Improving Deep Learning-based Diagnosis of Hepatic Tumors on Multi-phase Contrast-enhanced Ultrasound Images: Comparison of Padding Strategies

Taehoon Lee[a], Jaeyeol Lim[b], Nam-Joon Kim[c], Woo Kyoung Jeong[d], Won Jae Lee[e], Kyeonghun Kim[f], Hyuk-Jae Lee[g]

[a]*Department of Civil and Environmental Engineering, Seoul National University, Seoul, Republic of Korea*

[b]*Department of Landscape Architecture and Rural Systems Engineering, Seoul National University, Seoul, Republic of Korea*

[c]*Next-Generation Semiconductor, Seoul National University , Seoul, Republic of Korea*

[d]*Department of Radiology and Center for Imaging Science, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea*

[e]*Department of Radiology, Samsung Medical Center, Sungkyunkwan University Samsung Changwon Hospital, Seoul, Republic of Korea*

[f]*OUTTA, Seoul, Republic of Korea*

[g]*Department of Electrical and Computer Engineering, Seoul National University , Seoul, Republic of Korea*

Abstract

Clinical contrast-enhanced ultrasound (CEUS) data are often acquired at irregular temporal intervals, making automated analysis difficult. This challenge is especially pronounced in liver CEUS due to respiratory motion and variations in acquisition settings, which can degrade image consistency. In this study, we investigated ten different padding strategies to handle temporal irregularity, evaluating their effects on classification performance using two deep learning architectures: a 3D convolutional neural network (3D-CNN) and a convolutional recurrent neural network (CNN-LSTM), under five-fold cross-validation. The 3D-CNN achieved the highest F1 score (0.97) when using phase-level pre-padding with adjacent frames. The CNN-LSTM performed best (F1 score: 0.90) with blank-frame pre-padding. Statistical analysis using the Friedman test revealed that the choice of padding strategy significantly affected model performance in both architectures ($p < 0.01$). These findings highlight that, despite the inherent temporal irregularity in liver CEUS, high classification accuracy can still be achieved when appropriate preprocessing techniques are applied.

Keywords: Contrast-enhanced ultrasound, Liver lesion classification, Temporal padding, 3D-CNN, CNN-LSTM, Deep learning

1. Introduction

Liver cancer remains one of the leading causes of cancer-related death worldwide, with over 830,000 deaths reported in 2020 [1, 2]. Contrast-enhanced ultrasound (CEUS) is widely used in clinical settings for the detection and characterization of malignant liver lesions, such as hepatocellular carcinoma (HCC), due to its ability to visualize perfusion patterns in real time without exposing patients to ionizing radiation. The CEUS LI-RADS categorizes lesions based on features such as the type and degree of arterial phase (AP) enhancement, the presence of washout, and the timing and strength of washout [3]. Automating this classification using AI has become increasingly common in clinical research and applications. In recent years, artificial intelligence (AI) has been increasingly applied to automate CEUS-based lesion classification, with deep learning methods showing notable promise in handling complex imaging data.

Deep learning models for CEUS lesion classification can be broadly grouped into three categories. The first involves the application of attention mechanisms. With the rising popularity of transformer-based models, some studies have adopted temporal attention mechanisms to process sequential CEUS data [4], while others have applied local attention within individual images [5, 6]. However, these models often require large annotated datasets and are difficult to implement in routine clinical workflows, where temporal annotations are limited.

The second category incorporates multimodal approaches, where additional inputs such as B-mode ultrasound images [7, 8] or perfusion-related quantitative features extracted from CEUS videos [9] are used. While such methods enrich the model input, they increase data dimensionality and risk overfitting, especially when training on small clinical datasets. Liver CEUS also presents unique challenges compared to other organs, such as the thyroid, due to complex vasculature and frequent respiratory motion, which complicates the extraction of consistent hemodynamic information.

The third category utilizes relatively shallow architectures, such as 3D convolutional neural networks (3D-CNNs) and convolutional recurrent networks (CNN-LSTMs). These models are less prone to overfitting and are suitable for limited or heterogeneous data. CNN-LSTMs can model long-term temporal dependencies effectively [10, 11], but their performance is highly sensitive to how temporal irregularities in the input are handled. In contrast, 3D-CNNs are generally more robust to noise in sequence structure, as they extract spatiotemporal features in parallel from fixed-length clips. When the number of frames is properly controlled and phase-level alignment is ensured, 3D-CNNs can achieve performance comparable to or even exceeding that of recurrent models.

Accordingly, the method of inputting multiple frames into shallow architecture models plays a crucial role. Typically, two main approaches are used. The first is to extract regions of interest (RoIs) and feeding them into a deep learning model. For example, a fixed-size 3D volume (e.g., $32 \times 32 \times 20$) and a 3D convolutional neural network (3D-CNN) are used to classify lesions such as HCC and focal nodular hyperplasia (FNH) [12], or key frames and specific vessel phases are utilized in a multi-view framework [13, 14]. The second type uses a CNN-LSTM architecture to process the entire CEUS video to capture temporal dynamics between frames, which has demonstrated robust diagnostic performance [15, 16, 17].

However, these methods typically rely on normalizing the video input to a fixed number of frames, which conflicts with real clinical workflows. In a typical clinical setting, ultrasound examinations are performed by an operator who selectively records key moments. A radiologist later interprets these temporally sparse sequences. Given that CEUS studies often span over 8 minutes, storing all frames at full resolution is impractical due to storage and bandwidth constraints. As a result, the number of frames, phase distribution (arterial phase [AP], portal phase [PP], late phase [LP], and Kupffer phase [KP]), and interframe intervals vary significantly across examinations.

These challenges underscore the need for robust preprocessing strategies to normalize temporally irregular CEUS sequences. Such irregularities make it difficult to input data directly into deep learning models, necessitating normalization methods like padding or alignment. While some studies have proposed solutions such as zero-padding [18] or transformer-based models for irregular time-series data [19], these primarily address multivariate clinical signals (e.g., heart rate, blood pressure) and are not optimized for image-based CEUS data with high temporal resolution. Moreover, while CNN-LSTM models have been applied to variable-length ultrasound videos [20], their reliance on uniform compression through zero-padding may result in information loss and reduced capacity to capture fine-grained temporal patterns. To address these issues, there is a clear need for CEUS-specific preprocessing strategies that handle variable frame lengths while preserving clinically meaningful temporal information across different vascular phases.

In this study, we investigate ten padding-based temporal normalization strategies designed to handle irregular CEUS image sequences. Using both 3D-CNN and CNN-LSTM architectures, we systematically evaluate the impact of these strategies on classification performance, providing practical insights for robust CEUS-based deep learning in clinical settings. Our findings provide practical insights for developing robust AI models in liver CEUS, a domain particularly affected by motion artifacts, deep organ location, and signal attenuation. By identifying effective padding strategies, we demonstrate the feasibility of deep learning-based classification in real-world clinical scenarios.

2. Materials and methods

The overall pipeline of the proposed method is described in Figure 1. The process begins with manual annotation of lesion regions in CEUS and B-mode images, from which temporal information is extracted. Based on timestamps, frames are sorted into vascular phases: AP from contrast injection to 60 seconds, PP from 60 to 120 seconds, LP from 120 seconds to 8 minutes, and KP from 8 minutes onward. Ten padding strategies are then applied to normalize the sequence lengths. The resulting sequences are input into deep learning models to classify lesions as either HCC or FNH.

2.1. Description of the dataset

In this study, we introduce a CEUS abdominal dataset comprising 145 patients from Samsung Medical Center. The data acquisition and preprocessing pipeline is illustrated in Figure 2. The dataset consists of two distinct sources: **the CEUS Dataset with AP Videos (hereafter referred to as CEUS-D1)**, derived from AP videos containing additional frames from the PP, LP, and KP phases; and **the CEUS Dataset without Videos (hereafter referred to as CEUS-D2)**, a collection of discrete, non-sequential images covering all CEUS phases. Both datasets were acquired using the Logiq E9 (GE Healthcare, USA), Logiq E10 (GE Healthcare, USA), and Apollo 500 (Canon Medical Systems, Japan) ultrasound systems, with Sonazoid (GE Healthcare, USA) as the contrast agent. All images in the dataset

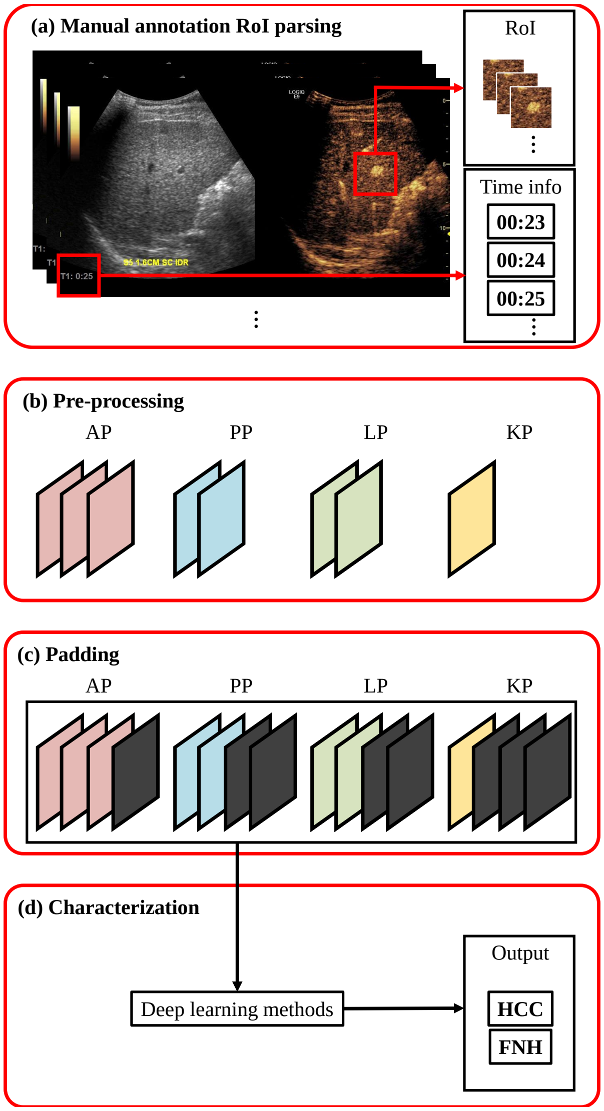

Figure 1: Overview of the proposed CEUS classification pipeline: (a) Lesion regions are cropped from the original ultrasound images, and temporal information is extracted; (b) Image sequences are sorted based on phase information derived from timestamps; (c) Phase-wise padding is applied using the selected padding method; (d) The resulting sequence is input into a deep learning model for classification.

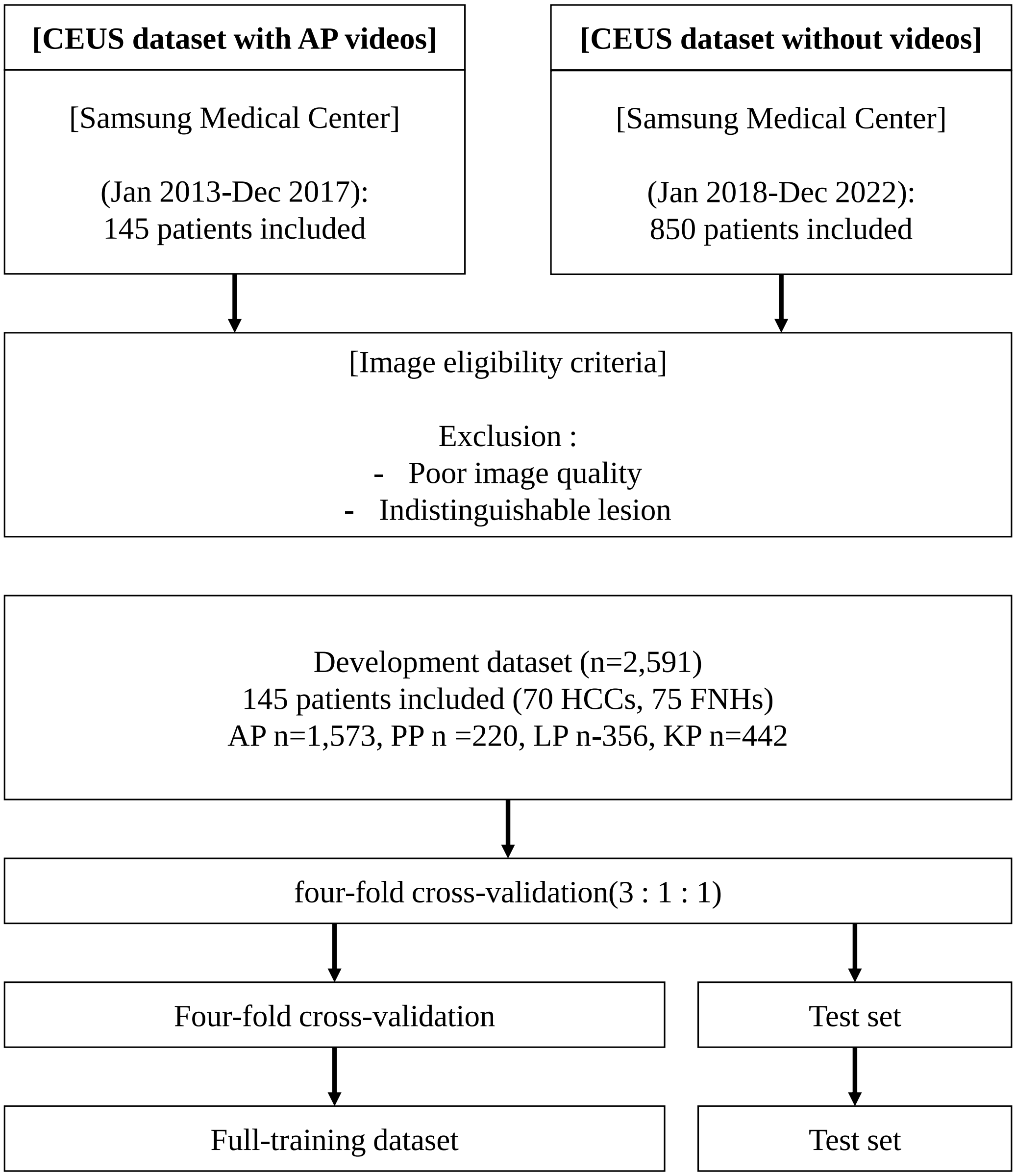
[CEUS dataset with AP videos]
[Samsung Medical Center]
(Jan 2013-Dec 2017):
145 patients included
[CEUS dataset without videos]
[Samsung Medical Center]
(Jan 2018-Dec 2022):
850 patients included
[Image eligibility criteria]
Exclusion :
- Poor image quality
- Indistinguishable lesion
Development dataset (n=2,591)
145 patients included (70 HCCs, 75 FNHs)
AP n=1,573, PP n =220, LP n-356, KP n=442
four-fold cross-validation(3 : 1 : 1)
Four-fold cross-validation
Test set
Full-training dataset
Test set

Figure 2: Overview of patient inclusion and data preparation.

were selected frames deemed clinically important by experienced radiologists. An exception is the AP in CEUS-D1, where images were uniformly sampled from the video starting at the point of initial lesion visibility.

CEUS-D1 was collected between January 2013 and December 2017. It includes 145 patients with liver lesions sized 1 to 3 cm, diagnosed as either HCC or FNH. Among them, 37 patients (33 HCC and 4 FNH) had accessible PP, LP, and KP frames for analysis. **CEUS-D2** was collected between January 2018 and December 2022. Among the 850 patients, 108 patients (37 HCC and 71 FNH) were confirmed with diagnosable HCC or FNH and had clearly distinguishable regions of interest (RoIs). Cases in which lesions were indistinguishable due to pre-enhanced backgrounds (e.g., retaken in KP) or poor image quality were excluded.

From both datasets, we selected a final cohort of 145 patients containing HCC or FNH diagnoses with available AP, PP, LP, and KP phase data (33 HCC and 4 FNH from videos; 37 HCC and 71 FNH from images), resulting in 70 HCC and 75 FNH cases in total. The dataset consists of 2,591 images, distributed as follows: AP – 1,573, PP – 220, LP – 356, and KP – 442. The number of images per patient ranges from 4 to 58. On average, patients have 15.2 AP images, 6.7 PP/LP images, and 4.1 KP images. Specifically, 29 patients lack PP, 7 lack LP, and 42 lack KP images. Additionally, 5 patients are missing both PP and LP, 3 are missing PP, LP, and KP, and 12 lack both PP and KP.

The original CEUS videos and images consist of 800×600-resolution images, including both CEUS and B-mode frames. Two radiologists manually annotated lesions with bounding boxes using microDicom Viewer. Bounding box coordinates were used to crop each lesion with an additional 20-pixel margin on all sides to include surrounding context. The cropped images were resized to 128×128 using nearest-neighbor interpolation for input standardization.

2.2. Frame padding method

To ensure compatibility with deep learning models, the number of frames in each video must be standardized. As illustrated in Figure 3, various padding strategies were designed to match the temporal dimension across samples. These methods rely on either black (blank) frames or replicated neighboring frames. When blank frames are used for padding, the strategies are categorized into four types based on the padding position. When

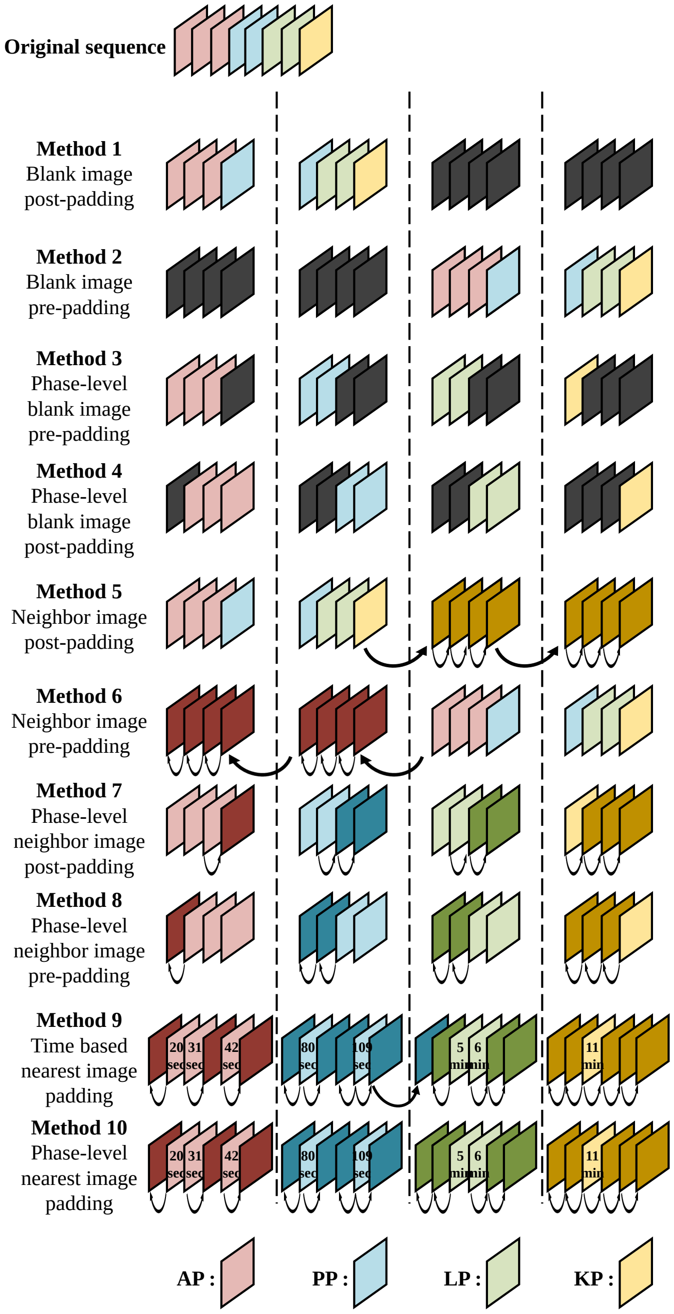

Figure 3: Schematic of the 10 padding strategies.

neighboring frames are duplicated, an additional two time-alignment-based methods are introduced, resulting in a total of six padding strategies.

Blank image post-padding (Method 1) arranges the original frames in chronological order at the beginning and appends black frames at the end to match the required length. Conversely, blank image pre-padding (Method 2) places black frames at the beginning, followed by the original frames. In neighbor image post-padding (Method 5) and pre-padding (Method 6), the last or first valid frame is duplicated instead of using black frames to fill the remaining space.

Given the clinical importance of phase information in CEUS, we further propose phase-level padding strategies. In this approach, videos are segmented by phases (e.g., AP, PP, LP, KP), and padding is applied within each phase. Phase-level blank image post-padding (Method 3) arranges phase-specific frames in chronological order and appends black frames at the end of each phase. Phase-level blank image pre-padding (Method 4) applies the padding at the beginning of each phase. Similarly, phase-level neighbor image post-padding (Method 7) and pre-padding (Method 8) repeat the last or first frame of each phase, respectively, instead of using blank frames.

In addition, time-aware padding strategies are introduced to further leverage temporal information. Time-based nearest image padding (Method 9) divides the video into temporal intervals and fills missing frames with the closest available frame in time. Phase-level nearest image padding (Method 10) extends this approach by considering both the temporal proximity and phase identity, filling the missing frames within each phase using the temporally nearest frame.

2.3. Deep learning architectures

2.3.1. 3D-CNN

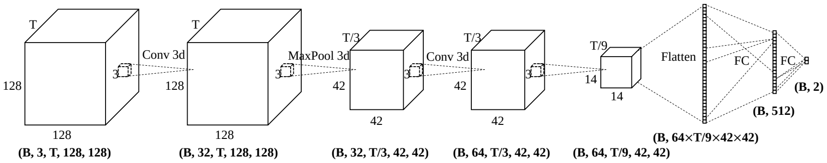

Figure 4: 3D-CNN architecture.

A 3D-CNN was implemented to jointly learn spatiotemporal features. The overall architecture of 3D-CNN is described in Figure 4. The input is a 5D tensor (B, C, T, H, W), where B is the batch size, C is the number of channels, T is the temporal axis, and H and W (128×128) represent the spatial dimensions of each CEUS frame. The architecture consists of two distinct types of 3D convolutional blocks (filters: 32 and 64), each comprising convolution, LeakyReLU, max pooling, and batch normalization. After passing through the convolutional blocks, features are flattened and passed through a 512-dimensional fully connected layer with dropout (0.25), followed by binary classification. This model is suitable for learning short-range spatiotemporal dynamics.

2.3.2. CNN-LSTM

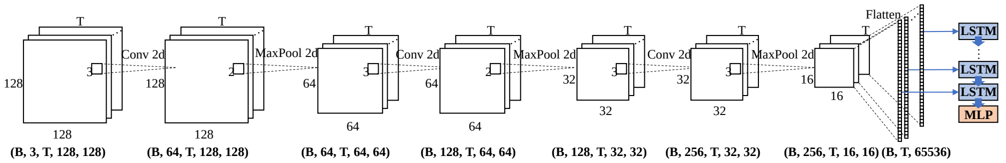

Figure 5: CNN-LSTM architecture.

To decouple spatial and temporal learning, a CNN-LSTM model was implemented. The overall architecture of CNN-LSTM is described in Figure 5. Each frame in the sequence (B, C, T, H, W) is passed through a shared 2D CNN backbone (3 convolutional layers with channels: 64, 128, 256), followed by a 300-dimensional embedding. The embeddings are then fed into a 3-layer LSTM (hidden size: 256), and classification is performed using the final hidden state. Two fully connected layers are used for the final prediction. This structure enables long-term temporal modeling with flexible sequence lengths.

2.4. Experimental setup

All experiments were conducted using an NVIDIA A100 80GB GPU. The dataset was split into 3:1:1 ratio for training, validation, and testing, respectively. We used a batch size of 8 for training. To improve generalization, we applied random shuffling, rotation and horizontal flipping. The models were trained for 300 epochs using the Adam optimizer with a learning rate of 0.0001 and weight decay of 0.0001. To select the optimal model, we monitored the ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) score on the validation set. The ROC-AUC evaluates classification performance by plotting the true positive rate against the false positive rate. A high-performing model has an AUC close to 1, while an AUC near 0.5 indicates random prediction.

Cross-entropy loss was used for training both models. It penalizes the negative log-likelihood of the true class, encouraging the model to assign high probability to the correct label:

$$L_{CE} = -\sum_{c=1}^{C} y_c \log(\hat{y_c}) \tag{1}$$

Here, y_c is the ground-truth label (1 if the sample belongs to class c, otherwise 0), and $\hat{y}^c$ is the predicted probability for class c produced by the softmax output.

To evaluate the performance of the proposed models, we used several standard classification metrics: Accuracy, Recall (Sensitivity), Precision, and F1-score.

Accuracy represents the proportion of correctly classified instances among all samples, and is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Recall measures the proportion of actual positive cases that were correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

Precision reflects the proportion of predicted positive instances that are truly positive:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

F1-score is the harmonic mean of precision and recall, providing a balanced metric:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

Here, TP = True Positives, FP = False Positives, TN = True Negatives, and FN = False Negatives. These metrics provide a comprehensive view of the model's effectiveness, especially in imbalanced classification settings.

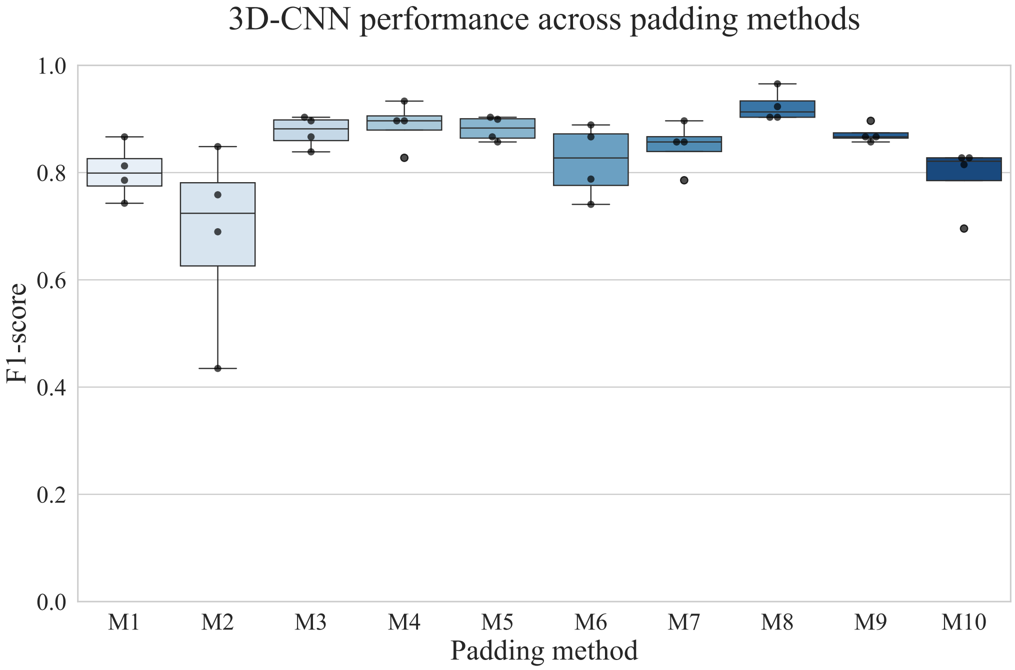

Figure 6: Performance of 3D-CNN across different padding methods using four-fold cross-validation.

No.	Padding method	Loss	AUC	Accuracy	Precision	Recall	F1
1	Pre blank (seq)	0.7577	0.9048	0.7586	0.7333	0.7857	0.7586
2	Post blank (seq)	1.8890	0.8762	0.8276	0.8462	0.7857	0.8148
3	Pre image (phase-aware)	0.3456	0.9619	0.8621	0.8125	0.9286	0.8667
4	Post image (phase-aware)	0.1634	0.9809	0.8966	0.9231	0.8571	0.8889
5	Post blank (seq)	0.4593	0.9333	0.8621	0.8571	0.8571	0.8571
6	Pre blank (seq)	0.2403	0.9714	0.9310	0.9286	0.9286	0.9286
7	Post image (phase-aware)	0.3885	0.9429	0.7931	0.7857	0.7857	0.7857
8	Pre image (phase-aware)	**0.1079**	**0.9905**	**0.9655**	**0.9333**	**1.0000**	**0.9655**
9	Time-based (sequence-level)	0.6564	0.9190	0.8621	0.8571	0.8571	0.8571
10	Time-based (phase-level)	0.8557	0.9190	0.7931	0.7857	0.7857	0.7857

Table 1: Performance of 3D-CNN across different padding methods using the full training dataset. Best performance per metric is highlighted in bold.

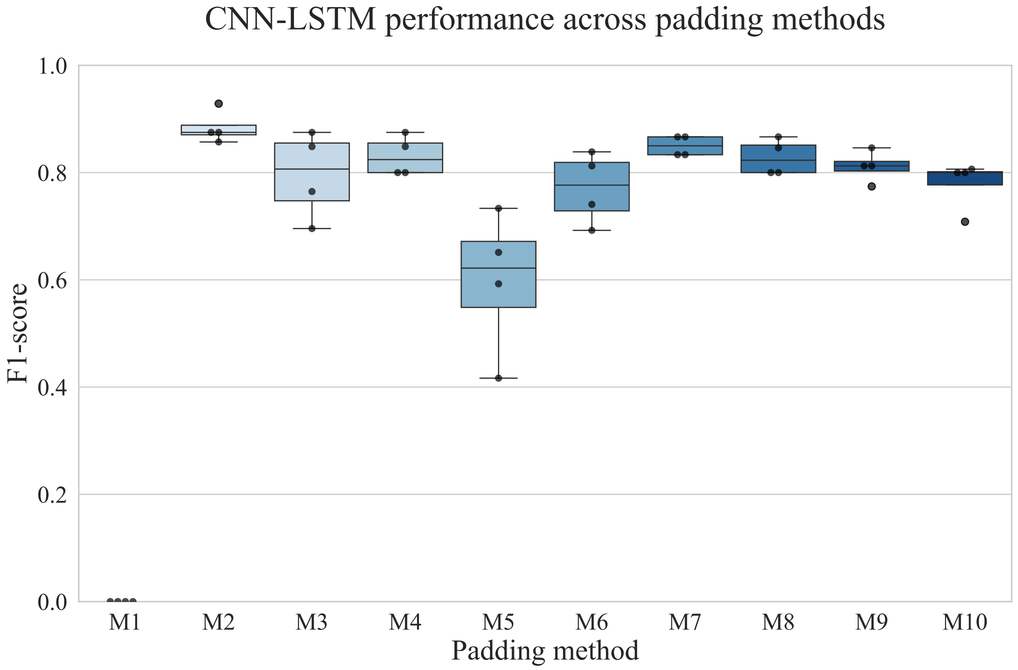

Figure 7: Performance of CNN-LSTM across different padding methods using four-fold cross-validation.

No.	Padding method	Loss	AUC	Accuracy	Precision	Recall	F1
1	Pre blank (seq)	**0.6930**	0.3690	0.5172	0.0000	0.0000	0.0000
2	Post blank (seq)	1.0045	**0.9476**	**0.8966**	0.8235	**1.0000**	**0.9032**
3	Pre image (phase-aware)	1.2262	**0.9476**	0.8621	0.8125	0.9286	0.8667
4	Post image (phase-aware)	1.0618	0.8952	0.8621	0.8125	0.9286	0.8667
5	Post blank (seq)	1.7490	0.4571	0.5517	0.5294	0.6429	0.5806
6	Pre blank (seq)	2.0490	0.9024	0.8276	0.7647	0.9286	0.8387
7	Post image (phase-aware)	0.9998	0.9333	**0.8966**	**0.8667**	0.9286	0.8966
8	Pre image (phase-aware)	0.7469	0.9238	0.8621	0.8571	0.8571	0.8571
9	Time-based (sequence-level)	1.5681	0.9429	0.7586	0.7059	0.8571	0.7742
10	Time-based (phase-level)	2.3319	0.8786	0.7931	0.7353	0.8929	0.8065

Table 2: Performance of CNN-LSTM across different padding methods using the full training dataset. Best performance per metric is highlighted in bold.

3. Results analysis

In this study, we investigated the impact of padding strategies on model performance for CEUS-based liver lesion classification by applying 10 different padding methods to two architectures: a 3D-CNN and a CNN-LSTM model. Both models were evaluated under identical conditions using Accuracy, Precision, Recall, and F1-score as performance metrics. Among these, the F1-score was used as the primary metric, as it effectively reflects the balance between precision and recall, particularly in imbalanced data scenarios. Figures 6 and 7 present the results for each padding method applied to the 3D-CNN and CNN-LSTM models, respectively. This experiment was conducted using four-fold cross-validation. Tables 1 and 2 summarize the experimental results of the four evaluation metrics obtained using the full-training dataset for the 3D-CNN and CNN-LSTM architectures, respectively. In conclusion, the 3D-CNN achieved the best performance with Method 8, while the CNN-LSTM showed superior results with Method 2 in both the cross-validation and full dataset evaluations.

3.1. Comparative analysis of performance by model architecture

The 3D-CNN model achieved an average F1-score of 0.8509, with average values for accuracy, precision, and recall of 0.8552, 0.8463, and 0.8571, respectively. The standard deviation of the F1-score was 0.066, indicating that the 3D-CNN maintained relatively high and consistent performance across all padding methods.

In contrast, the CNN-LSTM model achieved a lower average F1-score of 0.7390, indicating reduced performance compared to the 3D-CNN. The standard deviations for accuracy, precision, and recall were 0.1380, 0.2618, and 0.2955, respectively, showing that the model's performance varied substantially depending on the padding method. This suggests that the CNN-LSTM architecture is more sensitive to the choice of padding, likely due to its structural characteristics. While the 3D-CNN demonstrated stable and consistent performance across different padding methods, the CNN-LSTM model exhibit considerable performance fluctuations depending on the padding strategy applied.

3.2. Comparative analysis of performance by padding method

The 3D-CNN model achieved its highest performance with Method 8 (Phase-level neighbor image pre-padding), recording an F1-score of 0.9655.

In contrast, the CNN-LSTM model performed best with Padding Method 2 (Blank image pre-padding), yielding an F1-score of 0.9032.

To assess whether the choice of padding method had a statistically significant effect on the classification performance of the two deep learning models (3D-CNN and CNN-LSTM), we conducted the Friedman test, a non-parametric statistical test, using the F1-scores obtained from the 10 different padding methods for each model.

For the 3D-CNN model, the Friedman test revealed a statistically significant difference in performance across padding methods (χ^2 = 22.0947, p = 0.0086). Similarly, the CNN-LSTM model also showed a statistically significant variation in F1scores depending on the padding method used (χ^2 = 25.3674, p = 0.0026). These results indicate that both models are influenced by the choice of padding, with the CNN-LSTM model appearing more sensitive to such variations.

The standard deviation of the F1-scores for the 3D-CNN model was 0.06605, while for the CNN-LSTM model, it was considerably higher at 0.27594. This notable difference indicates that the CNN-LSTM model exhibited greater performance variability depending on the padding method. This result is consistent with the qualitative interpretation that the CNN-LSTM architecture is structurally more sensitive to changes in padding strategy, whereas the 3D-CNN tends to produce more stable outcomes across various padding approaches.

4. Discussion

4.1. Reasons for performance differences across padding methods

In the case of the 3D-CNN, Method 8 (Phase-level neighbor image pre-padding) exhibited the most outstanding performance. This result can be attributed to the model's reliance on the absolute temporal positioning of frames. Since 3D-CNNs process spatiotemporal patterns holistically, shifting images toward one direction at the phase level helped the model learn important frames in their correct temporal context. Compared to Method 7, the superior performance of Method 8 may be due to the characteristics of the AP phase. AP phase's early appearance of contrast enhancement is critical. Padding with earlier frames may have better preserved this initial contrast progression. therefore, Method 8 enhanced the model's ability to detect AP enhancement pattern.

In contrast, the CNN-LSTM model showed the best performance with Method 2 (Blank image pre-padding), which shifts all frames to the back. This can be explained by the nature of LSTM architectures, which are sensitive to redundant sequential input. Repeated frames or padding with similar images can lead to hidden state saturation or information loss due to the vanishing gradient problem. Hence, padding with blank images at the front, followed by unique frames at the back, provides a cleaner sequence for the LSTM to process. Methods such as 1 and 5, where padding is placed at the end, resulted in performance degradation because the model progressively lost meaningful information it learned from earlier frames. In contrast, Methods 2 and 6 involve padding at the front. The placement may have mitigated the loss of meaningful information during sequence processing, resulting in better performance than methods with rear padding. Overall, these findings suggest that front-loaded zero padding, which delays meaningful input until later in the sequence, is preferable to rear padding that risks information degradation. This supports the idea that introducing blank padding early in the sequence may be less harmful than introducing noise throughout via redundant or nonzero padding.

4.2. Reasons for performance differences between models

The 3D-CNN generally outperformed the CNN-LSTM. For the majority of the padding strategies, the 3D-CNN demonstrated consistently strong results. This can be attributed to the model's lower sensitivity to temporal order and its emphasis on learning overall spatiotemporal patterns. When frames are roughly aligned to phase-level positions, the 3D-CNN can effectively extract meaningful features regardless of the exact temporal sequence.

Moreover, the position of padding within the sequence—either at the beginning or the end—has minimal influence on the 3D-CNN's behavior. In contrast, LSTM-based architectures are more vulnerable to non-informative or irrelevant inputs at the end of the sequence. It can corrupt the learning process by introducing noise or causing information loss through hidden state saturation. This explains why the CNN-LSTM model achieved its lowest performance with Method 1, where zero-padding was placed at the end of the sequence.

Furthermore, as the input sequence length increases, LSTM models are prone to the vanishing gradient problem, which can hinder their ability to learn long-term dependencies. This accounts for the poor performance observed with Methods 9 and 10, which involve longer sequences due to their padding strategy.

These results suggest that while 3D-CNNs are relatively robust to different padding schemes due to their architectural characteristics, CNN-LSTM models require careful handling of padding strategy to maintain sequence integrity and prevent performance degradation.

4.3. Misclassification cases in 3D-CNN using Method 8

The 3D-CNN model extracts features based on the absolute temporal position of each frame. When a lesion's enhancement pattern deviates from the expected sequence—particularly during key diagnostic phases—the model may fail to correctly classify the lesion. Specifically, if early enhancement patterns are missing in the initial frames, or if wash-out is not apparent during the later phases, misclassification can occur.

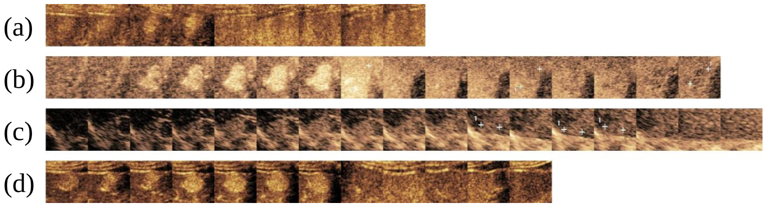

Figure 8: Misclassified FNH cases in the 3D-CNN.

Figure 8 presents cases where FNH was incorrectly classified as HCC. In cases (a), (b), and (d), typical early enhancement features such as centrifugal hyper-enhancement or spoke-wheel patterns were not observed, despite the ground truth being FNH. Notably, in case (d)—which was misclassified as HCC across all folds—there is visible contrast inflow at the lesion boundary in the first frame. This pattern may have been mistakenly interpreted by the model as the early arterial enhancement characteristic of HCC.

Additionally, the lesion area appears slightly darker in the LP and KP in multiple instances (a, b, c, d), mimicking hypoenhancement or washout—a hallmark feature of HCC. This emphasizes the diagnostic relevance of lesion behavior in both LP and KP, where the presence or absence of Kupffer cells can aid in distinguishing FNH from HCC lesions [21]. Therefore, such brightness variations likely contributed to misclassification. Notably, subtle hypoenhancement may also be attributed to central scar tissue in FNH, potentially confounding the 3D-CNN's decision-making [22]. Occasionally, lesions may also appear darker due to microbubble destruction. When a lesion is located close to the ultrasound transducer, the intensity of the ultrasound beam can be amplified by reverberation, causing Sonazoid microbubbles to collapse. Since CEUS relies on microbubble resonance to generate strong echo signals, their destruction can lead to localized signal loss that may mimic hypoenhancement or washout.

in particular, case (c) features a lesion with indistinct boundaries and subtle brightness variation. The lesion is located deep within the liver and is situated near the diaphragm, which introduces significant echo contrast. The depth increases the likelihood of image noise. These environmental complexities may have limited the model's ability to extract clear features.

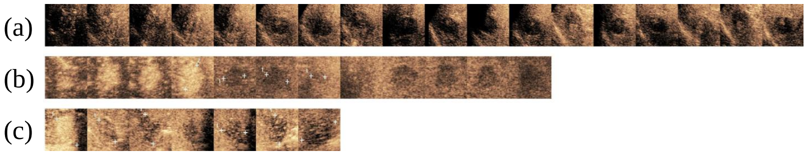

Figure 9: Misclassified HCC cases in the 3D-CNN.

Figure 9 lists all cases in which HCC was misclassified as FNH. These cases had an average of 3.33 images in the AP, which was insufficient to clearly capture the early enhancement pattern characteristic of HCC.

The enhancement pattern of HCC reflects the formation of unpaired arteries during tumor development, which increases arterial supply and results in AP hyperenhancement. However, early-stage or poorly differentiated HCCs may lack this typical appearance, potentially reducing diagnostic accuracy [21]. In case (a), the lesion is poorly distinguished during the AP phase, likely for this reason. And in case (c), the enhancement pattern of HCC is not identifiable from the single available AP image.

4.4. Misclassification cases in CNN-LSTM using Method 2

The CNN-LSTM model encodes information sequentially, determining which temporal features to retain based on frame order. With Method 2, all image frames are shifted toward the end of the sequence via blank image pre-padding, making the final frames—especially those capturing washout—critical for accurate classification.

Figure 10 shows examples where FNH was misclassified as HCC using the CNN-LSTM model. In most of these cases, the final frame exhibits a noticeable darkening of the lesion region, particularly in the KP. Cases (b),

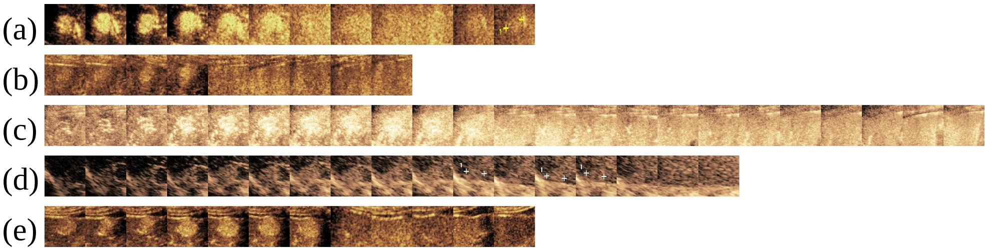

Figure 10: Misclassified FNH cases in the CNN-LSTM.

(c), (d), and (e) all demonstrate this subtle hypointensity in the central lesion area, which the model likely misinterpreted as wash-out. Notably, cases (b), (d), and (e) were also misclassified by the 3D-CNN model, suggesting that this visual cue is a common source of error across architectures.

Furthermore, in cases (a), (b), (d), and (e), the early arterial features of FNH—such as centrifugal enhancement or the spoke-wheel pattern—were absent. In (a) and (d), early enhancement at the lesion boundary is visible in the first frame, which may have been wrongly identified as HCC's AP enhancement. In particular, case (a) features a vessel adjacent to the right side of the lesion, which may have been interpreted as arterial enhancement of HCC extending from the periphery.

Case (e) features blurry lesion margins and minimal contrast variation, potentially preventing the model from recognizing meaningful patterns.

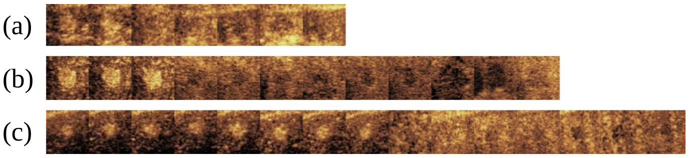

Figure 11: Misclassified HCC cases in the CNN-LSTM.

Figure 11 highlights cases where HCC was incorrectly classified as FNH. In cases (a) and (b), the number of AP phase images is 2 and 3, respectively, making it difficult for the model to capture the initial enhancement typical of HCC. In case (c), although eight AP images were available, the lesion shows a gradual increase in size and enhancement from the center outward—more similar to the typical pattern of FNH.

Additionally, misalignment between initial enhancement and wash-out may have disrupted temporal consistency, resulting in misclassification. In case (c), the lesion is small, which likely amplified the effect.

In case (b), the final frame appears abruptly brighter, likely because the preceding frames were darkened by obstruction of the ultrasound beam due to ribs or other anatomical structures.

5. Conclusion

This study presents a data preprocessing framework tailored to irregularly stored CEUS images commonly encountered in clinical practice. Using two deep learning models, 3D-convolutional neural network (3D-CNN) and CNN-LSTM, we systematically evaluated 10 padding strategies and found that combining phase-level pre-padding using adjacent frames with 3D-CNN achieved the highest classification performance, with an F1score of 0.97.

In particular, this result is highly relevant to real-world clinical environments where CEUS images are often recorded as temporally unbalanced static sequences without standardized frame alignment. Our results demonstrate that using appropriate phase-aware padding and an architecture optimized for spatiotemporal feature extraction enables accurate lesion classification even under these non-ideal conditions.

Moreover, the experimental results indicate that the 3D-CNN model consistently outperformed the CNN-LSTM model across multiple padding strategies, both in terms of accuracy and stability. Specifically, the 3D-CNN achieved high and stable F1-scores (mean: 0.85, std: 0.066), showing minimal variation across folds. In contrast, the CNN-LSTM model yielded lower and more variable performance (mean: 0.74, std: 0.28), with notable drops in F1-score when using less informative padding strategies. This indicates that while the 3D-CNN is more robust to variations in temporal input structure, the CNN-LSTM is significantly more sensitive to sequence continuity and padding-induced artifacts—likely due to its recurrent nature and strong dependence on temporal consistency.

Overall, our results highlight the critical role of architecture selection and preprocessing strategy in CEUS-based deep learning. The proposed approach provides a practical and effective foundation for future diagnostic applications, especially in heterogeneous or resource-constrained CEUS data collection environments.

References

[1] H. Sung, J. Ferlay, R. L. Siegel, et al., Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J Clin 71 (3) (2021) 209–249. doi:10.3322/caac.21660.

[2] R. L. Siegel, K. D. Miller, H. E. Fuchs, A. Jemal, Cancer statistics, 2021, CA Cancer J Clin 71 (1) (2021) 7–33. doi:10.3322/caac.21654.

[3] S. R. Wilson, et al., Ceus li-rads: algorithm, implementation, and key differences from ct/mri, Abdom Radiol 43 (2018) 127–142.

[4] F. Chen, et al., Joint segmentation and differential diagnosis of thyroid nodule in contrast-enhanced ultrasound images, IEEE Trans Biomed Eng (2023).

[5] P. Wan, et al., Ceus-net: Lesion segmentation in dynamic contrast-enhanced ultrasound with feature-reweighted attention mechanism, in: 2020 IEEE Int Symp on Biomedical Imaging (ISBI), IEEE, 2020, pp. 1629–1632.

[6] P. Wan, et al., Hierarchical temporal attention network for thyroid nodule recognition using dynamic ceus imaging, IEEE Trans Med Imaging 40 (6) (2021) 1646–1660.

[7] H. Dadoun, et al., Deep learning for the detection, localization, and characterization of focal liver lesions on abdominal us images, Radiol Artif Intell 4 (3) (2022).

[8] D. Mitrea, R. Badea, P. Mitrea, S. Brad, S. Nedevschi, Hepatocellular carcinoma automatic diagnosis within ceus and b-mode ultrasound images using advanced machine learning methods, Sensors 21 (2021).

[9] H. Yu, et al., Ln-net: Perfusion pattern-guided deep learning for lymph node metastasis diagnosis based on contrast-enhanced ultrasound videos, Ultrasound Med Biol 49 (5) (2023) 1248–1258.

[10] J. Mänttäri, T. Raiko, T. Salakoski, Interpreting video features: A comparison of 3d convolutional networks and convolutional lstm networks, in: Asian Conf on Computer Vision (ACCV), 2020, pp. 101–117.

[11] Q. Zhou, J. Li, K. Wang, et al., Cnn-lstm model for recognizing video-recorded actions in rehabilitation, J Biomed ResPMC10332470 (2023).

[12] F. Pan, Q. Huang, X. Li, Classification of liver tumors with ceus based on 3d-cnn, in: 2019 IEEE Int Conf on Advanced Robotics and Mechatronics (ICARM), IEEE, 2019, pp. 845–849.

[13] H. Zhou, J. Ding, Y. Zhou, et al., Malignancy diagnosis of liver lesion in contrast-enhanced ultrasound using an end-to-end method based on deep learning, BMC Med Imaging 24 (2024) 68.

[14] X. Feng, et al., Diagnosis of hepatocellular carcinoma using deep network with multi-view enhanced patterns mined in contrast-enhanced ultrasound data, Eng Appl Artif Intell 118 (2023) 105635.

[15] K. Schmiedt, G. Simion, C. D. Căleanu, Preliminary results on contrast enhanced ultrasound video stream diagnosis using deep neural architectures, in: 2022 Int Symp on Electronics and Telecommunications (ISETC), IEEE, 2022.

[16] N. J. Kim, W. J. Lee, H. J. Lee, Deep learning classification of focal liver lesions with contrast-enhanced ultrasound from arterial phase recordings, in: 2023 Int Conf on Electronics, Information, and Communication (ICEIC), IEEE, 2023.

[17] H. Zhou, et al., Malignancy diagnosis using cnn-lstm with multi-phase ceus, BMC Med Imaging 24 (1) (2024) 68.

[18] S. Nzamba Bignoumba, S. Ben Yahia, N. Mellouli, Deep padding and alignment strategies for irregular multivariate clinical time series, Procedia Comput Sci 233 (2024) 1124–1131.

[19] Y. Song, Y. Zhang, W. Li, et al., Trajgpt: Irregular time-series representation learning for health data, `https://arxiv.org/abs/2410.02133`, arXiv preprint arXiv:2410.02133 (2024).

[20] X. Cui, et al., Variable-frame cnn-lstm for breast nodule classification using ultrasound videos, `https://arxiv.org/abs/2502.11481`, arXiv preprint arXiv:2502.11481 (2025).

[21] Z. Huang, X. J. Lin, S. S. Li, H. C. Luo, K. Y. Li, Differentiating atypical hepatocellular carcinoma from focal nodular hyperplasia: The value of kupffer phase imaging with sonazoid-contrast-enhanced ultrasound compared to gadodiamide-enhanced mri, Eur J Radiol 184 (2025) 111991. doi:10.1016/j.ejrad.2025.111991.

[22] M. He, L. Zhu, M. Huang, L. Zhong, Z. Ye, T. Jiang, Comparison between sonovue and sonazoid contrast-enhanced ultrasound in characterization of focal nodular hyperplasia smaller than 3cm, J Ultrasound Med 40 (10) (2021) 2095–2104. doi:10.1002/jum.15589.