A 2.5D Deep Learning-Based Method for Drowning Diagnosis Using Post-Mortem Computed Tomography

Yuwen Zeng, *Student Member, IEEE*, Xiaoyong Zhang, *Senior Member, IEEE*, Yusuke Kawasumi, Akihito Usui, Kei Ichiji, *Member, IEEE*, Masato Funayama, and Noriyasu Homma, *Member, IEEE*

***Abstract*—It is challenging to diagnose drowning in autopsy even with the help of post-mortem multi-slice computed tomography (MSCT) due to the complex pathophysiology and the shortage of forensic specialists equipped with radiology knowledge. Therefore, a computer-aided diagnosis (CAD) system was developed to help with diagnosis. Most deep learning-based CAD systems only utilize 2D information, which is proper for 2D data such as chest X-ray images. However, 3D information should also be considered for 3D data like CT. Conventional 3D methods require a huge amount of data and computational cost when using 3D methods. In this article, we proposed a 2.5D method that converts 3D data into 2D images to train 2D deep learning models for drowning diagnosis. The key point of this 2.5D method is that it uses a subset to represent the whole case, covering this case as much as possible while avoiding other repetitive information. To evaluate the effectiveness of the proposed method, conventional 2D, previous 2.5D, and 3D deep learning-based methods were tested using an MSCT dataset obtained from Tohoku university. Then, to provide explainable diagnosis results, a visualization method called Gradient-weighted Class Activation Mapping was employed to visualize features relevant to drowning in CT images. Results on drowning diagnosis showed that our proposed method achieved the best performance compared to other 2D, 2.5D, and 3D methods. The visual assessment also demonstrated that our method could find the saliency regions corresponding to drowning.**

***Index Terms*—Computed tomography, drowning, deep learning, computer-aided diagnosis, explainability.**

Manuscript received 26 April 2022; revised 22 August 2022 and 17 October 2022; accepted 20 November 2022. Date of publication 29 November 2022; date of current version 6 February 2023. This work was supported by the Autopsy Imaging Center, Tohoku University Graduate School of Medicine, and JSPS KAKENHI under Grants JP18K19892, JP19H04479, and JP20K08012. *(Corresponding author: Yuwen Zeng.)*

Yuwen Zeng is with the Tohoku University Graduate School of Biomedical Engineering, Sendai 980-8575, Japan (e-mail: zeng.yuwen.s4@dc.tohoku.ac.jp).

Xiaoyong Zhang is with the National Institute of Technology, Sendai College, Sendai 989-3128, Japan, and also with the Institute of Development, Aging and Cancer, Tohoku University, Sendai 980-8575, Japan (e-mail: xiaoyong@ieee.org).

Yusuke Kawasumi, Akihito Usui, Kei Ichiji, Masato Funayama, and Noriyasu Homma are with the Tohoku University Graduate School of Medicine, Sendai 980-8575, Japan (e-mail: kawasumi@rii.med.tohoku.ac.jp; t7402r0506@med.tohoku.ac.jp; ichiji@tohoku.ac.jp; funayama@forensic.med.tohoku.ac.jp; homma@ieee.org).

Digital Object Identifier 10.1109/JBHI.2022.3225416

I. Introduction

ACCORDING to a report from World Health Organization, drowning is the third leading cause of unintentional death worldwide [1]. In forensic medicine, diagnosing via autopsy is challenging due to its non-specific pathophysiology [2]. Meanwhile, on routine autopsy, the macroscopical appearance of the cut surface of the drowned lung may be light red with small amounts of blood and large quantities of edema fluid. Such gross observation has limitations in the evaluation of pulmonary edema, but autopsy imaging like computed tomography (CT), can give reliable pulmonary information noninvasively [3]. Common features such as ground-glass opacity with thickened pulmonary interstitium can be found in CT, however, these findings are not specific for drowning either [3], [4]. Considering the challenges mentioned above and the shortage of forensic specialists who are also equipped with radiology knowledge, computer-aided diagnosis (CAD) systems can be developed to provide reference information.

It is hard to express lesions quantitatively by handcrafted features when classifying a medical image using a CAD system. Deep learning can extract representative features automatically through the training process. Thus, we can use those features for image classification. Our previous work on drowning diagnosis using post-mortem lung CT images showed the feasibility of the deep learning-based CAD system [5]. Then we further improved the CAD system with higher accuracy and visually explainable information [6]. These studies conducted slice-wise classification by training 2D-DCNNs with every single slice of each case, then calculated a case-wise result by averaging all slice-wise results. Such calculation was unsuitable for 3D data because information along the longitudinal axis was not included. Thus, training 3D models with 3D data might be a solution. Some studies have shown the feasibility of 3D-DCNNs in medical imaging, such as the glaucoma detection with 3D optical coherence tomography [7], or classifying liver tumors by using 3D magnetic resonance imaging (MRI) [8]. However, a 3D-DCNN has several times more parameters than its 2D-DCNN, which is also challenging due to the small-scale dataset and expensive computational cost.

2.5D methods that reshape 3D data into 2D data and use the reshaped data to train 2D-DCNNs were developed to find a compromise between computational cost and the performance

of models. There are already some works using 2.5D methods for medical image analysis. A study [9] converted 3D digital subtraction angiography data into 2D images and trained a 2D model for intracranial aneurysms classification. Alkadi et al. [10] used a 2.5D method for prostate cancer detection on MRI by putting every three consecutive slices of the input volume to the RGB dimension. A 2.5D lymph node detection method [11] was proposed by re-sampling the 3D volume of interest (VOI) centroid into 2D orthogonal views. The re-sampled views were then used to train a 2D-DCNN. These works obtained good classification results but have yet to go further on explainability. Due to the underlying black-box nature of the deep learning method, there needs to be more explicitly representing the knowledge for a given classification task. The lack of inspecting the behavior of models affects the use of deep learning in all domains, especially medical image diagnosis, where explainability and reliability are the key elements for trust by the end-user [12]. For example, Shi et al. [13] proposed an explainable attention-transfer classification model for COVID-19 automatic diagnosis to generate more understandable results. Instead of giving simple positive or negative classification results, explainable visualizations can show why the model gave such results and therefore make the model more reliable.

In this study, we developed an explainable 2.5D method to utilize 3D information of multi-slice CT (MSCT), which can provide case-wise classification rather than slice-wise. To evaluate the explainability of our method, Gradient-weighted Class Activation Mapping (Grad-CAM). [14] was employed to produce a coarse localization map highlighting important regions in the image for predicting a concept (such as drowning or non-drowning). Experiment results showed that clearer and more consistent attention could be achieved using the proposed method. We also discussed the coherency between the attention of models and the judgment of forensic specialists. The contributions of this study are:

1) This work presents the first case-wise explainable CAD system for drowning diagnosis.
2) We proposed a 2.5D method to convert 3D MSCT data into 2D images, which can be used to train 2D deep learning models. Considering the small amount of available data and the computational cost of 3D models, our study has provided a competitive way to deal with 3D medical data in deep learning.
3) Compared to most studies that only presented the development of CAD systems, we also evaluated its explainability by checking the correspondence between the attention of models and the judgement of forensic specialists.

II. Data and Methods

A. Dataset and Pre-Processing

Post-mortem MSCT scanning as part of the pre-autopsy screening was performed on an eight-channel scanner (Aquilion: Toshiba Medical Systems, Japan). We obtained axial conventional scan images of the chest with a protocol of 135 kVp, 190–250 mAs, M-sized field of view (FOV), and 1.0 mm slices

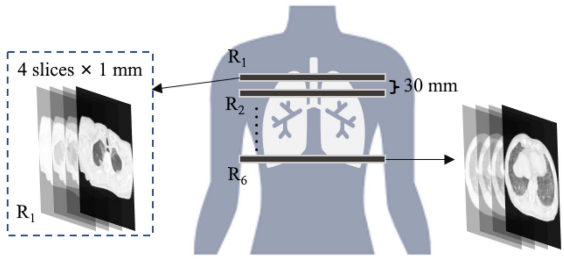
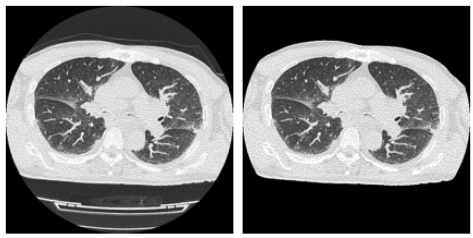


Fig. 1. The form of four-row MSCT image data for lung areas. Six or seven discrete regions with 30 mm intervals were obtained for drowning diagnosis. Each region consisted of four consecutive slices with 1 mm thickness. Here is an example with six regions (denoted by R_1-R_6).

TABLE I
Total Numbers of Different Cases for Each Class

Class	24-slice	28-slice
Non-drowning	40	113
Drowning	51	109



(a) original image (b) pre-processed image

Fig. 2. An original image (a) and the pre-processed image (b).

(size 512 × 512 pixels) every 30 mm through the chest in four-row multi-slice mode and processed with lung kernel settings. Every four slices compose a region, and there are six(24 slices) or seven regions (28 slices) in a case due to individual differences, such as stature. Fig. 1 shows an example of a 24-slice case. Though the MSCT data is not continuous and different from the helical CT, it is sufficient for forensic radiologists and used as a general pre-autopsy scanning. In total, we obtained 313 cases, including 153 for non-drowning and 160 for drowning, as listed in Table I. The test set consisted of 50 cases, including 25 for non-drowning and 25 for drowning. Each class had four cases with 24 slices and 21 cases with 28 slice case. We randomly split 15% of the training set into validation set, which would not be used for training or testing.

Pre-processing was applied to remove the background. Due to post-mortem factors such as putrefaction and rigor mortis, we segmented the body as foreground instead of the thoracic cavity. As shown in Fig. 2, given an original image (a), we first divided the image into foreground and background based on Otsu's method [15]. Then morphological processing was conducted to remove bags that covered the body and the bed under the body. Finally, we obtained the pre-processed image (b).

All cases used in this study had then undergone autopsy by a forensic specialist having autopsy experience for more than 30 years. The ground truth (drowning or non-drowning) was given

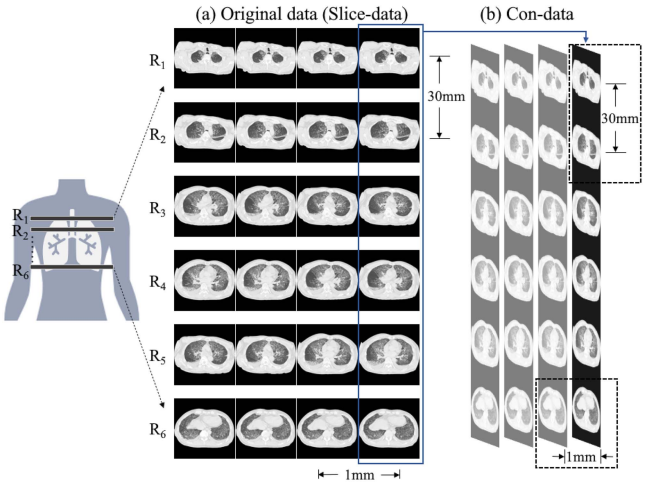


Fig. 3. Examples of 2.5D data generation. (a) An original CT case with 24 slices (six regions denoted by R_1, R_2,..., R_6, and four slices in each region). (b) Slices were selected in an equally-spaced way and concatenated into a new image named Con-data so that a 2D image could represent information about this case.

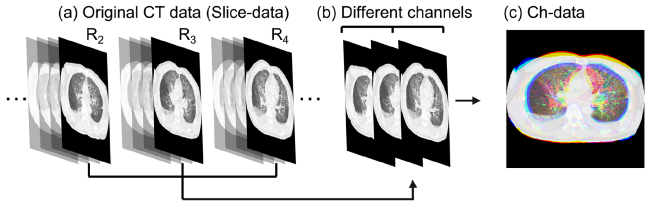


Fig. 4. An example of Ch-3 [10]. Three slices from consecutive regions (e.g., R_2, R_3, and R_4) of an original CT case (a) were put into different channels (b) and formed as a new color image (Ch-3) (c).

TABLE II
SHAPES AND THE NUMBER OF DIFFERENT DATA PER CASE

Data	Type	Shape H × W × C	Number per case (n)	
			24-slice	28-slice
Slice-data	2D	512 × 512 × 1	24	28
Ch-3 [10]	2.5D	512 × 512 × 3	16	20
Ch-6	2.5D	512 × 512 × 6	4	8
Con-data	2.5D	3072 × 512 × 1	4	8
3D-6	3D	512 × 512 × 6	4	8
3D-24	3D	512 × 512 × 24	1	2

based on a comprehensive judgment, including the pre-autopsy screening, the investigation of the death scene by the police, and medicolegal autopsies including drug screens and diatom tests. The use of MSCT images for this study was approved by the ethics board of Tohoku University (protocol number: 2021-1-495; date: 2018-09-18). Informed consent was not required for this research.

B. Generation of 2.5D Data

A single slice only provides information in the transverse plane, so 3D anatomical information was lost during the training process, which led to unreliable results. On the other hand, 3D deep learning-based methods have been widely used for 3D medical data like CT or MRI. Nevertheless, 3D-DCNNs often have much more parameters to be trained, and the small-scale data and expensive computational cost generally restrict their performance. Therefore, we proposed a 2.5D method that converts 3D data into 2D images, named Con-data, and trained them on 2D-DCNNs.

This method is context-aware as it can provide information in the transverse, coronal, and sagittal planes, making it possible to utilize all spatial information on 2D models. As shown in Fig. 3, (a) is an example of a 24-slice MSCT that contains six regions. Six slices of 512 × 512 pixels were selected out from each region with the same interval and concatenated into a new Con-data (b) of 3072 × 512 pixels. In addition, the receptive field of neurons of DCNNs ensures that convolutional filters produce a strong response to local input patterns. This means DCNNs are not sensitive to the absolute position in the field. Therefore, the way of concatenating images does not affect the results. Four subsets can be obtained for this case as shown in (b). For a 28-slice case that contains seven regions, we applied the same procedure in regions R_1-R_6, and regions R_2-R_7, so eight subsets would be generated.

There are two noteworthy points here. First, we selected 'six slices' here to cover the whole case (R_1-R_6) as much as possible while avoiding repetitive information (consecutive four slices in the same region). Second, we did not directly concatenate seven slices for 28-slice cases because to train both 24-slice and 28-slice cases together, we have to keep the input shape the same. If we resize an image from 3584 × 512 (R_1-R_7) into 3072 × 512, then there would be severe longitudinal deformation that makes the lung appear squashed. Furthermore, our data-generation method could be a flexible way to deal with other possible data (e.g., a 32-slice case) in the future.

For better evaluation, we also adopted another type of 2.5D data mentioned in a related study [10], named Ch-3, as shown in Fig. 4. Ch-3 (size 512 × 512 × 3) is constructed by rearranging three different MSCT slices into RGB channels (three primary colors: red, green, and blue) of a color image. To make a fair comparison with Con-data, we also extended Ch-3 into Ch-6 with size 512 × 512 × 6. For a Ch-3 image, the same as how we picked out slices for Con-data, taking a 24-slice case as an example, we took three slices from regions R_1-R_3, then R_2-R_4, R_3-R_5, and R_4-R_6, and finally we would obtain 16 subsets from this case.

To show the superiority of our 2.5D method, we carried out experiments using conventional 2D and 3D methods. The conventional 2D method is to train 2D models with every single slice (named Slice-data), as was done in previous works [5], [6]. The 3D method is to train 3D models with 3D data, named 3D-6 and 3D-24. The 3D-6 used the same method to pick out six slices from each case and reformed them into a sub-volume so as to make it comparable with Con-data. The 3D-24 used all 24 slices for a 24-slice case, or the first 24 slices and the last 24 slices for a 28-slice case.

Shapes in Height × Width × Channel (H × W × C) and the number (n) of generated data per case are shown in Table II. Here the number n can also represent how much the generated data

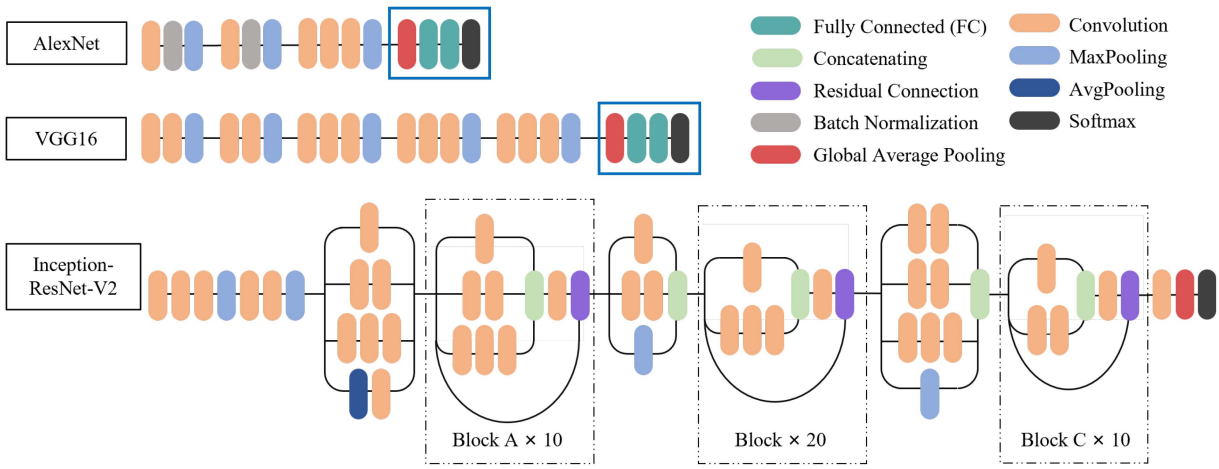


Fig. 5. Schematic diagrams of AlexNet, VGG16, and InceptionResNetV2 used in this study. Blue boxes mark out the modified layers. 2D and 3D versions of AlexNet and VGG16 have the same architecture except for different operators and input dimensions.

covers the overall case. The smaller the n is, the more regions it covers. $n = 1$ means that the data is the volumetric case itself.

C. DCNN Models

For all the 2D and 2.5D data, we adopted three 2D-DCNNs named AlexNet [20], VGG16 [21], and Inception-ResNet-V2 [22] (InResV2) in this study. AlexNet and VGG16 were used in previous studies [5], [6]. Meanwhile, to test whether better performance can be obtained from a deeper model, we chose the more advanced InResV2 as it combines the characteristics of InceptionNet and ResNet, with higher performance but lower computational cost. The original AlexNet and VGG16 have more than 61 million (M) and 138 M of parameters, which would lead to overfitting easily on our small-scale dataset. InResV2 has over 460 layers, 20 times more than VGG16, but only 55 M of parameters. A compressed view of the three models used in this work is given in Fig. 5. As the blue boxes show, to reduce the parameters of AlexNet and VGG16, the original layers on top of the last convolutional block were replaced by a global average pooling layer, two smaller fully-connected (FC) layers with 256 and 64 neurons, and a softmax layer. We determined the neuron numbers of FC layers empirically. For all the 3D data, we only adopted 3D versions of AlexNet and VGG16 to make parallel comparisons with 2D-DCNNs, and they had the same architecture as the 2D versions mentioned above. InResV2 was not included because it is too deep to be applied in 3D.

Notice that in Table II Ch-6 and 3D-6 have the same shape, and are actually the same data. The difference lies in how we treated them: 2D convolution for Ch-6 and 3D convolution for 3D-6. For a 2D convolution layer, when the input has more than one channel (e.g., a Ch-3 or Ch-6), the filter should match the number of input channels (three channels for a Ch-3, six channels for a Ch-6). Since they have the same number of channels, the convolutional filter moves along the x and y axis, thus called as a 2D convolution. To calculate the output, we perform convolution on each matching channel, then add the results together to get one feature map. In terms of a 3D convolution layer, the filter could have the different number of channels, and it moves along the x, y, and z axis, thus called as a 3D convolution. Instead of getting one feature map after the 2D convolution, the output shape of a 3D convolution is a 3D volume space such as a cube or cuboid.

When training models, transfer learning [24] is a widely used technique because it transfers the knowledge learned from ImageNet [25], a huge dataset containing 1.3 million natural images of 1,000 classes, to the target domain and fastens the convergence. Although some training processes could benefit from transfer learning, the transferability of features decreases due to the huge differences between the source domain (the large-scale natural images of ImageNet) and the target domain (our small-scale MSCT drowning dataset) [26]. Meanwhile, there is no the same available pre-trained model for 3D-DCNNs, so all models were trained from scratch.

To increase the diversity of our small-scale dataset, we carried out data augmentation by applying rotation, horizontal or vertical flipping, and height or width shifting to the original Slice-data. The same operation was done on those slices simultaneously for all the 2.5D and 3D data. The loss function was categorical cross-entropy, and the optimizer was Adam with a learning rate of 1e-5 and a decay rate of 1e-6. To determine whether the training was done, early stopping was applied to stop training when the validation loss was no longer decreasing in 10 epochs. Since changing these hyperparameters does not influence the conclusion of experiments in any significant way as long as the models were able to converge [27], [28], we used the same configuration mentioned above for all 2D and 3D models.

D. Saliency Visualization

To evaluate our proposed method and give explainable results, we employed a saliency visualization technique called Grad-CAM to show the highlighting areas of models on a given input image. In this way, we can compare the coherency between highlighting areas of models and the judgement of forensic specialists. A brief demonstration of how to apply Grad-CAM

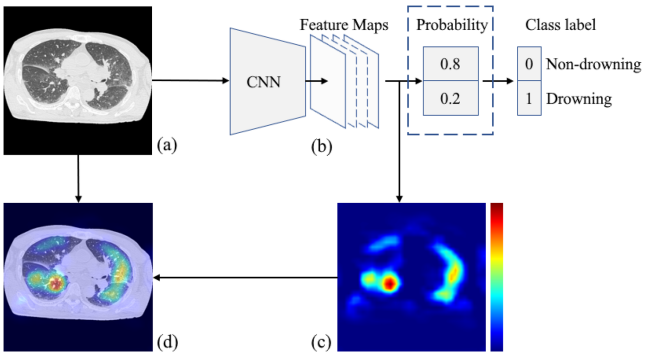


Fig. 6. A demonstration of saliency visualization. Given an input image (a), we compute the gradient of the score for a target class (e.g., drowning), with respect to feature maps (b) of the last convolutional layer. This gradient is then global-average-pooled to obtain neuron importance weights, which are then combined with feature maps to calculate the coarse localization map (c). Finally, the localization map is projected onto (a) to obtain the saliency visualization (d).

is shown in Fig. 6. After an image (a) is fed into a trained DCNN model, we compute the gradient of the score for a target class (e.g., drowning) with respect to feature maps (b) of the last convolutional layer. This gradient is subsequently global-average-pooled to obtain the neuron importance weights, which capture the importance of each feature map for the target class. Then a linear combination is performed on feature maps with weights to get a coarse localization map (c). Finally, we project the localization map onto the input image to obtain a saliency visualization (d). The color bar next to the localization map (c) depicts the score for the target class. Warmer (red) regions correspond to a higher score for the target class, showing model's stronger attention in this area. Based on the visualization, we can further evaluate whether the attention of the model match the human expert's knowledge. Details on Grad-CAM algorithm can be found in the original study by Selvaraju et al. [14].

III. Results

A. Drowning Diagnosis

As mentioned in Table II, the number n of different image data types varies for one case. Thus, we calculated the subset-wise and case-wise performance of each type of data in this study. As evaluation indexes, we used accuracy, false positive rate (FPR), false negative rate (FNR), receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC). Instead of using the default threshold of 0.5 for binary classification, we chose the maximum Youden's index J [29] to find out the optimal cut-off point of models. The J ranges from 0 to 1, and higher J means better performance of the dichotomous diagnostic test. Definitions of subset-wise and case-wise results are as follows:

1) *Subset-wise:* To decide whether the generated data is N (non-drowning) or P (drowning), we can directly check the class label outputted by the model. Then subset-wise results are calculated according to:

$$\text{Accuracy} = (\text{TN+TP})/(\text{TN+TP+FN+FP})$$

$$\text{FPR} = \text{FP}/(\text{FP+TN})$$

$$\text{FNR} = \text{FN}/(\text{TP+FN})$$

$$\text{J} = \text{TP}/(\text{TP+FN})+\text{TN}/(\text{TN+FP}) - 1$$

2) *Case-wise:* To decide whether a test case is N (non-drowning) or P (drowning), we calculated the sum of predicted probabilities prob_i of class labels for each image data of the same case (see Fig. 6). Then we can obtain case-wise results according to:

$$\text{case} = \begin{cases} N, & \text{if } \frac{1}{n}\sum_{i=1}^{n} \text{prob}_i \leq 0.5 \\ P, & \text{if } \frac{1}{n}\sum_{i=1}^{n} \text{prob}_i > 0.5 \end{cases}$$

where n is the number of generated data per case (Table II). After the classes of all cases were decided, we calculated the case-wise accuracy and FPR using the same definition as subset-wise results. Ideally, subset-wise and case-wise results of Con-data should be almost the same, as well as 3D-6, because they are designed to represent a case using a single piece of data.

Based on the subset-wise classification results, ROC curves and AUC values are given in Fig. 7 (case-wise ROC-AUCs are omitted because they are calculated from subset-wise results). Since values of TPR and FPR change as the threshold for the predicted probabilities varies, We used the maximum value of Youden's index J to select the optimum cut-off points rather than the default threshold of 0.5 for ROC curves. When using 2D-DCNN models, on the one hand it can be observed from (a), (b), and (d) that AUCs and Youden's index J were improved as the information contained in the input data increased. On the other hand, Ch-6 (c) had more information, but the models' performance was worse than that of Ch-3 (b) due to the overlap among too many channels. The best results on Con-data are not only because of the larger input size, which will be explained in the following subsection. Meanwhile, the Con-data (d) outperforms the Ch-6 (c) and 3D-6 (e) though they contain the same information, which proves the effectiveness of this 2.5D method. Regarding 3D-DCNN models, their performance decreased significantly because there were many more parameters to be trained and their huge computational cost. That is why in Fig. 7(e), (f) results of 3D-VGG were worse than that of 3D-AlexNet, as it has more than tens of times of parameters. Also, we can see from (c) and (f) that results on Ch-6 and 3D-24 were the worst out of the same type / all types of data, even though they contained the most information of cases.

Table III summarizes the subset-wise and case-wise classification accuracy, FPR, and FNR based on Youden's index J on test set, as well as the running time (second):

1) *2D, 3D vs. 2.5D:* Models performed their best on Con-data, on which InResV2 had reached the highest accuracy while maintaining the lowest FPR, FNR, and running time. VGG16 showed almost equal performance, suggesting that the classification results might not benefit from increasing the depth of networks. Although high accuracy was obtained on Ch-3, it also showed a higher FPR and FNR (smaller J). At the same time, although Ch-6 and 3D-6 ($512 \times 512 \times 6$) had the same input size as Con-data

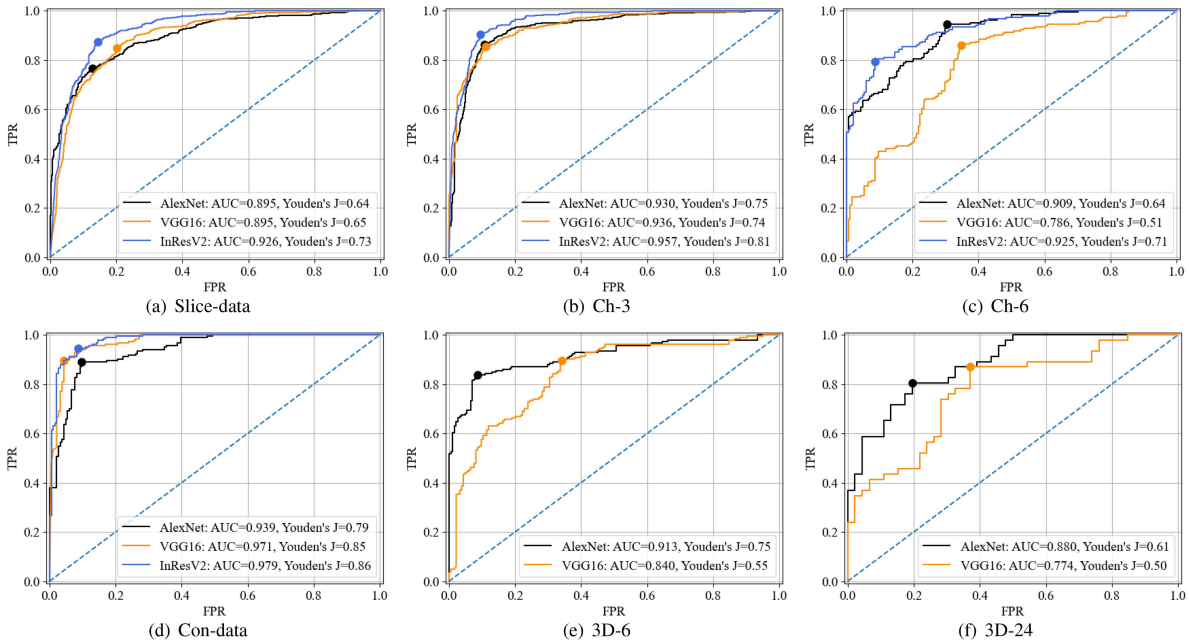


Fig. 7. Subset-wise ROC curves and AUC values of different models on each type of data. Points on the curves are the optimum cut-off points that are selected by the maximum value of the Youden's index J.

TABLE III
SUBSET-WISE AND CASE-WISE ACCURACY, FPR, FNR (BASED ON YOUDEN'S INDEX J), AND RUNNING TIME (S) ON THE TEST SET

Data	Model	Youden's J		Accuracy		FPR		FNR		Running Time	
		subset	case	subset	case	subset	case	subset	case	subset	case*
Slice-data (2D) [5], [6]	2D-AlexNet	0.638	0.720	0.819	0.860	0.127	0.120	0.235	0.160	1.9	45.6
	2D-VGG16	0.646	0.680	0.823	0.840	0.202	0.160	0.152	0.160	2.3	55.2
	2D-InResV2	0.728	0.800	0.864	0.900	0.145	0.080	0.127	0.120	3.9	93.6
Ch-3 (2.5D) [10]	2D-AlexNet	0.753	0.840	0.876	0.940	0.107	0.080	0.140	0.040	1.9	30.4
	2D-VGG16	0.741	0.760	0.871	0.880	0.112	0.040	0.147	0.040	2.3	36.8
	2D-InResV2	0.806	0.840	0.903	0.920	0.095	0.120	0.099	0.040	3.9	62.4
Ch-6 (2.5D)	2D-AlexNet	0.636	0.640	0.878	0.820	0.304	0.320	0.060	0.040	1.9	30.4
	2D-VGG16	0.505	0.400	0.753	0.700	0.348	0.440	0.147	0.160	2.3	36.8
	2D-InResV2	0.701	0.600	0.851	0.800	0.087	0.080	0.212	0.320	3.9	62.4
Con-data (2.5D, proposed)	2D-AlexNet	0.788	0.760	0.894	0.880	0.098	0.120	0.114	0.120	2.1	8.4
	2D-VGG16	0.848	0.800	0.924	0.900	0.043	0.040	0.109	0.160	**4.2**	16.8
	2D-InResV2	**0.853**	0.840	**0.927**	0.920	**0.087**	0.120	**0.060**	0.040	**4.1**	16.4
3D-6 (3D)	3D-AlexNet	0.745	0780	0.872	0.880	0.087	0.120	0.168	0.120	2.8	11.2
	3D-VGG16	0.549	0.520	0.774	0.760	0.342	0.440	0.109	0.040	4.2	16.8
3D-24 (3D)	3D-AlexNet	0.587	0.520	0.793	0.760	0.196	0.280	0.217	0.200	3.0	3.0
	3D-VGG16	0.478	0.520	0.739	0.760	0.370	0.440	0.152	0.040	5.6	5.6

* Here we take the n of 24-slice case to calculate case-wise times.

(3072 × 512), they had different information structures, which led to the difference in the result. Some information among slices of Ch-6 and 3D-6 was lost during the operation of 2D and 3D convolution and pooling, but these information was maintained by rearranging those slices into the same channel (Con-data).

2) *Difference between subset-wise and case-wise results:* It is worth noting that subset-wise and case-wise accuracy were almost the same on Con-data, 3D-6, and 3D-24, whereas huge differences appeared on Slice-data, Ch-3 and Ch-6. This means the proposed 2.5D method could represent a 3D case using a 2D image, and we could directly obtain case-wise classification using a single piece of Con-data.

3) *Running time:* Con-data could be more effective as it takes much less time when predicting a whole case. Although the prediction time per image for Con-data was the highest among 2D models, it is feasible to use a single piece of Con-data to represent a whole case based on 2). Thus, we can further reduce the case-wise prediction time to around 4 seconds.

B. Saliency Visualization

By applying the Grad-CAM, we could highlight areas that represent the significant areas on the image for drowning prediction. Fig. 8 shows the saliency visualization of Slice-data (b), Ch-3 (c), 3D-6 (d), and Con-data (e) on a 24-slice drowning case. The

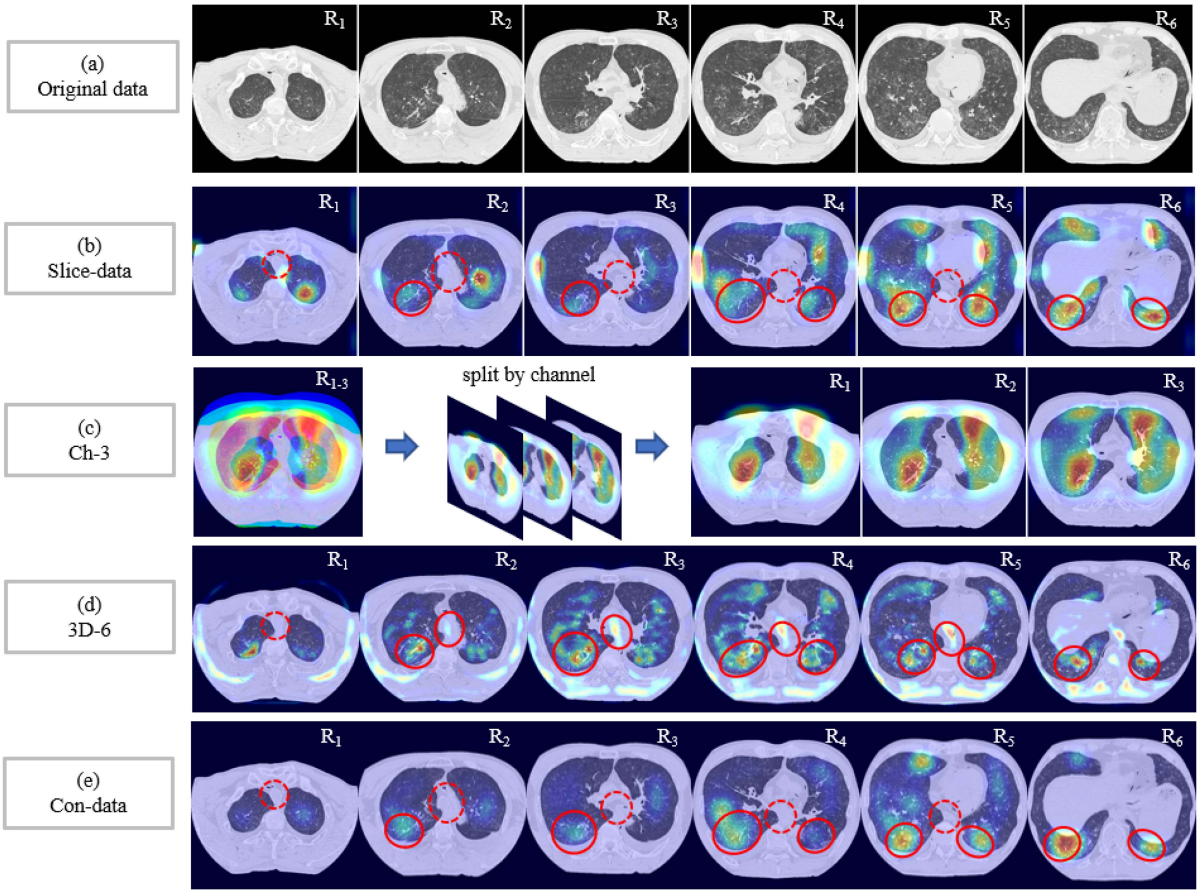


Fig. 8. Saliency visualization (horizontally displayed for a better view) of six slices of a 24-slice drowning case with true positive predictions on Slice-data, Ch-3, 3D-6 and Con-data using 2D and 3D-VGG16. Ch-6 and 3D-24 are omitted because saliency maps were the same for all channels of Ch-6, and little variation can be observed from slices in each region of 3D-24. The most concerned areas of forensic specialists are marked out with red circles. The solid/dashed circles represent the areas that the model had/had not paid attention to. (a) shows the pre-processed original data. When using Slice-data (b), many unrelated areas were activated, and attentions between slices were inconsistent. The visualization on Ch-3 (c) was not readable because of the overlap of different channels. 3D-6 (d) could provide more consistent and readable attention than Slice-data and Ch-3, but there were many messy and unrelated highlighting areas. Con-data (e) presented the clearest and the most consistent attention.

Ch-6 and 3D-24 are not displayed here because saliency maps were the same for all channels of Ch-6 and slices in each region were almost the same in 3D-24. Six original images from each region (R_1 to R_6, see Fig. 1) were displayed on the top and were all correctly predicted as drowning by InResV2. Features/areas that forensic specialists are concerned most during diagnosis are marked out in (a) with red circles (e.g., lung lesions, liquid in trachea and others). Corresponding highlighting areas of InResV2 model are also marked out in (b), (d), and (e), where the solid/dashed circles represent areas the model had/had not focused.

It can be observed from Slice-data (b) that the model had focused on correct areas but more on unrelated areas. Also, areas among slices were inconsistent, suggesting the model learned no contextual information. For Ch-3 (c), although high prediction accuracy can be achieved on it, the visualization was not even readable because of the overlap of different channels. Compared to other data, 3D models with 3D-6 (d) can capture the relative positions of the trachea and esophagus, and lung lesions. However, it is also because of 3D operators and the noncontinuous data leading to the unneeded activation on body, spine, and even artifact on the background. Finally, the clearest and most consistent attention was obtained on Con-data (d), with almost the same highlighting areas as forensic specialists' concerns, except those marked with dashed circles. In the autopsy process, solid or fluid in the trachea (unfocused areas) would be one of the considerations for drowning. However, it is not a specific nor discriminative factor, as such a situation can occur in subjects who suffer from heart failure or other diseases. It is reasonable for the model to predict only based on lung lesions because of the non-specific pathophysiology of drowning. Still, we expect models to see all features as much as possible while paying no attention to unrelated parts as forensic specialists do. Detailed findings in different data types will be discussed in the next section.

IV. Discussions

Although models have shown high classification performance, it is necessary to give explainable results considering the black-box nature of deep learning. Fig. 9 shows the same 24-slice drowning case as Fig. 1. The forensic specialists' concerned

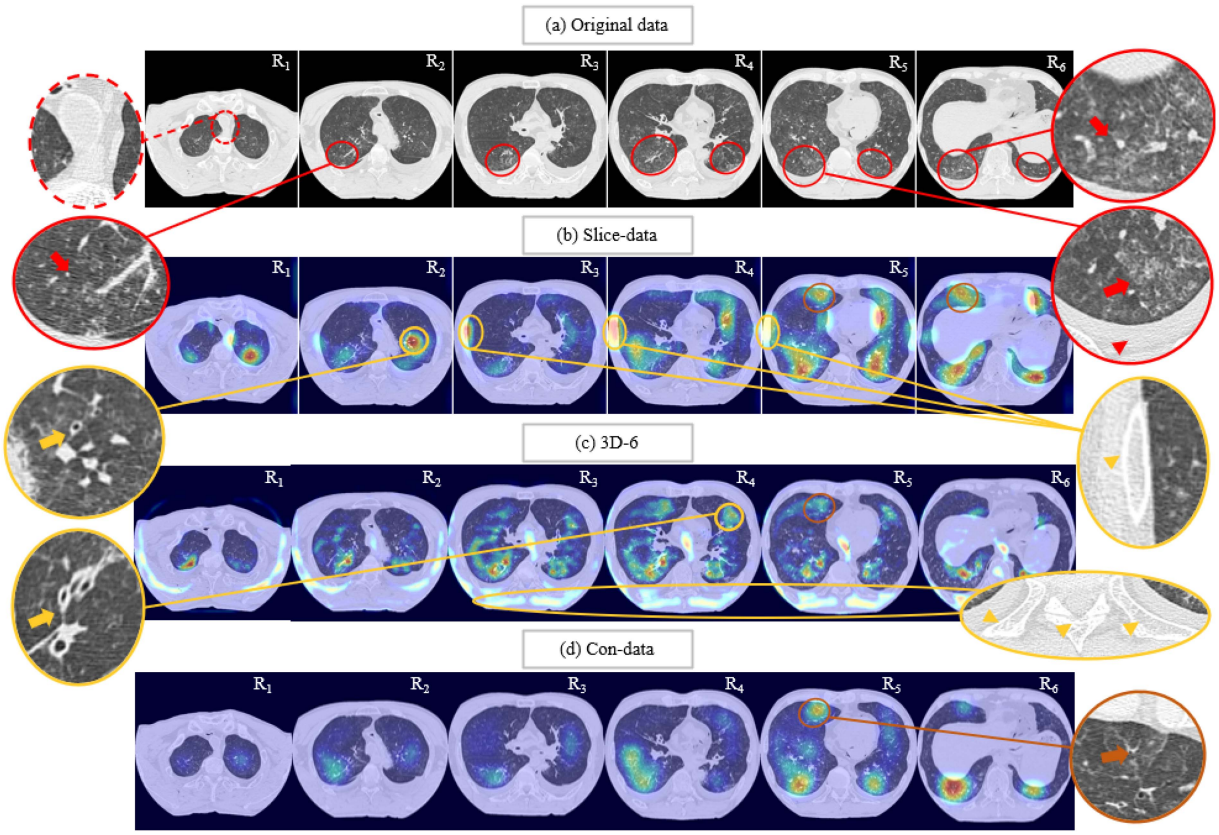


Fig. 9. Findings of a 24-slice drowning case with true positive predictions on Slice-data, 3D-6, and Con-data. The solid/dashed lines represent areas that the model had/had not paid attention to, respectively. We marked out findings that were most concerned by forensic specialists in (a) with red circles, some unrelated highlighting areas with yellow, and some not significant but correct highlighting areas with brown.

areas are marked out with red. These area included typical findings of liquid in the trachea and esophagus (dashed circle) in (a) R_1, granular opacities (arrow) in (a) R_2, reticular interstitial pattern (arrow) and pleural effusion (arrowhead) in (a) R_5, and ground-glass opacities (arrow) in (a) R_6. When using Slice-data and 3D-6, the model would focus on many other incorrect areas (yellow), such as vessels (arrow) in (b) R_2 and (c) R_4, ribs (arrowhead) in (b) R_3-R_5, and spines (arrowhead) in (c) R_3-R_6. Forensic specialists did not consider these unrelated areas during diagnosis and thus cannot prove the validity of prediction results. Highlighted areas of Con-data (d) corresponded to most findings marked in (a), and some other correct but not significant areas (brown) can also be found, such as ground-glass opacities in R_5. As we mentioned in the former section, the unfocused part of the trachea and esophagus in Con-data is not specific to drowning, so it is reasonable for the model to give prediction based on lung lesions. By showing the comparison above, Fig. 9 has proved that 2D models with Con-data could perform well and give explainable results that correspond to the drowning diagnosis. Nevertheless, we still would like to improve models in our future work to make the models' attention closer to how forensic specialists diagnose drowning (taking all factors into consideration whether they are specific or not).

However, models might fail to classify some hard cases that were difficult to diagnose even from the perspective of autopsy due to their complicated pathophysiology. For example, Fig. 10 shows a non-drowning case misclassified as drowning using Con-data. Three models in this study failed to classify it using all types of data. Models wrongly focused on the diaphragmatic surface of the liver in R_5 (yellow) but correctly focused on many typical findings in images, such as ground-glass opacities that are marked out with red in R_2, R_3, and R_5. However, these findings may also appear in other diseases or symptoms like edema. We can see edema-like pattern caused by the cardiopulmonary resuscitation in this non-drowning case. Such lung lesion and liquid in the airway made this case very similar to a drowning case even for forensic specialists. Considering the complex pathophysiology of drowning, in addition to referring to the results provided by MSCT, it is still necessary for forensic specialists to give a final diagnosis based on other information, such as the environmental situation and what circumstances bodies were found.

Besides, we can find some inconsistency among models' highlighted areas in Fig. 10. Although the three models gave the same prediction of drowning and focused on mostly the same areas, some inconsistent activation existed in R_4 and R_5. This might be caused by the differences in models' architecture. Such inconsistency problems should be handled carefully in future work by adding a strong constraint to models' attention to provide more convincing and explainable results. Furthermore, due to

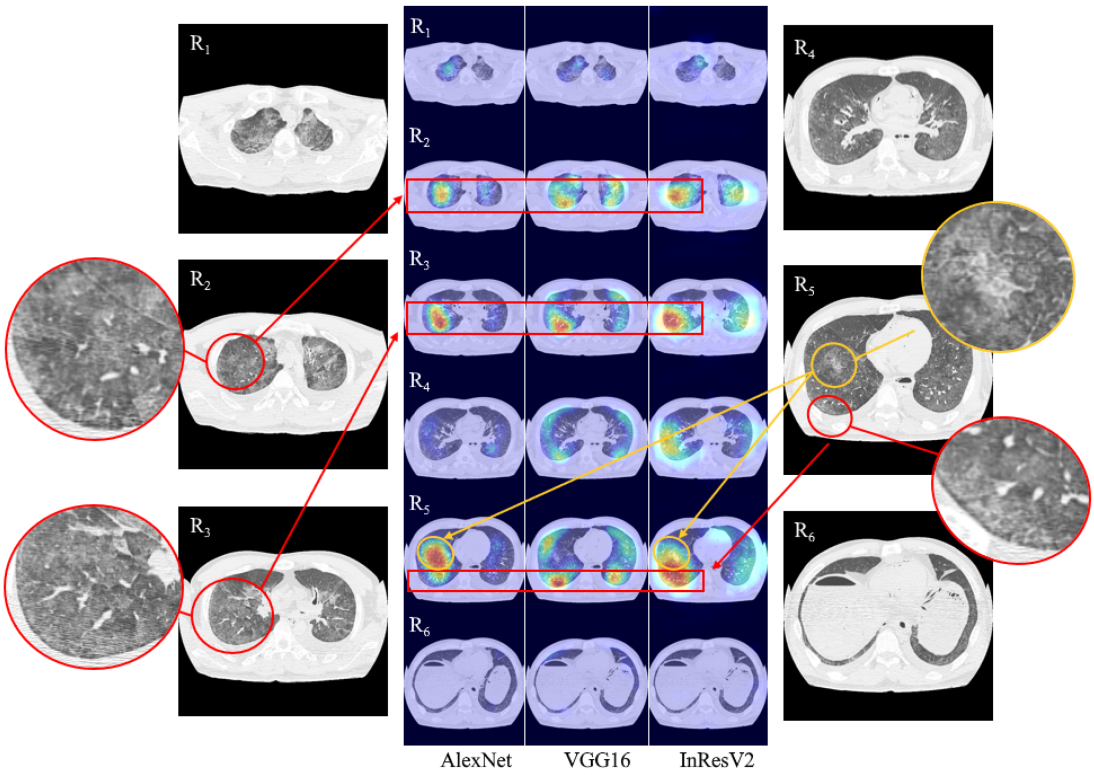


Fig. 10. A false positive example on Con-data of a 24-slice non-drowning case and its visualization on AlexNet, VGG16, and InResV2. Three models gave the same prediction of drowning. A common finding (e.g., ground-glass opacities) is marked out with red circles. An incorrect focusing of the diaphragmatic surface of the liver is marked out with the yellow circle.

the particularity of autopsy images, we did not perform external validation because it is tough to obtain reliable dissection-proved autopsy imaging data. However, we would like to conduct data normalization or generalized domain adaption to deal with data from different imaging and verify the generalizability of the system in our future studies.

V. Conclusion

In this study, we proposed a 2.5D deep learning-based CAD system for drowning diagnosis. Different 2D and 3D-DCNN models were trained using 2D data (Slice-data), 2.5D data (Ch-3, Ch-6 and Con-data), and 3D data (3D-6, 3D-24). Then the Grad-CAM was applied to obtain saliency visualization of models. Experiment results had shown the superiority and effectiveness of Con-data, as it could provide accurate case-wise classification and decent visual feedback to forensic specialists. Further research could be carried out to regularize the attention of models or utilize spatial information better by using channel-attention or depth-wise convolution to Ch-3, Ch-6, or other 3D data and evaluate whether features extracted by models agree with forensic knowledge.

References

[1] World Health Organization (WHO), "Global report on drowning: Preventing a leading killer," World Health Organization, Geneva, Switzerland, 2014.

[2] S. V. Plaetsen, E. D. Letter, M. Piette, G. V. Parys, J. W. Casselman, and K. Verstraete, "Post-mortem evaluation of drowning with whole body CT," *Forensic Sci. Int.*, vol. 249, pp. 35–41, 2015.

[3] A. Usui, Y. Kawasumi, M. Funayama, and H. Saito, "Postmortem lung features in drowning cases on computed tomography," *Japanese J. Radiol.*, vol. 32, no. 7, pp. 414–420, 2014.

[4] A. Christe, E. Aghayev, C. Jackowski, M. J. Thali, and P. Vock, "Drowning–post-mortem imaging findings by computed tomography," *Eur. Radiol.*, vol. 18, no. 2, pp. 283–290, 2008.

[5] N. Homma et al., "A deep learning aided drowning diagnosis for forensic investigations using post-mortem lung CT images," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2020, pp. 1262–1265.

[6] Y. Zeng et al., "Deep learning-based explainable computer-aided diagnosis of drowning for forensic radiology," in *Proc. IEEE 60th Annu. Conf. Soc. Instrum. Control Engineers Jpn.*, 2021, pp. 820–824.

[7] Y. George, B. J. Antony, H. Ishikawa, G. Wollstein, J. S. Schuman, and R. Garnavi, "Attention-guided 3D-CNN framework for glaucoma detection and structural-functional association using volumetric images," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 12, pp. 3421–3430, Dec. 2020.

[8] E. Trivizakis et al., "Extending 2-D convolutional neural networks to 3-D for advancing deep learning cancer classification with application to MRI liver tumor differentiation," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 3, pp. 923–930, May 2019.

[9] Y. Zeng et al., "Automatic diagnosis based on spatial information fusion feature for intracranial aneurysm," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1448–1458, May 2020.

[10] R. Alkadi, A. El-Baz, F. Taher, and N. Werghi, "A 2.5 D deep learning-based approach for prostate cancer detection on t2-weighted magnetic resonance imaging," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 734–739.

[11] H. R. Roth et al., "A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2014, pp. 520–527.

[12] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *J. Imag.*, vol. 6, no. 6, 2020, Art. no. 52.

[13] W. Shi, L. Tong, Y. Zh, and M. D. Wang, "COVID-19 automatic diagnosis with radiographic imaging: Explainable attention transfer deep neural networks," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, pp. 2376–2387, Jul. 2021.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
[15] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.
[16] J. Zhou et al., "Weakly supervised 3D deep learning for breast cancer classification and localization of the lesions in MR images," *J. Magn. Reson. Imag.*, vol. 50, no. 4, pp. 1144–1151, 2019.
[17] G. Huang, Z. Liu, L. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
[18] H. Mzoughi et al., "Deep multi-scale 3D convolutional neural network (CNN) for MRI gliomas brain tumor classification," *J. Digit. Imag.*, vol. 33, pp. 903–915, 2020.
[19] H. Zunair, A. Rahman, N. Mohammed, and J. P. Cohen, "Uniformizing techniques to process CT scans with 3D CNNs for tuberculosis prediction," in *Proc. Int. Workshop Predictive Intell. Med.*, 2020, pp. 156–168.
[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.
[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.
[22] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-V4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.
[23] K. Pasupa and W. Sunhem, "A comparison between shallow and deep architecture classifiers on small dataset," in *Proc. IEEE 8th Int. Conf. Inf. Technol. Elect. Eng.*, 2016, pp. 1–6.
[24] S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[25] D. Jia, D. Wei, R. Socher, L. Li, L. Kai, and F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[26] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Proc. 27th Int. Conf. Adv. Neural Inf. Process. Syst.*, 2014, vol. 27, pp. 3320–3328.
[27] M. L. Richter, W. Byttner, U. Krumnack, L. Schallner, and J. Shenk, "(Input) size matters for CNN classifiers," in *Proc. Int. Conf. Artif. Neural Netw.*, 2021, pp. 133–144.
[28] V. Thambawita, I. Strümke, S. A. Hicks, P. Halvorsen, S. Parasa, and M. A. Riegler, "Impact of image resolution on deep learning performance in endoscopy image classification: An experimental study using a large data set of endoscopic images," *Diagnostics*, vol. 11, no. 12, 2021, Art. no. 2183.
[29] R. Fluss, D. Faraggi, and B. Reiser, "Estimation of the youden index and its associated cutoff point," *Biometrical J: J. Math. Methods Biosci.*, vol. 47, no. 4, pp. 458–472, 2005.