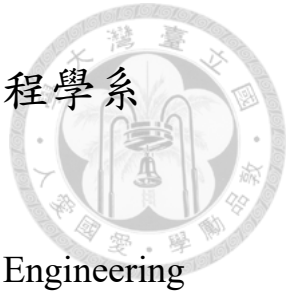


國立臺灣大學電機資訊學院資訊工程學系

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

使用 3-D 卷積神經網路於肺部 CT 結節診斷

Lung CT Nodule Classification using 3-D Convolutional Neural Networks

黃寅

Yin Huang

指導教授：張瑞峰 博士

Advisor: Ruey-Feng Chang, Ph.D.

中華民國 113 年 7 月

July 2024

doi:10.6342/NTU202401286

口試委員會審定書

國立臺灣大學碩士學位論文
口試委員會審定書
MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

使用 3-D 卷積神經網路於肺部 CT 結節診斷

Lung CT Nodule Classification using 3-D Convolutional Neural Networks

本論文係黃　寅君（學號 P09922002）在國立臺灣大學資訊工程學系完成之碩士學位論文，於民國 113 年 7 月 17 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Department of Computer Science and Information Engineering on 17 July 2024 have examined a Master's thesis entitled above presented by HUANG, YIN (student ID: P09922002) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

張瑞峰　　羅有銘　　黃耀賢

（指導教授 Advisor）

系主任/所長 Director: 陳祝嵩

doi:10.6342/NTU202401286

致謝

直到認識前女友後，才了解研究所比工作更有助提升能力，謝謝她的協助，如果沒有她，我不會來台大資工所就讀。從決定報考到最終入學遇到重重難關，而公司長官們表示支持，謝謝公司作為我就學的助力而非阻力。

謝謝李宏毅教授對教育的熱忱投入及奉獻，如果沒有機器學習這門課，我無法看懂專業論文、難以從事研究。張瑞峰教授作為我的指導教授，實驗室的研究主軸是我熱衷的深度學習，謝謝張教授給予機會，讓我透過醫學的題目，了解深度學習的能耐與侷限。

柯亞力的"一次一點，反轉憂鬱"一書，讓我理解腦科學，藉由改善生活習慣，化解壓力與憂鬱，增進學習與研究的動力，謝謝這本書的出版團隊。與朋友們偶爾聚會吃喝玩樂，成為我紓壓、繼續向前邁進的關鍵力量。謝謝朋友們的陪伴。

儘管教職繁重，黃耀賢博士仍不遺餘力地協助修改論文，引導我走完最後一哩路。謝謝黃博士的悉心指導。在這段充滿壓力的求學歲月裡，能擁有良好的居住環境，無後顧之憂全心投入學業，要謝謝父母的支持與協助。

doi:10.6342/NTU202401286

摘要

肺癌是全球癌症發生和死亡的主要原因之一。此外，肺癌的預後仍然不理想，大部分國家的五年存活率低於 20%。而低劑量電腦斷層掃描(low-dose computed tomography, LDCT)是一種廣泛使用的三維(three-dimensional, 3-D)篩檢方法，可以幫助醫生檢查肺結節，從而降低死亡率。實體結節(solid nodule, SN)、毛玻璃樣陰影(ground-glass opacity, GGO)和部分實體結節(part-solid nodule, PSN)是三種具有不同密度和良惡性程度的結節類型。準確的結節分類是一項具有挑戰性的工作，但可以通過基於深度學習的電腦輔助診斷(computer-aided diagnosis, CADx)系統得到改善。卷積神經網絡(convolutional neural network, CNN)和注意力機制現在是電腦輔助診斷系統設計中最常用的深度學習方法，因為它們具有強大的特徵提取和重新加權能力。因此，本研究提出了一種基於卷積神經網絡架構和注意力機制的三維電腦輔助診斷系統，用於結節類型分類。我們的電腦輔助診斷系統包括影像前處理和結節分類兩個部分。首先進行影像前處理，包括感興趣區域(volumes of interest, VOI)提取和影像縮放，以從低劑量電腦斷層掃描影像中提取結節及其周圍組織。然後，將提取的感興趣區域輸入到結節分類模型中，也就是我們提出的三維 SE-Inception 模型，它由 Inception-v4、Inception-ResNet-v2 和 squeeze-and-excitation(SE)模組構建而成，可以預測結節類型。此外，我們也提出 F1 引導的動態平衡訓練(F1-guided dynamic balance training, FDBT)方法，以實現更快的訓練和更好的性能。進行實驗時，從 8,789 個低劑量電腦斷層掃描影像提取的 34,898 個肺結節被使用於評估系

doi:10.6342/NTU202401286

統。根據實驗結果，我們的電腦輔助診斷系統達到 90.0%的 micro accuracy 和 83.6%的 macro F1-score，這證明了其結節分類的有效性。

關鍵詞：肺癌、肺結節、電腦斷層掃描、計算機輔助診斷系統、深度學習、卷積神經網絡、注意力機制

doi:10.6342/NTU202401286

Abstract

Lung cancer is a leading cause of cancer incidence and mortality worldwide. The prognosis for lung cancer remains unsatisfactory, with five-year survival rates below 20% in most countries. Low-dose computed tomography (LDCT) is a widely used three-dimensional (3-D) screening modality to help physicians check lung nodules to reduce mortality. Solid nodules (SNs), ground-glass opacities (GGOs), and part-solid nodules (PSNs) are three types of nodules with different densities and degrees of benignity and malignancy. Accurate nodule classification is challenging but could be improved with a computer-aided diagnosis (CADx) system based on deep learning. Convolutional neural networks (CNN) and attention mechanisms are now two of the most commonly used deep learning methods for CADx system design due to the powerful feature extraction and reweighting. Therefore, a 3-D CADx based on CNN architecture and attention scheme is proposed for nodule-type classification.

Our CADx system consists of image preprocessing and nodule classification. First, the image preprocessing composed of volume of interest (VOI) extraction and image resizing is performed to crop nodules and surrounding tissues from LDCT images. Then, the extracted VOIs are fed into the nodule classification model. Our classification model, the 3-D SE-Inception model, is built with Inception-v4, Inception-ResNet-v2, and squeeze-and-excitation (SE) modules to determine nodule types. Furthermore, the F1-

doi:10.6342/NTU202401286

guided dynamic balance training (FDBT) is introduced for faster training and better performance. In experiments, 34,898 pulmonary nodules from 8,789 3-D lung CT scans were used for system evaluation. According to the experiment, the CADx system achieved 90.0% micro accuracy and 83.6% macro F1-score, which proved its nodule classification effectiveness.

Keywords: Lung cancer, lung nodule, computed tomography, computer-aided diagnosis system, deep learning, convolutional neural network, attention mechanism

doi:10.6342/NTU202401286

Table of Contents

doi:10.6342/NTU202401286

doi:10.6342/NTU202401286

List of Figures

doi:10.6342/NTU202401286

List of Tables

doi:10.6342/NTU202401286

Chapter 1 Introduction

Lung cancer is the leading cause of cancer incidence and mortality worldwide, accounting for one-eighth of all cancer diagnoses and one-fifth of all cancer deaths [1]. According to the statistics, the prognosis for lung cancer is generally unsatisfactory, with five-year survival rates below 20% in most countries [2]. But, most lung cancer patients are diagnosed at a late stage, making curative treatment impossible [1]. Nevertheless, studies have shown that screening high-risk groups using low-dose computed tomography (LDCT) could significantly reduce lung cancer mortality, providing new hope for early detection and treatment [3, 4]. Once the screening detects a lung nodule, physicians will make subsequent decisions based on its imaging characteristics, including morphology, size, quantity, and density [5]. Based on the density, nodules can be classified into solid nodules (SNs), ground-glass opacities (GGOs), and part-solid nodules (PSNs). Among these types, GGOs appear the fuzziest and are typically benign. In contrast, SNs have the highest density, are most common, and are most likely malignant. PSNs contained both high and low-density regions, incorporating SN and GGO components. Accurate classification is a challenge for clinical decision-making and lung cancer diagnosis because of the complex characteristics of nodules. However, this challenge could be tackled with the assistance of computer-aided diagnosis (CADx) system.

Deep learning is now a powerful architecture in cancer detection [6], segmentation

doi:10.6342/NTU202401286

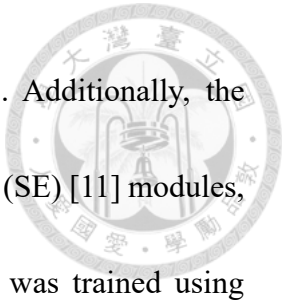
[7], and classification [8] due to its effective feature extraction. Convolutional neural networks (CNN) are a type of deep learning model with powerful capabilities in extracting image features due to their unique design [9]. Firstly, since the features of objects in an image are usually only related to their adjacency, CNN connects the neuron only to a small region of the image. Secondly, CNN uses shared neuron parameters as objects can appear in different positions and sizes within an image. These changes allow CNN to substantially reduce the parameters while effectively capturing the features of the same object in various images, drastically improving the performance. Meanwhile, to advance automated lung cancer screening technology, Pedrosa *et al*. organized the Lung Nodule Database (LNDb) challenge, which focused on the detection, segmentation, and classification of lung nodules, as well as predicting follow-up based on the 2017 Fleischner Society guidelines [10]. Although the LNDb challenge had remarkable results, it also had several limitations. The first limitation was that the dataset comprised only 294 CT scans, which was relatively small and led to insignificant differences in results [10]. In contrast, the PN9 dataset [11] was substantially larger and was used in our study. Additionally, the class imbalance made it challenging to classify specific nodules, such as PSN, thus affecting overall performance [10]. Direct training of a model on the class-imbalanced dataset with limited data will lead to bias and unsatisfactory performance. These problems can be mitigated through special sampling techniques to balance training

doi:10.6342/NTU202401286

data and the representative metric to select a balanced model during training. Therefore, this study proposed the F1-guided dynamic balance training (FDBT) method, which could address bias caused by class imbalance, mitigate overfitting due to limited data, reduce training time to one-fourth of the original, and enable the model to perform better. The second limitation was some participants employed primary ResNet architecture [12], which was less effective than the inception architecture [10, 13-15]. Notably, the Inception-v4 [14] and Inception-ResNet-v2 [14] modules were powerful in image classification and were used in our proposed 3-D SE-Inception model. The third limitation was that most models lacked attention mechanisms. In particular, channel attention can use global information to guide the model in focusing on essential features, and it is commonly used for image classification [16]. The squeeze-and-excitation (SE) [11] module was a famous attention mechanism, which was incorporated into our 3-D SE-Inception model.

In detail, this study proposed a CADx system comprised of image preprocessing and nodule classification. Initially, the image preprocessing extracts volumes of interest (VOI) containing nodule regions from lung CT scans and resizes them to a specified size. Afterward, the preprocessed VOIs are input into the proposed 3-D SE-Inception model for nodule classification. The model comprised extraction, inception, and attention stages. Specifically, the inception stage primarily consists of Inception-v4 [10] and Inception-

doi:10.6342/NTU202401286



ResNet-v2 [10] modules designed to extract multi-scaled features. Additionally, the attention stage of the model mainly comprises squeeze-and-excitation (SE) [11] modules, the most classic channel attention mechanism. Besides, the model was trained using FDBT to reduce training time and improve performance. These improvements could enhance the CADx system's performance in nodule classification, contributing to the advancement of automated lung cancer screening technology.

The subsequent chapters of this paper introduced the proposed CADx system in detail. Chapter 2 described the medical image dataset used in this study, including information on the data sources, quantities, nodule subtypes, and other statistics. Chapter 3 explained the technical details of the system, including image preprocessing steps, the architecture design of the 3D SE-Inception classification model, FDBT, data augmentation, and hyperparameter settings. Chapter 4 used standard evaluation metrics such as precision, recall, F1-score, accuracy, and specificity to evaluate the effectiveness of the proposed CADx system on the nodule classification task. Afterward, the experiment results and the drawbacks of the CADx were analyzed and discussed. Finally, in Chapter 5, the conclusion section summarizes this study's main findings, the model's strengths and weaknesses, and directions for future improvements.

doi:10.6342/NTU202401286

Chapter 2 Materials

The PN9 dataset [11], which comprised data collected from the clinics, hospitalization, and physical examination of 8,798 patients during 2015 and 2019, was used in this study. The dataset consisted of 8,798 3-D lung CT scans acquired from the CT scanners of 5 manufacturers, including GE Medical Systems, Siemens, Toshiba, United Imaging Healthcare, and Philips. According to the literature description [11], the PN9 was a preprocessed dataset. In each CT scan, the preprocess normalized the voxels to [0, 255] and reconstructed the slice thicknesses to a uniform 1×1×1 mm resolution. The categories comprised calcific nodules (CN), ground-glass opacities (GGO), part-solid nodules (PSN), and solid nodules (SN), while this study excluded CN. Hence, the dataset consisted of 34,898 pulmonary nodules ranging in size from 2 mm to 112 mm, with the mean and standard deviation of the nodule sizes being 7.62 and 5.47 mm, respectively. Fig. 2-1 and Table 2-1 shows the size distribution and the number of nodules in each subtype, respectively.

doi:10.6342/NTU202401286

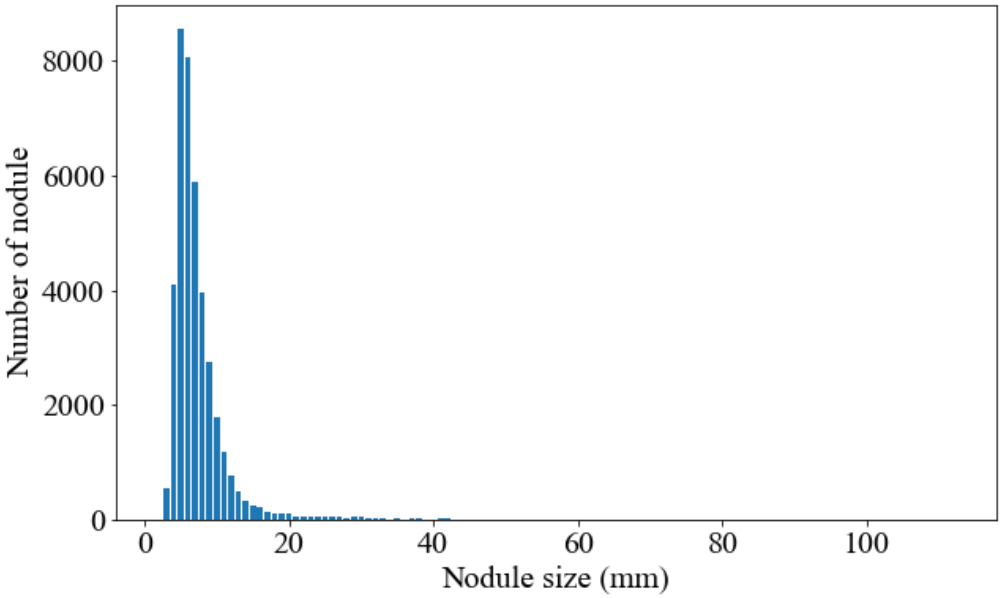


Fig. 2-1 The distribution of nodule sizes (mm).

Table 2-1 The number of each nodule class in the PN9 dataset.

Class			
GGO	PSN	SN	Total
11,931	2,929	20,038	34,898

doi:10.6342/NTU202401286

Chapter 3 Method

This study proposed a CADx system based on a 3-D SE-Inception model on CT scans for nodule classification. The system was composed of image preprocessing and nodule classification. The image preprocessing extracted the volumes of interest (VOIs) containing the nodules and surrounding tissue from the CT scan. Then, the extracted VOIs were resized to a fixed size to fit the model input size. After the image preprocessing, the resized VOIs were fed into the nodule classification to predict whether the nodule belonged to GGO, PSN, or SN. Fig. 3-1 illustrates the flowchart of the proposed CADx system.

doi:10.6342/NTU202401286

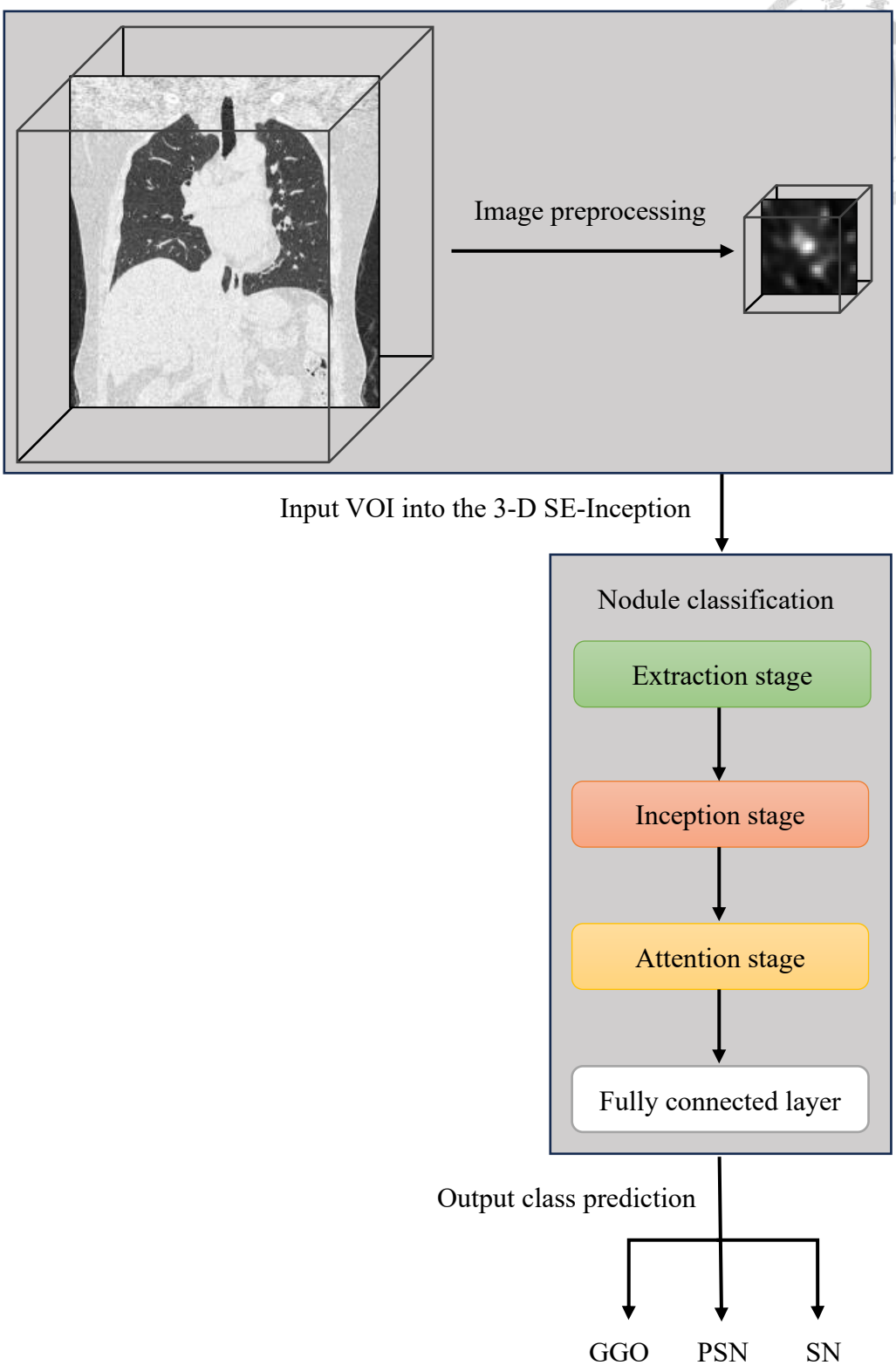
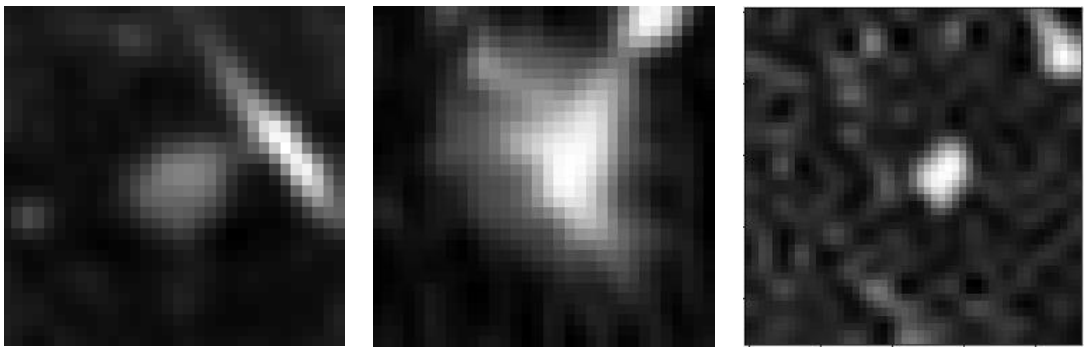


Fig. 3-1 The overall flowchart of the proposed method.

doi:10.6342/NTU202401286

3.1 Image Preprocessing

Because the goal of this study was nodule classification, image preprocessing was performed on the CT scans, including VOI extraction and resizing. In VOI extraction, starting from the center of the nodule, a 20 mm bounding box was extracted as VOI according to the nodule size distribution. In VOI resizing, the extracted VOI was resized to 32×32×32 voxels according to the model input size. Fig. 3-2 presented examples of GGO, PSN, and SN nodules after the image preprocessing.



(a) (b) (c)

Fig. 3-2 The nodule examples of different classes: (a) GGO, (b) PSN, and (c) SN.

3.2 Lung Nodule Classification

After the image preprocessing, the preprocessed VOI was fed into the following nodule classification model, 3-D SE-Inception, for subtype determination. Our 3-D SE-Inception architecture shown in Fig. 3-1 was constructed by incorporating the squeeze-and-excitation (SE) [17] attention module into the well-known Inception-ResNet-v2 [14].

doi:10.6342/NTU202401286

It comprised the extraction, the inception, and the attention stages to extract nodule features for nodule classification. The preprocessed VOIs were initially inputted into the extraction stage to capture fine-grained feature maps with reduced size. Subsequently, the fine-grained feature maps were directed to the inception stage, aiming to extract coarse-grained feature maps with increased filter dimensions. The resultant coarse-grained feature maps underwent the attention stage, where critical patterns were selected and enhanced. Finally, the outputs from the attention stage were fed into a fully connected layer to predict the nodule subtypes. The following sections describe the details of 3-D SE-Inception.

3.2.1 Extraction Stage

After the image preprocessing, the preprocessed VOIs were inputted into the extraction stage to capture minute details and local structure features. The extracted feature maps were referred to as fine-grained features, providing more information to help models better understand and classify VOIs. The convolutional layers should use smaller kernel sizes to focus on local image regions and capture fine-grained features. Therefore, the extraction stage comprised five 3-D convolutional layers (3-D Conv.) with 3×3×3 kernel sizes and progressively increased channel bandwidth. The initial 3-D Conv. layer employed a stride of 2 to reduce the size of the VOI by half. Subsequently, the second

doi:10.6342/NTU202401286

and third 3-D Conv. layers didn't use padding, which would generate feature maps slightly smaller than the input. There were two purposes for reducing the feature map sizes. The first was to reduce the computational burden, thereby enabling experiments to be conducted under limited GPU resources, and the second was to aggregate low-level features into higher-level features for better performance. Because the feature maps after the previous layers were small enough to match the requirements, the last two layers maintained consistent input and output sizes. Fig. 3-3 presented the structure of the extraction stage.

doi:10.6342/NTU202401286

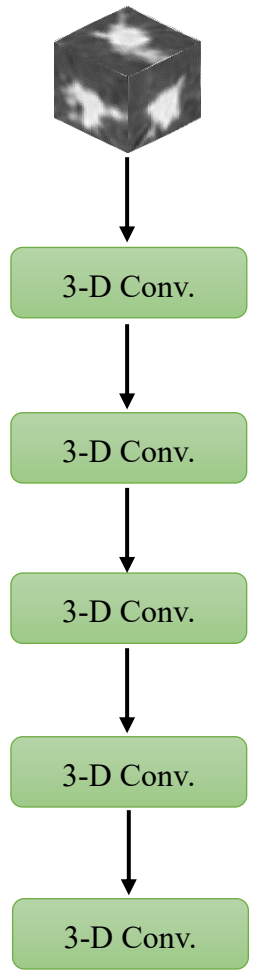


Fig. 3-3 The structure of the extraction stage.

3.2.2 Inception Stage

After the extraction stage, the extracted fine-grained features were delivered to the inception stage to refine features and capture the overall patterns of nodules. The captured feature maps were called coarse-grained features, containing higher-level information from broader scopes for distinguishing different classes of nodules. The inception stage was built using 3-D inception and 3-D residual inception modules to refine fine-grained features and capture coarse-grained features. These modules were modified from

doi:10.6342/NTU202401286

Inception-v4 [14] and Inception-ResNet-v2 [14], which were designed to classify large 2-D natural images. These modules could select patterns pertinent to target objects exhibiting varying scales. However, adopting these modules directly for small 3-D feature maps originating from nodule VOIs was inappropriate. Therefore, the primary Inception-v4 and Inception-ResNet-v2 modules were modified as the 3-D inception and 3-D residual inception modules by changing the dimension of convolution from 2-D to 3-D with 1×1×1 and 3×3×3 kernel sizes.

doi:10.6342/NTU202401286

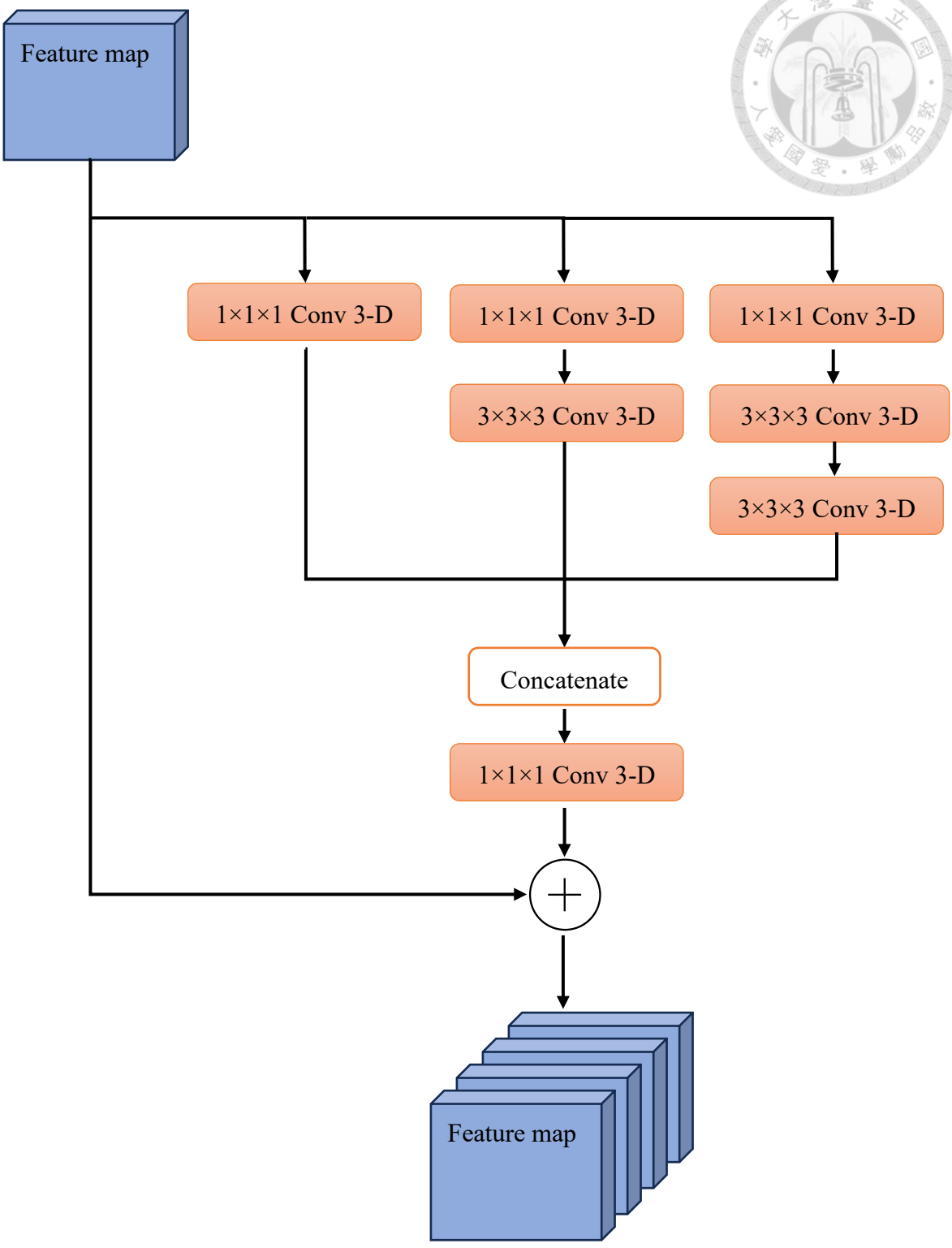


Fig. 3-4 The detail of the 3-D residual inception module.

Our inception stage comprised a carefully designed sequence of layer arrangements.

doi:10.6342/NTU202401286

Specifically, it began with one 3-D inception module followed by five 3-D residual inception modules, then another 3-D inception module, succeeded by ten 3-D residual inception modules, and finally, one more 3-D inception module followed by five 3-D residual inception modules. This intricate arrangement is illustrated in Fig. 3-5. Notably, this interleaved structure of Inception and Inception with residual modules achieved superior performance and training speed compared to inception modules alone [14].

doi:10.6342/NTU202401286

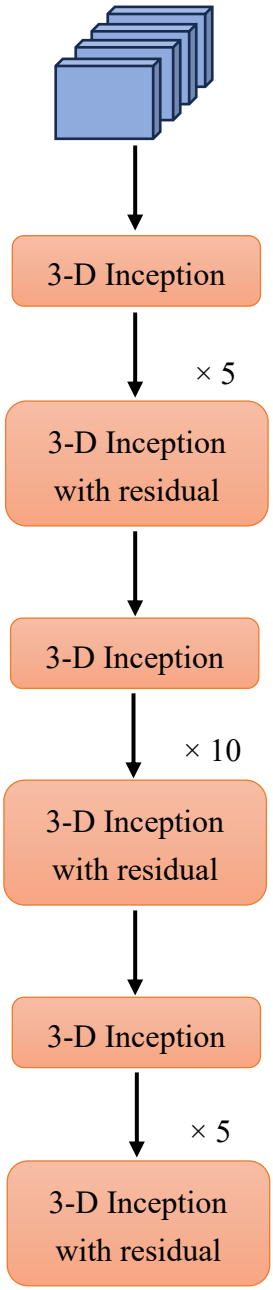


Fig. 3-5 The structure of the inception stage.

3.2.3 Attention Stage

After the inception stage, the feature maps transformed from fine-grained to coarse-grained and contained rich nodule information on different scales. To guide the model

doi:10.6342/NTU202401286

focusing more on essential features using global information of the coarse-grained feature maps, the attention [18] mechanism was helpful. Spatial and channel attention were the two major attention schemes in computer vision. Since the critical information for nodule classification was found on specific channels rather than being dependent on particular locations, it was more suitable to be handled by channel attention. Channel attention could be considered a feature selection, which learns the non-linear interdependencies between channels and dynamically adjusts the filter weights. This mechanism could selectively emphasize informative features while suppressing less useful ones for the current task.

The squeeze-and-excitation (SE) module [17] was a famous channel attention mechanism. The SE module initially employed global average pooling (GAP) to squeeze the feature map into a vector along the channel dimension. Next, it adopted the fully connected layers, rectified linear unit (ReLU) [19], and sigmoid [20] for the excitation. The feature map was passed through these layers to generate the channel weights, which were then multiplied with the original feature map. The flowchart is shown in Fig. 3-6. Hence, our attention stage illustrated in Fig. 3-7 comprised three SE modules and one 3-D Conv. layer to reweighted feature maps.

doi:10.6342/NTU202401286

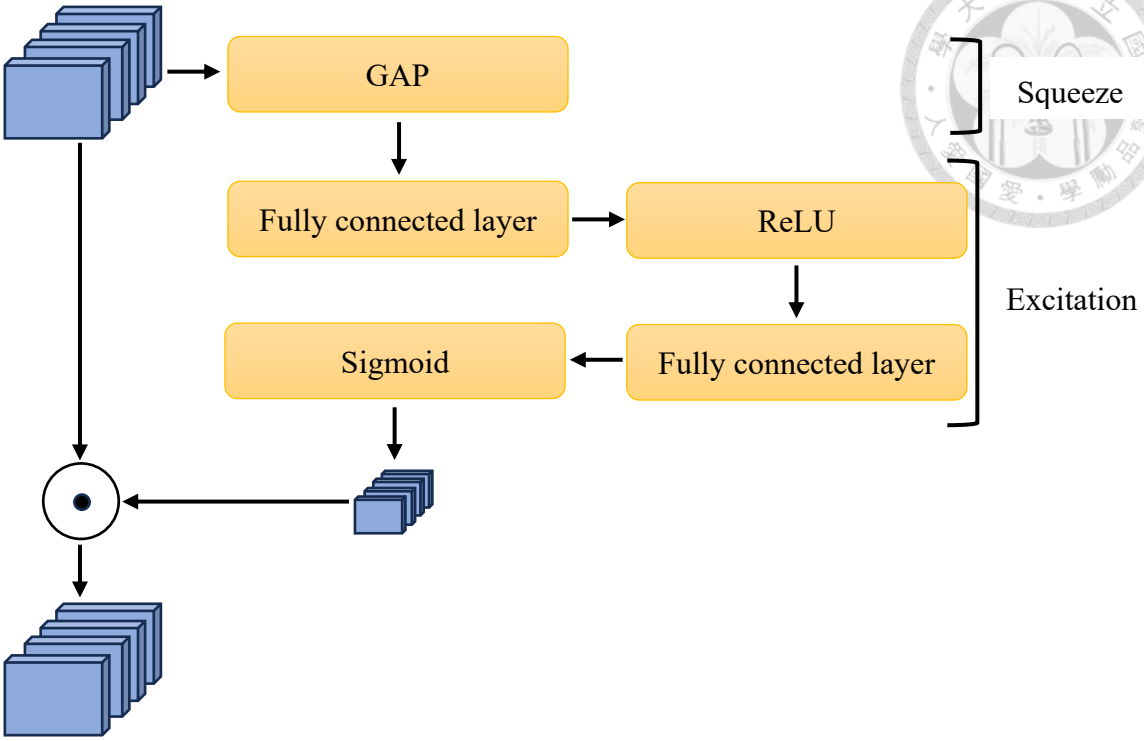


Fig. 3-6 The detail of the SE attention.

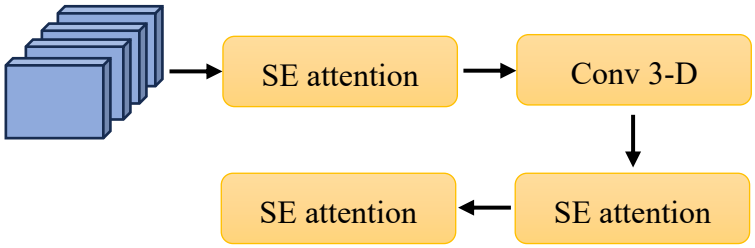
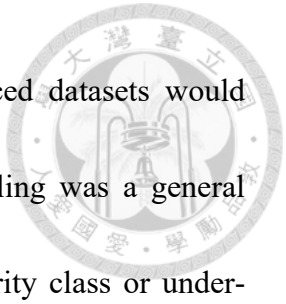


Fig. 3-7 The detail of the attention stage.

3.3 Model Training

Before training the proposed model, the problems of class imbalance and overfitting must first be addressed in this study. In class imbalance, the distribution of the PN9 dataset followed a long-tailed distribution, with the proportions of the GGO, PSN, and SN classes

doi:10.6342/NTU202401286



approximately 4:1:7. Training the model directly on the imbalanced datasets would introduce bias and result in ineffective performance. Data resampling was a general method to address this problem, either by over-sampling the minority class or under-sampling the majority class to obtain an equal quantity from each class [21]. However, over-sampling could lead to overfitting, while under-sampling might cause underfitting. Hence, this study used the FDBT, data augmentation, and hyperparameter tuning to address the problems and improve performance. Besides, CE loss and Adam [22] optimizer were used to train the model.

3.3.1 F1-guided Dynamic Balance Training

In the FDBT, each class was first under-sampled to match the quantity of the minority class and then resampled before each training epoch to mitigate overfitting and underfitting. Meanwhile, the worst validation F1-score among the three classes was used as the criterion throughout the training process. If the score of the current epoch surpassed that of the previous, the current model was considered an improvement and saved. This method could reduce the overall training time to only one-fourth. The trained model could also be less biased and more balanced, achieving better performance.

doi:10.6342/NTU202401286

3.3.2 Data Augmentation

Data augmentation was a standard solution to overfitting, and it could be implemented through data synthesis or transformation. However, due to the complex nature of patterns distinguishing different nodule classes, manual formula-based synthesis was impractical, and the dataset wasn't sufficient to train generative adversarial networks (GANs) [23] effectively for generating new data. Furthermore, most image transformations were unsuitable since they would alter key features like intensity and shape, which were crucial for distinguishing between different nodule classes. In response, a conservative approach, random rotation, was utilized in this study to overcome this issue. There were two steps to rotate a 3-D VOI through all possible orientations. First, position each of the six faces at the top in turn. Second, place each of the four horizontal sides at the front for every top face. As a result, there were 24 unique placements of a 3-D VOI after data augmentation.

3.3.3 Hyperparameters

According to our observation, the five hyperparameters listed in Table 3-1 significantly impacted model training and performance. The first hyperparameter was batch size. To address the class imbalance problem, each batch needed to contain an equal quantity of samples from the three classes, which implied it must be a multiple of three.

doi:10.6342/NTU202401286

Meanwhile, the 3-D data and model required large GPU memory, limiting the maximum batch size to 6. This batch sampling manner was used in the FDBT. The second hyperparameter was weight decay [24]. Increasing weight decay helped alleviate the overfitting problem by reducing the variance during training and increasing bias. According to our experiments, setting weight decay to 0.02 yielded the best performance. Fig. 3-8 shows the learning rate schedule used during the training, which consisted of warm-up and cosine decay.

Table 3-1 The hyperparameters used in training the proposed model.

Hyperparameters	Setting
Batch size	6
Optimizer	Adam [22]
Weight decay	0.02
Epoch	100
Loss function	Cross-entropy

doi:10.6342/NTU202401286

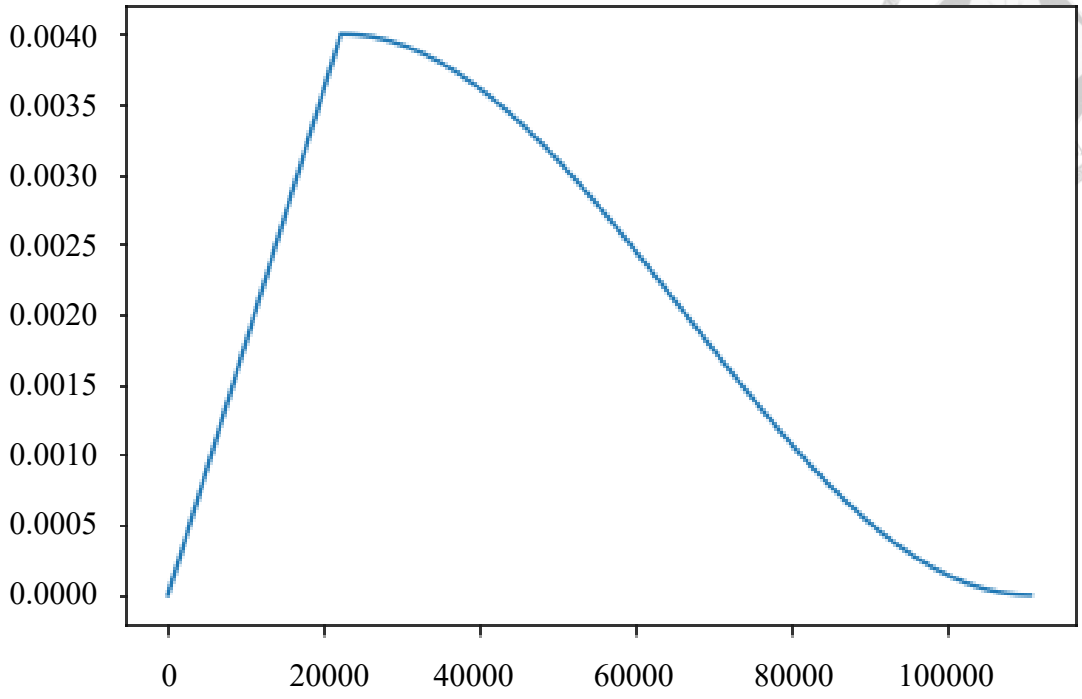
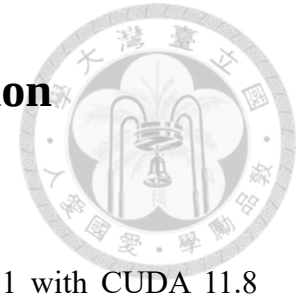


Fig. 3-8 The learning rate schedule used in training the proposed model.

doi:10.6342/NTU202401286



Chapter 4 Results and Discussion

4.1 Experiment Environment

The experiments were conducted using PyTorch version 2.0.1 with CUDA 11.8 support and Python version 3.8.3, running on a Windows 10 operating system. The hardware setup included an Intel i7-8700 CPU with a clock speed of 3.20GHz, 16.0 GB of RAM, and an NVIDIA TITAN RTX GPU with 24 GB of dedicated video memory.

4.2 Evaluation

For system evaluation, the criteria metrics, including precision, recall, F1-score, accuracy, and specificity, are utilized to assess the classification methods' performance. Each metric was calculated individually for each class, as micro-averaged across all classes and macro-averaged across all classes [25]. The *p*-value was also calculated using McNemar's Test [26] to evaluate whether the difference between the two methods was significant.

4.3 Experiment Result

This research conducted four experiments to assess the effectiveness of the proposed CADx system. Our experiments' baseline model consisted of only 3-D Inception-v4 and 3-D Inception-ResNet-v2 modules. The first was the ablation study to verify if the

doi:10.6342/NTU202401286

proposed 3-D SE-Inception, FDBT, and data augmentation could improve the performance. They were incorporated into the baseline model in sequence. The second was an experiment on the input dimension, comparing 2.5-D inputs constructed from three orthogonal slices of VOIs with 3-D inputs to determine whether the 2.5-D input dimension could preserve the complete 3-D spatial information and achieve comparable performance with the 3-D input dimension. The SE ResNet101 and SE-Inception were trained with FDBT and data augmentation using 2.5-D and 3-D input dimensions, respectively. The third was an experiment on four attention methods, including the SE module [17], convolutional block attention module (CBAM) [27], selective kernel (SK) [28] module, and efficient channel attention (ECA) module [29]. These attention modules were incorporated into the baseline model respectively. Then, these models were also trained using FDBT and data augmentation. The fourth was an experiment on model architectures. The proposed 3-D SE-Inception architecture was compared with four other architectures, including 3-D multiresolution statistical texture analysis (MSTA) [30], 3-D SE ResNet101 [12], 3-D Focal Transformer [31], and 3-D ConvNeXt-tiny [32].

4.3.1 Ablation Study

The first experiment conducted an ablation study to verify the effectiveness of the SE module, FDBT, and data augmentation. The model settings, training time, and results

doi:10.6342/NTU202401286

of the ablation study were listed in Table 4-1 and Table 4-2. According to the results, training the baseline with FDBT could mitigate model bias and improve the performance in classifying PSN, accounting for 5.3%, 16.5%, and 44.9% improvements of the macro F1-score, F1-score for PSN, and recall for PSN. Moreover, FDBT could reduce the overall training time to only one-fourth. Based on this setting, applying data augmentation through random rotation could produce more training data and alleviate overfitting, accounting for 1.7%, 3.3%, and 1.5% improvements of the macro F1-score, F1-score for PSN, and micro accuracy. Finally, the proposed method achieved the best metrics among the model settings, with 90.0% micro accuracy, 83.6% macro F1-score, 89.0% F1-score for GGO, 66.1% F1-score for PSN, and 95.6% F1-score for SN. These results verified the proposed 3-D SE-Inception with FDBT and data augmentation could improve the performance. Furthermore, to determine if there were significant differences between the results of the proposed method and those of other methods, McNemar's test was performed to calculate the *p*-values. The results are listed in Table 4-3. The difference between the two models was significant if the *p*-value was less than 0.05; otherwise, it was insignificant. The results showed that the proposed method significantly differed from other methods.

doi:10.6342/NTU202401286

Table 4-1 The model settings and training time of the ablation study.

Method	**FDBT**	**Data augmentation**	**SE module**	**Training time**
Baseline				> 57 hours
Baseline_FDBT	✓			< 15 hours
Baseline_FDBT_Aug.	✓	✓		< 15 hours
Proposed	✓	✓	✓	< 15 hours

Table 4-2 The results of the ablation study.

Method	Metrics	Micro	Macro	Class		
				GGO	PSN	SN
Baseline	Precision	0.878	0.810	0.812	**0.685**	0.934
	Recall	0.878	0.735	**0.926**	0.322	**0.958**
	F1-score	0.878	0.750	0.865	0.439	0.946
	Accuracy	0.878	0.919	0.911	0.909	0.936
	Specificity	0.939	0.930	0.904	**0.981**	0.906
Baseline_FDBT	Precision	0.872	0.796	**0.936**	0.497	0.955
	Recall	0.872	0.834	0.801	**0.771**	0.929
	F1-score	0.872	0.803	0.863	0.604	0.942
	Accuracy	0.872	0.915	0.921	0.889	0.933
	Specificity	0.936	0.939	**0.975**	0.904	0.940
Baseline_FDBT_Aug.	Precision	0.887	0.812	0.891	0.593	0.951
	Recall	0.887	0.831	0.873	0.688	0.933
	F1-score	0.887	0.820	0.882	0.637	0.942
	Accuracy	0.887	0.925	0.927	0.914	0.933
	Specificity	0.943	0.942	0.952	0.942	0.933
Proposed	Precision	**0.900**	**0.825**	0.903	0.603	**0.968**
	Recall	**0.900**	**0.851**	0.878	0.733	0.943
	F1-score	**0.900**	**0.836**	**0.890**	**0.661**	**0.956**
	Accuracy	**0.900**	**0.933**	**0.933**	**0.917**	**0.949**
	Specificity	**0.950**	**0.952**	0.958	0.940	**0.957**

doi:10.6342/NTU202401286

Table 4-3 The *p*-value for the proposed method and other methods.

Method	McNemar's Test
Baseline	<0.0001*
Baseline_FDBT	<0.0001*
Baseline_FDBT_Aug.	0.0007*

* The difference was statistically significant (*p*-value less than 0.05).

4.3.2 Comparison of Different Input Dimensions

The second experiment was the comparison of two different input types. The 2.5-D dimension was to extract and concatenate three orthogonal center slices, which could preserve some 3-D contextual information. By previous research, 2.5-D SE ResNet101 [12] was the first-place classification model in the nodule classification competition on the Lung Nodule Database (LNDb) Challenge [10]. Relative to 2.5-D, the 3-D dimension directly used the 3-D VOI as the model's input. The SE ResNet101 and the SE-Inception were trained with FDBT and data augmentation using 2.5-D and 3-D dimensions, respectively. Among the results in Table 4-4, 3-D dimensions generally performed better than 2.5-D on most evaluation metrics. Additionally, the *p*-value between 2.5-D and 3-D for each model in Table 4-5 showed statistically significant differences between different input dimensions.

doi:10.6342/NTU202401286

Table 4-4 The results of different input dimensions.

Method	Metrics	Micro	Macro	Class		
				GGO	PSN	SN
2.5-D SE ResNet101_FDBT_Aug.	Precision	0.836	0.761	0.882	0.447	0.953
	Recall	0.836	0.816	0.810	0.778	0.862
	F1-score	0.836	0.772	0.844	0.567	0.905
	Accuracy	0.836	0.891	0.907	0.870	0.895
	Specificity	0.918	0.925	0.951	0.881	0.941
3-D SE ResNet101_FDBT_Aug.	Precision	0.856	0.781	0.893	0.477	**0.971**
	Recall	0.856	0.842	0.829	**0.822**	0.876
	F1-score	0.856	0.795	0.860	0.604	0.921
	Accuracy	0.856	0.904	0.916	0.881	0.913
	Specificity	0.928	0.936	0.955	0.889	0.964
2.5-D proposed	Precision	0.849	0.773	0.895	0.467	0.957
	Recall	0.849	0.828	0.828	0.784	0.873
	F1-score	0.849	0.786	0.860	0.586	0.913
	Accuracy	0.849	0.899	0.916	0.878	0.904
	Specificity	0.924	0.931	0.956	0.890	0.946
3-D proposed	Precision	**0.900**	**0.825**	**0.903**	**0.603**	0.968
	Recall	**0.900**	**0.851**	**0.878**	0.733	**0.943**
	F1-score	**0.900**	**0.836**	**0.890**	**0.661**	**0.956**
	Accuracy	**0.900**	**0.933**	**0.933**	**0.917**	**0.949**
	Specificity	**0.950**	**0.952**	**0.958**	**0.940**	**0.957**

Table 4-5 The *p*-value between each model's 2.5-D and 3-D input dimensions.

Method	**McNemar's Test**
SE ResNet101_FDBT_Aug.	<0.0001*
Proposed	<0.0001*

* The difference was statistically significant (*p*-value less than 0.05).

doi:10.6342/NTU202401286

4.3.3 Comparison of Different Attention Mechanisms

This experiment selected the four most representative channel attention designs. The four attention modules were combined with the baseline model to compare the performances between the combined models and the baseline model. After observing the characteristics of these attention modules and conducting some experiments, each attention module was integrated into the ideal position within the baseline model for fair comparison. The SE and ECA modules were pure channel attention and were appended after the final three layers of the baseline model, respectively, because the feature maps at the last stage contained the richest channel information. The CBAM was appended after every convolutional layer within the baseline, which followed the original usage. The SK attention module had a multi-branch architecture like the Inception-v4 module. Hence, it was merged with the first three Inception-v4 modules of the baseline model. Afterward, the baseline model without any attention and the four combined with attention modules were trained using FDBT and data augmentation. Among the results in Table 4-6, the one combined with the SE attention module achieved the highest scores, with 90% micro accuracy, 83.6% macro F1-score, and 66.1% F1-score for PSN. In addition, the *p*-values between the results of SE attention and others were calculated. Among the results shown in Table 4-7, there were significant differences between SE attention and others except for CBAM.

doi:10.6342/NTU202401286

Table 4-6 The results of different attention modules.

Method	Metrics	Micro	Macro	Class		
				GGO	PSN	SN
Baseline_FDBT_Aug.	Precision	0.887	0.812	0.891	0.593	0.951
	Recall	0.887	0.831	0.873	0.688	0.933
	F1-score	0.887	0.820	0.882	0.637	0.942
	Accuracy	0.887	0.925	0.927	0.914	0.933
	Specificity	0.943	0.942	0.952	**0.942**	0.933
SE_FDBT_Aug.	Precision	**0.900**	**0.825**	0.903	**0.603**	0.968
	Recall	**0.900**	0.851	**0.878**	0.733	0.943
	F1-score	**0.900**	**0.836**	0.890	**0.661**	**0.956**
	Accuracy	**0.900**	**0.933**	**0.933**	**0.917**	**0.949**
	Specificity	**0.950**	**0.952**	0.958	0.940	0.957
CBAM_FDBT_Aug.	Precision	0.896	0.819	0.911	0.580	0.968
	Recall	0.896	0.853	0.877	0.746	0.934
	F1-score	0.896	0.832	**0.894**	0.652	0.951
	Accuracy	0.896	0.931	0.935	0.913	0.944
	Specificity	0.948	0.950	0.961	0.933	0.957
SK_FDBT_Aug.	Precision	0.891	0.814	0.911	0.574	0.958
	Recall	0.891	0.843	0.866	0.726	**0.936**
	F1-score	0.891	0.825	0.888	0.641	0.947
	Accuracy	0.891	0.927	0.932	0.911	0.939
	Specificity	0.945	0.946	**0.962**	0.933	0.944
ECA_FDBT_Aug.	Precision	0.888	0.812	**0.912**	0.544	**0.980**
	Recall	0.888	**0.866**	0.864	**0.822**	0.912
	F1-score	0.888	0.829	0.887	0.655	0.945
	Accuracy	0.888	0.925	0.932	0.905	0.938
	Specificity	0.944	0.950	**0.962**	0.915	**0.974**

doi:10.6342/NTU202401286

Table 4-7 The *p*-value for the proposed method and other methods.

Method	McNemar's Test
Baseline_FDBT_Aug.	0.0007*
CBAM_FDBT_Aug.	0.2357
SK_FDBT_Aug.	0.0129*
ECA_FDBT_Aug.	0.0002*

* The difference was statistically significant (*p*-value less than 0.05).

4.3.4 Comparison of Different Architectures

The final experiment compared the proposed 3-D SE-Inception with 3-D MSTA [30], 3-D SE ResNet101 [12], 3-D Focal Transformer [31], and 3-D ConvNeXt-tiny [32]. The MSTA framework was a classical radiomics algorithm that used 2-D ROI as input and SVM as the classifier in the original literature. It was changed to use 3-D VOI as input in this experiment and a fully connected layer as the classifier for better performance. Table 4-8 shows the results of these architectures trained using FDBT and data augmentation. According to the results, the proposed 3-D SE-Inception surpassed all others and achieved favorable results, with 90.0% micro accuracy, 83.6% macro F1-score, 89.0% F1-score for GGO, 66.1% F1-score for PSN, and 95.6% F1-score for SN. The 3-D Focal Transformer had underperforming scores because it could not converge in this experiment. The *p*-values between the proposed and other models were listed in Table 4-9, which showed significant differences between the proposed model and all other models.

doi:10.6342/NTU202401286

Table 4-8 The results of different architectures.

Method	Metrics	Micro	Macro	Class		
				GGO	PSN	SN
3-D MSTA_FDBT_Aug.	Precision	0.338	0.346	0.325	0.117	0.596
	Recall	0.338	0.349	0.605	0.226	0.217
	F1-score	0.338	0.298	0.422	0.154	0.319
	Accuracy	0.338	0.559	0.488	0.728	0.460
	Specificity	0.669	0.674	0.436	0.790	0.795
3-D SE ResNet101_FDBT_Aug.	Precision	0.856	0.781	0.893	0.477	**0.971**
	Recall	0.856	0.842	0.829	**0.822**	0.876
	F1-score	0.856	0.795	0.860	0.604	0.921
	Accuracy	0.856	0.904	0.916	0.881	0.913
	Specificity	0.928	0.936	0.955	0.889	**0.964**
3-D Focal Transformer_FDBT_Aug.	Precision	0.254	0.378	0.350	0.129	0.656
	Recall	0.254	0.388	0.042	0.872	0.251
	F1-score	0.254	0.221	0.076	0.224	0.363
	Accuracy	0.254	0.503	0.679	0.341	0.488
	Specificity	0.627	0.686	0.964	0.276	0.817
3-D ConvNeXt-tiny_FDBT_Aug.	Precision	0.767	0.694	0.819	0.349	0.915
	Recall	0.767	0.739	0.729	0.686	0.803
	F1-score	0.767	0.696	0.771	0.463	0.855
	Accuracy	0.767	0.844	0.866	0.825	0.842
	Specificity	0.883	0.889	0.927	0.843	0.897
Proposed	Precision	**0.900**	**0.825**	**0.903**	**0.603**	0.968
	Recall	**0.900**	**0.851**	**0.878**	0.733	0.943
	F1-score	**0.900**	**0.836**	**0.890**	**0.661**	**0.956**
	Accuracy	**0.900**	**0.933**	**0.933**	**0.917**	**0.949**
	Specificity	**0.950**	**0.952**	**0.958**	**0.940**	0.957

doi:10.6342/NTU202401286

Table 4-9 The *p*-value for the proposed method and other methods.

Method	McNemar's Test
3-D MSTA_FDBT_Aug.	<0.0001*
3-D SE ResNet101_FDBT_Aug.	<0.0001*
3-D Focal Transformer_FDBT_Aug.	<0.0001*
3-D ConvNeXt-tiny_FDBT_Aug.	<0.0001*

* The difference was statistically significant (*p*-value less than 0.05).

4.4 Discussion

This study proposed a 3-D CADx system for nodule classification in CT scans. The system comprised two primary components including image preprocessing and nodule classification. During the image preprocessing, VOIs containing nodules were extracted and resized to fit the model's input dimensions. These preprocessed VOIs were then fed into a 3-D SE-Inception for nodule classification. The 3-D SE-Inception model was constructed by integrating the SE attention mechanism with Inception-v4 and Inception-ResNet-v2 modules. Moreover, our system employed FDBT and data augmentation techniques to address the class imbalance and overfitting issues. As shown in Table 4-1, FDBT reduced the overall training time to one-fourth of the one directly trained on the original dataset, attributed to the undersampling technique of FDBT. Meanwhile, as shown in Table 4-2, FDBT also could effectively improve performance. Significant improvements were observed in the model's weakest area, with 16.5% and 44.9%

doi:10.6342/NTU202401286

improvements in the F1-score and recall for PSN, contributing a 5.3% increase in the macro F1-score. The performance improvement was attributed to the resampling technique and the criterion used to select the model of FDBT, which could mitigate the overfitting and bias issues, respectively. Based on FDBT and data augmentation, balanced and sufficient training data was provided, allowing the additional SE module to achieve its full capacities. As the results in Table 4-2, it achieved the best performance, with 90.0% micro accuracy, 83.6% macro F1-score, 89.0% F1-score for GGO, 66.1% F1-score for PSN, and 95.6% F1-score for SN. These results proved the effectiveness of the proposed methods.

In the LNDb nodule classification competition, the models with a 2.5-D input dimension scored higher than those with 3-D. Pedrosa *et al*. [10] attributed this performance disparity to two factors. First, the radiologists usually used 2-D slices instead of 3-D to detect and diagnose nodules in CT scans. Second, a 2.5-D input dimension allowed using 2-D models, which required fewer training parameters. These views implied that 2.5-D was superior to 3-D. However, Pedrosa *et al*. also mentioned that the small test set limited the interpretation of the nodule classification and the differences in results between 2.5-D and 3-D were actually insignificant [10]. Therefore, this study conducted a rigorous experiment comparing 2.5-D and 3-D input dimensions to determine which was better for nodule classification. In contrast to the LNDb, the PN9 dataset was

doi:10.6342/NTU202401286

substantially larger and provided a more robust empirical foundation. Furthermore, the variables were controlled by maintaining a consistent environment, ensuring that any performance differences could only be attributed to input dimensions. Among the second experimental results in Table 4-4, changing the input dimension from 2.5-D to 3-D could improve the performance, no matter which models were used. These findings challenge the previous implication about the superiority of 2.5-D over 3-D, suggesting that with a sufficiently large dataset and controlled experimental conditions, 3-D dimensions could leverage the complete spatial information of nodules, leading to more accurate classification outcomes.

Research has shown that attention mechanisms have received significantly increased focus in recent years, with different visual attention mechanisms exhibiting distinct characteristics [16]. Channel attention is a commonly used mechanism in image classification to select important feature channels. In contrast, spatial attention focuses on locating relevant regions within images to enhance object detection and segmentation performance [16]. Hence, in this study, channel attention was preferable to spatial attenuation since the image preprocessing removed the regions irrelevant to nodule classification. Table 4-6 showed that the model incorporated with SE attention achieved the highest scores in most evaluation metrics. Apparently, the CBAM used spatial attention, which was not helpful for nodule classification, and the additional parameters

doi:10.6342/NTU202401286

might even cause interference and exacerbate overfitting. Although the split operation helped extract multi-scale features for early feature maps, the effect of SK attention was inferior to that of the inception modules. On the other hand, although ECA used 1×1×1 convolutions to reduce the computational cost, it also sacrificed some performance compared to SE's use of FC layers. These findings suggested that attention mechanisms must be chosen specifically for the problem to leverage their advantages fully.

Finally, the fourth experiment compared different architectures and showed the results in Table 4-8. Among these results, the proposed method achieved favorable performance across most evaluation metrics, demonstrating that it was the most suitable architecture for nodule classification. In contrast, the 3-D MSTA [30] with classical radiomics algorithm had low performance because it might have lacked a proper step to remove the surrounding tissue from the 3-D VOI input. It directly applied radiomics feature extraction, which mixed the nodule-related information and surrounding tissue, consequently causing noisy information to degrade the performance. On the other hand, the 3-D SE ResNet101 was the second-best architecture in our experiments and the first-place classification model in the LNDb nodule classification competition [10]. However, our experiment and Kornblith *et al.*'s [13] experiments consistently showed that ResNet101 was inferior to inception. The possible reason was that ResNet101 lacked parallel paths with different-sized convolutional kernels to capture multi-scale features

doi:10.6342/NTU202401286

like inception. Conversely, the popular architectures Focal Transformer and ConvNeXt claimed to acquire outstanding performance in the original literature but had underperforming results, which might be attributed to several factors. Firstly, smaller image sizes would produce suboptimal performance for both Transformer-based architectures [33] and ConvNeXt [32]. Secondly, Transformer-based architectures [34] required a significantly large dataset like ImageNet [35] to perform well, thus underperforming on a relatively small PN9 dataset [11]. Lastly, when the target object in the image is smaller, Transformer-based architectures would have suboptimal performance because they focus more on global information and lack local perception [36].

Fig. 4-1 illustrates some examples diagnosed by the CADx system. In cases where the model misclassified GGO as PSN and PSN as SN, adjacent tissue appeared overly bright. Similarly, when the model misclassified GGO as SN, surrounding tissue appeared excessively bright and was separated from the central GGO. Consequently, the model focused on the peripheral tissue and resulted in misclassification. In contrast, PSN misclassification as GGO occurred when the overall appearance was dominated by the GGO component, with the brighter SN portion not prominently displayed. Additionally, the central GGO component's shape was atypical for a nodule and similar to the background, causing the model to focus on the peripheral tissue. Furthermore, SN

doi:10.6342/NTU202401286

misclassification as GGO might occur when the model interpreted the SN as other tissue and perceived the adjacent darker tissue as the nodule. Lastly, when the model misclassified SN as PSN, although the center appeared brighter, the nodule's surroundings were darker. As a result, the close integration of these two components led to this misclassification.

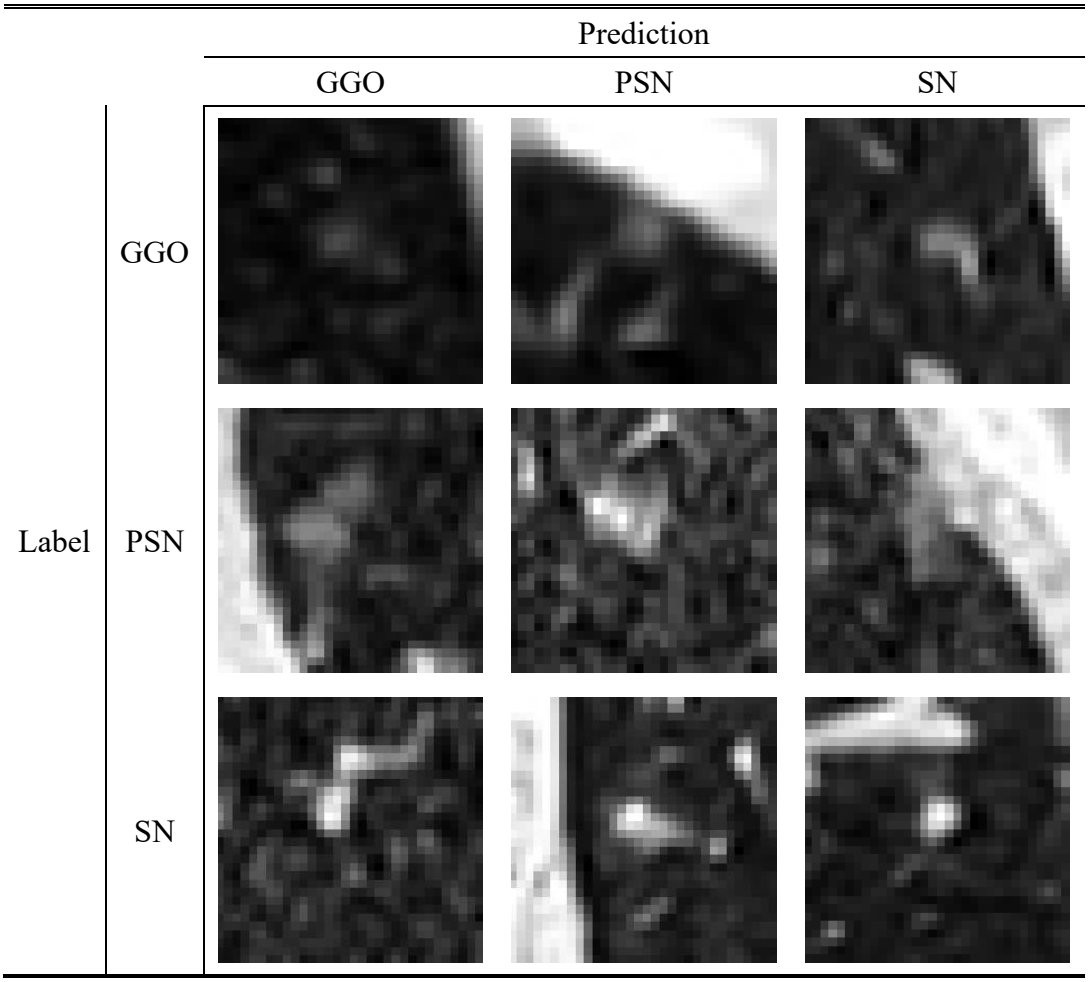


Fig. 4-1 Examples classified by our CADx system.

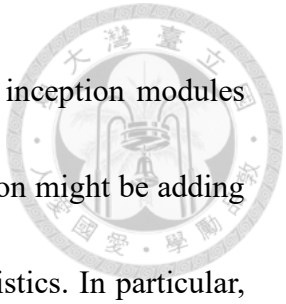
Although the proposed methods achieved favorable scores, there were still three

doi:10.6342/NTU202401286

drawbacks. The first drawback was that the random variation results made observing or reproducing the improvements in the model settings challenging. Specifically, if the performance difference between the two algorithms was slight, determining whether it represented an improvement or merely random variance was difficult. Hence, Renard *et al*. [37] and Bouthillier *et al*. [38] emphasized deep learning results' variability and reproducibility problems in light of this issue. Their research revealed that variability can stem from multiple sources, including datasets, parameter initialization, optimization processes, hyperparameters, and deep learning architectures. They proposed several solutions to address the challenge of variability and enhance the reproducibility and reliability of results, including offering detailed framework descriptions, constructing effective evaluation systems, analyzing and randomizing more sources of variability, and using bootstrapping instead of fixed test sets [37, 38].

The second drawback was that this study proposed the model mainly based on the inception modules with only minor modifications. Although they were relatively old architectures, it was worth noting that they were powerful in image classification [15]. This literature explored various architectural choices and revealed several design principles, ultimately demonstrating that inception modules were optimal for image classification [15]. Furthermore, Kornblith *et al*. evaluated different architectures on multiple datasets and also revealed that inception modules had excellent generalization

doi:10.6342/NTU202401286



abilities [13]. Therefore, any arbitrary changes to the sophisticated inception modules would probably degrade their performance. Instead, a possible solution might be adding some unique processes designed according to the nodule's characteristics. In particular, the examples in Fig. 4-1 suggested that the surrounding tissue might be the main reason for interfering with the model and causing misclassification. Hence, determining whether a given image was a nodule or surrounding tissue could be designed as a pre-train task for the model. Afterward, it might enhance the model's feature representation and improve its performance on the downstream nodule classification task.

The third drawback was the method's suboptimal performance in classifying the PSN class. As illustrated in Fig. 4-1, when darker or brighter surrounding tissues appear adjacent to SN or GGO, respectively, it became easier for the model to misdiagnose them as PSN. A possible solution to improve performance could be derived by observing the differences in characteristics. Liu *et al*. indicated significant differences in the growth characteristics between SN and PSN [39]. The volume doubling time (VDT) and mass doubling time (MDT) of SN were both shorter than those of PSN, meaning that SN grew significantly faster than PSN [39]. Based on this, if follow-up images of the same nodule could be obtained and used together, providing the growth rate as an additional feature to the model could further enhance diagnostic performance.

doi:10.6342/NTU202401286

Chapter 5 Conclusion

This study proposed a CADx system for the classification of lung CT nodules. In this system, image preprocessing was performed on the input 3-D lung CT scans to extract the nodules VOIs, which were resized to 32×32×32 voxels. Afterward, the system fed VOIs into the proposed 3-D SE-Inception for nodule classification. It consisted of the extraction stage, inception stage, attention stage, and a fully connected layer in order. When the VOIs went through these stages, the feature maps were captured and transformed from fine-grained to coarse-grained, with informative features emphasized and less useful ones suppressed by the SE attention. Then, the model outputs the prediction of the GGO, PSN, or SN nodules classes after the final fully connected layer. To tackle the class imbalance and overfitting problems, 3-D SE-Inception was trained with FDBT and data augmentation. The experiment showed that the proposed 3-D SE-Inception trained with FDBT and data augmentation surpassed other architectures and achieved favorable results, with 90.0% micro accuracy, 83.6% macro F1-score, 89.0% F1-score for GGO, 66.1% F1-score for PSN, and 95.6% F1-score for SN. Furthermore, FDBT reduced the overall training time to only one-fourth.

Although the proposed methods achieved the best evaluation scores, three drawbacks remained to be addressed. The first drawback was the random variation made it challenging to observe or reproduce the improvements in the model settings when their

doi:10.6342/NTU202401286

differences were minor. Hence, Renard *et al*. [37] and Bouthillier *et al*. [38] analyzed this issue and proposed several solutions to enhance reproducibility. The solutions included offering detailed framework descriptions, constructing effective evaluation systems, analyzing and randomizing more sources of variability, and using bootstrapping instead of fixed test sets. The second drawback was that this study proposed the model mainly based on the old inception modules with only minor modifications. However, the inception modules were the optimal architectures for image classification with excellent generalization abilities based on large-scale experiments. Hence, adding unique processes designed according to the nodule's characteristics would be more helpful in improving performance than making changes to the sophisticated inception modules. In particular, a possible solution was to pre-train the model to classify nodules or surrounding tissue. The third drawback was the suboptimal performance in classifying PSN nodules. A possible reason that PSNs are prone to be misdiagnosed is because they consist of GGO and SN components with random ratios. In addition, some examples showed that when darker or brighter surrounding tissues appear adjacent to SN or GGO, respectively, it became easier for the model to misdiagnose them as PSN. A feasible solution was to examine the misdiagnosed nodules with experienced radiologists and propose algorithms to tackle the noise characteristics accordingly. In particular, collecting the follow-up images of the same nodule to provide the additional growth rate feature for the model could be helpful.

doi:10.6342/NTU202401286

In summary, conducting future works on the three drawbacks and the corresponding improvement directions will bring more valuable contributions.

doi:10.6342/NTU202401286

References

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram *et al.*, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians,* vol. 74, no. 3, pp. 229-263, FEB 2024, doi: https://doi.org/10.3322/caac.21834.

[2] C. Allemani, T. Matsuda, V. Di Carlo, R. Harewood, M. Matz, M. Nikšić *et al.*, "Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37513025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries," *The Lancet,* vol. 391, no. 10125, pp. 1023-1075, JAN 2018, doi: 10.1016/S0140-6736(17)33326-3.

[3] "Lung Cancer Incidence and Mortality with Extended Follow-up in the National Lung Screening Trial," *Journal of Thoracic Oncology,* vol. 14, no. 10, pp. 1732-1742, 2019/10/01/, MAY 2019, doi: https://doi.org/10.1016/j.jtho.2019.05.044.

[4] H. J. d. Koning, C. M. v. d. Aalst, P. A. d. Jong, E. T. Scholten, K. Nackaerts, M. A. Heuvelmans *et al.*, "Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial," *New England Journal of Medicine,* vol. 382, no. 6, pp. 503-513, FEB 2020, doi: doi:10.1056/NEJMoa1911793.

[5] H. MacMahon, D. P. Naidich, J. M. Goo, K. S. Lee, A. N. Leung, J. R. Mayo *et al.*, "Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017," *Radiology,* vol. 284, no. 1, pp. 228-243, JUL 2017.

[6] B. Jiang, N. Li, X. Shi, S. Zhang, J. Li, G. H. de Bock *et al.*, "Deep Learning Reconstruction Shows Better Lung Nodule Detection for Ultra–Low-Dose Chest CT," *Radiology,* vol. 303, no. 1, pp. 202-212, NOV 2022, doi: 10.1148/radiol.210551.

[7] W. Chen, Y. Wang, D. Tian, and Y. Yao, "CT Lung Nodule Segmentation: A Comparative Study of Data Preprocessing and Deep Learning Models," *IEEE Access,* vol. 11, pp. 34925-34931, MAR 2023, doi: 10.1109/ACCESS.2023.3265170.

[8] S. Tomassini, N. Falcionelli, P. Sernani, L. Burattini, and A. F. Dragoni, "Lung nodule diagnosis and cancer histology classification from computed tomography data by convolutional neural networks: A survey," *Computers in Biology and Medicine,* vol. 146, p. 105691, 2022/07/01/, MAY 2022, doi: https://doi.org/10.1016/j.compbiomed.2022.105691.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, no. 7553, pp. 436-444, 2015/05/01, MAY 2015, doi: 10.1038/nature14539.

doi:10.6342/NTU202401286

[10] J. Pedrosa, G. Aresta, C. Ferreira, G. Atwal, H. A. Phoulady, X. Chen *et al.*, "LNDb challenge on automatic lung cancer patient management," *Medical image analysis,* vol. 70, p. 102027, MAR 2021.

[11] J. Mei, M. M. Cheng, G. Xu, L. R. Wan, and H. Zhang, "SANet: A Slice-Aware Network for Pulmonary Nodule Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 44, no. 8, pp. 4374-4387, AUG 2022, doi: 10.1109/TPAMI.2021.3065086.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, DEC 2016, pp. 770-778.

[13] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, JUN 2019, pp. 2661-2671.

[14] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, AUG 2017, vol. 31, no. 1.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, DEC 2016, pp. 2818-2826.

[16] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, and A. Mian, "Visual attention methods in deep learning: An in-depth survey," *Information Fusion,* vol. 108, p. 102417, APR 2024.

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, MAY 2018, pp. 7132-7141.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez *et al.*, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, AUG 2017.

[19] K. Fukushima, "Cognitron: A self-organizing multilayered neural network," *Biological cybernetics,* vol. 20, no. 3, pp. 121-136, FEB 1975.

[20] P.-F. Verhulst, *Deuxième mémoire sur la loi d'accroissement de la population*. Mémoires de l'Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique: Hayez, 1847.

[21] G. LemaÃŽtre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of machine learning research,* vol. 18, no. 17, pp. 1-5, SEP 2017.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv*

doi:10.6342/NTU202401286

preprint arXiv:1412.6980, JAN 2014.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair *et al.*, "Generative adversarial nets," *Advances in neural information processing systems,* vol. 27, JAN 2014.

[24] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, JAN 2017.

[25] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information processing & management,* vol. 45, no. 4, pp. 427-437, MAY 2009.

[26] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika,* vol. 12, no. 2, pp. 153-157, JUN 1947.

[27] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, JUL 2018, pp. 3-19.

[28] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, MAR 2019, pp. 510-519.

[29] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, APR 2020, pp. 11534-11542.

[30] J.-J. Qiu, J. Yin, W. Qian, J.-H. Liu, Z.-X. Huang, H.-P. Yu *et al.*, "A novel multiresolution-statistical texture analysis architecture: radiomics-aided diagnosis of PDAC based on plain CT images," *IEEE Transactions on Medical Imaging,* vol. 40, no. 1, pp. 12-25, JAN 2020.

[31] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan *et al.*, "Focal self-attention for local-global interactions in vision transformers," *arXiv preprint arXiv:2107.00641*, JUL 2021.

[32] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, MAR 2022, pp. 11976-11986.

[33] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, AUG 2021, pp. 10012-10022.

[34] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR),* vol. 54, no.

doi:10.6342/NTU202401286

10s, pp. 1-41, SEP 2022.

[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, AUG 2009: Ieee, pp. 248-255.

[36] W. Liu, C. Li, M. M. Rahaman, T. Jiang, H. Sun, X. Wu *et al.*, "Is the aspect ratio of cells important in deep learning? A robust comparison of deep learning methods for multi-scale cytopathology cell image classification: From convolutional neural networks to visual transformers," *Computers in biology and medicine,* vol. 141, p. 105026, NOV 2022.

[37] F. Renard, S. Guedria, N. D. Palma, and N. Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Scientific Reports,* vol. 10, no. 1, p. 13724, AUG 2020.

[38] X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto *et al.*, "Accounting for variance in machine learning benchmarks," *Proceedings of Machine Learning and Systems,* vol. 3, pp. 747-769, MAR 2021.

[39] M. Liu, J. Mu, F. Song, X. Liu, W. Jing, and F. Lv, "Growth characteristics of early-stage (IA) lung adenocarcinoma and its value in predicting lymph node metastasis," *Cancer Imaging,* vol. 23, no. 1, p. 115, DEC 2023.

doi:10.6342/NTU202401286