Clinical Hand Gesture Recognition in Intubated ICU Patients Using CNN-LSTM under Leave-One-Subject-Out Evaluation

Emy Setyaningsih
Department of Computer Systems Engineering, Akprind University, Indonesia
emysetyaningsih@akprind.ac.id (corresponding author)

Erma Susanti
Department of Informatics, Akprind University, Indonesia
erma@akprind.ac.id

Septiana Fathonah
Department of Emergency and Critical Care Nursing, Notokusumo School of Health Sciences, Yogyakarta, Indonesia
septiana.f@stikes-notokusumo.ac.id

Taukhit
Department of Nursing Management, Notokusumo School of Health Sciences, Yogyakarta, Indonesia
taukhit@stikes-notokusumo.ac.id

Received: 13 April 2026 | Revised: 19 May 2026 | Accepted: 3 June 2026

 DOI: https://doi.org/10.48084/etasr.19302

ABSTRACT

Non-verbal communication remains challenging for conscious intubated ICU patients who are unable to speak. This study investigates whether bedside hand gestures can be recognized from real ICU video recordings using a spatio-temporal deep learning framework. A private dataset comprising 20 videos from 10 intubated patients was organized into five clinically relevant gesture classes: Agreement, Discomfort, Neutral, RequestHelp, and Suction. Four ImageNet-pretrained CNN backbones, namely DenseNet121, ResNet50, MobileNetV2, and EfficientNetB0, were integrated with an LSTM layer to capture temporal gesture dynamics. Model performance was assessed under strict Leave-One-Subject-Out (LOSO) evaluation to measure generalization to unseen patients. Among the models evaluated, EfficientNetB0–LSTM achieved the best overall performance, with an accuracy of 0.40 and a micro-AUC of 0.5669. However, class-wise discrimination remained uneven, and all models showed limited sensitivity to minority and visually subtle gesture classes, with predictions frequently biased toward RequestHelp. These findings indicate that hand gesture recognition from real bedside ICU videos is considerably more challenging than recognition in controlled settings and provide an initial subject-independent benchmark for gesture-based communication support in critical care.

Keywords-clinical hand gesture recognition; CNN-LSTM; intubated ICU patients; leave-one-subject-out evaluation; non-verbal communication

I. INTRODUCTION

For intubated patients in the Intensive Care Unit (ICU), the inability to speak creates a major barrier to timely and effective communication. Even when consciousness is preserved, patients may struggle to express pain, discomfort, anxiety, or urgent care needs, which can complicate bedside interaction and increase the risk of misunderstanding in critical care practice [1-3]. Existing reviews have shown that conventional communication aids are not always sufficient, particularly when patients experience fatigue, fluctuating alertness, weakness, or impaired motor coordination [4-6]. Among observable bedside signals, hand gestures represent a practical non-verbal communication channel because they may convey intentional meaning without relying on speech, writing, or dedicated assistive devices [7, 8].

From a technical perspective, hand gesture recognition has advanced considerably with deep learning. Convolutional Neural Networks (CNNs) are widely used to extract spatial

representations, whereas temporal models are needed when gestures evolve across video frames [9-12]. Long Short-Term Memory (LSTM) networks remain relevant in this context because they aggregate motion-related information over time and help distinguish gestures that may appear similar in isolated frames. Accordingly, CNN-LSTM architectures have been applied in gesture-recognition and video-analysis tasks in which temporal continuity contributes to class discrimination [13-15]. Prior studies have also suggested that multimodal and adaptive learning strategies may improve robustness under variable gesture conditions [16, 17].

Despite these advances, most hand gesture recognition studies have been developed using controlled datasets. Such datasets typically involve healthy participants, fixed viewpoints, clearly defined gesture taxonomies, and relatively clean visual conditions [18-22]. These assumptions differ substantially from bedside ICU recordings. In real ICU environments, gestures may be weak, incomplete, or interrupted, and the visible hand region may be partially obscured by blankets, medical tubing, caregiver activity, or camera angle. In addition, recording conditions may vary across patients and sessions. As a result, models developed on non-clinical datasets cannot be expected to generalize reliably to bedside ICU videos characterized by occlusion, subtle motion, and pronounced inter-subject variability [8, 23, 24].

A further challenge concerns the evaluation protocol. In many vision-based recognition studies, random partitioning is still used to construct training and test sets. However, for subject-structured clinical data, this strategy may blur the distinction between learning gesture-related patterns and learning person-specific characteristics. This issue is particularly problematic in small datasets, where repeated exposure to samples from the same individual may lead to overly optimistic performance estimates. When the intended application involves previously unseen patients, subject-independent evaluation is more appropriate. Leave-One-Subject-Out (LOSO) evaluation provides a practical framework for this purpose because each fold reserves one subject for testing while the remaining subjects are used for model development [25, 26].

Subject-independent evaluation has also proven important in prior ICU-centered visual analysis. Earlier work in the same clinical setting showed that performance estimates became more conservative, yet more clinically realistic, when identity overlap between training and testing data was removed [27]. At the same time, facial observation alone may not fully capture non-verbal communication in intubated ICU patients, since visible expressions can be reduced and facial regions are often partially obscured by medical devices. This motivates the investigation of hand gestures as a complementary communication modality. In the broader direction of this research, facial and gesture analysis are treated as complementary components of a multimodal non-verbal communication model for intubated ICU patients. Despite this relevance, ICU hand gesture recognition under strict subject-independent evaluation remains insufficiently explored.

This study addresses this gap by evaluating hand gesture recognition in intubated ICU patients using a CNN-LSTM framework under strict subject-independent evaluation. The purpose of this work is not to introduce a new deep learning architecture, but to provide a realistic experimental reference for subject-independent gesture recognition under challenging bedside ICU conditions. A private clinical video dataset was organized into clinically meaningful gesture classes, and multiple ImageNet-pretrained CNN backbones were assessed within a shared temporal modeling framework so that their behavior could be compared under the same evaluation protocol.

The contributions of this study are fourfold. First, it presents a hand gesture dataset derived from real bedside recordings of intubated ICU patients. Second, it applies a strict LOSO protocol so that the reported results reflect generalization to unseen subjects. Third, it compares multiple CNN backbones within a unified CNN-LSTM framework to establish a fair spatio-temporal baseline. Fourth, it analyzes the resulting limitations to clarify the methodological difficulty of gesture recognition in ICU environments characterized by small data, class imbalance, occlusion, and inter-patient variability.

II. MATERIALS AND METHODS

This section describes the clinical dataset, preprocessing pipeline, CNN-LSTM architecture, training configuration, and evaluation protocol. Particular emphasis is placed on subject independence, fair backbone comparison, and reproducibility under real ICU bedside conditions.

A. Clinical Dataset

1) Data Acquisition

Clinical video data were collected from intubated patients receiving treatment in the ICU of Gadjah Mada University Academic Hospital (RSA UGM). The study protocol was approved by the institutional Research Ethics Committee (Ethics Approval No. 072/RSA/KEP/EC/2025), and informed consent was obtained from all patients or their legal representatives. All videos were anonymized before analysis to protect patient confidentiality.

Patient inclusion was restricted to subjects with Richmond Agitation-Sedation Scale (RASS) scores between −1 and +1, indicating sufficient alertness and motor responsiveness for observable hand gestures. Data collection was conducted in real bedside ICU conditions using a standard smartphone camera without controlling illumination, camera position, or background. The final dataset consists of 20 videos from 10 intubated ICU patients. Because the recordings contain sensitive clinical information, the dataset is not publicly released, although de-identified data may be provided upon reasonable request and institutional approval. Table I summarizes the distribution of the original expert-annotated labels.

TABLE I. DISTRIBUTION OF ORIGINAL EXPERT-ANNOTATED CLINICAL HAND GESTURE VIDEOS

Expert-annotated clinical hand gesture labels	Number of videos	Number of patients
Itching	2	2
Call	3	3
Request Help	4	4
Pointing to Pain	1	1
Neutral	2	1
Pain	1	1
Agreement	2	2
Suction	5	5
Total number of videos	20	—

2) Annotation and Class Mapping

The original expert annotations consisted of several fine-grained gesture labels. To reduce label fragmentation and improve analytical stability in this small-data setting, semantically related labels were consolidated into five final classes: Agreement, Discomfort, Neutral, RequestHelp, and Suction. Specifically, Itching and Pain were grouped into Discomfort, whereas Call, Request Help, and Pointing to Pain were grouped into RequestHelp. This mapping was intended to preserve clinical meaning while reducing class sparsity and overlap among visually related gesture categories. The final class mapping is presented in Table II.

TABLE II. MAPPING FROM ORIGINAL ANNOTATIONS TO FINAL GESTURE CLASSES

Final Class	Original Label(s)	No. of Videos	No. of Patients	Clinical Meaning
Agreement	Agreement	2	2	Affirmative or response gestures
Discomfort	Itching, Pain	3	3	Pain- or itch-related distress gestures
Neutral	Neutral	2	1	Baseline or no intentional movement
RequestHelp	Call, Request Help, Pointing to Pain	8	8	Gestures indicating a need for assistance
Suction	Suction	5	5	Gestures indicating airway discomfort or suction need
Total		20		

The final class distribution remains small and imbalanced, reflecting the natural composition of real bedside ICU recordings. These characteristics were retained to preserve clinical realism and to support evaluation under a strict subject-independent evaluation protocol. Accordingly, the objective of this study is not to maximize absolute predictive performance, but to establish a realistic baseline for cross-patient hand gesture recognition under clinically constrained conditions.

B. Video Preprocessing

Each video was converted into a fixed-length temporal sequence using uniform sampling. Specifically, every video was resampled into $T = 20$ frames, resized to 224×224 pixels, and normalized using ImageNet statistics. The resulting input tensor is defined as:

$$X \in R^{B\times T\times H\times W\times C} \quad (1)$$

where B denotes the batch size, $T = 20$ is the temporal length, $H = W = 224$ are the spatial dimensions, and $C = 3$ corresponds to RGB channels. Channel-wise normalization was then applied as:

$$\tilde{x}_{t,c} = \frac{x_{t,c}-\mu_c}{\sigma_c} \quad (2)$$

where $x_{t,c}$ is the pixel value of channel c at time step t, while μ_c and σ_c refer to the corresponding ImageNet. This preprocessing step ensures a consistent spatial and temporal representation before feature extraction and temporal modeling.

C. Proposed CNN–LSTM Framework

The proposed framework consists of three stages: subject-wise LOSO partitioning, spatio-temporal feature extraction through a CNN-LSTM pipeline, and aggregation of predictions across evaluation folds, as illustrated in Figure 1.

1) Phase 1: Subject-Wise LOSO Partitioning

As shown in the left block of Figure 1, the dataset consists of $S = 10$ subjects. In each fold, all samples from one subject are assigned to the test set, whereas the remaining subjects form the development set. This partitioning is defined as:

$$D = D_{train}^{(s)} \cup D_{test}^{(s)}, \quad D_{train}^{(s)} \cap D_{test}^{(s)} = \emptyset \quad (3)$$

where $D_{test}^{(s)}$ denotes the test data of subject s, and $D_{train}^{(s)}$ denotes the remaining data used for training and validation. An 80:20 split is then applied within the development set to create training and validation subsets. This procedure ensures that the model is always evaluated on a previously unseen patient.

2) Phase 2: CNN–LSTM Processing Pipeline

As illustrated in the upper-right block of Figure 1, each input video is represented as a temporal sequence:

$$X = \{x_t\}_{t=1}^{T}, \quad x_t \in R^{224\times224\times3}, \quad T = 20 \quad (4)$$

where T denotes the number of sampled frames, and each frame x_t is represented as an RGB image of size 224×224. Each frame is processed by a shared CNN encoder to extract spatial features. After Global Average Pooling (GAP), the resulting frame-level representation is defined as:

$$z_t = f(x_t) \in R^d \quad (5)$$

where $f(.)$ denotes the CNN feature extractor, and d is the feature dimension determined by the selected backbone. Four ImageNet-pretrained CNN backbones were evaluated within the same framework: DenseNet121, ResNet50, MobileNetV2, and EfficientNetB0.

The resulting sequence of feature vectors is then passed to an LSTM layer to capture temporal dependencies, as defined in

$$(h_t, c_t) = LSTM(z_t, h_{t-1}, c_{t-1}) \quad (6)$$

where h_t and c_t denote the hidden state and cell state, respectively. For completeness, the internal LSTM operations are defined in (7)–(12):

$$i_t = \sigma(W_i z_t + U_i h_{t-1} + b_i) \quad (7)$$

$$f_t = \sigma(W_f z_t + U_f h_{t-1} + b_f) \quad (8)$$

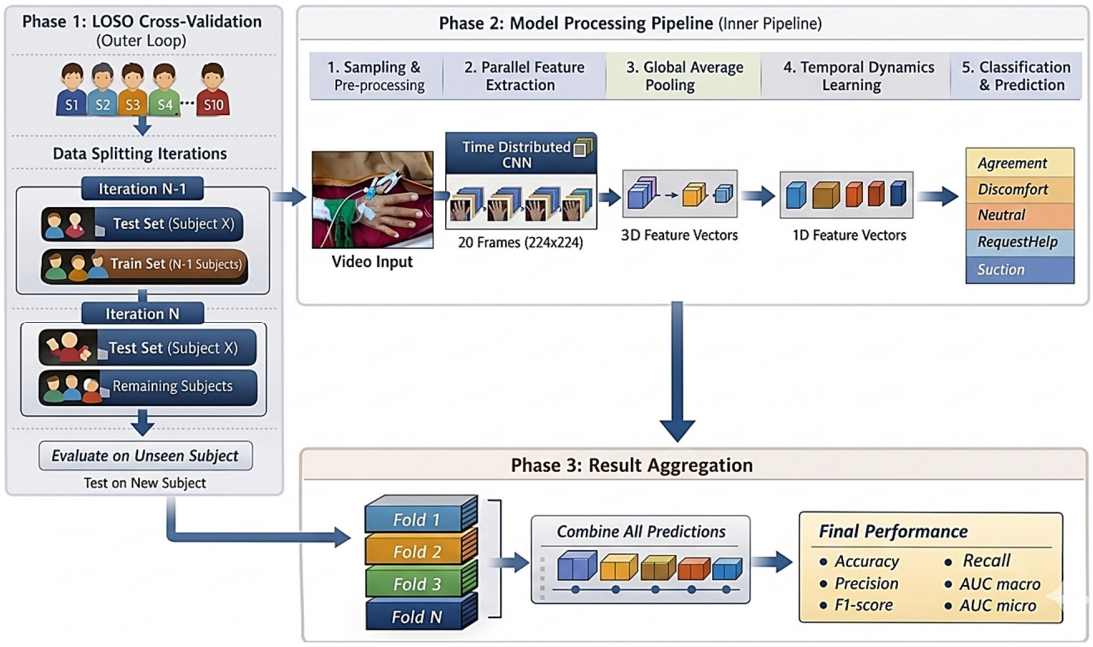

Fig. 1. Overall workflow of the proposed subject-independent hand gesture recognition framework, consisting of LOSO evaluation, model processing pipeline, and fold-wise result aggregation.

$$O_t = \sigma(W_o z_t + U_o h_{t-1} + b_o) \quad (9)$$

$$\tilde{c}_t = tanh(W_c z_t + U_c h_{t-1} + b_c) \quad (10)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (11)$$

$$h_t = o_t \odot tanh(c_t) \quad (12)$$

where $\sigma(.)$ denotes the sigmoid activation and $\odot$ denotes element-wise multiplication. Temporal modeling is particularly relevant because many ICU hand gestures are subtle and may not be reliably separated from single-frame appearance alone. Therefore, motion development across time provides additional discriminative information.

The final hidden state h_T is used as a compact representation of the video. Classification is then performed using a fully connected layer followed by a Softmax function:

$$o = W h_T + b \quad (13)$$

$$\hat{y} = softmax(o) \quad (14)$$

where $\hat{y} \in R^K$ and $K = 5$ is the number of gesture classes. The predicted label is obtained by:

$$\hat{y} = \arg\max_k \hat{y}_k \quad (15)$$

3) Phase 3: Fold-Wise Prediction Aggregation

Predictions from all LOSO folds are aggregated into a pooled prediction set:

$$\hat{y} = \bigcup_{s=1}^{S} \hat{y}^{(s)} \quad (16)$$

where $\hat{y}^{(s)}$ denotes the predictions obtained from fold s. Predictions from all LOSO folds are aggregated into a pooled set, from which all final evaluation metrics are computed to provide a global estimate of subject-independent performance across unseen patients.

D. Experimental Setup

1) Training Strategy

All models were trained under identical settings to ensure a fair comparison across CNN backbones. The Adam optimizer was used with categorical cross-entropy loss, a batch size of 4, 64 LSTM units, and a dropout rate of 0.4. Training followed a two-stage transfer learning strategy. In the first stage, all ImageNet-pretrained CNN layers were frozen, and only the LSTM and classifier layers were trained for 10 epochs using a learning rate of 1×10^{-4}. In the second stage, the final convolutional block was unfrozen and fine-tuned for 20 epochs with a reduced learning rate of 1×10^{-5} to adapt the spatial features to the clinical dataset while reducing the risk of catastrophic forgetting. This procedure was applied consistently to DenseNet121, ResNet50, MobileNetV2, and EfficientNetB0. No data augmentation was applied to preserve the original visual characteristics of bedside ICU recordings.

2) Hyperparameter Configuration

Table III summarizes the main hyperparameter settings used in all experiments. All backbone models were trained under identical configurations to support a fair comparison.

3) Implementation Details

All experiments were implemented in Python 3.12.13 using TensorFlow 2.20 and the Keras API. A hybrid computational setup was used in this study. Data preprocessing and initial script development were conducted locally on an Apple MacBook Air equipped with an M4 chip and 24 GB of unified memory, whereas the computationally intensive model training and evaluation phases were performed on Google Colab using an NVIDIA T4 Tensor Core GPU with 16 GB VRAM and approximately 12 GB of system RAM. A fixed random seed of

42 was applied to Python, NumPy, and TensorFlow to support reproducibility. Subject-wise LOSO partitions were generated before preprocessing to prevent data leakage. All videos were uniformly sampled into 20 frames, resized to 224×224 pixels, and normalized using ImageNet statistics. Identical preprocessing, training, and evaluation procedures were applied across all models.

TABLE III. TRAINING HYPERPARAMETER CONFIGURATION

Component	Parameter	Value / Description
Input	Sequence length	20 frames per video
	Resolution	$[224 \times 224]$
	Tensor form	$[B, 20, 224, 224, 3]$
Training	Batch size	4
Backbone	CNN models	DenseNet121, ResNet50, MobileNetV2, EfficientNetB0
Temporal module	LSTM units	64
Regularization	Dropout rate	0.4
Optimization	Optimizer	Adam
Optimization	Loss function	Categorical cross-entropy
Training	Phase 1 (Frozen)	10 epochs
Training	Phase 2 (Fine-tuning)	20 epochs
Evaluation	Subject-independent evaluation	LOSO across 10 subjects
Evaluation	Metrics	Accuracy, Precision, Recall, F1-score, Macro-AUC, Micro-AUC

E. Evaluation Metrics

Performance was evaluated using accuracy, precision, recall, F1-score, macro-AUC, and micro-AUC. Due to the limited number of test samples in each LOSO fold, predictions from all held-out folds were pooled before metric computation, providing an overall estimate of cross-subject generalization while preserving subject-independent evaluation.

III. RESULTS AND DISCUSSION

The experimental results indicate that subject-independent hand gesture recognition from bedside ICU videos is substantially more challenging than recognition under conventional data-splitting settings. Under the LOSO protocol, each test fold contains data from a previously unseen patient, requiring the models to generalize across inter-patient variability rather than familiar subject-specific patterns. Therefore, the reported results should be interpreted as conservative but clinically meaningful estimates of cross-patient performance.

A. Quantitative Performance under LOSO Evaluation

Table IV summarizes the quantitative performance of the evaluated CNN-LSTM models. EfficientNetB0-LSTM achieved the best pooled LOSO result, with an accuracy of 0.40 and a micro-AUC of 0.5669. ResNet50-LSTM and DenseNet121-LSTM showed lower but still above-baseline performance, whereas MobileNetV2-LSTM remained close to chance level in the five-class setting. Despite relative differences across backbones, the overall scores remained modest, indicating that reliable multi-class discrimination for previously unseen ICU patients remains challenging. The low macro-averaged metrics further suggest uneven class-wise performance, confirming that the main challenge lies not only in backbone selection but also in achieving robust generalization under realistic ICU variability.

TABLE IV. QUANTITATIVE PERFORMANCE OF CNN–LSTM MODELS UNDER LOSO EVALUATION (AGGREGATED TEST PREDICTIONS ACROSS ALL FOLDS)

	DenseNet121-LSTM	**ResNet50-LSTM**	**MobileNetV2-LSTM**	**EfficientNetB0-LSTM**
Accuracy	0.2500	0.3000	0.2000	0.4000
Precision	0.0625	0.0667	0.0533	0.0800
Recall	0.1250	0.1500	0.1000	0.2000
F1-score	0.0833	0.0923	0.0696	0.1143
AUC Macro	0.1641	0.1328	0.1790	0.1535
AUC Micro	0.5563	0.5656	0.5325	0.5669

Since each LOSO fold contains only a small number of test samples, the reported metrics should be interpreted cautiously. Following prior patient-independent ICU visual analysis, the 95% confidence intervals are reported as descriptive indicators of fold-wise variability rather than formal inferential statistics [28]. For the best-performing EfficientNetB0-LSTM model, the mean LOSO accuracy was 0.40, with a 95% confidence interval of 0.14–0.66 computed from the ten fold-wise accuracy values. This wide interval reflects substantial variability across held-out subjects, consistent with the small sample size, class imbalance, and heterogeneous bedside ICU conditions. Formal significance testing was not performed because LOSO folds are not fully independent, and the number of subjects is limited. Therefore, differences among backbone models are interpreted as overall trends rather than evidence of strict statistical superiority.

B. Error Pattern Analysis Using Confusion Matrices

The confusion matrices in Figure 2 show a recurrent tendency to overpredict RequestHelp, suggesting that broader and more visible gestures strongly influenced the learned decision boundaries. Discomfort and Suction were also frequently confused, likely because both involve subtle and clinically related motion patterns under similar bedside conditions. Neutral remained unstable due to limited intentional movement and the small number of samples. Overall, the models appeared more sensitive to broad motion salience than to fine-grained gesture differences, consistent with occlusion, weak articulation, and inter-subject variability in ICU recordings.

C. ROC-Based Analysis of Discriminative Behavior

The ROC analysis (Figure 3) reveals a clear discrepancy between the micro-AUC and macro-AUC values obtained by the evaluated models. While micro-AUC values ranged from 0.53 to 0.57, macro-AUC values remained much lower, typically between 0.13 and 0.18. This contrast is informative because it indicates that the models retain some ranking ability when predictions are aggregated across all samples, but fail to maintain comparable discriminative behavior across all gesture classes.

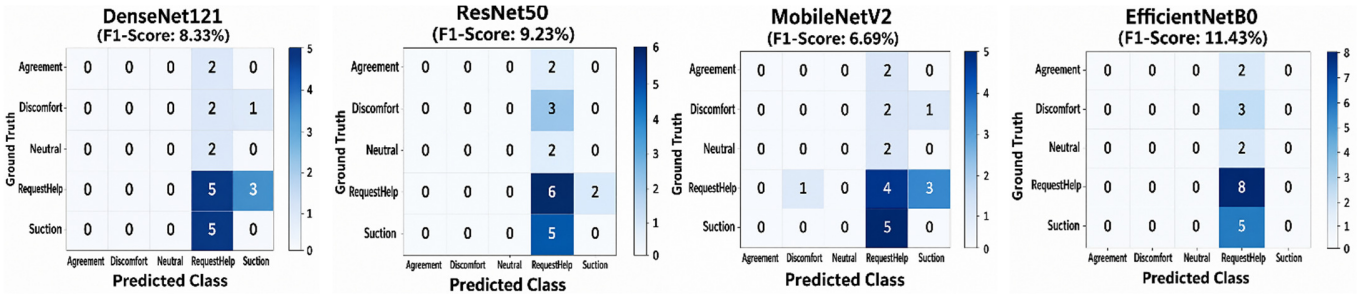

Fig. 2. Confusion matrices of the evaluated CNN–LSTM models under LOSO evaluation.

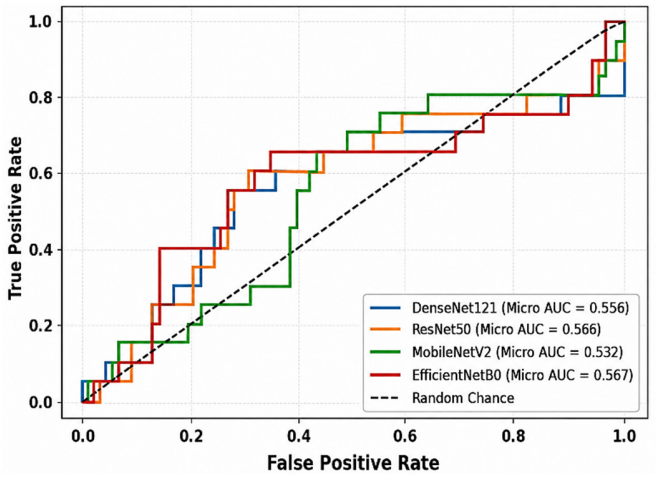

Fig. 3. Comparison of micro-average ROC curves across backbone architectures.

In practical terms, the higher micro-AUC suggests that some dominant or visually salient classes can still be ranked above chance. However, the much lower macro-AUC shows that this ability is not distributed evenly across the label space. Minority and ambiguous classes remain poorly separated, which is consistent with the observed confusion patterns and with the strong class imbalance of the dataset. Under LOSO evaluation, this imbalance becomes more pronounced because inter-subject differences introduce distribution shifts between the training and test data. Differences in patient condition, gesture execution style, camera viewpoint, and bedside context are therefore likely to affect underrepresented classes more severely than the dominant ones.

D. Impact of Dataset Size, Class Imbalance, and Subject Variability

The results of this study should be interpreted in light of several limitations related to dataset size, class imbalance, and subject variability. The dataset comprises only 20 videos from 10 intubated ICU patients, resulting in a limited number of training and test samples in each LOSO fold. Under this condition, performance estimates are sensitive to subject composition and gesture availability; therefore, the reported results should be regarded as an initial but clinically informative benchmark rather than a definitive estimate of deployment-level performance.

Class imbalance further affects model behavior. RequestHelp accounts for the largest proportion of samples, whereas Agreement and Neutral are represented by only a small number of videos. This distribution likely biases the learning process toward dominant classes and contributes to the low macro-averaged precision, recall, and F1-scores observed across all models. The confusion matrices in Figure 2 support this interpretation, showing a recurrent tendency to overpredict RequestHelp while minority and visually subtle classes remain difficult to distinguish.

The limited number of subjects also restricts the ability of deep learning models to learn subject-invariant representations. Gesture appearance may vary across patients due to differences in motor responsiveness, gesture amplitude, camera viewpoint, and bedside context. With only a small number of subjects, these sources of variability cannot be fully represented during training, which likely contributes to the modest subject-independent performance.

Accordingly, future studies should prioritize larger and more balanced datasets, preferably through multi-center data collection, to support more robust model development and more reliable evaluation under realistic ICU conditions.

E. Implications for Model Design

These results suggest that temporal modeling alone is not sufficient to fully address the challenges of clinical hand gesture recognition in ICU environments. Although the CNN-LSTM framework is more appropriate than frame-based analysis, performance remains constrained by weak spatial localization and limited sensitivity to subtle motion patterns. Under these conditions, gesture recognition depends not only on temporal aggregation, but also on the quality of the spatial representation extracted from clinically degraded video frames. Future improvements are likely to require more explicit modeling of hand regions and motion dynamics. Potential directions include hand detection and tracking, motion-aware representations such as optical flow or 3D convolution, more expressive temporal models including transformer-based architectures, and multimodal integration with physiological or contextual signals. However, such strategies will be meaningful only if they are evaluated under the same subject-independent protocol and supported by larger, more diverse bedside datasets.

TABLE V. COMPARISON WITH REPRESENTATIVE PRIOR WORKS

Ref.	Context	Data setting	Evaluation protocol	Relevance to this study
[8]	Healthcare-oriented gesture assistance	Healthcare gesture data, but not intubated ICU bedside videos	Not explicitly subject-independent	Demonstrates the clinical relevance of gesture-based interaction, but under less restrictive conditions
[16]	Multimodal gesture recognition	Sensor-based and multimodal datasets	Modality-dependent evaluation	Shows the potential benefit of combining complementary modalities, unlike the strictly vision-based setting used here
[17]	Adaptive deep learning for gesture recognition	General non-clinical gesture datasets	Conventional evaluation	Methodologically relevant, but not assessed under ICU-specific variability
[20]	Transformer-based dynamic gesture recognition	Controlled gesture dataset with clearly defined motions	Standard validation	Illustrates the potential of stronger temporal modeling under cleaner experimental conditions
This study	Non-verbal communication in intubated ICU patients	Real bedside ICU videos, 20 videos from 10 patients, imbalanced classes	Strict LOSO subject-independent evaluation	Provides a clinically grounded cross-patient benchmark under realistic ICU conditions

F. Positioning as a Clinically Grounded Benchmark

The findings should be interpreted in relation to the data source. Unlike many prior gesture-recognition studies conducted under controlled conditions, this work used real ICU bedside videos collected from intubated patients. Consequently, gesture amplitude, visibility, occlusion, and recording context vary substantially across samples. Under such conditions, even a spatio-temporal deep learning model may struggle to identify stable class-specific patterns. The main contribution of this study lies not in achieving high predictive accuracy but in clarifying the difficulty of gesture recognition under realistic ICU constraints and providing an initial subject-independent reference for future work.

G. Comparative Positioning Relative to Recent State-of-the-Art Methods

Recent State-Of-The-Art (SOTA) methods for hand gesture recognition have employed healthcare-oriented gesture assistance, transformer-based temporal modeling, multimodal sensor fusion, and adaptive deep learning strategies [8, 16, 17, 20]. These approaches have reported promising results, mostly on controlled datasets, healthy participants, clearly articulated gestures, or richer sensing modalities. However, subject-independent evaluation is not consistently enforced, making direct comparison with clinically oriented applications difficult. In contrast, this study evaluates gesture recognition from real bedside ICU videos under strict LOSO evaluation, where all test samples come from previously unseen patients. Under this more challenging setting, EfficientNetB0-LSTM achieved the best overall performance, with an accuracy of 0.40 and a micro-AUC of 0.5669. These modest results reflect the difficulty of the task rather than a limitation of the model architecture alone. Performance is constrained by severe class imbalance, subtle and incomplete gestures, frequent occlusion, and pronounced inter-subject variability. Accordingly, the main contribution of this study is not to outperform existing methods, but to provide a clinically grounded subject-independent benchmark for future approaches, including transformer-based, multimodal, and healthcare-oriented gesture-recognition models, under realistic ICU conditions.

Table V provides a concise comparison between this study and representative prior works in terms of study context, data setting, evaluation protocol, and practical relevance. Overall, the comparison indicates that the modest performance observed in this study is closely related to the stricter subject-independent protocol and the greater visual and clinical complexity of real ICU bedside data.

IV. CONCLUSION

This study evaluated vision-based hand gesture recognition for intubated ICU patients using a CNN-LSTM framework under strict LOSO subject-independent evaluation. Using real bedside ICU video data, the study provides a clinically realistic assessment of cross-patient generalization. EfficientNetB0-LSTM achieved the best overall performance, with an accuracy of 0.40, a micro-AUC of 0.5669, and a descriptive 95% confidence interval of 0.14–0.66, indicating substantial fold-wise variability across unseen subjects. The findings confirm that ICU hand gesture recognition remains challenging due to small and imbalanced data, subtle gestures, occlusion, and inter-subject variability. Thus, the modest performance should be interpreted as a consequence of the strict LOSO protocol and the complexity of real ICU bedside data, rather than as a limitation of the CNN-LSTM architecture alone.

The main contribution of this work is the establishment of an initial subject-independent baseline for hand gesture recognition in intubated ICU patients under realistic clinical conditions. Future work should focus on larger and more balanced datasets, improved hand localization, motion-aware representations, and multimodal integration to support more robust AI-assisted non-verbal communication in ICU environments.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to DRTPM (Directorate of Research and Community Service), Ministry of Higher Education, Science, and Technology (Kemdiktisainstek), for funding this work through the 2025 Applied Research Program under Research Grant Contract No. 126/C3/DT.05.00/PL/2025.

DATA AVAILABILITY

The clinical video dataset is not publicly available due to privacy and ethical restrictions. De-identified data may be provided by the corresponding author upon reasonable request and institutional approval. The source code, preprocessing scripts, and LOSO split definitions are also available upon reasonable request.

AI USE AND DECLARATION OF GENERATIVE AI USE

A generative AI tool was used only for limited assistance in improving language clarity, organization, and the conceptual layout of the manuscript's visual presentation. All scientific content, technical elements, workflow design, annotations, experimental data, results, clinical images, methodology, analysis, interpretation, and conclusions were independently developed, validated, and verified by the authors.

REFERENCES

[1] A. Anand, R. Panda, S. Kodamanchili, T. Gowthaman, R. R. Nair, and K. K. Bhardwaj, "Communication with Patients on Mechanical Ventilation: A Review of Existing Technologies," *Indian Journal of Critical Care Medicine*, vol. 26, no. 6, pp. 756–757, June 2022, https://doi.org/10.5005/jp-journals-10071-24225.

[2] S. Brambilla, D. Ausili, G. Locatelli, S. Di Mauro, G. Bellani, and M. Luciani, "Communication difficulties in mechanically ventilated voiceless patients in intensive care units: A qualitative study," *Nursing in Critical Care*, vol. 30, no. 3, May 2025, Art. no. e70037, https://doi.org/10.1111/nicc.70037.

[3] M. Kyranou, C. Cheta, and E. Pampoulou, "Communicating with mechanically ventilated patients who are awake. A qualitative study on the experience of critical care nurses in Cyprus during the COVID-19 pandemic," *PLOS ONE*, vol. 17, no. 12, Dec. 2022, Art. no. e0278195, https://doi.org/10.1371/journal.pone.0278195.

[4] C. Çelebi and K. Ö. Yeşilyurt, "Ensuring Effective Communication with Patients Receiving Mechanical Ventilation Support in Intensive Care Units: Current Communication Materials," *Cyprus Journal of Medical Sciences*, Aug. 2025, https://doi.org/10.4274/cjms.2025.2025-28.

[5] N. R. Kuruppu, W. Chaboyer, A. Abayadeera, and K. Ranse, "Augmentative and alternative communication tools for mechanically ventilated patients in intensive care units: A scoping review," *Australian Critical Care*, vol. 36, no. 6, pp. 1095–1109, Nov. 2023, https://doi.org/10.1016/j.aucc.2022.12.009.

[6] M. LaValley, T. Chavers-Edgar, M. Wu, R. Schlosser, and R. Koul, "Augmentative and Alternative Communication Interventions in Critical and Acute Care With Mechanically Ventilated and Tracheostomy Patients: A Scoping Review," *American Journal of Speech-Language Pathology*, vol. 33, no. 5, pp. 2667–2686, Sept. 2024, https://doi.org/10.1044/2024_AJSLP-23-00310.

[7] K. Aurangzeb, K. Javeed, M. Alhussein, I. Rida, S. I. Haider, and A. Parashar, "Deep Learning Approach for Hand Gesture Recognition: Applications in Deaf Communication and Healthcare," *Computers, Materials & Continua*, vol. 78, no. 1, pp. 127–144, 2024, https://doi.org/10.32604/cmc.2023.042886.

[8] H. P. J. Dutta, M. K. Bhuyan, D. R. Neog, K. F. MacDorman, and R. H. Laskar, "Patient Assistance System Based on Hand Gesture Recognition," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–13, 2023, https://doi.org/10.1109/TIM.2023.3282655.

[9] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, Mar. 2021, Art. no. 53, https://doi.org/10.1186/s40537-021-00444-8.

[10] M. Linardakis, I. Varlamis, and G. T. Papadopoulos, "Survey on Hand Gesture Recognition From Visual Input," *IEEE Access*, vol. 13, pp. 135373–135406, 2025, https://doi.org/10.1109/ACCESS.2025.3593428.

[11] M. Oudah, A. Al-Naji, and J. Chahl, "Hand Gesture Recognition Based on Computer Vision: A Review of Techniques," *Journal of Imaging*, vol. 6, no. 8, July 2020, Art. no. 73, https://doi.org/10.3390/jimaging6080073.

[12] R. Tripathi and B. Verma, "Survey on vision-based dynamic hand gesture recognition," *The Visual Computer*, vol. 40, no. 9, pp. 6171–6199, Sept. 2024, https://doi.org/10.1007/s00371-023-03160-x.

[13] L. I. Barona López, F. M. Ferri, J. Zea, Á. L. Valdivieso Caraguay, and M. E. Benalcázar, "CNN-LSTM and post-processing for EMG-based hand gesture recognition," *Intelligent Systems with Applications*, vol. 22, June 2024, Art. no. 200352, https://doi.org/10.1016/j.iswa.2024.200352.

[14] R. E. Nogales and M. E. Benalcázar, "Hand Gesture Recognition Using Automatic Feature Extraction and Deep Learning Algorithms with Memory," *Big Data and Cognitive Computing*, vol. 7, no. 2, May 2023, Art. no. 102, https://doi.org/10.3390/bdcc7020102.

[15] M. Ur Rehman *et al.*, "Dynamic Hand Gesture Recognition Using 3D-CNN and LSTM Networks," *Computers, Materials & Continua*, vol. 70, no. 3, pp. 4675–4690, 2022, https://doi.org/10.32604/cmc.2022.019586.

[16] H. G. Doan and N. T. Nguyen, "Fusion Machine Learning Strategies for Multi-modal Sensor-based Hand Gesture Recognition," *Engineering, Technology & Applied Science Research*, vol. 12, no. 3, pp. 8628–8633, June 2022, https://doi.org/10.48084/etasr.4913.

[17] A. O. Hashi, S. Z. M. Hashim, and A. B. Asamah, "Dynamic Adaptation in Deep Learning for Enhanced Hand Gesture Recognition," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15836–15841, Aug. 2024, https://doi.org/10.48084/etasr.7670.

[18] P. Bansal, T. Mishra, H. Solanki, and S. Patil, "Hand Gesture Recognition Using Machine Learning Technique," in *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, Feb. 2024, pp. 322–326, https://doi.org/10.1109/IC2PCT60090.2024.10486639.

[19] T. L. Dang, S. D. Tran, T. H. Nguyen, S. Kim, and N. Monet, "An improved hand gesture recognition system using keypoints and hand bounding boxes," *Array*, vol. 16, Dec. 2022, Art. no. 100251, https://doi.org/10.1016/j.array.2022.100251.

[20] H. H. Li and C. C. Hsieh, "Dynamic Hand Gesture Recognition Using MediaPipe and Transformer," in *IEEE ICEIB 2025*, Sept. 2025, Art. no. 22, https://doi.org/10.3390/engproc2025108022.

[21] C. N. Rang, P. Jerónimo, C. Mora, and S. Jardim, "Hand Gesture Recognition using Machine Learning," *Procedia Computer Science*, vol. 256, pp. 198–205, 2025, https://doi.org/10.1016/j.procs.2025.02.112.

[22] A. Sharma, A. Mittal, S. Singh, and V. Awatramani, "Hand Gesture Recognition using Image Processing and Feature Extraction Techniques," *Procedia Computer Science*, vol. 173, pp. 181–190, 2020, https://doi.org/10.1016/j.procs.2020.06.022.

[23] N. Al Mudawi *et al.*, "Innovative healthcare solutions: robust hand gesture recognition of daily life routines using 1D CNN," *Frontiers in Bioengineering and Biotechnology*, vol. 12, July 2024, Art. no. 1401803, https://doi.org/10.3389/fbioe.2024.1401803.

[24] Z. Gao, A. Sharma, M. Zheng, B. Planche, T. Chen, and Z. Wu, "Automated Patient Positioning with Learned 3D Hand Gestures," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Feb. 2025, pp. 3772–3781, https://doi.org/10.1109/WACV61041.2025.00371.

[25] A. Adin, E. T. Krainski, A. Lenzi, Z. Liu, J. Martínez-Minaya, and H. Rue, "Automatic cross-validation in structured models: Is it time to leave out leave-one-out?," *Spatial Statistics*, vol. 62, Aug. 2024, Art. no. 100843, https://doi.org/10.1016/j.spasta.2024.100843.

[26] D. Gholamiangonabadi, N. Kiselov, and K. Grolinger, "Deep Neural Networks for Human Activity Recognition With Wearable Sensors: Leave-One-Subject-Out Cross-Validation for Model Selection," *IEEE Access*, vol. 8, pp. 133982–133994, 2020, https://doi.org/10.1109/ACCESS.2020.3010715.

[27] S. Fathonah, E. Setyaningsih, E. Susanti, and Taukhit, "Cross-Patient Evaluation of CNN-Based Facial Expression Recognition for Intubated ICU Patients Using Leave-One-Patient-Out Validation," *Engineering, Technology & Applied Science Research*, vol. 16, no. 2, pp. 33187–33195, Apr. 2026, https://doi.org/10.48084/etasr.16931.

[28] E. Setyaningsih, S. Fathonah, E. Susanti, and Tauhkit "Patient-independent Evaluation of CNN-LSTM for Facial Expression Recognition in Intubated ICU Patients," *International Journal of Intelligent Engineering and Systems*, vol. 19, no. 5, pp. 1109–1123, 2026.