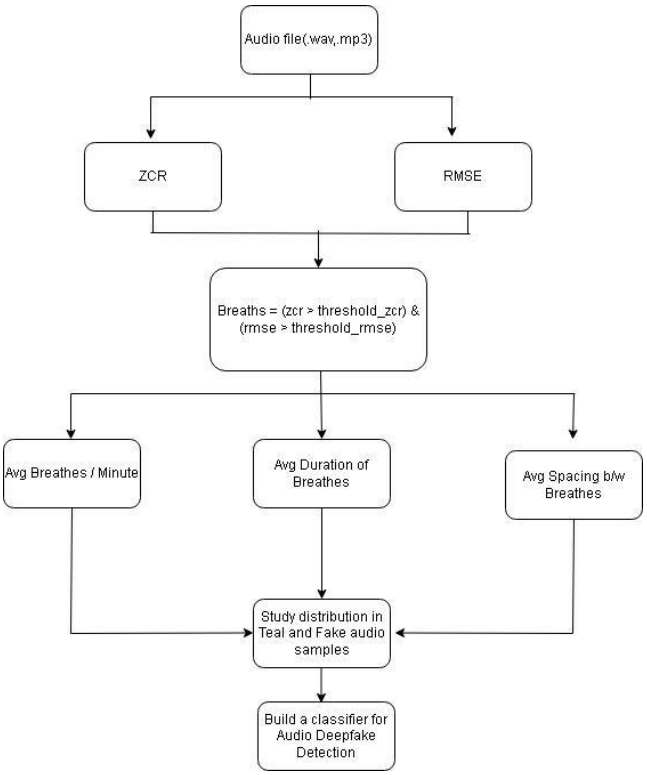

Fig 1: Methodology Diagram

3.1) Feature extraction & Statistical Analysis of Breath features

Here we aim to perform statistical study of how does breath as a feature is distributed in Fake and Real audio in these 10 datasets

Table 1 : Datasets used in study and their characteristics

Dataset	Characteristics
FA-1, FA-2, FA-3	Consists of 3.5 hours of spoken data collected from Frontier AI labs. With length of each audio file ranging from 15 to 20 seconds.
ASV Spoof 2021	Consists of data samples from ASV spoof LA evaluation set primarily target to deal with spoofing attacks. Predominantly length is less than 3 seconds
In the wild	Is a dataset capturing audio deepfakes from real-world scenarios with diverse and uncontrolled environments. Audio files have an average length of 7 seconds
Every Breath you do not take	Consists of real audio samples of Journalist reading articles and generated audio where a TTS system is used to generate the same spoken content.
Half Truth	Is a multilingual in bot English and Chinese dataset aimed to study characteristics of partially fake audio
Fake or Real	Consists 2 second clips of utterances of real humans and audio generated from TTS systems .
Deepvoice	Data consists of Fake audio samples that has been using generated using Retrieval based voice conversion process.
Wavefake	Combines generated audio data from multiple deepfake models like Melgan, Wavegan etc.

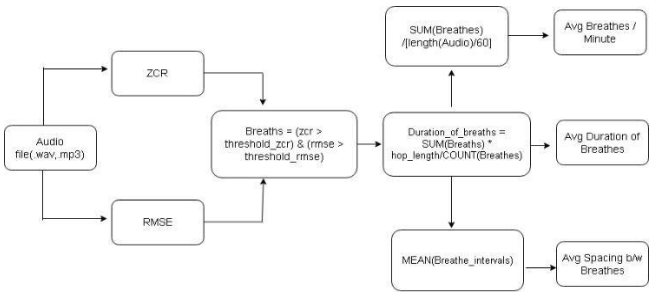

Fig 2: Pipeline to generate features

Here we extract 3 features from audio that is *"Average breaths per minute", "Average duration of breathes" and "Average spacing between breaths"*. and study their characteristics as mentioned in Fig 1

Table 2: Parameters for feature extraction as taken from Every Breath You Do Not Take

Parameter	Definition	Value
Window Length	Duration of audio segment that is analyzed at a time	0.02 s
Hop Length	Amount of time Analysis window is shifted forward	0.0025 s
Threshold ZCR	Minimum zero-crossing rate for significance.	0.1
Threshold RMSE	Minimum Root Mean Square Energy	0.1

Description of features and their measures

1. ***Average breaths per minute:*** This feature represents the estimated number of breaths taken per minute in the audio sample. It is calculated by counting the number of detected breath segments and normalizing over the total duration of the audio file in minutes.

   Nb=Number of detected Breaths, T=Total Duration of audio

$$ABPM = \frac{N_b}{T} \times 60$$

2. ***Average duration of breathes:*** This feature indicates the average length of each breath, measured in seconds. It is computed by determining the duration of each breath segment, then averaging these durations. (seconds)

   Nd=Number of detected Breath duration, D=Duration of each detected breath in seconds

$$ADB = \frac{1}{N_d} \sum_{i=1}^{N_d} D_i$$

3. ***Average Spacing between breathes:*** This feature measures the average time interval between successive breaths. After each breath's end, the time until the start of the next breath is calculated, and the mean of these intervals is taken. (seconds)

   Ni=Number of intervals between detected breaths

   Ij=Time interval between detected breaths

$$ASB = \frac{1}{N_i} \sum_{j=1}^{N_i} I_j$$

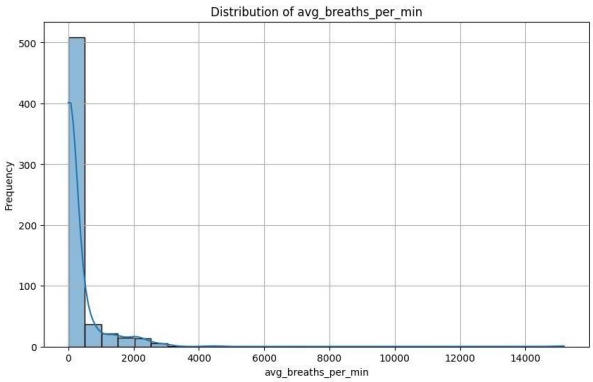

Fig 3: Distribution of "Average breath per minute" for Real audio samples

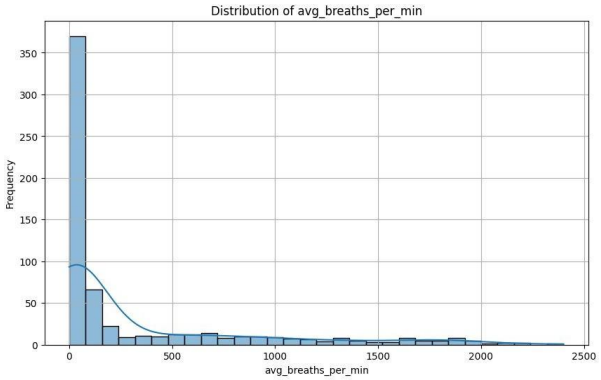

Fig 4: Distribution of "Average breath per minute" for Fake audio samples

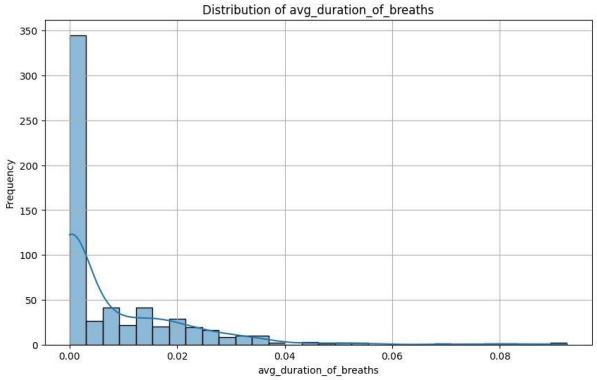

Fig 5: Distribution of "Average duration of breath" for Real audio samples

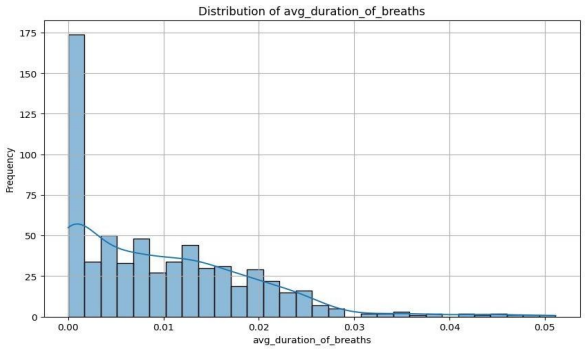

Fig 6: Distribution of "Average duration of breaths" for Fake audio samples

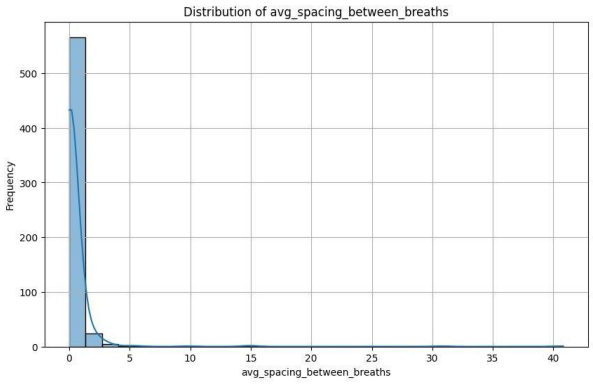

Fig 7: Distribution of "Average Spacing between breaths" for Real audio samples

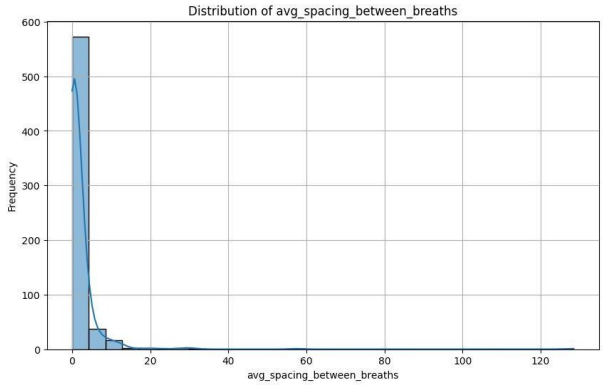

Fig 8: Distribution of "Average Spacing between breaths" for Fake audio samples

Figures 3-8 shows the histogram of distribution of Breath features for fake and real audio from the entire dataset.

MANN WHITENY U TEST: is a nonparametric test used to compare differences between two independent samples, especially when the sample distributions are not normally distributed. Mann Whitney U being a non-parametric test that does not assume a specific distribution of data. The test ranks all the data points from both groups and compares the ranks, which can provide insights into the differences in distributions without making strong assumptions about the underlying data. We apply this technique to study how the underlying distribution of breath features differ in the 10 datasets taken for study. The study is done for 75 Samples of Fake and Real class.

Null Hypothesis (H0): There is no significant difference in the breath features (average breaths per minute, average duration of breaths, and average spacing between breaths) between fake and real audio samples.

Alternative Hypothesis (H1) : There is a significant difference in the breath features between fake and real audio samples.

p-value < 0.05: This indicates that there is strong evidence against the null hypothesis, leading to its rejection in favor of the alternative hypothesis.

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

$$U = \min(U_1, U_2)$$

$$\begin{cases} \text{Reject } H_0 & \text{if } p \le U \\ \text{Fail to reject } H_0 & \text{if } p > U \end{cases}$$

R1: is the sum of the ranks for the first group.
R2: is the sum of the ranks for the second group.
n1: is the number of observations in the first group.
n2: is the number of observations in the second group.

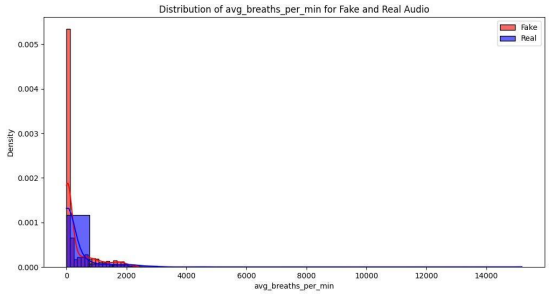

Fig 9: Distribution of "Avg_Breath_per_minute" for Fake and Real audio

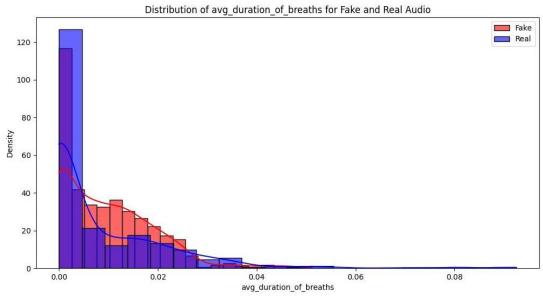

Fig 10: Dis

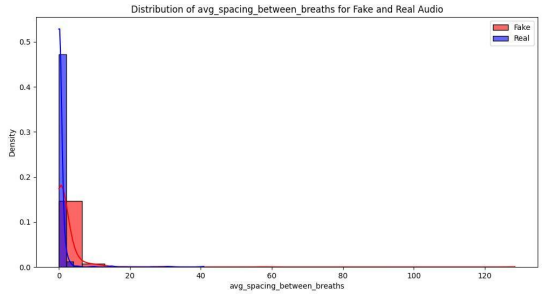

Fig 11: Distribution of Avg_Spacing_Between_breaths for Fake and Real audio

SUMMARY TABLE

Table 3: Result of Statistical analysis

Dataset	Average breathe per minute	Average duration of breaths	Average spacing between breaths	Overall assessment
FAI-1	No significant difference	No Significant difference	Significant difference	33%
FAI-2	Significant difference	No Significant difference	Significant difference	66%
FAI-3	No Significant difference	Significant difference	Significant difference	66%
ASVSpoof20	Significant	No Significant	No Significant	33%

21	difference	difference	difference	
In the wild	Significant difference	Significant difference	No significant difference	66%
Every breathe you do not take	Significant difference	Significant difference	Significant difference	100%
Half truth	Significant difference	No Significant difference	No significant difference	33%
Fake or Real	No Significant difference	Significant difference	No significant difference	33%
Deep voice	No Significant difference	No Significant difference	No Significant difference	0%
Wavefake	Significant difference	No Significant difference	Significant difference	66%

2) Classification with Statistical Algorithms

The results of Table 2 and Figures 9-11 shows the potential discriminative ability of breath features across multiple datasets which can be inferred by the difference in their respective distributions in Fake and Real audio.

In this study, we evaluate the efficacy of three breath features—Average Breaths Per Minute, Average Duration of Breaths, and Average Spacing Between Breaths—in classifying audio samples as "Fake" or "Real." by training statistical Machine learning models with the breath features in isolation.

The present study focuses on the FA-1, FA-2, and FA-3 datasets, as our preliminary analysis and existing literature, such as [14], suggest that the duration of audio recordings has a significant impact on the extracted breath features and subsequent results. Specifically, the length of audio samples can influence the accuracy and reliability of breath feature extraction, which in turn affects the performance of machine learning models. To mitigate this potential source of variability and ensure a more controlled environment for our experiments, we have deliberately selected these datasets, which provide a suitable range of audio lengths and characteristics.

Table 4: Classification with Statistical Machine learning and breath features

Model	Accuracy	EER in 'in the wild'
Logistic Regression	72%	84%
Random Forest	71%	90%
SVM RBF	72%	80%
KNN	71%	90%
Naive Bayes	67%	84%

The trained model is subsequently validated against a diverse "in the wild" dataset, which comprises a wide range of audio recordings from various sources and environments, to assess its generalizability and robustness in real-world scenarios.

Our findings from table 3 suggest that while the selected breath features are effective in capturing the distribution of the training data when trained in isolation, they exhibit limitations in generalization when applied to audio files generated by alternative methods as inferred by the high Equal Error Rates when generalizing on unseen "in the wild" dataset.

3) Classification with Audio Integrity Net

Further, we improve the classifier by incorporating a Convolutional Neural Network (CNN) architecture for detecting audio deepfakes, leveraging stacked features such as Mel Frequency Cepstral Coefficients (MFCC), Log-Spectrograms, and a binary breath sequence.

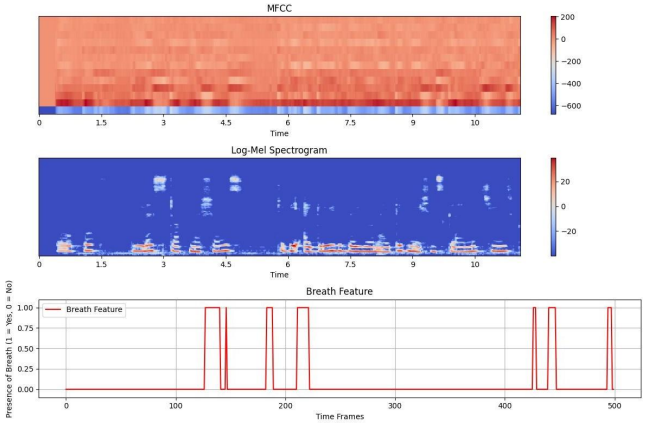

Fig 12: MFCC, LogSpec and Breath for Real Audio file, for a particular spoken sentence

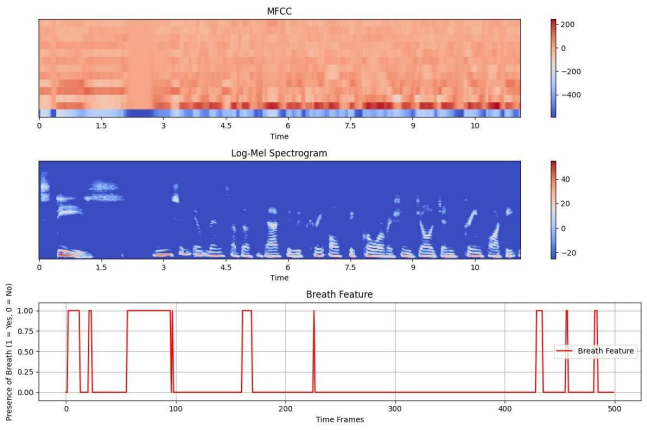

Fig 13: MFCC, LogSpec and Breath for Fake audio file, for the same spoken sentence as Fig 12

We employ MFCCs to extract the short-term power spectrum of sound and characterize human speech in both real and synthetic audio. Additionally, we utilize log-spectral features to analyze the frequency spectrum of audio signals over time, facilitating the identification of anomalies. The intersection of root-mean-square (RMS) and zero-crossing rate (ZCR) features, which

correspond to breath patterns, enables the detection of anomalies within the signal and aids in classifying incoming signals as either real or fake as represented in Fig 12 & 13. Furthermore, the fusion of these features enhances the generalizability of our approach.

Our proposed AudioIntegrityNet model is trained on the same FA-1, FA-2, FA-3 datasets. Following training, the AudioIntegrityNet model is then validated against a diverse "in the wild" dataset to asses it's generalizability.

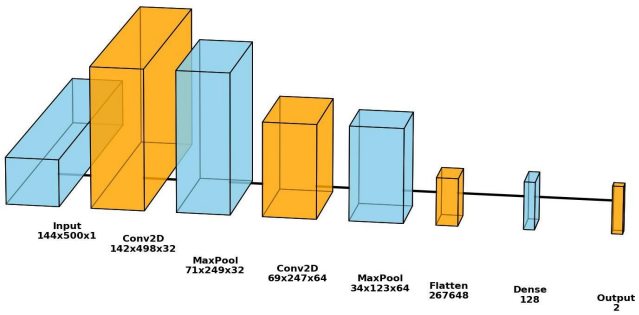

Fig 14 : Audio Integrity Net Architecture

The proposed CNN architecture as described in Fig 14, AudioIntegrityNet [22], is designed to effectively detect audio deepfakes by leveraging a series of convolutional and pooling layers for feature extraction, followed by dense layers for classification.

The model begins with an initial convolutional layer to capture essential audio features, followed by max pooling and dropout layers to reduce dimensionality and mitigate overfitting. This is succeeded by a second convolutional layer that further refines the feature representation. The output is flattened and processed through a dense layer, culminating in a final output layer that employs softmax activation to classify audio samples as either real or deepfake. The model is trained using categorical cross entropy as the loss function.

IV. RESULTS

Table 5: Performance metrics of AudioIntegrityNet

Model Evaluation Metrics for AudioIntegrityNet	*Results*
Training accuracy	92%
Validation Accuracy	82%
Precision	87%
Recall	74%
F1 Score	79%
AUC	90%

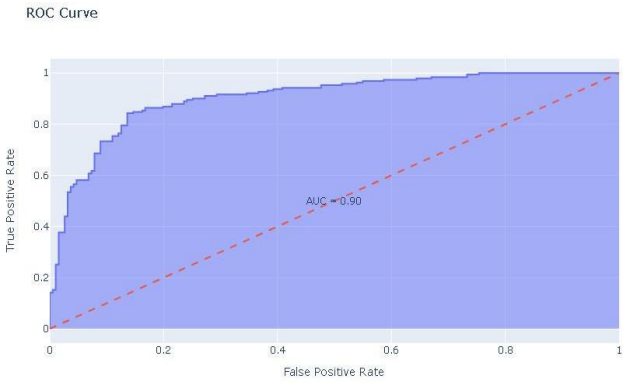

Fig 15: ROC curve and AUC

Table 6 : Equal Error Rate of AudioIntegrityNet on different datasets

Dataset	*Equal Error Rate (%)*
In The Wild (>4 seconds)	32.50
Every Breath you Do not Take	18.75
DeepVoice	25.00
WaveFake full band melgan	26.09
Validation set	11.79

Our AudioIntegrityNet [22] exhibits strong performance as a classifier when evaluated on our validation set, as illustrated in Table 5. The model achieves an impressive overall Area Under the Curve (AUC) score of 90%, indicating its robust classification capability and suggesting that it can effectively distinguish between genuine and manipulated audio. Furthermore, when independently validated against the "In the Wild" dataset, the model attains an Equal Error Rate (EER) of 32.50% (refer to Table 6), showcasing its reliability in real-world scenarios. Notably, the model performs even better on the "Every Breath You Do Not Take" dataset, achieving an EER of 18.75%.

These results underscore the competitive performance of AudioIntegrityNet in the realm of audio deepfake detection. When compared to existing models, such as the MesoInception model, which reports an EER of 37.4% on the "In the Wild" dataset, and RawNet2, which achieves an EER of 33.94% on the same dataset [20], our model demonstrates a measurable improvement. This advancement not only highlights the effectiveness of our approach, but also positions AudioIntegrityNet as a valuable tool for enhancing audio integrity verification in various applications.

V. CONCLUSION

In conclusion, our study provides evidence of a significant difference in the distribution of Root Mean Square (RMS) and Zero-Crossing Rate (ZCR) when

crossing a set threshold value between real and synthetic audio samples as inferred from the variation in their respective data distribution. We also infer that these features give the best results when trained in conjunction with other short term spectral features. The performance of our model, AudioIntegrityNet, which was trained on Mel-Frequency Cepstral Coefficients (MFCC), log spectrograms, and breath features, demonstrates its capability to effectively classify audio as real or fake. Achieving an Equal Error Rate (EER) of 32.50% during independent validation against an "in-the-wild" dataset. The study underscores the model's robustness and potential applicability in real-world scenarios. These findings contribute to the ongoing efforts in audio forensics and integrity verification, paving the way for future research aimed at enhancing the detection of audio deepfakes.

VI. SCOPES OF FURTHER RESEARCH

There remains significant areas for further research into the intersection of Root Mean Square (RMS) and Zero-Crossing Rate (ZCR) across various threshold values. Future studies should investigate how these distributions are influenced by diverse factors, including the nature of the language—as well as the impact of background noise on these metrics. Additionally, it is essential to explore how demographic characteristics, such as gender and age, affect RMS and ZCR values. Furthermore, examining the influence of respiratory conditions on these features could provide valuable insights into the nuances of audio deepfake detection.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Professor James KC for his invaluable guidance and support throughout the course of this research. I also wish to acknowledge the contributions of the M.Tech batch of 2023-25, whose assistance in the manual data collection process was essential. Their efforts in gathering both authentic and fabricated samples of spoken sentences from Frontier AI startups significantly contributed to the success of this study.

REFERENCES

[1] *Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C. Y., & Zhao, Y. (2023). Audio deepfake detection: A survey. arXiv preprint arXiv:2308.14970.*

[2] *World Economic Forum 'Year of elections: Lessons from India's fight against AI-generated misinformation' https://www.weforum.org/stories/2024/08/deepfakes-india-tackling-ai-generated-misinformation-elections/*

[3] *The Battle against AI-driven Identity fraud , Signicat. https://www.signicat.com/the-battle-against-ai-driven-identity-fraud*

[4] *Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., & Roberts, A. (2019). Gansynth: Adversarial neural audio synthesis. arXiv preprint arXiv:1902.08710.*

[5] *Donahue, C., McAuley, J., & Puckette, M. (2018). Adversarial audio synthesis. arXiv preprint arXiv:1802.04208.*

[6] *Caillon, A., & Esling, P. (2021). RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. arXiv preprint arXiv:2111.05011*

[7] *Kong, Z., Ping, W., Huang, J., Zhao, K., & Catanzaro, B. (2020). Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761.*

[8] *Tokuda, K. (1999). Speech synthesis based on hidden markov models.* 電子情報通信学会技術研究報告, *99(255 (SP99 55-61)), 47-54.*

[9] *Frank, J., & Schönherr, L. (2021). Wavefake: A data set to facilitate audio deepfake detection. arXiv preprint arXiv:2111.02813.*

[10] *Pham, L., Lam, P., Nguyen, T., Nguyen, H., & Schindler, A. (2024, September). Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models. In 2024 IEEE 5th International Symposium on the Internet of Sounds (IS2) (pp. 1-5). IEEE.*

[11] *Yang, Y., Qin, H., Zhou, H., Wang, C., Guo, T., Han, K., & Wang, Y. (2024, April). A robust audio deepfake detection system via multi-view feature. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 13131-13135). IEEE.*

[12] *Kawa, P., Plata, M., & Syga, P. (2022, December). Specrnet: Towards faster and more accessible audio deepfake detection. In 2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom) (pp. 792-799). IEEE.*

[13] *Liu, R., Zhang, J., Gao, G., & Li, H. (2023). Betray oneself: A novel audio deepfake detection model via mono-to-stereo conversion. arXiv preprint arXiv:2305.16353.*

[14] *Layton, S., De Andrade, T., Olszewski, D., Warren, K., Butler, K., & Traynor, P. (2024). Every Breath You Don't Take: Deepfake Speech Detection Using Breath. arXiv preprint arXiv:2404.15143.*

[15] *Doan, T. P., Nguyen-Vu, L., Jung, S., & Hong, K. (2023, June). BTS-E: Audio deepfake detection using breathing-talking-silence encoder. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.*

[16] *Sabry, A. H., Bashi, O. I. D., Ali, N. N., & Al Kubaisi, Y. M. (2024). Lung disease recognition methods using audio-based analysis with machine learning. Heliyon, 10(4).*

[17] *Ciftci, U. A., Demir, I., & Yin, L. (2020, September). How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals. In 2020 IEEE international joint conference on biometrics (IJCB) (pp. 1-10). IEEE.*

[18] *Bird, J. J., & Lotfi, A. (2023). Real-time detection of ai-generated speech for deepfake voice conversion. arXiv preprint arXiv:2308.12734.*

[19] *Choi, J. E., Schäfer, K., & Zmudzinski, S. (2024, June). Introduction to Audio Deepfake Generation: Academic Insights for Non-Experts. In Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation (pp. 3-12).*

[20] *Müller, N. M., Czempin, P., Dieckmann, F., Froghyar, A., & Böttinger, K. (2022). Does audio deepfake detection generalize?. arXiv preprint arXiv:2203.16263.*

[21] *Gasenzer, K., & Wolter, M. (2023). Towards generalizing deep-audio fake detection networks. arXiv preprint arXiv:2305.13033..*

[22] *Blue, L., Warren, K., Abdullah, H., Gibson, C., Vargas, L., O'Dell, J., ... & Traynor, P. (2022). Who are you (i really wanna know)? detecting audio {DeepFakes} through vocal tract reconstruction. In 31st USENIX Security Symposium (USENIX Security 22) (pp. 2691-2708).*

[23] *Prasanth 2025, Audio Deepfake Detection, Huggingspace: https://huggingface.co/spaces/2000prasanth/audio_deepfake_detection_audiointegritynet*

\