LUV-Net: Multi-Pattern Lung Ultrasound Video Classification through Pattern-Specific Attention with Efficient Temporal Feature Extraction

Jung Hoon Lee[1,2] crop2292@snu.ac.kr
Changi Kim [1,2] frz2yroom@snu.ac.kr
Jinwoo Lee[3] realrain7@snu.ac.kr
Si Mong Yoon[3] dostark1986@gmail.com
Kyung-Eui Lee[3] syku1@hanmail.net
Hyun-Jun Park[3] parkhjsnuh@gmail.com
Kwonhyung Hyung[3] herohkh@naver.com
Chang Min Park[1,2] morphius@snu.ac.kr

[1] *Department of Interdisciplinary Program in Bioengineering, Seoul National University.*

[2] *Department of Radiology, Seoul National University Hospital, Seoul National University College of Medicine.*

[3] *Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Seoul National University College of Medicine.*

Editors: Under Review for MIDL 2025

Abstract

Lung ultrasound (LUS) has emerged as a crucial bedside imaging tool for critical care, yet its interpretation remains challenging due to its artifact-based nature and high operator dependency. While deep learning approaches offer promising solutions for LUS pattern analysis, existing methods are limited by their focus on single-pattern recognition or disease-specific classification, and inadequate handling of temporal dynamics in video-based models. We propose LUV-Net (Lung Ultrasound Video Network), a novel deep learning model for multi-label classification of LUS patterns, combining pattern-specific attention mechanisms with temporal feature extraction. Our approach consists of two key modules: a spatial feature extraction module utilizing independent pattern-specific attention mechanisms, and a temporal feature extraction module designed to capture sequential relationships between adjacent frames. The model was evaluated using two distinct datasets: a development set of 341 LUS videos and a temporally separated validation set of 56 videos. Through 5-fold cross-validation, LUV-Net demonstrated superior performance in identifying all four LUS patterns (A-lines, B-lines, consolidation, and pleural effusion) compared to conventional video models, achieving higher AUC scores across patterns. The model's interpretability was validated through visualization of pattern-specific attention regions, providing insights into its decision-making process. The code is publicly available at $https://github.com/iamhxxn2/LungUS_Video$.

Keywords: Video Multi-label Classification, Lung Ultrasound, Pattern-Specific Attention, Efficient Temporal Feature

1. Introduction

Point-of-care ultrasound (POCUS) has progressively proven its significance as a useful bedside imaging modality, crucial for the assessment of critically ill patients and facilitating both diagnostic and therapeutic decision-making processes (Zieleskiewicz et al., 2021; Shrestha et al., 2018). Lung ultrasound (LUS) has been shown to have higher sensitivity for pneumothorax and pleural effusion than chest radiography (CXR) (Shrestha et al., 2018; Brogi et al., 2017), offering advantages of being non-invasive, cost-effective, and portable. Therefore, LUS has considerable potential as an important tool in low- and middle-income countries (LMICs) (Marini et al., 2021; Shrestha et al., 2018; Buonsenso and De Rose, 2022). However, LUS interpretation presents significant challenges due to its artifact-based nature rather than direct lung anatomy visualization, making it highly operator-dependent. Additionally, the lack of qualified ultrasound professionals and insufficient training programs are significant obstacles to the application of LUS in clinical practice (Marini et al., 2021; Nhat et al., 2023; Lim et al., 2017).

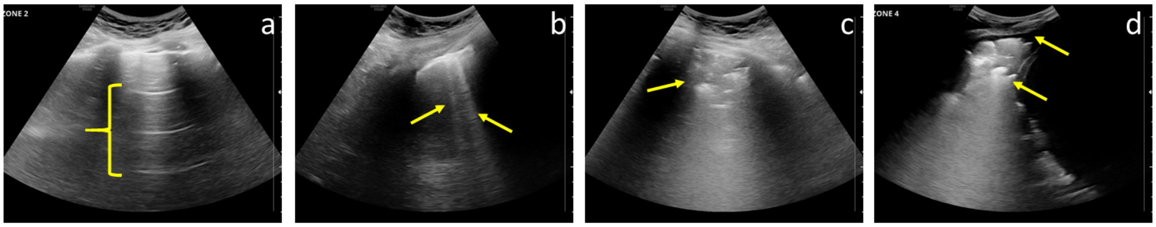

Figure 1: Example lung ultrasound frames and features: (a) A-lines, (b) B-line, (c) Consolidation, (d) Pleural effusion with consolidation

For LUS, there are several main patterns in lung ultrasound images, including A-line, B-line, consolidation, and pleural effusion, with healthy lungs typically exhibiting only A-lines while other patterns may emerge or coexist associated with different lung diseases(Ni et al., 2024), as shown in the examples in Figure 1. This characteristic inherently makes LUS pattern recognition a multi-label classification problem. However, recent research has focused mainly on the recognition of a single pattern (B-line) (Kerdegari et al., 2021; Arntfield et al., 2021) or, on multi-class classification for specific lung diseases(Nhat et al., 2023; Shea et al., 2023; Howell et al., 2024; Diaz-Escobar et al., 2021; Roy et al., 2020).

LUS examination typically involves collecting video clips from multiple lung zones for LUS pattern analysis. While deep learning methods have shown promise in automated video analysis (Kerdegari et al., 2021; Shea et al., 2023; De Rosa et al., 2022; Barros et al., 2021), applying video recognition techniques to LUS faces several challenges due to the fundamental differences between ultrasound and natural imagery. Current approaches primarily employ CNN+LSTM networks or 3D convolution-based architectures (Tran et al., 2015, 2018) to capture spatiotemporal features in LUS sequences (Shea et al., 2023; Barros et al., 2021; Dastider et al., 2021; Liu et al., 2024; Ebadi et al., 2021). These conventional approaches focus on learning temporal dependencies across the entire video sequence. Smith, D. H. et al. (Smith et al., 2023) challenge this methodology, arguing that models developed for human action recognition are not optimal in some practical scenarios involving medical ultrasound and that models assuming temporal independence demonstrate better sample efficiency. In the specific case of LUS data, we hypothesize that a hybrid approach considering both local temporal dependencies (relationships between a target frame and its neighboring frames) and frame-wise features (individual spatial features extracted from each frame independently) might be more effective for accurate pattern recognition in LUS video.

Contributions: We introduce the Lung Ultrasound Video Network (LUV-Net), a deep learning model designed to address multi-pattern recognition in LUS videos by combining pattern-specific attention with efficient temporal feature extraction. Based on prior research (Smith et al., 2023), our method is composed of an attention-based spatial feature extraction network designed for each LUS pattern and a temporal feature extraction network that considers the unique characteristics of LUS videos, ensuring a comprehensive and pattern-specific representation as shown in Figure 2. To evaluate its effectiveness, we compare LUV-Net's performance with conventional video models on both internal and temporally separated test sets and further enhance interpretability by visualizing the extraction of important frames for each LUS pattern.

2. Methods

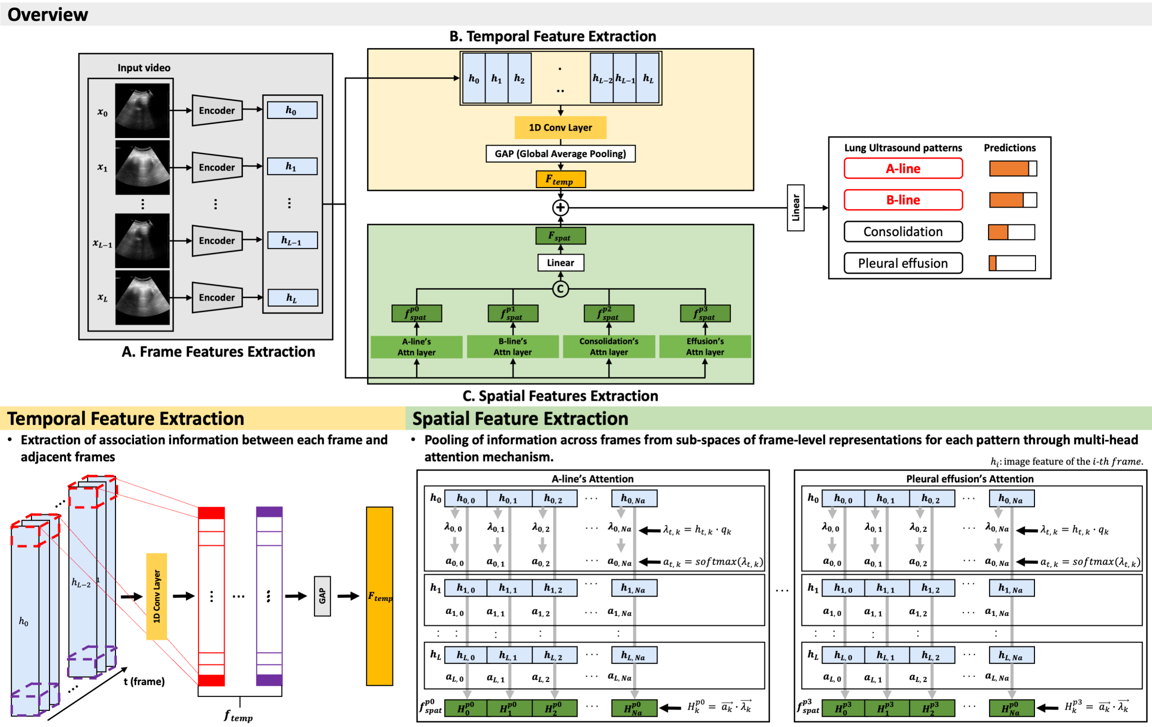
Overview
B. Temporal Feature Extraction
Input video
x_0
x_1
x_{L-1}
x_L
Encoder
h_0
h_1
h_{L-1}
h_L
h_2
h_{L-2}
1D Conv Layer
GAP (Global Average Pooling)
F_{temp}
F_{spat}
Linear
f_{spat}^{p0}
f_{spat}^{p1}
f_{spat}^{p2}
f_{spat}^{p3}
A-line's Attn layer
B-line's Attn layer
Consolidation's Attn layer
Effusion's Attn layer
A. Frame Features Extraction
C. Spatial Features Extraction
Linear
Lung Ultrasound patterns
Predictions
A-line
B-line
Consolidation
Pleural effusion
Temporal Feature Extraction
Extraction of association information between each frame and adjacent frames
t (frame)
1D Conv Layer
GAP
F_{temp}
f_{temp}
Spatial Feature Extraction
Pooling of information across frames from sub-spaces of frame-level representations for each pattern through multi-head attention mechanism.
h_i: image feature of the i-th $frame$.
A-line's Attention
Pleural effusion's Attention
$h_{0,0}$ $h_{0,1}$ $h_{0,2}$ $h_{0,Na}$
$\lambda_{0,0}$ $\lambda_{0,1}$ $\lambda_{0,2}$ $\lambda_{0,Na}$
$\lambda_{t,k} = h_{t,k} \cdot q_k$
$a_{0,0}$ $a_{0,1}$ $a_{0,2}$ $a_{0,Na}$
$a_{t,k} = softmax(\lambda_{t,k})$
$h_{1,0}$ $h_{1,1}$ $h_{1,2}$ $h_{1,Na}$
$a_{1,0}$ $a_{1,1}$ $a_{1,2}$ $a_{1,Na}$
$h_{L,0}$ $h_{L,1}$ $h_{L,2}$ $h_{L,Na}$
$a_{L,0}$ $a_{L,1}$ $a_{L,2}$ $a_{L,Na}$
H_0^{p0} H_1^{p0} H_2^{p0} H_{Na}^{p0}
$H_k^{p0} = \vec{a_k} \cdot \vec{\lambda_k}$
H_0^{p3} H_1^{p3} H_2^{p3} H_{Na}^{p0}
$H_k^{p3} = \vec{a_k} \cdot \vec{\lambda_k}$

Figure 2: Proposed framework of the Lung Ultrasound Video Network (LUV-Net).

The proposed Lung Ultrasound Video Network (LUV-Net), as shown in Figure 2, consists of four main components: frame feature extraction, temporal feature extraction, spatial feature extraction, and a feature fusion stage.

2.1. Frame Feature Extraction

The input clip X to the model is a sequence of L frames, $X = (x_1, x_2, \ldots, x_L)$. Each frame is fed into the CNN encoder individually, embedding the features into D-dimensional vectors.

The features for each frame are denoted as $h_1, h_2, \dots, h_L$, where $h \in \mathbb{R}^D$. These frame-level features are then used in both the temporal feature extraction and spatial feature extraction networks.

2.2. Temporal Feature Extraction

Our temporal feature extraction module is designed to capture the dynamic relationships between consecutive frames in LUS video sequences. This process utilizes a 1D convolution operation to analyze temporal patterns across the video timeline. For a sequence of frame features $h_1, h_2, \dots, h_L$, we apply a convolution kernel with size k that processes overlapping windows of consecutive frames. This operation generates temporally aware features:

$$f_{temp} = Conv1D(h_1, h_2, \dots, h_L) \tag{1}$$

where $f_{temp} \in \mathbb{R}^{L \times D}$ represents the extracted temporal features. The kernel size k determines how many adjacent frames are analyzed together, while the stride controls the step size between windows. To obtain the final video-level temporal features, we aggregate the temporal features using Global Average Pooling (GAP), which summarizes the temporal information into a single feature vector:

$$F_{temp} = GAP(f_{temp}) \tag{2}$$

2.3. Spatial Feature Extraction

Our approach employs a pattern-specific spatial attention mechanism that builds upon the work of (Smith et al., 2023), with a key modification to handle different LUS patterns independently. While prior study applied attention mechanisms uniformly across all features, we recognize that different LUS patterns (A-line, B-line, Consolidation, and Pleural effusion) may require attention to different frames within the video sequence. Therefore, we apply separate attention mechanisms for each pattern.

Each frame representation, denoted as $h_1, h_2, \dots, h_L$, is partitioned into N_a segments, corresponding to the number of attention heads. These segments are represented as $h_{t,k}$, where $t = 1, 2, \dots, L$ represents the frame index and $k = 1, 2, \dots, N_a$ represents the attention head index, and each $h_{t,k}$ has a dimensionality of $d_a = D/N_a$. Following (Smith et al., 2023), instead of comparing subspaces with other frames in the sequence, we compute attention scores using learned global query vectors. The use of global query vectors arises from our inductive prior that the recognition task involves locating key pieces of information at any point in the sequence, with the inferred query vectors representing that key information.

The attention mechanism computes dot products between frame representations and global query vectors: $\lambda_{t,k} = h_{t,k} \cdot q_k$. These scores are normalized using the softmax function to obtain attention vectors $a_{t,k}$ that is the attention vector derived from the normalized scores. The video-level representation from each head is computed as: $H_k^p = a_{t,k} \cdot \lambda_{t,k}$. For each pattern (p), we concatenate the outputs from all attention heads to obtain the pattern-specific spatial feature representation:

$$f_{spat}^p = concat([H_1^p, H_2^p, \dots, H_{N_a}^p]) \tag{3}$$

Finally, we combine the pattern-specific features and project them through a linear layer to match the temporal feature dimensionality:

$$F_{spat} = Linear(concat(f^{p0}_{spat}, f^{p1}_{spat}, f^{p2}_{spat}, f^{p3}_{spat})) \tag{4}$$

where superscripts p0, p1, p2, and p3 correspond to A-line, B-line, consolidation, and pleural effusion patterns respectively. This dimension reduction ensures balanced feature dimensionality between spatial and temporal representations, preventing bias in the final prediction.

2.4. Feature Fusion Stage

In the Feature Fusion Stage, the temporal and spatial features are combined to create a comprehensive video representation. This is achieved by summing the temporal feature representation F_{temp} with the spatial feature representation F_{spat}. Subsequently, The video-level prediction can then be computed using a fully-connected linear layer for multi-label classification of LUS patterns.

3. Materials and Experimental Setup

3.1. Datasets

Data collection and Annotation. The LUS scans were performed using a HS60 Ultrasound Machine (Samsung Healthcare, Republic of Korea) with a low–medium frequency (3–5 MHz) convex probe. Following the protocol recommended by the Bedside Lung Ultrasound in Emergency (BLUE) protocol (Lichtenstein and Meziere, 2008), three lung points—anterior, lateral, and posterior areas—were assessed by the operators, ensuring that a minimum of six videos were acquired per patient. From January to December 2023, we collected 370 LUS videos, each with a duration of approximately 5 seconds, from 36 patients in the ICU of Seoul National University Hospital(SNUH) for the model development set. Videos with poor quality due to blurring or darkness, which hindered the differentiation of LUS patterns, were excluded from the dataset. As a result, the final dataset used for model training consisted of 341 LUS videos from 35 patients. Additionally, for model validation, a temporally separated test set was collected from 11 patients in the SNUH ICU between January and December 2024, resulting in the acquisition of 56 LUS videos. Each video ranged from 5 to 8 seconds in duration, with a frame rate of 30 frames per second (fps) and a resolution of 924 × 1232 pixels. Each video was independently labeled by two clinicians with over 1 year of experience in lung ultrasound (LUS), who annotated the regions containing LUS patterns at the frame level. In cases where there was agreement between the two clinicians, their consensus label was adopted as the final label. For disagreements, a clinician with over 8 years of experience conducted the final review and provided the definitive label.

Preprocessing Stage. For training the model, we conducted several preprocessing steps on the video data. First, we segmented each video into one-second clips (30 frames) with a 20% frame overlap (6 frames) between consecutive frames. This preprocessing resulted in 2,588 clips from 370 videos for the model development set, and 366 clips from 56 videos for the temporally separated test set. Additionally, we downsampled all video frames to a uniform resolution of 256 × 256 pixels. To enhance data diversity, we implemented augmentation techniques including random horizontal flips and controlled rotations up to 10 degrees. The augmented images were subsequently converted into tensor format for deep learning processing. To ensure robust evaluation, we randomly selected 10% of the total data as a held-out test set and subjected the remaining 90% to 5-fold cross-validation, ensuring strict separation of patient data across all sets.

3.2. Implementation Details

In this study, we compared the performance of our proposed LUV-Net model with several conventional video models, including the USVN model proposed by Smith, D. H. et al. (Smith et al., 2023), C3D, R2Plus1D, and CNN+LSTM models. For a fair comparative analysis, all baseline models were trained on our dataset under the same experimental conditions. Additionally, we evaluated a frame-based method that processes frame-level features independently and derives final video predictions by pooling across the frames.

For our model architecture, we employed the ImageNet pre-trained DenseNet-161 as the encoder, which was also used in the CNN+LSTM and frame-based method. Based on a hyperparameter search (detailed in Table 7, Table 8, Appendix C), we optimized the kernel size and number of attention heads for both temporal and spatial feature extraction networks. The temporal feature extraction network incorporated a single 1D convolutional layer with a kernel size of 13. In the spatial feature extraction network, we implemented multiple attention heads with Na = 8. During training, we used the Adam optimizer with a learning rate of 1e-6, a batch size of 4, and trained the model for up to 150 epochs. We also implemented an early stopping mechanism with a patience value of 50 to prevent overfitting. Optimal thresholds for balancing sensitivity and specificity were determined at the epoch with the lowest validation loss. These thresholds were applied to both the development set and temporally separated set for performance evaluation. All model training was conducted using a single A100 GPU.

4. Results

The evaluation was conducted using 5-fold cross-validation on both development and temporally separated sets, with an additional comparative analysis between LUV-Net and its variant without temporal feature extraction. The main tables present the mean and standard deviation values for each metric, with bold and underlined values indicating the best and second-best performances, respectively. The detailed results for each fold of the 5-fold cross-validation, including P values (p <0.05) and 95% confidence intervals (CI), are presented in Tables 4, 5, and 6 in Appendix A.

4.1. Development Dataset

Table 1 shows the 5-fold cross-validation results for the development set. Our proposed LUV-Net demonstrates superior performance across most metrics, achieving the highest scores in B-line (AUC: 0.834±0.014), consolidation detection (AUC: 0.853±0.019), and overall performance (Micro AUC: 0.888±0.009, Macro AUC: 0.894±0.009). While the frame-based method shows competitive performance in A-line (AUC: 0.926±0.007) and pleural effusion detection (AUC: 0.973±0.004), LUV-Net maintains more consistent performance across all patterns. USVN achieves competitive results in pleural effusion (AUC: 0.931±0.037) and consolidation (AUC: 0.838±0.010), but shows high variance in B-line detection (0.770±0.137). C3D, R2Plus1D and CNN+LSTM show notably lower performance, particularly in A-line and B-line detection.

Input Type	Model	AUC				Avg	
		A-line	B-line	Consolidation	Pleural effusion	Micro	Macro
Frame (Image)	Frame-based	**0.926±0.007**	0.800±0.033	0.831±0.018	**0.973±0.004**	0.881±0.008	0.883±0.011
Video	C3D	0.789±0.062	0.626±0.038	0.788±0.023	0.930±0.002	0.791±0.025	0.792±0.023
	R2Plus1D	0.726±0.051	0.619±0.078	0.801±0.031	0.830±0.071	0.720±0.031	0.746±0.035
	CNN+LSTM	0.420±0.066	0.353±0.032	0.737±0.043	0.911±0.042	0.587±0.023	0.607±0.007
	USVN	0.879±0.080	0.770±0.137	0.879±0.080	0.931±0.037	0.846±0.062	0.856±0.058
	LUV-Net (ours)	0.918±0.013	**0.834±0.014**	**0.853±0.019**	0.966±0.010	**0.888±0.009**	**0.894±0.009**

Table 1: Results on the development set

4.2. Temporally Separated Dataset

On the temporally separated dataset (Table 2), LUV-Net maintains robust performance with the highest AUC in A-line detection (0.835±0.057) and overall metrics (Micro: 0.858±0.023, Macro: 0.844±0.015). While USVN achieves the highest performance in consolidation detection (0.846±0.022) and shows comparable B-line detection (0.848±0.039), LUV-Net demonstrates superior performance in other patterns, particularly in A-line and pleural effusion detection. C3D shows competitive performance in pleural effusion detection (0.882±0.032), but LUV-Net demonstrates more balanced performance across all patterns, confirming its effectiveness in multi-label LUS pattern classification.

Input Type	Model	AUC				Avg	
		A-line	B-line	Consolidation	Pleural effusion	Micro	Macro
Frame (Image)	Frame-based	0.763±0.025	**0.893±0.005**	0.772±0.019	0.812±0.023	0.809±0.031	0.813±0.014
Video	C3D	0.795±0.030	0.703±0.047	0.747±0.010	**0.882±0.032**	0.848±0.025	0.784±0.025
	R2Plus1D	0.659±0.056	0.634±0.010	0.684±0.027	0.868±0.011	0.734±0.030	0.713±0.014
	CNN+LSTM	0.366±0.092	0.491±0.047	0.756±0.042	0.754±0.084	0.602±0.069	0.593±0.029
	USVN	0.795±0.024	0.848±0.039	**0.846±0.022**	0.845±0.068	0.824±0.049	0.833±0.027
	LUV-Net (ours)	**0.835±0.057**	0.862±0.022	0.799±0.021	0.873±0.026	**0.858±0.023**	**0.844±0.015**

Table 2: Results on the temporally separated set

4.3. Effectiveness of Temporal Feature Extraction

Table 3 compares the performance of LUV-Net with and without temporal feature extraction on the development set. Incorporating temporal features consistently improved performance across all patterns, achieving higher Macro-AUC (0.894±0.009 vs 0.885±0.011) and Micro-AUC (0.888±0.009 vs 0.880±0.014), which demonstrates the effectiveness of temporal feature extraction in enhancing pattern recognition.

	AUC				Avg	
	A-line	B-line	Consolidation	Pleural effusion	Micro	Macro
LUV-Net (w/o temporal)	0.910±0.017	0.826±0.019	0.841±0.018	0.956±0.009	0.880±0.014	0.885±0.011
LUV-Net (w/ temporal)	**0.918±0.013**	**0.834±0.015**	**0.853±0.019**	**0.966±0.010**	**0.888±0.009**	**0.894±0.009**

Table 3: Temporal feature extraction with / without on development set

4.4. Qualitative Analysis

To further investigate the interpretability of the proposed LUV-Net model, we conducted a qualitative analysis by visualizing the attention scores across video frames for multiple patterns and extracting the top-k frames with the highest scores, as detailed in Appendix B. Figure 3 shows examples from the development set, including both multi-patterns and single-pattern clips (additional examples from the temporally separated set can be found in Appendix B). The top-3 frames for each pattern are marked with red dots, with (A), (B), and (C) denoting these frames. The green dashed lines represent the ground truth, where 1 indicates presence and 0 indicates absence of the pattern. Our model identifies critical frames and attends to the relevant regions of each pattern, both in multi-pattern and single-pattern clips.

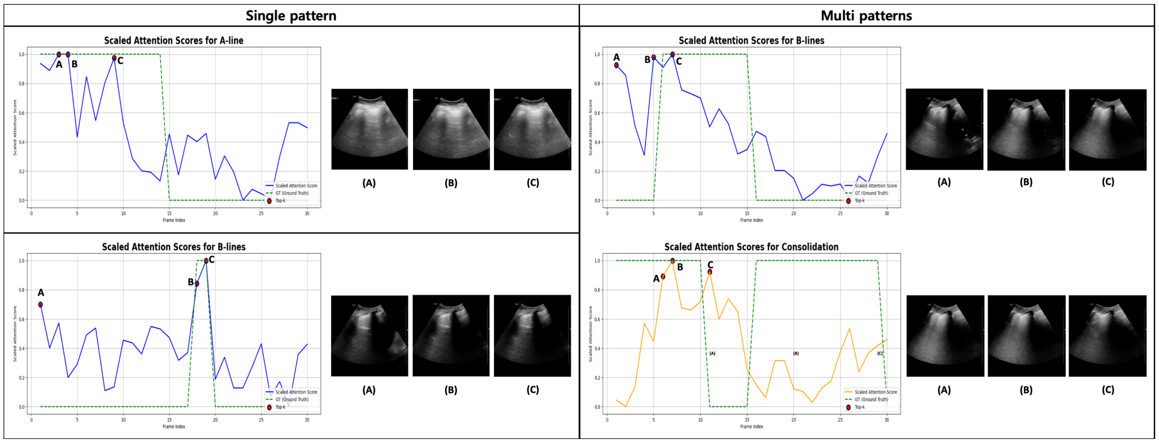

Figure 3: Visualization of attention scores and corresponding top-3 frames for different patterns.

5. Discussion and Conclusions

We propose LUV-Net, a deep learning model for multi-label classification of LUS patterns in ultrasound video sequences. Our 5-fold cross-validation results show superior performance compared to conventional video models and the USVN model across all four LUS patterns. The model integrates spatial attention mechanisms, focusing on relevant regions for each pattern, and temporal feature extraction to capture relationships between frames. This enhances classification performance and interpretability by identifying when and where specific patterns appear. However, the study has limitations: our dataset comes from a single institution, limiting generalizability, and the DenseNet-161 encoder could be optimized for real-time applications with lighter models. Finally, real-world validation through reader tests with clinicians is essential to assess practical effectiveness. In conclusion, LUV-Net represents a significant step toward automated LUS analysis, offering a robust solution for multi-label classification adaptable to diverse clinical applications.

References

Robert Arntfield, Derek Wu, Jared Tschirhart, Blake VanBerlo, Alex Ford, Jordan Ho, Joseph McCauley, Benjamin Wu, Jason Deglint, Rushil Chaudhary, et al. Automation of lung ultrasound interpretation via deep learning for the classification of normal versus abnormal lung parenchyma: a multicenter study. *Diagnostics*, 11(11):2049, 2021.

Bruno Barros, Paulo Lacerda, Celio Albuquerque, and Aura Conci. Pulmonary covid-19: learning spatiotemporal features combining cnn and lstm networks for lung ultrasound video classification. *Sensors*, 21(16):5486, 2021.

Etrusca Brogi, Elena Bignami, Anna Sidoti, Mohammed Shawar, Luna Gargani, Luigi Vetrugno, Giovanni Volpicelli, and Francesco Forfori. Could the use of bedside lung ultrasound reduce the number of chest x-rays in the intensive care unit? *Cardiovascular ultrasound*, 15:1–5, 2017.

Danilo Buonsenso and Cristina De Rose. Implementation of lung ultrasound in low-to middle-income countries: a new challenge global health? *European Journal of Pediatrics*, 181(1):1–8, 2022.

Ankan Ghosh Dastider, Farhan Sadik, and Shaikh Anowarul Fattah. An integrated autoencoder-based hybrid cnn-lstm model for covid-19 severity prediction from lung ultrasound. *Computers in Biology and Medicine*, 132:104296, 2021.

Laura De Rosa, Serena L'Abbate, Claudia Kusmic, Francesco Faita, et al. Applications of artificial intelligence in lung ultrasound: Review of deep learning methods for covid-19 fighting. *Artificial Intelligence in Medical Imaging*, 3(2):42–54, 2022.

Julia Diaz-Escobar, Nelson E Ordonez-Guillen, Salvador Villarreal-Reyes, Alejandro Galaviz-Mosqueda, Vitaly Kober, Raúl Rivera-Rodriguez, and Jose E Lozano Rizk. Deep-learning based detection of covid-19 using lung ultrasound imagery. *Plos one*, 16(8): e0255886, 2021.

Salehe Erfanian Ebadi, Deepa Krishnaswamy, Seyed Ehsan Seyed Bolouri, Dornoosh Zonoobi, Russell Greiner, Nathaniel Meuser-Herr, Jacob L Jaremko, Jeevesh Kapur, Michelle Noga, and Kumaradevan Punithakumar. Automated detection of pneumonia in lung ultrasound using deep video classification for covid-19. *Informatics in Medicine Unlocked*, 25:100687, 2021.

Lewis Howell, Nicola Ingram, Roger Lapham, Adam Morrell, and James R McLaughlan. Deep learning for real-time multi-class segmentation of artefacts in lung ultrasound. *Ultrasonics*, 140:107251, 2024.

Hamideh Kerdegari, Nhat Tran Huy Phung, Angela McBride, Luigi Pisani, Hao Van Nguyen, Thuy Bich Duong, Reza Razavi, Louise Thwaites, Sophie Yacoub, Alberto Gomez, et al. B-line detection and localization in lung ultrasound videos using spatiotemporal attention. *Applied Sciences*, 11(24):11697, 2021.

Daniel A Lichtenstein and Gilbert A Meziere. Relevance of lung ultrasound in the diagnosis of acute respiratory failure*: the blue protocol. *Chest*, 134(1):117–125, 2008.

Jang Sun Lim, Sanghun Lee, Han Ho Do, and Kyu Ho Oh. Can limited education of lung ultrasound be conducted to medical students properly? a pilot study. *BioMed Research International*, 2017(1):8147075, 2017.

Yiwen Liu, Wenyu Xing, Chao He, and Mingbo Zhao. Domain knowledge-enhanced integrated model for lus video scoring. In *2024 IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium (UFFC-JS)*, pages 1–3. IEEE, 2024.

Thomas J Marini, Deborah J Rubens, Yu T Zhao, Justin Weis, Timothy P O'connor, William H Novak, and Katherine A Kaproth-Joslin. Lung ultrasound: the essentials. *Radiology: Cardiothoracic Imaging*, 3(2):e200564, 2021.

Phung Tran Huy Nhat, Nguyen Van Hao, Phan Vinh Tho, Hamideh Kerdegari, Luigi Pisani, Le Ngoc Minh Thu, Le Thanh Phuong, Ha Thi Hai Duong, Duong Bich Thuy, Angela McBride, et al. Clinical benefit of ai-assisted lung ultrasound in a resource-limited intensive care unit. *Critical Care*, 27(1):257, 2023.

Yuanlu Ni, Yang Cong, Chengqian Zhao, Jinhua Yu, Yin Wang, Guohui Zhou, and Mengjun Shen. Active learning based on multi-enhanced views for classification of multiple patterns in lung ultrasound images. *Computerized Medical Imaging and Graphics*, 118:102454, 2024.

Subhankar Roy, Willi Menapace, Sebastiaan Oei, Ben Luijten, Enrico Fini, Cristiano Saltori, Iris Huijben, Nishith Chennakeshava, Federico Mento, Alessandro Sentelli, et al. Deep learning for classification and localization of covid-19 markers in point-of-care lung ultrasound. *IEEE transactions on medical imaging*, 39(8):2676–2687, 2020.

Daniel E Shea, Sourabh Kulhare, Rachel Millin, Zohreh Laverriere, Courosh Mehanian, Charles B Delahunt, Dipayan Banik, Xinliang Zheng, Meihua Zhu, Ye Ji, et al. Deep learning video classification of lung ultrasound features associated with pneumonia. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3103–3112, 2023.

Gentle S Shrestha, Dameera Weeratunga, and Kylie Baker. Point-of-care lung ultrasound in critically ill patients. *Reviews on recent clinical trials*, 13(1):15–26, 2018.

D Hudson Smith, John Paul Lineberger, and George H Baker. On the relevance of temporal features for medical ultrasound video recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 744–753. Springer, 2023.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

Laurent Zieleskiewicz, Alexandre Lopez, Sami Hraiech, Karine Baumstarck, Bruno Pastene, Mathieu Di Bisceglie, Benjamin Coiffard, Gary Duclos, Alain Boussuges, Xavier Bobbia, et al. Bedside pocus during ward emergencies is associated with improved diagnosis and outcome: an observational, prospective, controlled study. *Critical Care*, 25:1–12, 2021.

Appendix A. 5-Fold Cross Validation Results

A.1. Results on the development set

Tables 4 present the detailed results of 5-fold cross validation on the development. For each fold, we report the Area Under the Curve (AUC) scores with 95% confidence intervals (CI) and statistical significance (p<0.05) compared to LUV-Net performance. On the development dataset, LUV-Net demonstrated consistent performance across all folds, with the highest performance observed in Fold 3 (Micro-average: 0.903, Macro-average: 0.909). The model showed particularly stable performance in Pleural Effusion detection across all folds, maintaining AUC scores above 0.95. While frame-based method achieved marginally better performance in A-line detection for some folds (e.g., Fold 3: 0.934 vs 0.919), LUV-Net consistently outperformed all baselines in B-line detection and Consolidation patterns across most folds. Notable performance variations were observed across different folds, particularly for B-line detection, where AUC scores ranged from 0.821 to 0.855. This variation suggests that certain lung ultrasound patterns may be more challenging to detect consistently, possibly due to the inherent variability in their appearance or recording conditions. Conventional video models (C3D and R2Plus1D) showed significant performance degradation (p¡0.05) compared to LUV-Net across most folds, particularly in detecting B-lines and Consolidation patterns. The CNN+LSTM baseline demonstrated the most unstable performance, with significantly lower AUC scores across all patterns and folds, indicating the importance of our proposed architecture for temporal feature learning.

Fold #	Input Type	Model	AUC				Avg	
			A-line	B-line	Consolidation	Pleural effusion	Micro	Macro
Fold 0	Frame (Image)	frame-based	**0.917 (0.886-0.949)**	0.837 (0.800-0.875)	0.821 (0.775-0.866)	**0.979 (0.965-0.994)**	0.885	0.889
	Video	C3D	0.758 (0.708-0.808)*	0.640 (0.586-0.694)*	0.778 (0.733-0.822)*	0.931 (0.902-0.960)*	0.801	0.811
		R2Plus1D	0.707 (0.657-0.757)*	0.573 (0.513-0.632)*	0.802 (0.749-0.854)	0.911 (0.878-0.945)*	0.703	0.750
		CNN+LSTM	0.368 (0.315-0.421)*	0.347 (0.288-0.406)*	0.760 (0.705-0.815)*	0.924 (0.891-0.957)*	0.602	0.601
		USVN	0.902 (0.865-0.939)	0.834 (0.796-0.872)	0.829 (0.788-0.871)*	0.948 (0.926-0.969)*	0.868	0.880
		LUV-Net (ours)	0.900 (0.860-0.940)	**0.841 (0.802-0.880)**	**0.851 (0.810-0.892)**	0.974 (0.959-0.989)	**0.880**	**0.893**
Fold 1	Frame (Image)	frame-based	**0.924 (0.897-0.951)***	0.791 (0.747-0.832)	0.833 (0.783-0.883)*	**0.971 (0.958-0.984)**	0.875	0.880
	Video	C3D	0.762 (0.712-0.811)*	0.564 (0.509-0.618)*	0.814 (0.774-0.855)*	0.928 (0.899-0.956)	0.775	0.769
		R2Plus1D	0.653 (0.598-0.708)*	0.534 (0.479-0.590)*	0.805 (0.755-0.856)*	0.814 (0.750-0.878)*	0.686	0.703
		CNN+LSTM	0.478 (0.421-0.536)*	0.328 (0.274-0.382)*	0.724 (0.667-0.782)*	0.936 (0.904-0.967)	0.605	0.618
		USVN	0.729 (0.676-0.783)*	0.536 (0.479-0.594)*	0.848 (0.808-0.888)	0.866 (0.814-0.919)*	0.734	0.747
		LUV-Net (ours)	0.914 (0.886-0.943)	**0.821 (0.779-0.862)**	**0.860 (0.822-0.897)**	0.950 (0.927-0.972)	**0.882**	**0.887**
Fold 2	Frame (Image)	frame-based	0.930 (0.905-0.955)	0.747 (0.695-0.798)*	0.820 (0.774-0.866)	**0.971 (0.955-0.988)**	0.876	0.868
	Video	C3D	0.714 (0.663-0.766)*	0.648 (0.590-0.707)*	0.756 (0.708-0.805)*	0.929 (0.904-0.955)*	0.756	0.764
		R2Plus1D	0.737 (0.692-0.782)*	0.732 (0.674-0.790)*	0.747 (0.692-0.802)*	0.828 (0.767-0.888)*	0.743	0.763
		CNN+LSTM	0.350 (0.299-0.401)*	0.349 (0.290-0.408)*	0.779 (0.727-0.831)*	0.944 (0.917-0.970)*	0.585	0.607
		USVN	0.924 (0.896-0.951)	0.821 (0.780-0.862)	0.833 (0.792-0.874)	0.953 (0.932-0.974)	0.881	0.884
		LUV-Net (ours)	**0.933 (0.908-0.958)**	**0.827 (0.786-0.868)**	**0.834 (0.796-0.871)**	0.963 (0.944-0.983)	**0.890**	**0.890**
Fold 3	Frame (Image)	frame-based	**0.934 (0.910-0.958)**	0.823 (0.782-0.865)	0.859 (0.818-0.899)	0.969 (0.945-0.993)	0.894	0.897
	Video	C3D	0.859 (0.822-0.896)*	0.670 (0.615-0.725)*	0.795 (0.753-0.836)*	0.930 (0.902-0.957)*	0.820	0.815
		R2Plus1D	0.789 (0.746-0.832)*	0.676 (0.616-0.736)*	0.831 (0.786-0.876)*	0.869 (0.819-0.919)*	0.762	0.793
		CNN+LSTM	0.395 (0.340-0.450)*	0.332 (0.276-0.388)*	0.758 (0.703-0.813)*	0.924 (0.891-0.957)*	0.593	0.601
		USVN	0.924 (0.895-0.954)	0.849 (0.813-0.886)	0.852 (0.815-0.888)*	0.951 (0.930-0.972)*	0.884	0.895
		LUV-Net (ours)	0.919 (0.891-0.948)	**0.855 (0.817-0.893)**	**0.882 (0.848-0.916)**	**0.975 (0.958-0.992)**	**0.903**	**0.909**
Fold 4	Frame (Image)	frame-based	0.924 (0.897-0.952)	0.803 (0.757-0.848)	0.822 (0.779-0.864)	**0.977 (0.964-0.990)**	0.877	0.882
	Video	C3D	0.853 (0.815-0.891)*	0.606 (0.549-0.663)*	0.798 (0.756-0.839)*	0.934 (0.908-0.960)*	0.805	0.799
		R2Plus1D	0.745 (0.698-0.791)*	0.581 (0.524-0.638)*	0.819 (0.769-0.869)	0.730 (0.675-0.786)*	0.704	0.721
		CNN+LSTM	0.509 (0.447-0.570)*	0.409 (0.347-0.472)*	0.666 (0.605-0.726)*	0.836 (0.779-0.892)*	0.550	0.606
		USVN	0.916 (0.884-0.947)	0.810 (0.767-0.853)	0.830 (0.787-0.872)	0.938 (0.913-0.963)*	0.863	0.875
		LUV-Net (ours)	**0.924 (0.895-0.953)**	**0.824 (0.782-0.867)**	**0.837 (0.795-0.880)**	0.967 (0.949-0.985)	**0.886**	**0.890**

Table 4: Development set results, * indicates that the P-value is less than 0.05.

A.2. Results on the temporally separated validation set

Table 5 presents the performance of the proposed LUV-Net and five baseline models—USVN, C3D, R2Plus1D, CNN+LSTM, and frame-based method on the temporally separated validation set, using 5-fold cross-validation on temporally separated validation sets. LUV-Net maintained robust performance while showing some interesting patterns across different folds. Notably, the model demonstrated strong performance in detecting A-lines with AUC scores ranging from 0.771 to 0.895, achieving the highest score in Fold 3. This performance was particularly meaningful given the temporal gap between training and validation data. B-line detection showed consistently high performance across all folds (AUC: 0.835-0.890), though the frame-based method occasionally outperformed LUV-Net in this pattern (e.g., Fold 0: 0.899 vs 0.847, Fold 2: 0.898 vs 0.862). However, USVN showed competitive performance in detecting Consolidation patterns, achieving the highest AUC scores in several folds (Fold 0: 0.842, Fold 1: 0.844, Fold 4: 0.865). Conventional video models (C3D and R2Plus1D) showed statistically significant performance degradation (p¡0.05) in most cases, particularly in B-line and Consolidation detection. The CNN+LSTM baseline consistently underperformed, with AUC scores significantly lower than LUV-Net across all patterns and folds. Interestingly, while some performance metrics showed higher variance compared to the development dataset, LUV-Net maintained relatively stable micro and macro averages across folds (micro: 0.825-0.882, macro: 0.826-0.861).

Fold #	Input Type	Model	AUC				Avg	
			A-line	B-line	Consolidation	Pleural effusion	Micro	Macro
Fold 0	Frame (Image)	frame-based	0.763 (0.683-0.788)*	**0.899 (0.867-0.930)***	0.773 (0.717-0.829)	0.793 (0.656-0.931)*	0.807	0.801
	Video	C3D	0.796 (0.739-0.852)	0.735 (0.683-0.787)*	0.745 (0.670-0.820)	0.868 (0.793-0.942)	0.865	0.788
		R2Plus1D	0.680 (0.604-0.756)*	0.626 (0.571-0.682)*	0.716 (0.647-0.786)*	0.854 (0.761-0.947)	0.706	0.721
		CNN+LSTM	0.374 (0.295-0.452)*	0.470 (0.413-0.527)*	0.795 (0.742-0.848)	0.781 (0.666-0.895)*	0.550	0.606
		USVN	0.803 (0.759-0.847)	0.874 (0.838-0.910)*	**0.842 (0.801-0.883)***	0.775 (0.633-0.917)*	0.837	0.825
		LUV-Net (ours)	**0.819 (0.775-0.863)**	0.847 (0.809-0.885)	0.778 (0.723-0.833)	**0.877 (0.803-0.930)**	**0.882**	**0.832**
Fold 1	Frame (Image)	frame-based	0.790 (0.744-0.836)*	**0.889 (0.857-0.921)***	0.751 (0.687-0.815)*	0.851 (0.752-0.950)	0.824	0.821
	Video	C3D	0.785 (0.726-0.844)*	0.688 (0.636-0.740)*	0.749 (0.677-0.821)	0.877 (0.802-0.953)	**0.853**	0.777
		R2Plus1D	0.575 (0.496-0.655)*	0.638 (0.584-0.693)*	0.664 (0.591-0.736)*	0.875 (0.802-0.947)	0.710	0.690
		CNN+LSTM	0.328 (0.255-0.401)*	0.464 (0.407-0.521)*	0.759 (0.699-0.820)	0.832 (0.734-0.929)*	0.577	0.597
		USVN	0.774 (0.686-0.803)*	0.772 (0.726-0.818)*	**0.844 (0.803-0.885)**	0.838 (0.749-0.926)*	0.760	0.801
		LUV-Net (ours)	**0.893 (0.856-0.929)**	0.835 (0.794-0.875)	0.802 (0.746-0.859)	**0.907 (0.839-0.975)**	0.825	**0.861**
Fold 2	Frame (Image)	frame-based	0.775 (0.728-0.822)	**0.898 (0.867-0.930)***	0.796 (0.740-0.853)	0.805 (0.682-0.928)*	0.837	**0.838**
	Video	C3D	0.761 (0.695-0.828)	0.617 (0.562-0.672)*	0.733 (0.662-0.803)	0.835 (0.741-0.930)*	0.809	0.739
		R2Plus1D	0.718 (0.642-0.795)	0.638 (0.581-0.695)*	0.656 (0.583-0.730)*	0.875 (0.807-0.943)	0.723	0.724
		CNN+LSTM	0.229 (0.170-0.288)*	0.562 (0.505-0.619)*	0.794 (0.742-0.846)	0.822 (0.722-0.923)*	0.771	0.603
		USVN	**0.783 (0.739-0.826)**	0.835 (0.794-0.876)*	**0.806 (0.752-0.861)**	0.874 (0.788-0.959)	0.802	0.826
		LUV-Net (ours)	0.771 (0.725-0.818)	0.862 (0.825-0.899)	0.781 (0.728-0.834)	**0.884 (0.807-0.961)**	**0.848**	0.826
Fold 3	Frame (Image)	frame-based	0.748 (0.719-0.818)*	0.890 (0.865-0.927)*	0.763 (0.759-0.868)	0.799 (0.628-0.921)*	0.763	0.801
	Video	C3D	0.805 (0.746-0.864)*	0.741 (0.688-0.795)*	0.762 (0.692-0.833)*	**0.905 (0.846-0.964)***	0.841	0.805
		R2Plus1D	0.647 (0.566-0.728)*	0.645 (0.591-0.700)*	0.711 (0.644-0.777)*	0.858 (0.722-0.944)	0.786	0.717
		CNN+LSTM	0.444 (0.364-0.523)*	0.519 (0.462-0.577)	0.737 (0.677-0.797)*	0.754 (0.648-0.859)*	0.543	0.615
		USVN	0.804 (0.759-0.849)*	0.865 (0.828-0.901)	**0.874 (0.834-0.915)***	0.806 (0.684-0.927)	0.849	0.839
		LUV-Net (ours)	**0.895 (0.860-0.930)**	**0.878 (0.844-0.912)**	0.825 (0.733-0.877)	0.829 (0.727-0.932)	**0.874**	**0.858**
Fold 4	Frame (Image)	frame-based	0.739 (0.687-0.790)*	0.888 (0.845-0.922)	0.775 (0.711-0.838)	0.814 (0.691-0.938)*	0.813	0.805
	Video	C3D	**0.828** (0.778-0.878)	0.732 (0.679-0.784)*	0.745 (0.672-0.818)*	0.927 (0.880-0.974)*	0.870	0.810
		R2Plus1D	0.673 (0.584-0.751)*	0.621 (0.565-0.677)*	0.674 (0.601-0.747)*	0.880 (0.805-0.956)	0.745	0.714
		CNN+LSTM	0.453 (0.380-0.527)*	0.442 (0.385-0.500)*	0.693 (0.633-0.752)*	0.580 (0.456-0.704)*	0.571	0.544
		USVN	0.809 (0.766-0.853)	**0.893** (0.860-0.926)	**0.865 (0.824-0.905)***	**0.930 (0.880-0.979)***	**0.874**	**0.875**
		LUV-Net (ours)	0.798 (0.752-0.843)	0.890 (0.857-0.923)	0.811 (0.755-0.868)	0.870 (0.786-0.954)	0.859	0.844

Table 5: Temporally separated set results, * indicates that the P-value is less than 0.05

A.3. Results of temporal feature extraction study on development dataset

To validate the effectiveness of our temporal feature extraction module, we conducted an ablation study by comparing LUVM with its variant without temporal feature extraction (LUVM w/o temporal) across all five folds on the development dataset. Table 6 presents the detailed results of this comparison. The results demonstrate that the temporal feature extraction module generally contributes to improved performance, though the magnitude of improvement varies across different patterns. For A-line detection, both variants showed comparable performance, with LUVM showing slight improvements in Folds 0-2 (e.g., Fold 2: 0.933 vs 0.910) but marginally lower performance in Folds 3-4. This suggests that A-line patterns may be less dependent on temporal information for accurate detection. More notable improvements were observed in B-line detection, particularly in Folds 0 and 2, where LUVM achieved AUC scores of 0.841 and 0.827 compared to 0.827 and 0.797 for the non-temporal variant, respectively. The temporal feature extraction seemed particularly beneficial for Consolidation pattern detection, with consistent improvements across most folds and statistical significance observed in Fold 3 (0.882 vs 0.857, p¡0.05). Pleural Effusion detection showed interesting results, with the temporal feature extraction module contributing to statistically significant improvements in several folds (Folds 0 and 3). This suggests that temporal information plays a crucial role in accurately identifying this particular pattern. In terms of overall performance metrics, LUVM consistently achieved higher or comparable micro and macro averages across all folds compared to its non-temporal variant. The most substantial improvements were observed in Fold 2, where both Micro (0.890 vs 0.867) and Macro (0.890 vs 0.873) averages showed clear advantages of temporal feature extraction. These results validate the effectiveness of our temporal feature extraction module in capturing dynamic pattern characteristics while maintaining robust performance across different data splits.

Fold #	Model	AUC				Avg	
		A-line	B-line	Consolidation	Pleural effusion	Micro	Macro
Fold 0	LUV-Net (w/o temporal)	0.894 (0.855-0.934)	0.827 (0.786-0.869)	0.832 (0.793-0.871)	0.955 (0.935-0.975)*	0.867	0.879
	LUV-Net (w/ temporal)	**0.900 (0.860-0.940)**	**0.841 (0.802-0.880)**	**0.851 (0.810-0.892)**	**0.974 (0.959-0.989)**	**0.880**	**0.893**
Fold 1	LUV-Net (w/o temporal)	0.891 (0.849-0.933)*	0.819 (0.776-0.863)	**0.863 (0.827-0.900)**	0.944 (0.923-0.966)	**0.883**	0.881
	LUV-Net (w/ temporal)	**0.914 (0.886-0.943)**	**0.821 (0.779-0.862)**	0.860 (0.822-0.897)	**0.950 (0.927-0.972)**	0.882	**0.887**
Fold 2	LUV-Net (w/o temporal)	0.910 (0.875-0.944)*	0.797 (0.753-0.842)*	0.817 (0.774-0.860)	0.961 (0.940-0.981)	0.867	0.873
	LUV-Net (w/ temporal)	**0.933 (0.908-0.958)**	**0.827 (0.786-0.868)**	**0.834 (0.796-0.871)**	**0.963 (0.944-0.983)**	**0.890**	**0.890**
Fold 3	LUV-Net (w/o temporal)	**0.927 (0.898-0.955)**	0.845 (0.805-0.885)	0.857 (0.816-0.897)*	0.968 (0.952-0.985)*	0.901	0.901
	LUV-Net (w/ temporal)	0.919 (0.891-0.948)	**0.855 (0.817-0.893)**	**0.882 (0.848-0.916)**	**0.975 (0.958-0.992)**	**0.903**	**0.909**
Fold 4	LUV-Net (w/o temporal)	**0.927 (0.896-0.957)**	**0.840 (0.798-0.883)**	**0.838 (0.797-0.879**	0.952 (0.932-0.971)*	0.884	**0.891**
	LUV-Net (w/ temporal)	0.924 (0.895-0.953)	0.824 (0.782-0.867)	0.837 (0.795-0.880)	**0.967 (0.949-0.985)**	**0.886**	0.890

Table 6: Temporal feature extraction study results, * indicates that the P-value is less than 0.05

Appendix B. Qualitative Analysis

To further investigate the interpretability of the proposed LUV-Net model, we performed qualitative analysis by visualizing the attention scores across video frames for multiple labels and highlighting the most informative frames. The attention mechanism incorporated in our model provides a pathway to understand which frames contribute the most to the classification of each pattern. This section presents the results of this analysis, supported by both visual plots and mathematical expressions.

The proposed model employs a shared attention mechanism where each label $y_i \in \{y_1, y_2, \ldots, y_C\}$ (for $C = 4$, corresponding to 'A-line', 'B-lines', 'Consolidation', and 'Pleural effusion') is represented with its own set of attention query vectors $\mathbf{q}_i$. Specifically, the attention for the i-th label is computed as follows: $\alpha_i = \text{softmax}\left(\frac{\mathbf{H}\cdot\mathbf{q}_i}{\sqrt{d_k}}\right)$, where: $\mathbf{H} \in \mathbb{R}^{L\times d}$: represents the encoded feature matrix obtained from the encoder, with L denoting the number of frames and d the number of features per frame. The learnable attention query vector for the i-th label, $\mathbf{q}_i \in \mathbb{R}^{h\times d_k}$ is split into h attention heads, where each head has a dimensionality of $d_k = \frac{d}{h}$. The resulting attention scores, $\alpha_i \in \mathbb{R}^{L\times h}$, provide frame-level contributions for the i-th label. A scaling factor, $\sqrt{d_k}$, is applied to stabilize the computed attention scores.

After computing the attention weights $\{\alpha_i^h\}_{h=1}^H$ for each head h, where H is the total number of attention heads, the frame-wise attention scores are aggregated and normalized. First, we compute the raw attention sum for each frame i as:

$$\hat{\alpha}_i = \sum_{h=1}^{H} \alpha_i^h \tag{5}$$

Then, to ensure comparability across different sequences, we apply Min-Max normalization to scale the attention scores to the [0,1] range:

$$\alpha_i^{\text{norm}} = \frac{\hat{\alpha}_i - \min_j(\hat{\alpha}_j)}{\max_j(\hat{\alpha}_j) - \min_j(\hat{\alpha}_j)} \tag{6}$$

where α_i^{norm} represents the normalized attention score for frame i, indicating its relative contribution to the predicted label y_i. This normalization ensures that the most attended frame has a score of 1 and the least attended frame has a score of 0, while preserving the relative importance of each frame in the sequence.

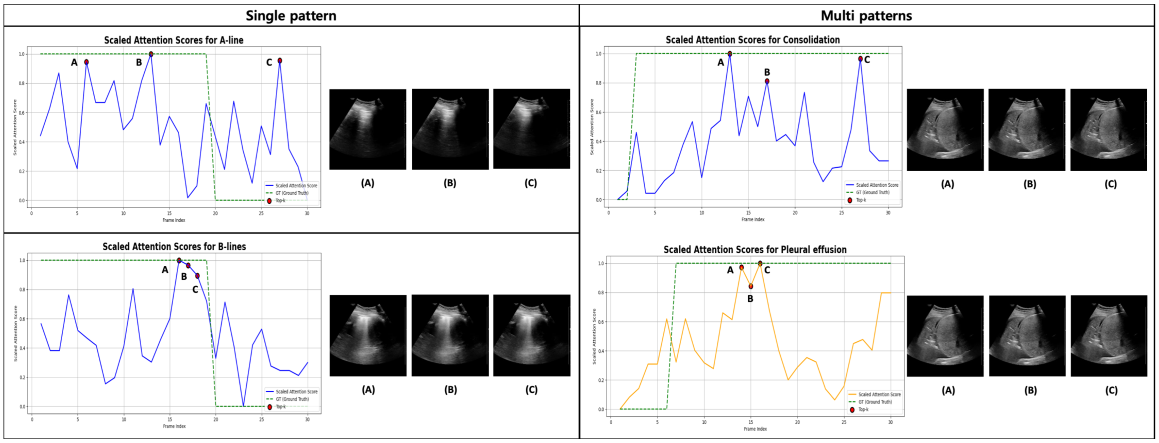

Figure 4: Visualization of attention scores and corresponding top-3 frames for different patterns on temporally separated set.

Appendix C. Ablation study

To evaluate the effectiveness of our proposed LUV-Net model and understand the impact of various architectural choices, we conducted extensive ablation studies focusing on four key aspects. First, we aimed to identify the optimal parameters for both the temporal feature extraction network and the spatial feature extraction network. Specifically, we experimented with the kernel size of the temporal feature extraction network and the number of attention heads in the spatial feature extraction network. Second, we investigated the effect of the input clip length on the model's performance. By varying the length of the input clips, we investigate the effect of various LUS video lengths on the model performance to classify LUS patterns accurately.

C.1. Effect of Kernel size of temporal feature extraction network

The analysis of different 1D kernel size (ranging from 1 to 29) on the development set revealed interesting patterns in model performance (Table 7). The results reveal that a kernel size of 13 achieves optimal overall performance. Specifically, with kernel size 13, we observe strong performance across all evaluation metrics: 0.900 for A-line detection, 0.853 for B-line detection, 0.882 for Consolidation detection, and 0.975 for Pleural effusion detection. The macro and micro averages at kernel size 13 are 0.908 and 0.902 respectively, indicating balanced performance across all classes. Based on these findings, we selected a kernel size of 13 as the most effective configuration for the temporal feature extraction network.

		Kernel size													
		1	3	5	7	9	11	13	15	17	19	21	23	25	29
AUC	A-line	0.918	0.928	0.912	0.909	0.906	0.909	0.917	0.900	0.912	0.909	0.901	0.926	0.918	0.922
	B-line	0.827	0.857	0.855	0.847	0.826	0.785	0.853	0.823	0.811	0.806	0.816	0.820	0.801	0.816
	Consolidation	0.862	0.856	0.861	0.846	0.876	0.870	0.882	0.880	0.884	0.884	0.860	0.897	0.892	0.881
	Pleural effusion	0.965	0.969	0.965	0.963	0.977	0.971	0.975	0.975	0.972	0.968	0.973	0.972	0.975	0.964
Avg	Micro	0.891	0.902	0.886	0.888	0.894	0.889	0.902	0.893	0.890	0.890	0.874	0.907	0.902	0.888
	Macro	0.894	0.904	0.899	0.893	0.898	0.885	0.908	0.896	0.896	0.893	0.889	0.905	0.898	0.897

Table 7: Development set Kernel size

C.2. Effect of number of attention heads of spatial feature extraction network

To investigate the optimal number of attention heads in our spatial feature extraction network, we conducted experiments varying the number of attention heads from 1 to 96. Table 8 presents the performance evaluation across different metrics for each attention head configuration. Our experimental results indicate that the number of attention heads has a relatively stable impact on model performance. When examining the results, we observe that using 8 attention heads achieves optimal performance across most metrics. With 8 attention heads, the model demonstrates strong results with AUC scores of 0.917 for A-line detection, 0.855 for B-line detection, 0.882 for Consolidation detection, and 0.975 for Pleural effusion detection. Furthermore, both micro and macro averages achieved the best overall performance (Micro: 0.903, Macro: 0.909) with 8 attention heads. Interestingly, increasing the number of attention heads beyond 8 does not yield significant performance improvements. Similarly, using fewer attention heads (1, 2, 4) shows marginally lower performance across most metrics.

		Attn head num						
		1	2	4	8	16	32	96
AUC	A-line	0.914	0.919	0.918	0.919	0.917	0.917	0.915
	B-line	0.852	0.854	0.851	0.855	0.852	0.853	0.850
	Consolidation	0.881	0.882	0.880	0.882	0.880	0.882	0.881
	Pleural effusion	0.974	0.975	0.973	0.975	0.975	0.975	0.976
Avg	Micro	0.901	0.902	0.897	0.903	0.901	0.902	0.901
	Macro	0.907	0.909	0.907	0.909	0.907	0.908	0.907

Table 8: Development set Number of Attn head

C.3. Performance comparison across number of frames

We conducted extensive experiments to analyze the model's performance using different numbers of input frames (40, 50, 60, and 70 frames) to determine the optimal temporal window for lung ultrasound pattern recognition. As shown in Table 9, our proposed LUV-Net consistently outperformed all baseline models (USVN, C3D, R2Plus1D, frame-based method, and CNN+LSTM) across different frame settings. Specifically, with 60 frames as input, LUV-Net achieved its best overall performance with a micro average of 0.908 and macro average of 0.905. At this setting, LUV-Net demonstrated robust performance across all pathology detection tasks, achieving AUC scores of 0.939 for A-line detection, 0.828 for B-line detection, 0.881 for Consolidation detection, and 0.963 for Pleural effusion detection. This suggests that 60 frames provide an optimal temporal window for capturing the dynamic characteristics of lung ultrasound patterns. When comparing across different frame settings, we observed that using fewer frames (40 frames) resulted in slightly decreased performance, particularly in detecting B-lines (AUC 0.829) and Consolidation (AUC 0.867). Conversely, increasing the frame count to 70 frames did not yield significant improvements and showed marginal performance degradation in some metrics (micro average decreasing from 0.908 to 0.883). Notably, the performance gap between LUV-Net and baseline models was particularly pronounced in challenging cases such as B-line detection, where LUV-Net consistently maintained superior performance across all frame settings.

Number of frames	Input Type	Model	AUC				Avg	
			A-line	B-line	Consolidation	Pleural effusion	Micro	Macro
40 frames	Video	C3D	0.805 (0.755-0.855)*	0.615 (0.549-0.680)*	0.790 (0.739-0.841)*	0.931 (0.898-0.963)*	0.789	0.788
		R2Plus1D	0.683 (0.626-0.739)*	0.732 (0.668-0.795)*	0.824 (0.771-0.876)	0.706 (0.619-0.794)*	0.679	0.738
		CNN+LSTM	0.311 (0.254-0.369)*	0.375 (0.303-0.448)*	0.747 (0.683-0.811)*	0.927 (0.875-0.978)	0.597	0.593
		USVN	0.905 (0.865-0.944)*	0.833 (0.789-0.878)	**0.874 (0.832-0.916)**	0.918 (0.865-0.971)*	0.862	0.884
		LUV-Net (ours)	**0.934 (0.906-0.962)**	0.829 (0.780-0.878)	0.867 (0.825-0.909)	**0.974 (0.957-0.990)**	**0.898**	**0.902**
	Frame (Image)	frame-based	0.924 (0.892-0.956)	**0.833 (0.785-0.880)**	0.847 (0.802-0.892)	0.950 (0.913-0.988)	0.881	0.890
50 frames	Video	C3D	0.781 (0.722-0.839)*	0.578 (0.506-0.651)*	0.763 (0.706-0.821)*	0.923 (0.888-0.958)*	0.768	0.764
		R2Plus1D	0.712 (0.650-0.774)*	0.722 (0.653-0.791)*	0.806 (0.751-0.861)	0.623 (0.510-0.736)*	0.667	0.719
		CNN+LSTM	0.335 (0.266-0.404)*	0.367 (0.287-0.447)*	0.676 (0.607-0.744)*	0.908 (0.858-0.959)*	0.583	0.574
		USVN	0.920 (0.877-0.962)	0.822 (0.771-0.873)	0.864 (0.819-0.910)	0.935 (0.895-0.975)	0.872	0.887
		LUV-Net (ours)	**0.937 (0.903-0.970)**	0.814 (0.758-0.870)	0.854 (0.808-0.900)	**0.958 (0.933-0.984)**	**0.894**	0.893
	Frame (Image)	frame-based	0.922 (0.888-0.957)	**0.841 (0.791-0.891)**	**0.863 (0.818-0.909)**	0.954 (0.914-0.993)	0.889	**0.897**
60 frames	Video	C3D	0.862 (0.815-0.909)*	0.649 (0.578-0.719)*	0.800 (0.747-0.852)*	0.940 (0.909-0.971)	0.823	0.815
		R2Plus1D	0.692 (0.627-0.758)*	0.698 (0.628-0.768)*	0.717 (0.648-0.785)*	0.639 (0.522-0.756)*	0.646	0.689
		CNN+LSTM	0.367 (0.298-0.436)*	0.397 (0.314-0.479)*	0.754 (0.683-0.825)*	0.875 (0.820-0.929)*	0.594	0.599
		USVN	0.912 (0.870-0.954)*	0.812 (0.753-0.870)	**0.885 (0.845-0.925)**	0.955 (0.925-0.986)	0.890	0.893
		LUV-Net (ours)	**0.939 (0.910-0.968)**	**0.828 (0.773-0.883)**	0.881 (0.839-0.923)	**0.963 (0.936-0.990)**	**0.908**	**0.905**
	Frame (Image)	frame-based	0.923 (0.889-0.956)	0.848 (0.798-0.899)	0.857 (0.810-0.904)	0.956 (0.921-0.991)	0.887	0.897
70 frames	Video	C3D	0.737 (0.663-0.811)*	0.664 (0.583-0.745)*	0.776 (0.712-0.839)*	0.931 (0.894-0.968)*	0.776	0.780
		R2Plus1D	0.670 (0.589-0.752)*	0.649 (0.567-0.730)*	0.671 (0.594-0.748)*	0.444 (0.311-0.577)*	0.618	0.612
		CNN+LSTM	0.316 (0.239-0.392)*	0.411 (0.320-0.501)*	0.693 (0.613-0.774)*	0.840 (0.768-0.911)*	0.573	0.567
		USVN	0.905 (0.854-0.955)	**0.821 (0.763-0.878)**	**0.868 (0.819-0.918)**	0.956 (0.927-0.985)	0.879	0.889
		LUV-Net (ours)	**0.939 (0.907-0.970)**	0.812 (0.744-0.880)	0.850 (0.793-0.907)	**0.971 (0.952-0.990)**	**0.883**	**0.896**
	Frame (Image)	frame-based	0.920 (0.879-0.961)	0.818 (0.754-0.881)	0.865 (0.814-0.916)	0.959 (0.923-0.994)	0.882	0.892

Table 9: Performance Comparison Across Number of frames, * indicates that the P-value is less than 0.05