Computers in Biology and Medicine 195 (2025) 110681

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A novel speech signal feature extraction technique to detect speech impairment in children accurately

Manisa Manoswini , Biswajit Sahoo , Aleena Swetapadma *

School of Computer Engineering, KIIT Deemed to Be University, Bhubaneswar, 751024, India

ARTICLE INFO

Keywords:
Artificial intelligence
Speech signal
Signal processing
Speech features
Deep learning
Speech impairment detection

ABSTRACT

Speech signal processing and extracting useful information from speech signal is necessary for speech language impairment (SLI) detection in children. Although different features has been suggested for SLI detection, there is still a scope exist for exploration of other methods. A comparative study of different techniques for feature extraction can be done to find the optimal feature extraction technique. In this work, a study has been carried out to obtain optimal feature extraction technique for SLI detection. Inputs used for SLI detection here are the speech signals recorded from children. Features are first extracted from the recorded speech signals using various feature extraction techniques. The feature extraction techniques that has been implemented are relative spectral transform - perceptual linear prediction (RASTA), wavelet packet transform (WPT), linear predictive coding (LPC), perceptual linear prediction (PLP), Mel-Frequency cepstral coefficients (MFCC), complex quantization cepstral coefficient (CQCC), perceptual noise cepstral coefficients (PNCC). The features extracted are then given to deep learning models namely transformer, temporal convolutional networks (TCN) and TabNet for SLI detection. The result obtained has highest accuracy of 100.00 % using PNCC feature combined with TabNet method. The novelty of the method is that the PNCC features has not been suggested for SLI detection previously. The proposed method can be used for speech impairment detection and monitoring by therapist and doctors.

1. Introduction

Speech signals are noteworthy because of their crucial role in human communication and their wide-ranging applications in areas such as speech recognition, speaker identification, and medical diagnostics [1]. These signals are fundamentally non-stationary which mirrors the dynamic essence of human speech. The non-stationary characteristics of speech signals means variations in frequency and amplitude over time [2]. The majority of a speech signal's energy is found in the lower frequency range. The fundamental frequency of speech or pitch varies among individuals, generally falling between 85 Hz and 255 Hz for adult males and females [3]. Extracting meaningful features from speech signals is crucial for various applications such as speech impairment SLI detection. Various techniques has been used for speech feature extraction and SLI detection as discussed below.

In [4], the SLI in children using has been detected using speech data. The feature extraction method involves MFCC for speech parameterization, which help in distinguishing healthy children from those affected by SLI. Additionally, Gaussian posteriograms learned from frame-level acoustic features are used to develop classifiers with extreme learning machines (ELM), with the kernel ELM achieving an accuracy of 99.41 % [5]. The paper focuses on categorizing children with SLI using pitch-based statistical features. The features are extracted from speech data and used with a weighted nearest neighbor (k-NN) classifier optimized through neighborhood component analysis (NCA). The model achieves an accuracy of 97.93 % with 3-NCA optimized features, demonstrating the effectiveness of pitch-based parameters for SLI classification [6]. The paper proposes a novel approach for detecting SLI in children using glottal source features extracted through glottal inverse filtering (GIF). Time- and frequency-domain glottal parameters, along with MFCC and openSMILE acoustic features, are extracted from speech utterances. The combination of glottal features with MFCC provides the best performance in detecting SLI when used with a feed-forward neural network (FFNN) classifier [7]. The paper focuses on detecting SLI in children using linear predictive coding (LPG) coefficients, which model the human vocal system. LPC features are extracted from children's speech and used with classifiers like Naïve Bayes (NB) and support vector machine (SVM). The study demonstrates that LPC parameters are

* Corresponding author.
E-mail address: aleena.swetapadmafcs@kiit.ac.in (A. Swetapadma).

https://doi.org/10.1016/j.compbiomed.2025.110681
Received 20 March 2025; Received in revised form 6 June 2025; Accepted 25 June 2025
Available online 29 June 2025

effective in distinguishing between children with SLI and healthy children, achieving classification accuracy of 97.9 % and 97.8 % [8].

The paper proposes a machine learning system for detecting SLI by analyzing the texture of speech utterances using spectrograms. Textural features such as Haralick's features and local binary patterns (LBPs) are extracted from these spectrograms to capture changes in audio quality. The combination of Haralick's and LBP features results in high classification accuracy, achieving up to 99 % with gender and speaker independence [9]. The paper proposes a new approach for detecting SLI by capturing the texture in pathological speech using LBP from the joint time-frequency representation. LBP extracts latent temporal information not typically captured by standard audio features. The system achieves improved classification accuracy, with a 13.1 % increase when LBP features are fused with traditional features, reaching an average accuracy of 97.36 % using 5-fold cross-validation [10]. The study proposes two approaches for detecting SLI in children using raw speech signals: a customized convolutional neural network (CNN) and a hybrid CNN-Long short-term memory (LSTM) model. These models automatically extract fuzzy-automated features from speech utterances through convolution filters, eliminating the need for dedicated feature extraction. The hybrid model achieved 100 % accuracy, outperforming the CNN alone, which achieved around 90 % accuracy [11]. This study proposes a novel feature extraction method for detecting SLI using vowels. It combines a favipiravir molecular structure patter, statistical feature extraction, and wavelet packet decomposition (WPD) to generate multilevel features. The most meaningful features are selected using Iterative INCA, and these features are classified using a SVM, achieving high classification accuracy [12].

This work focuses on predicting SLI in children using the improved conditional random fields (ICRF) approach. The feature extraction method involves converting sound signals from open databases into a corpus, followed by supervised learning to recognize and correct speech disability features. The ICRF segments are used to partition and eliminate negative factorized features, improving the accuracy of speech signal classification with a 89.14 % accuracy rate [13]. This study focuses on identifying SLI in children using speech utterances. The feature extraction methods include MFCC combined with a neural network and linear predictive coding (LPG) combined with a multi-layer perceptron (MLP). The results show that MFCC features combined with the MLP model yield the best performance, demonstrating the effectiveness of this approach in diagnosing SLI [8]. This paper introduces an interpretable deep learning model, the interpretable multi-band feature extraction network (IMBFN), for automatic pathological voice detection (APVD). The feature extraction method in IMBFN uses an amplitude-trainable SincNet (AT-SincNet) filter bank as the front-end frequency division network and a two-path one-dimensional depthwise separable convolutional neural network (CNN) to extract meaningful voice features, improving interpretability and generalization performance in APVD [14].

The existing literature demonstrates significant advancements in the detection of SLI in children using diverse feature extraction techniques and machine learning classifiers. Techniques such as MCC, LPC, glottal source features, and spectrogram-based methods have proven effective for capturing speech characteristics relevant to SLI detection. Additionally, classifiers like ELM, SVM, FFNN, and hybrid CNN-LSTM architectures have yielded impressive results, with some studies reporting near-perfect classification accuracy. However, there remains a need to systematically evaluate and compare the effectiveness of various feature extraction methods within a unified deep learning framework. To address this gap, our study proposes a comprehensive approach to SLI detection that integrates multiple feature extraction techniques with modern deep learning architectures. The primary objectives are to evaluate the performance of different speech feature extraction methods such as MFCC, LPC, PLP, WPT, and others-and to determine their effectiveness when used with advanced neural network models, specifically TCN and TabNet.

2. Techniques used

In this work, different techniques has been used for feature extraction from speech signal and classification of speech to healthy and impaired. Various techniques used for feature extraction from speech are RASTA, WPT, LPC, PLP, MFCC, CQCC, CFCC, PNCC, etc.

2.1. For feature extraction

2.1.1. Mel-frequency cepstral coefficients

MFCC are one of the most widely used feature extraction techniques in speech and audio processing. This method aims to represent the speech signal in a form that is more closely aligned with human auditory perception. The core principle behind MFCC is to approximate the human ear's response to different frequencies by using a series of filters applied to the power spectrum of a speech signal. These filters are spaced according to the mel-scale, a perceptual scale that mimics the non-linear way humans perceive pitch. For feature extraction, the filter bank size is 40 and frame length is 2048. The process of extracting MFCCs has been shown in Fig. 1 and it typically involves the following steps. The signal x(t) is passed through a pre-emphasis filter to emphasize high frequencies:

$$\acute{x}(t) = x(t) - \alpha x(t-1) - (1)$$

where α is typically set to 0.95. The signal is split into overlapping frames. Each frame is denoted as x(n), where n corresponds to the discrete time index. A window function ω(n), such as the Hamming window, is applied to each frame to minimize special leakage:

$$x_{\omega}(n) = x(n).\omega(n) - (2)$$

where the Hamming window is given by:

$$\omega(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) - (3)$$

where N is the number of samples in each frame. The fast Fourier transform (FFT) is applied to each windowed frame to convert it from the from the time domain to the frequency domains:

$$x(f) = FFT(x_{\omega}(n)) - (4)$$

The power spectrum is mapped to Mel scape using triangular Mel filters. The Mel scale frequency fm is given by:

$$f_m = 2595\log_{10}\left(1 + \frac{f}{700}\right) - (5)$$

where f is the frequency in Hertz. The Mel filter bank Hm(f) is then applied to the magnitude spectrum |X(f)|, producing the Mel frequency spectrum. Logarithmic scaling is the logarithm of the Mel spectrum as given in (6):

$$\log(M_f) = \log(H_m(f)) - (6)$$

where M_f is the Mel-frequency spectrum at each frequency bin. Finally, the discrete cosine transform (DCT) is applied to the log Mel-spectrum to obtain the MFCCs

$$c_k = \sum_{m=1}^{M}\log(M_m)\cos\left[\frac{\pi}{M}\left(m - \frac{1}{2}\right)k\right], k = 1, 2, ..., K - (7)$$

where ck are the MFCCS, M is the number of Mel frequency bins, K is the number of MFCC coefficients often 12 or 13), Mm is the m-th Mel-frequency bin, K is the index for the efficient. MFCCs have been successfully used in various speech-related applications, including speech recognition, speaker identification, and emotion detection, due to their ability to capture the relevant features of speech while minimizing the influence of noise. They are effective in handling validations in speaker

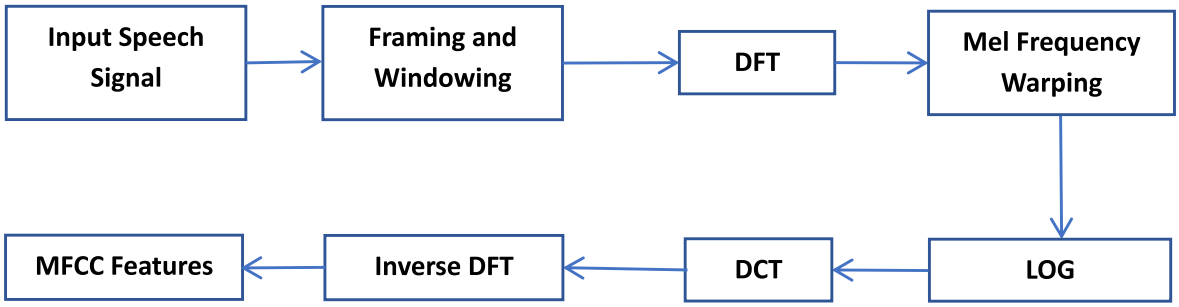

Fig. 1. Architecture of MFCC.

characteristics and environmental conditions, making them a robust feature for many audio classification tasks [16].

2.1.2. Complex quantization cepstral coefficient

Complex quantization cepstral coefficient (CQCC) is an advanced feature extraction method used for speech and audio signal processing, particularly for speaker recognition and speech emotion detection. CQCC is an enhancement of traditional cepstral feature extraction techniques such as MFCC, offering improved performance by better capturing phase information along with the magnitude spectrum of the signal. The CQCC method primarily focuses on the phase spectrum of the speech signal, which, although less emphasized in traditional methods, can provide valuable discriminative features. CQCC integrates both magnitude and phase information to create a more robust feature representation. This is achieved by applying a complex-valued approach to the quantization of the frequency components of the speech signal [18]. Fig. 2 shows the block diagram of obtaining the CQCC. The steps required for obtaining the CQCC are discussed as below.

Step 1: From the input speech signal, constant-Q transform (CQT) is obtained that will map the speech signal from time domain to frequency domain. To calculate CQT, the steps below are followed.

Step 1.1: The signal x(t) is passed through a pre-emphasis filter to enhance high frequencies:

$$x'(t) = x(t) - \alpha x(t-1) - (8)$$

where α is a constant typically set to 0.95, and x(t) is the original signal.

Step 1.2: The signal is divides into overlapping frames of length N, and a window function ω(n) such as a Hamming window is applied to each frame.

Step 1.3: The fast Fourier transform (FFT) is computed for each windowed frame:

$$X(k,n) = \sum_{m=0}^{N-1} x(n+m).e^{-km\frac{j2\pi}{N}}, k = 0,1,2,.....,n-1 - (9)$$

where X(k,n) is the complex-valued frequency bin at frame n, j is the imaginary unit, N is the number of samples in each frame.

Step 1.4: The spectrum |X(k,n)| and phase spectrum $\angle X(k,n)$ are obtained from the complex-valued FFT coefficients:

$$|X(k,n)| = \sqrt{\Re(X(k,n))^2 + \Im(X(k,n))^2} - (10)$$

$$\angle X(k,n) = \text{atan } 2(\Im(X(k,n)), \Re(X(k,n))) - (11)$$

Where $\Re(X(k,n))$ and $\Im(X(k,n))$ are the real and imaginary parts of X(k, n) respectively.

Step 1.5: A complex quantization technique is applied to the phase spectrum to capture the phase relationships. One possible method is to represent the quantize phase as:

$$\hat{\angle}X(k,n) = Q(\angle X(Kk,n)) - (12)$$

where Q is the quantization function that maps the continuous phase values to discrete values.

Step 2: Power spectrum of the signal is then calculated that describes the distribution of power into frequency components of that signal.

Step 3: The logarithm of the magnitude spectrum is computed to simulate the non-linear response of human hearing as given in (13).

$$\log|X(k,n)| = \log(|X(k,n)|) - (13)$$

Step 4: Both the log-magnitude and the quantized phase spectra are combined to form a joint representation called uniform resampling. The combined feature vector M(k,n) can be represented as:

$$M(k,n) = \log|X(k,n)| + \hat{\angle}X(k,n) - (14)$$

where M(k,n) represents both the magnitude and quantize phase

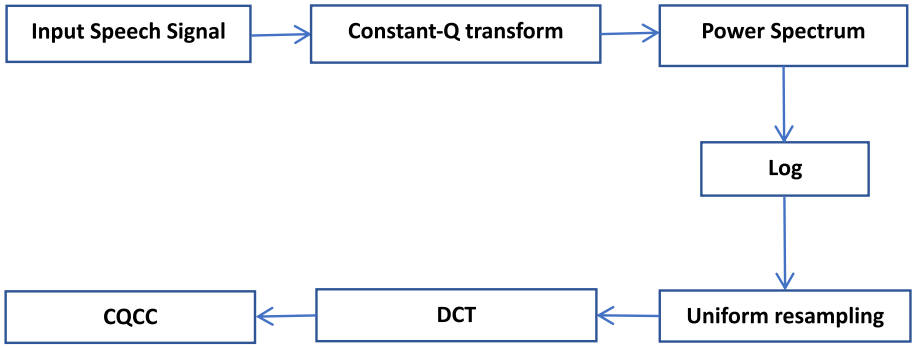

Fig. 2. Block diagram of obtaining CQCC.

information.

Step 5: The DCT is applied to the combined log-magnitude and phase information to obtain the CQCC. Let M(k,n) be the combined feature vector, the k-th CQCC coefficient at frame n is calculated as given in (15),

$$c_k(n) = \sum_{m=1}^{M} M_m(n).\cos\left[\frac{\pi}{M}\left(m-\frac{1}{2}\right)k\right], k = 1, 2, ..., k - (15)$$

where ck (n) is the k-th CQCC at frame n, Mm(n) is the combined feature vector at the m-th frequency bin, K is the number of CQCC coefficients (typically 12 or 13), M is the number of frequency bins used.

The general formula for extracting CQCC involves the following steps:

$$c_k(n) = DCT(\log|X(k,n)| + \widehat{\angle}X(k,n)) - (16)$$

CQCC based features has been shown to outperform traditional methods like MFCC based features in environments with noise or varying acoustic conditions. By incorporating phase information, CQCC offers enhanced robustness, particularly in challenging scenarios such as noisy recordings, low-quality microphones, and different speaker conditions. The approach has gained significant attention in the fields of speaker recognition, emotion detection, and other audio classification tasks [19].

2.1.3. Perceptual noise-cepstral coefficients

Perceptual noise-cepstral coefficients (PNCC) is a robust feature extraction technique used in speech and audio processing, designed to enhance speech recognition and speaker identification performance, especially in noisy environments. PNCC is an improvement over traditional features such as MFCCs by addressing the limitations of human perception and by incorporating aspects of noise modeling, making it particularly suitable for speech recognition systems working in challenging acoustic conditions [20]. The core idea behind PNCC is to enhance the robustness of speech features by modeling the perceptual properties of the human auditory system and accounting for the effects of noise. Unlike MFCC, which focuses primarily on capturing the spectral features of speech signals, PNCC incorporates a noise-robust pre--processing stage and models auditory masking effects, helping to distinguish speech from background noise more effectively. PNCC has been shown to outperform MFCC and other traditional feature extraction techniques in noisy environments, as it is specifically designed to preserve important speech features while minimizing the influence of noise. It has proven to be highly effective in automatic speech recognition (ASR) systems, especially in scenarios where speech is corrupted by background noise, reverberation, or other distortions [21]. Steps to obtain PNCC has been discussed here. The signal x(t) is passed through a pre-emphasis filter to enhance high frequencies. The STFT is then calculated from the signal. A perceptual filter bank is applied to the magnitude spectrum to simulate the frequency sensitivity of the human ear. This often uses a filter bank based on the Bark scale.

$$M_m(n) = \sum\nolimits_k H_m(k).|X(k,n)| - (17)$$

where Mm(n) is the output of the m-th filter at frame n, Hm(k) is the filter response for the m-th frequency band, |X(k,n)| is the magnitude of the FFT coefficients. Noise is filtered from the signal using techniques such as spectral subtraction or Wiener filtering, which helps remove unwanted noise components from the speech signal. A nonlinear transformation, often a logarithm function, is applied to the percectual features to model the nonlinear response of the human auditory system:

$$\widetilde{M}_m(n) = \log(M_m(n) + \epsilon) - (18)$$

where ϵ is a small constant added to avoid taking the logarithm of zero.

The DCT is applied to the log-transformed perceptual features to obtain the PNCC coefficients. The DCT is used to reduce the dimension and capture the most important features:

$$c_k(n) = \sum_{m=1}^{M} \widetilde{M}_m(n).\cos\left[\frac{\pi}{M}\left(m-\frac{1}{2}\right)k\right], k = 1, 2, ..., k - (19)$$

Where $c_k(n)$ is the k-th PNCC at frame n, $\widetilde{M}_m(n)$ is the transformed perceptual feature, M is the number of frequency bins, K is the number of PNCCs. The final PNCC coefficients are given by applying the DCT to the log-transformed perceptual features:

$$c_k(n) = DCT(\log(M_m(n))) - (20)$$

Where $M_m(n)$ is the perceptual feature extracted by the filter bank and noise filtering, and the DCT is applied to obtain the final PNCC coefficients.

2.1.4. Linear predictive coding

Linear predictive coding (LPC) is a fundamental technique in speech processing, primarily used for speech analysis, synthesis, and recognition. LPC is a method that models the speech signal as a linear combination of its past samples, capturing the spectral envelope of the signal efficiently. LPC is widely used due to its ability to represent speech features compactly, offering significant benefits in terms of storage, computational efficiency, and robust performance in speech-related applications. LPC works on the principle that speech signals can be approximated by predicting each sample based on a linear combination of previous samples. This prediction is achieved by minimizing the difference between the predicted and actual values of the signal. The method provides a set of coefficients, known as LPC coefficients, which represent the filter parameters that best approximate the speech signal [22,23].

2.1.5. Perceptual linear prediction and LPC

Perceptual linear prediction (PLP) is a feature extraction technique widely used in speech processing that aims to model the perceptual aspects of human hearing. It is a modification of LPC that incorporates the characteristics of human auditory perception, particularly the sensitivity of the human ear to certain frequencies and the non-linearities in how sound intensity is perceived. PLP is widely used in speech recognition, speaker identification, and other audio processing applications due to its ability to capture the perceptual features of speech more effectively than traditional LPC [24,25]. PLP enhances LPC by incorporating three main auditory characteristics. The human auditory system has a nonlinear response to different frequencies. High-frequency sounds are perceived with less resolution than low-frequency sounds. PLP applies a frequency warping function to simulate this non-linearity. The human ear processes sounds in “critical bands,” which are frequency bands where different sounds interfere with each other. In PLP, the spectrum is filtered using a set of critical band filters, mimicking the way the auditory system processes sound. The human ear exhibits a logarithmic response to intensity, meaning that a doubling of sound intensity is perceived as only a small increase in loudness. PLP applies a loudness compression function, often modeled using a cubic root law, to simulate this effect.

2.1.6. Relative spectral transform - perceptual linear prediction

Relative spectral transform - perceptual linear prediction (RASTA) is a powerful technique used in speech processing, particularly in speech recognition, to improve the robustness of speech features to distortions such as noise and channel variability. It enhances speech features by performing a spectral filtering process that emphasizes relative spectral changes over absolute spectral values, making it more sensitive to important speech information and less sensitive to variations in environmental conditions [26]. RASTA is based on the PLP feature

extraction method, with an additional spectral filtering step. The idea is to filter the log-energy speech features using a temporal filter that removes slow-changing, non-speech-related information (such as channel effects and noise), and preserves the fast-changing components that carry more relevant speech information. The key idea behind RASTA is to focus on the temporal dynamics of the speech signal, emphasizing relative spectral changes rather than absolute values. This filtering process helps to remove noise and channel distortions, which often vary slowly over time, while retaining the speech information, which typically exhibits faster temporal changes [27].

2.1.7. Wavelet packet transform

Wavelet packet transform (WPT) is an extension of the traditional wavelet transform (WT) used for signal processing, particularly in the analysis of non-stationary signals such as speech, audio, and biomedical signals. Unlike the discrete wavelet transform (DWT), which decomposes the signal into low and high-frequency components, WPT decomposes the signal at multiple levels of resolution, providing a more detailed representation of the signal's frequency content. This technique is particularly useful for tasks where precise frequency localization and time-frequency analysis are crucial [28]. Wavelet Packet Transform offers a more flexible and detailed decomposition of signals than the traditional wavelet transform by allowing both approximation and detail coefficients to be further decomposed at each level. This flexibility is beneficial for signal processing applications that require high precision in capturing both low-frequency and high-frequency components of the signal. WPT provides a multi-level hierarchical decomposition where each level decomposes the signal into a set of basis functions or wavelets, which can be finely tuned based on the specific frequency range of interest. This makes WPT well-suited for complex, non-stationary signals, such as speech signals with varying frequency characteristics over time [29].

2.2. For SLI detection

2.2.1. Temporal convolutional networks

Temporal convolutional networks (TCNs) are a type of deep learning architecture designed for sequence modeling tasks, combining the strengths of CNN with the ability to handle temporal dependencies. Unlike recurrent models like LSTM networks, TCNs rely on convolutional operations, which offer better parallelism and faster training while maintaining long-range dependency learning. The core architecture of a TCN is shown in Fig. 3 and the components includes causal convolutions, dilated convolutions and residual connections. Causal convolutions ensures that the output at any time step depends only on current and past inputs, preserving the temporal order. Dilated convolutions expands the receptive field of the network exponentially without increasing computational complexity, allowing TCN to capture long-range temporal dependencies efficiently. Residual connections facilitates deeper architectures by mitigating the vanishing gradient problem, improving model stability and training convergence. TCNs have shown excellent performance in various sequence modeling applications, such as time series forecasting, speech processing, and bioinformatics. For instance Ref. [30], demonstrated their effectiveness in action segmentation tasks, while [31] provided a comprehensive comparison of CNs with recurrent architectures, showcasing their superiority in terms of accuracy and computational efficiency.

2.2.2. TabNet

TabNet is a deep learning model specifically designed for tabular data, blending interpretable decision-making with the representational power of neural networks. It uses a sequential attention mechanism to select meaningful features at each decision step, thereby reducing redundancy and enhancing interpretability. Unlike traditional methods, TabNet employs sparse attention and learned masks to focus on relevant

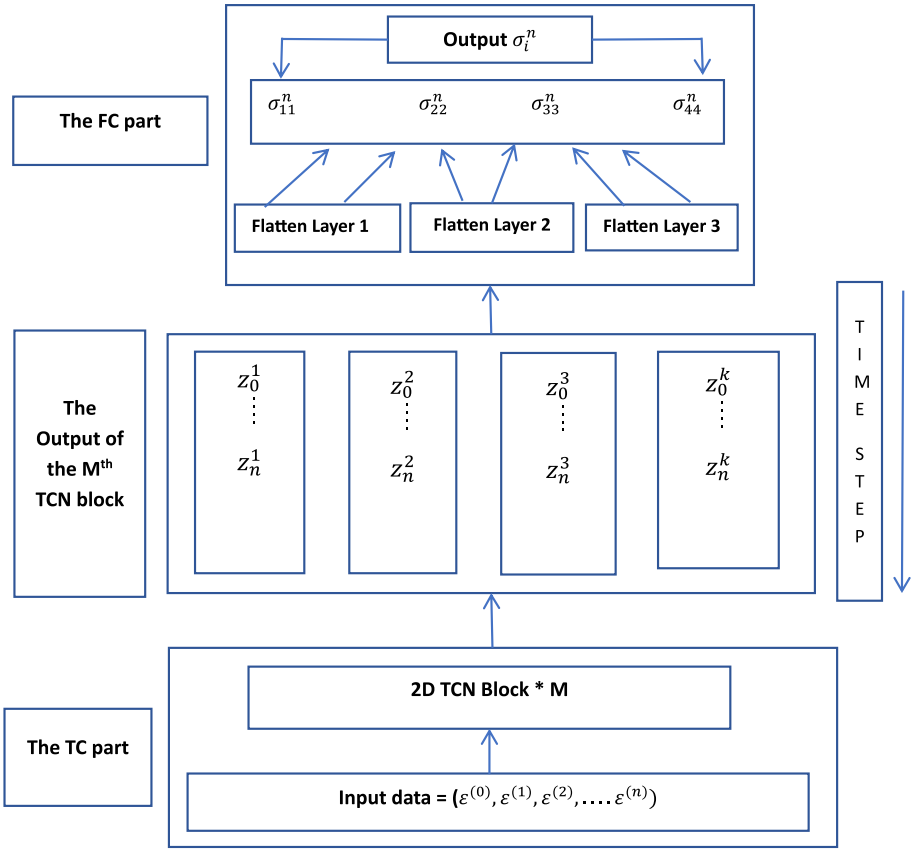

Fig. 3. Architecture of TCN.

features, enabling efficient learning and robust performance on high-dimensional datasets [32]. The architecture of TabNet consists of three main components: a feature transformer, an attentive transformer, and decision steps as shown in Fig. 4. The feature transformer pre-processes the input features, while the attentive transformer learns masks to focus on the most relevant information. Each decision step contributes to the final prediction, creating a balance between interpretability and accuracy. Additionally, TabNet integrates a unique regularization technique based on sparse entropy, promoting sparsity in feature selection. TabNet has shown competitive performance across various domains, often outperforming traditional methods like gradient boosting and random forests in tabular datasets. Its interpretability is particularly valuable for identifying the most critical features influencing predictions, making it an attractive choice for applications requiring explanation [33].

2.2.3. Transformer

Transformer is introduced by Vaswani et al. (2017), revolutionized sequence modeling by using only attention mechanisms, eliminating the need for recurrence or convolutions [11]. Fig. 5 shows the architecture of transformer. It employs a self-attention mechanism to model global dependencies in sequences, offering significant improvements in parallelization and long-range context capture compared to RNNs and CNNs. A typical transformer encoder consists of stacked layers, each with two main components: multi-head self-attention and a position-wise feed forward network. Each sub-layer uses residual connections followed by layer normalization. Multi-Head self-attention mechanism allows the model to focus on different positions of the sequence and capture various types of relationships in parallel. Feed-forward neural networks independently applied to each position for further transformation. To inject sequence information, positional encodings based on sine and cosine functions of different frequencies are added to the input embeddings. Recent studies have demonstrated the effectiveness of transformer models not only in natural language processing but also in various domains like audio classification, time-series forecasting.

3. Proposed method

The methodology involves extracting features from speech utterances using multiple established and novel techniques as shown in Fig. 6. These features will then be used to train transformer, TCN and TabNet models. The performance of each feature-model combination will be rigorously evaluated using standard metrics such as accuracy, precision, recall, and F-score. Additionally, it is aimed to analyze the interpretability of feature selection mechanism to provide insights into the most critical features for SLI detection. This study aims to establish a comprehensive benchmark for feature extraction method in SLI detection, ultimately contributing to the development of robust and interpretable diagnostic tools for doctors and therapist.

3.1. Inputs used

The LANNA speech data-set, developed by the laboratory of artificial neural network applications (LANNA) at the Czech technical university in Prague, is an important resource for detecting speech impairments. It consists of speech recordings from children aged 4–12 years, including 44 healthy children and 54 children diagnosed with SLI, totaling 3848 utterances. The data set includes various categories of speech such as vowels, consonants, one-to four-syllable words, and more complex words, all of which help assess speech difficulties in children [15]. These recordings were collected under the guidance of speech therapists and clinical psychologists from Motol University Hospital, ensuring high-quality data for analysis. For healthy children, recordings were made using Sony digital dictaphones at a 16 kHz sampling rate, while recordings for children with SLI were captured at 44.1 kHz using the same resolution. The data is stored in WAV format with a mono

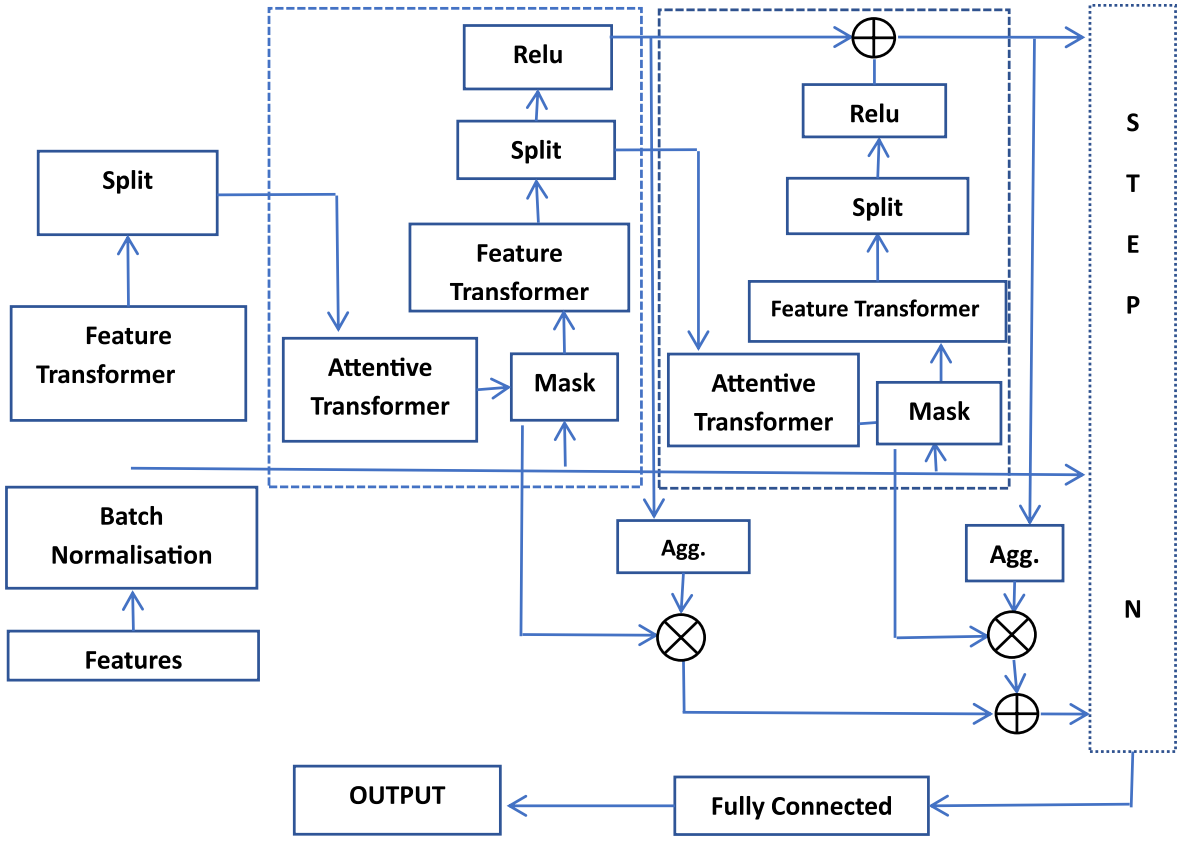

Fig. 4. Architecture of TabNet.

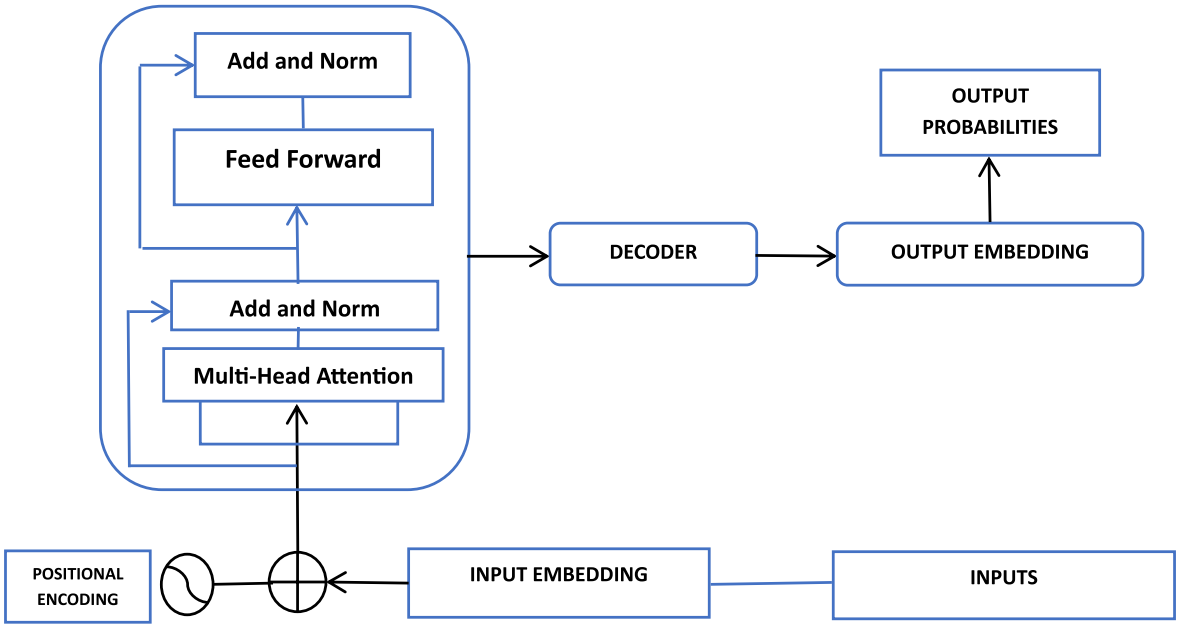

Fig. 5. Architecture of transformer.

configuration, ensuring compatibility with advanced speech analysis techniques.

The recordings were made in a controlled environment to reflect the natural speech behavior of children, with the recordings being done in classrooms or language therapist rooms to reduce environmental noise. Speech tasks were designed in collaboration with clinical psychologists and speech therapists, with a focus on mental and psychological development. The data-set primarily includes vowel, consonants, 1 syllables, two syllables, three syllables and difficult words which have been widely used in previous studies for pathological voice detection and SLI diagnosis. The proposed method have used all the form of speeches from vowel to difficult word. All the speeches are concatenated into one and given as input to the methods and the accuracy is computed on whole data. The sequence of combining various speeches does not have any impact on the method. The next step is to extract the features from the speech signals that is described in the next section.

3.2. Extracting features from speech signal

3.2.1. CQCC features

CQCC were employed to extract key features from speech signals to detect SLI. The signals were processed using the librosa library, which facilitated efficient management and transformation of the speech signals. Each speech file is used while preserving its original sampling rate to maintain its inherent characteristics. The CQCC coefficients are extracted from speech signals that effectively encapsulate the perceptually significant elements. The mean and standard deviation of the CQCC coefficients were computed which is the CQCC features. Fig. 7 shows the CQCC feature extraction process from a impaired speech. These CQCC features are then used as input to deep learning models to detect SLI.

3.2.2. PNCC features

PNCC is a robust feature extraction method that enhances noise resistance by emphasizing perceptually relevant components of speech signals. This makes it particularly suitable for audio classification tasks, such as distinguishing between healthy and patient speech recordings. From the logarithmic Mel-Spectrogram, MFCC were extracted to form the PNCC coefficients. To provide a fixed-length representation for each audio sample, the mean and standard deviation of the PNCC coefficients were computed. These statistical measures ensured that the features effectively summarized the temporal and spectral characteristics of the speech signals. This approach to feature extraction ensured the preservation of perceptually significant speech characteristics.

3.3. Speech language impairment detection

After extracting the features from speech signals, it has been given to different classifier for SLI detection. Two deep learning methods namely TCN and TabNet has been used for SLI detection in this work. Detail design of the SI detection method has been discussed below.

3.3.1. Using TCN method

For the task of detecting SLI, a TCN model have been utilized, taking advantage of its capability to capture long-range dependencies in sequential data. The data-set is divided into training and testing sets using an 80-20 split to assess the model's performance. To ensure the model's robustness against data scaling, feature normalization is applied. The TCN model was developed to capture the temporal dependencies present in speech signals. TCN comprised of 3 blocks of temporal convolutional layers (TCL), each utilizing different dilation rates (DR) of 1, 2, and 4 to enable the model to gather information across various time scales. Each TCL block has 2 convolutional layers, batch normalization, ReLU activation and dropout layer. A residual connection was implemented between each TCL block to facilitate deeper learning. The output layer consisted of a fully connected dense layer with a sig-moid activation function. The model uses Adam optimizer, binary cross-entropy loss and was trained for 50 epochs with a batch size of 32 and kernel size 3. Table 1 shows the training performance of method with different features used. Fig. 8 shows the training accuracy and loss performance of the method.

3.3.2. Using TabNet method

In this study, a deep learning architecture called TabNet has been used. The TabNet classifier was selected for its exceptional performance with tabular data, utilizing attention mechanisms to prioritize the most significant features. The TabNet consist of batch normalization layer, feature transformer (GLU blocks), attentive transformer, attention mask, TabNet decoder (feature transformer, followed by the fully connected layers) and output layer. The model was set up with a learning rate of

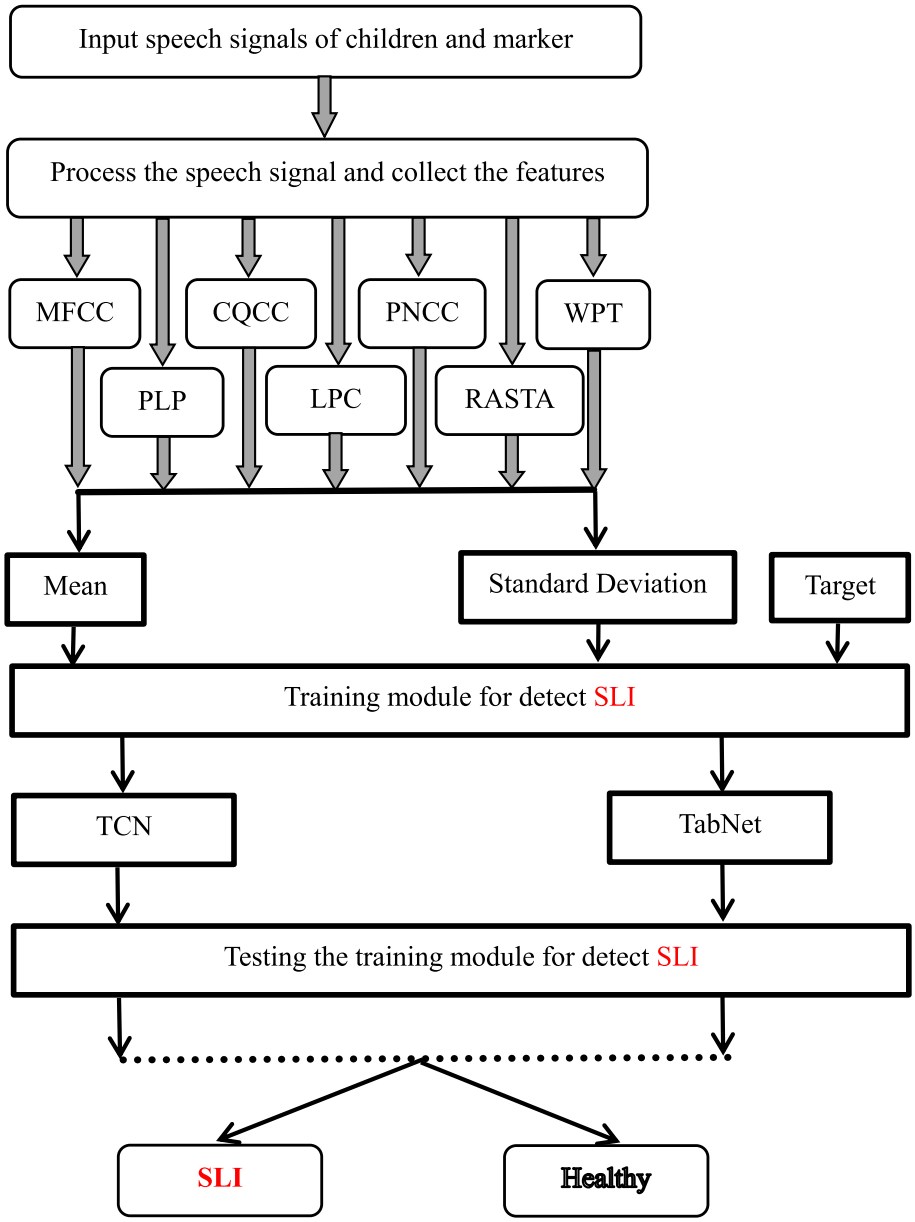

Fig. 6. Flowchart of the SLI detection method.

0.02, a batch size of 256, and a virtual batch size of 128. The activation function used is sig-moid and the optimizer used is Adam. These hyper-parameters were chosen to optimize both memory usage and computational efficiency while enhancing model performance. Training was carried out for a maximum of 200 epochs, with early stopping implemented to halt the training if there was no improvement in validation accuracy for 30 consecutive epochs, thereby preventing over-fitting. These parameters work together to ensure that the model is trained efficiently, with good performance, and evaluated properly. Table 1 shows the training performance of method with different features used. Fig. 7 shows the training accuracy and loss performance of the method.

3.3.3. *Using transformer*

A deep learning architecture called transformer also has been used for SLI detection. The model was set up with a batch size of 256. The activation function used is soft-max and the optimizer used is Adam. Training was carried out for a maximum of 200 epochs, with early stopping implemented to halt the training. Table 1 shows the training performance of method with different features used. Fig. 7 shows the training accuracy and loss performance of the method.

3.4. *Testing of the SLI detection module*

After the models are designed for SLI detection, it has been tested with new speech signals. To assess the model's performance, various metrics were employed, including accuracy, precision, recall, F-score, and AUC (area under the curve).

4. Result

The proposed SLI detection method has been implemented using python in i7,32 GB RAM computer in google colab. The performance has been measured in terms of precision, recall, F-score, AUC and accuracy. The results of the proposed SI detection method has been discussed in this section.

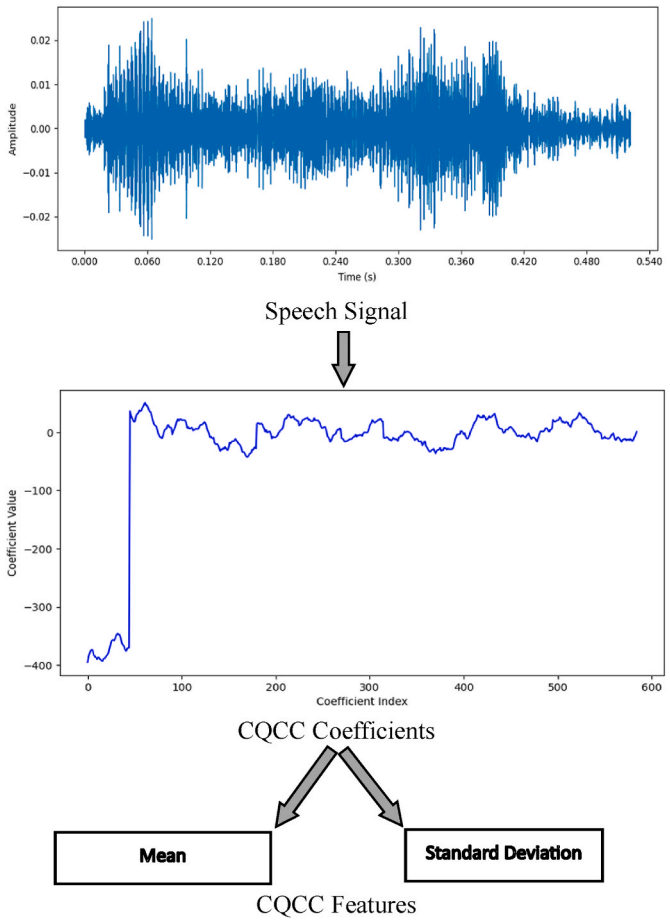

Fig. 7. Feature extraction process of CQCC feature.

4.1. Performance with WPT feature

The performance of the models was evaluated using WPT features for SLI detection. The results of the SLI detection with WPT features has been shown in Table 2. The transformer model has an accuracy of 70 % which is lower than other two methods. The TCN model achieved a precision of 0.81, meaning that it had a moderate rate of false positives. Its recall of 0.82 suggests that it correctly identified a large portion of relevant instances (true positives), but there were still some false negatives. The F-score of 0.82 indicates a fairly balanced trade-off between precision and recall. The accuracy of 76.15 % shows that TCN classified a substantial portion of the data correctly, though the performance is lower than other feature sets. TabNet performed slightly better than TCN in terms of precision (0.85), meaning that it had a lower rate of false positives. However, its recall of 0.80 indicates that it identified a slightly lower portion of relevant instances compared to TCN. The F-score of 0.82 reflects a balanced trade-off between precision and recall, but still below the optimal levels achieved with other features. The accuracy of 77.30 % shows that TabNet outperformed TCN slightly in terms of overall classification, but the performance is still moderate. Fig. 9 shows AUC of both methods. TabNet slightly outperformed TCN in terms of precision and accuracy, while TCN had a slightly higher recall. The overall performance suggests that WPT features, while useful, may not be as discriminative for this classification task compared to other features. Further exploration of alternative feature extraction methods could potentially lead to better results.

4.2. Performance with RASTA feature

The performance of the models was evaluated using RASTA features.

Table 1
Performance of the training method.

Method Used	Features used	Parameters used	Training accuracy (%)
TabNet	WPT	Maximum epochs = 200, optimizer = adam, activation = sigmoid	77.70 %
	RASTA	Maximum epochs = 200, optimizer = adam, activation = sigmoid	73.06 %
	PLP	Maximum epochs = 200, optimizer = adam, activation = sigmoid	77.29 %
	LPC	Maximum epochs = 200, optimizer = adam, activation = sigmoid	90.99 %
	CQCC	Maximum epochs = 200, optimizer = adam, activation = sigmoid	99.67 %
	MFCC	Maximum epochs = 200, optimizer = adam, activation = sigmoid	100.0 %
	PNCC	Maximum epochs = 200, optimizer = adam, activation = sigmoid	99.96 %
TCN	WPT	Epochs = 50, activation = sigmoid, optimizer = adam	77.58 %
	RASTA	Epochs = 50, activation = sigmoid, optimizer = adam	78.12 %
	PLP	Epochs = 50, activation = sigmoid, optimizer = adam	89.80 %
	LPC	Epochs = 50, activation = sigmoid, optimizer = adam	79.76 %
	CQCC	Epochs = 50, activation = sigmoid, optimizer = adam	99.84 %
	MFCC	Epochs = 50, activation = sigmoid, optimizer = adam	99.92 %
	PNCC	Epochs = 50, activation = sigmoid, optimizer = adam	100.00 %
Transformer	WPT	Epochs = 200, Optimizer = Adam, Activation function = soft-max	70.00 %
	RASTA	Epochs = 200, Optimizer = Adam, Activation function = soft-max	66.00 %
	PLP	Epochs = 200, Optimizer = Adam, Activation function = soft-max	74.00 %
	LPC	Epochs = 200, Optimizer = Adam, Activation function = soft-max	65.00 %
	CQCC	Epochs = 200, Optimizer = Adam, Activation function = soft-max	79.00 %
	MFCC	Epochs = 200, Optimizer = Adam, Activation function = soft-max	66.00 %
	PNCC	Epochs = 200, Optimizer = Adam, Activation function = soft-max	66.00 %

The accuracy of transformer method is found to be 66 %, TCN method is found to be 74.51 % and TabNet method is found to be 77.14 %. The transformer model achieved a precision of 0.33, recall of 0.50 and F-score of 0.40. The results of the TCN and TabNet method are shown in Fig. 10. The TCN model achieved a precision of 0.79, recall of 0.83 and F-score of 0.81. TabNet performed slightly better than TCN having a precision of 0.84, recall of 0.81 and F-score of 0.82. It shows that TabNet performed better than TCN but still lower than other models evaluated with different features. The overall performance indicates that RASTA features may not be as effective.

4.3. Performance with PLP feature

The performance of the method has been evaluated using PLP features and the results are shown in Table 3. Transformer model achieved a precision of 0.871, recall of 0.67, F-score of 0.68 and accuracy of 74 %. TCN model achieved a precision of 0.81, recall of 0.91, F-score of 0.85 and accuracy of 79.93 %. TabNet's performance is similar to TCN, with a precision of 0.82 and recall of 0.86. The F-score of 0.84 shows a

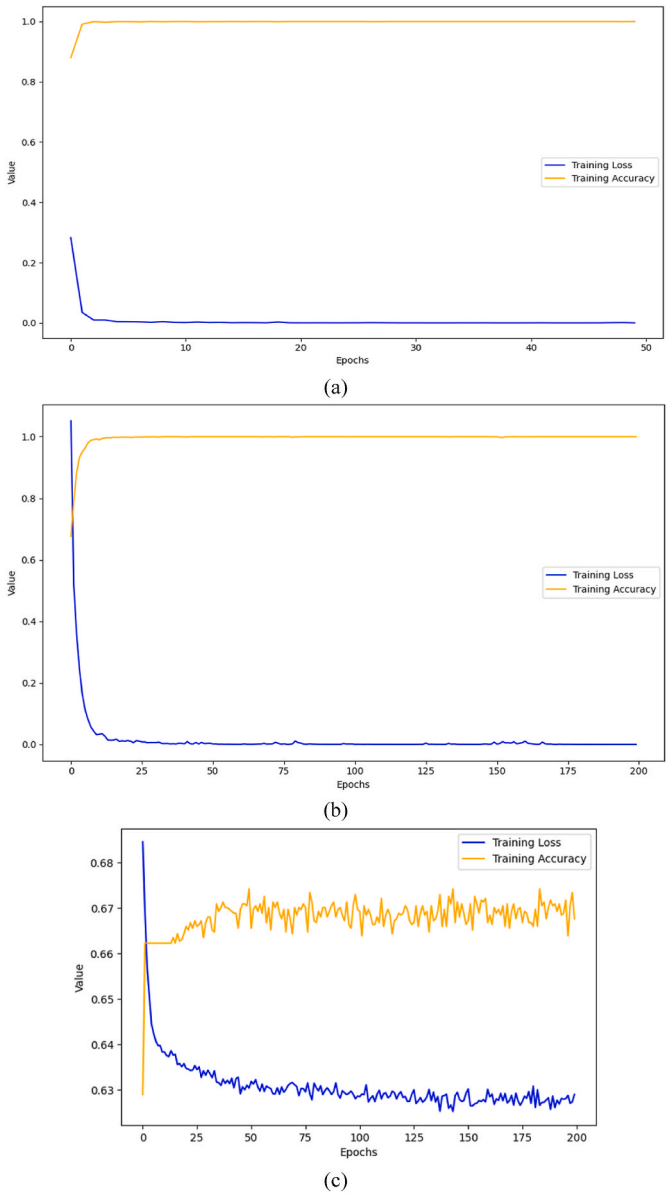

Fig. 8. Training performance (a) PNCC - TCN (b) PNCC - TabNet (c) PNCC - Transformer.

Table 2
Performance with WPT features.

Technique used	Precision	Recall	F-Score	Accuracy
TCN	0.81	0.82	0.82	76.15 %
Tabnet	0.85	0.80	0.82	77.30 %
Transformer	0.68	0.58	0.57	70.00 %

reasonable balance, although it is still lower than TCN's performance. The accuracy of 78.29 % of TabNet is slightly lower than TCN's, indicating that TabNet performed just a bit less effectively with PLP features.

4.4. Performance with LPC feature

The performance of the models was evaluated using LPC features. The transformer model achieved a precision of 0.50, recall of 0.50, F-score of 0.44. The TCN model achieved a precision of 0.91, recall of 0.93, F-score of 0.92. TabNet's has a precision of 0.89, recall of 0.94, F-score of 0.91. The accuracy of transformer is 65 %, TCN is 89.97 % and TabNet is 88.32 %. The results of TCN and TabNet method are shown in Fig. 11. The result suggests that LPC features may not be as effective in capturing the necessary information and alternative feature extraction methods might yield better results.

4.5. Performance with CQCC feature

The performance of the method have been evaluated using CQCC features for the SLI detection task. The results of the method are shown in Table 4. Transformer model achieved precision of 0.77, recall of 0.75, F-score of 0.76 and accuracy of 79 %. TabNet achieved precision of 0.99, recall of 1.00, F-score of 0.99 and accuracy of 99.01 %. Fig. 12(a) shows the AUC plot of Transformer model and Fig. 10(b) shows the AUC plot of TabNet. The accuracy of TCN method is 99.51 % further confirms that the model performed at a near-optimal level. Both TCN and TabNet models exhibit top-tier performance, making CQCC features a valuable tool for SLI detection.

4.6. Performance with MFCC feature

The performance of the method was evaluated using MFCC features and the results are shown in Table 5. The results of TCN and TabNet method indicate that both models achieved outstanding performance. The accuracy of 99.84 % further reinforces this result, demonstrating that TCN was able to detect SLI correctly. Fig. 13(a) shows the AUC plot of TCN and Fig. 13(b) shows the AUC plot of TabNet. The success of both models further suggests that MFCC features provide a robust representation of the data, making them a powerful tool for SLI detection.

4.7. Performance with PNCC feature

The performance of the method has been evaluated using PNCC features and the results are shown in Table 6. The TCN model achieved a perfect precision, recall, and F-score of 1.00, indicating that it was able to correctly classify all instances with no false positives or false negatives. The accuracy of 99.84 % further confirms this result, demonstrating that TCN performed at a near-optimal level in classifying the data. TabNet achieved flawless performance, with precision, recall, F-score, and accuracy all reaching 1.00 or 100 %. Fig. 14(a) shows the AUC plot of TCN and Fig. 14(b) shows the AUC plot of TabNet. The perfect or near-perfect results reflect the robustness of PNCC features, which effectively captured the necessary information for the task.

5. Discussion

In this work, different techniques has been used to extract features from speech signal for accurate detection of SLI. The results obtained from different features has been given in the previous section. Analysis of the results with different features has been discussed here.

5.1. Performance of the SLI detection method in absence of noise

The performance of the SLI detection method with various features in absence of noise using TabNet classifier has been recorded in Table 7. It can be observed that the MFCC features and PNCC features with TabNet have highest accuracy among all the features. The precision, recall and F-score is also high for both features. Further study is required to decide the optimal feature for SLI detection. In the next section, the proposed method has been tested in presence of noise in the signal to decide the optimal feature extraction method.

5.2. Performance of the SLI detection method in presence of noise

The SLI detection method has been tested in presence of noise in the signals. A 20 db noise has been added to the signals and then features are

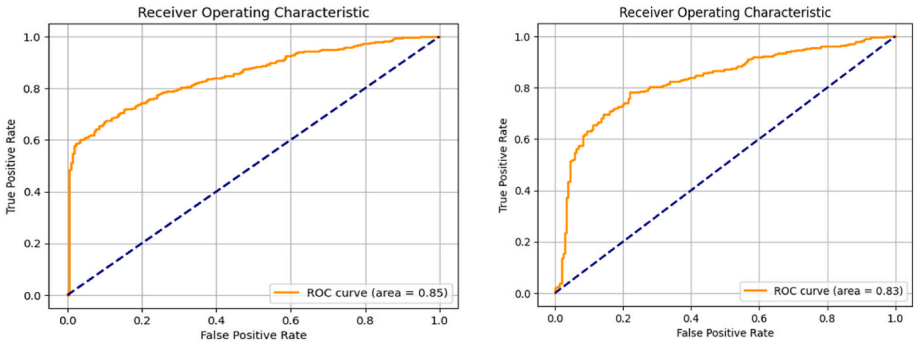

Fig. 9. AUC of both methods (a) AUC plot of TCN (b) AUC plot of TabNet.

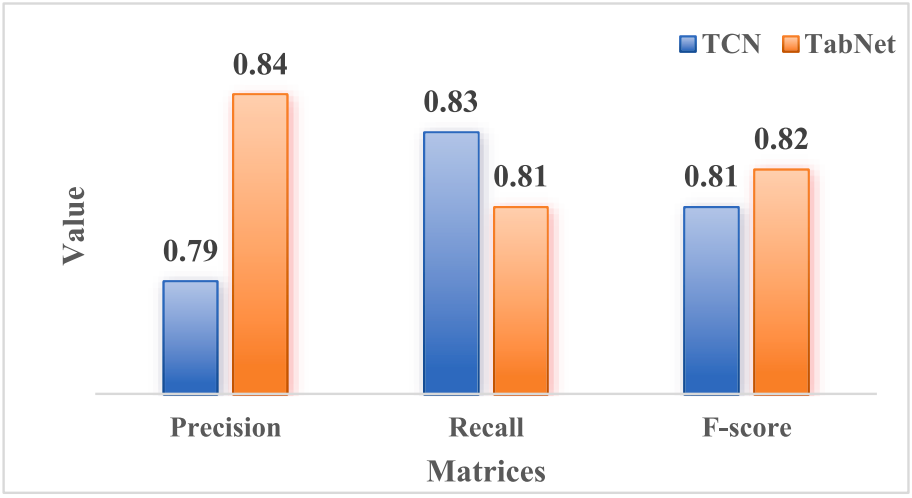

Fig. 10. Performance with RASTA features.

Table 3
Performance with PLP features.

Technique used	Precision	Recall	F-Score	Accuracy
TCN	0.81	0.91	0.85	79.93 %
Tabnet	0.82	0.86	0.84	78.29 %
Transformer	0.71	0.67	0.68	74.00 %

Table 4
Performance with CQCC features.

Technique used	Precision	Recall	F-Score	Accuracy
TCN	0.99	1.00	1.00	99.51 %
Tabnet	0.99	1.00	0.99	99.01 %
Transformer	0.77	0.75	0.76	79.00 %

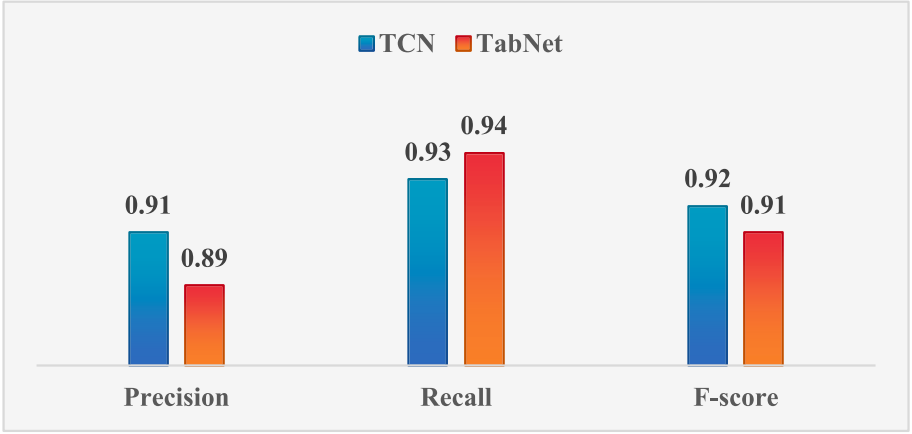

Fig. 11. Performance with LPC features.

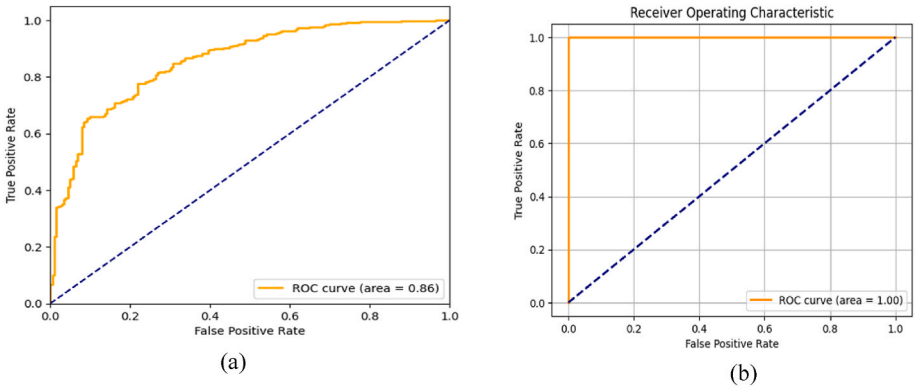

Fig. 12. AUC of both methods (a) AUC plot of Transformer (b) AUC plot of TabNet.

Table 5
Performance with MFCC features.

Technique used	Precision	Recall	F-Score	Accuracy
TCN	1.00	1.00	1.00	99.84 %
Tabnet	1.00	1.00	1.00	100.0 %
Transformer	0.58	0.51	0.43	66.00 %

extracted from the signals. The features are then tested with the TabNet based SLI detection method. The results of the method are shown in Table 8. It can be observed that MFCC based method is affected more to noisy signal than PNCC based method. The reason for this may be due to that fact that MFCC uses Mel filter bank and PNCC uses Gammatone filter bank. The MFCC uses logarithmic non-linearity and PNCC uses power law non-linearity. PNCC has more accuracy in presence of noise in signal but MFCC accuracy have decreased significantly. Hence, PNCC features can be chosen as optimal feature for SLI detection.

5.3. Validation with another data-set

The proposed SLI detection method has been validated with another data-set namely TORGO data-set [35,36]. It contains dysarthria speech disorder patients speech recordings. Dysarthria happens due to weakness of muscles used for speaking and causes slurred or slow speech that can be difficult to understand. It contains 2995 dysarthria speech and 5219 non-dysarthria speech. It has 6 female (3 normal and 3 dysarthria) and 9 male (5 normal and 4 dysarthria) individuals speech recording. The proposed method has been tested and the results are noted in Table 9. TabNet model perform better than transformer model and TCN model with every feature.TabNet has highest accuracy of 100 % with PNCC features.

5.4. Validation with paired t-test

To carry out the paired *t*-test, the average difference, the standard deviation of the difference and the sample size are required. The average score difference is 0.2, the standard deviation 0.979 and sample size is 10. Next, standard error for the score difference is calculated $= S\sqrt{n} = 0.979\sqrt{10} = 3.098$. Test statistic as t = Average difference/Standard Error $= 0.2/3.098 = 0.064$. The significance level, denoted by α, is set to 0.05. The degrees of freedom are based on the sample size is 9. Comparing the value of the statistic (0.064) to the t value. Because $0.064 < 16.912$, it cannot reject the idea that the PNCC feature based method is superior.

5.5. Comparison with other method

The comparison of various existing techniques with proposed SLI detection method has been shown in Table 10 considering different criteria. All the method compared has used the same data-set i.e. LANNA data-set. The advantage of the proposed method over other existing method can be outlined as follows:

I. Most of the method have not considered the whole data set for the validation of the method while proposed method has been validated with entire data-set that includes vowels, consonants, syllables, and challenging words.

Table 6
Performance with PNCC features.

Technique used	Precision	Recall	F-Score	Accuracy
TCN	1.00	1.00	1.00	99.84 %
Tabnet	1.00	1.00	1.00	100.0 %
Transformer	0.58	0.51	0.43	66.00 %

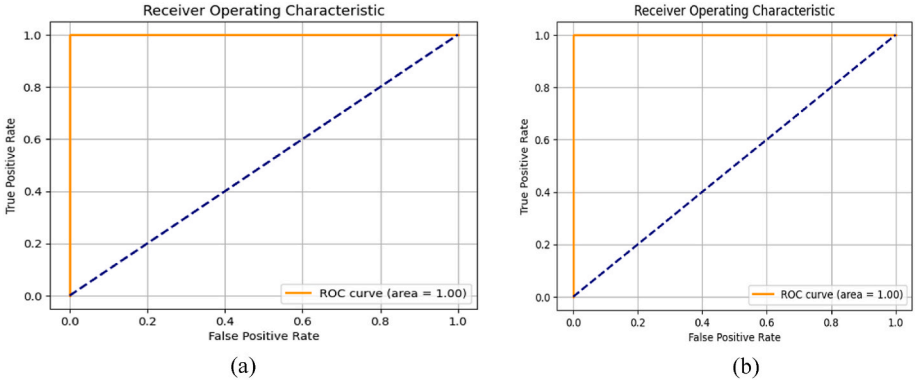

Fig. 13. AUC of both methods (a) AUC plot of TCN (b) AUC plot of TabNet.

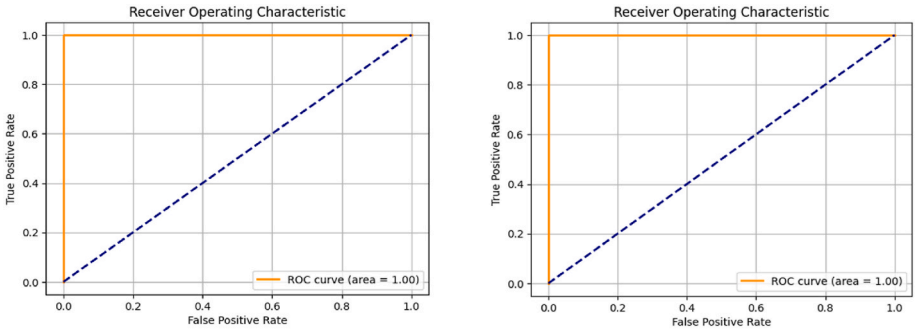

Fig. 14. AUC of both methods (a) AUC plot of TCN (b) AUC plot of TabNet.

Table 7
Performance of the method with different features used.

Features used	Precision	Recall	F-Score	Accuracy
RASTA	0.84	0.81	0.82	77.14 %
WPT	0.85	0.80	0.82	77.30 %
PLP	0.81	0.91	0.85	79.93 %
LPC	0.91	0.93	0.92	89.97 %
CQCC	0.99	1.00	1.00	99.51 %
MFCC	1.00	1.00	1.00	100.0 %
PNCC	1.00	1.00	1.00	100.0 %

Table 8
Performance of the method in presence of noise.

Speech Features Used	Precision	Recall	Specificity	F-score	AUC	Accuracy
MFCC	0.91	0.91	0.92	0.91	0.92	91.50 %
PNCC	0.97	0.98	0.97	0.97	0.99	98.50 %

II. The proposed method uses different features extraction method to explore an optimal feature extraction method for detecting SLI.
III. The proposed method suggested that PNCC feature would be a better approach to extract feature that has an accuracy of 100 % in detecting SLI without noise in signal and 98.5 % accurate with noise in signal.

Table 9
Performance with TORGO data-set.

Models used	Features used	Precision	Recall	F-Score	Accuracy
Transformer	RASTA	0.72	0.70	0.69	70.54 %
	WPT	0.67	0.59	0.55	59.69 %
	PLP	0.75	0.74	0.74	74.42 %
	LPC	0.68	0.67	0.67	67.44 %
	CQCC	0.64	0.64	0.64	64.34 %
	MFCC	0.78	0.78	0.78	78.29 %
	PNCC	0.75	0.75	0.75	75.19 %
TCN	RASTA	0.89	0.93	0.91	91.47 %
	WPT	1.00	1.00	1.00	100.0 %
	PLP	1.00	0.95	0.97	97.67 %
	LPC	0.96	0.88	0.92	93.02 %
	CQCC	0.98	0.98	0.98	98.45 %
	MFCC	0.99	1.00	0.99	99.99 %
	PNCC	1.00	1.00	1.00	100.0 %
TabNet	RASTA	0.97	0.94	0.95	95.35 %
	WPT	0.95	0.95	0.95	95.35 %
	PLP	1.00	0.97	0.98	98.45 %
	LPC	0.95	0. 92	0.94	93.80 %
	CQCC	0.99	1.00	0.99	99.99 %
	MFCC	0.99	1.00	0.99	99.99 %
	PNCC	1.00	1.00	1.00	100.0 %

Table 10
Comparison with other methods.

Authors	Features used	Techniques Used	Performance
Reddy et al. [7]	Glottal features and MFCC	Feed-forward Neural Network	Accuracy - 98.82 %
Sharma et al. [10]	Local binary patterns	SVMs, k-NN Complex Tree	Accuracy - 97.36 %
Sharma et al. [11]	CNN	Hybrid CNN-LSTM model	Accuracy −100.00 %
Barua et al. [12]	Favipiravir pattern, statistical feature extractor, and wavelet packet decomposition	SVM classifier	Accuracy 99.87 %
Proposed Method	PNCC features	TabNet	Accuracy - 100.00 % without noise Accuracy - 98.50 % with noise

IV. It can work with noisy data as the PNCC method is not affected by noisy signals while most method taken MFCC to process the signal.

6. Conclusion

In this work, a comparative study of various feature extraction method has been carried out to find an optimal method for accurate SLI detection. The accuracy of most of the feature extraction techniques demonstrated that it can used be effectively by doctors in SLI detection. Out of all the feature extraction techniques used, PNCC features can be chosen as optimal for SLI detection due to its perfect accuracy and work with noise signals also. The experimental results showed that combination of PNCC features with TabNet classifier performs better than the traditional methods combining all speech recordings. The proposed method contribute significantly to decide which features should be chosen from speech for accurate detection of SLI. The suggested method can be used effectively to detect SLI by speech pathologists and doctors. In future work, different techniques will be implemented to determine the speech age of the child so that the development of speech in children can be monitored effectively.

CRediT authorship contribution statement

Manisa Manoswini: Validation, Methodology, Software, Investigation, Writing – original draft, Resources, Data curation. **Biswajit Sahoo:** Writing – review & editing, Resources, Visualization, Project administration, Supervision. **Aleena Swetapadma:** Writing – review & editing, Validation, Investigation, Writing – original draft, Supervision, Formal

analysis, Visualization, Project administration, Conceptualization.

Ethical Statement for solid state ionics

Hereby, I Dr. Aleena Swetapadma consciously assure that for the manuscript "A novel speech signal feature extraction technique to detect speech impairment in children accurately" the following is fulfilled:

1) This material is the authors' own original work, which has not been previously published elsewhere.
2) The paper is not currently being considered for publication elsewhere.
3) The paper reflects the authors' own research and analysis in a truthful and complete manner.
4) The paper properly credits the meaningful contributions of co-authors and co-researchers.
5) The results are appropriately placed in the context of prior and existing research.
6) All sources used are properly disclosed.
7) All authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.

I agree with the above statements and declare that this submission follows the policies of Computers in Biology and Medicine as outlined in the Guide for Authors and in the Ethical Statement.

Declaration of competing interest

Authors do not have any competing interest.

References

[1] D.B. Fry, The Physics of Speech, Cambridge University Press, 1979.
[2] L.R. Rabiner, R.W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, 1987, 1978.
[3] J. Makhoul, Linear prediction: a tutorial review, Proc. IEEE 63 (1975) 561–580.
[4] T.F. Quatieri, Discrete-Time Speech Signal Processing: Principles and Practice, Prentice Hall, 2002.
[5] D. Ramarao, C. Singh, S. Shahnawazuddin, N. Adiga, G. Pradhan, Detecting developmental dysphasia in children using speech data, International Conference on Signal Processing and Communications (SPCOM) (2018) 100–104. Bangalore, India.
[6] Y. Sharma, B.K. Singh, Classification of children with specific language impairment using pitch-based parameters, IEEE Recent Advances in Intelligent Computational Systems (RAICS), Thiruvananthapuram, India (2020) 42–46.
[7] M.K. Reddy, P. Alku, K.S. Rao, Detection of specific language impairment in children using glottal source features, IEEE Access 8 (2020) 15273–15279.
[8] S. Safdar, S. Kausar, S. Tehsin, M. Mahmood, G.N. Alwakid, Prediction of specific language impairment in children using cepstral domain coefficients. International Conference on Business Analytics for Technology and Security (ICBATS), Dubai, United Arab Emirates, 2023, pp. 1–11.
[9] G. Sharma, D. Prasad, K. Umapathy, S. Krishnan, Screening and analysis of specific language impairment in young children by analyzing the textures of speech signal, Annu Int Conf IEEE Eng Med Biol Soc (2020) 964–967.
[10] G. Sharma, X.P. Zhang, K. Umapathy, S. Krishnan, Audio texture and age-wise analysis of disordered speech in children having specific language impairment, Biomed. Signal Process Control 66 (2021) 102471.
[11] Y. Sharma, B.K. Singh, One-dimensional convolutional neural network and hybrid deep-learning paradigm for classification of specific language impaired children using their speech, Comput. Methods Progr. Biomed. 213 (2022) 106487.
[12] P.D. Barua, E. Aydemir, S. Dogan, et al., Novel favipiravir pattern-based learning model for automated detection of specific language impairment disorder using vowels, Neural Comput. Appl. 35 (2023) 6065–6077.
[13] J. K, N. Deepa, Children specifically language impairment severity level prediction using improved conditional random fields and comparison with traditional models. 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM), 2023, pp. 1–6. Uttar Pradesh, India.
[14] D. Zhao, Z. Qiu, Y. Jiang, X. Zhu, X. Zhang, Z. Tao, A depthwise separable CNN-based interpretable feature extraction network for automatic pathological voice detection, Biomed. Signal Process Control 88 (2024) 105624.
[15] P. Grill, J. Tučková, Speech databases of typical children and children with SLI, PLoS One 11 (2016) e0150365.
[16] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust. Speech Signal Process. 28 (1980) 357–366.
[18] Z. Li, X. He, H. Xu, Complex quantization cepstral coefficients for robust speech recognition, IEEE Trans. Audio Speech Lang. Process. 30 (2022) 529–541.
[19] Z. Zhao, H. Zhang, F. Yang, Enhancement of speaker recognition systems using complex quantization cepstral coefficients, Speech Commun. 136 (2023) 51–64.
[20] Y. Chien, J. Lin, Enhancing speech recognition systems with perceptual noise-cepstral coefficients (PNCC), J. Acoust. Soc. Am. 151 (2022) 2247–2259.
[21] H. Lee, B. Kim, Noise-resilient speech recognition using PNCC features: a comparative study, Speech Commun. 128 (2021) 25–40.
[22] M. Rao, P. Gupta, Enhancing robustness of speech synthesis with LPC-based feature extraction, J. Acoust. Soc. Am. 152 (2022) 45–59.
[23] J. Li, L. Yang, Speech signal enhancement using LPC and deep neural networks for noise-resilient systems, Speech Commun. 125 (2021) 77–89.
[24] X. Huang, H. Yang, Comparing PLP and MFCC for deep learning-based speech emotion recognition, J. Acoust. Soc. Am. 152 (2022) 234–245.
[25] T. Li, Y. Wang, PLP and spectral features for speaker identification: a comparative analysis, Speech Commun. 131 (2021) 102–113.
[26] P. Miller, W. Zhang, A comparative study of RASTA and MFCC features in noisy speech recognition, J. Acoust. Soc. Am. 151 (2022) 2734–2745.
[27] R. Kumar, P. Gupta, RASTA-based feature extraction for noise-resilient speaker identification, Speech Commun. 130 (2021) 58–72.
[28] L. Wu, W. Yang, A wavelet packet transform-based approach for robust speech emotion recognition, IEEE Trans. Audio Speech Lang. Process. 31 (2023) 523–535.
[29] J. Zhang, L. He, Wavelet packet transform for time-frequency analysis of EEG signals in brain-computer interface systems, J. Neurosci. Methods 397 (2022) 108104.
[30] S. Bai, J.Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, arXiv preprint arXiv:1803.01271 (2018).
[31] C. Lea, M.D. Flynn, R. Vidal, A. Reiter, G.D. Hager, Temporal convolutional networks for action segmentation and detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 156–165.
[32] S.O. Arik, T. Pfister, TabNet: attentive interpretable tabular learning, Proc. AAAI Conf. Artif. Intell. 35 (2021) 6679–6687.
[33] H. Zhou, Y. Liu, J. Li, Enhanced tabular learning with TabNet for medical diagnostics, J. Mach. Learn. Res. 24 (2023) 110–124.
[35] F. Rudzicz, A.K. Namasivayam, T. Wolff, The TORGO database of acoustic and articulatory speech from speakers with dysarthria, Comput. Humanit. 46 (2012) 523–541.
[36] F. Rudzicz, Using articulatory likelihoods in the recognition of dysarthric speech, Speech Commun. 54 (2012) 430–444.