Do it Like the Doctor: How We Can Design a Model That Uses Domain Knowledge to Diagnose Pneumothorax

Glen Smith[1], Qiao Zhang[1] and Christopher J. MacLellan

Drexel University, Philadelphia, Pennsylvania, 19104, United States
[1]These authors contributed equally to this work.

Abstract

Computer-aided diagnosis for medical imaging is a well-studied field that aims to provide real-time decision support systems for physicians. These systems attempt to detect and diagnose a plethora of medical conditions across a variety of image diagnostic technologies including ultrasound, x-ray, MRI, and CT. When designing AI models for these systems, we are often limited by little training data, and for rare medical conditions, positive examples are difficult to obtain. These issues often cause models to perform poorly, so we needed a way to design an AI model in light of these limitations. Thus, our approach was to incorporate expert domain knowledge into the design of an AI model. We conducted two qualitative think-aloud studies with doctors trained in the interpretation of lung ultrasound diagnosis to extract relevant domain knowledge for the condition Pneumothorax. We extracted knowledge of key features and procedures used to make a diagnosis. With this knowledge, we employed knowledge engineering concepts to make recommendations for an AI model design to automatically diagnose Pneumothorax.

Keywords

Think-aloud, Pneumothorax, Domain Knowledge

1. Introduction

When building artificial intelligence (AI) models with limited data, we are often concerned with issues of low performance. This may be due to not having enough data for the model to effectively learn the relationships between the input and the output. Another example may be over-fitting, where the model learns the noise and nuance of the training data well, but performs poorly on unseen data. Further, for many classification tasks, we often lack sufficiently balanced datasets, which additionally leads to lowered model performance.

One such way to mitigate some of these issues is by incorporating subject matter expert knowledge, called "domain knowledge" [1], into the design of an AI model. In essence, we can ask an expert "how would you accomplish this task?" and extract key steps, milestones, and outcomes that we should consider in the model design. This approach allows us to build more targeted and robust models that focus on specific, expert-defined features. In this research, we

present two studies whose aims are to qualitatively extract domain knowledge from physicians in the POCUS (point-of-care ultrasound) domain for the task of detecting and diagnosing Pneumothorax.

An AI system capable of diagnosing Pneumothorax has many applications, one of which is assisting medics in the military. Battlefield medics are often required to make real-time diagnoses of multiple conditions in stressful, high-stakes environments. Ultrasound is one of the recommended ways for these medics to diagnose Pneumothorax, but training to interpret ultrasounds can be resource heavy. Thus, there is a need for automated support systems that can assist in making these diagnoses.

Previous studies show that incorporating expert knowledge into an AI model's design can produce higher and more robust performance. [2, 3, 4]. In this work, we analyze the process subject matter experts use to diagnose Pneumothorax in lung ultrasound videos and determine which features and artifacts in the videos are considered, the order in which they are considered, and the relative importance of these features.

We considered two main model design inspirations for our studies, robustness and user confidence. With a limited training dataset (62 ultrasound videos in our case), it is difficult for a machine learning model to extract meaningful distinctions between features to make an accurate diagnosis. To address this issue, we aimed to extract two kinds of domain knowledge from our medical experts: knowledge of key features and inference knowledge. Having a method to elicit important features from experts lets us develop a model design that hones in on those attributes, which creates a more robust model [5]. Further, by extracting the procedures for diagnosing Pneumothorax, we can similarly develop inference rules to incorporate into a model. This domain knowledge should result in AI models that require fewer training instances, less training time, and provide higher performance overall.

In this paper, we first provide some background on *cognitive task analysis* and the *think-aloud* method, which is the type of cognitive task analysis we employ for our work. We then present a brief overview of the POCUS domain to establish the context and need for AI technology to diagnose Pneumothorax. Next, we present two studies, that outline how we elicited domain knowledge from two medical experts and the results of these sessions. In the discussion, we provide recommendations for the design of an AI diagnosis system based on our findings from these studies. Finally, we discuss related works, our unique contributions to the space of knowledge engineering for designing AI models to help diagnose Pneumothorax, and conclude with ongoing and future work.

2. Background

2.1. Cognitive task analysis

Cognitive Task Analysis (CTA) is the process of extracting knowledge, processes, and patterns from individuals as they attempt to evaluate a scenario in a particular domain or problem space [6]. With this analysis, researchers can evaluate the cognitive load involved in performing a task, determine the knowledge and procedures an individual uses to complete a task, and the knowledge required to determine when a task is complete. Most often, subject matter experts are the focus of such studies as they usually contain the breadth of knowledge required to

navigate the problem space or their respective domains. However, in some cases it may be desirable to study how a novice might tackle a similar problem and what errors are present in their processes. The use of the CTA method is therefore to develop systems and processes based on expert and/or novice knowledge. Some examples of these systems and processes include training programs[7], and medical procedures [8].

There are many CTA methods used to facilitate knowledge engineering and each varies in terms of their knowledge representation, how information is extracted from participants, and the types of tasks the methods work best for. These facets are further divided into their own categories of methods for knowledge elicitation and representation [9]. When used for designing training systems, CTA is used two-fold, to first define the difficulty of the task in the problem space, determine common errors encountered by expert in the field, and produce ways to mitigate those errors [10]. This problem formulation stage focuses on the theory of the task and the cognitive processes required by participants to address it. The next stage examines the application of stage one, to determine technical requirements, reasons for performance issues, how a system might solve them, and what computational tools to employ [10].

Some specific CTA methodologies employed for knowledge elicitation include unstructured interviews, critical decision analyses, direct observation and questioning, and simulations [9]. In these approaches, researchers will either ask participants a series of questions, allowing for feedback and revising of further questions, or observe participants as they attempt to solve a problem or walk through a process, usually with some verbal component from either party. Often, a combination of some of these methods produce a robust qualitative method for extracting domain knowledge from subject matter experts.

According to [6] the CTA method is most effectively employed when the analysis contains "complex, ill-structured tasks" that may have multiple solutions, "dynamic, uncertain, and real-time environments" that the individual must navigate, or some level of multi-tasking where decisions are based on a variety of simultaneous conditions. Furthermore, they go on to cite that CTA is appropriate when the problem consists of complex perceptual learning and pattern recognition. All of these conditions apply to our use case, where battlefield medics must make real-time diagnoses of Pneumothorax in often stressful environments. Battlefield medics are often multi-tasking to tend to various injuries, and the constantly changing environment lends itself to uncertainty and places a load on the medics to make real-time decisions. This makes CTA a suitable method to extract physician knowledge for the purpose of developing an intelligent AI system to diagnose Pneumothorax. As explored in our two studies conducted with our physicians, we employed the Think-Aloud CTA strategy to inform the design of an AI model to diagnose Pneumothorax.

2.2. Think-Aloud Analysis

Think Alouds (TAs) are one specific type of cognitive task analysis that asks participants to verbalize each step in their problem-solving process [11]. In general, researchers conduct one-on-one sessions where they give a participant a problem to solve. The participant outlines their solution to the problem either "retrospectively" – thinking back on a previously solved problem – or "concurrently", where the problem is solved real-time with an accompanied verbalization [12]. Since the verbalization, or the "verbal report", is the output data of the study, the study is

transcribed and often recorded for later review. Transcripts are then analyzed to extract entities, concepts, decision flows, and more.

We make use of the Think-Aloud method for our studies for a myriad of reasons. Compared to other qualitative methods, the Think-Aloud method has been shown to be robust against various "errors" that would otherwise invalidate the data obtained from the process [13]. More precisely, Someren et. al. state that the act of verbalization can help resolve memory errors and allow for a slower, yet more complete cognitive flow. Think-Aloud studies are best utilized in scenarios of small participant sample sizes and simulation environments, where a participant works through a controlled "real-world" example of the problem [11].

3. POCUS Task: Pneumothorax Diagnosis

POCUS, which stands for Point-Of-Care Ultrasound, involves the use of an ultrasound device to answer specific diagnostic questions and to assess real-time physiological responses to treatment [14]. In the scope of POCUS AI, we aim to use this research to build a system that accepts ultrasound videos as input and classifies whether the video is an example of Pneumothorax [15]. A Pneumothorax, also known as collapsed lung or dropped lung, is the entry of air into the pleural space (the space between the lungs and chest wall). When air enters this area, the lung loses contact with the inside of the chest and "drops" down [16]. Figure 1 is an anatomical diagram that compares a normal lung with one conditioned with Pneumothorax. In most cases, a Pneumothorax is caused by a traumatic injury, such as a rib fracture or penetrating injury (stab or gunshot wound) that causes damage to the lung or chest [17].
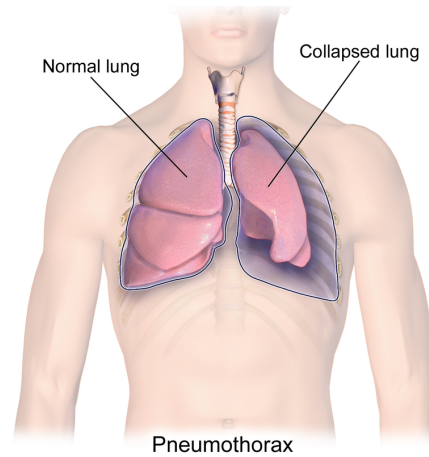

Figure 1: An anatomical diagram demonstrating the difference between a normal lung and a collapsed lung. [18]

The video data used for this study are provided by Brooke Army Medical Center (BAMC). Additionally, we utilized open-source ultrasound video data from the POCUS ATLAS [19] (examples are shown in Figure 2). For the BAMC data, we have 32 videos labeled with "sliding" and 30 videos labeled with "no sliding", where a label of "no sliding" is indicative of pneumothorax.

Each video is a 3 second short clip that contains 20 frames per second. While some of the ultrasound videos were from the same patient, each of the videos has a unique file name. Examples of "sliding" and "no sliding" snapshots from the POCUS ATLAS videos are provided in Figure 2.
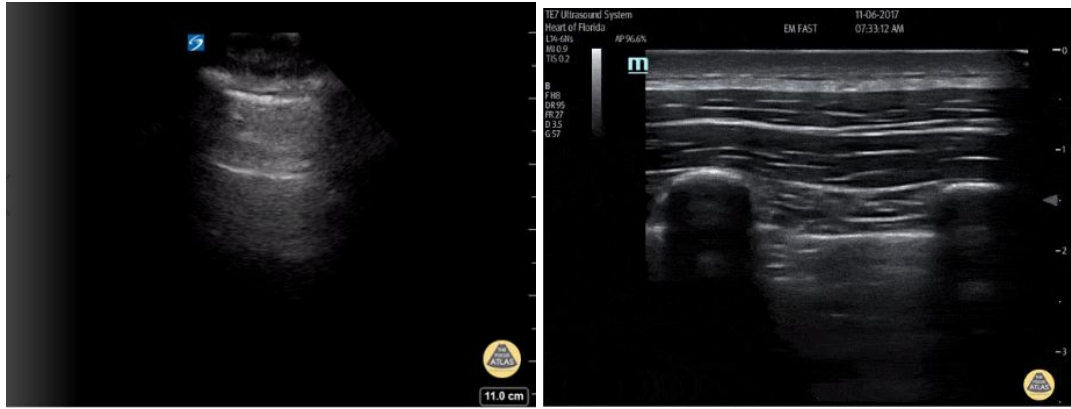

Figure 2: Two Ultrasound snapshots labeled with "sliding" (left) and "no sliding" (right). [Public domain], via the POCUS ATLAS. (https://www.thepocusatlas.com/lung/5l9jgyaszu0othj5tidg0miqxkmvyv) provided by Hannah Kopinski (MS4), Dr. Lindsay Davis and Matthew Riscinti, (https://www.thepocusatlas.com/lung/no-lung-sliding) provided by Francisco Norman.

Though we have limited training data for a complex machine learning system, the video data provide adequate information for experienced medical experts to diagnose Pneumothorax. Therefore, we were able to employ Think-Aloud analyses with our medical experts to aid in the design of a decision support system [20]. In this work, we captured the cognitive reasoning processes that the medical experts, identified in our studies with the pseudonyms Alex and Victor, use when diagnosing Pneumothorax. In the next two sections, we present the motivations, designs, analyses, and results of our two studies.

4. Study 1: Knowledge for Making Diagnoses

4.1. Motivation

To diagnose a condition such as Pneumothorax using ultrasounds, doctors must be aware of the condition's specific characteristics. They are knowledgeable of the features and artifacts that confirm, reject, or even make uncertain whether a patient has this condition. Further, it is expected that they can explain what these features are and why these features are relevant to a diagnosis. If we can elicit this knowledge as researchers, we can use it to design systems that can detect these specific features, make diagnoses using a similar process to the doctors, and generate explanations in terms that doctors can understand.

Therefore, the motivation of this study is to examine the first of the aforementioned expectations: defining the features of Pneumothorax in ultrasound videos. When a doctor makes a diagnosis, there is much we can learn about their process that would be useful for designing and building a robust AI model to detect features in an ultrasound video. We designed a Think-Aloud

study to extract this information from two doctors trained in the use and interpretation of point-of-care ultrasound captures. As we will see in the design and analysis sections, we were able to determine both the features of the ultrasound as well as the medical concepts they employed to make the diagnosis.

4.2. Design

For each doctor, we conducted two Think-Aloud sessions where we prompted them to "make a diagnosis of Pneumothorax". In the first session with each doctor, we asked them to diagnosis Pneumothorax within six ultrasound videos. This session served as a "warmup" study to familiarize the doctors with the Think-Aloud format and from which we might extract preliminary features of their diagnoses. The second session with each doctor repeated the process with six additional lung ultrasound videos. This approach of pairing a study with a "warmup" study is suggested by Someren et. al. [13]. The ultrasound videos used for the warmup sessions were open-source data from the POCUS ATLAS [19], and the videos used for the second set of sessions were from a dataset from the Brooke Army Medical Center.

Of the six videos in each session, three were examples of Pneumothorax and three were examples of non-Pneumothorax. We asked the doctors to verbalize their thinking process as they worked through the diagnosis. The videos were presented in a random order and each doctor was unaware of the true diagnosis of all videos until the end of the study. The Think-Aloud sessions occurred over Zoom and remote control of the computer's mouse was also provided so that the doctors could control the videos and visually annotate them.

With the doctor's consent, we recorded the entirety of each session for further review. During the Think-Aloud sessions, we were careful not to say things to the doctors to influence their decision making. Our interactions only consisted of a prompt to continue speaking if there was a pause of more than 3-5 seconds. This intervention was instrumental in ensuring all steps of the diagnosis process were verbalized and recorded. At the end of the session, we reviewed some keywords and concepts by asking the doctors to clarify some of their results. This ensured we captured the full breadth of knowledge presented in the Think Aloud. Finally, session transcripts were automatically generated from the recordings and these, along with the recordings, were used during our analysis.

It should be noted that in the original BAMC dataset, labels of "Sliding" and "No Sliding" were used as proxies for "No Pneumothorax" and "Pneumothorax" labels, respectively. In this context, sliding and no sliding refers to the movement of the pleural line, a key anatomical feature that doctors look for when diagnosing Pneumothroax in lung ultrasounds. Whether sliding is visible in the pleural line is one of the strongest indicators of their being no Pneumothorax. However, since sliding is not an entirely definitive proxy for a Pneumothorax diagnosis, this presented one limitation of the data. We conducted our analysis with this limitation in mind. Our study, including the use of the existing ultrasound data, was reviewed and approved by Drexel's Institutional Review Board.

4.3. Analysis

Alex and Victor made correct diagnosis on each of the six videos in Session 1 and Session 2. After we conducted our Think-Aloud studies with each doctor, we analyzed the transcripts to extract keywords and concepts that the doctors used. Transcripts were automatically generated by zoom, and prior to our analysis we cleaned the scripts by correcting grammar and punctuation, clarifying any misspelled vocabulary, and putting the dialogue in a standard format that grouped content by speaker and line. We then determined key themes and features present in the transcripts. Here is a snippet of one of the doctor's dialogue:

> *The issue here is that everything is shifting, and if I look at the line here, here, here, and here, I don't see independent movement. I don't see vertical artifacts.*

From this quote, we extracted two of our concepts seen in Tables 1 and 2, *movement* and *vertical artifacts*.

4.4. Results

We observed twelve key concepts in sessions one (see Table 1) and two (see Table 2). For each video, we label which doctors mention which concepts, using the square for Alex and triangle for Victor. For these studies, we simply mark if a concept is present in their discussion, but do not notate the number of times the concept is mentioned.

Through our analysis, we determined that there is a further distinction between the concepts. We define "features" of the ultrasound, which are objects or regions in the video, and "visual characteristics", which are characteristics the features exhibit, such as a type of movement. We noticed that some features and characteristics are discussed with higher frequency across all the videos compared to others such as *pleural line* and *movement.*

We summarized the keywords and concepts both doctors mentioned during the two sessions in Table 3. The *pleural line* is generally agreed upon by the two doctors as a critical feature to conduct analysis in both "sliding" and "no sliding" scenarios. Alex always used anatomical landmarks such as rib or muscle to locate the pleural line while Victor did not; we believe that this could be explained as personal diagnosing preference since anatomical landmarks are not a cause or sign of Pneumothorax. Lastly, Alex mentioned "lung pulse" and "vertical reverberation" when making diagnosis on "no sliding" videos, since these two features only appear in ultrasounds where the patients have a high likelihood of Pneumothorax.

As we will see in the discussion section, we are able to extract a large amount of value from this study and we present some ways this information can be utilized to design an AI system to diagnose Pneumothorax.

5. Study 2: Knowledge for Explaining Diagnoses

5.1. Motivation

While we have collected some important ultrasound features and visual characteristics from Study 1 that can help with diagnosing Pneumothorax, we wanted to further explore how domain

Table 1
Keywords summarized from Study 1 Session 1 (□ represents Alex and △ represents Victor)

Video labels	Video 1 Sliding	Video 2 No Sliding	Video 3 No Sliding	Video 4 Sliding	Video 5 No Sliding	Video 6 Sliding
Ultrasound Features						
Pleural line	□	□ △	□ △	□ △	□ △	□ △
Anatomical landmarks	□ △	□	□	□ △	□ △	□
B line	△		△			
Z line	□ △					
A line		□ △				
Visual Characteristics						
Lung pulse		□	□	□	□	
Lung point(s)		□ △	□ △		□	
Movement	□	□ △	△	△	□ △	□ △
Vertical artifacts	□ △	□				

Table 2
Keywords summarized from Study 1 Session 2 (□ represents Alex and △ represents Victor)

Video labels	Video 7 Sliding	Video 8 Sliding	Video 9 No Sliding	Video 10 Sliding	Video 11 No Sliding	Video 12 No Sliding
Ultrasound Features						
Pleural line	□ △	□ △	□ △	□ △	□ △	□ △
B line	□	△	△			□
Z line		△				
Visual Characteristics						
Lung pulse			△			□ △
Movement	□	□	□ △	□ △	△	□
Acoustic shadowing	□		△	□	□	
Horizontal sliding			□ △	△		□
Vertical reverberation	□ △	△	□	□	□	□

Table 3
Common features summarized from Session 1 and Session 2 in Study 1 (□ represents Alex and △ represents Victor)

Video labels	Sliding	No Sliding
Ultrasound Features		
Pleural line	□+6, △+5	□+6, △+5
B line	□+1, △+2	□+1, △+2
Z line	□+1, △+2	
Visual Characteristics		
Lung pulse	□+1	□+4, △+2
Movement	□+5, △+3	□+5, △+4
Vertical reverberation	□+3, △+3	□+4

experts produce a reasonable explanation of already-diagnosed conditions in ultrasound videos. This would help us to design for better better explainability and transparency in an AI system. We conducted this second study with the main purpose of capturing the domain experts' cognitive reasoning process when they see an ultrasound video paired with a previous diagnosis and are asked to produce a reasonable explanation of the diagnosis. The main difference between study 1 and study 2 is that study 1 observed how the experts generated a diagnosis of Pneumothorax, while study 2 observes how the experts explain a diagnosis of Pneumothorax. In study 2, we wanted to understand if and how the reasoning process changes when asked to confirm or reject a predetermined diagnosis and whether there is variation between the experts' diagnosis processes.

5.2. Design

We selected four videos from the BAMC data, two of them were labeled with "Sliding" and the other two were labeled with "No Sliding". The four videos were collected from four different patients. Similar to Study 1, we showed the videos to the medical experts in random order, recorded the entire session, and transcribed the recordings. In contrast to study 1, we presented the four videos with the accompanying diagnosis labels to the medical experts. To better observe how they extract key features and generate reasonable explanations, we flipped the labels (e.g., a "Sliding" label would be shown as "No Sliding" and vice versa) of one "Sliding" video and one "No sliding"video, thus providing two videos with the correct labels and the other two with incorrect labels. Ultimately, we have one video correctly labeled with "Sliding", one video correctly labeled with "No Slding", one video incorrectly labeled with "Sliding", and one video incorrectly labeled with "No Sliding".

The original ultrasound video labels and the labels that we showed the medical experts are displayed along with our analysis results in Table 4. Our study, including the use of the existing ultrasound data, was reviewed and approved by Drexel's institutional review board.

Table 4
Keywords summarized from Study 2 (□ represents Alex and △ represents Victor)

	Video 13	Video 14	Video 15	Video 16
Original label	Sliding	No Sliding	No Sliding	Sliding
Presented label	Sliding	Sliding	No Sliding	No Sliding
Ultrasound Features				
Pleural line	□	□ △	□	□ △
Anatomical landmark	□	□	□	□
Visual Characteristics				
Lung pulse	□ △			
Lung point(s)		□ △		□ △
Moving/Movement	□	□ △	□ △	□ △
Sliding	□	□ △		□ △
No sliding	□ △	□ △	□ △	□ △
Vertical reverberation artifact		□		□ △

5.3. Analysis

By providing two videos with the correct labels and the other two videos with the incorrect labels, we aimed to not only extract features that help explain the true diagnosis, but also cause the medical experts to question their diagnosis process and be more critical towards the demonstrated diagnosis results.

An interesting finding we have from this session is that both medical experts agreed to disagree with two videos. For Video 13, we showed them a video correctly labeled with "sliding", while both doctors mentioned that they would prefer recognizing the phenomenon as a lung pulse (a vertical motion of the pleura in sync with the cardiac rhythm [21]) instead of sliding. They agreed that this is suspicious of Pneumothorax, but were unable to definitively confirm it. For Video 14, we presented a video incorrectly labeled with "sliding" while the ground truth BAMC label was "No Sliding". Victor made an argument of a possibly incorrect diagnosis. Although he observed majority no-sliding, he stated that he would need more information to make the decision. Similar to Victor, Alex described the video to have 80% no-sliding and 20% sliding, and suspected Pneumothorax (which would be linked to no-sliding). Video 15 is correctly labeled with "No Sliding" and both experts made quick decisions to agree with the labeling. Video 16 is presented with the flipped label "No Sliding" and the ground truth BAMC label is "Sliding". Victor described what he saw as half sliding and half no sliding. He thought the video was suspicious of Pneumothorax (i.e., no sliding) but would need more information to make the diagnosis. Alex described what he observed as "clearly sliding".

Here is a snippet from one of the doctor's TAs:

> So the first thing I want to identify...I want to identify the pleural line, and so to identify the pleural line, I identify a rib... And so I know that the line, that is just underneath. And then I look at the movement because the question is "sliding" or "no sliding". I want to see an independent movement of sliding that could be seen here. So this is not a Pneumothorax for sure.

From this quote, we extracted three of the concepts displayed in Table 4, *pleural line*, *anatomical landmark* (referring to rib here), and *movement*.

5.4. Results

Upon completing the second Think-Aloud study, we analyzed the doctors' diagnoses in terms of sliding/no-sliding, Pneumothorax/no-Pneumothorax, and suspicions of Pneumothorax/no-Pneumothorax. Table 5 displays the original BAMC label, presented label, and the two medical experts' explanations for each video.

Compared to Study 1, where both doctors did not prioritize among the ultrasound features and visual characteristics, Alex used *inference rules* to construct his explanations of the previously labeled lung ultrasound videos. His cognitive reasoning process could be divided into four steps:

1. Alex stated that he always looks for pleural line first. To recognize the pleural line, he identifies anatomical landmarks such as ribs and muscle.
2. Next, he examined if there is any independent movement in the pleural line. If sliding is present along the entire pleural line, then he can confidently make a diagnosis of no Pneumothorax.

3. If he did not see sliding or only saw partial sliding, then he would look for lung pulse. Recognizing pulse would lead to a conclusion of no Pneumothorax.
4. Lastly, Alex would look for vertical artifacts. If there are vertical artifacts, then there is no Pneumothorax. However, if no vertical artifacts are observed, then more information is needed to make a definitive decision.

There's still a possibility that he cannot make a diagnosis after considering these four features sequentially. Poor image resolution or only seeing part of the lung were the common issues preventing him from making a diagnosis. To make a clinical decision in these cases, more ultrasound data would need to be collected to support decision making.

Table 5
Diagnosis analysis for Study 2 (□ represents Alex and △ represents Victor)

	Video 13	Video 14	Video 15	Video 16
Original label Presented label	Sliding Sliding	No Sliding Sliding	No Sliding No Sliding	Sliding No Sliding
Explanation				
Sliding				□
Reduced sliding	□			
No sliding	△		□ △	
Pneumothorax		□		□ △
No Pneumothorax	□			
Not sure about sliding/no sliding		□ △		△

6. Discussion

From our Think-Aloud studies, we determined that there are various ways in which we can incorporate lung ultrasound domain knowledge into the design of an AI system for this task.

We know that in order for a system to be capable of diagnosing Pneumothorax, it must be able to detect features that are *relevant* to the medical condition. One way researchers could achieve this is by building an object detection system that locates various features within the frames of an ultrasound video. However, not all features are relevant to Pneumothorax, so in order to construct a more robust and targeted model, researchers must be judicious in how certain features are weighted [22]. From our studies, we can not only derive the features of most importance to the doctors, but also an approximation of relative weight based on the frequencies the features were discussed across the video samples. Thus, one such model design could be an object (or feature) detector parameterized by the relative "weights" of those features. For example, we might build an object detector to identify the pleural line in an ultrasound video, so that subsequent analysis can focus on this feature.

An object detector is an excellent start, but only considers features of the ultrasound, not characteristics of those features. From the results of studies one and two, we know that "movement" plays an important role in determining Pneumothorax. In fact, it was stated by one doctor that without movement, it would be impossible to make a diagnosis for Pneumothorax.

Thus, the interpretation for a researcher would be that the model must not consider just one frame, but a series of sequential frames to determine the type of movement a feature exhibits. We, therefore, know that the type of model must consider multiple frames as an input to detect any movement.

Study two presents further insights that can help construct an AI model to diagnose Pneumothorax. In addition to "weighting" the features, we introduce the idea of "inference rules" by prioritizing the features we extracted. Inference rules suggest an order of detecting/recognizing ultrasound features and relevant visual characteristics, and provide the AI model with more knowledge for making classifications between Pneumothorax and no Pneumothorax. Similar to traversing a decision tree, the AI system would have knowledge about which feature to detect first, and whether to make a classification at that point or move on to the next feature. This approach could contribute to the accuracy of video classification, expedite the decision making process, and increase the transparency of AI-assisted image classification, which are crucial medical needs for battlefield diagnosis.

7. Related Work

There have been many other exciting works in the medical imaging space that seek to diagnose diseases using AI technology. In the scope of incorporating domain knowledge from medical experts, the closest work to ours is conducted by Guan et al., [23], where the researchers incorporated domain knowledge into the architecture design of the network for thorax disease. The proposed network in [23] has three branches, one for viewing the whole image, one for viewing the local areas and one for combining the global and local information together. One of the major differences between their approach and ours is that they performed different orders to train the network while we suggested weighting the features as well as using inference rules to prioritize the features extracted.

In [24], Liu et. al. designed a two-fold thyroid nodule classification system consisting of an ultrasound object detector to extract key features and a multi-branch convolutional neural network for classification of extracted features. In their study, the researchers utilize expert clinical knowledge to engineer feature attributes such as size and shape and place constraints on the model based on how these features are observed in practice. For example, thyroid nodule aspect ratio distributions were pre-computed based on the training set thereby ensuring detected regions would be appropriately scaled to true nodules sizes. Our proposed design is similar in that we would use multiple object detectors. However, we differ in that our object detector design uses expert knowledge to create an attention-based model due to weighting the features. Further, incorporation of inference rules creates a framework for a series of sequential object detectors, rather than simultaneous object detectors.

In other work conducted by Wang et al., [25], the researchers first used a segmentation subnetwork to locate the lung area, then the lesion areas, and finally the most discriminative features [20]. In our proposed approach to leverage inference rules in image classification, the order our AI system detect features will also match the frequency they were mentioned in the think-aloud sessions (e.g. "pleural line" is the top mentioned feature in all think-aloud sessions and it is also listed as the first feature to look at according to the inference rules). We argue

that our approach will encourage the AI model to first look at the most discriminative and supporting features, as compared to [25]. It would be interesting to further investigate how different orders of utilizing features would contribute to classification accuracy.

To the best of our knowledge, the proposed approach in [23], [24], and [25] were only tested on static Chest X-ray (CXR) or ultrasound images (not videos). Further, the features were extracted from CXR and ultrasound images. By utilizing ultrasound videos, we extracted both static and dynamic features, thus allowing for a model design capable of both object and motion detection.

8. Conclusions and Future Work

In studies one and two, we employed Think-Aloud studies to elicit domain knowledge from physicians trained in lung ultrasound interpretation. The ultimate goal of the studies was to identify the domain knowledge that doctors use, so that we can design an AI system that incorporates that knowledge. From study one, we extracted both static and dynamic features of the ultrasound videos, from which we suggested a system design focused around object detection as well as the notion of objects across multiple video frames. In study two, we examined the reasoning process the doctors utilized to explain previously generated diagnosis. By providing correct and incorrect video labels to the doctors, we sought a method to examine how their reasoning processes changed, if at all, in the presence of incorrect prior diagnoses.

We also analyzed if different knowledge was used to generate a diagnosis (study one) vs. explain a prior diagnoses (study two). Our results show that both doctors did not mention any new features in study two compared to study one. But one doctor used inference rules when explaining the diagnoses of previously labeled ultrasound videos. In this way, we were able to prioritize the features we extracted, providing more guidance and domain knowledge for the AI system we want to design.

Our Think-Aloud studies have laid the foundation for building an AI model that can automatically diagnose Pneumothorax from ultrasound videos. We envision a model that leverages the domain knowledge we have identified to reduce the amount of training data we need and that can explain the diagnoses it generates in terms that doctors and battlefield medics can understand. Moving forward, we plan to conduct more think-aloud studies so we can generalize our model to support diagnosis of multiple medical conditions from ultrasound videos.

Acknowledgments

References

[1] P. A. Alexander, Domain knowledge: Evolving themes and emerging concerns, Educational Psychologist 27 (1992) 33–51.

[2] J. Donahue, K. Grauman, Annotator rationales for visual recognition, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 1395–1402.

[3] M. Sharma, M. Bilgic, Learning with rationales for document classification, Machine Learning 107 (2018) 797–824.

[4] M. Sharma, M. Bilgic, Towards learning with feature-based explanations for document classification, in: IJCAI Workshop on BeyondLabeler-Human is More than a Labeler, 2016, pp. 1–7.

[5] 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 transactions on medical imaging (2021).

[6] C. E. Zsambok, G. Klein, Naturalistic decision making, Psychology Press, 2014.

[7] T. L. Seamster, R. E. Redding, Applied cognitive task analysis in aviation, Routledge, 2017.

[8] M. E. Sullivan, C. V. Brown, S. E. Peyre, A. Salim, M. Martin, S. Towfigh, T. Grunwald, The use of cognitive task analysis to improve the learning of percutaneous tracheostomy placement, The American journal of surgery 193 (2007) 96–99.

[9] G. Klein, L. Militello, 4. some guidelines for conducting a cognitive task analysis, in: Advances in human performance and cognitive engineering research, Emerald Group Publishing Limited, 2001.

[10] D. D. WOODS, et al., Cognitive task analysis: An approach to knowledge acquisition for intelligent system design, in: Studies in Computer Science and Artificial Intelligence, volume 5, Elsevier, 1989, pp. 233–264.

[11] M. E. Fonteyn, B. Kuipers, S. J. Grobe, A description of think aloud method and protocol analysis, Qualitative health research 3 (1993) 430–441.

[12] K. A. Ericsson, H. A. Simon, Verbal reports as data., Psychological review 87 (1980) 215.

[13] M. W. Van Someren, Y. F. Barnard, J. A. Sandberg, The think aloud method: a practical approach to modelling cognitive, London: AcademicPress 11 (1994).

[14] S. Damodaran, A. Alva, S. Kumar, M. Kanchi, Artificial intelligence in pocus: The vanguard of technology in covid-19 pandemic, Journal of Cardiac Critical Care TSS (2020).

[15] OUTREACH@DARPA.MIL, Researchers selected for point-of-care ultrasound program, 2021. URL: https://www.darpa.mil/news-events/2021-05-04.

[16] S. A. Sahn, J. E. Heffner, Spontaneous pneumothorax, New England Journal of Medicine 342 (2000) 868–874.

[17] W.-I. Choi, Pneumothorax, Tuberculosis and respiratory diseases 76 (2014) 99–104.

[18] W. contributors, Wikijournal of medicine/medical gallery of blausen medical 2014, 2020. URL: https://en.wikiversity.org/w/index.php?title=WikiJournal_of_Medicine/Medical_gallery_of_Blausen_Medical_2014&oldid=2187649.

[19] M. Macias, A collaborative ultrasound education platform, 2021. URL: https://www.thepocusatlas.com/.

[20] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, S. Yu, A survey on incorporating domain knowledge into deep learning for medical image analysis, Medical Image Analysis (2021) 101985.

[21] G. Volpicelli, M. Elbarbary, M. Blaivas, D. Lichtenstein, G. Mathis, A. Kirkpatrick, L. Melniker, L. Gargani, V. Noble, G. Via, et al., International liaison committee on lung ultrasound (ilc-lus) for international consensus conference on lung ultrasound (icc-lus). international evidence-based recommendations for point-of-care lung ultrasound, Intensive Care Med 38 (2012) 577–591.

[22] T. S. Hwang, Y. M. Yoon, D. I. Jung, S. C. Yeon, H. C. Lee, Usefulness of transthoracic lung ultrasound for the diagnosis of mild pneumothorax, Journal of veterinary science 19 (2018) 660–666.

[23] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, Y. Yang, Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification, arXiv preprint arXiv:1801.09927 (2018).

[24] T. Liu, Q. Guo, C. Lian, X. Ren, S. Liang, J. Yu, L. Niu, W. Sun, D. Shen, Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks, Medical Image Analysis 58 (2019) 101555. URL: https://www.sciencedirect.com/science/article/pii/S1361841519300970. doi:https://doi.org/10.1016/j.media.2019.101555.

[25] K. Wang, X. Zhang, S. Huang, F. Chen, X. Zhang, L. Huangfu, Learning to recognize thoracic disease in chest x-rays with knowledge-guided deep zoom neural networks, IEEE Access 8 (2020) 159790–159805.