Computers in Biology and Medicine 180 (2024) 109014

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

FLUEnT: Transformer for detecting lung consolidations in videos using fused lung ultrasound encodings

Umair Khan [a], Russell Thompson [b], Jason Li [c], Lauren P. Etter [d], Ingrid Camelo [e], Rachel C. Pieciak [f], Ilse Castro-Aragon [g], Bindu Setty [g], Christopher C. Gill [f], Libertario Demi [a,*], Margrit Betke [c]

[a] Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
[b] University of Massachusetts Dartmouth, MA, USA
[c] Department of Computer Science, Boston University, Boston, MA, USA
[d] University of Wisconsin Madison, Madison, WI, USA
[e] Augusta University, Pediatric Infectious Disease, Augusta, GA, USA
[f] Department of Global Health, Boston University School of Public Health, Boston, MA, USA
[g] Department of Radiology, Boston Medical Center, Boston, MA, USA

ARTICLE INFO

Keywords:
Pediatric lung ultrasound
Lung consolidation
Pneumonia
Fused encodings
Transformers

ABSTRACT

Pneumonia is the leading cause of death among children around the world. According to WHO, a total of 740,180 lives under the age of five were lost due to pneumonia in 2019. Lung ultrasound (LUS) has been shown to be particularly useful for supporting the diagnosis of pneumonia in children and reducing mortality in resource-limited settings. The wide application of point-of-care ultrasound at the bedside is limited mainly due to a lack of training for data acquisition and interpretation. Artificial Intelligence can serve as a potential tool to automate and improve the LUS data interpretation process, which mainly involves analysis of hyper-echoic horizontal and vertical artifacts, and hypo-echoic small to large consolidations. This paper presents, Fused Lung Ultrasound Encoding-based Transformer (FLUEnT), a novel pediatric LUS video scoring framework for detecting lung consolidations using fused LUS encodings. Frame-level embeddings from a variational autoencoder, features from a spatially attentive ResNet-18, and encoded patient information as metadata combiningly form the fused encodings. These encodings are then passed on to the transformer for binary classification of the presence or absence of consolidations in the video. The video-level analysis using fused encodings resulted in a mean balanced accuracy of 89.3 %, giving an average improvement of 4.7 % points in comparison to when using these encodings individually. In conclusion, outperforming the state-of-the-art models by an average margin of 8 % points, our proposed FLUEnT framework serves as a benchmark for detecting lung consolidations in LUS videos from pediatric pneumonia patients.

1. Introduction

Pediatric lung diseases are a common clinical problem around the world [1]. In 2019, a total of 740,180 lives under the age of five were claimed by pneumonia, making it the leading cause of death in children around the world [2]. To reduce the complications and mortality rate due to pneumonia, in-time diagnosis and assessment of disease severity are important. The significance of this increases in resource-limited settings such as remote rural healthcare facilities. Diagnosing and treating children with pneumonia are primarily based on the World Health Organization criteria [3], which do not involve imaging. However, several studies indicate that the lack of sensitivity and specificity in the criteria may lead to misdiagnosis and over-treatment [3–5]. The diagnostic process can be aided by a chest X-ray (CXR). Since resource-limited settings often lack radiographic imaging facilities, and lung ultrasound (LUS) is reported to be more sensitive than CXR for pneumonia screening among pediatric patients [6], LUS can be considered as an inexpensive alternative imaging modality.

LUS is an effective imaging tool to assess patient conditions, including in resource-limited settings, as has been shown by several

* Corresponding author.
E-mail address: libertario.demi@unitn.it (L. Demi).

https://doi.org/10.1016/j.compbiomed.2024.109014
Received 8 February 2024; Received in revised form 6 August 2024; Accepted 9 August 2024
Available online 19 August 2024

studies [7–12]. The primary motivation of our work is to support LUS interpretation in resource-limited settings with an AI tool that is sufficiently adapted to pediatric patient populations. To this end, we worked with a LUS dataset from a resource-limited hospital in Lusaka, Zambia. Another motivation for our work is the use of such an AI tool in environments that are not necessarily resource-limited but that use point-of-care ultrasound (POCUS) at the bedside or in emergency settings. Studies have shown that POCUS can reduce mortality and healthcare costs [13]. Its application in the United States, however, is limited mainly due to lack of training [14]. An effective POCUS framework requires a clinician able to acquire and interpret POCUS data. Not all residents are trained to read POCUS during an internal medicine or family medicine residency [14]. Artificial intelligence may increase the effectiveness of POCUS, by reducing the impact of poor confidence in data interpretation [15].

Interpreting LUS primarily includes analysis of hyper-echoic horizontal and vertical artifacts, and hypo-echoic small to large consolidations. Various AI-based methods have been developed to evaluate these patterns using a frame-wise analysis, with the majority of the work focusing on COVID-19-induced pneumonia in adults [16,17]. Extending to the task of video-level analysis, studies [18–20] proposed aggregation of CNN-based frame-level evaluations [16,19] to compute the video-level score. In particular, Roy et al. proposed a frame-level classification model based on a regularized spatial transformer network (STN) and a CNN (Reg-STN + CNN) to assess the severity of a lung abnormality due to COVID-19 induced pneumonia [16]. Predicted frame-level scores are then mapped and aggregated to a video-level score using a uninorm based function [16]. Utilizing the frame-level scores given by the DL model [16], the score at the video level is computed using aggregation methods based on threshold approach [18], decision tree-based classification of frequency-based [20] and cross-correlation-based [19] features of frame-level scores. These methods predict frame-level scores as an intermediary step for video-level analysis, potentially propagating classification errors and adding computational overhead to evaluating the Lung Ultrasound (LUS) score at the video level.

Expanding the application of ConvNets from 2D to 3D, Ebadi et al. proposed a two-stream Inflated 3D ConvNet (I3D) for the task of classifying LUS videos with pneumonia obtained using POCUS [21]. Similarly, 3D CNN-based methods are proposed to detect and localize vertical artifacts in LUS videos [22,23]. These methods allow to seamlessly learn spatio-temporal features, however, yield high computational costs with excessive memory usage. Moving towards the hybrid use of CNNs and RNNs, Barros et al. proposed a CNN-LSTM approach to classify spatio-temporal features of LUS videos to identify COVID-19-induced abnormalities [24]. Kerdegari et al. further extended it using temporal analysis networks with attention to automatically detect and localize vertical artifacts in ultrasound videos using deep neural networks trained in a weakly supervised manner [25]. These methods show promising results while relying on data-hungry and computationally expensive use of LSTMs. To this extent, Lum et al. proposed an attention-based video model for the identification of COVID-19 in lung ultrasound videos. The model used a 2D-CNN to extract frame-level spatial features and a transformer encoder-based network for temporal features [26]. Additional temporal information was extracted using a 1D-CNN layer over the neighboring features, which were further aggregated using pooling layers.

The above-mentioned techniques exhibited evaluation of LUS videos primarily from the adult COVID-19 patient population. Although some of the methods have been developed for frame-level classification of pediatric LUS data [27,28], to the best of the authors' knowledge, methods that evaluate an entire video directly have not been described in the literature yet. With limited research done in the field of pediatric LUS video data evaluation and considering the shortcomings of the existing methods for LUS video evaluation, the objective of this study is to propose a first-of-its-kind, computationally efficient pediatric LUS video classification model for detecting the presence or absence of consolidations due to pneumonia.

CNN-based models have proven to be powerful tools in the health care domain [29]. Diverse layers extending to large depths have enabled CNNs to extract deep features without human intervention. To further improve the identification of regions of interest and their extraction of features, spatial attention (SA) has been thoroughly studied in the recent literature, specifically in the medical domain [30]. To this extent, frame-level features as embeddings generated using a spatially-attentive convolutional neural network (SA-CNN) and a variational autoencoder (VAE), along with patient information encoded as metadata, can be fused together in a single encoding per frame. This work proposes a transformer-based pediatric LUS video scoring framework using fused lung ultrasound encodings (FLUEnT). The success of transformer-based models has been largely due to the self-attention mechanism, which allows the model to capture long-range relationships between the elements of the sequence [31]. In this regard, an integrated multi-head self-attention module is used that allows the framework to learn a video-level encoding representation that is then used to train a classifier, specifically a multi-layer perceptron (MLP), for detecting consolidations. In comparison to the traditional video vision transformer architectures [32–34], instead of taking images or videos as an input and encoding them into different patches, each image is treated as a single patch, encoded using convolutional neural network-based models. As the experiments show, the FLUEnT approach yields scalable and robust learning to detect consolidations in LUS videos of pediatric patients. The core contributions are as follows.

- This study presents FLUEnT, a transformer-based video scoring framework for detecting consolidations in lung ultrasound videos of pediatric patients.
- In contrast to vision transformers, which divide images into encoded patches, FLUEnT considers each frame of the video as a patch and encodes it using VAE and SA-CNN models.
- As a key part of FLUEnT, the study proposes a method of mid-level fusion of encodings of LUS video data.
- It also analyzes the significance of each encoding separately and its impact in a fused state within FLUEnT.
- To measure the effectiveness of FLUEnT, it is tested over a curated pediatric lung ultrasound dataset (the only available benchmark dataset) and showed that FLUEnT outperformed existing baseline video analysis models.
- The study also evaluates FLUEnT as a screening tool to assist clinicians for identifying patients with consolidation findings.

Based on the new international guidelines on LUS [35] for the application of artificial intelligence, our proposed method is a step towards the development of an AI-driven tool to assist clinicians in LUS data analysis.

Code will be available with the publication of the paper.

2. Methods

This section begins with a description of the pediatric dataset used in our study and follow with a block-wise description of the proposed Fused LUS Encoding-based Transformer (FLUEnT) framework. The overview in Fig. 1 illustrates the LUS collection process, sample LUS frames, and the three model blocks VAE, SA-CNN, and metadata processing that collectively yield fused embeddings that are interpreted by the fourth block, the video transformer.

2.1. Pediatric LUS dataset

This study used a LUS dataset consisting of 2400 videos from 200 pediatric patients (1–59 months old) [36]. These patients were enrolled from May 2020 to March 2021. Guardian informed consent was

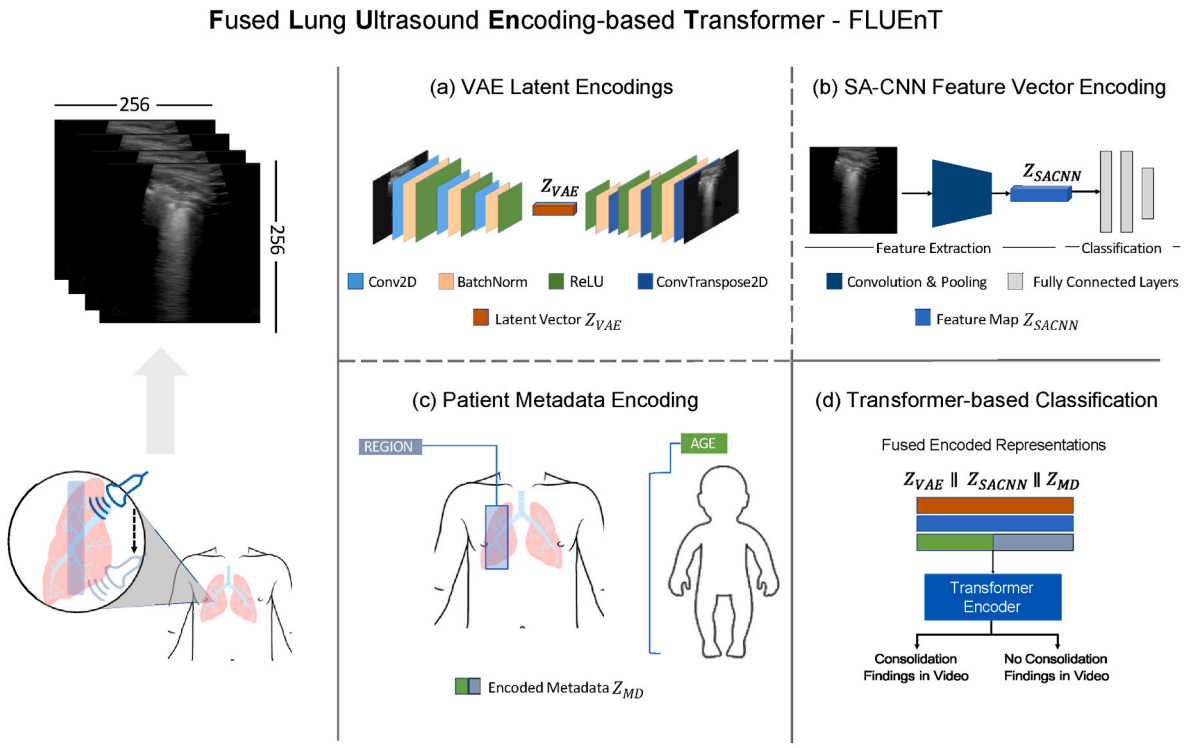

Fig. 1. Overview of the proposed pediatric LUS video scoring framework (FLUEnT). On the left, the data collection process and a few sample LUS frames are shown. Block (a) represents the VAE embeddings of an input video frame, block (b) represents the feature maps extracted using an SA-CNN, block (c) shows the encoding of the patient's information, and block (d) represents the transformer-based classification of fused encodings from blocks (a–c).

obtained for all patients. All of the patients showed symptoms in accordance with the WHO criteria for pneumonia (cough, difficulty breathing, plus lower chest indrawing) or severe pneumonia (pneumonia with lower chest indrawing plus any danger sign: central cyanosis, difficulty breastfeeding or drinking, vomiting, convulsions, lethargy, unconsciousness, or head nodding). For each patient, a total of twelve videos (left/right, anterior/posterior/lateral, sagittal/transverse) were collected using a handheld ultrasound probe (Butterfly Network, Inc., Burlington, MA, USA) in a sweep acquisition manner. It involves moving/sweeping the probe from the top to the bottom of the lung, as part of the protocol [36]. The data were acquired by a sonographer with over 6 years of formal training in ultrasound and intensive 2-week training to use the POCUS device. Acquisitions were performed according to a fixed protocol [36]. Based on this all patients were scanned using the pediatric lung preset with the focal point as a function of half the imaging depth, a factory setting in the POCUS probe [36]. As a result, in most of the cases, the imaging depth was 7 cm with a focal point at a depth of 3.5 cm, mechanical index (MI) set at 0.47, and thermal index (TIS) at 0.01. In the initial step of annotation [36], the two radiologists, with 14 and 18 years of experience, respectively, blindedly reviewed and annotated the LUS sweep acquisitions for consolidation findings. Differences in the interpretations of consolidation patterns from the initial step were discussed by the radiologists until they reached a consensus at the sweep level. This involved both radiologists convening and reviewing together only sweeps where disagreements occurred, to ultimately reach an agreement on the observed findings. In this way, all of the videos were annotated with the same label by the two radiologists. Consequently, of the 2400 collected videos, 178 videos had consolidation findings, according to the consensus of the two radiologists. The absence or presence of the appearance of consolidations in a video is referred to as a Score 0 or Score 1 label, respectively. The Score 0 label is used for videos having LUS frames showing fully aerated lungs with horizontal artifacts.

Score 0 videos may also show the first indications of an abnormality appearing as a non-continuous pleural line with vertical artifacts, as seen in the first two image samples from the left in Fig. 2. The initial focus of the study was on identifying cases with severe lung involvement by detecting the presence/absence of consolidations within the LUS videos. As a result, videos indicating abnormalities such as non-continuous pleural line along with vertical artifacts without consolidations were considered as negative cases. Score 1 is used for videos in which vertical artifacts may appear with small to large consolidations. These include small subpleural and segmental/lobar consolidations. Table 1 provides a concise summary of the scoring definitions. Sample frames with small subpleural and segmental/lobar consolidations are also shown in Fig. 2. The dataset videos range in length from 83 to 581 frames. For a video with consolidation findings, consolidations appear, on average, in 26.4 (±21.1) frames, i.e., 12.4 % (±9.9 %) of the frames per video. The ground truth provided by the radiologists is based on the interpretation of lung consolidation as anatomical findings and one of the most common sonographic signs of pneumonia. To this extent [35], served as the gold standard for analyzing LUS data for consolidations. The study was approved by the Boston University Medical Center IRB (H-39058). It was also approved locally by the University of Zambia Biomedical Research Ethics Committee (UNZABREC) (Ref No 310–2019). The study and data acquisition were in compliance to the WMA Helsinki declaration. The dataset along with the annotations can be made available on request.

2.2. *Types of encodings used by FLUEnT*

FLUEnt uses two types of encodings, described in detail below, to condense and project relevant image-level features onto a space-efficient representation such that a minimal sequential transformer can accurately classify them. Both encoding methods process LUS video frames and pass the encoded form into the transformer of FLUEnT as an input. FLUEnT uses a third type of encoding to represent metadata, which includes, for example, patient age and probe orientation.

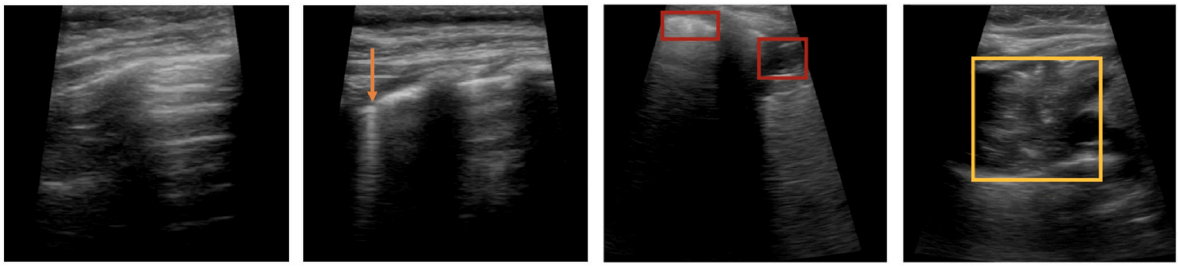

Fig. 2. Left to Right: Lung ultrasound image samples without consolidations with a continuous pleural line; with the non-continuous pleural line shown with orange arrow; with small subpleural consolidations (in red box) and segmental/lobar consolidations (in the yellow box).

Table 1
Scoring definition.

Score	Patterns
0	Video frames showing continuous pleural line with horizontal artifacts or non-continuous pleural line with vertical artifacts
1	Video frames with/without vertical artifacts along with consolidations

2.2.1. VAE latent encoding

VAEs are generative models that are able to learn complex distributions

$$p(x) = \int p(x|z)p(z)dz \tag{1}$$

over images x [37] by using a latent variable z with a prior distribution $p(z)$ (usually having a fixed normal unit distribution).

The VAE architecture FLUEnT uses is based on a Convolutional Variational Autoencoder. It focuses on the transformation of input data x into the probability distribution parameters mean μ and standard deviation σ of a Gaussian distribution. As a result, a continuous structured latent space is generated that can be used for classification purposes. The VAE comprises of the *Encoder, Bottleneck,* and *Decoder* components:

The ***Encoder*** is a convolutional neural network taking an RGB image of size 256 × 256 as input (as shown at the top left of Fig. 1). It defines the approximate posterior distribution $q(z|x)$, where x is the input image and z is the latent variable. With the distribution expressed as a Gaussian with a diagonal covariance matrix, the network generates μ and σ as its parameters. The encoder architecture consists of five 2D convolutional layers, each of which is followed by a batch normalization layer. Filters compute convolutions with a stride of 2 in the first three convolutional layers to reduce the dimensions of the feature maps. A stride of 1 is used for the remaining two. A Rectified Linear unit (ReLU) is used as the activation function in each convolutional layer. This results in the probability distribution vectors μ and σ of size 26 × 26, flattened to 1 × 676.

The ***Bottleneck***, referred to as the latent space representation, generates sample for the decoder by sampling from the latent distribution defined by the encoder's parameters μ and σ. A reparameterization technique is used to approximate the latent variable z. The technique samples from a standard normal distribution noise ϵ and then transforms these samples into samples from the desired distribution using encoder parameters μ and σ:

$$z = \mu + \sigma \odot \epsilon, \tag{2}$$

where μ and σ denote the mean and standard deviation of a Gaussian distribution, whereas ϵ can be conceived as random noise that is used to keep the latent space stochastic. As a result, the latent variable is generated as a function of μ, σ, and ϵ. With the sizes of μ and σ as 1 × 676, the sampled latent vector z is also of the size 1 × 676 and represents the size of the bottleneck layer. The sampled latent vector z, referred to as Z_{VAE} for each LUS video frame, represents the VAE latent encoding.

The ***Decoder*** of the VAE is a generative network. It functions by taking the latent vector z as input and returns the parameters for the conditional distribution $p(x|z)$ of the sample x as output. As a mirror architecture of the encoder network, the decoder consists of five 2D transposed convolutional layers, followed by batch normalization layers. To reconstruct the data up to the shape of the original input, a stride of 2 was used in transposed convolutional layers. A ReLU is used as an activation function in each convolutional layer.

2.2.2. SA-CNN feature vector encoding

The second building block of FLUEnt is a CNN with spatial attention (SA). This work uses a ResNet-18 model with the addition of spatial attention modules to exploit the spatial relationship between pixels. The architecture of the model is shown in Fig. 3. The spatial attention modules and their integration within each residual block in the ResNet architecture were inspired by the convolutional block attention module [38]. For the rest of our manuscript, ResNet-18 with spatial attention and SA-CNN will be used interchangeably. Given a LUS video frame as input, the SA-CNN extracts features as output of the final convolutional layer after passing the data through a series of convolutional layers, pooling layers, spatial attention blocks, and activation functions. The resulting features are then passed through fully connected layers for classification. A deep feature map for an input image x extracted through the last convolution process within the network is given as

$$C = f(x * W + b), \tag{3}$$

where "*" represents the convolution operation resulting in a feature map C of the image x. The feature map is a function of weights W of the convolution filter and bias b. The product of image features with weights and addition of bias b undergoes a non-linear activation function f, a sigmoid. The rectified feature map C then undergoes pooling, reducing the dimension of the feature map. Finally, the feature map is flattened and is recorded as Z_{SACNN} for each input instance x. For an input video frame of size, 256 × 256, Z_{SACNN} of size 1 × 512 represents the SA-CNN features for each input accordingly.

2.2.3. Video metadata encoding

For each patient, the LUS video collection involved a maximum of 12 manual "sweeps" of the ultrasound probe along the body. These sweeps were taken from anterior (A), posterior (P), and lateral (L) views of both left (Lf) and right (Rt) lungs. Furthermore, each sweep is characterized by the orientation of the probe, i.e., transverse (T) or sagittal (S). The patient's age, the imaged side of the lung, the view of the acquired sweep, and the probe orientation collectively form the video metadata. The metadata are mapped such that they contain only numerical values to be compatible with interpretation by standard neural networks. The mapping function converts the sweep acquisition information to three concatenated one-hot encoding vectors, where the value of 1 respectively indicates which lung was imaged, which body location was swept along, and which probe direction was used. The vector is also concatenated with the patient age in months, normalized by the oldest age

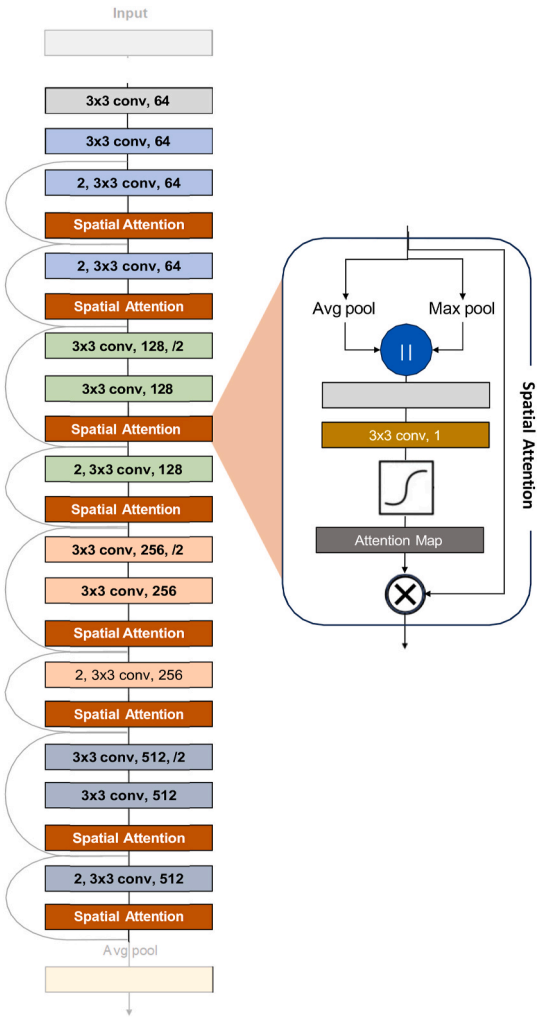

Fig. 3. Architecture of SA-CNN (ResNet-18 with spatial attention), along with an expanded structure of a spatial attention block.

present in the dataset, to keep the values of the entries of the encoding vector between 0 and 1. As a result, the metadata encoding for each LUS video is a vector of dimension 1 and length 8, see, for example, Fig. 4.

2.3. FLUEnT: fused lung-ultrasound encoding-based transformer

The complete end-to-end representation of FLUEnT architecture is shown in Fig. 5. FLUEnT is composed of a VAE, a SA-CNN, metadata, an embedding layer followed by a concatenation layer, a transformer encoder, and a classifier. This section focuses on the description of the transformer and its input and output. The objective of our proposed transformer model is to learn the mapping from the sequence of encoded images to the corresponding video label.

Side of Lung | View of Acquisition | Probe Orientation | Age

Lf	Rt	A	P	L	T	S	Age	1x8
1	0	1	0	0	1	0	0.5	

Fig. 4. Metadata vector representation of a LUS video (top row) with values for an example acquisition (bottom row). The sample video shows the left lung recorded during a sweep along the anterior of the lung with a transverse probe orientation. The patient age is encoded as 0.5 since the true age was 29 months and the oldest patient in the dataset was 58 months old.

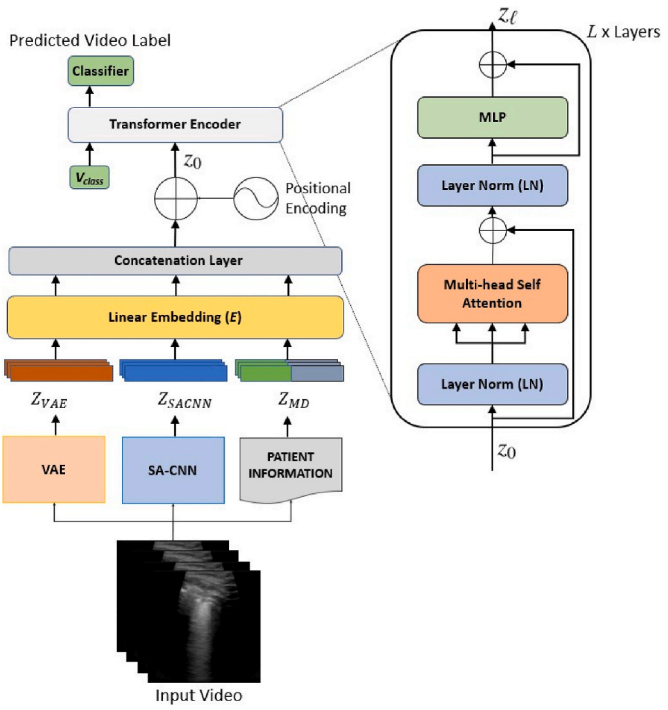

Fig. 5. Block-based architecture of proposed LUS video classification framework.

For a video V that consists of n frames of size $w \times h \times c$, where h is the height, w the width, and c the number of channels, a sequence of frames is extracted ($x_1, x_2, \ldots, x_n$). The notation $\{x_i, y_i\}_{i=1}^{n}$ denotes a set of lung ultrasound frames within the video, where x_i is a frame and y_i is its corresponding label. The spatial feature encodings from the frame sequence, extracted using the VAE and SA-CNN, along with the encoding of the patient metadata are linearly projected onto respective vectors of dimension $d = 200$, using a learned embedding matrix E. The value of d is tuned as a hyper-parameter during the training, over a search space that included values of 100, 200, 400, and 800. The three embedded representations are then concatenated row-wise, resulting in a single, unified embedding. The embedding is then combined with a learnable classification token v_{class}, required to perform the classification task. The embedded frames are viewed by the Transformer as a set of images without any notion of their order within the sequence. To keep the video frames arranged in time, the sinusoidal positional information E_{pos} is encoded and appended to the encoded video frame-level representations. The resulting embedded sequence of frames with the classification token v_{class} is given as

$$z_0 = [v_{\text{class}}; x_1E; x_2E; \ldots; x_nE] + E_{\text{pos}}, \tag{4}$$

where $E \in \mathbb{R}^d$ and $E_{\text{pos}} \in \mathbb{R}^{(n+1)\times d}$. The resulting sequence of embedded video frames z_0 is passed to the transformer encoder. As shown in Fig. 5, the transformer encoder is made of L identical layers. Each layer further consists of two key components. A multi-head self-attention block (MSA) and a fully connected feed-forward dense block (MLP). The latter block consists of two dense layers with a Gaussian Error Linear Unit (GELU) activation in between. Each of the two sub-components of the encoder employs residual skip connections preceded by a normalization layer (*LN*):

$$z'_\ell = MSA(LN(z_{\ell-1})) + z_{\ell-1}, \quad \ell = 1, \ldots, L, \text{ and} \tag{5}$$

$$z_\ell = MLP(LN(z_\ell)) + z'_\ell, \quad \ell = 1, \ldots, L. \tag{6}$$

At the last layer of the encoder, FLUEnT takes the first element in the sequence $z_L^{(0)}$ and passes it to the classifier for predicting the class label

$$y = LN\left(z_L^{(0)}\right).$$

2.4. Data evaluation procedure

To measure the effectiveness of FLUEnT, the dataset is divided into three non-overlapping splits based on their use in the evaluation of FLUEnT. These patient-level splits are called *Split-A*, *Split-B*, and *Split-C*. The patient-wise split was determined in a random selection manner, ensuring an unbiased representation across all the data splits. The number of patients per split and their consolidation findings are listed in Table 2. A patient has a consolidation finding if in one or more of the 12 sweeps the radiologists found at least one consolidation. *Split-A* is used as the training set for the generation of LUS encodings using VAE and CNN, *Split-B* serves as the validation set for hyper-parameter tuning of the VAE, and *Split-C* is used to train and evaluate the proposed FLUEnT framework. Videos with consolidation findings are, on average, longer than videos without consolidation frames and have a higher standard deviation in length. To feed the videos as input to an automated video classification model, all the videos in $Split - C$ (both with and without consolidations) were normalized to an equal length of 400 frames. In this regard, shorter videos were padded with frames of zeros, and longer videos were shortened by dropping frames toward their end without the loss of class-relevant information. This process was enabled by the frame-level annotations provided by the radiologists. These annotations identified the start and end of frames featuring consolidations within the videos.

2.5. VAE training

The training of the VAE calibrates the encoder parameters to ensure that Eq. (2) fits the distribution of training images x. This can be achieved by minimizing the loss

$$\mathscr{L}(x) = \mathscr{L}_R(x) + \mathscr{L}_{KL}(x), \tag{7}$$

where $\mathscr{L}_R$ the pixel-wise mean squared error (mse) between the original image x and the reconstructed image x', given as

$$\mathscr{L}_R(x) = \frac{1}{N}\sum_{i=0}^{N-1} \left\|\mathbf{x}_i' - \mathbf{x}_i\right\|^2, \tag{8}$$

and $\mathscr{L}_{KL}$ is the Kullback–Leibler (KL) divergence loss, introduced as a loss function for VAE in Ref. [39] for the first time.

$$\mathscr{L}_{KL}(x) = D_{KL}(q(z|x)\|p(z)). \tag{9}$$

The VAE is trained on $Split - A$ with the task to encode the LUS video frames into a latent vector that allows reconstruction with the minimum reconstruction error $\mathscr{L}_R$. Furthermore, the training aims to reduce the divergence between the learned prior and the posterior distributions $\mathscr{L}_{KL}$. $Split - B$ is used to fine-tune the hyper-parameters and to choose the best model for encoding spatial information of video frames. The tuned hyper-parameters include batch size, learning rate, number of epochs, and optimization algorithm. As a result, the VAE model is trained using the Adam optimizer with a learning rate of *1e-3* with a batch size of 16 for a total of 200 epochs. Frame-level encodings for video frames from *Split-C* are then processed by the proposed FLUEnT model for video-level scoring. It is to note that the trained VAE, in the proposed study, is not used to generate LUS data but only to encode the LUS video frames.

Table 2
Number of patients and consolidations per evaluation split.

Split	No. of Patients	Without Consolidations	With Consolidations
A	70	56	14
B	8	4	4
C	122	84	38

2.6. SA-CNN training

The SA-CNN was trained from scratch with the pediatric LUS data from *Split-A*. Likewise to the training of VAE, the *Split-B* was used as a validation set to fine-tune the hyper-parameters and to choose the best-performing model. The model was trained and validated on $Split - A$ and $Split - B$ respectively, with a learning rate of $1e - 4$ and stochastic gradient descent (SGD) as the optimizer [40]. The cross-entropy loss was computed and back-propagated over a batch of 64 frames. Likewise for the VAE, the SA-CNN encodings, generated for the video frames from *Split-C*, are passed to FLUEnT for video-level scoring.

2.7. FLUEnT training

Frame-level encodings for the LUS video data of Split-C are collected as encoded representations from the VAE encoder and spatial features from the SA-CNN, both trained and validated on $Split - A$ and $Split - B$ respectively. Patient metadata is also extracted for each patient to fuse with the frame-level encodings. To identify the individual contribution of each encoded representation, we followed a compositional approach by training and evaluating our proposed architecture for each encoding separately and their combinations accordingly, yielding three ablations VAE, SA-CNN, and VAE + SA-CNN, as well as the final FLUEnT approach i.e., VAE + SA-CNN + metadata. All the ablations along with the full FLUEnT approach utilized the same training data, i.e. encoded representations of $Split - C$, for the training of the transformer model. The training strategy for the ablations and FLUEnT approach followed the leave-one-out cross-validation (LOOCV) approach. LOOCV is a configuration of k-fold cross-validation where k is set to the number of examples in the dataset. Here the value of k is set to the number of patients in *Split-C*, i.e. 122, making sure that there is no data from the same patient in the train and test sets. This is important because training on a video of a patient and then testing on another video of the same patient may bias the model. FLUEnT is trained at a learning rate of *1e-4* with a Cross Entropy Loss. Videos in a batch size of 16 were used to train over a total of 30 epochs, across which the best-performing model was taken. To mitigate the impact of data imbalance during the training of VAE, SA-CNN, and FLUEnT, class weighting technique was employed.

2.8. System specifications

The VAE, SA-CNN, and the proposed FLUEnT architecture were trained and evaluated on an NVIDIA A100 GPU.

2.9. Evaluation metrics and statistics

To measure how each of the ablation models and FLUEnT perform on the utilized pediatric LUS dataset, the employed metrics are described below to evaluate the reconstruction of the VAE and the classification abilities of the SA-CNN and the transformer.

2.9.1. Image reconstruction (PSNR & SSIM)

To evaluate how well the image information is encoded in the latent space by the VAE, its corresponding reconstructed image by the decoder is analyzed. To measure the image reconstruction error the peak signal-to-noise ratio (PSNR) and the structure similarity (SSIM) index is utilized. The PSNR computes the ratio between the maximum possible value (power) of an image and the power of the distorting noise that affects the quality of its representation. Mathematically it is given as

$$\text{PSNR} = 10\log_{10}\left(\frac{M - 1^2}{\text{MSE}}\right), \tag{10}$$

where M refers to the maximum possible intensity value in an image

(255 in this case), and MSE is the mean squared error defined by Eq. (8). The structure similarity SSIM is the measure of similarity between the original and reconstructed image and considers image degradation as the change perceived in structural information s while incorporating perceptual information such as luminance l and contrast c. Mathematically it is given as

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = [l(\mathbf{x},\mathbf{y})]^{\alpha}\,[c(\mathbf{x},\mathbf{y})]^{\beta}\,[s(\mathbf{x},\mathbf{y})]^{\gamma}, \tag{11}$$

where α, β, and γ control the significance of each of the information, and their values are set as 1 [41].

2.9.2. SA-CNN classification (F1-Score)

The F1-Score is the harmonic mean of precision and recall. Taking into account both precision and recall of the classification model, it is considered as a strong metric for a classification task without biasing predictions towards the majority class. Since the dataset is highly imbalanced with the majority of frames without consolidations, this metric is used for a fair analysis. The F1-score is defined as

$$\mathrm{F1} = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP} + \mathrm{FN})}, \tag{12}$$

where TP, FP, and FN are the number of true positive, false positive, and false negative classifications, respectively.

2.9.3. Video classification (balanced accuracy)

To evaluate the video classification performance of FLUEnT, *balanced accuracy* (BA) is used, which is computed as the average of sensitivity and specificity. Since the ground truth at the video level is highly imbalanced between consolidation and non-consolidation findings, BA is a valuable measure of performance. Mathematically, the balanced accuracy is given as

$$\mathrm{BA} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}\right). \tag{13}$$

3. Results

3.1. Image reconstruction & SA-CNN classification results

Starting with the visualization of images reconstructed by the VAE in Fig. 6, it is observed that the VAE decoder is able to capture the significant image patterns when sampled from the latent space. The reconstructed images yielded a PSNR of 33.5 dB and an SSIM of 0.8842 for the *Split-B*. These results show that the VAE is able to generalize the reconstruction of significant patterns with little loss of information among frames taken from LUS videos of different patients. The SA-CNN-based frame-level classification of *Split-B* data resulted in a binary F1-Score of 0.61. This score is relatively low, however, our primary focus is to see how effectively the SA-CNN extracted features may contribute to the LUS video classification task.

3.2. Video classification results of FLUEnT and ablation studies

Ablation studies are conducted to evaluate if the reconstruction quality of the VAE and the classification performance of the spatially attentive ResNet-18 are respectively able to translate in effective video-level scoring. It is also evaluated if the patient metadata aids in further improving model performance. Video-level classification results are summarized in Table 3, which shows them broken down based on the contributing encoding methods in the final fused encoded representation of the video frames. Table 3 reports the mean, minimum, and maximum values of the employed video-level evaluation metric *balanced accuracy (BA),* and associated values of *sensitivity* and *specificity*. Additionally, the negative predictive value (NPV) and positive predictive value (PPV) are also reported in order to ascertain a complete picture of model performance on both positive and negative samples.

In (Table 3, rows 5–8), encodings generated individually from the VAE and from the CNN (ablation 1 and 2) resulted in a mean *BA* of 85.9 % and 80.6 %, respectively. Fusion of both of the neural network-based encoding methods (ablation 3) resulted in an improved video-level classification performance achieving a mean *BA* of 87.1 %. The full FLUEnT model, which combines the encodings from the VAE and SA-CNN with the patient metadata, outperformed the rest of the combinations with the highest mean *BA* of 89.3 % with a sensitivity of 83.8 % and specificity of 94.2 %. This indicates that the classification of these encodings performs better when fused together than when classified individually. In efforts towards investigating the correlation between consecutive frames, the classification performance of FLUEnT is also evaluated by testing it on downsampled video data. By dropping every other frame within a LUS video and giving it as an input, FLUEnT showed a decrease in classification performance with a mean balanced accuracy of 78 %. This drop in performance indicates that the consecutive frames within the video may not be significantly correlated and removing them can result in the loss of class-relevant information. This is also reflected by the sweeping nature of data acquisition.

3.3. Performance comparison with baseline models

To draw a fair comparison with the state-of-the-art transformer-based methods, the video vision-transformer ViViT [32], the TimeSformer [33], and the Swinformer [34] are considered as the baselines

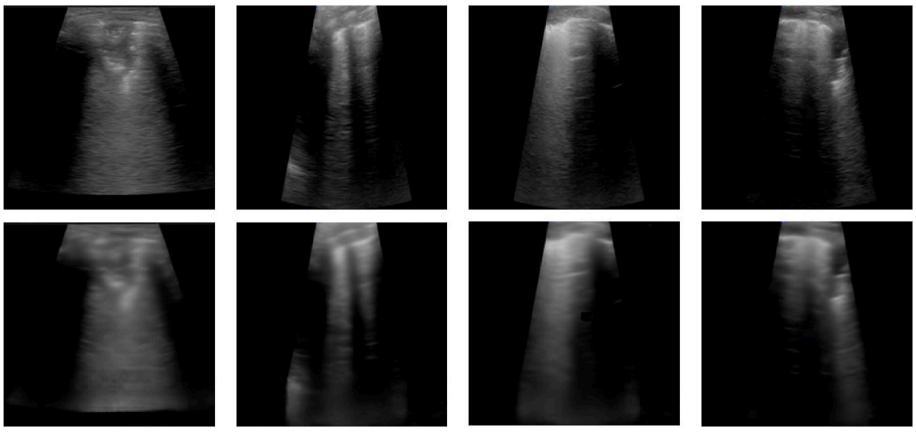

Fig. 6. VAE Reconstruction Results: Samples of original images (top) and corresponding reconstructed images (bottom). The images in the 1st and 4th columns show the presence of segmental/lobar consolidation; no consolidations are visible in the remaining images.

Table 3
Results for Baseline Models, FLUEnT, and Alblation Studies. The metrics, apart from Average Computational Time, represent the minimum-mean-maximum values for all the folds in LOOCV (%).

Model	Sensitivity	Specificity	NPV	PPV	Balanced Accuracy	Average Computational Time (ms)
Baseline 1: ViViT	59-70.3-81.6	54.31-88.5-100	62.5-86.0-100	62.58-74.8-87.02	57.1-79.4-100	33.33
Baseline 2: TimeSFormer	50.54-81.2-100	69.64-85.6-100	72.92-85.1-97.28	55.38-81.8-100	60.1-83.4-100	43.10
Baseline 3: Swinformer	46.04-84.3-100	76.19-90.4-100	66.53-87.9-100	51.37-87.4-100	61.2-87.4-100	32.46
Baseline 4: CNN + LSTM	0-19.67-59.59	8.32-57.72-100	7.0-55.0-100	0-6.0-20.0	22.14-61.82-100	22.72
Ablation 1: VAE	53.3-76.6-100	84-75-95-100	72.13-95.7-100	55.9-74.5-93.1	69-85.9-100	3.47
Ablation 2: SA-CNN Features	63.09-74.5-85.91	64.55-85.7-100	56.76-85.2-100	57.68-75.2-92.72	64.35-80.6-96.85	3.47
Ablation 3: VAE + SA-CNN Features	64.1-79.2-94.3	69.56-95.3-100	78.27-94.1-100	62.1-82.9-100	66.7-87.1-100	4.03
FLUEnT: **VAE + SA-CNN Features + Metadata**	**66.43–83.8–100**	**80.78–94.2–100**	**80.15–91.6–100**	**77.56–88.5–99.94**	**74–89.3–100**	**4.09**

models, trained on the same task. The CNN-LSTM approach proposed in [24], is also evaluated to extend the comparative analysis with our proposed method. The classification results of the baseline models on the pediatric LUS dataset are shown in Table 3, rows 1–4. ViViT achieved a mean *BA* of 79.4 %, with a sensitivity of 70.3 % and specificity of 88.5 %. Video-level classification using TimeSformer showed improved performance compared to ViViT with a mean *BA* of 83.4 % and sensitivity and specificity of 81.2 % and 85.6 %, respectively. Among the three transformer-based models, Swinformer performed the best with a mean *BA* of 87.4 % and sensitivity and specificity of 84.3 % and 90.4 %, respectively. CNN-LSTM showed a comparatively lower performance with a mean *BA* of 61.82 % and sensitivity and specificity as 19.67 % and 57.72 % respectively. Considering the best performance among the baseline approaches, results showed that the ablation study 3, achieved a comparable performance with the Swinformer. Whereas, the full FLUEnT model outperformed Swinformer by almost 2 % points.

3.4. Computational complexity

To evaluate if the performance gain of FLUEnT over the baseline models is achieved at the cost of increasing computational complexity, the computational complexity of the FLUEnT model is analyzed. Since the highest order time complexity present in the FLUEnT model is the self-attention module of the transformer, the overall time complexity of FLUEnT can be summarized as $O(n^2)$, where n is the number of input frames and k is a constant proportional to the image size as well as the convolutional kernel size and stride. Considering the patch processing, our method considers each frame in a video as a single token. Therefore, in comparison to the baseline approaches that process N patches or tokens per frame over a length of video T (resulting in $N \times T$ tokens), our proposed method only processes T tokens. To quantify the analysis, the computational time of FLUEnT and the ablations are compared against the state-of-the-art video classification models. For this purpose, the average computational time per frame is reported as shown in Table 3. FLUEnT demonstrated the capability to process a single frame in 4.09 ms, compared to the computational time of 33.3, 43.10, 32.46, and 22.72 ms per frame by ViViT, TimeSformer, Swinformer, and CNN + LSTM architecture, respectively.

4. Discussion

The study proposed FLUEnT, the first video-level classification architecture that can detect the presence or absence of consolidations due to pneumonia in lung ultrasound videos of pediatric patients. FLUEnT is inspired by the architectures of ViViT, TimeSformer, and Swinformer, which divide video frames into a collection of patches and linearly project them into a 1D space for action classification. Instead, it considered each image in the video as a single patch and used convolutional neural network-based embeddings. In this regard, it used the VAE and ResNet-18 with spatial attention-based encoding methods to generate feature representations in a low dimensional space. It also integrated the patient and acquisition information as metadata to add domain knowledge to the classification task. To this extent, information about the patient's age, the imaged side of the lung, the view direction of the acquired sweep, and the orientation of the ultrasound probe were utilized.

Results show that although ResNet-18 with spatial attention performed the frame-level classification task with a relatively low F1-Score of 0.61, it is able to capture class-relevant features that are useful for LUS video classification (*BA* – 80.6 %). The variational autoencoder, on the other hand, is also able to capture the frame-level features that correspond to accurate reconstructions resulting in a promising video-level classification performance (*BA* – 85.9 %). Combining the neural network-based embeddings from the two, an improved video-level classification (*BA* – 87.1 %) can be achieved. Furthermore, another layer of distinction can be added in the form of patient metadata, which can effectively improve the training of FLUEnT, achieving the highest video-level classification performance with a mean balanced accuracy of 89.3 %.

Inspired by the state-of-the-art transformer-based architectures of ViViT [32], TimeSformer [33], and Swinformer [34], our study evaluated the three architectures as the baselines for comparison. Results showed that the ViViT, TimeSformer, and Swinformer classified the LUS video data with mean balanced accuracies of 79.4 %, 83.4 %, and 87.4 % respectively. Although the Swinformer performed better than the TimeSformer, FLUEnT outperformed the prior with 2 % points using the fused feature encodings. The results indicate that although the transformers are designed to capture both the spatial and temporal information from the video data, the convolutional feature-based spatial information enhances the feature embeddings, thus resulting in an overall improved classification performance. In addition, the gain in the classification performance by FLUEnT is achieved without any additional computational overhead. On the contrary, instead of evaluating self-attention between the frame patches/tokens of the video sequence, FLUEnT computes the self-attention among the frames as a whole. This reduces the number of tokens processed by FLUEnT in comparison to the ones processed by ViViT, TimeSformer, and Swinformer.

4.1. FLUEnT as a potential patient screening tool?

To study whether FLUEnT can be used as a screening tool in identifying patients with possible consolidation findings, it is to evaluate how accurately the proposed model identifies patients with and without consolidations. In our setting, a minimum of one consolidation finding in one of the (up to 12) videos recorded per patient, flags the patient as "having consolidation." Flagging of patients, based on consolidation findings in at least one out of a maximum of 12 sweeps (videos), is set as the screening threshold. The model is able to correctly flag 37 out of 38

patients with consolidations, mis-classifying only one patient as a false negative, thus showing a sensitivity of 97.3 %. For the 84 patients with no consolidations in any of their 12 sweeps (videos), FLUEnT flagged 14 patients with consolidation findings (specificity of 83.3 %). Overall results show that the FLUEnT is able to screen patients with a balanced accuracy of 90.33 % for the presence/absence of consolidation.

Lung ultrasound (LUS) videos from the anterior left and anterior right views are acquired by positioning the probe at the midclavicular line and scanning down to the liver in a sweep acquisition manner as part of the acquisition protocol [36]. These views can therefore include anatomical structures such as the heart and liver, respectively. As a result, these structures can potentially be misclassified as lung consolidations. Upon further investigation, the model demonstrated a specificity of 94.7 % for the anterior right and 90 % for the anterior left view. Given that these specificity values are close to the overall specificity of 94.2 %, it appears that the presence of these organs does not substantially impact the model's performance. However, to validate this further, it is essential to employ methods that introduce explainability to the model's behavior. This would help to understand which parts of the negative case videos the model mistakenly identifies as consolidations.

Overall, the analysis suggests the potential utility of the proposed model to assist clinicians in the initial screening of patients for consolidations. A characteristic of the dataset, however, calls for caution: Clinicians recorded videos without being blinded to the patient's symptoms and may have slowed down the speed of the probe during data acquisition when they observed a suspicious artifact. This could explain the difference in the distributions of video length between videos with and without consolidations (Section 2.1) and could potentially result in an inductive bias of the model. Moreover, the distributions of video lengths are not only affected by the pace of the sweep but also by the size of the patient. In the utilized dataset, children aged 1–59 months were imaged, and thus their sizes likely differed significantly. Regarding the training of FLUEnT, it is trained on LUS data acquired from a single probe in one medical center. Training on multi-center pediatric LUS data will likely expose the model to a broader range of variations, thus enabling it to learn more comprehensive and discriminative features. Without training and evaluating the performance of FLUEnT on additional datasets, it may be too early to consider it as a patient screening tool, and further investigation may be required.

5. Conclusion

Results showed that FLUEnT is able to accurately detect consolidation findings in a LUS video with a mean balanced accuracy of 89.3 %. The analysis highlighted that the fusion of spatial information extracted from a spatially attentive convolutional neural network (ResNet-18 with spatial attention) and a variational autoencoder with video metadata was effective in capturing class-relevant information (consolidation present or not in the video). This resulted in a successful training of FLUEnT achieving high video-level classification performance. FLUEnT also outperformed models like ViViT, TimeSformer, and Swinformer which are the state-of-the-art transformer-based methods in non-medical video analysis, e.g., of human day-to-day activities. The effectiveness of FLUEnT as a potential tool to screen patients for consolidations was also analyzed and it was found that FLUEnT showed promising screening performance with a sensitivity and specificity of 97.3 % and 83.3 %, respectively.

6. Limitations and future work

It is to be noted that the utilized pediatric LUS dataset is relatively small compared to other video datasets traditionally used for image and video classification. Its issues of size and class imbalance are common among medical image data sets. It is suggested that the way FLUEnT overcomes these limitations, using a fusion model, might be helpful in other data-constrained fields as well. FLUEnT is trained on LUS data acquired from a single butterfly probe. This may limit the generalization capability of the model on other datasets. In this regard, training and evaluation of FLUEnT on large pediatric datasets acquired from multiple probes and assessing its ability to generalize to other pediatric patient populations will be tested in future work. The focus of our study was limited to identifying pneumonia cases with severe lung involvement by detecting the presence/absence of consolidations within the LUS videos. Sonographic LUS features such as thickened pleural line and white lung can indicate an abnormal lung. As a future study, we also aim to include these features as characteristic patterns to improve the sensitivity of the proposed method toward identifying patients with pneumonia. Integrating other meta parameters like a symptom score accounting for vital signs and the presence of clinical symptoms associated with WHO clinical pneumonia can be a potential future direction. We also aim to expand on the correlation analysis between the consecutive frames by introducing methods like optical flow analysis or frame differencing to gain better understanding of temporal dynamics and correlation within video data. Another promising direction can be the development of a hybrid approach using CNN-based and Vision Transformer-based encodings that can leverage both the local and global context of the extracted features. Furthermore, methods to introduce the explainability of the proposed model are also worth investigating to gather a better understanding of the model's behavior. Methods will also be investigated to effectively deploy the proposed model on handheld devices such as smartphones and tablets to provide clinical assistance in resource-constrained environments.

CRediT authorship contribution statement

Umair Khan: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Russell Thompson:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation. **Jason Li:** Methodology. **Lauren P. Etter:** Writing – review & editing, Writing – original draft, Data curation. **Ingrid Camelo:** Writing – review & editing, Writing – original draft, Data curation. **Rachel C. Pieciak:** Writing – review & editing, Writing – original draft, Data curation. **Ilse Castro-Aragon:** Writing – review & editing, Data curation. **Bindu Setty:** Writing – review & editing, Data curation. **Christopher C. Gill:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Funding acquisition, Data curation. **Libertario Demi:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Margrit Betke:** Writing – review & editing, Writing – original draft, Supervision, Resources, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Libertario Demi is the co-founder of UltraAI. The rest of authors declare no conflict of interests.

Acknowledgments

This work was partially funded by the U.S. National Institutes of Health (R21TW011343). Part of this work was supported by a research grant provided in compliance with Mission 6/component 2/Investment: 2.1 "Strengthening and enhancing biomedical research in the NHS", financed by the European Union - NextGenerationEU, CUP: E63C22003260001.

References

[1] T. Bobomuratov, U. Yusupova, Identification of clinical and laboratory changes in acute pneumonia in young children living in an unfavorable ecological environment (in Khorezm region), Journal of Intellectual Property and Human Rights 2 (1) (2023) 14–20.

[2] Pneumonia in children. https://www.who.int/news-room/fact-sheets/detail/pneumonia. (Accessed 27 January 2024).

[3] D.R. House, S. Rijal, S. Adhikari, M.L. Cooper, C.M. Hohl, Prospective evaluation of World Health Organization guidelines for diagnosis of pneumonia in children presenting to an emergency department in a resource-limited setting, Paediatr. Int. Child Health 40 (4) (2020) 227–230.

[4] World-Health-Organization, Revised WHO classification and treatment of pneumonia in children at health facilities: quick reference guide, Tech. rep. (2014).

[5] M.A. Chavez, N. Naithani, R.H. Gilman, J.M. Tielsch, S. Khatry, L.E. Ellington, J. J. Miranda, G. Gurung, S. Rodriguez, W. Checkley, Agreement between the world health organization algorithm and lung consolidation identified using point-of-care ultrasound for the diagnosis of childhood pneumonia by general practitioners, Lung 193 (2015) 531–538.

[6] D.S. Balk, C. Lee, J. Schafer, J. Welwarth, J. Hardin, V. Novack, S. Yarza, B. Hoffmann, Lung ultrasound compared to chest x-ray for diagnosis of pediatric pneumonia: a meta-analysis, Pediatr. Pulmonol. 53 (8) (2018) 1130–1139.

[7] F. Mento, U. Khan, F. Faita, A. Smargiassi, R. Inchingolo, T. Perrone, L. Demi, State of the art in lung ultrasound, shifting from qualitative to quantitative analyses, Ultrasound Med. Biol. 48 (12) (2022) 2398–2416.

[8] J.-E. Bourcier, J. Paquet, M. Seinger, E. Gallard, J.-P. Redonnet, F. Cheddadi, D. Garnier, J.-M. Bourgeois, T. Geeraerts, Performance comparison of lung ultrasound and chest x-ray for the diagnosis of pneumonia in the ED, Am. J. Emerg. Med. 32 (2) (2014) 115–118.

[9] F. Cortellaro, S. Colombo, D. Coen, P.G. Duca, Lung ultrasound is an accurate diagnostic tool for the diagnosis of pneumonia in the emergency department, Emerg. Med. J. 29 (1) (2012) 19–23.

[10] L.E. Ellington, R.H. Gilman, M.A. Chavez, F. Pervaiz, J. Marin-Concha, P. Compen-Chang, S. Riedel, S.J. Rodriguez, C. Gaydos, J. Hardick, et al., Lung ultrasound as a diagnostic tool for radiographically-confirmed pneumonia in low resource settings, Respir. Med. 128 (2017) 57–64.

[11] H. Xin, J. Li, H.-Y. Hu, Is lung ultrasound useful for diagnosing pneumonia in children?: a meta-analysis and systematic review, Ultrasound Q. 34 (1) (2018) 3–10.

[12] X.-l. Liu, R. Lian, Y.-k. Tao, C.-d. Gu, G.-q. Zhang, Lung ultrasonography: an effective way to diagnose community-acquired pneumonia, Emerg. Med. J. 32 (6) (2015) 433–438.

[13] K.G. Carrera, G. Hassen, G.P. Camacho-Leon, F. Rossitto, F. Martinez, T.K. Debele, The benefits and barriers of using point-of-care ultrasound in primary healthcare in the United States, Cureus 14 (8) (2022).

[14] D. Ramgobin, V. Gupta, R. Mittal, L. Su, M.A. Patel, N. Shaheen, S. Gupta, R. Jain, POCUS in internal medicine curriculum: quest for the holy-grail of modern medicine, J. Community Hosp. Intern. Med. Perspect. 12 (5) (2022) 36–42.

[15] H. Wang, A.M. Uraco, J. Hughes, Artificial intelligence application on point-of-care ultrasound, J. Cardiothorac. Vasc. Anesth. 35 (11) (2021) 3451–3452.

[16] S. Roy, W. Menapace, S. Oei, B. Luijten, E. Fini, C. Saltori, I. Huijben, N. Chennakeshava, F. Mento, A. Sentelli, et al., Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound, IEEE Trans. Med. Imag. 39 (8) (2020) 2676–2687.

[17] O. Frank, N. Schipper, M. Vaturi, G. Soldati, A. Smargiassi, R. Inchingolo, E. Torri, T. Perrone, F. Mento, L. Demi, et al., Integrating domain knowledge into deep networks for lung ultrasound with applications to COVID-19, IEEE Trans. Med. Imag. 41 (3) (2021) 571–581.

[18] F. Mento, T. Perrone, A. Fiengo, A. Smargiassi, R. Inchingolo, G. Soldati, L. Demi, Deep learning applied to lung ultrasound videos for scoring COVID-19 patients: a multicenter study, J. Acoust. Soc. Am. 149 (5) (2021) 3626–3634.

[19] U. Khan, S. Afrakhteh, F. Mento, N. Fatima, L. De Rosa, L.L. Custode, Z. Azam, E. Torri, G. Soldati, F. Tursi, et al., Benchmark methodological approach for the application of artificial intelligence to lung ultrasound data from COVID-19 patients: from frame to prognostic-level, Ultrasonics 132 (2023) 106994.

[20] L.L. Custode, F. Mento, F. Tursi, A. Smargiassi, R. Inchingolo, T. Perrone, L. Demi, G. Iacca, Multi-objective automatic analysis of lung ultrasound data from COVID-19 patients by means of deep learning and decision trees, Appl. Soft Comput. (2022) 109926.

[21] S. Ebadi, D. Krishnaswamy, S. Bolouri, D. Zonoobi, R. Greiner, N. Meuser-Herr, J. Jaremko, J. Kapur, M. Noga, K. Punithakumar, Automated detection of pneumonia in lung ultrasound using deep video classification for COVID-19, Inform. Med. Unlocked 25 (2021) 100687, https://doi.org/10.1016/j.imu.2021.100687.

[22] C. Baloescu, G. Toporek, S. Kim, K. McNamara, R. Liu, M.M. Shaw, R.L. McNamara, B.I. Raju, C.L. Moore, Automated lung ultrasound b-line assessment using a deep learning algorithm, IEEE Trans. Ultrason. Ferroelectrics Freq. Control 67 (11) (2020) 2312–2320.

[23] R.T. Lucassen, M.H. Jafari, N.M. Duggan, N. Jowkar, A. Mehrtash, C. Fischetti, D. Bernier, K. Prentice, E.P. Duhaime, M. Jin, et al., Deep learning for detection and localization of b-lines in lung ultrasound, IEEE Journal of Biomedical and Health Informatics 27 (9) (2023) 4352–4361.

[24] B. Barros, P. Lacerda, C. Albuquerque, A. Conci, Pulmonary COVID-19: learning spatiotemporal features combining CNN and LSTM networks for lung ultrasound video classification, Sensors 21 (16) (2021) 5486.

[25] H. Kerdegari, N.T.H. Phung, A. McBride, L. Pisani, H.V. Nguyen, T.B. Duong, R. Razavi, L. Thwaites, S. Yacoub, A. Gomez, et al., B-line detection and localization in lung ultrasound videos using spatiotemporal attention, Appl. Sci. 11 (24) (2021) 11697.

[26] T. Lum, M. Mahdavi, O. Frenkel, C. Lee, M.H. Jafari, F.T. Dezaki, N. V. Woudenberg, A.N. Gu, P. Abolmaesumi, T. Tsang, Imaging biomarker knowledge transfer for attention-based diagnosis of COVID-19 in lung ultrasound videos, in: Simplifying Medical Ultrasound: Second International Workshop, ASMUS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 2, Springer, 2021, pp. 159–168.

[27] X. Fang, W. Li, J. Huang, W. Li, Q. Feng, Y. Han, X. Ding, J. Zhang, Ultrasound image intelligent diagnosis in community-acquired pneumonia of children using convolutional neural network-based transfer learning, Frontiers in Pediatrics 10 (2022) 1063587.

[28] B. Nti, A.S. Lehmann, A. Haddad, S.K. Kennedy, F.M. Russell, Artificial intelligence-augmented pediatric lung pocus: a pilot study of novice learners, J. Ultrasound Med. 41 (12) (2022) 2965–2972.

[29] D. Sarvamangala, R.V. Kulkarni, Convolutional neural networks in medical image understanding: a survey, Evolutionary Intelligence 15 (1) (2022) 1–22.

[30] N.S. Shaik, T.K. Cherukuri, Multi-level attention network: application to brain tumor classification, Signal, Image and Video Processing 16 (3) (2022) 817–824.

[31] S. Khan, M. Naseer, M. Hayat, S.W. Zamir, F.S. Khan, M. Shah, Transformers in vision: a survey, ACM Comput. Surv. 54 (10s) (2022) 1–41.

[32] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, ViVit: a video vision transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6836–6846.

[33] G. Bertasius, H. Wang, L. Torresani, Is space-time attention all you need for video understanding? ICML 2 (2021) 4.

[34] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[35] L. Demi, F. Wolfram, C. Klersy, A. De Silvestri, V.V. Ferretti, M. Muller, D. Miller, F. Feletti, M. Wełnicki, N. Buda, et al., New international guidelines and consensus on the use of lung ultrasound, J. Ultrasound Med. 42 (2022) 309–344.

[36] L. Etter, M. Betke, I.Y. Camelo, C.J. Gill, R. Pieciak, R. Thompson, L. Demi, U. Khan, A. Wheelock, J. Katanga, et al., Curated and annotated dataset of lung us images in zambian children with clinical pneumonia, Radiology: Artif. Intell. 6 (2) (2024) e230147.

[37] D.P. Kingma, M. Welling, An introduction to variational autoencoders, Foundations and Trends in Machine Learning 12 (4) (2019) 307–392.

[38] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 3–19.

[39] D.P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv: 1312.6114 (2013).

[40] R. Thompson, U. Khan, J. Li, L.P. Etter, I. Camello, R.C. Pieciak, I. Castro-Aragon, B. Setty, C.C. Gill, L. Demi, et al., Investigating effective transfer of deep learning models from adults to children for lung ultrasound data analysis, in: 2023 IEEE International Ultrasonics Symposium (IUS), IEEE, 2023, pp. 1–4.

[41] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.