JTIM : Jurnal Teknologi Informasi dan Multimedia

p-ISSN : 2715-2529
e-ISSN : 2684-9151
https://journal.sekawan-org.id/index.php/jtim

Wavelet-Based MFCC and CNN Framework for Automatic Detection of Cleft Speech Disorders

Muhammad Hilmy Herdiansyah [1], Syahroni Hidayat[2] and Nur Iksan[2]

1 Department of Computer Engineering, Universitas Negeri Semarang, Indonesia
2 Department of Electrical Engineering, Universitas Negeri Semarang, Indonesia

* Correspondence: hilmyherdiansyah@students.unnes.ac.id

Citation: Herdiansyah, M. H.; Hidayat, S.; Iksan, N. (**2025**). Wavelet-Based MFCC and CNN Framework for Automatic Detection of Cleft Speech Disorders. JTIM: Jurnal Teknologi Informasi Dan Multimedia, 7(3), 652-663. https://doi.org/10.35746/jtim.v7i3.780

Received: 16-06-2025
Revised: 14-08-2025
Accepted: 21-08-2025

Abstract: Cleft Lip and Palate (CLP) is a congenital condition that often results in atypical speech articulation, making automatic recognition of CLP speech a challenging task. This study proposes a deep learning-based classification system using Convolutional Neural Networks (CNN) and Wavelet-MFCC features to distinguish speech patterns produced by CLP individuals. Specifically, we investigate the use of two wavelet families Reverse Biorthogonal (rbio1.1) and Biorthogonal (bior1.1)—with three decomposition strategies: single-level (L1), two-level (L2), and a combined level (L1+2). Speech data were collected from 10 CLP patients, each pronouncing nine selected Indonesian words ten times, resulting in 900 utterances. The audio signals were processed using wavelet-based decomposition followed by Mel-Frequency Cepstral Coefficients (MFCC) extraction to generate time-frequency representations of speech. The resulting features were input into a CNN model and evaluated using 5-fold cross-validation. Experimental results show that the combined L1+2 decomposition yields the highest classification accuracy (92.73%), sensitivity (92.97%), and specificity (99.04%). Additionally, certain words such as "selam", "kapak", "baju", "muka", and "abu" consistently achieved recall scores above 0.94, while "lampu" and "lembab" proved more difficult to classify. The findings demonstrate that integrating multi-level wavelet decomposition with CNN significantly improves the recognition of pathological speech and offers promising potential for clinical diagnostic support.

Keywords: cleft lip and palate; speech recognition; wavelet-mfcc; convolutional neural network;

1. Introduction

Cleft lip and/or palate (CLP) is one of the most common congenital anomalies worldwide, characterized by a separation in the upper lip and/or the roof of the mouth. This structural deformity often results in significant speech production difficulties, including resonance and articulation disorders. In Indonesia, the prevalence is notably high, with approximately 50.53% of recorded cases exhibiting both cleft lip and palate simultaneously [1]. Speech produced by individuals with CLP typically manifests atypical vocal quality, distorted formant transitions, and abnormal spectral features, rendering their communication distinct from that of individuals without cleft anomalies [2].

Previous research has explored various approaches for analyzing speech signals from individuals with cleft lip and palate (CLP), with a predominant focus on feature extraction techniques such as Discrete Wavelet Transform (DWT) paired with traditional machine learning classifiers like K-Nearest Neighbors (KNN). For instance, Yusuf and Dinata (2024) demonstrated that wavelets like rbio1.1 and dmey can effectively extract statistical features (e.g., mean, median, skewness) from CLP speech signals, achieving up to 93%

accuracy [3]. However, these methods rely heavily on hand-crafted features and shallow classifiers, which are often inadequate for capturing the complex and non-stationary nature of pathological speech. Given this, modern deep learning models—particularly Convolutional Neural Networks (CNNs) offer a promising alternative due to their ability to learn hierarchical representations directly from raw or minimally processed time-frequency features, such as MFCCs derived from wavelet-transformed signals [4], [5].

Nevertheless, current research remains limited by its reliance on traditional machine learning paradigms and has yet to fully leverage the capabilities offered by modern deep learning architectures [6]. Given the complexity and time-variant nature of speech signals from individuals with cleft conditions characterized by extensive variations across both temporal and frequency domains—traditional approaches relying solely on hand-crafted features and simple classifiers are often inadequate in capturing the intricate patterns inherent in such speech signals [7]. Therefore, there exists a pressing need to explore more robust, data-driven approaches capable of autonomously learning feature representations, thereby potentially offering improved generalization capabilities and higher performance [8].

This study aims to develop an automated speech classification system for individuals with cleft lip and palate (CLP) using Convolutional Neural Networks (CNN) and wavelet-based Mel-Frequency Cepstral Coefficients (MFCC). By focusing exclusively on rbio1.1 and bior1.1 wavelets, we streamline the feature extraction process while evaluating the impact of multi-level decomposition on classification accuracy. The research seeks to provide a robust, deep learning-driven solution for CLP speech recognition, offering potential applications in clinical diagnostics and assistive technologies.

2. Materials and Methods

The study follows a systematic five-phase pipeline (illustrated in Figure 1). First, we acquire speech samples exclusively from CLP-affected individuals to establish our experimental dataset. Subsequently, the raw audio undergoes preprocessing to enhance signal quality through noise reduction and normalization techniques. The third phase implements our novel feature extraction approach, combining wavelet decomposition with MFCC analysis to capture both time-frequency characteristics and perceptual speech features. These processed features then feed into our deep learning architecture, where we employ CNN networks for pattern recognition in different experimental setups. The study culminates in rigorous performance evaluation, utilizing standard metrics (accuracy, sensitivity, specificity) to quantify each model's capability in identifying CLP-specific articulatory patterns.

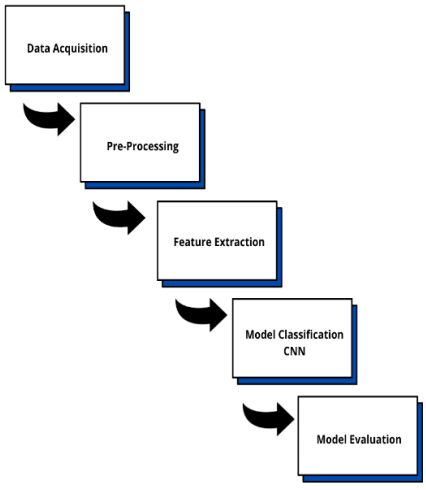

Figure 1. Overview of the proposed research workflow for CLP speech classification.

2.1. Data Acquisition

This research employed speech recordings obtained from ten participants clinically diagnosed with cleft lip and palate (CLP). The audio corpus consists of nine carefully selected Indonesian words - "Abu", "Atap", "Baju", "Kapak", "Lampu", "Lembab", "Muka", "Paku", and "Selam" - chosen specifically for their inclusion of bilabial plosives (/p/, /b/) and nasals (/m/), which are particularly challenging for CLP speakers to articulate clearly. Each participant repeated every word ten times, yielding a robust dataset of 900 utterances (10 participants × 9 words × 10 repetitions). All recordings followed strict acquisition parameters: 8 kHz sampling rate, 16-bit PCM encoding, single-channel capture, and storage in uncompressed WAV format to ensure optimal signal quality for subsequent feature extraction and analysis.

2.2. Preprocessing

The preprocessing pipeline begins with pre-emphasis filtering, a crucial step for boosting high-frequency components in speech signals [9]. These frequencies are naturally weakened during speech production, especially in CLP cases where anatomical differences further reduce their prominence. The filter compensates for this attenuation by enhancing rapid signal variations between consecutive samples, thereby preserving critical phonetic information that might otherwise be lost. The operation is mathematically defined as:

$$y[n] = x[n] - \alpha * x[n - 1] \tag{1}$$

where y[n] is the output signal, x[n] is the input signal, and α is the pre-emphasis coefficient, typically set in the range of 0.95 to 0.97.

After pre-emphasis, the speech signals undergo amplitude normalization to achieve consistent dynamic range across all recordings. This critical step eliminates amplitude variations caused by factors like vocal loudness, microphone positioning, or recording conditions, preventing potential biases in downstream processing. The normalization algorithm scales each signal sample by dividing it by the maximum absolute amplitude value found in the entire recording, as expressed by:

$$S_{norm} = \frac{s[n]}{max|s[n]|} \backslash label \tag{2}$$

where *s[n]* represents the original signal sample and max(|*s[n]*|) denotes the peak amplitude value.

2.3. Feature Extraction

Following preprocessing, the speech signals undergo Discrete Wavelet Transform (DWT) to perform multiresolution analysis. The DWT provides a time-frequency representation of the signal through scaled and translated wavelet basis functions, making it particularly suitable for analyzing non-stationary speech signals [10], [11] .

For a discrete-time signal *s[n]*,the DWT decomposition produces two types of coefficients at each level:

1. Approximation coefficients (*cA*): Represent the signal's low-frequency components that characterize its overall shape and trends
2. Detail coefficients (*cD*): Capture high-frequency components containing information about rapid transitions and fine structural details

In this study, both *cA* and *cD* coefficients are utilized at each decomposition stage. The selected wavelet families are rbio1.1 and bior1.1, applied under three decomposition strategies:

a. Level 1 Decomposition (L1)

The signal decomposed into:

- cA1 (approximation, level 1)
- cD1 (detail, level 1)

MFCC features are computed separately from cA1 and cD1. The resulting MFCC vectors are then concatenated to form a single feature set representing L1.

b. Level 2 Decomposition (L2)

The signal undergoes two-level decomposition:

- First level: cA1 and cD1
- Second level: cA2 (approximation of cA1) and cD2 (detail of cD1)

Only the second-level coefficients (cA2 and cD2) are used for MFCC extraction. MFCCs are computed separately for cA2 and cD2, and the resulting vectors are concatenated to form the L2 feature set.

c. Combined Multi-Level (L1+L2)

Features from L1 and L2 are merged by concatenation:

- MFCC(cA1) + MFCC(cD1) + MFCC(cA2) + MFCC(cD2)

This combined vector retains both broad spectral patterns (from L1) and fine-grained temporal details (from L2), providing the CNN with a richer representation of CLP speech characteristics.

The one-level decomposition can be mathematically expressed as:

$$cA_1[n] = \sum_k s[k] \cdot g[2n - k] \tag{3}$$

$$cD_1[n] = \sum_k s[k] \cdot h[2n - k] \tag{4}$$

The decomposition process utilizes two complementary filters: $g[n]$ (low-pass) and $h[n]$ (high-pass), which are intrinsic to each wavelet's mathematical formulation. This investigation specifically employs rbio1.1 and bior1.1 wavelets - selected for their superior performance in preliminary tests - across three distinct decomposition schemes: single-level (L1), two-level (L2), and a combined multi-resolution approach (L1+L2).

These particular wavelet families were chosen due to their:

1. Demonstrated efficacy in representing time-frequency structures in biomedical signals
2. Complementary filter characteristics that effectively capture both smooth trends and abrupt transitions
3. Proven performance in speech pathology applications

The resulting coefficient sets (approximation cA and detail cD) from each decomposition level form the foundational inputs for subsequent MFCC feature extraction. This multi-scale analysis framework is particularly crucial for characterizing the non-stationary articulatory patterns characteristic of CLP speech, as it simultaneously preserves both:

- Broad spectral trends (via *cA* coefficients)
- Transient phonetic features (via *cD* coefficients)

The experimental design systematically evaluates how decomposition depth (L1 vs. L2 vs. combined) interacts with wavelet type to influence classification accuracy, with particular attention to the optimal representation of pathological speech characteristics.

Building upon the wavelet-derived coefficients, we implement Mel-Frequency Cepstral Coefficients (MFCC) - the gold standard for speech feature extraction that mimics human auditory perception. The transformation process applies uniformly to both approximation (*cA*) and detail (*cD*) coefficients through four systematic stages:

1. Spectral Transformation: Each coefficient set undergoes FFT conversion to obtain power spectra, transitioning from time-domain to frequency-domain representation.
2. Perceptual Warping: The resulting spectra pass through a Mel-scaled filter bank that replicates the human ear's nonlinear frequency sensitivity, emphasizing perceptually relevant ranges.
3. Dynamic Range Compression: Logarithmic scaling of filter energies simulates the logarithmic sensitivity of human loudness perception while normalizing amplitude variations.
4. Decorrelation: A Discrete Cosine Transform condenses the log-filter energies into compact cepstral coefficients, with the k-th MFCC feature calculated as:

$$\mathrm{MFCC_k} = \sum_{m=1}^{M} \log(E_m) \cdot \cos\left[\frac{\pi k(m-0.5)}{M}\right] \tag{5}$$

where *Em* represents the m-th filter's energy, M*M* the total filters, and k*k* the coefficient index. The final feature vectors concatenate MFCCs from both cA and cD components, preserving multi-resolution speech characteristics essential for CLP pattern recognition.

This hybrid approach leverages wavelet decomposition's temporal precision with MFCC's psychoacoustic fidelity, creating optimal features for subsequent deep learning analysis. The combined representation proves particularly effective for capturing the atypical spectral patterns in CLP speech while maintaining robustness to individual articulation variations.

2.4. CNN Architecture

Convolutional Neural Networks (CNNs) have become a cornerstone in deep learning for analyzing spatial patterns in signal and image processing. They have demonstrated particular efficacy in speech recognition tasks[12]. By interpreting spectral representations such as MFCCs and wavelet transforms as two-dimensional speech "images," CNNs can effectively leverage their inherent spatial processing capabilities to identify phonetic and articulatory patterns.

This approach is especially valuable for analyzing Cleft Lip and Palate (CLP) speech, where the network can autonomously learn discriminative features from wavelet-enhanced spectral representations without relying on manual feature engineering. The hierarchical architecture of CNNs naturally captures both local and global patterns in the time–frequency domain. This makes them particularly adept at detecting characteristic articulatory distortions and non-uniform spectral features present in CLP speech.

Such spatial processing advantages allow CNNs to outperform traditional methods in identifying the unique acoustic signatures of CLP-related speech disorders. At the same time, they maintain robustness against the variability inherent in pathological speech production. Furthermore, the architecture's ability to model structural relationships across

frequency bands and temporal segments provides a powerful framework for analyzing the complex, non-stationary patterns that distinguish CLP speech from typical speech production.

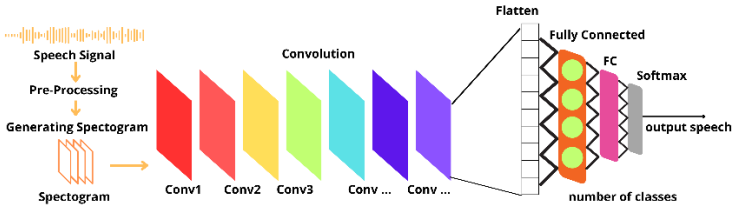

Figure 2. Convolutional Neural Network (CNN) architecture designed for CLP speech recognition using Wavelet-MFCC features.

This study employs a Convolutional Neural Network (CNN) to analyze two-dimensional Wavelet-MFCC features, capturing spatial patterns in speech spectra that are critical for identifying CLP-related articulatory distortions [13]. The network architecture is structured as follows:

1. Input Layer: Processes MFCC matrices of size ($N_{koe}f$, T), where $N_{koe}f$ is the number of coefficients and T represents time frames.
2. Convolutional Layers: Two to three layers with compact 3×3 kernels and ReLU activation, extracting localized spectral features linked to phonetic distortions.
3. Pooling Layers: Max pooling (2×2) reduces dimensionality while retaining perceptually significant features.
4. Flatten Layer: Converts the hierarchical feature maps into a 1D vector for classification.
5. Dense Layer: Consolidates learned features for class discrimination.
6. Softmax Output: Generates probabilistic class predictions.

This architecture is tailored to detect CLP-specific articulatory anomalies particularly in bilabial and nasal consonants by leveraging the spatial relationships in wavelet-enhanced spectral representations. The design emphasizes efficiency in processing non-stationary speech patterns while maintaining discriminative power for pathological speech classification.

2.5. Model Evaluation

This study employs a rigorous evaluation framework to assess the performance of Convolutional Neural Networks (CNNs) in classifying speech patterns associated with Cleft Lip and Palate (CLP). Given the diagnostic significance of accurate articulation analysis, the evaluation prioritizes both classification accuracy and generalization capability [14]. To ensure reliable performance estimation while mitigating overfitting, we implement 5-fold cross-validation - a robust validation technique where the dataset is partitioned into five equal subsets. In each evaluation cycle, the model trains on four subsets and validates on the remaining hold-out set, with this process systematically rotated across all folds [15]. The final performance metrics represent averaged results across all validation folds, yielding a comprehensive and unbiased assessment of model generalizability. This approach offers an optimal compromise between computational practicality and statistical reliability, particularly for medium-sized datasets, as established in previous machine learning research. The cross-validation strategy not only validates model effectiveness but also ensures the findings are representative of the broader CLP population, crucial for developing clinically applicable diagnostic tools.

The model's effectiveness is quantified through four essential metrics, each providing unique diagnostic insights critical for clinical applications:

1. Average Accuracy (cross validated)

$$MeanAccuracy = \frac{1}{K}\sum_{i=1}^{K} \text{Accuracy}_i \tag{6}$$

2. Core Classification Metrics

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{Sensitivity (Recall)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{9}$$

where:

TP = Correct CLP identifications
TN = Correct normal speech classifications
FP = False CLP detections
FN = Missed CLP cases

The cross-validated approach provides robust performance estimates while mitigating dataset bias, particularly crucial given the clinical consequences of both false positives and negatives in speech pathology assessment [16].

3. Results

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

3.1. CNN Performance

Table 1. Performance of CNN at Decomposition Level 1

Wavelet Type	Accuracy %	Sensitivity%	Specificity%
Rbio1.1	91.44 ± 0.88	91.90 ± 0.94	98.93 ± 0.09
Bior1.1	92.00 ± 0.86	92.02 ± 0.90	99.00 ± 0.09

The bior1.1 wavelet achieved the highest accuracy (92.00%) and specificity (99.00%) at level 1, demonstrating its suitability for CLP speech classification. The rbio1.1 wavelet also performed well, albeit with slightly lower metrics.

Table 2. Performance of CNN at Decomposition Level 2

Wavelet Type	Accuracy %	Sensitivity%	Specificity%
Rbio1.1	90.02 ± 0.93	91.04 ± 0.95	98.78 ± 0.10
Bior1.1	89.00 ± 0.97	89.58 ± 0.96	98.63 ± 0.12

The Rbio1.1 wavelet balanced superioriry across all metrics at level 2 despite lower performance than the previous level.

Table 3. Performance of CNN at Decomposition Levels 1 and 2

Wavelet Type	Accuracy %	Sensitivity%	Specificity%
Rbio1.1	92.73 ± 0.85	92.88 ± 0.89	99.04 ± 0.08
Bior1.1	92.33 ± 0.85	92.97 ± 0.89	99.04 ± 0.08

The combined-level wavelet decomposition (Level 1+2) demonstrates superior performance compared to single-level approaches, achieving peak classification metrics for both wavelet types. Notably, Rbio1.1 and Bior1.1 yield identical accuracy (92.33% ± 0.85) and specificity (99.04% ± 0.08), with Bior1.1 showing marginally better sensitivity (92.97% ± 0.89 vs. 92.88% ± 0.89).

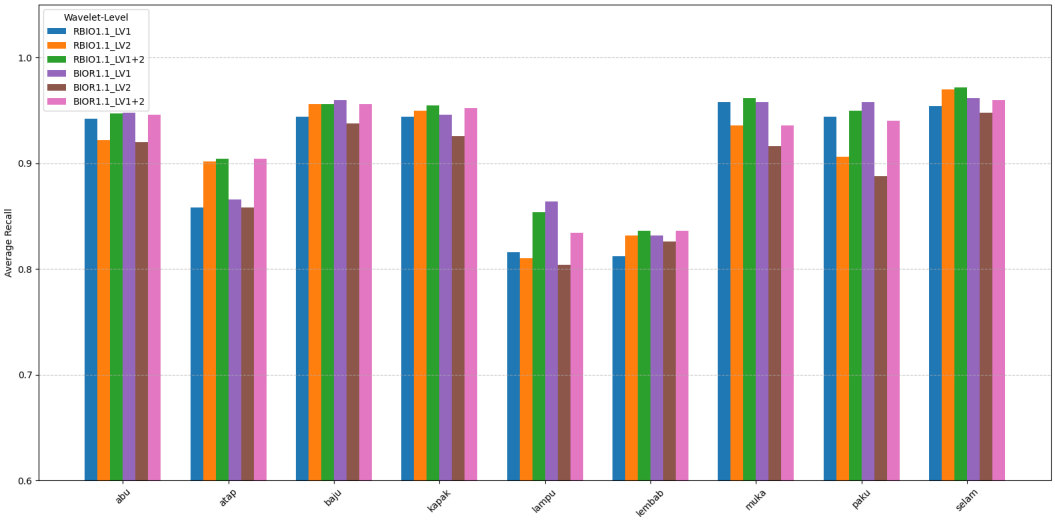

Figure 3. Average Recall per Words for Different Wavelet Decomposition Levels

Based on the average recall visualization, it can be concluded that the Wavelet-MFCC configuration combining decomposition levels (L1+2) consistently yields better performance across most words. Additionally, words such as "selam", "kapak", and "baju" demonstrate consistently high recall, while "lampu" and "lembab" remain challenging to classify accurately. The performance differences between RBIO1.1 and BIOR1.1 further suggest that the choice of wavelet type significantly influences the effectiveness of the speech recognition system for individuals with CLP.

4. Discussion

The experimental results reveal several important insights about wavelet-based CLP speech classification:

1. The combined Level 1+2 approach consistently outperformed single-level decompositions, achieving the highest accuracy (92.33%) and specificity (99.04%) for both wavelet types. This demonstrates that integrating multiple resolution levels provides more comprehensive feature representation, capturing both broad spectral trends (Level 1) and fine-grained details (Level 2) essential for identifying CLP speech patterns.
2. The exceptional specificity scores (>99%) across all configurations indicate remarkable reliability in distinguishing normal speech from CLP cases. This is particularly valuable for clinical applications where false positives could lead to unnecessary interventions. The high sensitivity (~93%) further confirms the model's ability to detect genuine pathology.
3. While both wavelets performed comparably in the combined configuration, Bior1.1 showed marginally better sensitivity (92.97% vs 92.88%), suggesting slightly better detection of true CLP cases. The minimal performance difference between wavelets in Level 1+2 indicates that decomposition strategy may be more critical than wavelet selection when using multi-level approaches.

4. The tight confidence intervals (±0.85-0.89) demonstrate robust model stability, particularly important for clinical deployment. This consistency holds across all evaluation metrics, suggesting reliable performance on unseen data.
5. The words “Selam”, “baju”, “kapak”, “muka”, and “abu” consistently achieved recall scores above 0.94 across nearly all wavelet and decomposition configurations. This indicates that the Wavelet-MFCC features effectively capture strong and distinctive patterns in these words, regardless of the wavelet type or decomposition level used.
6. "Lampu" and "lembab" exhibited the lowest recall scores, falling below 0.90 in most configurations and approaching 0.80 in some cases. This may be attributed to the articulation complexity experienced by individuals with CLP when pronouncing double consonants or nasal-vowel combinations, or due to acoustic similarities between certain word classes.

5. Conclusion

This study presented a deep learning-based approach for classifying speech from individuals with cleft lip and palate (CLP) using Convolutional Neural Networks (CNN) and Wavelet-MFCC features derived from rbio1.1 and bior1.1 wavelets. The experimental results demonstrate that multi-level wavelet decomposition, particularly the combination of level 1 and level 2 (L1+2), consistently enhances classification performance across most target words. Notably, words such as “selam”, “baju”, “kapak”, “muka”, and “abu” exhibited recall scores exceeding 0.94 across nearly all configurations, indicating the robustness of the proposed feature extraction method in capturing salient speech patterns. Conversely, words like “lampu” and “lembab” posed greater classification challenges, potentially due to complex articulation or acoustic similarity between classes.

The findings further highlight that the choice of wavelet significantly affects recognition accuracy, with rbio1.1 and bior1.1 each showing strengths depending on the word and decomposition level used. Overall, the integration of wavelet-based MFCC and CNN proves to be a promising framework for pathological speech recognition, offering potential applications in automated screening tools and speech therapy assistance for individuals with CLP.

References

[1] C. I. Alois and R. A. Ruotolo, “An overview of cleft lip and palate,” *JAAPA*, vol. 33, no. 12, pp. 17–20, Dec. 2020, https://doi.org/10.1097/01.JAA.0000721644.06681.06.

[2] F. R. Larangeira *et al.*, “Speech nasality and nasometry in cleft lip and palate,” *Braz. J. Otorhinolaryngol.*, vol. 82, no. 3, pp. 326–333, May 2016, https://doi.org/10.1016/j.bjorl.2015.05.017.

[3] S. A. A. Yusuf and M. I. Dinata, “Features Extraction on Cleft Lip Speech Signal using Discrete Wavelet Transformation,” *JTIM J. Teknol. Inf. Dan Multimed.*, vol. 6, no. 2, Art. no. 2, July 2024, https://doi.org/10.35746/jtim.v6i2.545.

[4] M. Telmem, N. Laaidi, Y. Ghanou, S. Hamiane, and H. Satori, “Comparative study of CNN, LSTM and hybrid CNN-LSTM model in amazigh speech recognition using spectrogram feature extraction and different gender and age dataset,” *Int. J. Speech Technol.*, vol. 27, no. 4, pp. 1121–1133, Dec. 2024, https://doi.org/10.1007/s10772-024-10154-0.

[5] R. B. Pittala, B. R. Tejopriya, and E. Pala, “Study of Speech Recognition Using CNN,” in *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, Feb. 2022, pp. 150–155. https://doi.org/10.1109/ICAIS53314.2022.9743083.

[6] P. N. Sudro, R. K. Das, R. Sinha, and S. R. M. Prasanna, “Significance of Data Augmentation for Improving Cleft Lip and Palate Speech Recognition,” Oct. 02, 2021, *arXiv*: arXiv:2110.00797. https://doi.org/10.48550/arXiv.2110.00797.

[7] M. Geng *et al.*, “Spectro-Temporal Deep Features for Disordered Speech Assessment and Recognition,” in *Interspeech 2021*, ISCA, Aug. 2021, pp. 4793–4797. https://doi.org/10.21437/Interspeech.2021-60.

[8] A. Subasi, “Chapter 3 - Machine learning techniques,” in *Practical Machine Learning for Data Analysis Using Python*, A. Subasi, Ed., Academic Press, 2020, pp. 91–202. https://doi.org/10.1016/B978-0-12-821379-7.00003-5.

[9] M. Labied, A. Belangour, M. Banane, and A. Erraissi, “An overview of Automatic Speech Recognition Preprocessing Techniques,” in *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, Chiangrai, Thailand: IEEE, Mar. 2022, pp. 804–809. https://doi.org/10.1109/DASA54658.2022.9765043.

[10] A. Koduru, H. B. Valiveti, and A. K. Budati, “Feature extraction algorithms to improve the speech emotion recognition rate,” *Int. J. Speech Technol.*, vol. 23, no. 1, pp. 45–55, Mar. 2020, https://doi.org/10.1007/s10772-020-09672-4.

[11] Y. Huang, K. Tian, A. Wu, and G. Zhang, "Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition," *J. Ambient Intell. Humaniz. Comput.*, vol. 10, no. 5, pp. 1787–1798, May 2019, https://doi.org/10.1007/s12652-017-0644-8.
[12] D. Baker, "Mahmood A. & Köse U. / AAIR vol 1:1(2021) 6-12," vol. 1, 2021.
[13] Department of Computer Sciences, Ajayi Crowther University, Oyo, Nigeria. and J. A. Ayeni, "Convolutional Neural Network (CNN): The architecture and applications," *Appl. J. Phys. Sci.*, vol. 4, no. 4, pp. 42–50, Dec. 2022, https://doi.org/10.31248/AJPS2022.085.
[14] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Mach. Learn. PYTHON*, 2011.
[15] R. T. Nakatsu, "Validation of machine learning ridge regression models using Monte Carlo, bootstrap, and variations in cross-validation," *J. Intell. Syst.*, vol. 32, no. 1, Jan. 2023, https://doi.org/10.1515/jisys-2022-0224.
[16] Department of Computer Science and Informatics, University of Energy and Natural Resources, Sunyani, Ghana, I. K. Nti, O. Nyarko-Boateng, and J. Aning, "Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation," *Int. J. Inf. Technol. Comput. Sci.*, vol. 13, no. 6, pp. 61–71, Dec. 2021, https://doi.org/10.5815/ijitcs.2021.06.05.