ABC-CapsNet: Attention based Cascaded Capsule Network for Audio Deepfake Detection

Taiba Majid Wani
Sapienza University of Rome,
Italy
majid@diag.uniroma1.it

Reeva Gulzar
Sapienza University of Rome,
Italy
gulzar.1958711@studenti.uniroma1.it

Irene Amerini
Sapienza University of Rome,
Italy
amerini@diag.uniroma1.it

Abstract

In response to the escalating challenge of audio deepfake detection, this study introduces ABC-CapsNet (Attention-Based Cascaded Capsule Network), a novel architecture that merges the perceptual strengths of Mel spectrograms with the robust feature extraction capabilities of VGG18, enhanced by a strategically placed attention mechanism. This architecture pioneers the use of cascaded capsule networks to delve deeper into complex audio data patterns, setting a new standard in the precision of identifying manipulated audio content. Distinctively, ABC-CapsNet not only addresses the inherent limitations found in traditional CNN models but also showcases remarkable effectiveness across diverse datasets. The proposed method achieved an equal error rate EER of 0.06% on the ASVspoof2019 dataset and an EER of 0.04% on the FoR dataset, underscoring the superior accuracy and reliability of the proposed system in combating the sophisticated threat of audio deepfakes.

1. Introduction

Automatic speaker verification (ASV) systems are crucial in speech processing for verifying a speaker's identity through their voice characteristics. These systems aim to confirm the authenticity of a speaker's utterance. Despite significant advancements in ASV technology, the emergence of sophisticated voice conversion (VC) and text-to-speech (TTS) techniques has introduced new vulnerabilities, making ASV systems prone to deepfake attacks [11]. Deepfake audio poses a considerable threat to privacy, social security, and authenticity. While significant advancements have been made in detecting deepfake videos, the challenges posed by audio spoofing and malicious deepfakes require the creation of specialized models [1]. The field of pure audio-based deepfake detection is less explored than image and video-based approaches, which utilize both audio and spatio-temporal information from videos to train deep learning models [4]. However, the importance of classifiers that focus exclusively on audio for detection is crucial, emphasizing the need for research in this area to enhance the security of ASV systems and address the threats posed by audio deepfakes [21].

The ASVspoof2015 dataset [23] marked the initiation of concerted efforts in the research of automatic speaker verification spoofing and countermeasures, managing to reduce the equal error rate (EER) by up to 1.5%. Certain types of spoofing attacks were found to have a high EER of about 50%, but rates for novel, unseen attacks could be approximately five times higher. The subsequent ASVspoof2017 [6] focused on enhancing the detection of replay spoofing attacks, achieving an EER of 6.73% and showing that Instantaneous Frequency Cosine Coefficients (IFCC) significantly improved the effectiveness of countermeasures. Then, ASVspoof2019 [25] concentrated more on strengthening countermeasures against spoofing in the context of automatic speaker verification, particularly in identifying spoofed audio. Moreover, deep learning architectures like convolutional neural networks (CNNs) have been extensively applied in the field of audio deepfake detection, leveraging their proficiency in handling spectrogram-based analyses to identify audio deepfakes [22, 24]. Beyond the standard convolutional neural network framework, a spectrum of CNN variants including Light Convolutional Neural Networks (LCNN), Temporal Convolutional Networks (TCN), and Spatial Transformer Networks (STN) have been em-

ployed in the nuanced task of detecting audio deepfakes [5, 27]. These advanced models extend the capabilities of traditional CNNs, using their specialized structures to enhance feature extraction from audio signals. Additionally, the strategy of transfer learning has been adopted to further refine the detection process, taking advantage of pre-learned patterns from vast datasets to improve the model's performance in the detection of audio deepfakes [22].

Despite their strengths, CNN-based models face limitations due to their inherent nature. The pooling operations often employed in CNNs can lead to a loss of temporal information, which is crucial in audio analysis. Such information loss may necessitate a vast amount of training data and can increase the susceptibility of the model to adversarial attacks [17, 18]. These adversarial vulnerabilities and the time-intensive training processes present substantial challenges in employing CNNs for audio deepfake detection. Moreover, CNNs often fail to recognize the variations in position, texture, and other deformations within an image, which is vital for accurately classifying manipulated content. This inability is due to their invariance property as they do not maintain spatial hierarchies between the high-level and low-level features, leading to suboptimal recognition capabilities [13].

In response to these limitations, the introduction of Capsule Networks (CapsNets) has been proposed as a more capable alternative. CapsNets, conceptualized by Hinton et al. [3], offer a paradigm shift by processing information in vector form rather than the scalars used in CNNs. This architectural choice allows capsules to maintain the spatial relationships and hierarchies between features, which is crucial for detecting complex manipulations in images and potentially in audio spectrograms. Unlike CNNs, CapsNets are designed to be equivariant, i.e., they can recognize and adjust to changes in the input data, such as rotations and tilts, without losing the integrity of the detection process [15]. Each capsule within a CapsNet is a collection of neurons that identifies and processes a specific feature of the input data, encapsulating both the probability of the feature's presence and its instantiation parameters [12]. This duality provides the network with the ability to recognize an entity by first understanding its constituent parts, ensuring that the detection is not just invariant but also equivariant to input transformations. CapsNets hold promise for a more nuanced approach to audio deepfake detection, where recognizing the subtle deformations and inconsistencies in synthesized audio can be critical . By preserving the intricate patterns and temporal dynamics within the spectrogram data, CapsNets aim to outperform CNNs in the accuracy and reliability of deepfake identification, setting a new standard for the field [7].

Now, in this study, we introduce the novel ABC-CapsNet (Attention-Based Cascaded Capsule Network) architecture, marking a significant advancement in audio deepfake detection. Leveraging the nuanced processing capabilities of Mel spectrograms and the robust feature extraction offered by the VGG18 model, our approach emphasizes the critical role of identifying key audio characteristics. Following the feature extraction phase, an attention mechanism is employed to prioritize and amplify the most salient features identified by VGG18. This mechanism plays a pivotal role, acting as a refined filter that sharpens the model's focus, allowing for a more concentrated and effective analysis of potential deepfakes. The data, enriched and focused through this process, is subsequently analyzed by cascaded capsule networks. This design is meticulously crafted to boost detection accuracy through a deeper, more nuanced examination of the complex patterns that characterize audio deepfakes. The strategic cascading of capsule networks affords a thorough analysis of the audio data, leveraging the unique capabilities of capsule technology to identify subtle manipulations that conventional models might miss. The robustness and efficacy of the ABC-CapsNet architecture have been rigorously validated across two extensive audio deepfake datasets, FoR [14] and ASVspoof 2019 [19], showcasing its effectiveness against a wide array of deepfake audio challenges.

1.1. Contributions

The proposed ABC-CapsNet methodology introduces several key contributions to the domain of audio deepfake detection, each playing a critical role in advancing the detection capabilities and understanding of audio forgeries. The three main contributions of this methodology are:

Advanced Feature Extraction: We integrate Mel spectrograms with the VGG18 model to harness both the perceptual accuracy of Mel scales and the deep learning power of VGG18, enhancing the extraction of detailed audio features.

Focused Feature Analysis with Attention Mechanism: An attention mechanism is employed post-feature extraction to highlight and prioritize key features, allowing for more targeted analysis of potential audio manipulations.

Depth Analysis via Cascaded Capsule Networks: We utilize cascaded capsule networks to delve deeper into the structural intricacies of audio data, providing a novel approach to detecting complex patterns typical of manipulated audio.

Together, these contributions embody a comprehensive and sophisticated approach to audio deepfake detection, establishing the ABC-CapsNet methodology as a state-of-the-art solution in the ongoing effort to combat digital audio manipulation.

2. Related Works

The rising prevalence of audio deepfakes challenges voice biometric systems and societal trust, with current audio forensic techniques showing limited success in detecting them. This section reviews recent studies that utilize machine learning (ML) and deep learning (DL) algorithms across various datasets for audio deepfake detection. We specifically focus on the effectiveness of capsule networks, convolutional neural networks, and transfer learning in differentiating genuine from fake audio samples.

A.Luo et. al., [7] proposed capsule network architecture specifically designed with a modified dynamic routing algorithm to improve the generalization of the detection system.The feature extraction module utilized the logarithm of the power spectrum and linear frequency cepstral coefficients (LFCC) to capture information from the speech signal. The proposed capsule network demonstrated superior performance compared to other state-of-the-art methods. For the LA dataset, the capsule network achieved an equal error rate (EER) of 3.19% and a tandem detection cost function (t-DCF) of 0.0982 on the evaluation set.

Q. Ma et. al., [8] proposed ConvNeXt-based neural network (CNBNN) designed by revising the ConvNeXt network architecture to suit audio anti-spoofing tasks. The model incorporated a Res2Net style block and a modified efficient channel attention (MECA) layer to focus on the most informative sub-bands of speech representations and hard-to-classify samples. The model achieved an equal error rate (EER) of 0.64% and a minimum tandem detection cost function (min-tDCF) of 0.0187 on the ASVSpoof 2019 LA evaluation dataset.

T. Mao et. al., [9] proposed deep capsule network model consisting of a convolutional neural network (CNN) architecture followed by a capsule architecture. The CNN extracted hierarchical features from the input cepstrum features, and the capsule network used dynamic routing to classify the input as bona fide or spoofed speech. Experimental results on the ASVspoof 2019 Logical Access (LA) evaluation set demonstrated that the proposed deep capsule network significantly improved the baseline algorithms' tandem detection cost function (t-DCF) and equal error rate (EER) scores by 31% and 37% respectively.

M. Mcuba et.al., [10] employed various deep learning models, including custom architectures and VGG-16, to analyze audio features extracted through MFCC, Mel-spectrum, Chromagram, and spectrogram representations for forensic investigators in distinguishing between synthetic and real voices by evaluating the effectiveness of different deep learning approaches, FG-LCNN and ResNet. The results showed that the VGG-16 architecture performs best for the MFCC image feature with accuracy of 86.906% on Baidu Silicon Valley AI Lab.

A.Hamza et. al., [2] explored various machine learning (ML) and deep learning (DL) algorithms, including support vector machine (SVM), gradient boosting, and the VGG-16 deep learning model, to classify deepfake audio based on the extracted MFCC features. The SVM model achieved accuracies of 98.83% and 97.57% on the for-rece and for-2-sec of FoR dataset, respectively. The gradient boosting model performed well on the for-norm dataset with 92.63% accuracy, while the VGG-16 model achieved 93% accuracy on the for-original dataset.

A. Ustubioglu et. al., [20] proposed CNN architecture with data augmentation to classify Mel spectrogram images into two classes, original and forged. Mel spectrograms were generated from the Arabic Speech Corpus and the TIMIT speech database and achieved an accuracy of 99% and 91% respectively.

R. Yen et.al., [26] utilized a standard 34-layer ResNet with multi-head attention pooling to learn discriminative embeddings for fake audio and spoof detection. The proposed system employed data augmentation techniques, such as noise addition, compression, and frequency conversion, to improve robustness. The classification network consisted of two fully-connected layers and a 2-dimensional softmax layer and achieved an equal error rate (EER) of 10.1% in Track 3.2 of ADD 2022.

Taking inspiration from the state-of-the-art works, this work proposes a novel architecture, ABC-CapsNet (Attention-Based Cascaded Capsule Network), designed to tackle the challenges of audio deepfake detection. By synthesizing the strengths of advanced feature extraction through VGG18 with Mel spectrograms, and integrating an attention mechanism for focused analysis, ABC-CapsNet advances the field beyond the current limitations of traditional models. The cascaded capsule networks further distinguishes this methodology, enabling a deeper and more nuanced exploration of audio data for identifying manipulation.

3. Proposed Methodology

In this research, we propose a novel methodology for detecting audio deepfakes, utilizing ABC-CapsNet (Attention-based Cascaded Capsule Networks) as shown in Fig. 1. Our approach leverages the strengths of VGG18, an attention mechanism, and capsule networks to effectively identify and differentiate between real and fake audio samples. We selected VGG18 for its proven capability in capturing distinctive audio characteristics crucial for deepfake detection. While VGG18 is a suitable choice for our methodology, other models with potential for feature extraction in this context could also be considered. The methodology comprises several pivotal steps explained in this section.

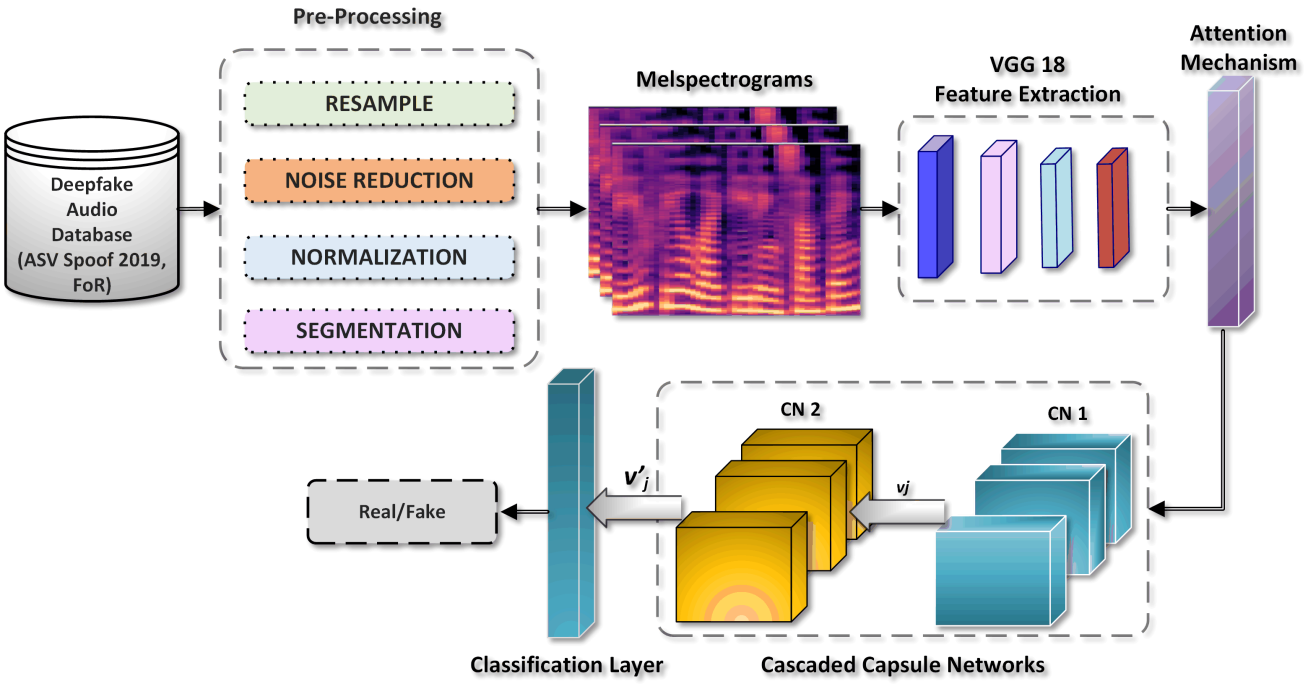


Figure 1. Proposed architecture of ABC-CapsNet for Audio Deepfake Detection; CN1 is Capsule Network 1 and CN2 is Capsule Network 2

3.1. Preprocessing

In the preprocessing stage, we prepare audio samples from the ASVSpoof 2019 and FoR datasets by first resampling them to a uniform 16 kHz to align frequency content and compatibility with feature extraction techniques. We then apply noise reduction algorithms to enhance clarity by removing background noise, followed by normalizing the amplitude to a consistent range of -1 to 1 for uniform volume levels across samples. Finally, segments of silence are removed to focus the analysis on relevant audio content. Post-preprocessing, Mel spectrograms are generated using the Mel scale to better represent the human ear's response to frequenciesThe transformation from frequency to Mel scale can be described using the formula:

$$m = 2595 \log_{10} \left(1 + \frac{f}{700}\right) \tag{1}$$

where m is the Mel scale frequency and f is the frequency in Hertz. This makes Mel spectrograms particularly useful for audio analysis tasks, as they provide a more perceptually relevant representation of the sound. By using Mel spectrograms, we can capture the essential characteristics of the audio, such as timbre and pitch, which are crucial for distinguishing between real and fake audio samples in our deepfake detection methodology. In this study, we generated Mel spectrograms of size 224 × 224 × 3, using the Hanning window with a size of 2048 and a hop length of 512. The number of Mel filter banks used was 224. The generation of Mel spectrograms serves as a critical step in our approach, as it prepares the audio data for the subsequent feature extraction and analysis stages.

3.2. Feature Extraction using VGG18

We employ VGG18 for feature extraction from audio Mel spectrograms. Originally designed for image recognition, VGG18 features 16 convolutional layers and 3 fully connected layers, making it adept at capturing complex patterns. This depth allows it to effectively extract rich features from audio data through successive convolutional operations, pooling, and non-linear activations.

$$F_i(x) = \sigma(W_i * x + b_i) \tag{2}$$

where W_i and b_i represent the weights and biases of the convolutional filters in layer i,x denotes the input to the layer, and $F_i(x)$.

The selection of VGG18 for feature extraction from audio Mel spectrograms is motivated by its proven effectiveness in capturing a wide spectrum of features in image data. By treating spectrograms as images, we leverage VGG18's robustness in identifying patterns that are indicative of the authenticity of audio signals. The extracted features encompass both local and global characteristics of the Mel spectro-

gram, providing a comprehensive representation that is essential for accurately detecting deepfake audio. The application of VGG18 introduces the concept of transfer learning to our methodology and allows us to utilize the rich feature set learned by VGG18 from its pre-training on extensive image datasets, thereby reducing the need for large-scale audio-specific training data.

3.3. Attention Mechanism

Following the extraction of features from the Mel spectrograms, the deployment of an attention mechanism plays a pivotal role in refining the feature set for the detection task. The attention mechanism operates on the principle of selectively focusing on parts of the input that are most pertinent to the task at hand, thereby enhancing the model's sensitivity to features that are indicative of deepfake audio.

The mathematical formulation of the attention mechanism can be represented as follows:

Let F denote the set of features extracted by VGG18, where $F = \{f_1, f_2, \ldots, f_n\}$ and each f_i is a feature vector. The attention mechanism assigns a weight w_i to each feature vector f_i, with the weights being determined by a trainable attention layer. The output of the attention mechanism, F', is a weighted sum of the feature vectors, given by:

$$F' = \sum_{i=1}^{n} w_i \cdot f_i \tag{3}$$

The weights w_i are computed using a softmax function over the scores assigned to each feature vector by the attention layer, as follows:

$$w_i = \frac{e^{s(f_i)}}{\sum_{j=1}^{n} e^{s(f_j)}} \tag{4}$$

where $s(f_i)$ is the score assigned to feature vector f_i by the attention layer, which is typically implemented as a fully connected layer with a single output unit. The softmax function ensures that the weights sum up to 1, allowing them to be interpreted as probabilities that indicate the importance of each feature vector in the context of the detection task.

The incorporation of the attention mechanism significantly improves the model's ability to identify and prioritize features that are most indicative of audio authenticity. By dynamically adjusting the focus on different parts of the audio Mel spectrogram, the attention mechanism allows the model to adapt to the varying characteristics of real and fake audio samples.

3.4. Cascaded Capsule Network Architecture

The architecture consists of two main capsule networks connected in series, where the output of the first serves as the input to the second. The cascading structure allows for a refined processing pipeline that accentuates pertinent features and temporal relationships inherent in audio signals, which are crucial for classifying the authenticity of the content. The initial capsule network in the cascade focuses on extracting fundamental patterns and relationships, laying the groundwork for the subsequent network to delve into more complex and subtle features indicative of deepfakes. Through this progressive refinement, the architecture embodies a deep understanding of the audio data, setting the stage for a robust detection mechanism.

3.4.1 Capsule Network 1 (CN1)

Capsule Network 1, being the foundational component of the proposed cascading architecture, leverages the innovative concept of capsules—groups of neurons that activate for various properties of a particular entity type, thus maintaining the spatial and feature hierarchy. Unlike traditional neural layers that scalarize feature presence, capsules retain multidimensional information that represents various properties and orientation of the features. This design allows CN1 to capture and preserve complex phenomena within the audio data, providing a robust platform for higher-order feature analysis by subsequent layers and networks. The CN1 consists of following layers.

Input Layer: The input layer of CN1 receives a feature vector, denoted as u_i, which is the output from the preceding attention layer. This layer encodes and emphasizes the most salient features within the Mel spectrogram data, such as specific frequency bands and temporal characteristics that are essential for differentiating between real and fake audio.

Convolutional Layers: Following the input layer, there are two convolutional layers. These layers apply a series of learnable filters to the input data, capturing local feature patterns such as edges or texture in the Mel spectrogram that may signal possible manipulation. This is done prior to the feature vector being fed into the primary capsules.

Primary Capsule Layer: This is the first layer of the capsule hierarchy in CN1. Each primary capsule contains a small group of neurons that specialize in identifying specific features within the received feature map from the convolutional layer. These capsules output a set of prediction vectors, which are lower-dimensional representations of the input data, encapsulating the probability of feature presence and their spatial orientations. Prediction vectors are computed for each capsule i in layer l using transformation matrices W_{ji}:

$$u_{ji} = squash(W_{ji} u_i) \tag{5}$$

- $squash$ function is a non-linear function that shrinks the length of the vector to be between 0 and 1, which can be

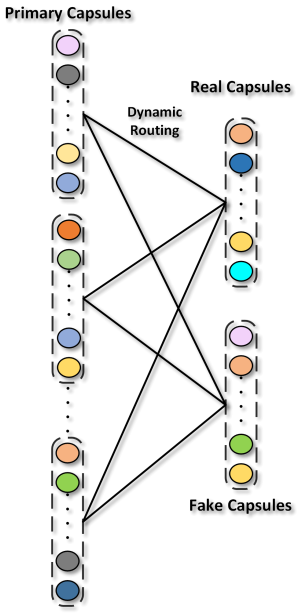


Figure 2. Capsule layer

given as:

$$squash(v) = \frac{||v||^2}{1 + ||v||^2} \frac{v}{||v||} \tag{6}$$

Each primary capsule represents a pose as a vector within its specific area as illustrated in Fig. 2.

Dynamic Routing: This is not a layer but a process between the primary capsules and the higher-level capsules. Dynamic routing is an algorithm that allows capsules to communicate and send information to higher-level capsules [16]. It determines the connections between capsules based on the current input data, ensuring that the network focuses on the spatial hierarchies in the data and involves computing the coupling coefficients c_{ij} between capsules. The dynamic algorithm is shown in Algorithm 1.

Algorithm 1 Dynamic Routing Algorithm

Initialization: For all capsule pairs i, j: $b_{ij} \leftarrow 0$
for r routing iterations **do**
  For all capsule pairs i, j: compute coupling coefficients c_{ij} via softmax:
  $c_{ij} \leftarrow \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})}$
  For each capsule j in layer $l+1$: compute total input s_j:
  $s_j \leftarrow \sum_i c_{ij} u_{ji}$
  Squash to get the output vector of capsule j:
  $v_j \leftarrow \text{squash}(s_j)$
  Update the log probabilities for the next iteration:
  $b_{ij} \leftarrow b_{ij} + u_{ji} \cdot v_j$
end for

Digit Capsule Layer: A digit capsules layer follows the primary capsules. This layer represents more complex combinations of the features detected by the primary capsules. For audio deepfake detection, this could include detecting irregularities in frequency patterns that do not correspond to natural human speech variations.

Output: The output from CN1 consists of a set of activity vectors, each from a capsule in the last capsule layer of CN1. These vectors encapsulate the instantiation parameters of the features detected in the audio data. The length of each vector indicates the probability that a certain feature is present in the input data, while its orientation in the vector space represents the instantiation parameters of that feature. The output of CN1 then serves as the input for CN2, where the process of feature abstraction and representation continues at an even higher level.

3.4.2 Capsule Network 2 (CN2)

Capsule Network 2 (CN2) functions as the subsequent stage in the cascaded architecture, where the primary goal is to process and interpret the complex features relayed by Capsule Network 1 (CN1). CN2 is specifically designed to take the output vectors v_j from CN1 and subject them to a secondary phase of transformation and dynamic routing. This is achieved through a secondary capsule layer that fine-tunes the feature detection process.

Input Layer: The input to CN2 consists of the output vectors v_j from CN1. These vectors encapsulate high-level feature information and the likelihood of their presence as determined by the previous network's dynamic routing.

Secondary Capsule Layer: Here, the dynamic routing algorithm is executed again. However, the prediction vectors u'_{ji} created in CN2 are derived from the outputs v_j received from CN1, which means they are based on already processed and interpreted feature data. This layer employs transformation matrices similar to those in CN1 to generate prediction vectors for each higher-level capsule. The dynamic routing process then iteratively updates the coupling coefficients between the capsules based on the "agreement" between their predictions, refining the representation of the data.

Dynamic Routing in CN2: The dynamic routing mechanism in CN2 iteratively refines the coupling coefficients between the secondary capsules, with the objective of identifying and emphasizing complex, high-level features that are most indicative of audio authenticity or tampering.

Output Layer: The output of CN2, denoted by v'_j comprises the final activity vectors of the secondary capsules. These vectors represent the network's distilled knowledge and conviction regarding the presence of intricate patterns and relationships within the audio data, which are potential markers of deepfake manipulation.

3.5. Classification and Marginal Loss

The culmination of the capsule network's processing is the classification stage, where the authenticity of audio samples is determined. This stage utilizes the lengths and orientations of the digit capsules' output vectors to classify audio samples as real or fake. The geometric properties of these vectors serve as a powerful mechanism for distinguishing between real and fake audio content.

Margin Loss Function: The margin loss function is critical for guiding the network towards accurate classification. It is formulated to penalize incorrect classifications while providing a balance between sensitivity and specificity [13]. This balance is crucial for audio deepfake detection, where the cost of false positives and negatives can be significant. The mathematical expression for margin loss is

$$L_k = T_k \max(0, m^+ - ||v_k||)^2 + \lambda(1 - T_k) \max(0, ||v_k|| - m^-)^2 \quad (7)$$

where T_k is 1 if the class k is present and 0 otherwise, m^+ and m^- are the margins for correct and incorrect classifications, respectively, and λ is a down-weighting term for the absence of a class.

4. Experimental Setup and Results

4.1. Datasets

ASVspoof 2019: The ASVspoof 2019 dataset addresses three primary forms of spoofing attacks—replay, speech synthesis, and voice conversion, within a unified competition framework. This dataset is divided into two distinct scenarios: logical access (LA) and physical access (PA), each offering a unique dataset tailored to its respective access control challenges. In our research, we have concentrated on the LA scenario. The LA subset encompasses a diverse range of 17 spoofing attacks, from A01 to A19, blending various speech synthesis and voice conversion techniques. It includes high-quality speech synthesis with WaveNet (A01), the WORLD vocoder for scenarios with limited WaveNet data (A02), and the Merlin toolkit for ease of TTS system construction (A03). Waveform concatenation via MaryTTS (A04) emphasizes natural speech, while A05 and A06 introduce neural network-based VC, leveraging a Variational AutoEncoder and a transfer-function approach, respectively. The remaining attacks (A07 to A19) explore a diverse blend of methods, including Generative Adversarial Networks and neural source-filter models, targeting various aspects of speech synthesis and conversion to challenge automatic speaker verification systems effectively.

Fake or Real dataset (FoR): FoR dataset contains over 198,000 utterances from state-of-the-art TTS algorithms and real speech. It's published in four versions: for-original, for-norm, for-2seconds, and for-rerecorded to simulate real-world attacks.

for-original: Contains raw audio files from TTS systems and real human speech, providing a baseline for comparison.

for-norm: Features audio files that have been normalized to ensure consistent volume levels across the dataset.

for-2seconds: Includes audio clips shortened to 2 seconds to focus on short-duration speech analysis.

for-rerecorded: Consists of the original dataset audio played and recorded in a real environment to simulate real-world conditions and background noises.

In our research, we have employed each version of the dataset both individually and by combining them into a unified, extensive dataset for analysis.

4.2. Experimental Setup

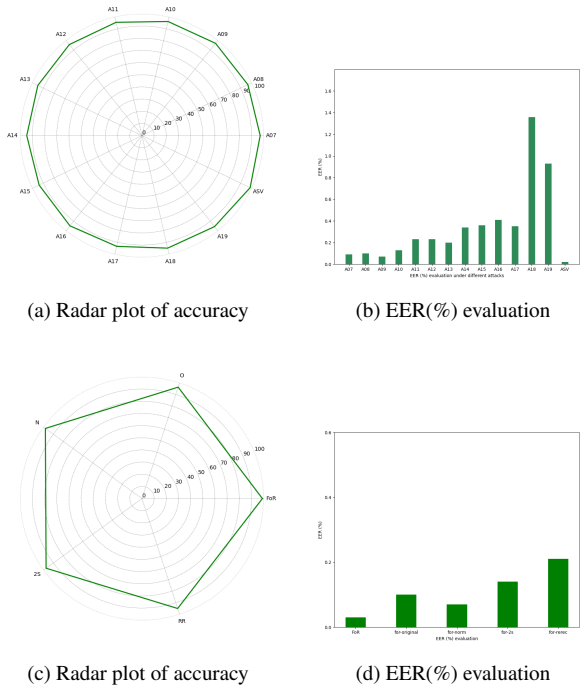
In our study on audio deepfake detection, the experimental setup for the ABC-CapsNet integrates VGG18 for initial feature extraction and an attention mechanism for precise feature refinement, followed by a novel implementation of two cascaded capsule networks for detailed analysis. The training was conducted with a batch size of 32 across 100 epochs, utilizing the Adam optimizer for its adaptability and efficiency with a learning rate of 0.0001 for both the feature extraction phase and the first capsule network. The cross-entropy loss function was chosen for its effectiveness in binary classification tasks, critical for distinguishing between real and fake audio samples. Evaluation metrics focused on accuracy and the Equal Error Rate (EER), providing a dual perspective on the model's performance by assessing both its precision and its ability to balance false positives and negatives.

4.3. Results on ASVspoof 2019 (LA)

We conducted a series of experiments on the LA senario of ASVspoof 2019 dataset, testing individually on attacks A07 to A19 and on the dataset as a whole. The accuracies obtained were high, with the lowest being 95.5% for attack A17 and the highest reaching 98.1% for the entire LA senario , as depicted Fig. 3a. Such high accuracy demonstrates the model's proficiency in correctly classifying audio samples across a range of attack conditions.

The EER(%) values present a more nuanced understanding of the model's performance as depicted in Fig. 3b. EER(%) spans from as low as 0.06% for the full ASVspoof2019 dataset to a peak of 1.36% for attack A18. The majority of individual spoofing attacks (A07 to A17) maintain an EER below 0.41%, demonstrating the model's robustness. However, the EER does rise significantly for

attacks A18 and A19, with A18, in particular, reaching 1.36%. This suggests that the features presented by attack A18 are more challenging for the model to discern, pointing to potential areas for further model refinement and training.



(a) Radar plot of accuracy

(b) EER(%) evaluation

(c) Radar plot of accuracy

(d) EER(%) evaluation

Figure 3. Evaluation of ASVspoof2019 (a) and (b) and FoR dataset (c) and (d). (where O is for-original, N is for-norm, 2s is for-2seconds and RR is for-rerecording)

4.4. Results on FoR dataset

We carried out a comprehensive set of experiments across all four versions of the FoR dataset, as well as on the dataset as a whole, to evaluate the ABC-CapsNet architecture's proficiency in audio deepfake detection. The results, illustrated in Figs. 3c and 3d, demonstrate the system's proficiency, with an EER of only 0.04% and an accuracy of 99% on For dataset. Analysis of the individual dataset versions revealed consistently low EERs: 'for-original' at 0.09%, 'for-norm' at 0.07%, and 'for-2s' at 0.13%. The 'for-rerec' version, which presents a more complex challenge, recorded a higher EER of 2.2%. Nonetheless, the accuracy remained robust across all scenarios, notably 97.3% for 'for-original', 98.8% for 'for-norm', 98.0% for 'for-2seconds', and 96.3% for 'for-rerecording'.

Fig. 3, illustrates the robust capability of ABC-CapsNet in accurately identifying and classifying deepfake audio content across a spectrum of manipulations, solidifying its standing as a state-of-the-art system in this critical aspect of digital security.

Study	Dataset	Architecture	EER (%)
[7]	ASVspoof2019	Capsule Network	1.07
[9]	ASVspoof2019	MFCC Capsule	9.21
[9]	ASVspoof2019	CQCC Capsule	5.09
Our Method	FoR	ABC-CapsNet	**0.04**
Our Method	ASVspoof2019	ABC-CapsNet	**0.06**

Table 1. Comparison with state-of-the-art methods

4.5. Benchmarking

In the benchmarking landscape of audio deepfake detection, our work pioneers the use of cascaded capsule networks, representing a novel approach in the field. Prior studies exploring the use of capsule networks for this purpose are sparse. For instance, [7]. implemented a capsule network-based model and achieved an EER of 1.07%, while [9] experimented with two distinct models: MFCC-capsule and CQCC-capsule, which yielded EERs of 9.21% and 5.09%, respectively. In contrast, our ABC-CapsNet architecture has set a new benchmark, obtaining an EER of 0.04% on the FoR dataset and 0.06% on the ASVspoof 2019 dataset. This substantial improvement underscores the efficacy of cascaded capsule networks in classifying audio authenticity, positioning our model at the forefront of current audio deepfake detection technologies.

5. Conclusion

In this study, we proposed ABC-CapsNet, a novel architecture for audio deepfake detection, integrating Mel spectrograms, VGG18 feature extraction, and attention mechanisms with cascaded capsule networks. This comprehensive approach demonstrated superior efficacy across diverse datasets, including ASVspoof 2019 and FoR, achieving unprecedented low EERs. In future, we intend to extend our model to accommodate a broader array of audio manipulation techniques and explore adaptive mechanisms to enhance model efficiency and reduce computational demands, ensuring broader applicability and scalability.

Acknowledgments

This study has been partially supported by SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union – NextGenerationEU and Sapienza University of Rome project 2022–2024 "EV2" (003 009 22).

References

[1] Zaynab Almutairi and Hebah Elgibreen. A review of modern audio deepfake detection methods: Challenges and future directions. *Algorithms*, 15(5):155, 2022.

[2] Ameer Hamza, Abdul Rehman Rehman Javed, Farkhund Iqbal, Natalia Kryvinska, Ahmad S Almadhor, Zunera Jalil, and Rouba Borghol. Deepfake audio detection via mfcc features using machine learning. *IEEE Access*, 10:134018–134028, 2022.

[3] Geoffrey E Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with em routing. In *International conference on learning representations*, 2018.

[4] Zahra Khanjani, Gabrielle Watson, and Vandana P Janeja. Audio deepfakes: A survey. *Frontiers in Big Data*, 5: 1001063, 2023.

[5] Janavi Khochare, Chaitali Joshi, Bakul Yenarkar, Shraddha Suratkar, and Faruk Kazi. A deep learning framework for audio deepfake detection. *Arabian Journal for Science and Engineering*, pages 1–12, 2021.

[6] Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. 2017.

[7] Anwei Luo, Enlei Li, Yongliang Liu, Xiangui Kang, and Z Jane Wang. A capsule network based approach for detection of audio spoofing attacks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6359–6363. IEEE, 2021.

[8] Qiaowei Ma, Jinghui Zhong, Yitao Yang, Weiheng Liu, Ying Gao, and Wing WY Ng. Convnext based neural network for audio anti-spoofing. *arXiv preprint arXiv:2209.06434*, 2022.

[9] Terui Mao, Diqun Yan, Yongkang Gong, and Randing Wang. Identification of synthetic spoofed speech with deep capsule network. In *International Conference on Frontiers in Cyber Security*, pages 257–265. Springer, 2021.

[10] Mvelo Mcuba, Avinash Singh, Richard Adeyemi Ikuesan, and Hein Venter. The effect of deep learning methods on deepfake audio detection for digital investigation. *Procedia Computer Science*, 219:211–219, 2023.

[11] Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263*, 2022.

[12] Chenbin Pan and Senem Velipasalar. Pt-capsnet: A novel prediction-tuning capsule network suitable for deeper architectures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11996–12005, 2021.

[13] Mensah Kwabena Patrick, Adebayo Felix Adekoya, Ayidzoe Abra Mighty, and Baagyire Y Edward. Capsule networks–a survey. *Journal of King Saud University-computer and information sciences*, 34(1):1295–1310, 2022.

[14] Ricardo Reimao and Vassilios Tzerpos. For: A dataset for synthetic speech detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–10. IEEE, 2019.

[15] Fabio De Sousa Ribeiro, Kevin Duarte, Miles Everett, Georgios Leontidis, and Mubarak Shah. Learning with capsules: A survey. *arXiv preprint arXiv:2206.02664*, 2022.

[16] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017.

[17] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. Attacking convolutional neural network using differential evolution. *IPSJ Transactions on Computer Vision and Applications*, 11:1–16, 2019.

[18] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

[19] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*, 2019.

[20] Arda Ustubioglu, Beste Ustubioglu, and Guzin Ulutas. Mel spectrogram-based audio forgery detection using cnn. *Signal, Image and Video Processing*, 17(5):2211–2219, 2023.

[21] Kishan Vyas, Preksha Pareek, Ruchi Jayaswal, and Shruti Patil. Analysing the landscape of deep fake detection: A survey. *International Journal of Intelligent Systems and Applications in Engineering*, 12(11s):40–55, 2024.

[22] Taiba Majid Wani and Irene Amerini. Deepfakes audio detection leveraging audio spectrogram and convolutional neural networks. In *International Conference on Image Analysis and Processing*, pages 156–167. Springer, 2023.

[23] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, and Junichi Yamagishi. Asvspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *Training*, 10(15):3750, 2014.

[24] Jun Xue, Cunhang Fan, Zhao Lv, Jianhua Tao, Jiangyan Yi, Chengshi Zheng, Zhengqi Wen, Minmin Yuan, and Shegang Shao. Audio deepfake detection based on a combination of f0 information and real plus imaginary spectrogram features. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, pages 19–26, 2022.

[25] Junichi Yamagishi, Massimiliano Todisco, Md Sahidullah, Héctor Delgado, Xin Wang, Nicholas Evans, Tomi Kinnunen, K Aik Lee, Ville Vestman, and Andreas Nautsch. Asvspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *ASV Spoof*, 13, 2019.

[26] Rui Yan, Cheng Wen, Shuran Zhou, Tingwei Guo, Wei Zou, and Xiangang Li. Audio deepfake detection system with neural stitching for add 2022. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9226–9230. IEEE, 2022.

[27] Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi, Jose Patino, and Nicholas Evans. An initial investigation for detecting partially spoofed audio. *arXiv preprint arXiv:2104.02518*, 2021.