A Multimodal Deep Learning Framework for Amyotrophic Lateral Sclerosis Diagnosis using Clinical and Audio Morphology Features

[1]I Nyoman Switrayana*, [2]Tomi Tri Sujaka, [3]Imelda Silpiana Putri
[1,2,3]Computer Science, Faculty of Engineering, Bumigora University
[1,2,3]Ismail Marzuki Street no, 22, Mataram, Nusa Tenggara Barat, Indonesia
*e-mail: nyoman.switrayana@universitasbumigora.ac.id

(***received***: 24 September 2025, ***revised***: 13 November 2025, ***accepted***: 14 November 2025)

Abstract

Amyotrophic Lateral Sclerosis (ALS) is a highly progressive neurodegenerative disease that impairs motor and speech function. Conventional diagnostic methods, both invasive and non-invasive, are often time-consuming and produce limited sensitivity. This leads to delays in treatment and worsening disease progression. This study proposes a multimodal deep learning framework that utilizes and integrates invasive medical records with non-invasive morphological features of patient speech audio extracted into Mel-Spectrograms. Unlike previous studies that focused solely on speech or clinical features, this study introduces an integrated multimodal diagnostic framework that effectively combines both data sources to achieve reliable diagnostic accuracy. The study included two experimental scenarios. In the first scenario, the audio-trained model used a Convolutional Neural Network (CNN) and was systematically optimized by testing variations in network depth, feature fusion techniques, and layer dropout probabilities to improve model generalization and stability. From the experimental results of the first scenario, the CNN achieved the best performance, achieving 80.33% accuracy in classification using audio data alone from all the tested model variations. In the second experimental scenario, when the best model was trained by incorporating clinical data, the model demonstrated improved diagnostic performance, achieving 100% accuracy. This finding highlights the importance of combining data modalities or sources from various domains, both invasive and non-invasive, to achieve optimal model performance for early ALS detection.

Keywords: amyotrophic lateral sclerosis, audio morphology, clinical records, convolutional neural network, multimodal

1 Introduction

Amyotrophic Lateral Sclerosis (ALS) is a progressive neurodegenerative disease that primarily affects motor neurons in the brain and spinal cord. This condition ultimately leads to muscle weakness and paralysis [1], [2]. As the disease progresses, individuals with ALS experience a gradual decline in voluntary muscle function. This is characterized by impairments in speech, mobility, swallowing, and even breathing [3]. Given its fatal nature and the lack of a definitive cure, early diagnosis is crucial to allow for therapeutic interventions that can slow disease progression and improve quality of life [[4], [5], [6]. Despite its great urgency, ALS diagnosis still relies heavily on conventional clinical assessment and electromyography (EMG). Both are invasive, expensive, and require specialized expertise [7]. Furthermore, these diagnostic methods are often time-consuming and inaccessible in many clinical settings. This results in delays in diagnosis and subsequent treatment initiation [8]. Such delays are detrimental, as interventions tend to be less effective once the disease has reached an advanced stage.

Another additional challenge in diagnosing ALS is its complex and multifactorial nature. It not only affects motor function but also impairs a person's ability to speak due to damage to the motor neurons that control vocal articulation. As a result, ALS sufferers often experience tremors when speaking [9]. In this context, speech analysis has recently emerged as a promising non-invasive biomarker for the early detection of ALS. This analysis offers the potential for early screening through detectable acoustic and prosodic changes in the patient's voice [10], [11], [12]. However, most previous studies on ALS detection systems suffer from critical limitations. These studies have used only one type of data, either invasive or non-invasive data [13-29]. This limited approach ignores the multidimensional nature of ALS symptom data. Multidimensionality means that ALS has diverse clinical and physiological symptoms, where all of these data can provide complementary insights when analyzed within an integrated framework. The lack of an approach to combining these multi-modal data in current diagnostic systems represents a missed opportunity to leverage all ALS-related indicators to produce a more accurate and reliable early identification or detection system.

To address these issues, this study proposes a novel diagnostic system approach. The proposed diagnostic system integrates both invasive and non-invasive data modalities. Specifically, this study combines clinical information, such as genetic markers and other neurological assessments, with voice-based audiomorphological features represented through Mel-Spectrograms. This representation is then processed using a convolutional neural network (CNN) model. The CNN model processes the data and recognizes patterns within the two datasets. To produce a high-performance model, the model is trained and validated using K-Fold Cross-Validation. It is then optimized using the Adam Optimizer and Early Stopping to ensure optimal generalization and prevent overfitting. Through this integrated and structured framework, this study aims to develop an ALS diagnostic system by improving classification performance and enhancing system robustness. This research contributes to the field of deep learning and medical diagnostics for ALS. This research approach not only strengthens diagnostic capacity by leveraging diverse data sources but also encourages the development of non-invasive, efficient, and scalable diagnostic tools. It is hoped that it can be developed and applied to a wider clinical sector.

2 Literature Review

Recently, ALS detection models have been extensively explored through various modalities or data types and machine learning techniques. For example, Antunes et al. [13] proposed the use of surface electromyography (sEMG) features processed using the AdaBoost model. Their research showed promising results in early ALS identification. In Rong et al. [14], they investigated how sEMG and acoustic sound features were combined. The results showed that multimodal data integration significantly improved the model's performance in ALS detection. Similarly, in Cebola et al. [15], audio signals underwent feature extraction by segmenting through windowing and testing several machine learning models. Their research showed that the Support Vector Machine (SVM) model outperformed other models.

The use of other types of data from ALS symptoms is used, such as the identification approach by models with input data from MRI. MRI itself is a feature for ALS detection extracted from brain imaging data. In the studies of Tafuri et al. [16], Jamrozy et al. [17], and Kocar et al. [18], they used MRI data and showed that the SVM model was the most effective in ALS detection. Furthermore, research conducted by Tena et al. [19], they tried to utilize time-frequency features and phonatory sounds trained on a random forest model. Research by Kurmi et al. [20] even introduced wavelet time scattering transforms on acoustic features and compared the performance of several deep learning architectures. The deep learning models used included Deep Neural Networks (DNN), Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN). Their findings revealed that DNN and CNN provided superior performance.

Another study in acoustic-based ALS detection is the work of Simmatis et al. [21]. Here they explored various acoustic features to distinguish between ALS patients and those without. Then, the study by Bhattacharjee et al. [22], they compared pitch and Mel-Frequency Cepstral Coefficients (MFCC). The results confirmed that the pitch feature has robustness in noisy environments. The study by Wang et al. [23] introduced an architecture named Longitudinal Speech Transformer for temporal analysis of speech in ALS patients. Meanwhile, the study by An KH et al. [24] who extracted statistical features from intelligible speech and trained it on CNN mode was successful for ALS detection at an early stage. Recent studies have also shown how to integrate advanced pre-trained models and equipped with hybrid feature extraction strategies. The study by Jayakumar et al. [25] used the HuBERT model for hypernasality detection and ALS patient detection using a Dense Neural Network deep learning model. Another study by Ilias et al. [26], they used voice features extracted by MFCC and used a hypernetwork variant model based on AlexNet. Research by [27] used neural networks to predict ALS progression by utilizing ALS Functional Rating Scale (ALSFRS) data. Also research conducted by Faghri et al. [28], they used clinical data to personalize ALS subtypes. In a different type of modality, Ngo TD et al. [29] proposed the use of Electroencephalogram (EEG) signals and eye movement recordings to assess ALS-related motor dysfunction. These features are relatively new and different from other features. However, they are still obtained from observable symptoms experienced by ALS patients. So from there we can see that there are many types of data that can be obtained from people with ALS.

Of all the studies discussed above, there are still significant limitations in each study. This limitation is that most approaches focus on a single data modality, whether it is data obtained invasively (e.g., clinical, MRI, EEG) or non-invasive (e.g., voice, sEMG). These studies have not considered integrating both types of data into an integrated diagnostic system. Also considering that the complexity of ALS symptoms comes from the entire motor, respiratory, and vocal systems [30-31]. By integrating various data sources, it is hoped that it can provide a more comprehensive and accurate diagnostic representation. Because the learning model will have many features or variables that need to be learned and considered just to detect whether someone has ALS or not. Assessments in the form of clinical data [32], MRI imaging [33], muscle ultrasound [34], and voice signals each capture unique aspects of ALS symptoms. Although the potential of each modality has been proven, previous studies still lack multimodal integration, thus strengthening the gap in previous studies. The hypothesis of this study is that a diagnostic system that processes clinical and speech data simultaneously can offer more holistic and reliable predictive capabilities. Therefore, to address the existing gaps in previous research, this study proposes a multimodal ALS detection framework, which combines clinical data (invasive) and Mel-Spectrogram representations of speech signals (non-invasive), into a CNN-based model. In addition to proposing the integration of these data, this study will also introduce several technical contributions that can be observed from the workflow. These contributions include, first, deep feature extraction from Mel-Spectrogram using various CNN configurations (number of layers, fixed kernel size, max/average pooling). Second, the use of varying dropout probabilities for regularization. And third, is implementing performance optimization through K-Fold Cross-Validation, Adam Optimizer, and Early Stopping. Thus, this study proposes an ALS diagnostic system with a novel approach. This not only addresses the practical need for a scalable and non-invasive ALS screening tool but also opens new directions for future research in multimodal health informatics and intelligent neurodiagnostics.

3 Research Method

The research methodology, or stages, of this study were designed to ensure a systematic and organized workflow. A flowchart of the research stages is presented in Figure 1. Figure 1 shows the sequential steps to be undertaken in this study. This research begins with data collection, followed by data preprocessing, and then feature extraction. The data preprocessing stage aims to clean and prepare the data. The feature extraction stage aims to capture relevant characteristics of the data used before modeling. The next stage is the modeling and model evaluation stage.

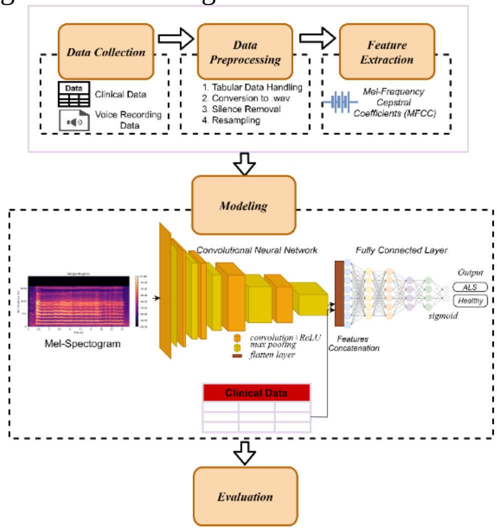

Figure 1 Research Flowchart

ISSN:2302-8149
e-ISSN:2540-9719

A. In the Data Collection stage, this stage includes obtaining medical record datasets and voice recordings for ALS diagnosis as explained in the research [35]. There are several parameters or clinical features that can be obtained to distinguish between healthy individuals and those suffering from ALS. These features include genetic markers (C9orf72, SOD1, FUS, TARDBP) [36], [37], [38], pulmonary function (FVC%), difficulty swallowing score, Revised ALS Functional Assessment Scale (ALSFRS-R), King's Clinical Stage, Medical Research Council (MRC) Score, namely muscle strength score, and Penn Upper Motor Neuron Score (PUMNS). Then for voice data collected from ALS patients and healthy controls. Data collected were obtained from 153 subjects, with 102 of them being ALS patients and 51 healthy controls.

B. Data preprocessing is the stage for preparing tabular (clinical) data and voice data before feature extraction to ensure the quality and consistency of each data. For tabular data, preprocessing involves handling missing values and maintaining the consistency of clinical data using Min-Max scaling. The value scaling technique on data using Min-Max is defined in Equation 1 [39]. For audio data, the data is converted to WAV format (16-bit, mono) to be compatible with the subsequent processing flow. Then, silence removal will be performed to remove silent or noisy segments in the voice signal. Resampling of the voice signal is also performed to standardize the sample rate across recordings at 22.05 kHz.

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

Where,

X_{min} : the minimum value in the dataset
X_{max} : the maximum value in the dataset
X : the data value to be scaled
X_{scaled} : the data value after scaling

C. The feature extraction stage is a process carried out to obtain important information/characteristics from voice data. This stage aims to represent the voice signal in a more concise and meaningful form without losing its essential characteristics. The method used in this study is MFCC [40], [41], [42], with the procedure illustrated in Figure 2.

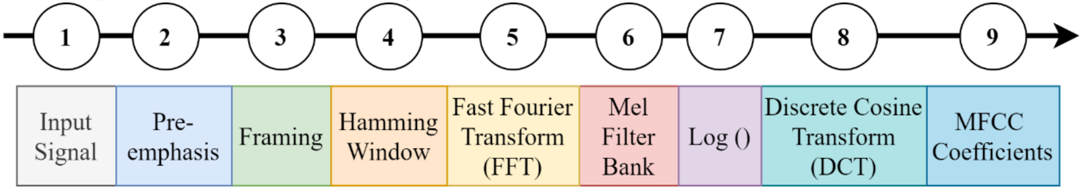

Figure 2 MFCC processing steps

MFCC processing begins with pre-emphasis to apply a high-pass filter to the speech signal. This process aims to amplify the signal and enhance its high-frequency components. Next, the signal is segmented into short frames, with or without overlap. This process is called framing. After the framing process, the next process is windowing. The Hamming window used in this study functions to reduce spectral distortion and optimally capture signal features. The application of the Hamming window is explained in Equation (2).

$$W(n, a) = (1 - a) - a\cos\left(\frac{2\pi n}{N - 1}\right), \qquad (2)$$
$$where\ 0 \leq n \ \leq N - 1$$

$W(n)$ represents the Hamming window and S(n) enotes the signal frame from n = 0 to n = N -1. Fast Fourier Transform (FFT) is applied to convert the signal from the time domain to the frequency domain, as defined in Equation (3). Here, $\mathrm{S_i(n)}$ represents the time-domain signal, $\mathrm{z_i(k)}$ denotes the frequency-domain signal, $h(n)$ is the window of length N samples, and k corresponds to the FFT length.

$$z_i(k) = \sum_{n=1}^{N} S_i(n)h(n)e^{\frac{-2\pi}{N}} \qquad (3)$$

After the FFT process, the next process is the application of the Mel Filterbank which is useful for adjusting the frequency representation to suit human hearing sensitivity. The method is by converting the linear frequency f (Hz) to the Mel scale f_{mel}) according to Equation (4). Each

f_{mel}) is then mapped to its corresponding filter in the filterbank. Then, the Logarithm process is applied to the filterbank output which is calculated to approximate human perception of sound intensity. Finally, the Discrete Cosine Transform (DCT) is applied to convert the signal back from the frequency domain to the time domain. This process produces the MFCC coefficients as a feature vector, as defined in Equation (5). Here, C_m represents the DCT results for m = 1, 2, ..., 12. N is the number of filterbanks, and E_k represents the logarithmic output. The output of the MFCC process will be used as a feature, which can be visually represented in the form of a Mel Spectrogram.

$$f_{mel} = 2595\, log_{10}\left(1 + \frac{f}{700}\right) \qquad (4)$$

$$C_m = \sum_{k=1}^{N} \cos\left(m(k - 0.5)\frac{\pi}{N}\right) E_k \qquad (5)$$

D. In the Modeling stage, CNN is designed and built to process Mel-Spectrogram. CNN is used because of its ability to capture spatial patterns and complex features, especially in the form of images [43]. The CNN architecture will include a convolutional layer with ReLU activation followed by a pooling layer that functions for feature dimensionality reduction. There is an optional dropout layer for regularization. And at the end of the CNN there will be a Fully Connected Layer (FCL) used to process Mel-Spectrogram features with combined clinical data. In the multimodal data fusion process, audio features extracted from Mel-Spectrogram are combined with clinical data using a concatenate layer. Where each modality first needs to be processed in its respective neural network branch before being combined for multimodal classification. The output layer of this architecture uses Sigmoid activation which is suitable for binary classification with Binary Cross-Entropy Loss as its loss function. To develop a model with multimodal data input, initial investigations of CNN architecture were carried out first by varying the number of convolutional layers, pooling strategy, and dropout configuration. This systematic exploration is summarized in Table 1. These architectural variations were designed to identify the most effective architectural setup for capturing relevant patterns in Mel-Spectrogram data before integration with clinical features.

Table 1 CNN architecture scenarios for melspectogram modeling

Research Scenario	Scenario Code	#Conv Layers	Filters & Kernel Size	Pooling Type
S1	S1A	1	Conv(16, 3x3)	MaxPooling(2x2)
	S1B	1	Conv(16, 3x3)	AveragePooling(2x2)
S2	S2A	2	Conv(16, 3x3) → Conv(32, 3x3)	MaxPooling(2x2)
	S2B	2	Conv(16, 3x3) → Conv(32, 3x3)	AveragePooling(2x2)
S3	S3A	3	Conv(16, 3x3) → Conv(32, 3x3) → Conv(64, 3x3)	MaxPooling(2x2)
	S3B	3	Conv(16, 3x3) → Conv(32, 3x3) → Conv(64, 3x3)	AveragePooling(2x2)
S4	S4A	4	Conv(16, 3x3) → Conv(32, 3x3) → Conv(64, 3x3) → Conv(128, 3x3)	MaxPooling(2x2)
	S4B	4	Conv(16, 3x3) → Conv(32, 3x3) → Conv(64, 3x3) → Conv(128, 3x3)	AveragePooling(2x2)
S5	S5A	5	Conv(16, 3x3) → Conv(32, 3x3) → Conv(64, 3x3) → Conv(128, 3x3) → Conv(256, 3x3)	MaxPooling(2x2)
	S5B	5	Conv(16, 3x3) → Conv(32, 3x3) → Conv(64, 3x3) → Conv(128, 3x3) → Conv(256, 3x3)	AveragePooling(2x2)

The model was trained using K-Fold Cross Validation to optimize performance and minimize bias [44], as illustrated in Figure 3. For example, if K = 4, the dataset is divided into four folds. This means the model is trained on three folds and tested on the remaining folds. This process is repeated four times, with each fold serving once as the test dataset. The final evaluation results are obtained by averaging the performance across the four iterations. This results in a more stable evaluation of the model and reduces bias in the model performance assessment. In this study, the evaluation of the training and testing split/model ratios was performed using a 5-fold cross-validation scheme.

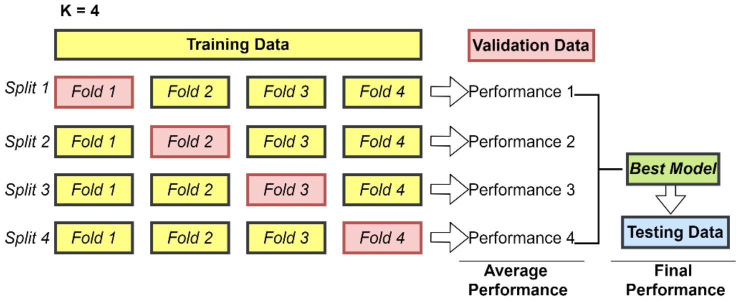

Figure 3 Schematic representation of k-fold cross validation

Next, optimization during the training process is performed using the Adam optimizer. Early stopping is also applied to prevent overfitting. The Adam optimizer works adaptively by adjusting the learning rate of each parameter, as illustrated in Figure 4. With this optimization, the model is able to avoid local minima and approach global minima. This approach can speed up and stabilize the training process [45].

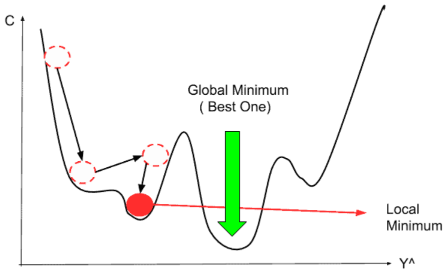

Figure 4 Schematic representation of the adam optimizer

Early Stopping works by stopping model training when performance on validation data begins to decline. This phenomenon is illustrated in Figure 5. This technique prevents overfitting and ensures the model remains able to generalize to previously unseen data [46]. All procedures in this study were implemented using Python with the TensorFlow framework.

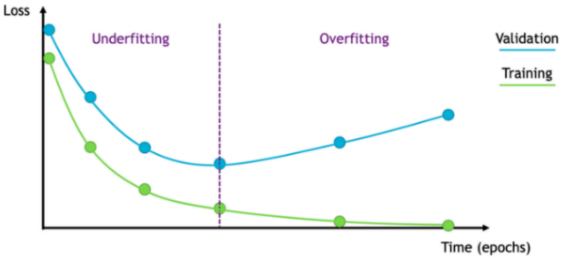

Figure 5 Schematic representation of the early stopping mechanism

E. Evaluation is the stage where the trained model is then tested with test data to assess its performance. Evaluation in this study was conducted using four main metrics, namely Precision, Recall, F1 Score, and Accuracy. The calculation of each evaluation metric is defined in Equation (6–9) [47]. Here, TP is the number of correct predictions for the positive class. TN is the number of correct predictions for the negative class. Then FP is the number of incorrect predictions for the positive class. And finally FN is the number of incorrect predictions for the negative class.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

$$\text{F1} - \text{Score} = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \tag{8}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{9}$$

4 Results and Analysis

A. Data Collection

The clinical dataset used in this study consists of 35 features that comprehensively capture the demographic, genetic, functional, and neurological aspects of ALS patients and healthy controls. The main features in the dataset include diagnostic metrics and disease duration (DiagnosticDelay, DiseaseDuration), pulmonary function (FVC%), and functional ability assessed using the ALS Functional Rating Scale Revised (ALSFRS-R). Because the ALSFRS-R contains several attribute columns, these features will be further broken down into subscores to measure the ability to speak, salivate, swallow, handwriting, cut food, dress and personal hygiene, turning over in bed, walking, climbing stairs, dyspnea, orthopnea, and respiratory insufficiency. The dataset also includes a feature in the form of motor function evaluated using the Medical Research Council (MRC) score for head, upper limb, and lower limb muscles. Along with the Penn Upper Motor Neuron Score (PUMNS) for bulbar, upper limb, and lower limb involvement. There is a disease stage feature recorded through the King Clinical Stage and the severity of dysarthria is assessed using the Cantagallo Questionnaire. Additionally, genetic information for key genes in patients with ALS (SOD1, TARDBP, C9orf72, FUS) was collected. Region of disease onset and treatment status were also documented. The Revised El Escorial Criteria were included to categorize disease probability. Overall, this dataset reflects a diverse range of clinical presentations, spanning early and advanced disease states. Therefore, with this clinical data, models trained with a rich multimodal foundation approach can be developed. Examples of clinical data are shown in Table 2.

Table 2 Sample of clinical records including als and healthy control (HC) subjects

ID	Category	Genetic Test	FVC%	ALSFRS-R_TotalScore	…	Cantagallo Questionnaire
CT001	HC	-	-	-	…	7,0
CT004	HC	-	-	-	…	0,0
CT010	HC	-	-	-	…	22,0
CT069	HC	-	-	-	…	11,0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
PZ111	ALS	C9ORF72 expansion	104	43	…	49,0
PZ112	ALS	negative	64	20	…	89,0
PZ114	ALS	negative	98	40	…	18,0
PZ115	ALS	negative	92	40	…	3,0

The voice data used in this study were systematically collected from HC and ALS subjects. The voice recordings included phonated vowels (A, I, U, E, O) and rhythmic syllables (KA, PA, TA). Each phonation or syllable type contained a total of 153 recordings, with a distribution of 51 recordings from HC individuals and 102 recordings from ALS patients. Overall, the dataset contained 408 recordings for HC and 816 recordings for ALS. These recordings were later transformed into Mel-Spectrogram representations during the CNN-based feature extraction and modeling stages. The distribution of the recording data in the form of a pie chart for each phonation and rhythmic syllable is shown in Figure 6. This distribution provides an overview of the balance of the dataset and sample coverage across the various vocal tasks in this study.

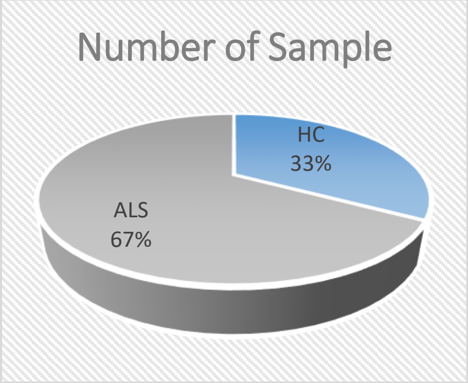

Figure 6 Distribution of voice recordings for HC and ALS subjects across phonation vowels and rhythmic syllables.

B. Preprocessing

Clinical and voice data were preprocessed to ensure data quality and consistency for subsequent data processing and modeling. Two types of preprocessing were performed due to the different data types. First, for clinical data preprocessing, missing values in all numeric columns were handled using SimpleImputer with a constant strategy of filling in 0. This step ensured that no missing values would interfere with subsequent analysis and model training. The imputed numeric features were then combined with categorical features that had been previously converted to numeric representation through encoding. This data then formed a complete tabular feature matrix. To ensure that all features were on a consistent scale and compatible with machine learning algorithms, the feature matrix was normalized using MinMaxScaler. This technique works by scaling each value in a feature column to the range 0–1. This approach ensures uniform feature scaling and reduces bias due to differences in magnitude. This approach aims to improve the stability and predictive performance of the model. Table 3 shows the results of clinical data preprocessing.

Table 3 Overview of clinical data preprocessing including missing value imputation and feature normalization

ID	Category	Genetic Test SOD1	Genetic Test TARDBP	Genetic Test C9ORF72 expansion	Genetic Test negative	FVC %	ALSFRS-R Total Score	…	Cantagallo Questionnaire
CT001	HC	0	0	0	0	0	0	…	0,0538
CT004	HC	0	0	0	0	0	0	…	0
CT010	HC	0	0	0	0	0	0	…	0,1692
CT069	HC	0	0	0	0	0	0	…	0,0846
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
PZ111	ALS	0	0	1	0	0,8253	0,9347	…	0,3769
PZ112	ALS	0	0	0	1	0,5079	0,4347	…	0,6846
PZ114	ALS	0	0	0	1	0,7777	0,8695	…	0,1384
PZ115	ALS	0	0	0	1	0,7301	0,8695	…	0,0230

The second preprocessing is to prepare the audio data. The initial sound recording was first converted to 16-bit mono-channel PCM and resampled to 22.05 kHz. The sound signal in the recording was then processed with silence removal. Figure 7 illustrates an example of a sound waveform before and after silence removal. Silence removal in the sound benefits the extraction process by focusing only on relevant sound segments. It will also impact the computation process of Mel-Spectrogram features and CNN-based modeling more efficiently. As shown in the figure, the original signal, which was originally 14 seconds long, is now reduced to 12 seconds after silence removal.

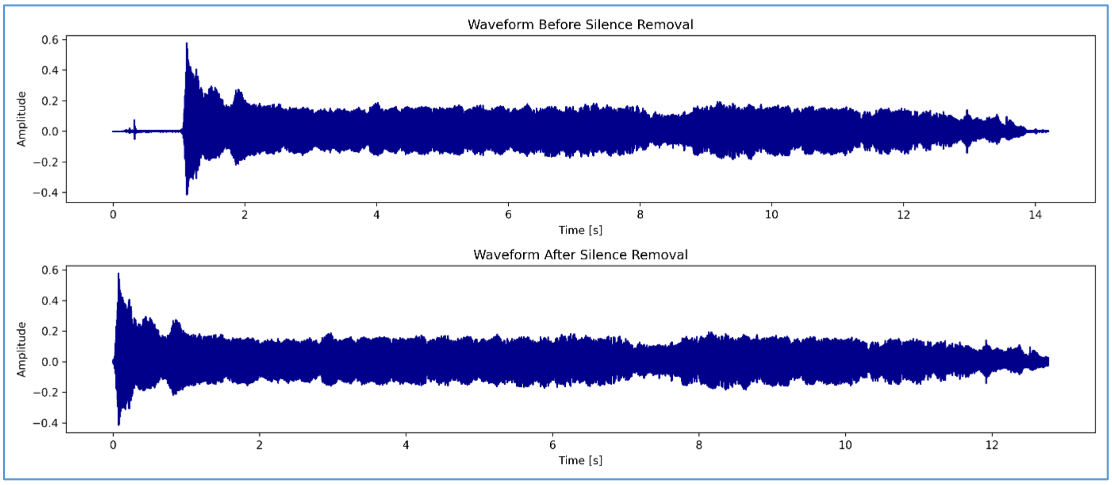

Figure 7 Voice signal before and after silence removal

C. Feature Extraction

The features extracted from preprocessed voice recordings are then represented as Mel-Spectrograms. Mel-Spectrograms are the features of a voice that characterize each voice signal in the time and frequency domains. Figure 8 shows an example of the results of feature extraction in the form of Mel-Spectrograms for audio of healthy controls (HC) and ALS patients when pronouncing U phonation. From the Mel-Spectrogram image, a clear difference in spectral patterns is visible that reflects the voice disorders of patients with ALS and healthy ones. In the Mel-Spectrogram, different spectral patterns between healthy controls (HC) and ALS patients can be observed where the HC shows clearer and more stable harmonics. The energy distribution is also more uniform over time. This condition occurs because the vibration of the vocal cords in healthy individuals is more normal and the phonation is stable (no vibration or oscillation). In contrast, the ALS Mel-Spectrogram sample has an irregular and less stable harmonic structure, with more variability in amplitude and frequency over time. This reflects that there is a disturbance in voice production due to neuromuscular degeneration in ALS patients. These differences explain to the model that the potential of Mel-Spectrogram features to capture the characteristics of ALS-related dysarthria can be effectively utilized later by CNN-based deep learning models.

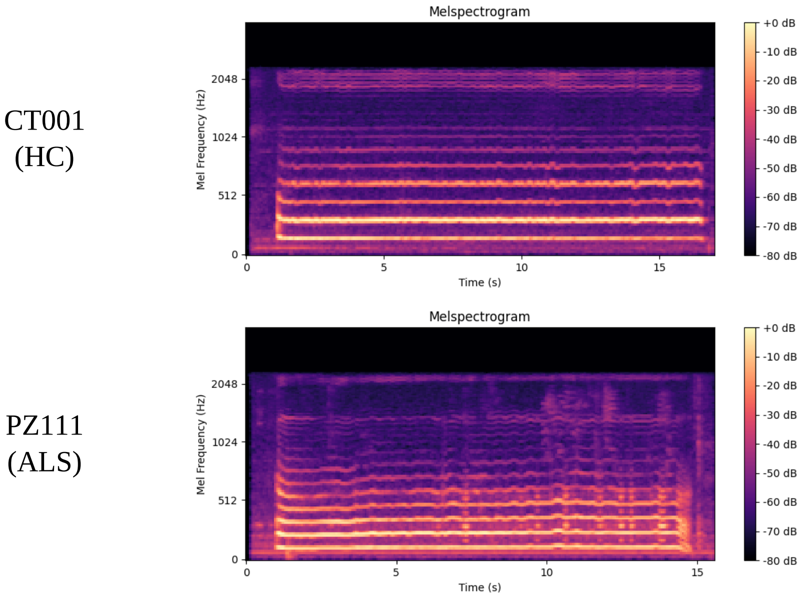

Figure 8 Mel-Spectrogram of voice recordings for healthy controls (HC) and ALS patients on phonation "U"

D. Modeling

The convolutional neural network models were developed according to the experimental scenarios outlined in Table 1, varying in the number of convolutional layers, filter sizes, and pooling types. Among all tested scenarios, the S1B configuration, which consists of a single convolutional layer with 16 filters of size 3×3 followed by AveragePooling (2×2), demonstrated the best performance and was therefore adopted as the CNN backbone for the multimodal model. The multimodal architecture, illustrated in Figure 9, integrates features from both the CNN-based Mel-Spectrogram extractor and tabular clinical data. The image branch includes a Conv2D layer with 16 filters, AveragePooling, Dropout (0.3), Flatten, and a fully connected layer with 256 neurons, while the tabular branch contains fully connected layers with 64 and 32 neurons, interleaved with BatchNormalization and Dropout (0.3). Features from both branches are concatenated and passed through a Dense layer with 64 neurons before the final softmax classification layer. The model was trained with an input image size of 128×128, batch size of 32, and up to 100 epochs, using 5-fold stratified cross-validation. Training included early stopping, learning rate reduction on plateau, and checkpointing to save the best model based on validation loss. This multimodal design allows the network to jointly learn from spectral and clinical features, enhancing its discriminative capability for ALS detection..

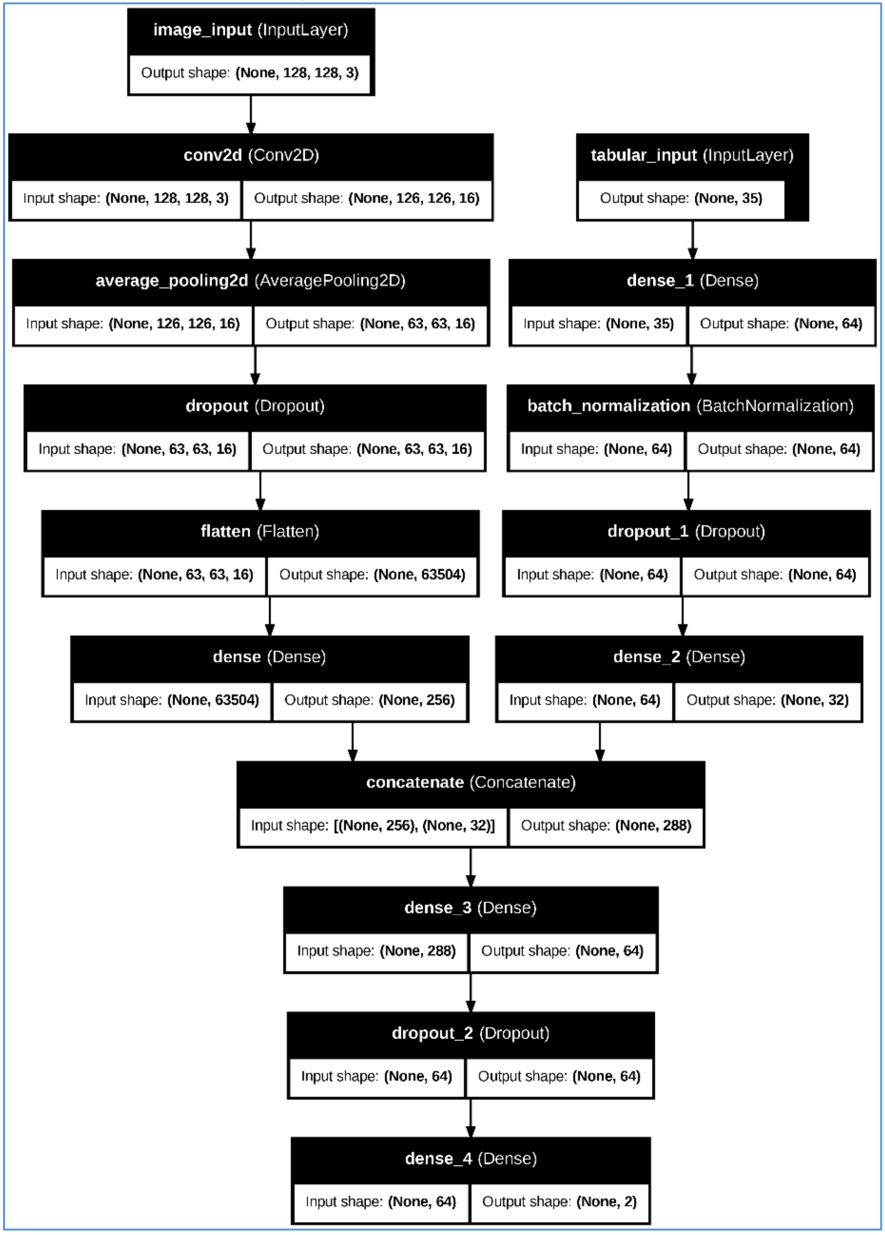

Figure 9 Architecture of the multimodal model combining Mel-Spectrogram-based CNN features and tabular clinical data for ALS classification

E. Evaluation

1) Performance Evaluation Results

The performance of the CNN models under various experimental scenarios and dropout rates is summarized in Table 4. Accuracy was calculated for all configurations and used as the primary metric to compare scenarios and guide model selection. Among all scenarios, S1B consistently achieved the highest performance, with a peak accuracy of 80.33 percent at a dropout rate of 0.3, outperforming deeper architectures and alternative pooling types. Although the models were also evaluated using precision, recall, and F1 score, the table reports only accuracy as a summary metric for clarity. Detailed evaluation including all four metrics, accuracy, precision, recall, and F1 score, will be presented later for the best-performing S1B model and the finalized multimodal model.

Table 4 Summary of model accuracy for all experimental CNN scenarios

Scenario Code	**Accuracy**						
	Dropout Rate (0,5)	**Dropout Rate (0,4)**	**Dropout Rate (0,3)**	**Dropout Rate (0,2)**	**Dropout Rate (0,1)**	**No Dropout**	**Average (Each Scenario)**
S1A	77,96%	77,46%	79,10%	77,55%	77,87%	77,87%	**77,97%**
S1B	78,37%	76,33%	**80,33%**	78,37%	76,64%	77,05%	77,85%
S2A	66,94%	66,94%	69,26%	73,88%	72,95%	76,23%	71,03%
S2B	66,94%	70,90%	68,44%	72,95%	75,92%	73,36%	71,42%
S3A	66,94%	66,94%	66,94%	66,80%	75,51%	74,18%	69,55%
S3B	66,94%	67,76%	68,16%	70,08%	72,95%	77,87%	70,63%
S4A	66,94%	66,94%	66,94%	66,94%	68,57%	75,10%	68,57%
S4B	66,94%	66,94%	66,94%	66,94%	69,80%	73,06%	68,44%
S5A	66,94%	66,94%	66,94%	66,94%	66,94%	76,23%	68,49%
S5B	66,94%	66,94%	66,94%	66,94%	67,35%	70,90%	67,67%
Average (Each Dropout and No Dropout)	69,19%	69,41%	70,00%	70,74%	72,45%	**75,19%**	

2) Effect of Dropout and Pooling Layer

Figure 10 illustrates the impact of varying levels of dropout layer probability on the performance of all the CNN models tested. Among all the model architecture configurations trained and evaluated, the model with scenario code S1B consistently achieved the highest accuracy. Its accuracy peaked at around 80.33% at a dropout probability of 0.3. It is concluded that the architecture consisting of a single layer of CNN network with AveragePooling effectively captures the discriminative features of the Mel-Spectrogram. This model is also able to maintain strong generalization under the regularization strategy of the dropout layer. In contrast, deeper models or their architectures with more and more complex ones (from S2-S5) show lower accuracy. The decreasing accuracy occurs especially at higher levels of dropout probability. This could be due to the increase in model complexity causing overfitting or difficulty learning from the relatively limited and less complex Mel-Spectrogram data. At low dropout layer probabilities (0-0.2), the performance of the deeper models shows a slight improvement but does not surpass that of S1B. Higher dropout values, from 0.4 to 0.5, lead to a decrease in performance in almost all models. This can also be explained by the regularization technique that uses dropout layers; excessive deactivation of neurons during training leads to underfitting. As a result, simpler models are sufficient for this task. Furthermore, increasing model complexity while simultaneously reducing the number of dropouts can result in counterproductive models. Deeper architectures are generally more sensitive to regularization and may fail to fully learn from the data. In this study, experiments with code S1A implementing MaxPooling performed worse than S1B implementing AveragePooling. Therefore, average pooling provides a more stable feature representation for ALS speech classification.

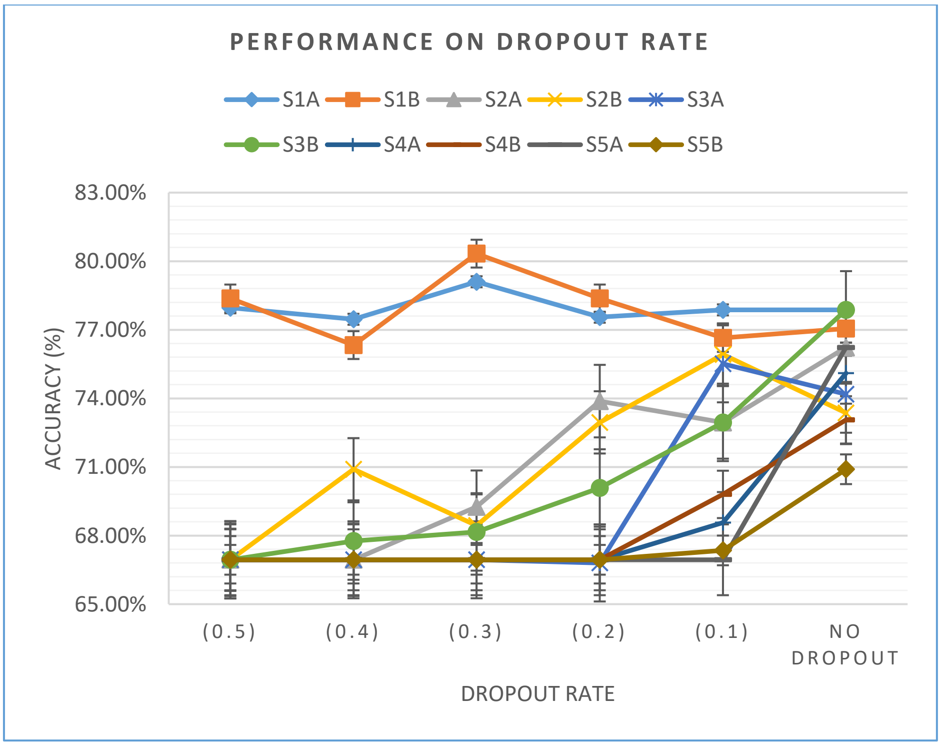

Figure 10 Effect of varying dropout rates on the accuracy of different CNN model configurations (S1–S5)

The average performance of all CNN configurations across varying dropout rates is illustrated in Figure 11. In the graph, the scenario with code S1 consistently outperforms other models with deeper architectures. It achieved an average accuracy of 77.97% with MaxPooling as its pooling layer. Furthermore, the model using AveragePooling as its pooling layer achieved 77.85%. Conversely, deeper architectures with 2-5 layers (S2-S5) showed increasingly lower accuracy. Here, the experiment with scenario code S5 only achieved 68.49% (for MaxPooling layer and 67.67% when using AveragePooling. The results show that increasing the depth or in this case adding layers and the number of kernels in the CNN model does not necessarily improve the performance of Mel-Spectrogram-based ALS sound classification. This could be because the spectral features in the audio are not too complex and can be captured effectively by a simpler architecture. In addition, for the influence of the type of pooling layer used in the model, MaxPooling is slightly superior to AveragePooling in the S1 code scenario. It can be concluded that the strategy of using the type of pooling layer has a small influence when compared to the overall model complexity. The findings in this study strengthen that a simpler CNN model with moderate pooling is sufficient to achieve stable and high performance on the dataset used in this study.

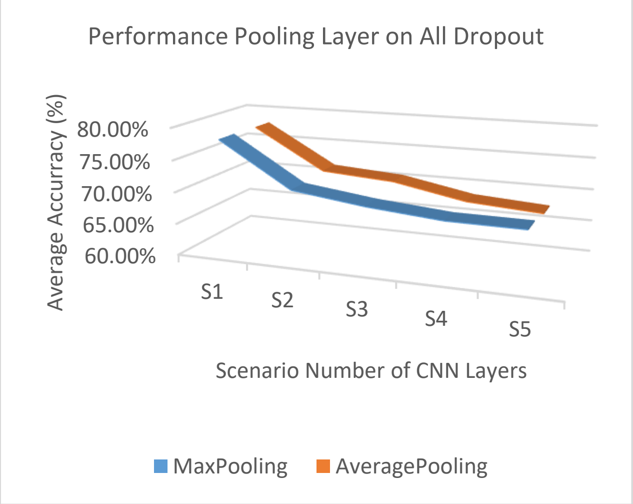

Figure 11 Performance of pooling layer

3) Multimodal Fusion Analysis

The evaluation results of the model based on the number of features used in the dataset are presented in Table 5. It can be seen that using only audio features, the model achieved an accuracy of 80.33%, with precision, recall, and F1-score ranging from 79–80%. This indicates that the voice features represented in the Mel-Spectrogram form provide sufficient discriminatory information for the ALS classification task. However, some misclassification errors still occur, so the performance is still not perfect. In contrast, by combining features from audio and clinical data, the model can achieve excellent performance in all metrics (accuracy, precision, recall, and F1-score of 100%). This phenomenon indicates that the integration of multimodal data substantially improves the model's discriminatory ability. These results also provide several important insights. First, although audio features are informative, the incorporation of complementary clinical data allows the model to resolve ambiguities that cannot be captured by the voice signal alone. Therefore, the model is able to clearly produce class boundaries between ALS and HC that are indeed completely separable. Second, the performance gap between single and multimodal inputs underscores the superiority of multimodal approaches in distinguishing ALS patients. This also suggests that even when heterogeneous information sources describe the same object, these data synergistically improve the predictive performance of models with multiple variables/features to consider during training and prediction. Finally, the findings of this study emphasize that for practical applications, models need to combine clinical indicators with voice features or even other features that could provide highly reliable predictions for ALS detection.

Table 5 Model performance using single and multimodal features

Feature	Performance			
	Precision	Recall	F1 Score	Accuracy
Audio	80,15%	80,33%	79,29%	80,33%
Audio & Clinical	100%	100%	100%	100%

Figure 12 presents the confusion matrix of the S1B model trained on audio features alone, while Figure 13 shows the confusion matrix of the same model trained on a combination of audio recordings and clinical data. These confusion matrices provide clearer insights into the misclassification patterns and the level of misprediction within each class. For the audio-only model, misclassification is clearly visible, with 12 ALS samples predicted as HC, and 36 HC samples misclassified as ALS. This indicates that while the Mel-Spectrogram feature captures useful discriminatory patterns for model training, the model still experiences ambiguity in distinguishing between ALS and HC. This is likely due to overlapping voice characteristics between some ALS patients and healthy individuals. In contrast, the model trained on multimodal data achieved perfect classification, with all ALS and HC samples correctly predicted. This demonstrates that integrating clinical features with audio signals provides complementary information, resolving ambiguity and significantly improving the model's ability to distinguish between ALS patients and healthy individuals. These results highlight that the advantage of a training approach using multimodal data in ALS detection (heterogeneous data sources) can produce fully separable class boundaries and maximize the predictive performance of a model.

ISSN:2302-8149
e-ISSN:2540-9719

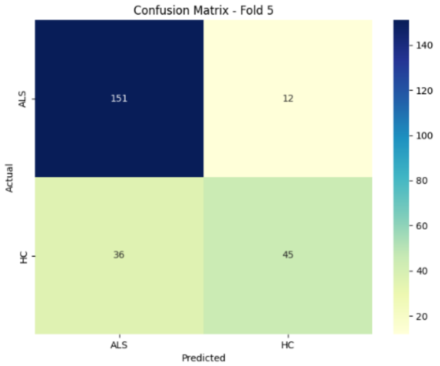

Figure 12 Confusion Matrix of the S1B model using audio features only

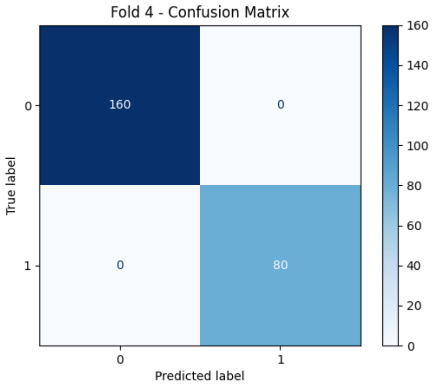

Figure 13 Confusion Matrix of the S1B model using a combination of audio and clinical records

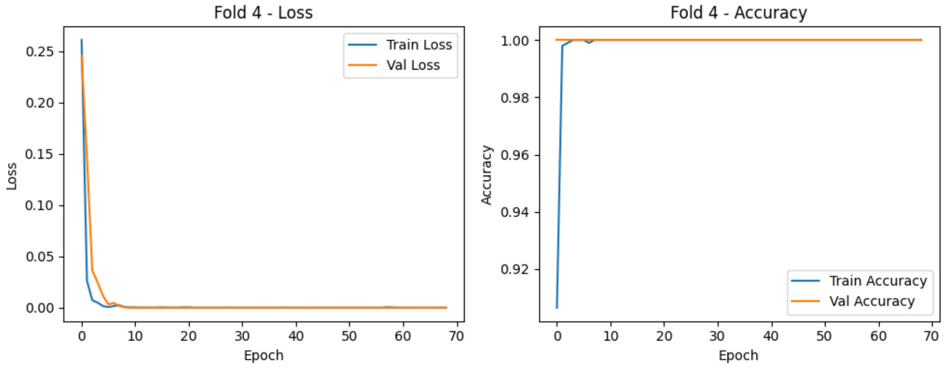

Figure 14 Loss and accuracy curves of the multimodal model on Fold 4

The multimodal model exhibits stable convergence across training epochs, as shown in Figure 14. The left figure illustrates the training loss, and the right figure shows its accuracy for Fold 4. This is because Fold 4 is the fold that yields the best model performance. As can be seen, the training and validation loss curves decrease rapidly and stabilize near zero after about ten epochs. This indicates effective learning without any indication of overfitting. Similarly, the accuracy curve remains consistently high, approaching 100% accuracy throughout the training process. These results confirm that the multimodal data fusion approach (Mel-Spectrogram and clinical data) can effectively achieve excellent generalization performance across validation folds. However, it is also worth noting that this near-perfect accuracy can also be attributed to the influence of clinical features that are highly informative for ALS cases. Conversely, some clinical features are missing or uninformative for Healthy Controls (HC), as these particular indicators or features are unique to ALS patients. This imbalance in feature richness gives the model a more robust classification capability for ALS detection.

5 Conclusion

Based on the study results, the diagnosis of Amyotrophic Lateral Sclerosis (ALS) using a deep learning model is highly effective. Training and evaluation using CNNs show that performance is highly dependent on the architecture selection based on the number of layers, dropout, and training scenario. The best model accuracy on audio-only data reached 80.33% with a CNN with one convolutional layer, average pooling, and a dropout probability of 0.3. This indicates the importance of designing a robust CNN for speech-based ALS detection. Furthermore, when combined with clinical data, the model achieved 100% accuracy. The main scientific contribution of this study lies in the establishment of an ALS diagnostic framework using multimodal data, namely integrating non-invasive speech biomarkers with clinical data, with results that significantly improve diagnostic accuracy. These findings confirm that while audio morphology provides valuable non-invasive biomarkers, integration with clinical data results in a more robust diagnostic system. However, challenges remain, particularly the limited number of healthy controls and the missing clinical features that are unique to ALS patients. Future research recommendations include expanding the size and balance of datasets, improving the comprehensiveness of clinical data, particularly for healthy patients, and exploring other, more sophisticated deep learning architectures. These architectures include attention-based models or transformer/transfer learning models validated using real-world data. This is expected to result in systems developed for ALS diagnostics that are more scalable and practical, and integrate explainable AI models to ensure the interpretability of diagnostic decisions.

Acknowledgement

The authors would like to thank the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia through the Directorate of Research, Technology, and Community Service for the financial support provided to support the implementation of this research. This research received a Beginner Lecturer Research grant in Fiscal Year 2025. The authors also thank the Institute for Research and Community Service (LPPM) of Bumigora University for the assistance provided during the implementation of the research.

References

[1] L. Migliorelli, L. Scoppolini Massini, M. Coccia, L. Villani, E. Frontoni, and S. Squartini, “A Deep Learning-based Telemonitoring Application to Automatically Assess Oral Diadochokinesis in Patients with Bulbar Amyotrophic Lateral Sclerosis,” *Comput. Methods Programs Biomed.*, Vol. 242, No. October, p. 107840, 2023, DOI: 10.1016/j.cmpb.2023.107840.

[2] G. Lauria, R. Curcio, and P. Tucci, “A Machine Learning Approach for Highlighting microRNAs as Biomarkers Linked to Amyotrophic Lateral Sclerosis Diagnosis and Progression,” *Biomolecules*, Vol. 14, No. 1, 2024, DOI: 10.3390/biom14010047.

[3] T. P. Umar *et al.*, “Artificial Intelligence for Screening and Diagnosis of Amyotrophic Lateral Sclerosis: A Systematic Review and Meta-Analysis,” *Amyotroph. Lateral Scler. Front. Degener.*, Vol. 25, No. 5–6, pp. 425–436, 2024, DOI: 10.1080/21678421.2024.2334836.

[4] M. Vidovic, L. H. Müschen, S. Brakemeier, G. Machetanz, M. Naumann, and S. Castro-Gomez, “Current State and Future Directions in the Diagnosis of Amyotrophic Lateral Sclerosis,” *Cells*, Vol. 12, No. 5, pp. 1–24, 2023, DOI: 10.3390/cells12050736.

[5] K. Imamura *et al.*, “Prediction Model of Amyotrophic Lateral Sclerosis by Deep Learning with Patient Induced Pluripotent Stem Cells,” *Ann. Neurol.*, Vol. 89, No. 6, pp. 1226–1233, 2021, DOI: 10.1002/ana.26047.

[6] M. Neumann *et al.*, “Investigating the Utility of Multimodal Conversational Technology and Audiovisual Analytic Measures for the Assessment and Monitoring of Amyotrophic Lateral Sclerosis at Scale,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, Vol. 4, pp. 3061–3065, 2021, DOI: 10.21437/Interspeech.2021-1801.

[7] F. Fernandes *et al.*, “Biomedical Signals and Machine Learning in Amyotrophic Lateral Sclerosis: A Systematic Review,” *Biomed. Eng. Online*, Vol. 20, No. 1, pp. 1–22, 2021, DOI: 10.1186/s12938-021-00896-2.

[8] K. G. Gwathmey *et al.*, “Diagnostic Delay in Amyotrophic Lateral Sclerosis,” *Eur. J. Neurol.*, Vol. 30, No. 9, pp. 2595–2601, 2023, DOI: 10.1111/ene.15874.

[9] R. Chauhan and U. Sharma, “Exploiting Speech Tremors: Machine Learning for Early Diagnosis of Amyotrophic Lateral Sclerosis,” *Eng. Res. Express*, Vol. 6, No. 4, 2024, DOI: 10.1088/2631-8695/ad7d62.

[10] B. G. Schultz *et al.*, “Automatic Speech Recognition in Neurodegenerative Disease,” *Int. J. Speech Technol.*, Vol. 24, No. 3, pp. 771–779, 2021, DOI: 10.1007/s10772-021-09836-w.

[11] M. Bowden *et al.*, “A Systematic Review and Narrative Analysis of Digital Speech Biomarkers in Motor Neuron Disease,” *npj Digit. Med.*, Vol. 6, No. 1, 2023, DOI: 10.1038/s41746-023-00959-9.

[12] R. Cave and S. Bloch, “The Use of Speech Recognition Technology by People Living with Amyotrophic Lateral Sclerosis: A Scoping Review,” *Disabil. Rehabil. Assist. Technol.*, Vol. 18, No. 7, pp. 1043–1055, 2023, DOI: 10.1080/17483107.2021.1974961.

[13] M. Antunes *et al.*, “A Morphology-based Feature Set for Automated Amyotrophic Lateral Sclerosis diagnosis on Surface Electromyography,” *Biomed. Signal Process. Control*, Vol. 79, No. P1, p. 104011, 2023, DOI: 10.1016/j.bspc.2022.104011.

[14] P. Rong, L. Heidrick, and G. L. Pattee, “A Multimodal Approach to Automated Hierarchical Assessment of Bulbar Involvement in Amyotrophic Lateral Sclerosis,” *Front. Neurol.* , Vol. 15, No. May, 2024, DOI: 10.3389/fneur.2024.1396002.

[15] R. Cebola, D. Folgado, A. Carreiro, and H. Gamboa, “Speech-based Supervised Learning Towards the Diagnosis of Amyotrophic Lateral Sclerosis,” Vol. 4, No. Biostec, pp. 74–85, 2023, DOI: 10.5220/0011694700003414.

[16] B. Tafuri *et al.*, “Machine Learning-based Radiomics for Amyotrophic Lateral Sclerosis Diagnosis,” *Expert Syst. Appl.*, Vol. 240, No. November 2023, p. 122585, 2024, DOI: 10.1016/j.eswa.2023.122585.

[17] M. Jamrozy, E. Maj, M. Bielecki, M. Bartoszek, M. Golebiowski, and M. Kuzma-Kozakiewicz, “Machine Learning Classificatory as a Tool in the Diagnosis of Amyotrophic Lateral Sclerosis using Diffusion Tensor Imaging Parameters Collected with 1.5T MRI Scanner: A Case study,” *Electron. J. Gen. Med.*, Vol. 20, No. 6, 2023, DOI: 10.29333/ejgm/13536.

[18] T. D. Kocar, A. Behler, A. C. Ludolph, H. P. Müller, and J. Kassubek, “Multiparametric Microstructural MRI and Machine Learning Classification Yields High Diagnostic Accuracy in Amyotrophic Lateral Sclerosis: Proof of Concept,” *Front. Neurol.*, Vol. 12, No. November, pp. 1–7, 2021, DOI: 10.3389/fneur.2021.745475.

[19] A. Tena, F. Clarià, F. Solsona, and M. Povedano, “Detecting Bulbar Involvement in Patients with Amyotrophic Lateral Sclerosis based on Phonatory and Time-Frequency Features,” *Sensors*, Vol. 22, No. 3, 2022, DOI: 10.3390/s22031137.

[20] O. P. Kurmi, M. Gyanchandani, N. Khare, and A. Pillania, “Comparative Analysis on Classification Efficiency of Deep Learning Models on Amyotrophic Lateral Sclerosis Patients using Speech Signals,” Vol. 7, No. 1, pp. 1–10, 2024.

[21] L. E. R. Simmatis, J. Robin, M. J. Spilka, and Y. Yunusova, “Detecting Bulbar Amyotrophic Lateral Sclerosis (ALS) using Automatic Acoustic Analysis,” *Biomed. Eng. Online*, Vol. 23, No. 1, pp. 1–13, 2024, DOI: 10.1186/s12938-023-01174-z.

[22] T. Bhattacharjee *et al.*, “Effect of Noise and Model Complexity on Detection of Amyotrophic Lateral Sclerosis and Parkinson’s Disease using Pitch and MFCC,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, Vol. 2021-June, pp. 7313–7317, 2021, DOI: 10.1109/ICASSP39728.2021.9413997.

[23] L. Wang *et al.*, “Automatic Prediction of Amyotrophic Lateral Sclerosis Progression using Longitudinal Speech Transformer,” pp. 2000–2004, 2024, DOI: 10.21437/interspeech.2024-158.

[24] K. H. An *et al.*, “Automatic Early Detection of Amyotrophic Lateral Sclerosis from Intelligible Speech using Convolutional Neural Networks,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, Vol. 2018-Septe, No. August, pp. 1913–1917, 2018, DOI: 10.21437/Interspeech.2018-2496.

[25] A. Jayakumar *et al.*, “Classification Between Patients with Amyotrophic Lateral Sclerosis and Healthy Individuals using Hypernasality in Speech : A Low Complexity Approach.”

[26] L. Ilias and D. Askounis, “Recognition of Dysarthria in Amyotrophic Lateral Sclerosis Patients using Hypernetworks,” 2025, [Online]. Available: http://arxiv.org/abs/2503.01892.

[27] C. Pancotti *et al.*, “Deep Learning Methods to Predict Amyotrophic Lateral Sclerosis Disease Progression,” *SCI. Rep.*, Vol. 12, No. 1, pp. 1–10, 2022, DOI: 10.1038/s41598-022-17805-9.

[28] F. Faghri *et al.*, “Identifying and Predicting Amyotrophic Lateral Sclerosis Clinical Subgroups: A Population-based Machine-Learning Study,” *Lancet Digit. Heal.*, Vol. 4, No. 5, pp. e359–e369, 2022, DOI: 10.1016/S2589-7500(21)00274-0.

[29] T. D. Ngo *et al.*, “An EEG & Eye-Tracking Dataset of ALS Patients & Healthy People During Eye-Tracking-based Spelling System Usage,” *SCI. Data*, Vol. 11, No. 1, pp. 1–11, 2024, DOI: 10.1038/s41597-024-03501-y.

[30] E. L. Feldman *et al.*, “Amyotrophic Lateral Sclerosis,” *Lancet*, Vol. 400, No. 10360, pp. 1363–1380, 2022, DOI: https://doi.org/10.1016/S0140-6736(22)01272-7.

[31] F. G. Vieira *et al.*, "A Machine-Learning based Objective Measure for ALS Disease Severity," *npj Digit. Med.*, Vol. 5, No. 1, 2022, Doi: 10.1038/s41746-022-00588-8.

[32] T. Segura *et al.*, "Symptoms Timeline and Outcomes in Amyotrophic Lateral Sclerosis using Artificial Intelligence," *SCI. Rep.*, Vol. 13, No. 1, pp. 1–10, 2023, DOI: 10.1038/s41598-023-27863-2.

[33] A. Behler, H. P. Müller, A. C. Ludolph, and J. Kassubek, "Diffusion Tensor Imaging in Amyotrophic Lateral Sclerosis: Machine Learning for Biomarker Development," *Int. J. Mol. SCI.*, Vol. 24, No. 3, 2023, DOI: 10.3390/ijms24031911.

[34] K. Fukushima *et al.*, "Early Diagnosis of Amyotrophic Lateral Sclerosis based on Fasciculations in Muscle Ultrasonography: A Machine Learning Approach," *Clin. Neurophysiol.*, Vol. 140, No. June, pp. 136–144, 2022, Doi: 10.1016/j.clinph.2022.06.005.

[35] R. Dubbioso *et al.*, "Voice Signals Database of ALS Patients with Different Dysarthria Severity and Healthy Controls," *SCI. Data*, Vol. 11, No. 1, pp. 1–14, 2024, DOI: 10.1038/s41597-024-03597-2.

[36] A. Catanese *et al.*, "Multiomics and Machine-Learning Identify Novel Transcriptional and Mutational Signatures in Amyotrophic Lateral Sclerosis," *Brain*, Vol. 146, No. 9, pp. 3770–3782, 2023, DOI: 10.1093/brain/awad075.

[37] C. Vildan, D. Sule, B. Turker, U. Hilmi, and K. B. Sibel, "Genetic Alterations of C9orf72, SOD1, TARDBP, FUS, and UBQLN2 Genes in Patients with Amyotrophic Lateral Sclerosis," *Cogent Med.*, Vol. 6, No. 1, p. 1582400, 2019, DOI: 10.1080/2331205x.2019.1582400.

[38] S. Edgar *et al.*, "Mutation Analysis of SOD1, C9orf72, TARDBP and FUS Genes in Ethnically-Diverse Malaysian Patients with Amyotrophic Lateral Sclerosis (ALS)," *Neurobiol. Aging*, Vol. 108, No. xxxx, pp. 200–206, 2021, DOI: 10.1016/j.neurobiolaging.2021.07.008.

[39] I. N. Switrayana, R. Hammad, P. Irfan, T. T. Sujaka, and M. H. Nasri, "Comparative Analysis of Stock Price Prediction using Deep Learning with Data Scaling Method," Vol. 7, No. 1, pp. 78–90, 2025.

[40] R. D. Alamsyah and S. Suyanto, "Speech Gender Classification using Bidirectional Long Short Term Memory," pp. 646–649, 2023.

[41] S. Chaudhary and D. Kumar Sharma, "Gender Identification based on Voice Signal Characteristics," pp. 869–874, 2018.

[42] I. N. Switrayana, S. Hadi, and N. Sulistianingsih, "A Robust Gender Recognition System using Convolutional Neural Network on Indonesian Speaker," Vol. 13, pp. 1008–1021, 2024.

[43] I. N. Switrayana and N. U. Maulidevi, "Collaborative Convolutional Autoencoder for Scientific Article Recommendation," *Proc. - 2022 9th Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2022*, pp. 96–101, 2022, DOI: 10.1109/ICITACEE55701.2022.9924130.

[44] 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, 2021, DOI: 10.5815/ijitcs.2021.06.05.

[45] E. Hassan, M. Y. Shams, N. A. Hikal, and S. Elmougy, The Effect of Choosing Optimizer Algorithms to Improve Computer Vision Tasks: A Comparative Study, Vol. 82, No. 11. Multimedia Tools and Applications, 2022.

[46] Y. Bai *et al.*, "Understanding and Improving Early Stopping for Learning with Noisy Labels," *Adv. Neural Inf. Process. Syst.*, Vol. 29, No. NeurIPS, pp. 24392–24403, 2021.

[47] Ž. Vujović, "Classification Model Evaluation Metrics," *Int. J. Adv. Comput. SCI. Appl.*, Vol. 12, No. 6, pp. 599–606, 2021, DOI: 10.14569/IJACSA.2021.0120670.