EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY Open Access

Automatic dysarthria detection and severity level assessment using CWT-layered CNN model

Shaik Sajiha[1], Kodali Radha[1,2], Dhulipalla Venkata Rao[1], Nammi Sneha[1], Suryanarayana Gunnam[1] and Durga Prasad Bavirisetti[3*]

Abstract

Dysarthria is a speech disorder that affects the ability to communicate due to articulation difficulties. This research proposes a novel method for automatic dysarthria detection (ADD) and automatic dysarthria severity level assessment (ADSLA) by using a variable continuous wavelet transform (CWT) layered convolutional neural network (CNN) model. To determine their efficiency, the proposed model is assessed using two distinct corpora, TORGO and UA-Speech, comprising both dysarthria patients and healthy subject speech signals. The research study explores the effectiveness of CWT-layered CNN models that employ different wavelets such as Amor, Morse, and Bump. The study aims to analyze the models' performance without the need for feature extraction, which could provide deeper insights into the effectiveness of the models in processing complex data. Also, raw waveform modeling preserves the original signal's integrity and nuance, making it ideal for applications like speech recognition, signal processing, and image processing. Extensive analysis and experimentation have revealed that the Amor wavelet surpasses the Morse and Bump wavelets in accurately representing signal characteristics. The Amor wavelet outperforms the others in terms of signal reconstruction fidelity, noise suppression capabilities, and feature extraction accuracy. The proposed CWT-layered CNN model emphasizes the importance of selecting the appropriate wavelet for signal-processing tasks. The Amor wavelet is a reliable and precise choice for applications. The UA-Speech dataset is crucial for more accurate dysarthria classification. Advanced deep learning techniques can simplify early intervention measures and expedite the diagnosis process.

Keywords Dysarthria, Severity assessment, TORGO, UA-Speech, CWT-layer, Amor wavelet, Morse wavelet, Bump wavelet

1 Introduction

Speech-based effective communication is essential to human contact and has a significant impact on an individual's quality of life. Nonetheless, some people experience difficulties producing speech due to neurological problems, which can present as motor speech difficulties such as dysarthria [1–3]. The motor-speech system is characterized by weakness, paralysis, or coordination deficiencies in dysarthria, which reduces the lucidity, naturalness, and effectiveness of verbal communication [4, 5]. The creation of ADD and ADSLA using raw speech data has garnered significant attention in recent decades, with the potential to transform dysarthria diagnostic and treatment methods [6]. Raw waveform models offer a

*Correspondence:
Durga Prasad Bavirisetti
durga.bavirisetti@ntnu.no
[1] Department of Electronics & Communication Engineering, Velagapudi Ramakrishna Siddhartha Engineering College, Kanuru, Vijayawada 520007, Andhra Pradesh, India
[2] Division of Pediatric Neurology, Department of Pediatrics, University of Tennessee Health Science Center, Memphis 38163, TN, USA
[3] Department of Computer Science, Norwegian University of Science and Technology, 7034 Trondheim, Norway

potential way to categorize dysarthric speech. They provide a complete view of the speech signal, making it easier to distinguish between normal and dysarthric speech patterns [7]. Raw waveform models avoid the requirement for manually extracting features, which speeds up the procedure and reduces the possibility of biased feature selection. This improves objectivity and saves time [8]. These models also perform well with end-to-end techniques like CNN models, which improve classification accuracy by directly learning complex representations based on raw waveform data [9, 10]. In contrast to traditional methods that depend on intricate feature extraction procedures [11–14], raw waveform models examine the speech signal directly, making them easier to apply and comprehend [15]. There are several benefits to incorporating CWT-layer in CNN models, one of which is the ability to capture local and global aspects of the signals that are input at various wavelets such as Amor, Morse, and Bump. Analyzing signals in both the temporal and frequency domains at the same time makes CWT a useful tool in signal processing, especially when it comes to irregular signals like speech. The CNN may extract pertinent characteristics at various resolutions from the input speech data by applying CWT, which allows it to pick up both small details and larger patterns in the signal. In the CNN architecture, one typical method is to arrange CWT-layer next to conventional convolutional layers. As a result, the network can acquire hierarchical visualizations of the voice signal it receives, gradually extracting increasingly abstract and discriminative properties from the raw waveforms. CWT-layer aids in the dimensionality reduction of the input data while maintaining pertinent information, improving CNN generalization performance and learning efficiency. Improved interpretability of the learned representations is made possible by the incorporation of CWT-layer into CNN designs.

The article's following sections are organized as follows: the research that has been done on ADD and ADSLA is shown in Section 2. CWT-layered CNN model is explained in Section 3. The datasets used and experimental results of ADD and ADSLA are reported in Section 4, demonstrating great accuracy in both tasks. The main conclusions of the recommended work are finally summarized in Section 5, along with their consequences for additional field investigation.

2 Motivation and relevant work

In speech recognition and healthcare, ADD and ADSLA utilizing raw speech presents substantial obstacles [16]. Dysarthria is a neurological condition marked by defective articulation resulting from muscle weakness or paralysis. It has been linked to many neurological illnesses, including cerebral palsy, Parkinson's disease, and stroke [17]. For these disorders to be properly diagnosed and treated, dysarthria must be accurately detected and assessed [18–20]. However, a variety of approaches have been used in recent research to identify dysarthria and evaluate the degree of it in two reference datasets: UA-Speech and TORGO.

Raw waveform modeling offers a potential approach for ADD and ADSLA [21]. Upon analyzing the TORGO corpus, Millet et al. [9] used raw waveforms and applied long-short-term memory networks (LSTM), and an attention model, obtaining an accuracy of 75.63% for dysarthria detection. Similarly, in Hernandez et al. [22] using rhythm and voice quality or prosody, random forest (RF) achieved an accuracy of 81.50%, while support vector machine (SVM) achieved an accuracy of 82.30% for dysarthria identification. Furthermore, Yue et al. [6] employed raw magnitude-based feature extraction using the cascade convolution model and achieved an accuracy of 88.0% for dysarthria detection and 86.00% for severity assessment. In contrast, Radha et al. [23] used raw waveform modeling using short-time Fourier transform (STFT) layered CNN and achieved an accuracy of 94.62% in dysarthria detection and 90.15% for severity assessment. Switching the attention to the UA-Speech dataset, Narendra and Alku [24] worked on raw glottal flow waveforms, implementing CNN and multi-layered perceptron (MLP), leading to an accuracy of 87.93%. Kachhi et al. [25] used CWT scalograms with CNN, respectively, obtaining an accuracy of 95.17% for severity evaluation and 87.93% for dysarthria detection. Similarly, Joshy and Rajan [26] used mel spectrograms for the squeeze excitation (SE) CNN model and achieved an accuracy of 87.93%. Moreover, Radha et al. [23] employed raw waveform modeling using an STFT-layered CNN model and achieved an accuracy of 99.80% for dysarthria detection and 94.67% for severity assessment.

In contrast to conventional feature-based techniques, raw waveform offers an additional detailed description of an audio signal strength, making it possible to record the precise temporal and spectral properties required for evaluation. Developments in deep learning frameworks for automatic evaluation have been made easier by recent technological advances in clinical settings [27–29]. Researchers have developed novel methods for utilizing sound characteristics to assess the severity of dysarthria [30]. They have investigated innovative paths in dysarthria assessment by categorizing speech into distinct severity levels inside datasets such as the TORGO and the UA-Speech datasets. Using the UA-Speech dataset, research has employed CNN approaches based on raw waveform to differentiate between individuals with dysarthria and healthy individuals [24, 26, 31, 32]. These models achieve accurate categorization by making use of the fine features present in raw speech waveforms. This study aims to improve the identification of dysarthric persons in the UA-Speech datasets by utilizing deep learning techniques. The majority of current techniques in this sector rely on intricate deep-learning algorithms or feature-based models. To provide simple yet efficient models that do not require feature extraction and allow the model to learn directly, the proposed work presents a variable CWT-layered CNN model. The paper makes three primary contributions:

- Proposed a method with a variable CWT-layered CNN model for ADD and ADSLA
- Experimented on different CWT-layer wavelets to improve accuracy
- TORGO and UA-Speech datasets were used to verify the efficiency of the suggested method

3 Proposed methodology

The CWT-layered CNN model integrates CWT layers and CNN structures as shown in Fig. 1, offering a potent method for signal processing tasks. Beginning with a concise overview, this architecture combines the versatility of CWT in CNN architecture as shown in Section 3.1. The customized CWT-layer extracts frequency information across different wavelets, such as Amor, Morse, and Bump. By integrating the hierarchical feature learning of CNNs, it refines the traditional CWT, adapting it to the specific requirements of the CNN. This adaptation enhances the compatibility and effectiveness of the architecture. The details of this process are explained in Section 3.2. Following this, the CNN layers utilize the transformed representations from the CWT layers to perform feature extraction, and hierarchical abstraction is shown in Section 3.3. Finally, the classification stage of the CWT-based CNN architecture involves utilizing the extracted features to classify signals. This stage demonstrates the model's ability to learn discriminative representations and make accurate predictions. The details are explained in Section 3.4.

3.1 CWT-layered CNN architecture

CWT-layered CNN architecture is a type of neural network that uses wavelet transformations in its layers. This design is especially helpful for applications like time series data, audio processing, and analysis where it is necessary to analyze signals at various scales or resolutions [29]. The input layer is where the data is received. The CWT-layer creates a series of feature maps that depict the signal at various scales and orientations by applying a series of wavelet filters to the input data. The dimension of the analysis frame and the temporal shift have no bearing on the CWT's time and frequency resolutions. Instead, to break down the voice signal, the CWT considers a basic function known as wavelet. Convolution of the signal using shifted and compressed wavelet versions is how this process is carried out. The CWT has been defined formally as

$$C_{wt}(q,r) = \frac{1}{\sqrt{r}}\int_{-\infty}^{\infty} y(t)\phi^{*}\left(\frac{t-q}{r}\right)dt \tag{1}$$

where $y(t)$ is the speech signal, q and r are the scale and shift factors, respectively, and $C_{wt}(q,r)$ is the primary function, which in this study is the Amor wavelet.

Two 2D convolutional layers were used after the wavelet transform layer to extract spatial information from the processed data. These convolutional layers apply filters to the input feature maps to identify the speech patterns and features pertinent to the ADD and ADSLA tasks. In an attempt to downsample the feature maps and reduce their spatial dimensions while preserving crucial

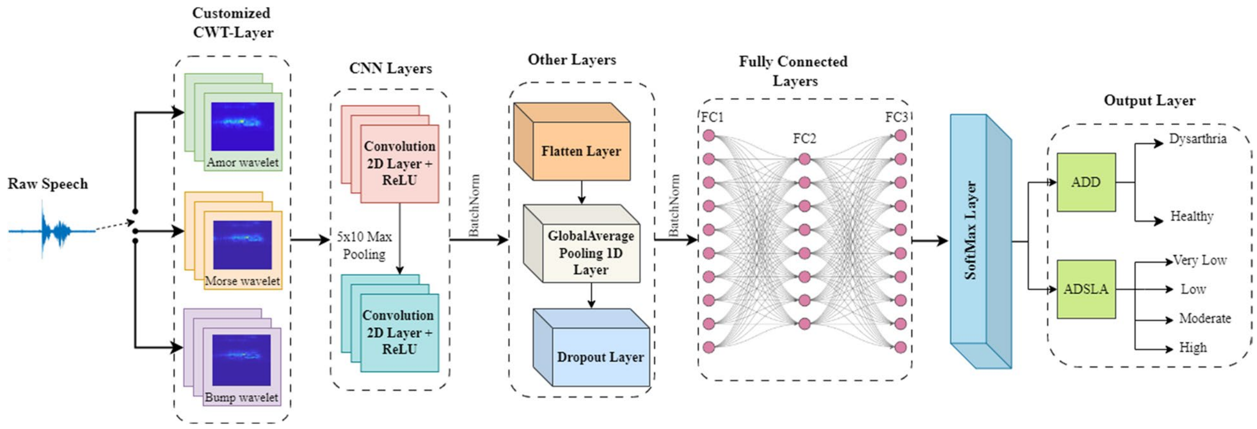


Fig. 1 Flow diagram of ADD and ADSLA using CWT-layered CNN model

information, pooling layers are frequently included following convolutional layers. After the convolutional and pooling layers, the network expanded with fully connected layers depending on the number of classes. To provide predictions or classifications, these layers aggregate and integrate the spatial characteristics that the convolutional layers retrieved. The network's last layer generates the intended output, which is regression result, classification label, or other type of prediction. Network details have been depicted in the Table 1. The integration of wavelet transforms into the network layers is the primary innovation of the CWT-layered CNN architecture. This can help the network better catch intricate patterns and changes in the data by enabling it to assess signals at numerous scales at once.

3.2 Customized CWT-layer

Customized first layers are essential for optimizing the visualization of signals converted with the CWT in the CNN architecture with layers. These specialized layers allow for efficient feature extraction because they are designed to deal with the multi-scale properties of CWT-transformed signals. The architecture can take advantage of the special qualities of the CWT by customizing input layers for CWT-transformed data. This improves the architecture's capacity for analyzing and modeling signals at different wavelets. Modifying learnable parameters in preliminary layers allows customization for optimizing efficiency and performance. Overall, the CWT-layered CNN architecture's incorporation of customizable initial layers improves its applicability for tasks such as multi-scale representations are essential for reliable and precise processing in time-series analysis and audio processing. The following sections will discuss various wavelets and their mathematical representations.

3.2.1 Amor wavelet

Integrating the analytic Morlet (Amor) wavelet within the CWT-layer of audio scalogram analysis is a viable method for deriving significant features from audio data in the time-frequency domain [33]. The Amor wavelet is a popular choice for capturing the complex dynamics of audio sources because of its capacity for localizing both frequency and temporal information. A neural network, like CNN, may dynamically train a hierarchical model of audio information at various scales and frequencies by incorporating the Amor wavelet into its CWT-layer. The network can handle audio scalograms, which give a thorough depiction of the time-frequency information of the signal, efficiently due to this integration. The network's ability to separate discriminative attributes from audio scalograms is improved by the use of the Amor wavelet in the CWT-layer. This allows the network to identify and categorize audio patterns more reliably and accurately. Amor wavelet is defined in the frequency domain as

$$\hat{\Psi}(\omega) = 2e^{(-\omega-6)^2/2}\hat{V}(\omega) \tag{2}$$

Table 1 Network configuration of CWT-layered CNN model with learnables

Layer indices	Type of layer	Layer parameters	Learnables
1	Sequence Input Layer	MinLengh: 4097, Name: Input, Normalization: zscore	-
2	Convolution 1D Layer	Filter size: 5, Number of filters: 1 Name: stride	Weights: 5 × 1 × 1 Bias: 1 × 1
3	CWT-layer	Signal length: 1023, IncludeLowPass: True, Wavelet: Amor, Morse, Bump	Weights: 1 × 1 × 8921
	Maxpooling2D Layer	Pooling window size: 5 × 10	-
4	Convolution 2D layer	Filter size: 5 × 10; Number of filters: 5	Weights: 5 × 10 × 1 × 5, Bias: 1 × 1 × 5
	BatchNormalization	-	Offset: 1 × 5, Scale: 1 × 5
	ReLu Layer	-	-
	Maxpooling2D Layer	Pooling window size: 5 × 10	-
5	Convolution 2D Layer	Filter size: 5 × 10; Number of filters: 10	Weights: 5 × 10 × 5 × 10, Bias: 1 × 1 × 10
	BatchNormalization	-	Offset: 1 × 10, Scale: 1 × 10
	ReLu Layer	-	-
	Maxpooling2d Layer	Pooling window size: 5 × 10	-
-	Flatten Layer	-	-
-	GlobalAveragePooling1d Layer	-	-
-	Dropout Layer	Rate: 50%	-
-	Fully Connected Layer	Size: No. of output classes	-
-	SoftMax Layer	-	-

where the $\hat{V}(\omega)$ is the unit step in frequency. As a result, the wavelet can only be analytic. For the time being, ignoring the unit step portion of the inverse Fourier transformation, the equation will be as follows,

$$\hat{\Psi}(\omega) = 2e^{(-\omega-6)^2/2} \text{ is } \sqrt{\frac{2}{\pi}} e^{-t^2/2} e^{i6t}$$

This provides the fundamental basis for Amor. Naturally, to maintain technical accuracy, we have an inverse Fourier transform of the unit step convolution with the time domain wavelet given above. This is how the equation will be obtained:

$$\hat{\Psi}(\omega) = \sqrt{\frac{2}{\pi}} e^{-t^2/2} e^{i6t} * \frac{i}{2t\Pi} \tag{3}$$

Equations 2 and 3 show that the Amor wavelet is suitable for the real and imaginary valued signals. This means the Amor wavelet in the CWT-layer of the CNN model can help improve audio processing. This enhances the model's ability to analyze audio scalograms and recognize important characteristics.

3.2.2 Morse wavelet

Using Morse wavelets in conjunction with CWT for audio scalogram analysis effectively captures complex frequency-time representations [33]. CWT can effectively interpret audio scalograms and visual representations of audio data in the time-frequency domain by utilizing Morse wavelets. These wavelets offer a versatile framework for evaluating signals over a range of scales and frequencies. Morse wavelets are an excellent choice for tasks like audio categorization and audio event identification. They are versatile in capturing both local and global properties in audio signals. When paired with CNNs, Morse wavelets improve the network's capacity to gather unique characteristics from audio scalograms. CNNs are highly skilled at learning hierarchical models from input data. This integration allows CNNs to automatically identify and learn intricate patterns from audio data. As a result, their performance on tasks requiring in-depth time-frequency analysis, such as voice recognition, music genre classification, and environmental sound categorization, is enhanced. All things considered, using Morse wavelets in CNN architectures to analyze audio scalograms improves the network's capacity to efficiently handle audio data and derive significant features for a range of audio-related applications.

A set of precisely analytic wavelets is called Morse wavelets. Complex-valued wavelets with support for Fourier transform only on their positive real axis are known as analytical wavelets. They can be used to analyze signals that have both frequency and amplitude variations over time, known as modulated signals. They are helpful in the analysis of localized discontinuities as well.

The generalized Morse wavelet's Fourier transform is

$$\psi_{a,b}(\omega) = V(\omega)_{a,b}\omega^{\frac{a^2}{b}} e^{-\omega^b} \tag{4}$$

where $V(\omega)$ is the unit step, $X_{a,b}$ is a normalizing constant, a^2 is the time-bandwidth product, and b characterizes the symmetry of the Morse wavelet. Much of the literature about Morse wavelets uses δ, which can be viewed as a decay or compactness parameter, rather than the time-bandwidth product $a^2 = b\delta$. The equation for the Morse wavelet in the Fourier domain parameterized by δ and b is given as

$$\psi_{\delta,b}(\omega) = V(\omega)X_{a,b}\omega^{\frac{\delta}{b}} e^{-\omega^b} \tag{5}$$

Analytic wavelets with distinct characteristics and behaviors can be created by varying the time-bandwidth combination and symmetry components of a Morse wavelet. One advantage of Morse wavelets is that they may be generalized to create several commonly used analytic wavelets. Morse wavelet supports real-valued signals but shares the same properties with the Amor wavelet, making Morse as second considerable wavelet.

3.2.3 Bump wavelet

To analyze audio scalograms, the Bump wavelet can be modified for usage within the CWT-layer. The Bump is a useful wavelet for capturing localized information in signals because of its smoothness attributes and compact support [34]. The Bump wavelets can be used in the convolutional layer of the neural network to assess audio scalograms by combining them with an input signal at various scales. Through this procedure, the network can extract pertinent information from the scalograms for tasks like audio categorization and speech recognition. The network can efficiently capture specific features within the time-frequency domains of audio signals by utilizing a Bump wavelet in the CWT-layer. This allows for reliable and accurate analysis of audio data. This method can be especially helpful for situations requiring in-depth time-frequency analysis, such as pinpointing individual sound occurrences or spotting irregularities in audio files.

$$\varphi(k\omega) = e^{\left(1-\frac{1}{h}\right)}.I_{\left[\frac{\eta-\theta}{k},\frac{\eta+\theta}{k}\right]} \tag{6}$$

Consider $h = 1 - \frac{(k\omega-\eta)^2}{\Delta^2}$, k and ω are scaling and frequency parameter, whereas Δ is a positive constant. $I_{[\frac{\eta-\theta}{k},\frac{\eta+\theta}{k}]}$ is the indicator function on the interval

$[\frac{\eta-\theta}{k}, \frac{\eta+\theta}{k}]$. The indicator function is 1 on the interval and zero elsewhere. The term $1-\frac{1}{h}$ in the exponent controls the smoothness of the transition from zero to non-zero and back to zero.

The equation defines a function called φ which is related to Bump wavelets. Bump wavelets are a type of wavelet function that smoothly transitions from zero to a non-zero value and then back to zero. The parameter η controls the center of the Bump and θ controls the spread of the Bump.

3.3 CNN layers

In the CWT-based approach, the CNN layers are crucial for feature extraction from the first few layers [35]. In this model, it is essential to extract hierarchical characteristics from the input signals that the CWT layer has processed. The output wavelet coefficients from the CWT layers, which convert the raw waveform into a multi-scale model that captures frequency information, are often passed into the layers of a CNN for additional analysis.

Convolutional, pooling, and activation techniques make up the CNN layers, which use the coefficients of wavelets to extract features. Convolutional filters capture spatial correlations and patterns at various scales by convolving over the modified signals [36]. The feature maps are then down-sampled by pooling layers, which lowers computational complexity while keeping crucial data. Moreover, the network may learn intricate mapping between the input data and their related characteristics because of the non-linearity that activation functions introduce. The flow diagram in Fig. 1 illustrates various layers used in this methodology.

To prevent overfitting, a dropout layer is added to the network. Dropout randomly eliminates a portion of neurons throughout each training cycle, pushing the network to develop stronger and more comprehensive representations. By using this regularization strategy, the network is kept from becoming overly dependent on any one feature and is encouraged to explore a variety of network routes.

At last, a fully connected layer is connected to the output of the dropout layer. This layer creates connections between every neuron and the layers that come after it, allowing complete information to spread throughout the network. See Table 1 for additional details regarding the network design.

3.4 Classification layer

The final predictions are determined by the classification layers of the CNN model, which is based on CWT. A softmax layer processes the fully connected layer's output and determines the probability for each class. This guarantees that the expected probability sum equals one. The network additionally employs a classification layer to determine the classes present in the dataset. This layer uses the probability and learned representations from the preceding levels to assist the network in accurately labeling input samples (Fig. 2).

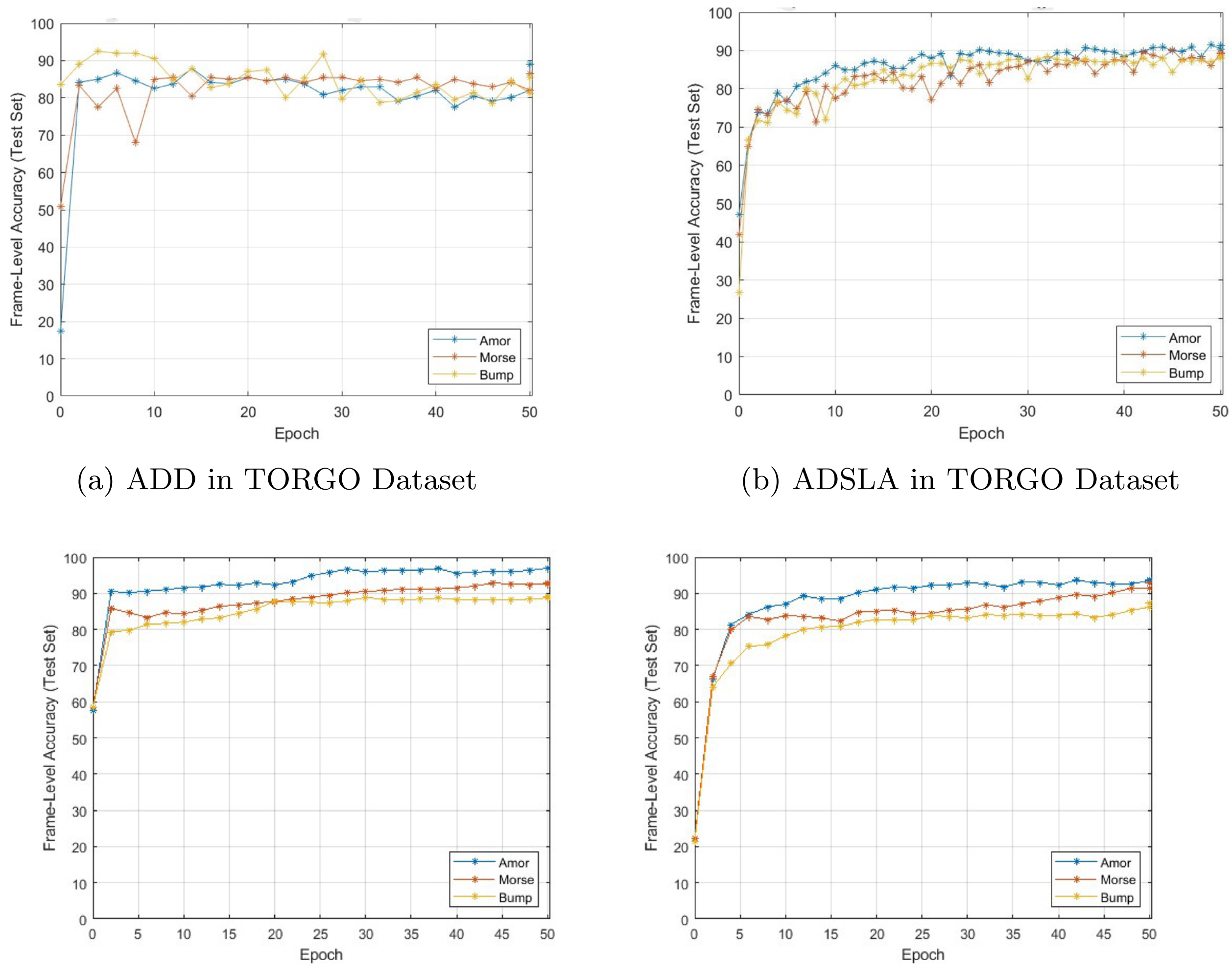
4 Experimental results

The analysis of a computerized network that uses raw speech information from two different datasets, TORGO and UA-Speech, to identify and score dysarthria is presented in this section. The system's goal in this experiment was to categorize dysarthric speech from speech that was in healthy management, thereby differentiating between two binary classes.

The CNN with CWT-layer was used by the system. In particular, three variants of the CNN layer were assessed and trained to detect dysarthria. The TORGO and UA-Speech datasets were split into sets for training (80%) and test sets (20%) to conduct the evaluation. The new or unknown speech signals were then classified as dysarthric or healthy management using the trained variations of the CWT-layered CNN. The evaluation parameter used to evaluate the classifier's performance was overall accuracy.

4.1 Investigating variations of the CWT-layered CNN model

The work aims to analyze raw waveform data for ADD and ADSLA with various CWT-layered CNN model modifications. Time-frequency representations for both dysarthric and healthy speakers are shown in Fig. 3. In addition, the CNN's first layer's wavelets and time-frequency representation scalograms were taken out for comparison. By highlighting the variations between high-level dysarthria and healthy control speech, these differences enable a graphical representation in the time-frequency domain. It was observed that a healthy subject spoke the word "*Delete*" in just 1.8 s, displaying a broad range of frequency components typical of natural speech. In contrast, the high-level dysarthria subject required almost 4.5 s, with the speech signal exhibiting significant energy at lower frequencies. The scalograms reveal that healthy speech exhibits more consistent and continuous frequency content over time, reflecting regular and stable speech patterns. In contrast, dysarthric speech shows more localized and less stable frequency content, indicating irregularities and disruptions in speech production. The CWT layer is highly effective in analyzing signals jointly in both time and frequency domains, allowing for a detailed examination of speech patterns and the identification of transient features. CWT's proficiency in localizing transients further enhances its utility in distinguishing between the stable frequency content of healthy speech and the disrupted patterns observed in



(a) ADD in TORGO Dataset

(b) ADSLA in TORGO Dataset

(c) ADD in UA-Speech Dataset

(d) ADSLA in UA-Speech Dataset

Fig. 2 Performance curves of CWT-layered CNN architecture with various wavelets for ADD and ADSLA tasks, measured in terms of frame-level accuracy (%)

dysarthric speech. This analysis was conducted using all the wavelets. Bump wavelets, due to their design, usually have higher frequency components compared to other wavelets. This is because Bump wavelets are intended to be extremely localized in terms of both time and frequency, and achieving this requires a greater frequency concentration.

However, the Amor and Morse wavelets are both designed with real and imaginary components. This feature makes them effective in analyzing signals across different domains [33]. The real components of these wavelets capture amplitude variations, while the imaginary components convey phase information. This provides a comprehensive time-frequency representation of the signal. Unlike the STFT, which uses a fixed window size limiting its resolution, wavelets in the CWT adapt in scale. This adaptation allows for superior analysis of non-stationary signals with varying frequency content over time. The two-component structure of the Amor and Morse wavelets allows for breaking down signals into their parts and providing information about both temporal and spectral properties at the same time. By utilizing the capacity of these wavelets to capture both phase and amplitude dynamics, we can more accurately and deeply perform complex signal processing tasks such as time-frequency analysis, feature extraction, and denoising. The scalograms of Amor and Morse wavelets shown in Fig. 3 have almost similar characteristics which imply that both the wavelets show similar behavior for speech processing.

On the other hand, the Bump wavelet is specifically designed to represent real-valued signals and functions. The wavelet is naturally suited for this purpose due to its fast decay and compact support. Unlike several wavelets that combine real and imaginary

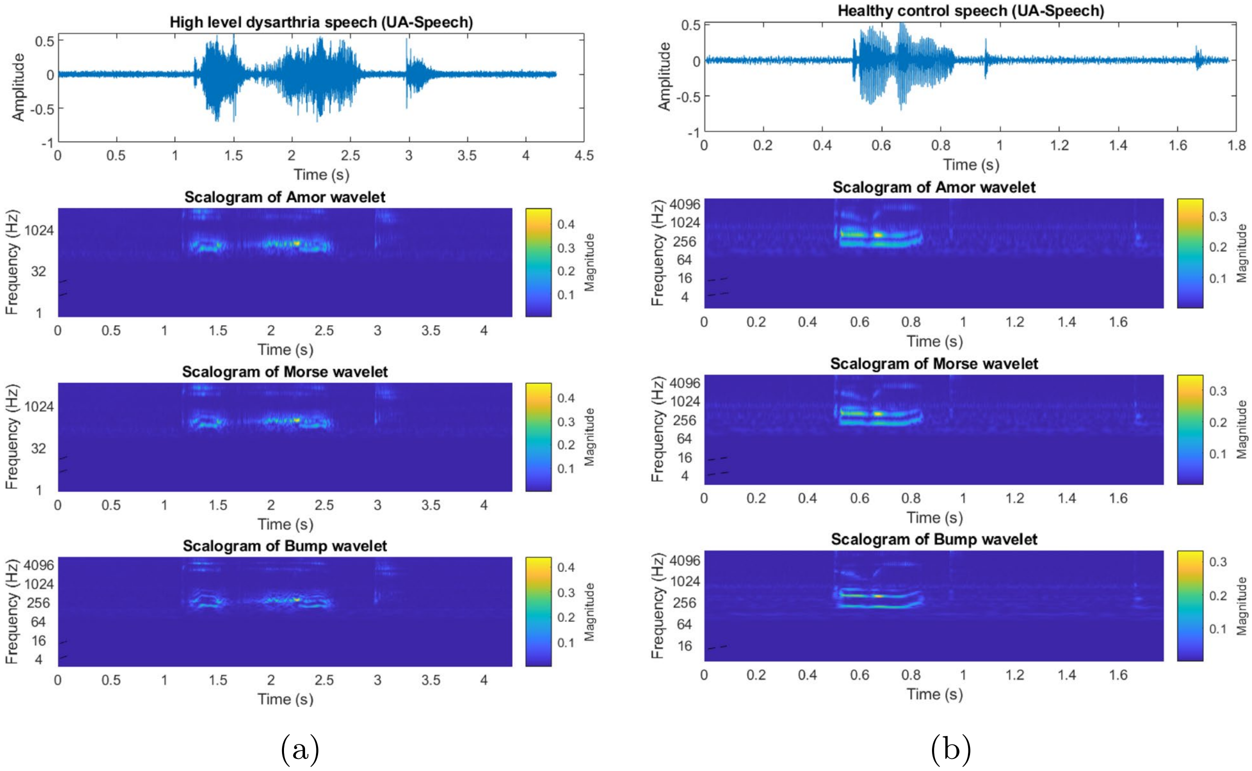


Fig. 3 Examining CNN variants with CWT layers to distinguish between **a** high-level dysarthria and **b** healthy control speech on the word "*Delete*" for UA-Speech dataset

components, the Bump wavelet functions exclusively in the real domain [34]. This restriction is warranted by the wavelet's main purpose, which is to analyze and dissect real-world signals that usually only display real-valued characteristics. The Bump wavelet ensures effective and accurate signal representation by limiting itself to the real components and avoiding needless complexity. This eliminates the need to take imaginary components into account. Moreover, this real-only characteristic fits well with a wide range of real-world signal analysis applications, where the focus is frequently on deriving insights and significant characteristics from real-valued data.

4.2 Dataset

Two datasets are included in ADD and ADSLA. A comprehensive analysis of the performance of the suggested methods is made possible by the presence of both moderate as well as large datasets. Larger datasets are very beneficial to neural networks because, as data-hungry models, they enable the acquisition of a greater variety of patterns and variances found in speech data. As a result, neural networks that have been trained on larger datasets perform better and can more successfully adapt to new data.

4.2.1 TORGO

A well-liked dataset for researching dysarthria, a speech disability that impairs articulation, is the TORGO database [37]. It includes measures of speech motions in great detail as well as audio recordings of both dysarthria sufferers and a group of controls of healthy people. Table 2 lists the database's data sources: eight dysarthric speakers and seven healthy control speakers, all of whom are between the ages of sixteen and fifty.

4.2.2 UA-Speech

Thirteen healthy control speakers and fifteen dysarthric speakers with cerebral palsy have recordings in the UA-Speech dataset [38]. The collection contains 455 distinct terms, including frequent and uncommon keywords, computer commands, and international radio alphabets and numerals. To make sure that each speaker had been covered in every category, three-word blocks were recorded. Each speaker recorded a total of 765 distinct words. The dataset includes speech intelligence

Table 2 Details of TORGO and UA-Speech datasets in ADD and ADSLA tasks

Dataset	Specifications of dataset	ADD		ADSLA			
		Dysarthria	Healthy control	Very low	Low	Moderate	High
TORGO	Number of speakers	8	7	3	2	3	-
	Gender distribution	3Female 5Male	3Female 4Male	2Female 1Male	1Female 1Male	3Male	-
	Speaker identities	F1,F3,F4, M1,M2,M3, M4,M5	FC1,FC3,FC4, MC1,MC2,MC3, MC4	F3,F4,M3	F1,M5	M1,M2,M4	-
	No. of raw audio files	1000	1000	1050	403	547	-
UA-Speech	Number of speakers	15	13	5	3	3	4
	Gender distribution	4Female 11Male	4Female 9Male	1Female 4Male	1Female 2Male	1Female 2Male	1Female 3Male
	Speaker identities	F2,F3,F4, F5,M1,M4, M5,M7,M8, M9,M10,M11, M12,M14,M16	CF2,CF3,CF4, CF5,CM1,CM4, CM5,CM7,CM8, CM9,CM10,CM12 CM13	F5,M8,M9, M10,M14	F4,M5, M11	F2,M7, M16	F3,M1, M4,M12
	No. of raw audio files	73950	64260	26775	15300	15300	16575

assessments for each dysarthric speaker provided by five impartial listeners. These results provide an unbiased evaluation of the speaker's comprehensibility.

4.3 Automatic dysarthria detection

For ADD, we applied a binary classification technique to differentiate between dysarthria and healthy speech samples. In ADD, it has been observed that the UA-Speech dataset presents better results as compared to TORGO. Initially, using the UA-Speech dataset the CNN's first layer is replaced with the CWT-layer with the Bump wavelet, resulting in an accuracy of 88.98%. Subsequently, by using the Morse wavelet, it has been observed that the model achieved an accuracy of 92.72%, which is considerable. The model accuracy is further improved to 97.00% using the Amor wavelet. In comparison, for the TORGO dataset, the Bump wavelet exhibited an accuracy of 85.42%, which is increased to 87.69% by using the Morse wavelet. To further improve the accuracy of the model, we have used the Amor wavelet, which achieved the highest accuracy of 88.50%. The higher amount of data in UA-Speech has led to an increased accuracy in ADD. After analyzing the performance disparities among the wavelet transforms, we found that the characteristics of the Amor wavelet, such as its ability to capture both temporal and spectral information effectively, contributed significantly to its superior performance for both datasets. Additionally, the Amor wavelet's flexibility and adaptability in representing complex speech signals likely played a pivotal role in enhancing the accuracy of dysarthria detection. The Amor wavelet and Morse wavelet make it easier to differentiate between dysarthria and healthy speech signals and enhance the way lower-intensity frequency components are represented.

Upon analyzing the confusion matrix presented in Fig. 4a and c, we found that 23 dysarthria classes were incorrectly identified as healthy patients, and 23 classes of healthy were identified as dysarthria within the TORGO dataset. Despite this, because of the UA-Speech dataset's larger size, even the cases in which 522 healthy

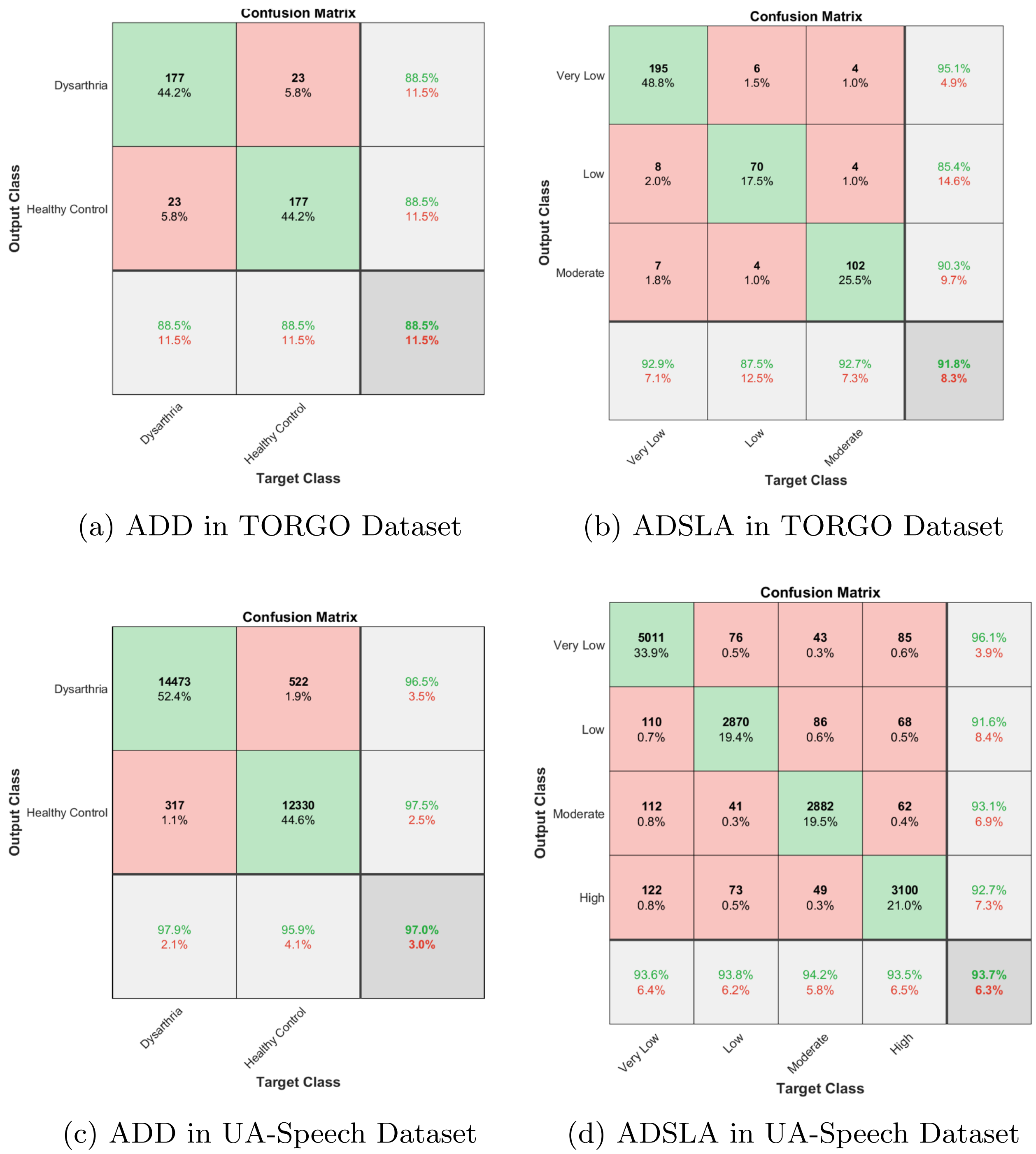
Table 3 Performance assessment of CWT-layered CNN models using the UA-Speech and TORGO datasets in terms of accuracy (%)

Model	Wavelets	TORGO		UA-Speech	
		ADD (binary class)	ADSLA (multi-class)	ADD (binary-class)	ADSLA (multi-class)
CWT-layered CNN	Amor	88.50	91.80	97.00	93.70
	Morse	87.69	89.7	92.72	92.77
	Bump	85.42	88.4	88.98	87.43

patients were improperly categorized as dysarthria and 317 dysarthria subjects incorrectly identified as healthy did not considerably impair the model's overall accuracy. Consequently, with the Amor wavelet, the ultimate accuracy reached 97.00%, as shown in Table 3. Additionally, Fig. 5 illustrates the effectiveness of using Amor wavelet in improving ADD performance.

4.4 Automatic dysarthria severity level assessment

ADSLA plays a critical role in diagnosing and monitoring speech disorders, with accurate and reliable assessments being essential for effective treatment planning and intervention strategies. The UA-Speech dataset has four different classes: very low, low, moderate, and high as shown in Table 2. Using the UA-Speech dataset ADSLA achieved



	Dysarthria	Healthy Control	
Dysarthria	177 44.2%	23 5.8%	88.5% 11.5%
Healthy Control	23 5.8%	177 44.2%	88.5% 11.5%
	88.5% 11.5%	88.5% 11.5%	88.5% 11.5%

(a) ADD in TORGO Dataset

	Very Low	Low	Moderate	
Very Low	195 48.8%	6 1.5%	4 1.0%	95.1% 4.9%
Low	8 2.0%	70 17.5%	4 1.0%	85.4% 14.6%
Moderate	7 1.8%	4 1.0%	102 25.5%	90.3% 9.7%
	92.9% 7.1%	87.5% 12.5%	92.7% 7.3%	91.8% 8.3%

(b) ADSLA in TORGO Dataset

	Dysarthria	Healthy Control	
Dysarthria	14473 52.4%	522 1.9%	96.5% 3.5%
Healthy Control	317 1.1%	12330 44.6%	97.5% 2.5%
	97.9% 2.1%	95.9% 4.1%	97.0% 3.0%

(c) ADD in UA-Speech Dataset

	Very Low	Low	Moderate	High	
Very Low	5011 33.9%	76 0.5%	43 0.3%	85 0.6%	96.1% 3.9%
Low	110 0.7%	2870 19.4%	86 0.6%	68 0.5%	91.6% 8.4%
Moderate	112 0.8%	41 0.3%	2882 19.5%	62 0.4%	93.1% 6.9%
High	122 0.8%	73 0.5%	49 0.3%	3100 21.0%	92.7% 7.3%
	93.6% 6.4%	93.8% 6.2%	94.2% 5.8%	93.5% 6.5%	93.7% 6.3%

(d) ADSLA in UA-Speech Dataset

Fig. 4 Confusion-matrices of ADD and ADSLA for Amor wavelet-based CWT-layered CNN architecture

the highest accuracy of 93.70% for the Amor wavelet, 92.77% for the Morse wavelet, and 87.43% for the Bump wavelet. The performance validation curve of ADSLA for the UA-Speech dataset has been depicted in Fig. 2d. Another experiment of ADSLA on the TORGO dataset has three different classes: very low, low, and moderate. To improve accuracy, we performed data augmentation, increasing the data size from 1000 files to 2000 files in dysarthria speech. Details of speaker identities and number of raw audio files are shown in Table 2. For the TORGO dataset, the Amor wavelet achieved the highest accuracy of 91.8%, followed by the Morse wavelet at 89.7%, and the Bump wavelet at 88.4% which can be referred to in Table 3. The performance curves of the ADSLA using the TORGO dataset can be referred to in Fig. 2b, and the confusion matrix as shown in Fig. 4b represents the mapping between output and target classes. This finding underscores the effectiveness of the Amor

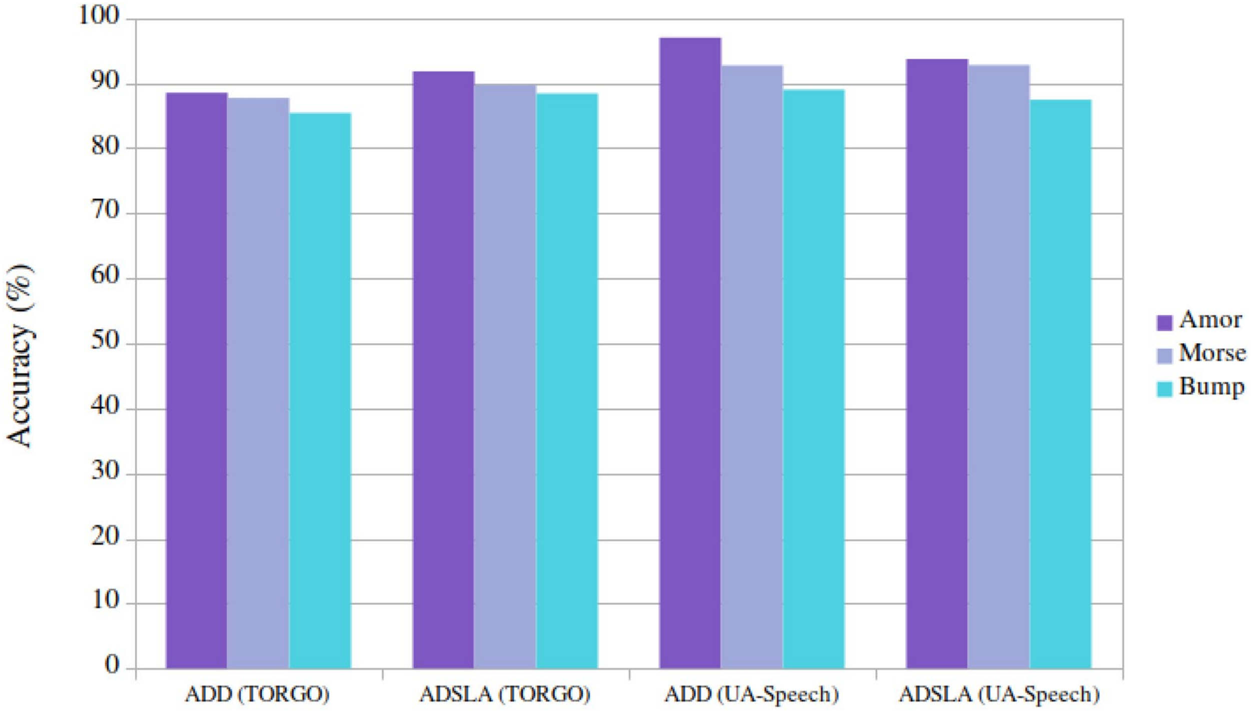


Fig. 5 Performance analysis of the different wavelets of CWT-layered CNN model for TORGO and UA-Speech datasets

Table 4 Performance comparison of the suggested model with earlier methods using the TORGO dataset

Year/author	Dataset used	Feature extraction	Model used	Performance
2019 [9] Millet and Zeghidour	TORGO	Raw waveform	LSTM and attention	Highest accuracy: 75.63%
2020 [22] Hernandez et al.	TORGO	Rhythm metrics	RF and SVM	RF: 81.50%, SVM: 82.30%
2023 [6] Yue et al.	TORGO	Raw magnitude	Cascade convolution	Dysarthria detection: 88.00% Dysarthria severity: 86.00%
2024 [23] Radha et al.	TORGO	Raw waveform	STFT layer	Dysarthria detection: 94.62% Severity level Assessment: 90.15%
2024 [10] Sajiha et al.	TORGO	Raw waveform	SincNet layer	Dysarthria detection: 97.00% Dysarthria Speaker identification: 88.00%
Proposed methodology	TORGO	Raw waveform	CWT layer (Amor)	ADD: 88.50% ADSLA: 91.80%

Table 5 Performance comparison of the suggested model with earlier methods using the UA-Speech dataset

Year/author	Dataset used	Feature extraction	Model used	Performance
2021 [24] Narendra and Alku	UA-Speech	Raw glottal Flow waveforms	CNN and MLP	Dysarthria detection: 87.93%
2021 [25] Kachhi et al.	UA-Speech	CWT scalograms	CNN	Severity level Assessment: 95.17%
2023 [26] Joshy and Rajan	UA-Speech	Mel-spectrograms	SE CNN	Dysarthria detection: 87.93%
2024 [23] Radha et al.	UA-Speech	Raw waveforms	STFT layer	Dysarthria detection: 99.89% Severity level Assessment: 94.67%
Proposed methodology	UA-Speech	Raw waveforms	CWT layer (Amor)	ADD: 97.00% ADSLA:93.70%

wavelet in accurately assessing dysarthria severity levels in both datasets. The superior performance of the Amor wavelet can be attributed to its ability to capture both temporal and spectral features efficiently, thereby enabling more comprehensive representations of speech signals.

To understand why the Amor wavelet performed better than the Morse and Bump wavelets, it is important to consider the unique properties of each wavelet transform. The Amor wavelet is known for its adaptability and flexibility in representing complex signals, allowing it to capture the intricate temporal and spectral variations in dysarthric speech. On the other hand, while the Morse and Bump wavelets have their strengths, such as their ability to localize features in the time-frequency domain, they may not be as effective as the Amor wavelet in capturing the nuanced characteristics of dysarthric speech.

4.5 Comparative analysis

The prior method concentrates on pertinent segments of the input sequence employing filterbanks and LSTM attention mechanisms to capture significant speech elements [9]. Other models, including ADD and ADSLA, are classic machine learning methods, like RF and SVM, that are employed for classification problems [22]. When processing voice data, CNNs and GRUs each handle different tasks: CNNs record local patterns, GRUs represent sequential dependencies [39], and cascade convolution models frequently combine multiple convolutional layers.

The suggested method uses the CWT-layered deep learning model with raw waveform and a variety of wavelets, including Amor, Morse, and Bump. However, despite slightly lower performance metrics, the CWT's detailed signal analysis offers potential for broader applications and deeper insights, enhancing the accuracy and robustness of speech analysis models. Unlike the STFT, which uses a fixed window size, the CWT adapts to different frequencies, effectively capturing transient and non-stationary features in dysarthric speech. Results show that with the Amor wavelet on the UA-Speech dataset, the CWT achieved accuracy rates of 97.00% for ADD and 93.70% for ADSLA, and on the TORGO dataset, 88.50% for ADD and 91.80% for ADSLA. Using the Amor wavelet, the CWT achieved comparable accuracy rates on the UA-Speech and TORGO datasets, demonstrating its equivalence to existing techniques. CWT's nuanced view of signals is crucial for developing robust dysarthria diagnosis models. Additional details regarding the existing methods can be found in Table 4 for TORGO and Table 5 for UA-Speech, along with a comparison of the proposed method.

5 Conclusion

The proposed CWT layered CNN model showed that the Amor wavelet performed consistently better than the other wavelets in both the ADD and ADSLA. This result demonstrated the resilience and generalizability of the Amor wavelet in this field and remained consistent for both the TORGO and UA-Speech datasets. The significance of this outcome lies in its implications for real-world applications. This holds promise for enhancing diagnostic processes, facilitating early intervention, and ultimately improving the quality of life for individuals affected by dysarthria. The findings of this paper contribute to the advancement of ADD and ADSLA, shedding light on the pivotal role of wavelet selection in enhancing model performance. By embracing the Amor wavelet within the CWT-layered CNN architecture, we pave the way for more accurate, efficient, and reliable diagnostic tools for dysarthria evaluation, thus offering enhanced benefits to both clinicians and individuals impacted by this speech disorder.

Authors' contributions

SS and NS were responsible for the model architecture design and experimentation, participated in sequence alignment, and drafted the manuscript. KR and DVR contributed to drafting, reviewing, and supervision. SG reviewed the results and the manuscript. DPB supervised the work and corresponded with the editor. All authors read and approved the final manuscript.

Funding

Open access funding provided by Norwegian University of Science and Technology This research received no external funding.

Availability of data and materials

The open access TORGO data that support the findings of this study are available from the Kaggle repository. The University of Illinois team provided the UA-Speech data upon request. More details about the data are given in Section 4.2.

Declarations

Competing interests

The authors declare no competing interests.

Received: 8 April 2024 Accepted: 17 June 2024
Published online: 25 June 2024

References

1. M.J. Vansteensel, E. Klein, G. van Thiel, M. Gaytant, Z. Simmons, J.R. Wolpaw, T.M. Vaughan, Towards clinical application of implantable brain-computer interfaces for people with late-stage ALS: Medical and ethical considerations. J. Neurol. **270**(3), 1323–1336 (2023)
2. S.M. Shabber, M. Bansal, K. Radha, in *2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)*. Machine learning-assisted diagnosis of speech disorders: A review of dysarthric speech (IEEE, Roorkee, India, 2023), pp. 1–6
3. S.M. Shabber, M. Bansal, K. Radha, in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. A review and classification of amyotrophic lateral sclerosis with speech as a biomarker (IEEE, Delhi, India, 2023), pp. 1–7
4. M. Carl, E.S. Levy, M. Icht, Speech treatment for hebrew-speaking adolescents and young adults with developmental dysarthria: A comparison of mSIT and Beatalk. Int. J. Lang. Commun. Disord. **57**(3), 660–679 (2022)
5. V. Mendoza Ramos, The added value of speech technology in clinical care of patients with dysarthria. Ph.D. thesis, University of Antwerp (2022)
6. Z. Yue, E. Loweimi, H. Christensen, J. Barker, Z. Cvetkovic, in *INTERSPEECH*. Dysarthric speech recognition from raw waveform with parametric CNNs. (IEEE, Incheon, Korea, 2022), pp. 31–35
7. N. Tavabi, D. Stück, A. Signorini, C. Karjadi, T. Al Hanai, M. Sandoval, C. Lemke, J. Glass, S. Hardy, M. Lavallee et al., Cognitive digital biomarkers from automated transcription of spoken language. J. Prev. Alzheimer Dis. **9**(4), 791–800 (2022)
8. K. Radha, M. Bansal, Towards modeling raw speech in gender identification of children using sincNet over ERB scale. Int. J. Speech Technol. **26**(3), 651–663 (2023)
9. J. Millet, N. Zeghidour, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Learning to detect dysarthria from raw speech (IEEE, Brighton, UK, 2019), pp. 5831–5835
10. S. Sajiha, K. Radha, D.V. Rao, V. Akhila, N. Sneha, in *2024 National Conference on Communications (NCC)*. Dysarthria diagnosis and dysarthric speaker identification using raw speech model (IEEE, Chennai, India, 2024)
11. K. Radha, M. Bansal, Feature fusion and ablation analysis in gender identification of preschool children from spontaneous speech. Circ. Syst. Signal Process. **42**(10), 6228–6252 (2023)
12. K. Radha, M. Bansal, Audio augmentation for non-native children's speech recognition through discriminative learning. Entropy **24**(10), 1490 (2022)
13. K. Radha, M. Bansal, S.M. Shabber, in *2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP)*. Accent classification of native and non-native children using harmonic pitch (IEEE, Amaravati, India, 2022), pp. 1–6
14. K. Radha, M. Bansal, R. Sharma, in *2023 10th International Conference on Signal Processing and Integrated Networks (SPIN)*. Whitening transformation of i-vectors in closed-set speaker verification of children (IEEE, Noida, India, 2023), pp. 243–248
15. A.K. Jardine, D. Lin, D. Banjevic, A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mech. Syst. Signal Process. **20**(7), 1483–1510 (2006)
16. S. Latif, J. Qadir, A. Qayyum, M. Usama, S. Younis, Speech technology for healthcare: Opportunities, challenges, and state of the art. IEEE Rev. Biomed. Eng. **14**, 342–356 (2020)
17. P. Enderby, Disorders of communication: Dysarthria. Handb. Clin. Neurol. **110**, 273–281 (2013)
18. S.K. Maharana, A. Illa, R. Mannem, Y. Belur, P. Shetty, V.P. Kumar, S. Vengalil, K. Polavarapu, N. Atchayaram, P.K. Ghosh, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Acoustic-to-articulatory inversion for dysarthric speech by using cross-corpus acoustic-articulatory data (IEEE, Toronto, Ontario, Canada, 2021), pp. 6458–6462
19. B. Suhas, D. Patel, N.R. Koluguri, Y. Belur, P. Reddy, A. Nalini, R. Yadav, D. Gope, P.K. Ghosh, in *INTERSPEECH*. Comparison of speech tasks and recording devices for voice based automatic classification of healthy subjects and patients with amyotrophic lateral sclerosis. (IEEE, Graz, Austria, 2019), pp. 4564–4568
20. K.M. Yorkston, Treatment efficacy: Dysarthria. J. Speech Lang. Hear. Res. **39**(5), S46–S57 (1996)
21. H. Chandrashekar, V. Karjigi, N. Sreedevi, Spectro-temporal representation of speech for intelligibility assessment of dysarthria. IEEE J. Sel. Top. Signal Process. **14**(2), 390–399 (2019)
22. A. Hernandez, E.J. Yeo, S. Kim, M. Chung, in *INTERSPEECH*. Dysarthria detection and severity assessment using rhythm-based metrics. (IEEE, Shanghai, China, 2020), pp. 2897–2901
23. K. Radha, M. Bansal, V.R. Dulipalla, Variable STFT layered CNN model for automated dysarthria detection and severity assessment using raw speech. Circ. Syst. Signal Process. **43**, 3261–3278 (2024). https://doi.org/10.1007/s00034-024-02611-7
24. N. Narendra, P. Alku, Glottal source information for pathological voice detection. IEEE Access **8**, 67745–67755 (2020)
25. A. Kachhi, A. Therattil, P. Gupta, H.A. Patil, in *International Conference on Speech and Computer*. Continuous wavelet transform for severity-level classification of dysarthria (Springer, Gurugram, India, 2022), pp. 312–324
26. A.A. Joshy, R. Rajan, Dysarthria severity classification using multi-head attention and multi-task learning. Speech Commun. **147**, 1–11 (2023)
27. C. Divakar, R. Harsha, K. Radha, D.V. Rao, N. Madhavi, T. Bharadwaj, in *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. Explainable AI for CNN-LSTM network in PCG-based valvular heart disease diagnosis (IEEE, Noida, India, 2024), pp. 92–97
28. K. Radha, D.V. Rao, K.V.K. Sai, R.T. Krishna, A. Muneera, in *2024 International Conference on Green Energy, Computing and Sustainable Technology (GECOST)*. Detecting autism spectrum disorder from raw speech in children using STFT layered CNN model (IEEE, Miri, Sarawak, Malaysia, 2024), pp. 437–441
29. K. Radha, M. Bansal, R. Sharma, Raw waveform-based custom scalogram CRNN in cardiac abnormality diagnosis. IEEE Access. **12**, 13986–14004 (2024). https://doi.org/10.1109/ACCESS.2024.3356075
30. C. Bhat, B. Vachhani, S.K. Kopparapu, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Automatic assessment of dysarthria severity level using audio descriptors (IEEE, New Orleans, USA, 2017), pp. 5070–5074
31. J. Fritsch, M. Magimai-Doss, Utterance verification-based dysarthric speech intelligibility assessment using phonetic posterior features. IEEE Signal Process. Lett. **28**, 224–228 (2021)
32. D. Korzekwa, R. Barra-Chicote, B. Kostek, T. Drugman, M. Lajszczak, in *INTERSPEECH*. Interpretable deep learning model for the detection and reconstruction of dysarthric speech. (IEEE, Graz, Austria, 2019), pp. 3890–3894. https://doi.org/10.21437/Interspeech.2019-1206

33. P. Gupta, P.K. Chodingala, H.A. Patil, in *2022 30th European Signal Processing Conference (EUSIPCO)*. Morlet wavelet-based voice liveness detection using convolutional neural network (IEEE, Belgrade, Serbia, 2022), pp. 100–104
34. P. Gupta, S. Gupta, H. Patil, in *9th International Conference on Pattern Recognition and Machine Intelligence*. Voice liveness detection using bump wavelet with CNN (Springer, Kolkata, India, 2021)
35. K. Radha, M. Bansal, R.B. Pachori, Speech and speaker recognition using raw waveform modeling for adult and children's speech: A comprehensive review. Eng. Appl. Artif. Intell. **131**, 107661 (2024)
36. K. Radha, M. Bansal, R.B. Pachori, Automatic speaker and age identification of children from raw speech using sincNet over ERB scale. Speech Commun. **159**, 103069 (2024)
37. F. Rudzicz, A.K. Namasivayam, T. Wolff, The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Lang. Resour. Eval. **46**, 523–541 (2012)
38. H. Kim, M. Hasegawa-Johnson, A. Perlman, J.R. Gunderson, T.S. Huang, K.L. Watkin, S. Frame, in *INTERSPEECH*. Dysarthric speech database for universal access research, vol. 2008. (IEEE, Incheon, Korea, 2008), pp. 1741–1744
39. D.H. Shih, C.H. Liao, T.W. Wu, X.Y. Xu, M.H. Shih, Dysarthria speech detection using convolutional neural networks with gated recurrent unit. Healthcare **10**(10), 1956 (2022)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.