Cross-Patient Evaluation of CNN-Based Facial Expression Recognition for Intubated ICU Patients Using Leave-One-Patient-Out Validation

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

Emy Setyaningsih
Faculty of Science and Information Technology, Akprind University, Indonesia
emysetyaningsih@akprind.ac.id (corresponding author)

Erma Susanti
Faculty of Science and Information Technology, Akprind University, Indonesia
erma@akprind.ac.id

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

Received: 13 December 2025 | Revised: 9 January 2026 | Accepted: 17 January 2026

 DOI: https://doi.org/10.48084/etasr.16931

ABSTRACT

Intubated Intensive Care Unit (ICU) patients experience severe communication limitations due to medical device occlusions and altered physiological conditions, which make facial expression interpretation challenging for healthcare professionals. Facial Expression Recognition (FER) based on Convolutional Neural Network (CNN) offers a promising solution; however, its reliability in real ICU environments critically depends on the model's ability to generalize across patients with heterogeneous clinical characteristics. This study evaluates the cross-patient performance of a CNN-based FER system using a two-stage transfer learning approach and a rigorous patient-independent validation protocol, Leave-One-Patient-Out Cross-Validation (LOPOCV). Four ImageNet backbones, ResNet50, DenseNet121, MobileNetV2, and EfficientNetB0, were trained on 33 videos from 10 intubated ICU patients mapped into three FER classes. Conventional frame-level training yielded high accuracies, with DenseNet121 achieving 100% under non-patient-independent evaluation. An ablation study showed that partially unfreezing the final backbone layers produced the most stable configuration for small-scale clinical datasets. However, LOPOCV revealed a marked performance decrease (mean accuracy ≈ 45%), highlighting identity leakage inherent in frame-level evaluation and limited cross-patient generalization under high occlusion and inter-patient variability. These findings establish a more realistic, patient-independent performance baseline for FER in intubated ICU patient settings and underscore the need for temporal architectures and multimodal strategies to improve robustness and clinical reliability.

Keywords-convolutional neural network; facial expression recognition; cross-patient evaluation; intubated ICU patients; leave-one-patient-out cross-validation; transfer learning

I. INTRODUCTION

Some intubated Intensive Care Unit (ICU) patients are conscious but unable to communicate verbally due to mechanical ventilation, such as an endotracheal tube or tracheostomy cannula. Only a limited proportion of nonverbal communication can be accurately interpreted by healthcare professionals [1]. Inability to speak leads to frustration,

anxiety, and depression because patients cannot clearly communicate basic needs, pain, or discomfort [1-3]. This condition negatively affects patients' psychological and physiological well-being, reduces the effectiveness of care, and makes it difficult for medical personnel to assess patients' subjective condition [3]. Therefore, a more effective nonverbal communication approach is needed to improve understanding between patients and healthcare professionals. An alternative for assessing patient condition is by observing facial expressions [2-5]. However, manual observation of facial expressions by healthcare professionals is prone to subjectivity, limited observation frequency, and variation among healthcare professionals. One potential approach to address these issues is the use of artificial intelligence-based Facial Expression Recognition (FER) technology.

The development of artificial intelligence has enabled various technological applications, including support for human-machine interaction and medical systems [4-9]. In this context, Convolutional Neural Networks (CNNs) have become the dominant approach for FER due to their ability to extract discriminative spatial features from facial images automatically [8-12]. Numerous studies using public FER datasets, such as FER2013, CK+, and KDEF, have shown that CNN-based models can recognize seven basic facial expressions, namely happy, sad, fear, anger, surprise, disgust, and neutral, with high accuracy [10, 13-16]. When initialized with ImageNet pretrained weights, CNN architectures including VGG, Inception, ResNet, EfficientNet, MobileNet, and DenseNet typically achieve accuracies ranging from 88% to 97% under controlled non-occluded experimental conditions [10, 17, 18].

Several studies have further explored FER using occluded facial images or newly collected datasets. In [19], a two-stage CNN approach for occluded facial image verification achieved accuracies of up to 97.3% on public datasets, but this work focused on face verification rather than FER. In [11], a laboratory-collected dataset of 700 facial images from 20 individuals was employed, reporting an average accuracy of approximately 95% using a ResNet-50 architecture with batch normalization and ReLU under 10-fold cross-validation. FER research has also been extended to video-based real-time emotion prediction using pretrained CNN models such as EfficientNet, ResNet, and VGGNet on the FER2013 dataset, achieving accuracies of up to 82% [8].

Although several studies have reported high accuracy, they predominantly rely on public datasets of non-intubated facial images collected under controlled laboratory conditions, which differ substantially from real clinical environments, particularly ICUs. In practice, CNN-based FER for intubated ICU patients faces significant challenges, as large facial regions are frequently partially occluded by medical devices, such as endotracheal tubes, ventilator masks, and monitoring sensors, thereby reducing the availability of discriminative spatial features required for accurate expression recognition [5, 19, 20]. Additional difficulties arise from illumination variability caused by fluctuating ICU lighting, shadows from medical equipment, and changes in patient positioning, which introduce luminance distortions and degrade the reliability of visual feature extraction [6]. These effects are further compounded by uncontrolled head pose variations associated with clinical procedures and physiological conditions, often resulting in extreme viewing angles that are difficult to handle using conventional frame-level FER models [7]. Consequently, FER research under real-world ICU conditions remains limited. At the same time, the demand for AI-based support for nonverbal communication in intubated ICU patients continues to increase and has high clinical relevance [2].

Early clinical FER studies focused on pain assessment in conscious ICU patients using CNN-BiLSTM [5] or hybrid architectures such as DenseNet201-RBF-ELM evaluated on fully exposed facial images [7]. More recent approaches employed Vision Transformer (ViT) and Swin Transformer architectures to detect facial action units in the presence of medical artifacts [4]. Although these studies reported promising results, they were primarily limited to patients with minimal facial occlusion, without explicitly assessing cross-patient generalization, potentially leading to identity bias due to frame-level evaluation strategies.

Another critical challenge in clinical FER is the limited availability of annotated clinical datasets. Facial data collection in ICU settings requires ethical approval and strict medical oversight to address privacy and anonymity concerns, substantially limiting dataset size and increasing the risk of overfitting [4-6]. Furthermore, achieving reliable cross-patient generalization remains a fundamental challenge for CNN-based FER in clinically heterogeneous ICU environments. Most existing studies continue to rely on conventional evaluation schemes, such as random splits or frame-based k-fold cross-validation, which allow samples from the same patients to appear in both training and test sets. This practice can lead to identity leakage and overly optimistic performance estimates that do not reflect true generalization to unseen patients.

In contrast, patient-based evaluation using Leave-One-Patient-Out Cross-Validation (LOPOCV) provides a more representative assessment of cross-patient generalization by enforcing strict subject independence between training and testing data. As demonstrated in [21], LOPOCV is particularly suited for small-scale clinical datasets, as it maximizes the utilization of training data while preserving the independence of the test set. Thus, LOPOCV constitutes a methodological prerequisite for evaluating CNN-based FER systems intended for deployment in real-world ICU environments characterized by high variability, medical device occlusion, subtle expression dynamics, and limited clinical data.

Based on the reviewed literature, several research gaps can be identified. First, most FER studies have not specifically targeted intubated ICU patients with high levels of facial occlusion and complex physiological variations. Second, many studies have not employed patient-based evaluation protocols such as LOPOCV, leading to a high risk of identity bias in reported performance. Third, although FER models achieve high accuracy on public datasets, their performance has not been evaluated under extreme clinical conditions characterized by severe facial occlusion, small-scale and heterogeneous datasets, and medical artifacts.

To address these gaps, this study makes the following contributions: (i) Emphasizes the importance of a cross-patient generalization evaluation framework for clinical FER in intubated ICU patients by applying LOPOCV to address identity bias that often arises in frame-level evaluations; (ii) Integrates a more adaptive two-stage transfer learning strategy for small-scale clinical datasets with high levels of occlusion to reduce overfitting in the clinical domain; and (iii) Provides a methodological and conceptual baseline for the development of an artificial intelligence-based nonverbal communication system for intubated ICU patients. This contribution is expected to be the basis for further research that integrates temporal modeling and multimodal fusion to improve the accuracy and stability of AI-based nonverbal communication systems for intubated ICU patients in hospitals.

II. METHODOLOGY

Figure 1 illustrates the overall workflow of the research, including stages from data preprocessing to LOPOCV evaluation.

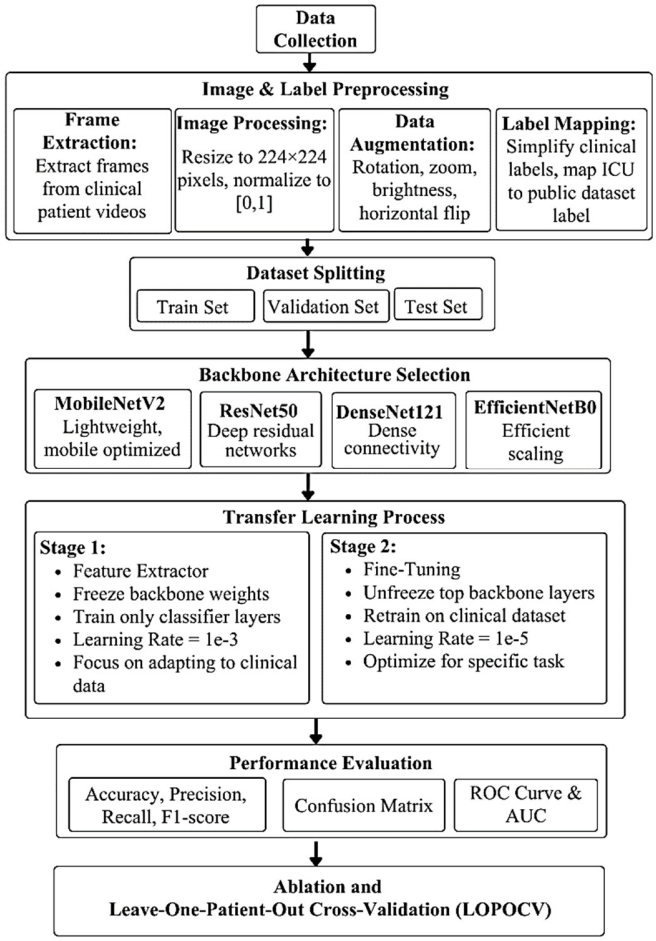

Fig. 1. Research workflow.

A. Data Collection

Clinical data were collected from intubated patients treated in the ICU of Gadjah Mada University Academic Hospital (RSA UGM). The data collection protocol was approved by the Research Ethics Committee of RSA UGM (Ethics Approval No. 072/RSA/KEP/EC/2025), and informed consent was obtained from all patients or their legal representatives. Samples were selected using purposive sampling to ensure the inclusion of patients capable of producing facial expressions that could be reliably interpreted. Inclusion criteria consisted of: (i) compos mentis patients with a Glasgow Coma Scale (GCS) score $\geq$ 13, indicating that the patient is ensuring sufficient consciousness for voluntary facial muscle control; (ii) intubated and hemodynamically stable, to minimize involuntary facial movements resulting from acute physiological stress; and (iii) an age range of 18 to 72 years to reduce bias associated with extreme age-related facial variability. Meanwhile, patients with neurological disorders, impaired facial mobility, or tracheostomy were excluded. These criteria were established to obtain a homogeneous and reliable dataset while balancing ethical considerations, clinical feasibility, and data validity in the ICU setting. The final cohort consisted of 10 intubated ICU patients (five males and five females) aged 42–72 years, with ICU stays ranging from 1 to 14 days. A total of 33 anonymized videos (30–60 s each) were recorded at 30 fps. The recorded clinical facial expressions were initially categorized into ten clinical labels, as summarized in Table I, and subsequently validated by on-duty ICU medical personnel in accordance with ethical regulations. These labels were then mapped into three target FER classes for model training and evaluation.

This research also used the Roboflow Medical Emergency Detection dataset [22], which contains the DiscomfortFace, NormalFace, and NaturalBody classes. The dataset was not used for training, validation, or testing the clinical FER models. Instead, it was employed solely as a reference dataset to validate the consistency of label mapping between clinically annotated expressions and commonly used public FER categories. All LOPOCV evaluations and performance analyses were conducted exclusively on the clinical ICU dataset to avoid data leakage and preserve patient independence.

TABLE I. DISTRIBUTION OF EXPERT-ANNOTATED CLINICAL FACIAL EXPRESSION VIDEOS FROM INTUBATED ICU PATIENTS.

Expert-annotated clinical FER labels	Number of videos	Number of patients	Patient
Uncomfortable	2	2	P003, P005
Asking for Something	1	1	P002
Reject	2	2	P005, P006
Nauseous	1	1	P008
Neutral	9	7	P001, P002, P003, P004, P008, P009, P010
Painful	6	6	P001, P006, P007, P008, P009, P010
Agree	2	2	P002, P006
Suction	4	3	P003, P005, P009
Disagree	1	1	P002
Sleep	5	5	P001, P002, P006, P007, P008
Total	**33**	**10**	

B. Image and Label Preprocessing

The preprocessing stage was carried out to ensure consistent clinical image quality, which is often affected by medical device occlusion, unstable lighting, and head position. The steps in image and label preprocessing were as follows:

1) Frame Extraction

This stage aimed to perform initial preprocessing by extracting frames from clinical patient video data collected using 3 fps sampling. The next step was face detection using MTCNN, followed by cropping, focusing on the facial area of intubated ICU patients.

2) Image Processing

This stage aimed to standardize the image size of the frame extraction results by resizing each frame to 224×224 pixels. Next, pixel normalization (mean–std ImageNet) was performed from the [0, 255] to [0, 1] range to ensure model stability and convergence during training.

3) Data Augmentation

Following the data augmentation stage, facial data variability was increased using visual augmentation and contextual augmentation (clinical simulation). Visual augmentation consisted of random rotation (±15°), translation and zoom (≤0.1), random in/out zooming (±10%), and horizontal flipping. Additional operations included brightness and contrast adjustment, mild Gaussian blurring, and occlusion simulation through partial facial masking. Contextual augmentation simulated ICU conditions by overlaying medical devices (e.g., tubes and oxygen masks), adding ICU-related background noise, and simulating low-light conditions.

4) Label Mapping

This stage aimed to simplify clinical labels by mapping the clinical dataset (Table I) into three FER classes (DiscomfortFace, NaturalBody, NormalFace) derived from public datasets, as shown in Table II.

TABLE II. MAPPING CLINICAL EXPRESSION CATEGORIES INTO FINAL FER CLASSES

Public classes	Clinical datasets	Number of videos	Number of patients	Unique patients
Clutching Chest	-	-	-	-
Discomfort Face	Uncomfortable, Nauseous, Painful, Suction	13	8	P001, P003, P005, P006, P007, P008, P009, P010
Natural Body	Asking for Something, Reject, Agree, Disagree	6	3	P002, P005, P006
Normal Face	Neutral, Sleep	14	9	P001, P002, P003, P004, P006, P007, P008, P009, P010

C. Dataset Splitting

After data augmentation, the dataset was split into 70% training, 15% validation, and 15% testing sets. Public dataset samples were used exclusively for label mapping validation and were excluded from the LOPOCV-based clinical evaluation. Table III summarizes the class-wise distribution.

TABLE III. DISTRIBUTION OF CLINICAL AND PUBLIC DATASETS ACROSS TRAINING, VALIDATION, AND TESTING SETS

Classes	Clinical datasets			Public datasets		
	Train	Validation	Test	Train	Validation	Test
ClutchingChest	0	0	0	1831	523	262
DiscomfortFace	557	119	120	57	16	9
NaturalBody	218	46	48	305	87	44
NormalFace	473	101	103	1430	408	206

D. Backbone Architecture Selection

The ImageNet-pretrained architectures evaluated as backbone networks in this study included ResNet50, DenseNet121, MobileNetV2, and EfficientNetB0. These backbones were selected based on prior studies that reported a favorable trade-off between classification performance, representational capacity, and computational efficiency, making them suitable for small-scale learning scenarios [10, 11, 15, 16]. At this stage, the original fully connected layers were replaced with a Global Average Pooling (GAP) layer, followed by dropout and a softmax-based classification layer corresponding to the number of target classes.

E. Transfer Learning Process

This two-stage transfer learning strategy was applied to ensure stable and practical model training with limited clinical data. As illustrated in Figure 1, the transfer learning process consisted of two sequential stages.

- Stage 1 (Feature Extraction): In this stage, the ImageNet-pretrained backbone was adapted to the clinical domain by freezing all backbone weights to prevent retraining. Only the classification head was trained using a learning rate of 1×10^{-3} for five epochs.
- Stage 2 (Fine-tuning): This stage aimed to further optimize feature representations for recognizing facial expression classes of intubated ICU patients while mitigating overfitting. The final 50% of the backbone layers were unfrozen, and the model was retrained on the clinical dataset using a lower learning rate of 1×10^{-5} for 10 epochs.

F. Performance Evaluation

This stage aimed to determine the best model for detecting facial expression patterns in intubated ICU patients. The four backbones were tested using a clinical dataset to identify the best model applicable to real-world environments, namely intubated ICU patients.

G. Ablation Study and LOPOCV Evaluation

This stage tested the model's sensitivity to the training configuration (an ablation study) and assessed its generalization across patients using LOPOCV. Testing was conducted on the best pre-trained ImageNet architecture, with the following evaluation steps.

1) Dataset Preparation for Ablation and LOPOCV

The same dataset from the first stage was reused, but the augmentation was re-controlled, and the backbone was selected based on the first-stage testing results that showed the most stable performance.

2) Ablation Study

Ablation was performed to evaluate the model's sensitivity to training configurations. The goal was to identify the most stable backbone configuration, test whether regularization can improve robustness in clinical domains, and assess the architecture's limitations on small datasets. This stage involved testing six configurations:

- A1: Augmentation ON, dropout 0.3, unfreeze half
- A2: Augmentation OFF, no dropout
- A3: Dropout 0.5
- A4: Unfreeze none (feature extractor only)
- A5: Unfreeze all (full fine-tuning)
- A6: L2 regularization 1e–4

Each configuration was trained using the same transfer learning procedure to assess how changes in hyperparameters affected training stability and feature adaptation.

3) LOPOCV Evaluation

LOPOCV was used as the primary evaluation of clinical generalization with high inter-patient variation. The goal of evaluation using LOPO is to prevent identity leakage by ensuring that no patient images in the training set appear in the test set, and to produce evaluation conditions as close as possible to a real ICU environment, where the model must recognize patient expressions never seen during training. The LOPO steps are [23]:

- Data is split based on patient identity, with one patient designated as the left-out one at each evaluation iteration.
- All data from patients other than the dropped patient is used as training data.
- The best CNN backbone model is trained using a two-stage transfer learning strategy, which includes a feature extraction stage and partial fine-tuning in the final backbone layer.
- The trained model is tested exclusively on dropped patients to assess cross-patient generalization.
- The process is repeated until each patient has been assigned to the test data, and all performance metrics (accuracy, precision, recall, F1-score, and AUC) are aggregated to obtain a realistic, identity-leakage-free performance estimate.

III. RESULTS AND DISCUSSION

This section analyzes the proposed approach using a clinical dataset with three class labels: DiscomfortFace, NaturalBody, and NormalFace. Performance is evaluated across four ImageNet-pretrained CNN backbones, namely ResNet50, DenseNet121, MobileNetV2, and EfficientNetB0, which are widely used in facial expression recognition due to their strong capability in learning discriminative spatial representations [10, 18].

Model training followed a two-stage transfer learning strategy, comprising feature extraction and partial fine-tuning, to improve training stability and reduce overfitting on small-scale clinical data, which is particularly suitable for medical imaging scenarios characterized by limited data availability and high inter-subject variability [5, 16]. Performance evaluation was conducted using two schemes: a conventional train-validation-test split and a patient-independent LOPOCV protocol, allowing for a comparative analysis between intra-patient performance and cross-patient generalization under realistic ICU conditions.

A. Comparison of Public-Only and Clinical Fine-Tuning Models

Before analyzing the performance of the CNN backbone after clinical fine-tuning, an initial comparison was conducted between a model trained solely on public datasets and one fine-tuned on clinical data. Table IV presents a comparison of test accuracy across public and clinical datasets to assess the impact of clinical-domain adaptation on FER performance.

TABLE IV. COMPARATIVE RESULTS BETWEEN THE PUBLIC AND THE CLINICAL DATASET

Backbone	Public dataset test accuracy	Clinical dataset test accuracy
BaseCNN	0.5555	-
ResNet50	0.5106	0.9963
DenseNet121	0.7025	1.0000
EfficientNetB0	0.5029	0.3801
MobileNetV2	0.6583	0.8930

The results in Table IV show a significant difference in performance when testing the models on the clinical dataset. The ResNet50 and DenseNet121 models show a drastic increase in accuracy after fine-tuning on clinical data, achieving 99.63% and 100%, respectively. This indicates that both architectures are capable of adapting to complex visual characteristics in clinical settings. MobileNetV2 also showed a significant performance increase, although not as high as ResNet50 and DenseNet121. In contrast, EfficientNetB0 showed a decrease in performance on the clinical dataset, indicating its limitations in handling occlusions from medical devices and very subtle facial expressions. These findings confirm that models trained solely on public datasets cannot be directly generalized to the ICU clinical domain without sufficient adaptation.

B. Backbone Architecture Performance in Conventional Evaluation

Further evaluation was conducted using conventional data partitioning, with 70% for training, 15% for validation, and 15% for testing. Table V presents the performance results after fine-tuning using a clinical dataset.

TABLE V. RESULTS AFTER FINE-TUNING ON CLINICAL DATASETS

Backbone	Precision	Recall	F1-score	AUC	Accuracy
ResNet50	0.9963	0.9963	0.9963	1.0000	0.9963
DenseNet121	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
MobileNetV2	0.9165	0.8930	0.8945	0.9932	0.8930
EfficientNetB0	0.1445	0.3801	0.2093	0.5034	0.3801

As shown in Table V, the DenseNet121 backbone demonstrated the highest performance, with precision, recall, F1-score, accuracy, and AUC reaching 100% under conventional frame-level evaluation; however, this result does not reflect cross-patient generalization capability. The ResNet50 backbone also achieved exceptionally high performance, albeit slightly lower than DenseNet121, with all metrics around 99.63%. AUC values exceeding 99% for both models indicate strong intra-patient classification performance and the ability to recognize subtle facial expressions in intubated ICU patients, rather than true cross-patient generalization

The trained models were evaluated on 271 clinical samples across three classes: DiscomfortFace (120), NaturalBody (48), and NormalFace (101). As illustrated in Figure 2, the DenseNet121 backbone achieved the best performance, with a very low misclassification rate. The ResNet50 model also demonstrated strong performance, with only a single misclassification where a NormalFace sample was incorrectly classified as DiscomfortFace. In contrast, MobileNetV2 and EfficientNetB0 produced more misclassifications, particularly for classes associated with subtle facial expressions.

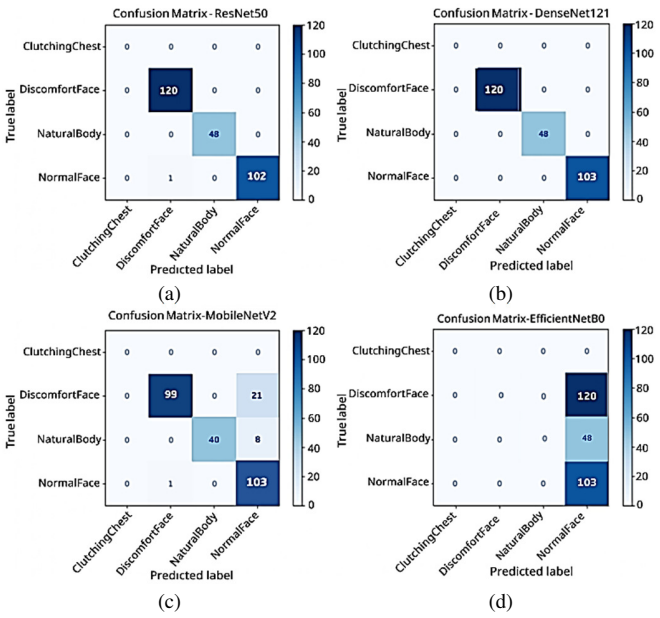

Fig. 2. Confusion matrices on testing: (a) ResNet50; (b) DenseNet121; (c) MobileNetV2; (d) EfficientNetB0

Further tests were conducted to evaluate training stability and potential overfitting by analyzing the validation accuracy and validation loss curves, as shown in Figure 3. Based on these test results, the DenseNet121 model demonstrated the best and most stable performance, as evidenced by a validation accuracy of 100% and a validation loss that remained very low and steadily decreased. Although the evaluation results show very high performance, particularly with DenseNet121 architectures, they do not fully reflect the model's ability to handle real-world clinical conditions. The train-validation-test split still involves frames from the same patient in both the training and test sets, potentially introducing identity leakage and leading to overly optimistic performance estimates. In the context of ICU applications, the model must be able to recognize patients' facial expressions that have never been seen before, with varying physiological characteristics, levels of medical device occlusion, and visual conditions.

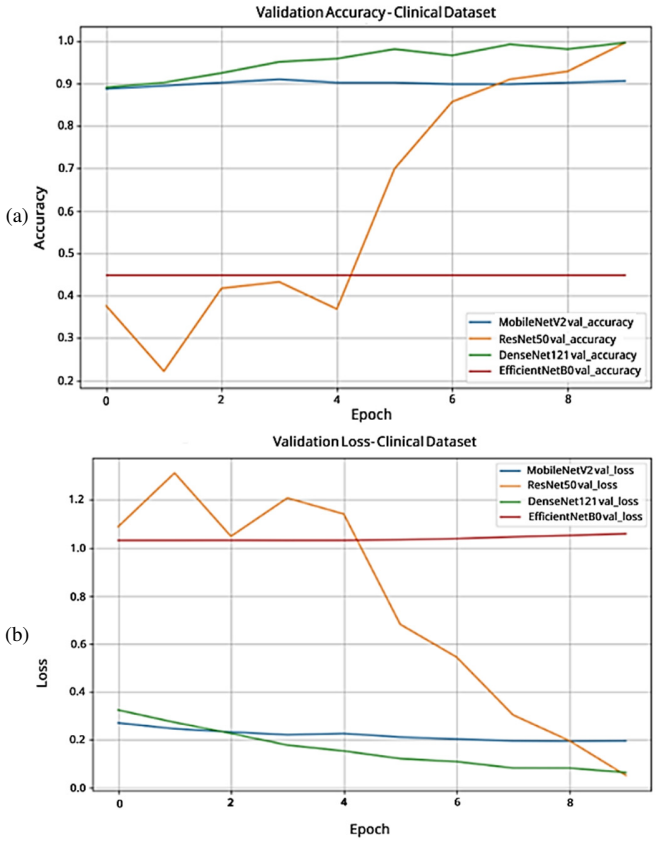

Fig. 3. Validation curves of the four backbone architectures: (a) validation accuracy and (b) validation loss.

C. Ablation Study of Training Configurations

Before assessing cross-patient generalization ability, an ablation study was conducted to identify the most stable and representative training configuration on a small clinical dataset. This ablation study analyzed the model's sensitivity to freezing-unfreezing strategies, regularization, and data augmentation, to ensure that high performance in conventional evaluation was not solely due to overly specific training settings. Based on the previous testing results, the DenseNet121 backbone with the most stable configuration was selected as the baseline model for further evaluation. Table VI shows test results for six configurations of the DenseNet121 backbone, including variations in dropout, augmentation, number of unfrozen layers, and L2 regularization. These results show that all configurations using the half-unfreeze strategy achieved an accuracy of 1.000, regardless of the dropout variation, augmentation, or L2 regularization. This finding confirms that partially unfreezing the final layers of the backbone is the most stable setting for small clinical datasets. Conversely, the

complete fine-tuning configuration achieved slightly lower accuracy (0.996), suggesting instability when all backbone parameters are adjusted with limited data. Meanwhile, the unfreeze-none configuration performed the lowest (0.867), demonstrating that standard ImageNet features are insufficient to handle the visual complexity of intubated ICU patients.

TABLE VI. ABLATION STUDY RESULTS FOR DIFFERENT TRAINING CONFIGURATIONS OF DENSENET121

Experiment	Test accuracy
aug_on_do0.3_unfreeze_half_l2_0	1.000
aug_off_do0.0_unfreeze_half_l2_0	1.000
aug_on_do0.5_unfreeze_half_l2_0	1.000
aug_on_do0.3_unfreeze_all_l2_0	0.996
aug_on_do0.3_unfreeze_none_l2_0	0.867
aug_on_do0.3_unfreeze_half_l2_1e-4	1.000

Hyperparameter variations did not yield significant performance differences under frame-level evaluation conditions, indicating that the datasets in the conventional evaluation were relatively homogeneous. Overall, the ablation study confirmed that the half-unfreeze configuration is optimal.

D. Cross-Patient Generalization Using LOPOCV

This study further applied LOPOCV το obtain a more realistic evaluation of the generalization ability across patients. This approach ensures that the test data comes entirely from patients who were not included in the training process, thus the evaluation results more closely reflect the model's performance in real-world clinical scenarios.

LOPOCV evaluation reveals substantial variability in model performance across patients, indicating limited cross-patient generalization of the CNN-based FER model in the intubated ICU setting. Table VII reports a statistical summary of the aggregated LOPOCV performance metrics.

TABLE VII. STATISTICAL SUMMARY OF LOPOCV PERFORMANCE ACROSS TEN ICU PATIENTS.

Metric	Min	Max	Mean	Std Dev
Accuracy	0.1583	0.6426	0.4511	0.1505
F1-score	0.2187	0.6860	0.4544	0.1379
Precision	0.1933	0.6593	0.4572	0.1415
Recall	0.3167	0.9358	0.6097	0.2077
ROC-AUC	0.4383	0.4383	0.6182	0.1662

As shown in Table VII, accuracy ranged from 0.1583 to 0.6426, with a mean of 0.4511 and a relatively high standard deviation (0.1505), reflecting pronounced inter-patient variability. A similar trend was observed for the F1-score, which varied from 0.2187 to 0.6860, further indicating heterogeneous classification performance.

Figure 4 shows patient-wise accuracy and F1-score distributions under the LOPOCV protocol. Patient P008 achieved the highest accuracy (0.6426), followed by P003 (0.6137) and P009 (0.5932), whereas the lowest performance was observed for P004 (0.1583) and P006 (0.2649). These variations highlight the difficulty of recognizing facial expression patterns in intubated ICU patients, where severe medical device occlusion and visual heterogeneity are typical. Discrepancies between accuracy and F1-score are also evident; for example, patient P007 shows relatively low accuracy (0.3167) but a higher F1-score (0.5716), suggesting class imbalance and uneven prediction confidence across expression categories.

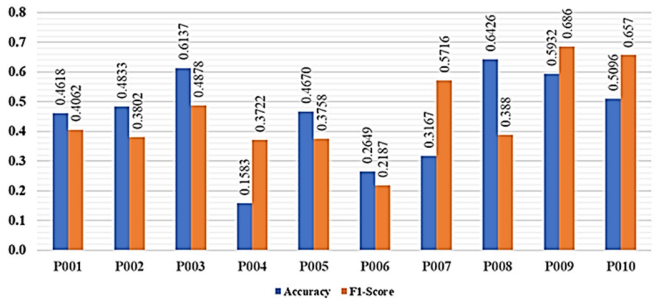

Fig. 4. LOPOCV performance of the DenseNet121 model across 10 patients, with accuracy and F1-score per fold.

Figure 5 presents the patient-wise ROC-AUC values obtained under the same LOPOCV protocol. Despite identical evaluation settings, ROC-AUC scores vary considerably across patients, indicating differences in discriminative capability associated with patient-specific conditions. Higher ROC-AUC values suggest more stable class separability, whereas lower values reflect increased difficulty due to occlusion, low expression intensity, and visual noise. Overall, these results confirm that individual patient characteristics strongly influence FER performance in ICU settings.

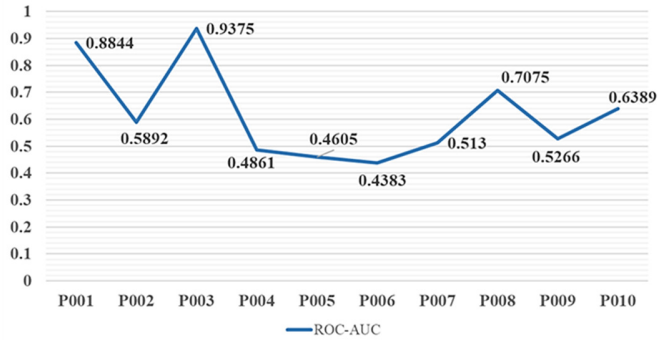

Fig. 5. ROC-AUC distribution of the DenseNet121 model across ten LOPOCV folds.

Compared to conventional train-validation-test evaluation, LOPOCV results show an apparent performance decline, underscoring the challenge of cross-patient generalization. This performance gap is indicative of identity leakage, as described in patient-independent evaluation guidelines [21], where overly optimistic results may arise when training and testing samples originate from the same patient. Similar degradation under patient-independent evaluation has been reported in prior ICU-based FER studies, motivating the use of LOPOCV in this work to obtain a more realistic assessment of model generalization under clinical conditions.

E. Comparison with Prior ICU-Based Studies

To contextualize the LOPOCV results, the findings of this study were compared with representative ICU-based studies on facial expression and pain recognition. Table VIII provides a concise overview, summarizing differences in task formulation, dataset characteristics, evaluation protocols, and reported performance.

TABLE VIII. COMPARISON WITH PRIOR ICU-BASED FACIAL EXPRESSION AND PAIN RECOGNITION STUDIES

Summary	[4]	[5]	This research
Task	Facial AU detection	Pain classification (image & video)	Facial expression recognition
Dataset & Subjects	49 ICU patients, 76,388 frames	63 ICU patients, 746 video clips	10 intubated ICU patients
Evaluation	Patient-wise train-test split	Patient-wise train-test split	LOPOCV (patient-independent)
Performance	F1 up to 0.88; Acc. up to 0.85	Acc. ≈0.56 (multi-class); 0.77–0.88 (binary)	Mean Acc. 0.45; Mean F1 0.45
Key Notes	Transformer-based; sensitive to AU prevalence	CNN & BiLSTM; higher performance in binary tasks	Strict generalization; high inter-patient variability

Previous studies reported moderate to high performance on data from critically ill patients. In [5], accuracy values of approximately 0.56 were achieved for multi-class facial expression recognition and 0.77–0.88 for binary pain classification. In [4], F1-scores of up to 0.88 were reported for facial action unit detection. These results demonstrate the feasibility of facial analysis in ICU environments, but were obtained using patient-wise train-test splits or reference-based evaluation schemes, which may reduce inter-subject variability and lead to optimistic generalization estimates.

In contrast, this study employed an LOPOCV protocol to assess patient-independent generalization. Under this protocol, the model achieved a lower mean accuracy (0.45) and substantial inter-patient variability, reflecting the inherent difficulty of generalizing FER models to unseen ICU patients. The observed performance gap with previous studies is therefore primarily attributed to differences in evaluation protocols rather than to model limitations.

Challenges such as medical device occlusion, reduced facial expressiveness due to sedation, and heterogeneous physiological conditions are inherent to real-world ICU data and have been consistently reported in previous studies. Consequently, benchmark facial expression datasets acquired under controlled conditions are not directly comparable to the present clinical evaluation and were not used as substitutes for clinical testing, but only to verify label consistency. Overall, when protocol differences are taken into account, the results of this work align with prior ICU-based research and provide a more conservative yet clinically relevant assessment of FER performance.

IV. CONCLUSION

Building on the cross-patient evaluation results, this study provides a comprehensive assessment of CNN-based FER for intubated ICU patients, a clinical domain characterized by high visual and physiological complexity. By employing a two-stage transfer learning strategy and comparing models trained on public datasets with those adapted to clinical data, the results demonstrate that clinical domain adaptation is a critical factor for improving FER performance in ICU environments. DenseNet121 and ResNet50 achieved substantial performance gains after clinical fine-tuning, whereas EfficientNetB0 showed limitations in handling occlusions from medical devices and subtle facial expressions. Although conventional frame-level evaluation yielded very high performance, with DenseNet121 achieving up to 100% accuracy, such results should not be interpreted as indicators of clinical readiness. Ablation analysis further revealed that training configuration significantly affects model stability, with the half-unfreeze strategy providing the most effective balance between feature adaptation and overfitting prevention on small-scale clinical datasets.

Nevertheless, conventional evaluation remains prone to overly optimistic estimates when frames from the same patient appear in both the training and the test sets. In contrast, patient-based evaluation using LOPOCV offers a more realistic assessment of cross-patient generalization. The LOPOCV results showed a marked performance decline, with an average accuracy of approximately 45% and substantial inter-patient variability, highlighting the limitations of frame-level CNN approaches in capturing subtle, gradual, and occluded clinical expressions.

Overall, this study underscores the importance of patient-based evaluation and establishes a realistic methodological baseline for clinical FER in ICU settings. Future work will focus on temporal modeling to address the inherent limitations of frame-level CNN architectures in capturing dynamic facial expressions, multimodal integration with physiological and clinical signals, expansion of clinical datasets, and the development of real-time FER systems to support reliable AI-based nonverbal communication in ICU environments.

ACKNOWLEDGMENT

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

REFERENCES

[1] C. Çelebi and K. Öztepe 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.

[2] M. Danielis, A. Povoli, E. Mattiussi, and A. Palese, "Understanding patients' experiences of being mechanically ventilated in the Intensive Care Unit: Findings from a meta-synthesis and meta-summary," *Journal of Clinical Nursing*, vol. 29, no. 13–14, pp. 2107–2124, July 2020, https://doi.org/10.1111/jocn.15259.

[3] S. Fathonah, E. Ernawati, and M. S. Hakim, "Communication Experiences of Conscious Indonesian Patients Who Underwent Mechanical Breathing in an Intensive Care Unit: A Phenomenological Approach," *Jordan Medical Journal*, vol. 58, no. 4, Nov. 2024, https://doi.org/10.35516/jmj.v58i4.1601.

[4] S. Nerella, K. Khezeli, A. Davidson, P. Tighe, A. Bihorac, and P. Rashidi, "End-to-End Machine Learning Framework for Facial AU

Detection in Intensive Care Units." arXiv, 2022, https://doi.org/10.48550/ARXIV.2211.06570.

[5] C. L. Wu *et al.*, "Deep Learning-Based Pain Classifier Based on the Facial Expression in Critically Ill Patients," *Frontiers in Medicine*, vol. 9, Mar. 2022, Art. no. 851690, https://doi.org/10.3389/fmed.2022.851690.

[6] S. Gkikas and M. Tsiknakis, "Automatic assessment of pain based on deep learning methods: A systematic review," *Computer Methods and Programs in Biomedicine*, vol. 231, Apr. 2023, Art. no. 107365, https://doi.org/10.1016/j.cmpb.2023.107365.

[7] S. Alphonse, S. Abinaya, and N. Kumar, "Pain assessment from facial expression images utilizing Statistical Frei-Chen Mask (SFCM)-based features and DenseNet," *Journal of Cloud Computing*, vol. 13, no. 1, Sept. 2024, Art. no. 142, https://doi.org/10.1186/s13677-024-00706-9.

[8] D. Bhagat, A. Vakil, R. K. Gupta, and A. Kumar, "Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN)," *Procedia Computer Science*, vol. 235, pp. 2079–2089, 2024, https://doi.org/10.1016/j.procs.2024.04.197.

[9] A. Liu and H. Yue, "Facial Expression Recognition Based on CNN-LSTM," in *Proceedings of the 2023 7th International Conference on Electronic Information Technology and Computer Engineering*, Oct. 2023, pp. 486–491, https://doi.org/10.1145/3650400.3650480.

[10] Q. Zhu, H. Zhuang, M. Zhao, S. Xu, and R. Meng, "A study on expression recognition based on improved mobilenetV2 network," *Scientific Reports*, vol. 14, no. 1, Apr. 2024, Art. no. 8121, https://doi.org/10.1038/s41598-024-58736-x.

[11] B. Li and D. Lima, "Facial expression recognition via ResNet-50," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 57–64, June 2021, https://doi.org/10.1016/j.ijcce.2021.02.002.

[12] R. A. Elsheikh, M. A. Mohamed, A. M. Abou-Taleb, and M. M. Ata, "Improved facial emotion recognition model based on a novel deep convolutional structure," *Scientific Reports*, vol. 14, no. 1, Nov. 2024, Art. no. 29050, https://doi.org/10.1038/s41598-024-79167-8.

[13] A. Talukder and S. Ghosh, "Facial Image expression recognition and prediction system," *Scientific Reports*, vol. 14, no. 1, Nov. 2024, Art. no. 27760, https://doi.org/10.1038/s41598-024-79146-z.

[14] C. J. Meryl, K. Dharshini, D. Sujitha Juliet, J. Akila Rosy, and S. S. Jacob, "Deep Learning based Facial Expression Recognition for Psychological Health Analysis," in *2020 International Conference on Communication and Signal Processing (ICCSP)*, July 2020, pp. 1155–1158, https://doi.org/10.1109/ICCSP48568.2020.9182094.

[15] R. Angeline and A. A. Nithya, "Deep Human Facial Emotion Recognition: A Transfer Learning Approach Using EfficientNetB0 Model," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 8, pp. 3664–3670.

[16] M. Bie, H. Xu, Q. Liu, Y. Gao, K. Song, and X. Che, "DA-FER: Domain Adaptive Facial Expression Recognition," *Applied Sciences*, vol. 13, no. 10, May 2023, Art. no. 6314, https://doi.org/10.3390/app13106314.

[17] R. Grover and S. Bansal, "Efficient Facial Expression Recognition Through Lightweight CNN Technique on Public Datasets," *SN Computer Science*, vol. 6, no. 1, Dec. 2024, Art. no. 15, https://doi.org/10.1007/s42979-024-03557-y.

[18] T. Alrimy, A. Alloqmani, A. Alotaibi, N. Aljohani, and S. Kammoun, "Facial Expression Recognition Based on Well-Known ConvNet Architectures," *Journal of King Abdulaziz University: Computing and Information Technology Sciences*, vol. 12, no. 1, July 2023, https://doi.org/10.4197/Comp.12-1.5.

[19] W. Zhe *et al.*, "A Research on Two-Stage Facial Occlusion Recognition Algorithm based on CNN," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 18205–18212, Dec. 2024, https://doi.org/10.48084/etasr.8736.

[20] S. Nerella, J. Cupka, M. Ruppert, P. Tighe, A. Bihorac, and P. Rashidi, "Pain Action Unit Detection in Critically Ill Patients," in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, July 2021, pp. 645–651, https://doi.org/10.1109/COMPSAC51774.2021.00094.

[21] T. J. Bradshaw, Z. Huemann, J. Hu, and A. Rahmim, "A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging," *Radiology: Artificial Intelligence*, vol. 5, no. 4, July 2023, Art. no. e220232, https://doi.org/10.1148/ryai.220232.

[22] "Medical Emergency Detection - v2." Roboflow, [Online]. Available: https://universe.roboflow.com/sangeevan-workspace/medical-emergency-detection/dataset/2.

[23] T. Winger, C. Ozdemir, S. L. Narasimhan, and J. Srivastava, "Time-Adaptive Machine Learning Models for Predicting the Severity of Heart Failure with Reduced Ejection Fraction," *Diagnostics*, vol. 15, no. 6, Mar. 2025, Art. no. 715, https://doi.org/10.3390/diagnostics15060715.