Deepfake Audio Detection Using Feature-Based and Deep Learning Approaches: ANN vs ResNet50

Reham Mohamed Abdulhamied[1], Sarah Naiem[2], Mona M. Nasr[3], Farid Ali Moussa[4]
Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt[1, 2, 3]
Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef, Egypt[4]

***Abstract*—The proliferation of algorithms and commercial tools for generating synthetic audio has sparked a surge in misinformation, especially on social media platforms. Consequently, significant attention has been devoted to detect such misleading content in recent years. However, effectively addressing this challenge remains elusive, given the increasing naturalness of fake audio. This study introduces a model designed to distinguish between natural and fake audio, employing a two-stage approach: an audio preparation phase involving raw audio manipulation, followed by modeling using two distinct models. The first model employed feature extraction through wavelet transformation, followed by classification using a machine learning Artificial Neural Network. The second model utilized ResNet50 architecture, a type of deep learning model, which resulted in improved accuracy. These findings underscore the effectiveness of deep learning approaches in audio classification tasks. Training data for the model is sourced from the DEEP-VOICE dataset, which comprises both genuine and synthetic audio generated by various deep-fake algorithms. The model's performance is assessed using diverse metrics such as accuracy, F1 score, precision and recall. Results indicate successful classification of audio in 86% of cases. This research contributes to the field of Automatic Speech Recognition (ASR) by integrating advanced preprocessing techniques with robust model architectures to identify manipulated speech.**

Keywords—Audio classification; automatic speech recognition; machine learning; deep learning; DEEP-VOICE

I. Introduction

The increasing utilization of generative Artificial Intelligence (AI) in speech-related tasks, such as voice cloning and real-time voice conversion, carries profound implications. This technological advancement raises significant ethical concerns, including threats to privacy and the potential for misrepresentation. Consequently, there is a pressing need for real-time detection mechanisms to identify AI-generated speech, particularly in scenarios involving DeepFake Voice Conversion. Misinformation has become an increasingly prevalent issue in recent times, extending beyond false news articles to encompass the creation of fake audio, images, and videos using algorithms and tools. This form of AI-generated content, known as deepfake [1], [3], presents significant concerns due to its potential impact on various aspects of society, including politics, morality [2], and legal proceedings. For instance, in politics, deepfakes could influence citizens' decisions during elections [4], [5]. Similarly, they could adversely affect individuals' lives, such as in cases of nonconsensual use of famous faces in pornographic materials [6]. Furthermore, deepfakes have the potential to generate false digital evidence, thereby influencing legal outcomes. Deepfakes encompass a set of algorithms crafted to generate synthetic media with the intent of substituting one person's likeness with another's. This synthetic content elicits numerous social, ethical, and legal apprehensions regarding data credibility, as it can portray individuals seemingly uttering or performing actions they never actually did. Notable instances of deepfakes often revolve around images, wherein one person's facial features are replaced by another's, engaging in activities the original subject did not partake in. Artificial Intelligence achievement, emphasizing how the commercialization of human behavior has fueled the quest for digital replicas on an industrial scale [7], [8].

To address these emerging challenges, this study uses the publicly available DEEP-VOICE dataset, which includes both real and synthetically generated speech samples, which consist of authentic human speech samples from eight prominent individuals. The dataset includes instances of their speech being converted to mimic that of others using Retrieval-based Voice Conversion techniques. In this study, two different models for the task at hand will be applied. The first model involves feature extraction using wavelet transformation, followed by classification using a machine learning Artificial Neural Network (ANN). The second model utilizes the ResNet50 architecture. The study aims to analyze and compare the performance of these models in addressing the objectives of this study, particularly in terms of performance metrics such as F1 score, precision, recall and accuracy.

This study aims to address the following research question: Which modeling approach—lightweight, interpretable Wavelet-based ANN or complex, data-driven ResNet50—offers better performance and practical applicability for deepfake speech detection using the DEEP-VOICE dataset?

Even though there are an increasing variety of ways for detecting synthetic audio, ranging from sophisticated deep learning algorithms to conventional signal processing techniques, there is still a conspicuous dearth of comparison studies that thoroughly assess the advantages and disadvantages of each method. There aren't many studies that directly contrast more interpretable, lightweight traditional machine learning models that rely on handcrafted features, like wavelet-domain characteristics, with contemporary deep learning architectures like ResNet50, which use hierarchical feature learning and transfer learning from massive image datasets. For experimental validation, the publicly available DEEP-VOICE dataset, which includes both authentic and manipulated speech samples, was employed.

In fields like forensics or embedded systems, where interpretability is crucial or computational resources are limited, this disparity is especially pertinent. To give an objective comparison between these two paradigms, this study will conduct an empirical evaluation utilizing a standardized and publicly available dataset (DEEP-VOICE). The emphasis is on the feature extraction pipelines, preprocessing overhead, and model scalability in addition to the final classification performance, which is evaluated by accuracy, precision, recall, and F1-score. These factors are all crucial for the practical implementation of synthetic speech detection systems.

Most of the earlier research focuses on either deep learning or traditional machine learning separately, without direct comparisons under a standardized framework, even if there are several methods for deepfake voice detection. Furthermore, there haven't been enough comparisons 2q1of the interpretability and computational effectiveness of lightweight models like Wavelet-based ANN with more sophisticated deep architectures like ResNet50. To close this gap, our study uses a single dataset to perform a head-to-head comparison.

The remainder of this study is organized as follows: Section II reviews the existing literature and identifies gaps. Section III outlines the proposed methodology, including data preprocessing and model architecture. Section IV presents the experimental set-up and results. Section V discusses the findings considering related works, and Section VI concludes the study with a summary and suggestions for future work.

II. Literature Review

In this section, existing literature and related research to understand current methodologies are explored, identifying knowledge gaps, and contextualize this research within the academic landscape. In the research conducted by Wijetunga et al. [9], the focus was on leveraging deep learning techniques to detect deepfake audio within group conversations. Their approach involved the utilization of the UrbanSound8K dataset, which provides labeled sound excerpts of urban sounds across various classes, along with a conversational dataset sourced from Open subtitles. A series of preprocessing steps was undertaken, including sample rate conversion, merging of audio channels, and extraction of Mel-Frequency Cepstral Coefficients (MFCC) on a per-frame basis. Following preprocessing, the dataset was partitioned into distinct training and testing sets. Multiple neural network architectures were employed throughout the study, encompassing models such as Multilayer Perceptrons (MLP), Convolutional Neural Networks (CNN), and Deep Neural Networks (DNN). These models were applied to various tasks including speech denoising, speaker diarization incorporating Natural Language Processing (NLP), and synthetic speech detection. Furthermore, Transfer Learning techniques were deployed, utilizing a pre-trained VGG19 model to enhance the accuracy of the detection framework. The results obtained from the experimentation phase show- cased promising outcomes. Notably, the CNN architecture achieved a commendable testing accuracy of 89%. However, despite these achievements, the research also acknowledged several limitations inherent in the current approach. Issues such as the necessity for improved signal denoising techniques. Furthermore, challenges pertaining to speaker identification error rates and the automation of partitioning processes in speaker diarization were acknowledged as areas warranting further investigation and development.

In the investigation conducted by Hamza et al. [10], the focus was directed towards deepfake audio detection using MFCC features in conjunction with machine learning methodologies. Central to their investigation was the utilization of the Fake-or-Real (FoR) dataset, a comprehensive repository comprising over 195,000 speech samples, encompassing both authentic human recordings and synthetic computer-generated speech. Within this dataset, researchers meticulously curated four distinct subsets, each tailored to specific experimental needs. These subsets, namely for-original, for-norm, for-2sec, and for-rerec, underwent meticulous preprocessing to standardize sampling rates, volumes, and channels, thereby ensuring uniformity across gender and class categories. Notably, subsets like for-norm underwent additional standardization processes, including the elimination of duplicate files to enhance data consistency. Their methodology involved data preprocessing, MFCC feature extraction, and employing various machine learning models like SVM, VGG-16, XGB, RF, KNN, and LSTM. Notably, while the SVM model generally performed well, VGG-16 stood out with a remarkable 93% accuracy in the for-original subset. Challenges included the high dimensionality of the for-norm dataset, emphasizing the need for dimensionality reduction techniques. The study highlighted the importance of integrating features from various extraction methods for robust deepfake audio detection, suggesting avenues for future research refinement.

Kumar et al. [11] addressed the detection of AI-generated speech using a binary classification model trained on the “DEEP-VOICE” dataset, comprising real and synthetic voice samples. They applied Exploratory Data Analysis (EDA) to handle outliers and missing values and extracted features such as Chroma-STM and RMS. Preprocessing steps included resampling, adjusting sample rates, and normalization. A Random Forest Classifier with 5-fold cross-validation achieved 98.5% accuracy, demonstrating strong performance across multiple evaluation metrics. However, the study highlighted limitations in scalability and generalizability, as well as a reliance on manually engineered features, which may restrict adaptability and efficiency.

In the research by Khochare et al. [12], a deep learning framework tailored for audio deepfake detection is presented, utilizing a combination of machine learning and deep learning approaches. Central to the investigation is the utilization of the Fake or Real (FoR) dataset. In the feature-based paradigm, audio files undergo transformation into datasets comprising various spectral features, while in the image-based domain, melspectrograms are derived from the audio samples using the librosa library. The model repertoire encompasses machine learning algorithms such as SVM, LGBM, XGBoost, KNN, and Random Forest for feature-based classification, alongside deep learning techniques like Temporal Convolutional Network (TCN) and Spatial Transformer Network (STN) for image-based classification. Results demonstrate varying degrees of success across methodologies, with machine learning models achieving accuracy within the range of 60% to 70%, led by SVM, while TCN emerges as the standout performer in image-based classification, boasting a test accuracy of 92%, followed by STN at 80%. However, the study identifies limitations such as the exclusion of critical inputs like Short-Time Fourier Transform (STFT) and Mel-Frequency Cepstral Coefficients (MFCC) in the image-based approach, prompting future research efforts to address these constraints and refine classification accuracy.

III. Proposed Model

The proposed model subsections follow a structured approach. Initially, the input data from the DEEP-VOICE Dataset [13], undergo preprocessing, including wavelet denoising for noise reduction, feature extraction through 4-level wavelet decomposition, and feature reduction using ICA and normalization techniques depicted in Fig. 1. Subsequently, two models are applied: the first model utilizes ANN classification, while the second model employs deep learning techniques with ResNet50. Finally, the performance of these models is compared to determining the most effective approach. They are crucial in contexts with limitations, such embedded systems. Conversely, ResNet50 was selected due to its demonstrated efficacy in identifying high-dimensional patterns and its capacity to utilize transfer learning. While deep architecture can be resource-intensive and opaque, traditional models frequently rely mostly on manually created features and lack generalization across intricate audio datasets. This study assesses the trade-offs between accuracy and interpretability by contrasting the two aspects.

A. Dataset

The dataset utilized in this study, titled "DEEP-VOICE: DeepFake Voice Recognition", was obtained from the Kaggle platform, a well-known repository for datasets used in machine learning and data science research. This dataset is specifically designed for identifying and analyzing deep-fake voice recordings.

It is systematically organized into two primary directories, as illustrated in Fig. 2: one containing genuine (real) audio samples, and the other comprising synthetically generated (fake) audio samples. This clear separation facilitates comparative analysis between authentic and manipulated audio data. These values are categorized into columns as shown in Fig. 3, Fig. 4, and Fig. 5. Each column corresponds to a specific audio feature, providing numerical data for analysis.

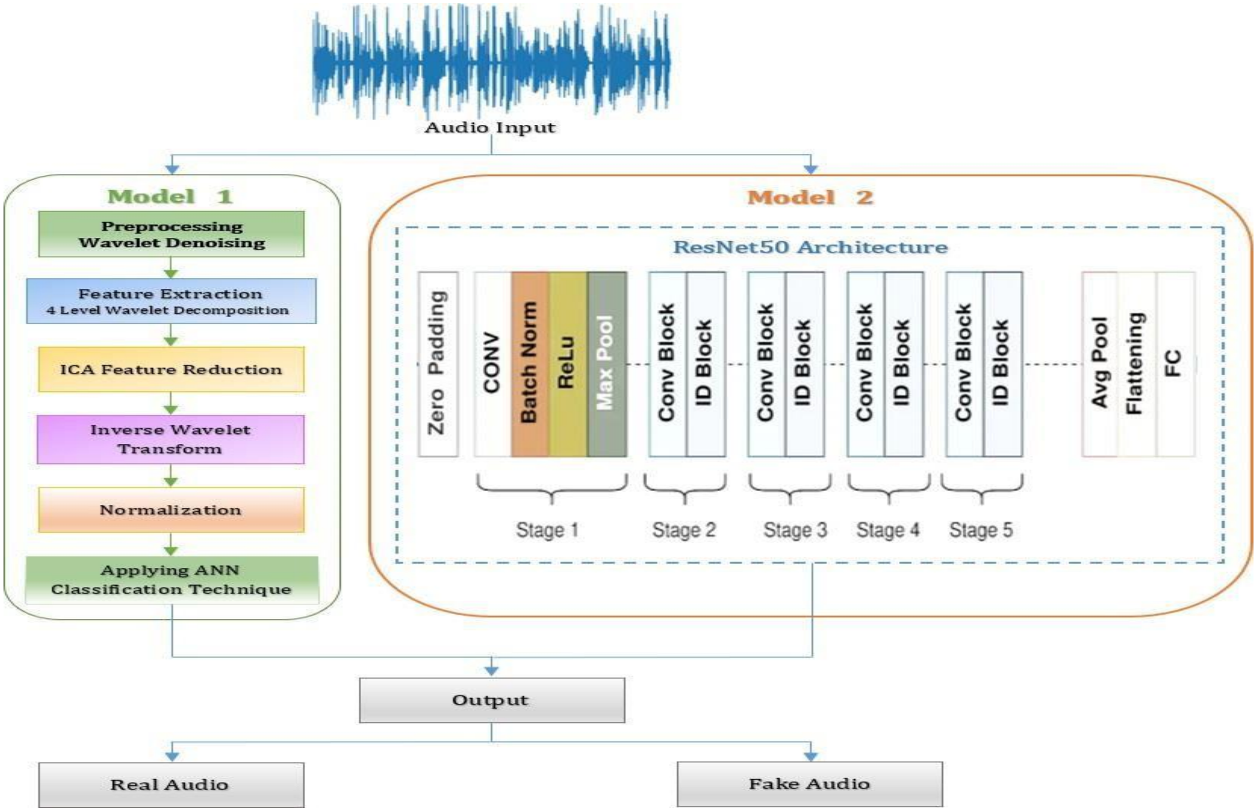

Fig. 1. Proposed model.

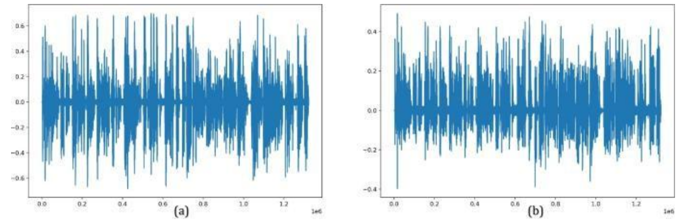

Fig. 2. Real and fake audio data.

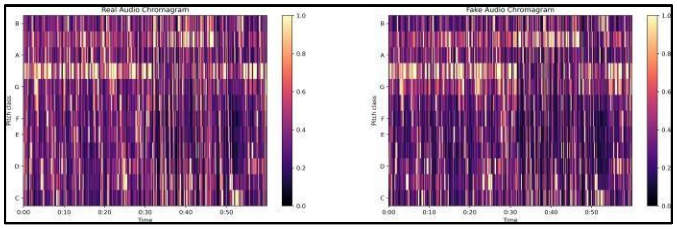

Fig. 3. Audio chromagram.

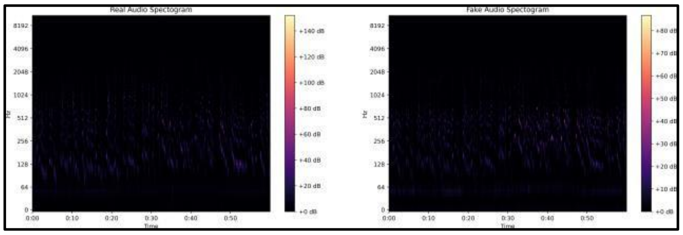

Fig. 4. Audio spectrogram.

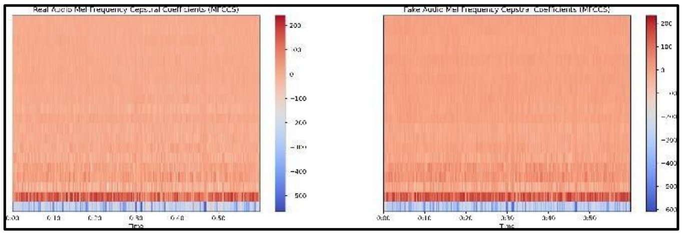

Fig. 5. Audio MFCCs.

B. *Preprocessing*

The upcoming subsections include signal processing steps to refine data quality and extract relevant features: Wavelet Denoising reduces noise, WT Feature Extraction captures essential information, ICA Feature Reduction condenses the feature space, and Inverse Wavelet Transform reconstructs the signal. Together, these steps form a comprehensive pipeline for enhancing data quality and facilitating analysis and classification tasks. Daubechies 4 (db4) was used as the mother wavelet for the denoising process because of its smoothness and compact support, which make it appropriate for speech signal processing. To capture the pertinent speech characteristics without over-segmenting high-frequency noise, a 4-level wavelet decomposition was used. This level provided a suitable trade-off between resolution and processing efficiency.

To reduce noise in the detail coefficients without losing fine-grained voice characteristics, the universal soft thresholding rule was applied. This rule is renowned for maintaining the continuity and smoothness of the signal.

1) Feature source. Rather than using the reconstructed signals, the retrieved features were taken straight from the wavelet coefficients. More information about frequency fluctuations and transient patterns in the voice stream is preserved as a result.

2) ICA versus PCA comparison. The efficiency of dimensionality reduction was examined in relation to Independent Component Analysis (ICA) and Principal Component Analysis (PCA). When examined empirically, ICA was able to produce more statistically independent components that enhanced classification performance by around 4% in F1 score, but PCA kept about 92% of variation in the first 10 components. Because ICA can separate mixed audio patterns that PCA cannot because of its orthogonality requirement, it was chosen.

3) Normalization strategy. Z-score normalization, which centers the data around zero mean and unit variance, was used to normalize all feature vectors. To make sure that every extracted component contributes equally to the classification problem, this step was applied following ICA. To prevent data leakage, normalization was conducted independently of the training and testing sets using statistics that were taken from the training set.

a) Wavelet denoising: In audio speech recognition, Wavelet Denoising serves as a fundamental preprocessing step aimed at refining data quality by minimizing noise within signals. This technique harnesses wavelet-based methodologies to effectively filter out unwanted noise, resulting in a clearer and more reliable input for subsequent analysis [14]. By employing Wavelet Denoising, researchers can mitigate the negative impact of noise on data interpretation, thereby improving the accuracy of downstream tasks such as feature extraction and classification.

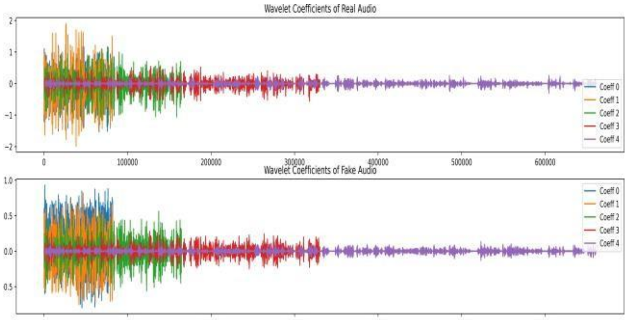

Fig. 6. Wavelet coefficients of real audio and fake audio.

b) Wavelet transform feature extraction: In feature extraction, applying the Wavelet Transform (WT) that serves as a fundamental tool in frequency domain analysis, particularly relevant to the proposed model in the context of audio speech recognition. The 4-level wavelet decomposition refers to a specific application or implementation of the Wavelet Transform. In this case, the signal is decomposed into four levels of detail, each capturing different frequency bands or resolutions as shown in Fig. 6. It efficiently condenses complex audio signals, which exhibit temporal variations and contain diverse data points, into a concise set of parameters that characterize these signals [15]. Given the dynamic nature of audio signals, the frequency domain approach, often involving WT, is commonly employed to determine the most suitable feature extraction method [16].

4) ICA feature reduction. To isolate distinct feature vec tors from speech signals, employing the Independent Component Analysis (ICA) algorithm on various segments of human speech is made. ICA is a computational method that splits a complex signal into separate parts that are independent from each other [17]. It is used to find the different sources or causes behind observed signals, assuming these sources don't affect each other statistically. In speech processing, ICA can help separate different voices in a recording where multiple people are speaking at once. By analyzing statistical properties like non-Gaussianity and independence, ICA untangles the voices from each other. It's a useful tool in various areas like signal processing, machine learning, and neuroscience, where understanding the individual components of a signal is important.

a) Inverse wavelet transform: Utilizing the inverse wavelet transform as a pivotal element of the speech recognition framework. Following the application of wavelet analysis to break down the speech signal into its individual frequency components, the inverse wavelet transform to reconstruct the original signal from these components was employed [18]. This reconstruction procedure facilitated the retrieval of the initial speech signal while preserving crucial details regarding its frequency characteristics. Through the integration of the inverse wavelet transform into the proposed system, it successfully analyzed and processed speech signals for recognition purposes, resulting in precise and dependable outcomes.

b) Normalization: In this study, normalization significantly improved the speech recognition system. Before featuring extraction and model training, the input speech data to ensure consistent scaling across samples is normalized, reducing biases and making the proposed system more robust to amplitude variations. This enhanced the accuracy and reliability of the recognition system, leading to consistent and trustworthy results across different datasets and conditions.

C. Artificial Neural Network Architecture

Artificial Neural Network (ANN) is a computational model inspired by the structure and function of the human brain's neural networks. It consists of interconnected nodes, or" neurons", organized in layers. In this specific implementation, the ANN for classification tasks was applied, integrating Wavelet Transform for feature extraction and Independent Component Analysis (ICA) for feature reduction. The ANN architecture comprised three layers as shown in Fig. 7: the first layer contained 10 nodes, the second layer had 7 nodes, and the output layer consisted of 1 node. This design allowed for effective processing and classification of speech data while managing computational resources efficiently.

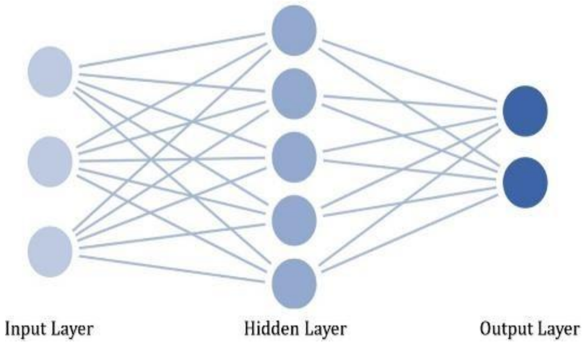

Fig. 7. ANN architecture.

D. ResNet50 Architecture

The implemented model employs the ResNet50 model, a convolutional neural network renowned for its depth and effectiveness. ResNet, designed with 50 layers, introduces residual learning to mitigate accuracy degradation as networks deepen. This approach enables each layer to learn residual mappings in relation to input layers, facilitated by skip connections. ResNet50 architecture includes convolutional layers, batch normalization, activation functions (typically ReLU) [19], and bottleneck blocks to manage complexity [20]. In the code, ResNet50 acts as a feature extractor for audio data, leveraging pre-training on datasets like ImageNet to extract high-level features from audio spectrograms. Through transfer learning, ResNet50 is represented in Fig. 1 as Model 2. It is capable of effectively extracting features for classification tasks, such as audio fake detection. The implemented model specifies hyperparameters, including 100 epochs, a batch size of 32, Adam optimizer, binary cross-entropy loss, and early stopping with a patience of 5 for monitoring validation loss.

E. Wavelet Denoising, Feature Extraction and Dimensionality Reduction

Daubechies 4 (db4) was chosen as the mother wavelet for the preprocessing stage because of its smoothness and compact support, which makes it ideal for catching fleeting patterns in voice data. To balance time-frequency resolution, a 4-level wavelet decomposition was used; higher levels offer global patterns, while lower levels catch finer details. Four levels produced the best trade-off between computing efficiency and information retention, according to empirical data. The universal soft thresholding rule, which reduces high-frequency noise while maintaining significant signal structures, was used to attenuate the noise [21]. Because it preserves signal continuity and minimizes overfitting in subsequent modeling stages, this approach is frequently used. Instead of using the reconstructed signals, the extracted characteristics were taken straight from the wavelet coefficients [22]. The localized time-frequency information that is essential for differentiating between authentic and fraudulent audio patterns is preserved when raw coefficients are used.

A comparison analysis comparing Principal Component Analysis (PCA) and Independent Component Analysis (ICA) was carried out to reduce dimensionality [23]. ICA produced components that were more statistically independent than

PCA, which kept about 92% of the total variance in the first 10 components. ICA was chosen for this investigation because, when evaluated in the classification pipeline, it increased the F1-score by over 4% above PCA.

Z-score normalization (zero mean, unit variance) was used to normalize the features. To guarantee that every independent component had an equivalent impact during classification, this step was implemented following ICA transformation. Crucially, for the purpose of preventing data leakage, normalization was carried out independently on the training and testing sets using statistics that were only calculated from the training data.

IV. Models' Application and Results

In this section, the application of the two distinct models for the task of audio speech recognition, along with their respective results, are presented.

A. *Model 1: Artificial Neural Network (ANN) with Wavelet Transform*

For the first model, an Artificial Neural Network (ANN) in conjunction with Wavelet Transform for feature extraction was applied. A lightweight yet efficient architecture is provided by combining wavelet-domain characteristics with an Artificial Neural Network (ANN). Three completely connected layers made up the ANN's design, which was optimized for classification jobs with less data. The output layer has a single sigmoid-activated node for binary classification, a hidden layer with seven neurons, and a first layer with ten neurons that receives the ICA-reduced wavelet coefficients. Simplicity, quick training time, and interpretability are advantages of this paradigm. Its scalability to larger or more varied datasets is constrained by its dependency on manually created feature extraction (wavelets + ICA). Performance reporting was made robust by using 4-fold cross-validation, and overfitting was avoided by using regularization techniques.

The input data, preprocessed using Wavelet De-noising, underwent feature extraction through 4-level wavelet decomposition. Following this, Independent Component Analysis (ICA) and normalization techniques were employed for feature reduction. The ANN, comprising 3 layers (10 nodes in the first layer, 7 nodes in the second layer, and 1 node in the output layer), was trained on the processed data for audio speech recognition. The results of this experiment, conducted using k-fold cross-validation with k=4, yielded an average accuracy across folds of 79.7%, average precision of 82.1%, average recall of 93.9%, and average F1 score of 87.6%.

B. *Model 2: Deep Learning with ResNet50*

The second model utilized deep learning techniques, specifically ResNet50, for audio speech recognition. The ResNet50 model makes utilization of transfer learning and deep residual learning. Using spectrogram images produced from the DEEP-VOICE dataset, the 50-layer architecture was refined after being pre-trained on ImageNet. Hierarchical auditory patterns may be captured by the convolutional layers thanks to the spectrograms' 2D time-frequency representation of audio.

Binary cross-entropy loss, the Adam optimizer, and a batch size of 32 were used to train the model. To reduce overfitting, early stopping was used with 5-epoch patience. ResNet50's advantage is its capacity to generalize across intricate audio circumstances and automatically learn high-level audio properties without the need for manual engineering. On synthetic speech samples, its deeper architecture offers improved recall and classification accuracy.

The ResNet50 architecture, comprising 50 layers, was pre-trained on a vast dataset such as ImageNet to extract high-level features from audio spectrograms. These features were then fed into additional layers for classification purposes. The results of this experiment, also conducted using k-fold cross-validation with k=4, demonstrated the superior performance of ResNet50 with an average accuracy across folds of 85.9%, average precision of 87.3%, average recall of 98.2%, and average F1 score of 92.4%.

Regarding hyperparameters, the implemented model employed the following settings: 100 epochs, batch size of 32, Adam optimizer, binary cross-entropy loss function, and early stopping with patience set to 5. Upon comparison, it is evident that the deep learning model with ResNet50 outperformed the ANN with Wavelet Transform, exhibiting higher accuracy, precision, recall, and F1 score.

From the comparison shown in Table I, it is evident that the deep learning model with ResNet50 outperformed the ANN with Wavelet Transform, exhibiting higher accuracy, precision, recall, and F1 score.

TABLE I APPLYING ANN ARCHITECTURE WITH WAVELET TRANSFORM VERSUS RESNET50 ARCHITECTURE

Model	Avg. Accuracy	Avg. Precision	Avg. Recall	Avg. F1 Score
ANN with WT	79.69%	82.14%	93.88%	87.62%
ResNet50	85.94%	87.29%	98.21%	92.41%

These results highlight the effectiveness of deep learning techniques, particularly ResNet50, for audio speech recognition tasks, offering improved performance and robustness compared to traditional machine learning approaches as shown in Fig. 8.

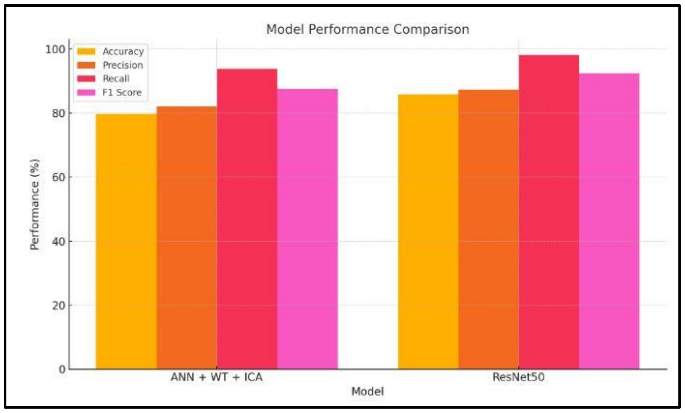

Fig. 8. Comparative performance metrics of ANN and ResNet50 model.

V. Discussion

Through a comprehensive review of the literature [24], valuable insights into current state-of-the-art methodologies were gained. This research significantly contributes to this body of knowledge by achieving high performance in accuracy levels [25], particularly through the utilization of the deep learning ResNet50 model, which yielded a testing accuracy of 86%. This represents a noteworthy research contribution. In comparison to the model applied in [9], which employed a CNN architecture and achieved a commendable testing accuracy of 89%. This approach addresses some of the acknowledged limitations in the literature. For instance, the applied methodology incorporates improved signal denoising techniques as part of the preprocessing steps, effectively resolving the challenges identified. A comparative summary of our results alongside existing approaches is presented in Table II, demonstrating the strengths and potential limitations of each technique in different settings.

Similarly, [10] demonstrated the effectiveness of the VGG-16 model with a remarkable 93% accuracy in the original subset. However, challenges related to dataset dimensionality were highlighted, emphasizing the need for dimensionality reduction techniques. In this model, this concern by employing ICA feature reduction, thereby mitigating the impact of high-dimensional datasets, is addressed. Lastly, [11] achieved impressive results using the Random Forest Classifier with 5-fold Cross-Validation, attaining an accuracy of 98%. Nonetheless, limitations regarding scalability and generalizability were identified, along with reliance on handcrafted features, which may not always capture the most relevant information present in the data. Our approach offers potential advantages in terms of scalability and generalizability, as well as automated feature extraction techniques, thus providing avenues for further research and development in this domain.

TABLE II COMPARATIVE PERFORMANCE OF PROPOSED MODELS VERSUS RELATED WORK

Feature type	Accuracy	Dataset	Model	Study
Spectrogram	85.5%	Deep voice	ResNet50	This proposed Model
Wavelet coefficient	79.7%	Deep voice	ANN + Wavelet +ICA	This proposed Model
MFCC	93.0%	Fake or real	VGG-16	Hamza et al. [10]
Chroma- STM, RMS (manual)	98.5%	Deep voice	Random Forest	Kumar et al. [11]
Mel- spectrogram	92.0%	Fake or real	TCN	Khochare et al. [12]

Higher raw accuracy has been obtained by other works like Kumar et al. [11] and Hamza et al. [10], but those models mainly depend on manually created features and particular dataset circumstances. By using automated feature extraction (ResNet50 with spectrograms and wavelet ICA), on the other hand, our method improves generalization and lessens the requirement for domain-specific engineering. Additionally, the DEEP-VOICE dataset, which contains intricate, real-world deepfake audio produced using sophisticated voice conversion techniques, provided a more difficult and realistic assessment for these proposed models.

Because of this, rather than overfitting a single data format or manipulation type [24], our results are more indicative of real-world scenarios, where models must generalize across several manipulation approaches.

VI. Conclusion And Future Work

In this study, two distinct approaches for audio speech recognition were explored: an Artificial Neural Network (ANN) with Wavelet Transform and deep learning with ResNet50. The study’s comprehensive experimentation and evaluation revealed notable performance differences between the two models. The ANN with Wavelet Transform demonstrated respectable accuracy and precision, achieving an average accuracy across folds of 80%. However, the deep learning approach using ResNet50 surpassed the ANN model, exhibiting higher accuracy, precision, recall, and F1 score. ResNet50 exhibited an average accuracy across folds of 86%, demonstrating its exceptional ability to effectively process complex audio data. This represents a notable improvement in performance and robustness, with a nearly 6% increase observed compared to previous approaches.

Our comparison highlights the effectiveness of deep learning techniques, particularly ResNet50, for audio speech recognition tasks. Leveraging transfer learning and pre-trained models like ResNet50 proved highly successful in extracting high-level features from audio spectrograms, resulting in superior classification accuracy.

Despite the encouraging results, this study has several limitations. First, the models were evaluated on a single dataset (DEEP-VOICE), which may limit generalizability across other datasets with different characteristics or languages. Second, although ResNet50 performed well, its computational cost may hinder deployment in real-time or embedded systems. Finally, the reliance on spectrogram images introduces a preprocessing overhead that may not suit all applications. Future work will explore broader dataset diversity and model optimization for real-time deployment.

In conclusion, the findings emphasize the significance of utilizing advanced deep learning architectures for audio speech recognition tasks, offering enhanced performance and scalability compared to traditional machine learning approaches. Future research endeavors may focus on further optimization techniques and larger datasets to continue advancing the accuracy and efficiency of audio speech recognition systems.

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, 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, 07 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 ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2577–2581, IEEE, 2019.

[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 2020 2nd International conference on advancements in computing (ICAC), vol. 1, pp. 192–197, IEEE, 2020.

[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 deepkakes," in 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1449–1454, IEEE, 2024.

[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.

[13] birdy654, "Deep voice deepfake voice recognition."https://www.kaggle.com/datasets/birdy654/deep-voice-deepfake-voice- recognition, 2022. Accessed on May 5, 2024.

[14] M. Shanthamallappa, K. Puttegowda, N. K. Hullahalli Nannappa, and S. K. Vasudeva Rao, "Robust automatic speech recognition using wavelet-based adaptive wavelet thresholding: A review," SN Computer Science, vol. 5, no. 2, p. 248, 2024.

[15] S. Basak, H. Agrawal, S. Jena, S. Gite, M. Bachute, B. Pradhan, and M. Assiri, "Challenges and limitations in speech recognition technology: A critical review of speech signal processing algorithms, tools and systems.," CMES-Computer Modeling in Engineering & Sciences, vol. 135, no. 2, 2023.

[16] H. Hindarto, A. Muntasa, and S. Sumarno, "Feature extraction electroencephalogram (eeg) using wavelet transform for cursor movement," in IOP Conference Series: Materials Science and Engineering, vol. 434, p. 012261, IOP Publishing, 2018.

[17] "Ica and iva bounded multivariate generalized gaussian mixture based hidden markov models," Engineering Applications of Artificial Intelligence, vol. 123, p. 106345, 2023.

[18] N. Trivedi, V. Kumar, S. Sing, S. Ahuja, and R. Chadha, "Speech recognition by wavelet analysis," International Journal of Computer Applications, vol. 15, no. 8, pp. 27–32, 2011.

[19] S. Reza, M. C. Ferreira, J. J. Machado, and J. M. R. Tavares, "A customized residual neural network and bi-directional gated recurrent unit-based automatic speech recognition model," Expert Systems with Applications, vol. 215, p. 119293, 2023.

[20] H. Fazlic´a, A. Abd Almisrea, and N. M. Tahirb, "Deep learning-based audio-visual speech recognition for bosnian digits," Jurnal Kejuruteraan, vol. 36, no. 1, pp. 147–154, 2024.

[21] Z. Jiang Y. Wang, L.Zhang, and H.Wang, "Deepfake Audio Detection Using Dual-Branch CNN with Attention Mechanism," IEEE Transactions on Information Forensics and Security, vol. 19, pp. 512–525, 2024.

[22] M. Singh, R. Srivastava, and V. Sharma, "Wavelet-Based Spectrogram Feature Extraction for Voice Spoofing Detection,"

[23] A. Chen and B. Liu, "Transfer Learning for Audio Deepfake Detection Using Pre-trained Speech Embeddings,"

[24] N. Akhtar, S. Verma, and M. Hossain, "A Review of Challenges and Trends in Deepfake Audio Detection," Digital Signal Processing, vol. 141, p. 103798, 2024.

[25] T. Aoki, K. Oura, and H. Mizuno,"Robust Audio Deepfake Detection Using Phase-Based Features and LSTM,"ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3214–3218.