Deep learning-based volition detection and action potential extraction for fully automated diagnosis of neuromuscular disease using needle electromyography signals

Soomin Chung, Ilhan Yoo, Jinkyu Lee, Dongmin Kim, Kwangsoo Kim, Keewon Kim and Seung-Bo Lee

Abstract— *Objective*: This study aimed to develop a deep learning-based volition-detection model to automate the diagnostic process and improve neuromuscular disease classification using needle electromyography (nEMG) signals. *Methods*: The model was developed using 376 nEMG signals from 57 subjects and independently evaluated on an external dataset of 751 nEMG signals from 115 subjects at a single tertiary medical center. The proposed model directly processed raw nEMG signals to automatically extract volition signals (motor unit action potentials), eliminating the need for physician-dependent preprocessing. *Results*: The optimal segment length for volition detection was 0.060 s, and nEMGNet demonstrated the best overall performance. Model-extracted volition signals yielded higher segment-wise classification performance than raw nEMG signals, with area under the receiver operating characteristic curve (AUROC) values of 0.806 for myopathy, 0.819 for neuropathy, and 0.819 for normal cases, compared with 0.761, 0.728, and 0.733, respectively. At the patient-wise level, AUROC improved by 20.10% for myopathy, 23.15% for neuropathy, and 23.15% for normal cases compared with physician-detected volition data. *Conclusion*: The proposed volition-detection model significantly enhances neuromuscular disease classification performance while enabling a fully automated and less labor-intensive diagnostic workflow.

Index Terms— **Action potential extraction, Deep learning algorithm, Electromyography, Volition detection**

1 Introduction

ELECTROMYOGRAPHY (EMG) is an electrophysiological test that records electrical signals generated by skeletal muscles. Needle electromyography (nEMG) is an invasive form of EMG performed by inserting a needle electrode directly into the muscle. It is an important test for diagnosing neuromuscular diseases. Each nEMG signal represents the electrical activity generated during muscle contraction and is recorded in digital form. The signal amplitude reflects contraction strength, with larger amplitudes indicating stronger contractions while smaller amplitudes indicating weaker activity. nEMG waveforms consist of three components: volition signals, spontaneous activity, and noise (Supplemental Figure 1). The signal observed during voluntary muscle contraction is referred to as the volition signal or the motor unit action potential (MUAP). Neuromuscular diseases are broadly categorized into neuropathies and myopathies, and volition signals exhibit characteristic patterns that aid in distinguishing between these conditions and normal muscle activity [1]. Abnormal spontaneous activity can support disease progression assessment but does not substantially contribute to differentiating neuropathy from myopathy [2]. Noises arising from patient movement, needle displacement, or environmental factors have no diagnostic value and must be excluded. Consequently, accurate identification of volition signals is essential for neuromuscular disease classification. Neuropathies are typically characterized by large amplitude, long duration MUAPs with reduced recruitment, whereas myopathies exhibit MUAPs with smaller amplitudes, shorter durations, and early recruitment [1]. Normal MUAP durations range from 0.005 to 0.015 s, while myopathic MUAPs are generally shorter and neuropathic MUAPs may extend up to 0.030 s [2]. Because clinical

• *Soomin Chung is with the Interdisciplinary Program in Bioengineering, Seoul National University, Seoul 08826, Republic of Korea. (e-mail: soomin.chung.9910@gmail.com)*
• *Ilhan Yoo is with the Department of Neurology, Nowon Eulji Medical Center, Eulji University School of Medicine, Nowon-gu 01830, Seoul, Republic of Korea. (e-mail: kjfdf1@gmail.com)*
• *Jinkyu Lee is with the Department of Rehabilitation Medicine, Seoul National University Hospital, Seoul 03080, Republic of Korea. (e-mail: jkl3921@gmail.com)*
• *Dongmin Kim is with the Biomedical Research Institute, Seoul National University Hospital, Seoul 03080, Republic of Korea. (e-mail: dmk2436@gmail.com)*
• *Kwangsoo Kim is with the Department of Transdisciplinary Medicine, Institute of Convergence Medicine with Innovative Technology, Seoul National Univer-sity Hospital, Seoul 03080, Republic of Korea, and Department of Medicine, Seoul National University, Seoul 03080, Republic of Korea. (e-mail: kksoo716@gmail.com)*
• *Keewon Kim is with the Department of Rehabilitation Medicine, Seoul National University Hospital, Seoul 03080, Republic of Korea and Office of Vision, Seoul National University College of Medicine, Seoul 03080, Republic of Korea. (e-mail: keewonkimm.d@gmail.com)*
• *Seung-Bo Lee is with the Department of Medical Informatics, Keimyung Uni-versity School of Medicine, Daegu 42601, Republic of Korea. (e-mail: ko-reateam23@gmail.com)*
Please provide a complete mailing address for each author, as this is the address the 10 complimentary reprints of your paper will be sent

*Please note that all acknowledgments should be placed at the end of the paper, before the bibliography (**note that corresponding authorship is not noted in affiliation box, but in acknowledgment section**).*

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

symptoms alone often fail to distinguish between these disorders, precise identification of these MUAP features is critical for accurate diagnosis. However, MUAP interpretation is challenging: waveform patterns are transient during nEMG examinations, and their interpretation depends heavily on examiner experience, leading to variability in diagnostic outcomes.

Recent studies have explored artificial-intelligence (AI)–based approaches for neuromuscular disease classification using nEMG signals. Most prior work relied on conventional machine learning methods with handcrafted feature extraction [3],[4],[5],[6],[7]. Although some studies employed deep learning models, they did not use raw signals as input. Instead, they typically transformed raw nEMG signals into image-based representations, such as mel-spectrograms, before classification [8],[9]. In contrast, directly using raw nEMG signals offers several advantages: streamlined preprocessing improves computational efficiency; one-dimensional convolutions reduce complexity compared to two-dimensional operations; and direct signal processing avoids potential information loss from format conversion [10]. Moreover, learning directly from raw nEMG data preserves temporal continuity, enabling more effective modeling of intrinsic signal patterns and dynamics.

Yoo et al. [11] demonstrated successful neuromuscular disease classification using the signal format of nEMG in conjunction with a one-dimensional convolutional neural network. However, this approach required physician involvement during the data preprocessing stage. Specifically, a certified electromyographer manually extracted MUAPs from raw nEMG recordings by removing non-informative signal components. This procedure is not only labor-intensive and time-consuming but also introduces potential variability due to subjective examiner judgment.

Building on this work, we proposed a deep learning-based volition-detection model that automatically extracts MUAP components directly from raw nEMG signals. By eliminating the need for manual preprocessing, the proposed approach enables a fully automated end-to-end pipeline encompassing volition detection and subsequent neuromuscular disease classification using raw nEMG data.

2 METHODS

2.1 Data and Study Design

We used two datasets: a Development dataset derived from our previous study [11] and an External Validation dataset. The Development dataset consisted of 376 nEMG recordings from 57 subjects, with diagnostic labels for neuromuscular diseases—myopathy (M), neuropathy (N), and normal (NL)—assigned at the patient-wise level. Signals were acquired from both proximal and distal muscles, reflecting routine clinical practice in which electromyographers examine multiple muscle groups to establish comprehensive diagnostic patterns. The dataset also contained annotations specifying the start and end points of volition and non-volition segments for each signal. The Development dataset was randomly split at the patient-wise level into two subsets: 80% for model development and 20% for testing. Within the development subset, five-fold cross-validation was performed to optimize hyperparameters and select the final model. To assess the generalization performance of the proposed volition-detection model, we constructed an External Validation dataset consisting of 751 nEMG recordings from 115 subjects who underwent nEMG examinations at Seoul National University Hospital between August 2010 and November 2021 (Supplemental Figure 2). A certified electromyographer assigned diagnostic labels—myopathy, neuropathy, or normal—to each subject based on clinical records.

This study was approved by the institutional review board (No. 2008-055-1147) and conducted in accordance with the Declaration of Helsinki and its later amendments. Informed consent was waived due to the retrospective nature of the study. Data supporting the findings of this study are available from the corresponding author upon reasonable request.

The proposed framework consists of two sequential stages: Stage 1 performs binary volition detection to identify volition segments containing MUAP activity from raw nEMG signals. Stage 2 applies three-class neuromuscular disease classification to the detected volition segments (Figure 1). Both Stage 1 and Stage 2 models were trained using the Development dataset, whereas the External Validation dataset was used exclusively for independent testing. The following sections describe the preprocessing, model architecture, and experimental setup for each stage.

2.2 Preprocessing

The number and length of nEMG signals vary across patients due to clinical factors such as disease severity and muscle selection. In addition, interpretive criteria differ among examiners when determining whether signals are normal or abnormal and whether abnormalities indicate neuropathy or myopathy. Variability in individual nEMG signal characteristics also influences the number of muscles examined. Electromyographers select muscles based on patient history and suspected diagnosis: distal muscles are preferentially examined for neuropathies (e.g., first dorsal interosseous muscle in motor neuron disease), proximal muscles for myopathies where pathological changes are more pronounced, and specific myotomes for radiculopathies. When signal patterns are ambiguous, examiners may extend the examination to additional muscles for clarification. Consequently, signal segmentation was required to standardize variable-length nEMG recordings for input into deep learning models. Different segmentation strategies were applied for Stage 1 (volition detection) and Stage 2 (disease classification).

For volition detection, the optimal segment length was determined by evaluating model performance across segment durations ranging from 0.015 s to 0.080 s in increments of 0.005 s. This step size was chosen to balance fine-grained performance analysis with computational feasibility. During training, a hop size of 0.01 s was used to enable fine-resolution data augmentation. During inference, non-overlapping segmentation (hop size equal to segment

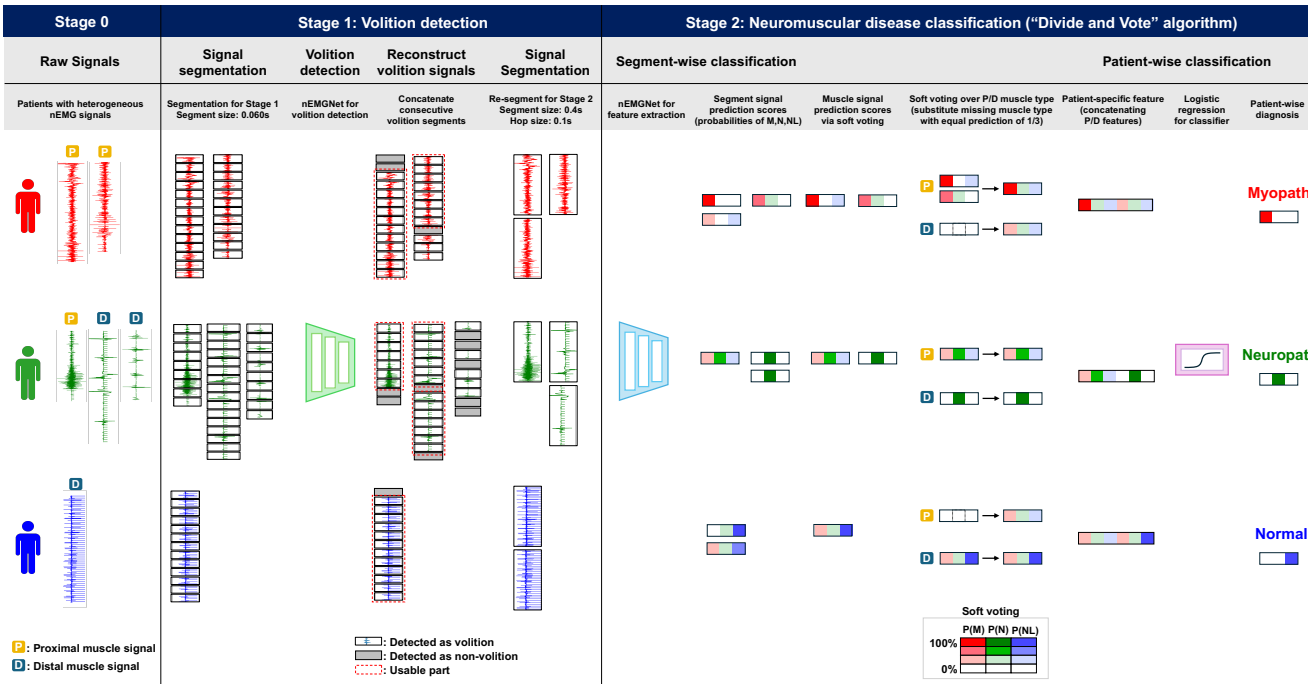

Fig. 1. Automated volition-detection-based pipeline for neuromuscular disease classification. Stage 1: volition detection and signal reconstruction using nEMGNet; Stage 2: hierarchical neuromuscular disease classification via the 'Divide-and-Vote' algorithm. Colored boxes in the prediction scores represent disease class probabilities (red: myopathy, green: neuropathy, blue: normal), with color intensity indicating probability magnitude. (M=Myopathy, N=Neuropathy, NL=Normal)

length) was applied to identify contiguous volition regions. Consecutively detected volition segments were then concatenated to reconstruct volition signals for subsequent disease classification. Segment labels were assigned based on whether they originated from volition or non-volition portions of the signal.

For neuromuscular disease classification, a segment length of 0.4 s and a hop size of 0.1 s were selected based on our previous optimization study [11], which demonstrated an optimal balance between capturing complete MUAP waveforms and preserving sufficient temporal resolution for classification. To reduce computational complexity while maintaining essential signal morphology, nEMG signals were downsampled from 48 kHz to 10 kHz. Segment labels corresponded to the patient-wise diagnosis.

Each patient had a single diagnostic label (myopathy, neuropathy, or normal), and all signals recorded from the same patient shared this label. Each signal contained multiple volition and non-volition parts. Segments inherited the patient-wise disease label and were additionally annotated as volition or non-volition depending on the signal region from which they were extracted.

2.3 Model

For volition detection, we compared the performance of several convolutional neural network (CNN) architectures to determine the optimal model. All CNN architectures were adapted to use one-dimensional convolutions, enabling direct processing of raw nEMG signals without requiring time-frequency transformations or signal reshaping. We evaluated representative models from the ResNet [12], DenseNet [13], and SqueezeNet [14] families, which are general-purpose architectures widely used in image classification, as well as nEMGNet [11], a model specifically designed for nEMG signal analysis. nEMGNet is a one-dimensional CNN designed for raw nEMG signal classification, inspired by VGGNet and ResNet architectures. The network consists of two main block types: (1) spatial reduction blocks that progressively downsample the input signal to capture hierarchical features and (2) residual blocks that incorporate skip connections to stabilize training and facilitate gradient flow (Supplemental Table I). The architecture processes the input signal through 12 convolutional blocks with progressively increasing channel dimensions (64 to 1024), followed by five fully connected layers (FC-512 to FC-3) with ReLU activations and a final softmax layer for three-class classification (Supplemental Table II). The overall performance of the full model families is summarized in Supplemental Table III, and the best-performing configuration from each family is presented in Figure 5.

For neuromuscular disease classification, we employed our previously proposed 'Divide-and-Vote' algorithm [11], which was designed to handle the heterogeneous data structure of nEMG examinations, wherein each patient may have a different number of signals from various muscle types with varying signal lengths. In this framework, the CNN-based feature extractor processes individual signal segments and produces segment-wise prediction scores for three classes (myopathy, neuropathy, and normal), referred to as the segment-wise result. Segment-wise prediction scores from segments belonging to the same muscle signal are then aggregated via soft voting to

generate a muscle-wise prediction. Muscle-wise predictions from all examined muscles of a single patient are subsequently combined using soft voting and concatenation to construct a patient-specific feature representation. This representation is then passed to a logistic regression classifier to produce the final patient-wise diagnostic label, referred to as the patient-wise result.

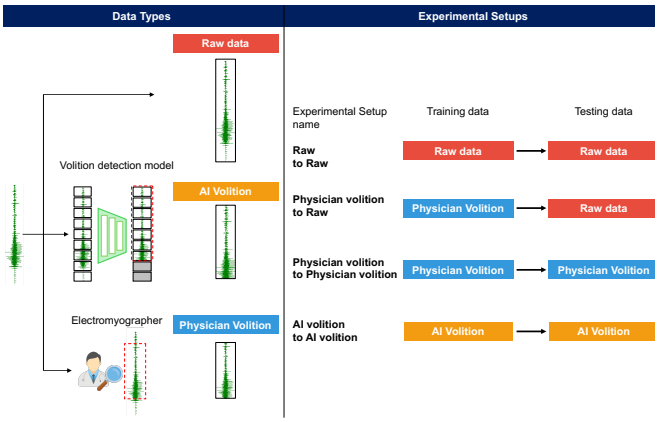

Fig. 2. Data types and experimental setups with corresponding train/test configurations for neuromuscular disease classification. Both the nEMGNet and logistic regression classifier are trained and tested according to each configuration. (Raw data= the unprocessed signal that contains both the volition and non-volition parts, Physician volition= volition signal manually obtained from physicians, AI volition= volition signal obtained through the volition-detection model.)

2.4 Experiments Setup

For neuromuscular disease classification experiments, the data were organized into three types according to the source of volition information: (1) Raw data, consisting of unprocessed nEMG signals that include both volition and non-volition regions; (2) Physician volition, comprising volition signals manually extracted by physicians; and (3) AI volition, consisting of volition signals automatically detected by the proposed volition-detection model (Figure 2). Using these data types, we designed multiple experimental configurations to address key research questions: (1) whether volition extraction improves neuromuscular disease classification performance, (2) whether AI volition outperforms raw signals, and (3) whether AI volition achieves performance comparable to Physician volition. Experimental setups were defined according to the combination of training and testing data types and are summarized in Figure 2. The corresponding performance comparisons are presented in Results Sections 3.2, 3.4, and 3.5, respectively. To ensure fair comparison across experimental setups, the same patient-wise level splits were used for training and testing in all experiments.

Hyperparameters were optimized using grid search within a nested five-fold cross-validation, with search spaces detailed in Supplemental Table IV. All models were trained for up to 100 epochs, with early stopping applied to prevent overfitting. Training employed the Adam optimizer [15] and cross-entropy loss. All hyperparameters were selected empirically, and all training was conducted on an NVIDIA RTX A6000 GPU.

For the volition-detection model, a learning rate of 1e-4 and a batch size of 32 were used. The optimal training epoch for each hyperparameter configuration was selected based on model sensitivity (TP/(TP + FN)), defined as the proportion of true volition segments correctly identified. The final model was chosen based on the highest average sensitivity across the five validation folds. For the neuromuscular disease classification model, separate hyperparameter searches were performed for each experimental setup. Model accuracy was used as the criterion for selecting the optimal training epoch, and the final model for each setup was chosen based on the highest average accuracy across the five validation folds. The resulting hyperparameter configurations are reported in Supplemental Table V. To mitigate class imbalance, class weights were applied in inverse proportion to the frequency of signal segments for each diagnostic category.

2.5 Statistical Analysis

Python 3.8.0 (Python Software Foundation, Wilmington, DE, USA) was used for signal preprocessing, model development and validation, statistical testing, and visualization. Model performance was evaluated using the area under the receiver operating characteristic curve (AUROC), accuracy, sensitivity (recall), positive predictive value (PPV; precision), and F1 score. AUROC comparisons were performed using the DeLong test. This test was applied to compare AUROCs across experimental setups, specifically between Raw-to-Raw and Physician-to-Raw, between Raw-to-Raw and Physician-to-Physician (Fig. 3), between Physician-to-Physician and AI-to-AI (Table III), and between Raw-to-Raw and AI-to-AI (Fig. 7). Signal quality metrics comparing Physician volition and AI volition (Supplemental Table VI) were analyzed using the Wilcoxon signed-rank test because the data were non-normally distributed. A two-sided p-value < 0.05 was considered statistically significant. Volition detection was formulated as a binary classification task and was evaluated using segment-wise classification results. Neuromuscular disease classification was treated as a three-class classification task. Performance metrics were computed using a one-versus-rest strategy for each class and are reported at both the segment-wise and patient-wise levels.

TABLE I
SUMMARY OF NEEDLE ELECTROMYOGRAPHY DATASET

		Myopathy	Neuropathy	Normal	Total
Development Dataset	Number of subjects	19	19	19	57
	Number of signals	121	160	94	375
	Total signal length (s)	312.84	422.78	203.5	939.12
External Validation Dataset	Number of subjects	18	7	90	115
	Number of signals	135	121	495	751
	Total signal length (s)	577.91	472.27	2041.99	3092.16

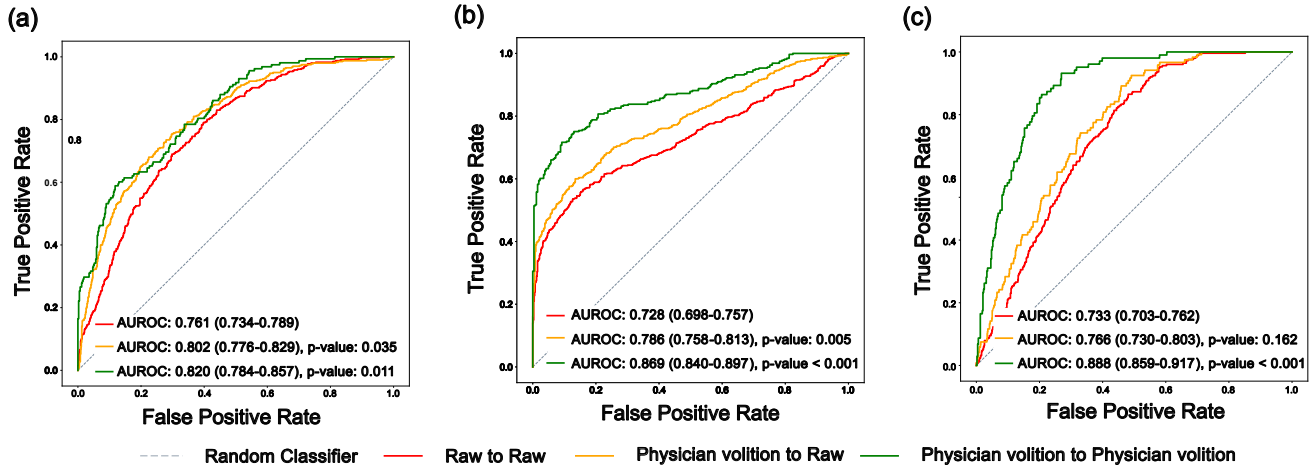

Fig. 3. ROC Curves for neuromuscular disease classification results for three different experimental setups. The three experimental setups are Raw to Raw: trained on raw data and tested on raw data (red); Physician volition to Raw: trained on physician volition and tested on raw data (yellow); and Physician volition to Physician volition: trained on physician volition and tested on physician volition (green). (a) ROC curve for myopathy class, (b) ROC curve for neuropathy class, and (c) ROC curve for normal class. The p-values were obtained using the DeLong test. (ROC=receiver operating characteristics, AUROC=area under the receiver operating characteristic curve).

3 Results

3.1 Data Characteristics

Table I summarizes the characteristics of the Development and External Validation datasets. The mean signal length was 2.498 ± 0.969 s in the Development dataset and 4.117 ± 1.363 s in the External Validation dataset.

The number of 0.4 s segments used as input for neuromuscular disease classification differed across data types: 4,295 segments for Raw data, 2,805 segments for Physician volition (manually extracted by certified electromyographers), and 1,981 segments for AI volition (automatically extracted by the volition-detection model). A detailed breakdown of segment counts is provided in Table II. Quantitative comparison of signal quality metrics between Physician volition and AI volition segments revealed no significant differences in voltage, signal-to-noise ratio, or signal variability (Supplemental Table VI). These results support the reliability of AI-based volition detection.

TABLE II
Number of Segments by Data Type

	Data Type	Myopathy	Neuropathy	Normal	Total
Development Dataset	Raw data	1,386	1,791	1,118	4,295
	Physician volition	936	1,273	596	2,805
	AI volition	685	877	419	1,981
External Validation Dataset	Raw data	1,821	1,432	6,318	9,571
	AI volition	830	464	1,569	2,863

(Raw data= the unprocessed signal that contains both the volition and non-volition parts, Physician volition= volition signal manually obtained from physicians, AI volition= volition signal obtained through the volition-detection model.)

3.2 Neuromuscular disease classification performance comparison between Raw data and Physician volition data

We compared three experimental setups—Raw to Raw (R-to-R), Physician volition to Raw, and Physician volition to Physician volition (P-to-P)—to evaluate the impact of volition extraction on neuromuscular disease classification performance.

In the R-to-R setup, segment-wise AUROCs were 0.761 (95% confidence interval [CI], 0.734–0.789) for myopathy, 0.728 (95% CI, 0.698–0.757) for neuropathy, and 0.733 (95% CI, 0.703–0.762) for normal. Physician volition to Raw test setup demonstrated higher performance than R-to-R, despite both setups yielding results on the same raw test data. AUROCs were 0.802 (95% CI, 0.776–0.829) for myopathy, 0.786 (95% CI, 0.758–0.813) for neuropathy, and 0.869 (95% CI, 0.840–0.897) for normal. Improvements for myopathy and neuropathy were statistically significant ($p < 0.05$). The P-to-P setup achieved the highest performance among all configurations, with AUROC of 0.820 (95% CI, 0.784–0.857) for myopathy, 0.869 (95% CI, 0.840–0.897) for neuropathy, and 0.888 (95% CI, 0.859–0.917) for normal. When comparing R-to-R and P-to-P, statistically significant differences were noted across all three classes with $p < 0.05$ (Figure 3).

3.3 Grid search for volition-detection model

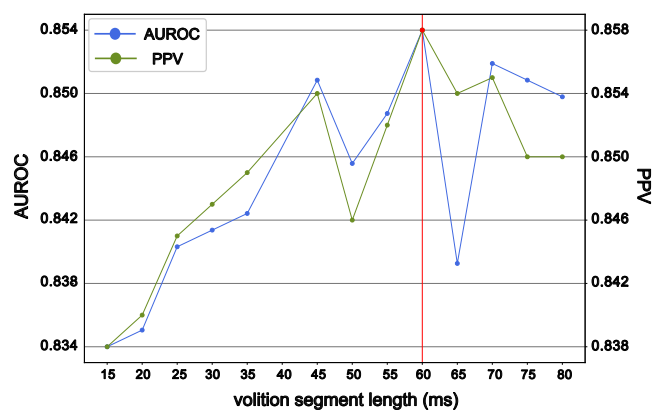

Fig. 4. Five-fold validation average AUROC and PPV for volition classification based on different volition segment length. (AUROC= Area under the receiver operating characteristic curve, PPV= positive predictive value)

Figures 4 and 5 show the results of the segment-wise binary classification, which determines whether an input segment is volition or non-volition.

Figure 4 shows that the signal segment length of 0.060 s achieved the highest performance in terms of both AUROC and PPV. Five-fold validation results demonstrate a gradual improvement in classification performance as the segment length increases from 0.015 s to 0.060 s, reaching peak performance at 0.060 s (AUROC, 0.854; PPV, 0.858). Performance declined beyond this point, and thus 0.060 s was selected as the final input segment length. After preprocessing, the dataset contained 74,622 volition segments and 40,318 non-volition segments.

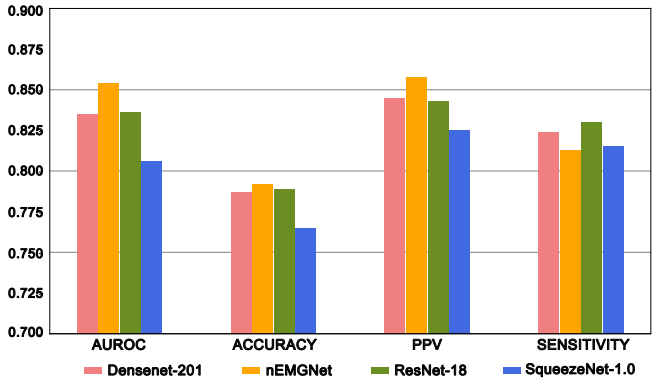

Fig. 5. 5-fold validation average sensitivity, AUROC, accuracy, and PPV for volition classification based on different model architectures. (AUROC= area under the receiver operating characteristic curve, PPV= positive predictive value)

Figure 5 presents the five-fold validation performance of different model architectures using the optimal segment length of 0.060 s. For each model family, the best-performing configuration is reported. nEMGNet consistently outperformed all other architectures across evaluation metrics. DenseNet-201 and ResNet-18 showed competitive performance, whereas SqueezeNet-1.0 achieved the lowest performance. Based on these results, nEMGNet was selected as the final architecture for the volition-detection model. Full performance results for each model family are provided in Supplemental Table III.

Figure 6 illustrates representative examples of classification results from the final volition-detection model. Figure 6a shows correctly identified volition segments, and Figure 6b presents segments classified as non-volition. All examples were randomly selected from the Development test set.

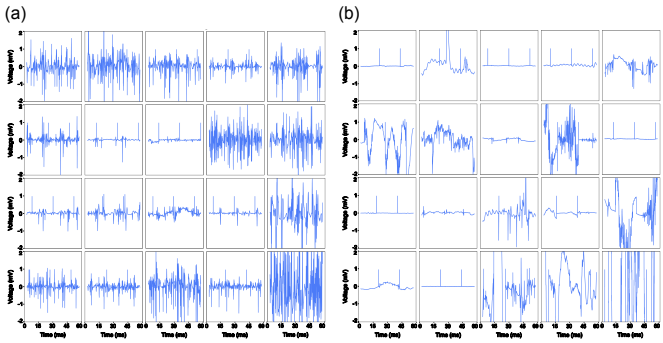

Fig. 6. Volition signal detected by volition-detection model. (a) shows the plots of 0.060 s segments detected as volition, and (b) shows the plots of 0.060 s segments detected as non-volition.

3.4 Comparison of neuromuscular disease classification performance between Raw data and AI volition

Figure 7 compares neuromuscular disease classification performance between R-to-R and AI volition to AI volition (A-to-A) on both the Development test set and External Validation dataset. On the Development test set, neuromuscular disease classification using AI volition (A-to-A) outperformed R-to-R across all three classes (myopathy: AUROC 0.806, 95% CI 0.746–0.865 vs. 0.761, 95% CI 0.734–0.789; neuropathy: AUROC 0.819, 95% CI 0.774–0.864 vs. 0.728, 95% CI 0.698–0.757; normal: AUROC 0.819, 95% CI 0.773–0.865 vs. 0.733, 95% CI 0.703–0.762). The improvements for neuropathy ($p < 0.001$) and normal ($p = 0.002$) were statistically significant (Figure 7a–c).

In the External Validation dataset, A-to-A also outperformed R-to-R for myopathy (AUROC 0.711, 95% CI 0.690–0.732 vs. 0.693, 95% CI 0.679–0.707) and neuropathy (AUROC 0.728, 95% CI 0.701–0.755 vs. 0.664, 95% CI 0.647–0.681). However, for the normal class, R-to-R showed slightly better performance (AUROC 0.602, 95% CI 0.581–0.623 vs. 0.644, 95% CI 0.633–0.656) (Figure 7d–f).

3.5 Comparison of neuromuscular disease classification performance between Physician volition and AI volition

TABLE III
SEGMENT-WISE AND PATIENT-WISE NEUROMUSCULAR DISEASE CLASSIFICATION RESULTS.

		Segment-wise		Patient-wise	
		Physician volition to Physician volition	AI volition to AI volition	Physician volition to Physician volition	AI volition to AI volition
Myopathy	Accuracy	0.796	0.738	0.750	0.833
	F1	0.612	0.511	0.667	0.667
	Recall	0.595	0.687	0.750	0.550
	AUROC	0.820	0.806	0.781	0.938
	p-value	0.682		0.361	
Neuropathy	Accuracy	0.801	0.786	0.833	0.917
	F1	0.798	0.786	0.667	0.857
	Recall	0.713	0.695	0.500	0.750
	AUROC	0.869	0.819	0.812	1.000
	p-value	0.071		0.350	
Normal	Accuracy	0.765	0.735	0.750	0.917
	F1	0.581	0.582	0.400	0.857
	Recall	0.424	0.463	0.250	0.750
	AUROC	0.888	0.819	0.812	1.000
	p-value	0.013		0.217	

Physician volition to Physician volition represents the results when using data obtained manually from electromyographer in the training and testing on the Development dataset, and AI volition to AI volition represent the results when using data obtained by the volition model in the training and testing on the Development dataset. The p-value was used to compare AUROCs between Physician volition to Physician volition and AI volition to AI volition. The p-values were obtained using the DeLong test. (ACC= Accuracy, AUROC= Area under the Receiver

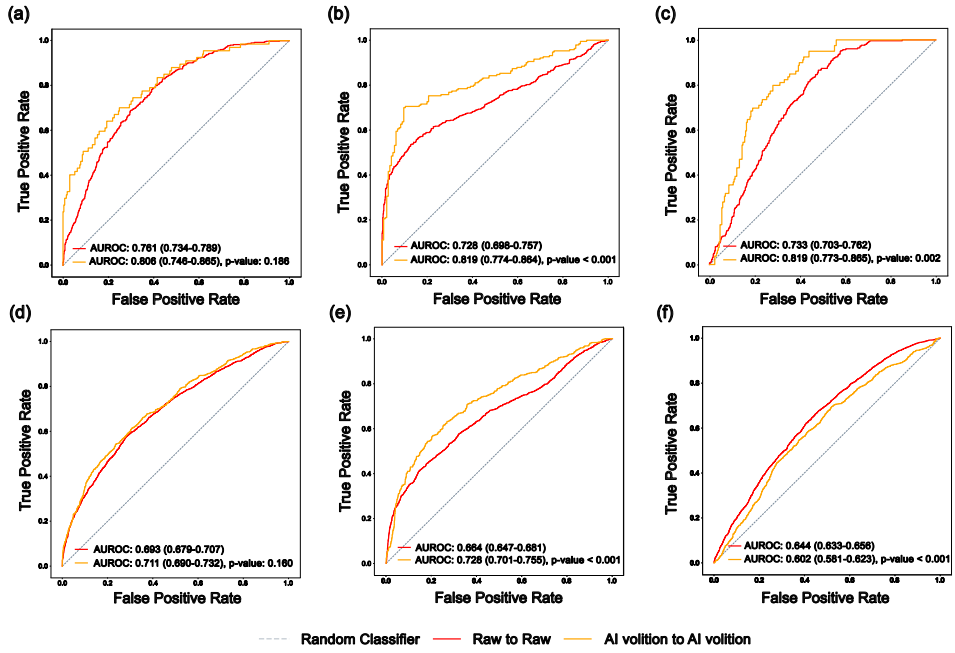

Fig. 7. ROC curve of neuromuscular disease classification results on Development test set and External Validation dataset. Comparison of performance between Raw to Raw: trained on raw data and tested on raw data (red); and AI volition to AI volition: trained on AI volition data and tested on AI volition data (yellow). (a)-(c): Results of myopathy class, neuropathy class, and normal class in Development test set, (d)~(f): Results of myopathy class, neuropathy class, and normal class in External Validation dataset. The p-values were obtained using the DeLong test. (ROC=receiver operating characteristics, AUROC=area under the receiver operating characteristic curve, AI= artificial intelligence)

Operating Characteristic curve, AI= artificial intelligence)

Table III compares the neuromuscular disease classification performance between P-to-P and A-to-A. In the segment-wise results, P-to-P generally achieved higher performance than A-to-A across most evaluated metrics. However, statistically significant differences were observed only for the normal class. In contrast, A-to-A outperformed P-to-P across most metrics in the patient-wise results.

To further investigate these differences, we examined the distribution of segment-wise prediction probabilities for each data type (Table IV). Within the Development test set, AI volition segments exhibited lower mean probabilities and reduced variability compared to Physician volition segments. Specifically, the average segment-wise probabilities for AI volition were 0.537 ± 0.285 for neuropathy, 0.682 ± 0.239 for myopathy, and 0.374 ± 0.133 for normal, whereas those for Physician volition were 0.542 ± 0.443 for neuropathy, 0.832 ± 0.304 for myopathy, and 0.573 ± 0.343 for normal. Statistically significant differences were observed for the myopathy and normal classes ($p < 0.05$).

TABLE IV
THE AVERAGE SEGMENT-WISE PROBABILITY VALUES ON THE DEVELOPMENT TEST SET

	AI volition to AI volition	Physician volition to Physician volition	p-value
Myopathy	0.682 ± 0.239	0.832 ± 0.304	<0.001
Neuropathy	0.537 ± 0. 285	0.542 ± 0.443	0.435
Normal	0.374 ± 0.133	0.573 ± 0.343	<0.001

Physician volition to Physician volition represents the results when using data obtained manually from an electromyographer in the training and testing on the Development dataset, and AI volition to AI volition represent the results when using data obtained by the volition model in the training and testing on the Development dataset. The p-values were calculated using the Wilcoxon rank sum test.

3.6 Early detection potential analysis

To assess the early detection potential of the proposed model, we applied it to 12,576 previously excluded nEMG signals that did not show obvious typical findings in the External Validation dataset (Supplemental Figure 2). For each disease class, the top 50 signals with the highest softmax output probability of the predicted class were selected and compared with clinical re-examination results. Each selected signal contained an average of 6.41 segments. The concordance rates were 54% for myopathy, 24% for neuropathy, and 98% for normal (Table V). Among the neuropathy predictions, 76% (38/50) were misclassified as normal rather than myopathy.

TABLE V
EARLY DETECTION ANALYSIS RESULTS

Predicted class	Sample size	Clinical re-examination results			Concordance rate
		Myopathy	Neuropathy	Normal	
Myopathy	50	27	1	22	54%
Neuropathy	50	0	12	38	24%
Normal	50	1	0	49	98%

4 Discussion and Conclusion

In this study, we developed an efficient volition-detection model capable of extracting volition signals from raw nEMG recordings with high performance.

Comparisons between R-to-R and P-to-P setups demonstrated that training and testing on refined Physician volition data resulted in significantly better neuromuscular disease classification performance than using Raw data. Moreover, results from Physician volition to Raw and P-to-P experiments indicated that even when using the same Physician volition data for training, testing on Physician volition data led to better performance compared to testing on Raw data (Figure 3). These findings confirm the critical role of volition signals in enhancing AI-based neuromuscular disease classification.

Importantly, within the context of using volition signals, A-to-A outperformed P-to-P in patient-wise results, which directly determine the final diagnostic outcome (Table III). Whereas Physician volition data requires considerable time and effort, AI volition data enables a fully automated end-to-end diagnostic pipeline. This pipeline takes raw nEMG signals as input, extracts volition signals, and subsequently performs neuromuscular disease classification without manual intervention. Collectively, these results indicate that AI-derived volition signals can provide both improved diagnostic performance and increased operational efficiency.

With respect to time efficiency, our quantitative analysis revealed a substantial potential reduction in processing time. Manual volition extraction by a certified electromyographer required an average of approximately 2 minutes per signal, whereas the proposed volition-detection model processed each signal in only 0.0003 seconds on average. Actual time savings in clinical practice may vary due to the complex nature of nEMG examinations, including patient instruction, environmental noise, and artifacts arising from needle insertion or involuntary movements [16]. Nevertheless, the proposed model may assist clinicians by facilitating more efficient identification of volition signals under such challenging conditions. Prospective clinical studies will be required to validate real-world time savings and clinical utility.

Notably, despite Raw data containing approximately 1.5 times more segments than Physician volition data in the Development dataset (Table II), classification performance using Raw data was inferior (Figure 7). This observation contrasts with the commonly held expectation that larger datasets inherently improve deep learning performance. Similarly, while there are 1.4 times more Physician volition segments in the Development dataset compared to AI volition segments, patient-wise neuromuscular disease classification using AI volition surpassed that using Physician volition (Table III). Together, these findings emphasize that data quality is critical for performance enhancement rather than sheer quantity, and highlight the importance of high-quality volition signal extraction in neuromuscular disease classification [17],[18].

To identify an optimal configuration for volition detection, we evaluated performance across different segment lengths and model architectures. Grid search analysis indicated that the optimal segment length for volition detection was 0.060 s (Figure 4). Given that MUAP durations range from 0.005-0.030 s across different neuromuscular disease types [2], a segment length of 0.060 s is sufficient to capture complete MUAP waveforms while limiting contamination from adjacent non-volition regions. Therefore, 0.060 s represents a reasonable and physiologically appropriate segment duration for volition detection.

Among the evaluated architectures, SqueezeNet, although efficient and effective for image-based tasks, exhibited comparatively lower performance when applied to nEMG signals (Figure 5). DenseNet and ResNet share the common design principle of reusing features from earlier layers. However, they differ in their layer-to-layer connection methods and information-integration techniques, with DenseNet using dense connections and ResNet using residual connections. In the volition-detection task, models incorporating residual connections demonstrated slightly superior performance relative to those using dense connections (Figure 5). nEMGNet was inspired by the structures of ResNet and another high-performance model. It was fine-tuned through experimentation to optimize the number and order of blocks to effectively extract rich information from nEMG signals. Its superior performance can therefore be attributed to its ability to leverage the strengths of the underlying model architectures, including ResNet, while being specifically tailored for nEMG signal processing (Figure 5).

The proposed volition-detection model effectively identified volition signals across a wide range of amplitudes (Figure 6a) and remained robust to diverse non-volition patterns, including noise and spontaneous activity (Figure 6b). Furthermore, neuromuscular disease classification using AI volition on the External Validation dataset outperformed that on Raw data, indicating that the model generalized well to data with previously unseen distributions (Figure 7d–f). Notably, the two datasets differed substantially in their curation methods and class distributions, reflecting key differences between controlled development environments and real-world clinical settings. The Development dataset was curated by a single expert electromyographer and exhibited relatively balanced class proportions (myopathy : neuropathy : normal = 685 : 877 : 419 in AI volition segments), whereas the External Validation dataset comprised consecutive, unselected clinical recordings with diagnostic labels derived from clinical reports and a markedly skewed distribution (myopathy : neuropathy : normal = 830 : 464 : 1,569 in AI volition segments), in which the normal class constituted the majority (54.8%). Despite these case-mix differences, AI volition consistently outperformed Raw data on the External Validation dataset, demonstrating that the proposed framework retains clinical utility even under the more demanding conditions of unselected clinical data.

While P-to-P achieved higher performance at the segment-wise level, A-to-A demonstrated superior performance at the patient-wise level (Table III). This observation may be explained by the mechanics of the Divide-and-Vote framework, which aggregates segment-wise predictions into signal-wise level and patient-wise level

representations and subsequently applies logistic regression to derive the final diagnosis. Examination of segment-wise prediction probability distributions (Table IV) revealed greater variability in Physician volition segments, including both excessively high probabilities and values near zero. This pattern suggests that Physician volition extraction may have included noise that was not adequately filtered out, leading to unreliable segment-wise predictions. Such severely misestimated segment probability values are likely to have influenced the training of the logistic regression model, resulting in reduced patient-wise performance of P-to-P. Through this analysis, we confirmed that the volition-detection model was more stable and accurate than physician-based detection. These findings support the growing body of evidence that AI-assisted data preprocessing can enhance robustness and downstream predictive performance in biomedical signal analysis [19],[20],[21],[22].

The early detection analysis demonstrated the potential of the proposed model to identify subtle pathological changes in clinically ambiguous cases (Table V). The high concordance rate for normal cases (98%) and strong inter-disease specificity indicates reliable model behavior in distinguishing non-pathological signals. Clinical review showed that normal signals misclassified as neuropathy predominantly contained large-amplitude artifacts that were not fully excluded during preprocessing. This occurred due to challenges at both the volition detection and classification stages, where short signal segments containing artifacts could resemble clean volition signals, and the model was not explicitly trained to distinguish such noise patterns from neuropathic activity. The comparatively stronger performance in myopathy detection suggests that the model developed higher sensitivity to interference-pattern characteristics, whereas neuropathy classification relied more heavily on amplitude-related features. Collectively, these findings highlight the model's potential to detect early pathological changes that may not be readily apparent during routine clinical assessment—particularly for myopathy—while emphasizing the need for improved artifact detection to further enhance neuropathy classification performance.

This study has several limitations. First, the dataset was relatively limited. It was derived from a single institution and did not include the full spectrum of neuropathies and myopathies. The study population consisted only of patients who underwent nEMG examinations at the hospital, thereby excluding individuals with very mild conditions who were unlikely to seek medical attention as well as severely ill patients who were unable to visit the hospital. In addition, rare neuromuscular diseases with very low prevalence were excluded, and the dataset primarily comprised patients with more common disease subtypes. The muscles examined were also restricted to those typically selected by clinicians for specific suspected conditions, which may not fully represent the diversity of neuromuscular presentations. Therefore, future studies should analyze more comprehensive datasets that include a broader range of patients across different disease severities and rare neuromuscular subtypes. Second, this study was limited by its retrospective design. Retrospective data collection may introduce selection bias toward patients with more severe presentations or toward disease types commonly evaluated at tertiary medical centers. Moreover, as a training hospital, the electromyographers conducting nEMG examinations change annually, which may result in variability in data quality, examination protocols, data storage methods, and criteria for data retention. These factors may limit the generalizability of our findings to broader clinical populations and practice settings. Accordingly, prospective studies are necessary to further validate the clinical utility of the proposed volition-detection model. Third, although our model demonstrated substantial computational efficiency—processing each signal in approximately 0.0003 seconds compared with an average of 2 minutes for manual extraction by physicians—the actual time savings in clinical practice require prospective validation. Real-world nEMG examinations involve complex factors, including patient cooperation, examination difficulty, and environmental conditions, which may influence overall examination time. Fourth, the early detection analysis highlighted areas for improvement, particularly in artifact detection and neuropathy classification accuracy. Enhancing the model's ability to exclude large-amplitude artifacts is expected to significantly improve neuropathy classification performance. Future research that more closely aligns model design with clinical diagnostic reasoning may further strengthen the system's early detection capability.

Despite these limitations, the proposed volition-detection model significantly improved neuromuscular disease classification performance. Unlike manual signal inspection and labeling by physicians, which is time-consuming and labor-intensive, the proposed approach enables a fully automated, end-to-end pipeline from raw nEMG signals to diagnostic classification, resulting in substantial time savings. These findings indicate that accurate automated volition extraction is an effective strategy for enhancing neuromuscular disease classification. With further validation and development, the proposed model has the potential to contribute meaningfully to AI-assisted EMG diagnostic systems in clinical practice.

ACKNOWLEDGMENT

This work was supported by the Technology Innovation Program (20016225, Development and Dissemination of Standard Reference Data) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea) and by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2022-00143911, AI Excellence Global Innovative Leader Education Program).
Corresponding author: Keewon Kim and Seung-Bo Lee.
Soomin Chung and Ilhan You are co-first authors.

REFERENCES

[1] K. R. Mills, “The basics of electromyography,” *J. Neurol. Neurosurg. Psychiatry,* vol. 76, no. suppl_2, pp. ii32-ii35, 2005-06-01, 2005.

[2] D. C. Preston, and B. E. Shapiro, *Electromyography and neuromuscular disorders e-book: clinical-electrophysiologic-ultrasound correlations*: Elsevier Health Sciences, 2020.

[3] H. Nodera, Y. Osaki, H. Yamazaki, A. Mori, Y. Izumi, and R. Kaji, “Classification of needle-EMG resting potentials by machine learning,” *Muscle Nerve,* vol. 59, no. 2, pp. 224-228, 2019-02-01, 2019.

[4] E. Gokgoz, and A. Subasi, “Comparison of decision tree algorithms for EMG signal classification using DWT,” *Biomed. Signal Process Control,* vol. 18, pp. 138-144, 2015.

[5] M. Kefalas, M. Koch, V. Geraedts, H. Wang, M. Tannemaat, and T. Bäck, "Automated machine learning for the classification of normal and abnormal electromyography data." pp. 1176-1185.

[6] M. U. Khan, S. Aziz, M. Bilal, and M. B. Aamir, "Classification of EMG signals for assessment of neuromuscular disorder using empirical mode decomposition and logistic regression." pp. 237-243.

[7] A. Subasi, E. Yaman, Y. Somaily, H. A. Alynabawi, F. Alobaidi, and S. Altheibani, “Automated EMG signal classification for diagnosis of neuromuscular disorders using DWT and bagging,” *Procedia Comput. Sci.,* vol. 140, pp. 230-237, 2018.

[8] H. Nodera, Y. Osaki, H. Yamazaki, A. Mori, Y. Izumi, and R. Kaji, “Deep learning for waveform identification of resting needle electromyography signals,” *Clin. Neurophysiol.,* vol. 130, no. 5, pp. 617-623, 2019.

[9] S. Nam, M. K. Sohn, H. A. Kim, H.-J. Kong, and I.-Y. Jung, “Development of artificial intelligence to support needle electromyography diagnostic analysis,” *Healthc. Inform. Res.,* vol. 25, no. 2, pp. 131-138, 2019.

[10] V. Alan, W. Ronald, and J. Buck, "Discrete-time signal processing: Prentic e Hall," USA, 2010.

[11] J. Yoo, I. Yoo, I. Youn, S.-M. Kim, R. Yu, K. Kim, K. Kim, and S.-B. Lee, “Residual one-dimensional convolutional neural network for neuromuscular disorder classification from needle electromyography signals with explainability,” *Comput. Methods Programs Biomed.,* vol. 226, pp. 107079, 2022.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition." pp. 770-778.

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." pp. 4700-4708.

[14] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size,” in International Conference on Learning Representations (ICLR), Toulon, France, 2017.

[15] D. P. Kingma, and J. Ba, “Adam: A method for stochastic optimization,” in International Conference on Learning Representations (ICLR), San Diego, CA, USA, 2014.

[16] M. Boyer, L. Bouyer, J.-S. Roy, and A. Campeau-Lecours, “Reducing noise, artifacts and interference in single-channel EMG signals: a review,” *Sensors,* vol. 23, no. 6, pp. 2927, 2023.

[17] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, “Understanding deep learning requires rethinking generalization,” in International Conference on Learning Representations (ICLR), Toulon, France, 2017.

[18] C. Scott, G. Blanchard, and G. Handy, "Classification with asymmetric label noise: Consistency and maximal denoising." pp. 489-511.

[19] E. Almazrouei, G. Gianini, N. Almoosa, and E. Damiani, "A deep learning approach to radio signal denoising." pp. 1-8.

[20] C. T. Arsene, R. Hankins, and H. Yin, "Deep learning models for denoising ECG signals." pp. 1-5.

[21] I. A. Khowailed, and A. Abotabl, “Neural muscle activation detection: A deep learning approach using surface electromyography,” *Journal of biomechanics,* vol. 95, pp. 109322, 2019.

[22] T. Pang, H. Zheng, Y. Quan, and H. Ji, "Recorrupted-to-recorrupted: Unsupervised deep learning for image denoising." pp. 2043-2052.

Soomin Chung received her MS degree from the Interdisciplinary Program in Bioengineering, Seoul National University in 2025, and her BS in Mechanical and Biomedical Engineering from Ewha Womans University in 2022. Her research interests include medical AI, bio signal and deep learning.

Ilhan Yoo received an MD degree from Ajou University, Suwon, Korea, in 2011. He worked as an assistant professor with the School of Medicine, The Eulji University, Korea, from 2022. His research interests include electromyography, amyotrophic lateral sclerosis, and machine learning.

Jinkyu Lee is currently a senior researcher in Department of Rehabilitation Medicine, Seoul National University Hospital. He received his BS in Computer Engineering from Sogang University in 2009 and Ph.D. in Mechanical Engineering from Sogang University in 2019. His research field includes neuromuscular biomechanics of human posture and movement, rehabilitation engineering and robotics, prosthetics, and orthotics.

Dongmin Kim received a BE degree in computer science from Sogang University in 2024. He worked at Seoul National University Hospital from 2020 to 2022. His interest lies in deep learning-friendly computer systems and web applications.

Kwangsoo Kim received an BS, MS and PhD degrees in Industrial Engineering from Korea University, Seoul, Korea in 2004, 2006 and 2011 respectively. He is currently a associate professor of Department of Transdisciplinary Medicine, Institute of Convergence Medicine with Innovative Technology, Seoul National University Hospital and a adjunct associate professor of Department of Medicine, College of Medicine, Seoul National University. His research interests include Biomedical Informatics, Digital Health, and Advanced Analytics.

Keewon Kim received his BS and MD degree from Seoul National University College of Medicine in 2002, MS degree with Interdisciplinary Program in Bioinformatics from Seoul National University in 2009, and PhD degree with Department of Biomedical Engineering from Seoul National University in 2016. He is currently a clinical professor of the Department of Rehabilitation medicine, Seoul National University Hospital, adjunct professor of Department of Intelligence and Information Graduate School of Convergence Science and Technology Seoul National University, and adjunct researcher of Medical Big data research center, Seoul National University College of Medicine. His research field includes neurophysiology, electromyography, intraoperative monitoring, motion analysis and musculoskeletral rehabilitation.

Seung-Bo Lee is currently an assistant professor in the Department of Medical Informatics at the Keimyung University School of Medicine. He received his PhD in Brain and Cognitinve Engneering from Korea University in the year 2020. His primary research focus encompasses various aspects of medical artifical intelligence, pariculary in diagnostic applications. he primarrily conducts deep learning-based analyses using physiological signals, but he is also expanding hsih interests to modalities such as ultrasound, CT, and MR imaging.