Uncertainty-guided Graph Attention Network for Parapneumonic Effusion Diagnosis

Jinkui Hao\textsuperscript{a,h}, Jiang Liu\textsuperscript{c}, Ella Pereira\textsuperscript{d}, Ri Liu\textsuperscript{b}, Jiong Zhang\textsuperscript{a}, Yangfan Zhang\textsuperscript{b}, Kun Yan\textsuperscript{b}, Yan Gong\textsuperscript{f}, Jianjun Zheng\textsuperscript{b}, Jingfeng Zhang\textsuperscript{b,*}, Yonghuai Liu\textsuperscript{d}, Yitian Zhao\textsuperscript{a,f,g,*}

\textsuperscript{a}Cixi Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China
\textsuperscript{b}Hwa Mei Hospital, University of Chinese Academy of Sciences, Ningbo, China
\textsuperscript{c}Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China
\textsuperscript{d}Department of Computer Science, Edge Hill University, Ormskirk, UK
\textsuperscript{e}Ningbo Eye Hospital, Ningbo, China
\textsuperscript{f}Zhejiang International Scientific and Technological Cooperative Base of Biomedical Materials and Technology, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China
\textsuperscript{g}Zhejiang Engineering Research Center for Biomedical Materials, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China
\textsuperscript{h}University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:

Keywords: Parapneumonic effusion, Uncertainty, Graph

ABSTRACT

Parapneumonic effusion (PPE) is a common condition that causes death in patients hospitalized with pneumonia. Rapid distinction of complicated PPE (CPPE) from uncomplicated PPE (UPPE) in Computed Tomography (CT) scans is of great importance for the management and medical treatment of PPE. However, UPPE and CPPE display similar appearances in CT scans, and it is challenging to distinguish CPPE from UPPE via a single 2D CT image, whether attempted by a human expert, or by any of the existing disease classification approaches. 3D convolutional neural networks (CNNs) can utilize the entire 3D volume for classification: however, they typically suffer from the intrinsic defect of over-fitting. Therefore, it is important to develop a method that not only overcomes the heavy memory and computational requirements of 3D CNNs, but also leverages the 3D information. In this paper, we propose an uncertainty-guided graph attention network (UG-GAT) that can automatically extract and integrate information from all CT slices in a 3D volume for classification into UPPE, CPPE, and normal control cases. Specifically, we frame the distinction of different cases as a graph classification problem. Each individual is represented as a directed graph with a topological structure, where vertices represent the image features of slices, and edges encode the spatial relationship between them. To estimate the contribution of each slice, we first extract the slice representations with uncertainty, using a Bayesian CNN: we then make use of the uncertainty information to weight each slice during the graph prediction phase in order to enable more reliable decision-making. We construct a dataset consisting of 302 chest CT volumetric data from different subjects (99 UPPE, 99 CPPE and 104 normal control cases) in this study, and to the best of our knowledge, this is the first attempt to classify UPPE, CPPE and normal cases using a deep learning method. Extensive experiments show that our approach is lightweight in demands, and outperforms accepted state-of-the-art methods by a large margin. Code is available at https://github.com/iMED-Lab/UG-GAT.

© 2021 Elsevier B. V. All rights reserved.

*Corresponding authors: jingfengzhang73@163.com; yitian.zhao@nimte.ac.cn
1. Introduction

A parapneumonic effusion (PPE) is an accumulation of exudative pleural fluid that occurs in association with a pulmonary infection, which is present in 20% to 40% of hospitalized patients with pneumonia [Wu et al., 2017]. Based on fluid characteristics and pathogenesis, PPEs are classified as uncomplicated or complicated (i.e., UPPE or CPPE) [Sahn, 2007]. Compared with UPPE, the occurrence of CPPE increases significantly the risk of morbidity and mortality if the pleural space is not adequately drained [Colice et al., 2000]. Generally, UPPE can be cured with antibiotic treatment alone, while patients with CPPE requires the drainage of the pleural space [Light, 2006]. Thus, the identification of UPPE and CPPE in the early stages is of great importance for the clinicians, in order to develop personalized treatment plans.

At present, the gold standard in the diagnosis of UPPE and CPPE is to derive them from positive findings in pathogenic and pathological examinations [Ren et al., 2019]. However, pathogenic diagnosis using pleural effusion analysis or bacterial cultures results in low positivity rates and/or long culturing times [Light, 2006]. In clinical practice, imaging techniques, e.g., thoracic ultrasound, X-ray, and computed tomography (CT), have also been demonstrated to be effective in the diagnosis and management of PPE [Ferreiro et al., 2015, 2018]. CT is the most widely used imaging technique. Its high resolution and three-dimensional (3D) view makes CT the best imaging technology to display the pleura, allowing lesion observation and location in the underlying lung [Ferreiro et al., 2018]. A recent study shows that CT has higher sensitivity in predicting a CPPE compared with other imaging techniques [PZ et al., 2017]. Existing CT-based PPE-aided diagnosis (CAD) methods mainly rely on custom-specified features, such as pleural thickening, split pleura sign, microbubbles, etc. [Porcel et al., 2017; Tsujimoto et al., 2015; Kalkanis et al., 2017; Bugalho et al., 2014]. However, the mere appearance of these features may not guarantee the presence of the disease, thus potentially leading to sub-optimal diagnostic performance.

In recent years, deep learning approaches have shown increasing success in various medical image analysis tasks [Shin et al., 2016; Zhao et al., 2016; Jin et al., 2017; Shen et al., 2017a; Ghesu et al., 2017; Fan et al., 2020; Qian et al., 2020; Zhou et al., 2021; Lyu et al., 2021]. Although there has been as yet no application in PPE detection yet, deep convolutional neural networks (CNNs) have been applied to develop advanced CAD methods for other lung related diseases, e.g., lung cancer detection [Ardila et al., 2019; Alakwaa et al., 2017], classification of lung nodules [Song et al., 2017; Cui et al., 2020] and pneumonia screening [Ouyang et al., 2020; Han et al., 2020; Zhao et al., 2018]. Most existing works have made use of 2D CT slices as the input of CNNs: these 2D slices lack depth information, and make it impossible to capture the true spatial distribution of the lesion in the lungs.

The imaging appearance of CPPE and UPPE have considerable similarity, and that, in any individual case, the CPPE cannot be distinguished from UPPE solely based on 2D CT slice. For instance, CPPE and UPPE have some image features in common, such as the presence of bilateral pleural effusions, and liquid-like homogeneous density. In clinical practice, it
usually requires thoracentesis for pleural effusion analysis or bacterial culture (Light, 2006). The main difference of imaging appearance between the CPPE and UPPE is that patients with CPPE may have pleural microbubbles, thickened visceral pleural membranes and thickened parietal pleural membranes (Ferreiro et al., 2015). However, these symptoms are difficult to be identified or characterized in 2D images due to the lack of spatial information. Precise characterization of the spatial morphology of the lesions is essential to distinguish the two infection types by CT imaging. 3D CNN may achieve this by utilizing the entire 3D volume (Alakwaa et al., 2017; Han et al., 2020; Yin et al., 2019). However, because of a large number of parameters, obtaining a well-trained and powerful 3D CNN model requires a large-scale annotated training dataset, which is a challenge for the PPE classification task due to ethics issues and the expensive imaging involved. Therefore, the development of an approach that does not rely on a large-scale training dataset, but can nonetheless effectively model the distribution of lesions is highly desirable.

A graph structure approach may model the structural continuity across and interaction between different CT slices (Zhou et al., 2021), while graph neural network (GNN) is parameter efficient compared with the 3D CNN (Parisot et al., 2018; Zhang et al., 2019; Song et al., 2021). Most existing graph neural network (GNN)-based methods utilize all subjects to construct the graph, in which a given node on the graph represents a single subject’s imaging data and the edges establish the interactions between each pair of subjects (Wang et al., 2021; Liang et al., 2021). This means that the topological structure of the graph is limited by the number and characteristics of the subjects. If a new subject needs to be diagnosed, then the GNN has to be re-trained.

Uncertainty measures can provide information about the reliability and contents of the images, enabling the model to proactively select data with high confidence and reducing reliance on training samples. To this end, in this paper, we propose a novel uncertainty-guided graph-based framework for the classification of UPPE, CPPE, and normal control classes. We use uncertainty to represent the reliability of image features derived from each slice of a CT scan: this enables GNN to adaptively fuse certainty-varied image features of the CT scan to assist in decision making, and thus to facilitate the classification process.

Compared with 2D and 3D CNNs, our approach not only bypasses the requirement for a large-scale annotated dataset, and avoids the over-fitting defect of 3D CNNs caused by high-dimensional volume input, but also utilizes the spatial relationship between different slices. This is achieved by the combination of graph and uncertainty modeling: each CT volume is represented as a directed graph with a fixed topology, where nodes are composed of the image features of slices with uncertainty, and the edges encode the relationship between slices to ensure that information about surrounding nodes is taken into account when updating graph nodes and classifying the graph. The introduction of uncertainty modeling for each node enables the model to make reliable decisions with limited training data through proactively focusing on key slices and symptoms with lesser uncertainty. Specifically, we apply Bayesian CNN to obtain the slice representation and corresponding uncertainty, which are then used to construct the graph, and re-weight the node features for information fusion during the graph classification.

The main contributions of our work may be summarized as follows:

- To the best of our knowledge, this is the first work to distinguish and classify UPPE, CPPE and normal cases using a deep learning method.
- We propose an uncertainty-guided graph attention network (UG-GAT) to capture and represent the spatial information and rich contextual information of the given 3D volumetric data, so as to improve the classification performance.
- We propose uncertainty measurement for the guidance of the model in order to reduce data noise and uncertainty, and further concentrate on the most salient slices and symptoms, which help the network to bypass the requirement for a large-scale annotated dataset.
2. Related Work

To our best knowledge, the existing works on PPE classification are mainly based on conventional methods (Kalkanis et al., 2017; Bugalho et al., 2014; Tsujimoto et al., 2015; Porcel et al., 2017), and deep learning-based methods for PPE classification is relatively unexploited. For example, Tsujimoto et al. (Tsujimoto et al., 2015) proposed to use the split pleura sign and the amount of pleural effusion on thoracic CT to differentiate complicated parapneumonic effusion from parapneumonic effusion. Porcel et al. (Porcel et al., 2017) designed a CT scoring system for discriminating between CPPE and UPPE based on the logistic regression. In the following subsections, we briefly review previous works on computer-aided methods for chest-related, and graph- and uncertainty- incorporated disease diagnosis, respectively.

2.1. CNN-based lung disease diagnosis

We have witnessed the rapid development of computer-aided methods for chest-related disease diagnosis, as evidenced by several reviews (Hu et al., 2018; Bhattacharya et al., 2021). Recently, CNN-based methods have been used to extract high-level features in a data-driven manner for lung disease diagnosis. For example, Shen et al. (2017b) presented a Multi-crop Convolutional Neural Network (MC-CNN) to automatically extract nodule information by employing a novel multi-crop pooling strategy to classify lung nodules by degree of suspected malignancy. Lyu and Ling (2018) developed a multi-level convolutional neural network (ML-CNN) to investigate the problem of lung nodule malignancy classification. Xie et al. (2019) proposed a semi-supervised adversarial classification (SSAC) model that can be trained using both labeled and unlabeled data for benign–malignant lung nodule classification. Xie et al. (2019) introduced a multi-view knowledge-based collaborative (MV-KBC) deep model to separate malignant from benign nodules using limited CT data of the chest.

However, these approaches make use of 2D CT slices for classification, overlook 3D spatial structure continuity, and are not able to make a case-based prediction. To this end, some recent studies proposed to utilize 3D volumetric data. Tafti et al. (2018) developed a multi-scale 3D CNN in an end-to-end manner for categorizing lung CT images into cancerous or non-cancerous groups. Alakwaa et al. (2017) first segmented lung regions and then fed them into 3D CNNs to classify the CT scans as positive or negative for lung cancer. Dey et al. (2018) applied several two-pathway 3D CNNs to perform the diagnostic classification between benign and malignant lung nodules in CT images. Ouyang et al. (2020) proposed a 3D CNN model to diagnose COVID-19 from CAP, in which a novel online attention module is combined with a dual-sampling strategy. The online attention module focuses on the infected regions when making diagnostic decisions. Compared with 2D CNNs, 3D CNNs require a large-scale annotated dataset to learn the representative appearance of a lesion, training a large number of parameters in order to obtain a relatively higher classification performance.

2.2. Graph neural network- and uncertainty-based disease diagnosis

Veličković et al. (2018) proposed the graph attention network (GAT) method, in which an attention mechanism is incorporated in order to attend selectively to neighboring vertices of each vertex. Recently, the GNN has been employed for the automated disease diagnosis due to its superiority in handling complex graphic data (Parisot et al., 2018; Zhang et al., 2019; Song et al., 2021). Wang et al. (2021) firstly applied multiple differently configured CNNs to extract features: a GCN is then utilized to fuse these features and create a relation-aware representation for the final Covid-19 classification. Liang et al. (2021) proposed a COVID-19 graph in a GCN to incorporate multiple datasets that differentiate COVID-19 infected cases from normal controls. These GNN-based methods use all subjects to construct the graph, where specific nodes on the graph represent imaging data for a subject and edges establish interactions between each pair of subjects. This indicates that the topology of the graph is limited by the number of subjects: if a new subject needs to be diagnosed, then the GNN needs to be retrained.
Fig. 2: Overall architecture of the proposed method, including graph construction phase and the uncertainty guided graph attention network. The Bayesian CNN is used to generate image representation vectors and corresponding uncertainty, which are used to obtain the graph representation and re-weight the node features for information fusion during the graph classification. The UG-GAT takes the graph and uncertainty as inputs, to learn the graph reasoning with the uncertainty-guided graph attentional convolution, in order to improve the classification accuracy of UPPE and CPPE.

Uncertainty measures can provide the human or autonomous model with valuable information to assess which subjects require further detailed inspection, and can help to reduce the risk of misclassification. Leibig et al. (2017) applied uncertainty obtained by a Bayesian CNN to refer a subset of difficult cases for further inspection, and improved the detection performance of diabetic retinopathy (DR) from fundus images. Ghoshal and Tucker (2020) also achieved performance improvement of COVID-19 classification from X-ray images by selecting training images according to uncertainty. Cicalese et al. (2020) used uncertainty to filter noisy data, and improved the performance of kidney kevel kupus nephritis classification. In medical image classification, uncertainty is mainly used in the selection of training samples to reduce training noise, so as to improve the classification accuracy.

3. Proposed Method

In this section, we show the details of the proposed method. It includes the graph construction phase and the uncertainty guided graph attention network. The overall architecture of the proposed method is shown in Fig. 2.

3.1. Graph construction

The precise distinguishing power of the graph relies on the accurate representation and inter-dependency characterization of image features (Parisot et al., 2018), and the CNNs may be used to extract representations from original CT images. The consideration and measure of uncertainty is essential for decision making: knowing the confidence with which we can trust the image representation will facilitate, and explain the subsequent graph classification. Consequently, we first apply a Bayesian CNN to obtain image features and their corresponding uncertainty for each CT image, these data are then used to construct the graph for each subject.

3.1.1. Bayesian convolutional neural networks

It is hard for CNNs to provide reliable prediction results, despite the fact that they have the ability to extract general data-driven highly expressive features (Shridhar et al., 2019).
A Bayesian neural network (BNN) can overcome this issue via probabilistic interpretation of model parameters. Apart from prediction uncertainty estimation, BNNs offer robustness to overfitting and can be efficiently trained on small data sets (Wang and Yeung, 2016; Wang et al., 2018). However, neural networks that apply Bayesian inference can be computationally expensive. While it can be approximated, we make use of the Bayesian equivalent of CNNs proposed in (Gal and Ghahramani, 2015) in this work to obtain the image representations and their corresponding uncertainty, and this is based on Monte Carlo Dropout (MC Dropout) method. MC Dropout is a powerful method used to measure uncertainty and has been employed extensively in many applications (Roy et al., 2019; Harper and Southern, 2020).

Given the training data \( \{x, y\} \), where \( x \) is a set of CT images with label \( y \in \{0, 1, 2\} \) and \( \{0, 1, 2\} \) represent UPPE, CPPE, and normal control, respectively, a dropout network \( \theta(w) \) can be trained, where \( w \) is a finite set of random variables that are determined by dropout. \( \{x^*, y^*\} \) is a new sample with corresponding prediction. Obtaining model uncertainty for a given image is as simple as keeping the dropout mechanism switched on at test time and performing multiple predictions. The width of the distribution of predictions is a reasonable proxy for the model uncertainty. Specifically, we approximate the integral by Monte Carlo sampling and compute the predictive mean as

\[
\mu_{\text{pred}} \approx \hat{\mu}_{\text{pred}} = \frac{1}{T} \sum_{t=1}^{T} p(y^* | x^*, \theta(\hat{w}_t)),
\]

where \( T \) is the total number of Monte Carlo samplings and \( \hat{w}_t \) represents different dropped unit states of the \( t_{th} \) sampling. \( \theta(\cdot) \) and \( p(\cdot) \) represent the parameters and the probability output of Bayesian CNN, respectively. The predictive standard deviation as a proxy for the uncertainty associated with this prediction is then defined as

\[
\sigma_{\text{pred}} \approx \hat{\sigma}_{\text{pred}} = \frac{1}{T - 1} \left[ \sum_{t=1}^{T} (p(y^* | x^*, \theta(\hat{w}_t)) - \hat{\mu}_{\text{pred}})^2 \right]^{1/2}
\]

We follow (Leibig et al., 2017) and set \( T = 100 \) to improve the quality of the uncertainty estimation.

3.1.2. Graph Representation

We apply ResNet (He et al., 2016) as Bayesian CNN backbone, and CT images and their corresponding labels to train a classification model to diagnose UPPE, CPPE and normal control cases at slice-level. Then we use the well-trained Bayesian CNN to extract all image features and their uncertainty for the graph construction. As shown in Fig. 2, we use the feature vector after the global average pooling (GAP) layer as the representation of each slice, where each node feature \( \vec{h} \) in the graph is a 512 \( \times \) 1 dimensional feature vector.

The graph structure provides a broader view field, filtering the value of the node features with respect to their neighbors rather than treating each feature in isolation. In order to accurately model the interaction between node features, the graph topology needs to be designed carefully. The graph topology of our method is shown as Fig. 3 (B). Every CT scan of a subject is represented as a directed graph: vertices are composed of slice features, and edges encode the spatial relationship between these slice features. We describe the graph as \( G = (V,A) \), where \( V \) contains \( N+1 \) vertices, \( V \in \mathbb{R}^{(N+1) \times F} \), and \( N \) is the number of CT slices per subject, \( F \) is the dimensionality of each node feature \( \vec{h} \). \( A \in \{0, 1\}^{(N+1) \times (N+1)} \) is a sparse adjacency matrix, representing the edge connections between vertices, where \( A_{i,j} = 1 \) means that vertex \( V_i \) connects to vertex \( V_j \) with an edge, and \( A_{i,j} = 0 \) that it does not. The graph nodes of two adjacent slices in the CT volume are connected to each other with directed edges, and each image node is connected to the center node by an edge. The blue and green nodes in Fig. 3 (B)
represent the slice features and the central node feature, respectively. In our implementation, due to the size of CT volume being 224 × 224 × 64 pixels (see Section 4.1 for details), the whole graph is composed of 64 slice nodes and 1 central node. Thus, there are 65 vertices and 127 edges (i.e., 63 bi-directed edges and 64 uni-directed edges) altogether for one graph.

The center vertex provides a central and effective mechanism to encode the sequence information of the CT volume effectively within the graph, and aids the graph neural networks in processing and integrating information from different slices during the graph classification. Following empirical testing, the feature of the center vertex is initialized using a vector of all zeros. Compared with an undirected graph, our proposed directed graph can ensure that the center node can take account of all the other nodes when updating vertex features, while the other nodes only consider their neighboring nodes, and are not affected by irrelevant nodes.

3.2. Uncertainty-guided graph attention network (UG-GAT)

After graph construction, the CT data of each subject is represented as a graph $G$, and the uncertainty of the graph nodes obtained by the Bayesian CNN is stored in a vector $\bar{U}$ with size of $(N + 1)$. $(N + 1)$ is the number of graph nodes. Since the central node is used to process and integrate information, we set its uncertainty to 0. In order to make use of uncertainty information to facilitate the decision making, we propose an uncertainty-guided graph attention network (UG-GAT). This differs from a graph attentional layer [Veličković et al., 2018], in that in addition to performing self-attention, we also weight each vertex with uncertainty when updating the node features.

Here we explain a single uncertainty-guided graph attentional layer. Taking a constructed graph $G = (V, A)$ which has a set of node features, $h = \{h_1, h_2, ..., h_{N+1}\}$, $h_i \in \mathbb{R}^F$ as input, the UG-GAT layer produces a new node feature, i.e., $h' = \{h_1', h_2', ..., h_{N+1}'\}$, $h_i' \in \mathbb{R}^{F'}$, and $F'$ is the dimension of the new feature at each node. This learning process is detailed as follows.

An initial step, a shared linear transformation parameterized by a weight matrix $W \in \mathbb{R}^{F' \times F}$, is applied to every node. Then attention coefficients $e_{ij}$, that indicate the importance of node $j$’s feature to that of node $i$, is computed by using a self-attention mechanism:

$$e_{ij} = \text{LeakyReLU}(\bar{a}^T [W h_i || W h_j]),$$  
where $||$ is the concatenation operation, and $\bar{a}^T \in \mathbb{R}^{2F'}$ a parameterized weight vector implemented by a fully connected layer with the LeakyReLU nonlinearity (with negative input slope $\alpha = 0.2$). And $E \in \mathbb{R}^{(N+1) \times (N+1)}$ is the attention coefficient matrix. Then the uncertainty matrix is used to re-weight the attention coefficients to ensure that the node features with large uncertainty have less impact on other nodes, while allowing the nodes with smaller uncertainty to play a greater role in the decision-making process. To achieve this, we first normalize the previously obtained uncertainty vector $\bar{U}$ to set its element values in between 0 and 1, as:

$$\bar{U}' = \frac{\bar{U} - \min(\bar{U})}{\max(\bar{U}) - \min(\bar{U}) + \delta},$$  
where $\delta$ is a small positive number, set to $1 \times 10^{-8}$ for stability, to avoid the case that the denominator is zero. $\min(\cdot)$ and $\max(\cdot)$ represent the maximum and minimum values of the vector, respectively. The reweighting operation is defined as:

$$E' = E \cdot \text{Diag}(1 - \bar{U}'),$$  
where $\text{Diag}(\cdot)$ is the diagonalization operation. After that, the softmax function is then utilized to normalize the coefficient $e_{ij}'$ to make it comparable across different nodes:

$$\alpha_{ij} = \text{Softmax}(e_{ij}') = \frac{\exp(e_{ij}')}{\sum_{k \in N_i} \exp(e_{ik}')},$$  
where $N_i$ is the neighborhood of node $i$ in the graph, which means that only $e_{ij}'$ for neighboring node $j \in N_i$ is considered to avoid involving any irrelevant node. The normalized attention coefficients $\alpha_{ij}$ are finally used to compute a weighted sum of the associated features, to serve as the final output features for every node:

$$\overline{h}_i = \text{ELU}(\sum_{j \in N_i} \alpha_{ij} W h_j),$$  
where $\text{ELU}$ represents the exponential linear unit (ELU) non-linearity. Following [Veličković et al., 2018], we employ multi-head attention to stabilize the learning process of self-attention.
To avoid performance deterioration due to the over-smoothing issue (Chen et al., 2020), our UG-GAT consists of three hidden layers, in which the first layer is an uncertainty-guided graph attentional layer and the others are graph attentional layers. Instead of performing averaging of all node features in the last layer, we take the features of the central node as the finally integrated feature: this is passed through a softmax activation layer for subject classification.

4. Experimental Results

4.1. Data description

In this study, we retrospectively reviewed all adult patients with a final diagnosis of PPE who underwent both a diagnostic thoracentesis and a chest CT, at Hwa Mei Hospital, University of Chinese Academy of Sciences, Ningbo, China from year 2018 to 2020. Finally, 198 met the criteria, namely 99 with UPPE and 99 with CPPE. There is no statistically significant difference between the ages of the UPPE and CPPE subjects ($P > 0.1$). Besides, 104 normal control cases were collected in this study. CT examinations of all the enrolled patients were performed on a ScintCare CT16 (Minfound Inc, China) with standard chest imaging protocols. The normal control is the subject without the presence of PPE, which was diagnosed using CT scans. Patients presenting with PPE will have a pleural effusion analysis to further determine if it is CPPE or UPPE: in order to discriminate CPPE from UPPE, generally accepted criteria were pleural fluid pH ($ < 7.20$) or pleural fluid glucose ($ < 600mg/L$) or pleural fluid LDH ($ > 1000U/L$) or culture evidence of pleural fluid microorganisms. The conducted case-control study was approved by Hwa Mei Hospital Review Board and adhered to the Declaration of Helsinki.

Each CT scan has a dimension of $512 \times 512$ pixels and the depth size varies from about 50 to 100 slices. To have the same input for 3D CNNs, we set the depth to a fixed value of 64 slices: if the CT scans had a depth of less than 64, we repeatedly added the last slice until the target depth was reached. For CT scans with depth greater than 64 slices, because the first and last slices may not contain useful lung regions, the middle 64 slices are taken as input. During training, we find that an input of size $512 \times 512 \times 64$ pixels for the selected comparison methods (Khawaldeh et al., 2018; Zhou et al., 2019) results in GPU memory overflow. Thus, we resize it to $224 \times 224$ pixels on the slice level, and then stack them into a 3D volume. Eventually, to ensure a fair comparison between different methods, we kept an input size of $224 \times 224 \times 64$ pixels for all our experiments with the same data augmentation methods (i.e., color jitter and random rotation) during the training stage.

The CT volumes of each subject are labeled as 0, 1 or 2, representing UPPE, CPPE, and normal control, respectively. We evaluated our UG-GAT in 5-fold cross-validation across all 302 cases for more reliable analysis over this relatively small dataset. We split these cases into 5 subsets, and kept the number of subjects in each subset similar. For each fold, we used 4 subsets for training and the remaining subset for testing. In addition, in order to train the Bayesian CNN, so as to obtain the slice representations of each subject and their corresponding uncertainty, a total of 2441 slices with lesion in UPPE and CPPE volumes (i.e., 1254 UPPE and 1187 CPPE) were labeled by a professional radiologist. We also selected 1590 normal control slices. Finally, a total of 4031 slices were used to train the Bayesian CNN. Notably, these slices were not from the 302 candidates used to evaluate UG-GAT, which ensured that the data were not seen during the model training phase when calculating uncertainty.

4.2. Baseline and implementation details

The Bayesian CNN was trained to minimize cross-entropy loss using an Adam optimizer with an initial learning rate of 0.0001. We utilized a batch size of 4, and a maximum epoch of 500. In addition, online data enhancement was employed to enlarge the training dataset. The first layer of UG-GAT has 4 attention head computing $F^* = 256$ dimensional features. The last two layers have a single attention head that computes 128 dimensional features. We trained UG-GAT using an Adam optimizer for 100 epochs with an initial learning rate of 1e-7. We implemented the proposed method in Pytorch on a PC with two GPUs (NVIDIA GeForce GTX 1080 Ti 11GB). We trained the Bayesian CNN and UG-GAT separately. Specifically, we firstly
trained the Bayesian CNN, which was then used to produce slice features and corresponding uncertainty. After that, constructed graph was used to train the UG-GAT.

To evaluate the proposed method, a comparison was made between the UG-GAT and several approaches previously proposed for medical image classification using 3D data. Most of these methods [Khawaldeh et al., 2018; Chang et al., 2018; Zhou et al., 2019; Liu et al., 2020; Zunair et al., 2020] took 3D volumetric data as input and apply 3D CNNs with different architectures to extract features for classification. [Hao et al., 2021] firstly extracted features using CNN and then utilized ConvLSTM to fuse the information and obtain the probability of different categories.

4.3. Evaluation Metrics

Following commonly-used metrics in the classification task, the weighted sensitivity (SEN), specificity (SPE), and accuracy (ACC) were calculated. These metrics are defined as:

$$\text{SEN} = \sum_{i=1}^{N_c} w_i \frac{TP_i}{TP_i + FN_i}, \quad \text{SPE} = \sum_{i=1}^{N_c} w_i \frac{TN_i}{TN_i + FP_i},$$

$$\text{ACC} = \sum_{i=1}^{N_c} w_i \frac{TN_i + TP_i}{TN_i + FP_i + FN_i + TP_i},$$

where $TP_i$, $TN_i$, $FP_i$, and $FN_i$ denote, respectively, the true positive, true negative, false positive, and false negative values for the $i$-th category; and $w_i$ represents the percentage of cases whose ground truth labels are $i$. $N_c$ denotes the number of total classes.

In order to reflect the trade-off between sensitivity and specificity, and evaluate the quality of our classification results more reliably, F1-score (F1) was also calculated:

$$\text{PRE} = \sum_{i=1}^{N_c} w_i \frac{TP_i}{TP_i + FP_i}, \quad F1 = 2 \cdot \frac{\text{PRE} \cdot \text{SEN}}{\text{PRE} + \text{SEN}}.$$ 

We also calculated the area under receiver operating characteristic curve (AUC) in classifying UPPE and CPPE. The AUC is calculated based on all possible pairs of SEN and 1-SPE obtained by changing the threshold performed on the classification scores yielded by the trained networks.

4.4. Results

In this section, extensive experiments were conducted to evaluate the performance of the proposed method. Firstly, we compare our method with several case-based medical image classification approaches [Khawaldeh et al., 2018; Chang et al., 2018; Zhou et al., 2019; Liu et al., 2020; Zunair et al., 2020; Hao et al., 2021]. Briefly, Khawaldeh et al. (2018), Chang et al. (2018) and Zhou et al. (2019) used 3D volumetric data as input and employed 3D AlexNet, 3D ResNet and 3D SE-DenseNet to extract features for classification, respectively. Liu et al. (2020) and Zunair et al. (2020) specifically designed 3D CNN models for 3D data classification. Hao et al. (2021) first extracted features using CNN, then utilized ConvLSTM to fuse the information and finally obtained probability of different categories. In addition, we performed ablation studies of the proposed method to analyze the impact of uncertainty, graph construction strategy and size of input on its performance.

4.4.1. Diagnostic performance

We first performed the experiments for the classification of UPPE, CPPE and normal control. Table 1 shows the results obtained by the competing methods and our UG-GAT. Our method has the lowest number of parameters and achieves the highest scores in all the metrics. Several observations can be summarized from Table 1: 1) The specifically designed 3D CNNs (i.e., [Liu et al., 2020; Zunair et al., 2020]) outperform the models directly converted from 2D CNNs (i.e., [Khawaldeh et al., 2018; Chang et al., 2018]) due to parametric optimization. Although the 3D SE-DenseNet model applied an attention mechanism, it does not produce as good results as the original technique due to overfitting caused by the small amount of data in our experiments. This shows that 3D CNNs rely heavily on large-scale datasets, which limit their application in many medical imaging areas. 2) The 2D CNN + ConvLSTM approach [Hao et al., 2021] yields better classification results than 3D CNNs-based methods. The main reason is likely to be that ConvLSTM is less dependent on the amount of training data while utilizing the spatial information of the CT volume compared to 3D CNNs, which makes the generalization ability of
Table 1: Classification results of UPPE, CPPE and Normal using different methods. The input size for all the methods in the experiments was 224 × 224 × 64.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Params</th>
<th>Year</th>
<th>SEN</th>
<th>SPE</th>
<th>ACC</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khawaldeh et al. (3D AlexNet)</td>
<td>390.01M</td>
<td>2018</td>
<td>0.7093</td>
<td>0.7113</td>
<td>0.7119</td>
<td>0.6177</td>
</tr>
<tr>
<td>Chang et al. (3D ResNet)</td>
<td>46.06M</td>
<td>2018</td>
<td>0.7398</td>
<td>0.7412</td>
<td>0.7417</td>
<td>0.6542</td>
</tr>
<tr>
<td>Zhou et al. (3D SE-DenseNet)</td>
<td>473.26M</td>
<td>2019</td>
<td>0.7191</td>
<td>0.7212</td>
<td>0.7219</td>
<td>0.6288</td>
</tr>
<tr>
<td>Liu et al. (3D CNN)</td>
<td>33.09M</td>
<td>2020</td>
<td>0.7650</td>
<td>0.7674</td>
<td>0.7682</td>
<td>0.6807</td>
</tr>
<tr>
<td>Zunair et al. (3D CNN)</td>
<td>34.20M</td>
<td>2020</td>
<td>0.7512</td>
<td>0.7540</td>
<td>0.7550</td>
<td>0.6646</td>
</tr>
<tr>
<td>Hao et al. (2D CNN + ConvLSTM)</td>
<td>40.67M</td>
<td>2021</td>
<td>0.7983</td>
<td>0.8006</td>
<td>0.8013</td>
<td>0.7242</td>
</tr>
<tr>
<td>Ours (Bayesian CNN + UG-GAT)</td>
<td>3.95M*</td>
<td>-</td>
<td>0.8889</td>
<td>0.8903</td>
<td>0.8907</td>
<td>0.8428</td>
</tr>
</tbody>
</table>

* The parameter number of our method includes two parts: Bayesian CNN (3.16M) and UG-GAT (0.79M).

Fig. 4: Confusion matrices of different methods. (A) - (F) are the results of Khawaldeh et al. [2018], Chang et al. [2018], Zhou et al. [2019], Zunair et al. [2020], Hao et al. [2021] and ours, respectively. The numbers in the confusion matrices denote the percentage of the predicted class.

the model stronger. 3) Compared with the other methods, our proposed graph-based method demonstrates outstanding performance across all the metrics. For example, our model exhibits a large advantage over 2D CNN + ConvLSTM [Hao et al., 2021] by increases of ACC and F1-score of about 0.0894 and 0.1186, respectively. Fig. 4 shows the confusion matrices of different methods, and these results further imply the superiority of our approach. Unlike those evaluation metrics in Table 1 that show the average values, Fig. 4 clearly shows the classification performance of the different methods in each category. It can be seen that our method obtains relative high accuracy in each category, especially in UPPE and CPPE, which indicates that our model has stronger discriminating ability for different categories. In addition, as shown in Table 1, our model also requires much less parameter size when compared with other state-of-the-art methods. That is because compared to other methods, our model does not use any 3-dimensional convolution operations, and our method is based on GNN, which is an parameter efficient method. In consequence, our method can achieve superior performance with a smaller number of parameters.

As noted in Section 1, the difficulty of PPE diagnosis is the differentiation between UPPE and CPPE. Accordingly, we conducted experiments on their direct classification. Results are
Table 2: Classification results of different methods in classifying UPPE and CPPE. The input size for all the methods in the experiments was $224 \times 224 \times 64$. P-value is calculated by Delong’s test.

<table>
<thead>
<tr>
<th>Method</th>
<th>SEN</th>
<th>SPE</th>
<th>ACC</th>
<th>F1-score</th>
<th>AUC (P-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khawaldeh et al. (2018)</td>
<td>0.5051</td>
<td>0.9091</td>
<td>0.7071</td>
<td>0.6329</td>
<td>0.7346 ($P &lt; 0.001$)</td>
</tr>
<tr>
<td>Chang et al. (2018)</td>
<td>0.6162</td>
<td>0.8587</td>
<td>0.7374</td>
<td>0.7011</td>
<td>0.7861 ($P &lt; 0.001$)</td>
</tr>
<tr>
<td>Zhou et al. (2019)</td>
<td>0.6263</td>
<td>0.7980</td>
<td>0.7121</td>
<td>0.6851</td>
<td>0.7565 ($P &lt; 0.001$)</td>
</tr>
<tr>
<td>Liu et al. (2020)</td>
<td><strong>0.8283</strong></td>
<td>0.6466</td>
<td>0.7374</td>
<td>0.7593</td>
<td>0.7592 ($P &lt; 0.001$)</td>
</tr>
<tr>
<td>Zunair et al. (2020)</td>
<td>0.5758</td>
<td>0.9196</td>
<td>0.7677</td>
<td>0.7125</td>
<td>0.8117 ($P = 0.0024$)</td>
</tr>
<tr>
<td>Hao et al. (2021)</td>
<td>0.6667</td>
<td>0.9293</td>
<td>0.7980</td>
<td>0.7674</td>
<td>0.8585 ($P = 0.0122$)</td>
</tr>
<tr>
<td>Ours</td>
<td>0.8082</td>
<td><strong>0.9331</strong></td>
<td><strong>0.8688</strong></td>
<td><strong>0.8602</strong></td>
<td><strong>0.9096</strong></td>
</tr>
</tbody>
</table>

Fig. 5: ROC curve and AUC of different approaches in classifying UPPE and CPPE.

Even though our method produced a slightly lower SEN than the method in (Liu et al., 2020) by 0.0201, it achieved the best performance in terms of SPE, ACC, F1-score and AUC, which are 0.0708, 0.0928 and 0.0511 higher than those of the second best 2D CNN + ConvLSTM method in (Hao et al., 2021), respectively.

4.4.2. Importance of uncertainty

We argue that the uncertainty of the image representation is of great significance for graph classification, reducing the impact of unexpressive representations and making predictions more reliable. Fig. 6 shows the distribution of prediction uncertainty values for several slices of UPPE and CPPE. The class with the highest softmax output of the mean of the predictive distribution is considered as the prediction, and the predictive entropy of the output distribution is used as its estimated uncertainty (computed by Equation 2). Depending on the input image, the Bayesian model can determine its decision with high or low confidence, represented by the predictive posterior distribution. The wider the posterior distribution of the output, the lower the confidence of the model in its prediction. This is because the uncertainty in the weight space captured by the posterior has been incorporated into the predictive uncertainty, giving us a way for the model to quantify the confidence.

In this group of experiments, we evaluated the effectiveness of uncertainty as well as the effect of different uncertainty quantification methods on the classification results of UPPE and CPPE. Specifically, after having constructed the graph using the acquired image representation, experiments were firstly conducted without using the uncertainty by replacing the uncertainty-guided graph attentional layer with a normal graph attentional layer. In addition, two different uncertainty quantification methods (i.e., MC Dropout (Gal and Ghahramani, 2015) and Deep Ensembles (Lakshminarayanan et al., 2017)) were implemented to verify their impact on the performance of the proposed method. As shown in Table 3, we can see that our uncertainty-guided graph attentional layer effectively improved the classification performance. This implies that uncertainty information plays a key role in the process of feature updating and fusion, obtaining a better diagnostic performance. Intuitively, the graph neural network is effective in characterizing the feature interactions between different slices and modelling the distribution of lesions. Uncertainty information can assist the GNN to select key slices and symptoms, producing a model with more attention to nodes with low uncertainty when making decisions, thus increasing the reliability of the decision. Moreover, it was found that the UG-GAT is not
Table 3: PPE prediction results of the proposed graph classification model without and with uncertainty. We also compare the effect of different uncertainty quantification methods on the classification results.

<table>
<thead>
<tr>
<th>Methods</th>
<th>SEN</th>
<th>SPE</th>
<th>ACC</th>
<th>F1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without uncertainty</td>
<td>0.6566</td>
<td>0.7778</td>
<td>0.7172</td>
<td>0.6989</td>
<td>0.7860</td>
</tr>
<tr>
<td>Deep Ensembles</td>
<td>0.8080</td>
<td>0.9293</td>
<td>0.8636</td>
<td>0.8556</td>
<td>0.9045</td>
</tr>
<tr>
<td>MC Dropout</td>
<td>0.8082</td>
<td>0.9331</td>
<td>0.8688</td>
<td>0.8602</td>
<td>0.9096</td>
</tr>
</tbody>
</table>

very sensitive to the used of different uncertainty quantification methods.

4.4.3. Graph construction strategy

As introduced in Section 3.1.2, the graph structure is expected to have a strong impact on graph classification performance. Therefore, we investigate the effect of different graph structures (i.e., undirected graph and directed graph) on diagnostic accuracy in cases of UPPE and CPPE. The corresponding experimental results are presented in Fig. 7, from which we may observe that the undirected graph performed worse than the directed one, with decrements of 0.0859, 0.0403 and 0.1208 in ACC, AUC and F1-score, respectively. This shows that the use of a directed graph ensures that all surrounding nodes can influence the central node when updating features, while other nodes consider neighboring nodes only, and are not influenced by irrelevant nodes, allowing for better integration of information from different slices.

4.4.4. Influence of the size of the image

In this group of experiments, we investigate the influence of the size $S \times S \times 64$ pixels of the CT images on the classification performance of UPPE and CPPE achieved by our method, where $S$ is the size of the single slice. Specifically, we incre-
mentally selected $S$ as 128, 224, 448, 512 pixels and recorded the corresponding results. The classification performance quantified in SEN, SPE, ACC, AUC and F1-score are summarized in Fig. 8. Generally, we can see that the values of all metrics are increased when changing $S$ from 128 to 512. For example, we have $\text{ACC} = 0.8333$, $\text{AUC} = 0.8811$ and $\text{F1-score} = 0.8156$ when $S = 128$, while $\text{ACC} = 0.8687$, $\text{AUC} = 0.9193$ and $\text{F1-score} = 0.8571$ when $S = 512$. This implies that the larger the size and the higher the resolution of the image, the more easily the lesion information contained is captured by the model, which is beneficial for classification tasks.

### 4.4.5. Effectiveness of parameter $T$

We further investigated the sensitivity of hyper-parameter to the entire model: the number of Monte Carlo samplings $T$. We selected $T$ from [15, 30, 50, 70, 90, 100, 120] when calculating the uncertainty, and reported the corresponding results. As can be seen from Table 4, our method is not very sensitive to the number of Monte Carlo samplings in a wide range (i.e., 50 - 120). The performance of model using smaller $T$ (i.e., $T = 15$) is slightly decreased, this scenario is consistent with the conclusion of published work, larger $T$ can improve the quality of uncertainty estimation, thus reducing the noise when training the UG-GAT model and improving the classification performance.

### 5. Discussion and Conclusion

#### 5.1. Limitations

Although our approach achieves the best results in automatic classification of the CPPE, UPPE and normal control, there are still some limitations in this work. Firstly, the dataset we used is relatively small for deep neural network to learn deep-level and discriminative features, and the model is not evaluated on an external dataset. To our best knowledge, there is no publicly available dataset for PPE classification task, and collecting data from multiple centers that meet the requirements is difficult due to various conditions. It is our intention to evaluate the performance of our model on external datasets in future. Secondly, we applied ResNet34 as the backbone of Bayesian CNN, which can be replaced by any existing state-of-the-art CNN feature extractor. We believe that careful selection of the extractor may further improve classification performance, however, due to it is not the main focus of this work, no additional experiments were performed to compare the influence of different CNN backbone.

#### 5.2. Analysis on failure cases

We show some failure cases of CPPE and UPPE classifications by our method in Fig. 9. According to the radiologist’s analysis, these are difficult samples due to relatively small lesions, which are not clearly characterized, and it is indeed difficult to distinguish them by CT alone, and a combined analysis
by other image modalities (e.g. ultrasound images) may be able to identify them more accurately.

5.3. Conclusion

The identification of UPPE and CPPE in their early stages is of great importance for clinicians in developing personalized treatment plans and reducing the mortality rate in patients with a parapneumonic effusion. Although artificial intelligence (AI) shows increasingly successful applications in various medical image analysis tasks, there is as yet no application in PPE diagnosis yet. In this study, we proposed a novel graph-based approach for parapneumonic effusion diagnosis in CT scans.

In conventional volumetric medical image classification approaches, using 2D convolutional neural networks (CNNs) is common to deal with the individual slices. This deliberately discards the spatial information and results in poor performance of the intended task. While the prediction results are slice-based, case-based prediction results are more valuable in clinical settings. 3D CNNs can compensate for the shortcomings of 2D CNNs, both by exploiting the spatial distribution information of lesions and by producing case-based prediction results. However, 3D CNNs impose a heavy memory and computational cost, and training a model with strong generalizability requires large-scale datasets, which are usually difficult to collect. These shortcomings limit their application in the real world.

Graphs, by contrast, provide a powerful and intuitive way of modelling volumetric data and associations between different slices. Our proposed method characterizes uncertainty information through Bayesian CNNs, and demonstrated a significant performance without dependence on a large amount of data. We formalize the medical image differentiation as a graph classification problem. Each subject is represented as a directed graph with a fixed topology, where vertices represent image slices, edges encode and propagate the spatial adjacency relationship and information between these slices, and the center node integrates dynamically all the information from the whole CT volume for final classification. The graph neural network is powerful in capturing the feature associations between different slices and in modeling effectively the distribution of lesions. Uncertainty information is essential in assisting the GNN in selecting and focusing on key slices and symptoms, making the model pay more attention to nodes with low uncertainty when making final decisions, thus increasing the reliability of the decision. This also reduces reliance on large-scale datasets, allowing us to obtain models with good robustness and accuracy using only a small number of training samples.

Our method overcomes the heavy memory and computational requirements of 3D CNNs and also leverages the depth information of CT volumes by comparison to 2D CNNs. In order to improve the accuracy of graph classification, we designed UG-GAT to exploit the uncertainty of image features to reduce impact of unexpressive representations and make the predictions more reliable. The major advantage of our proposed graph-based method for volumetric medical data classification...
is that a robust and powerful model can be obtained using a small amount of training data. In this work, we used a small dataset containing 302 subjects (99 UPPE, 99 CPPE and 104 normal control cases) but achieved an ACC and F1-score of 0.8907 and 0.8428, respectively.

Although we only evaluated our method on the CT dataset of PPE, it can be adapted to any other 3D medical image classification problems, such as benign-malignant lung nodule classification from chest CT, and the identification of Alzheimer’s disease. Lastly, our method can also be used to construct graphs using both image data and non-image data, to further improve the accuracy of diagnosis.

6. Acknowledgement

This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China (LZ19F010001 and LY19H120001), in part by the Youth Innovation Promotion Association CAS (2021298), in part by the Key Research and Development Program of Zhejiang Province (2020C03036), in part by the Ningbo 2025 S&T Megaprojects (2019B10033 and 2019B1006), in part by the Ningbo Natural Science Foundation (202002N7034 and 2019A61053).

References


