diagnostics

MDPI

Article

Enhanced Malignancy Prediction of Small Lung Nodules in Different Populations Using Transfer Learning on Low-Dose Computed Tomography

Jyun-Ru Chen [1,†], Kuei-Yuan Hou [2,3,†], Yung-Chen Wang [2,4], Sen-Ping Lin [2], Yuan-Heng Mo [2,4], Shih-Chieh Peng [2] and Chia-Feng Lu [1,*]

[1] Department of Biomedical Imaging and Radiological Sciences, National Yang Ming Chiao Tung University, Taipei 112, Taiwan; r1c15j.be08@nycu.edu.tw
[2] Department of Radiology, Cathay General Hospital, Taipei 106, Taiwan; circle@cgh.org.tw (K.-Y.H.); edwardwg@cgh.org.tw (Y.-C.W.); lsp@cgh.org.tw (S.-P.L.); cgh06058@cgh.org.tw (Y.-H.M.); cgh415823@cgh.org.tw (S.-C.P.)
[3] Department of Biomedical Imaging and Radiological Sciences, Chung Shan Medical University, Taichung 402, Taiwan
[4] Department of Medicine, School of Medicine, Fu Jen Catholic University, Taipei 242, Taiwan
* Correspondence: alvin4016@nycu.edu.tw; Tel.: +886-2-2826-7308
† These authors contributed equally to this work.

Abstract: Background: Predicting malignancy in small lung nodules (SLNs) across diverse populations is challenging due to significant demographic and clinical variations. This study investigates whether transfer learning (TL) can improve malignancy prediction for SLNs using low-dose computed tomography across datasets from different countries. **Methods:** We collected two datasets: an Asian dataset (669 SLNs from Cathay General Hospital, CGH, Taiwan) and an American dataset (600 SLNs from the National Lung Screening Trial, NLST, America). Initial U-Net models for malignancy prediction were trained on each dataset, followed by the application of TL to transfer model parameters across datasets. Model performance was evaluated using accuracy, specificity, sensitivity, and the area under the receiver operating characteristic curve (AUC). **Results:** Significant demographic differences ($p < 0.001$) were observed between the CGH and NLST datasets. Initial models trained on one dataset showed a substantial performance decline of 15.2% to 97.9% when applied to the other dataset. TL enhanced model performance across datasets by 21.1% to 159.5% ($p < 0.001$), achieving an accuracy of 0.86–0.91, sensitivity of 0.81–0.96, specificity of 0.89–0.92, and an AUC of 0.90–0.97. **Conclusions:** TL enhances SLN malignancy prediction models by addressing population variations and enabling their application across diverse international datasets.

Keywords: small lung nodule; malignancy prediction; population variations; deep learning; transfer learning

check for updates

Academic Editor: Ioannis M. Tsougos

Received: 23 April 2025
Revised: 1 June 2025
Accepted: 5 June 2025
Published: 8 June 2025

Citation: Chen, J.-R.; Hou, K.-Y.; Wang, Y.-C.; Lin, S.-P.; Mo, Y.-H.; Peng, S.-C.; Lu, C.-F. Enhanced Malignancy Prediction of Small Lung Nodules in Different Populations Using Transfer Learning on Low-Dose Computed Tomography. *Diagnostics* **2025**, *15*, 1460. https://doi.org/10.3390/diagnostics15121460

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The 5-year survival rate of lung cancer patients could be improved by 30% by the patients receiving appropriate treatments at an early stage [1,2]. To achieve early-stage treatments, an effective screening approach for lung nodules is required. However, previous studies using tissue biopsy assessment have reported a more than 13% reduced diagnostic accuracy for small lung nodules (SLNs, less than 10 mm in diameter) compared to large lung nodules (with sizes between 10 and 30 mm) [3,4]. Considering the high prevalence [5]

and low malignancy rate of SLNs [6], predicting the malignancy of SLNs is critical to reducing the waste of medical resources and significantly enhancing patient prognosis.

Deep-learning (DL) algorithms, including architectures like ResNet and U-Net, have demonstrated significant potential in extracting subtle image features from low-dose computed tomography (LDCT) to enhance malignancy prediction in SLNs [7,8]. These models leverage convolutional neural networks (CNNs) to capture complex patterns that may be imperceptible to human observers, thereby enhancing diagnostic accuracy in early lung cancer detection. However, despite these promising results, studies have shown a concern in terms of the poor international application for DL models [9,10], which indicated that DL models trained on data from one country often perform poorly when applied to datasets from different countries, reflecting the challenge of application across populations with varying demographic and clinical characteristics. For instance, Asian and American populations have inconsistent risk factors for lung cancer, including age at diagnosis, smoking history, and SLN types [11–16]. Accordingly, the application of malignancy prediction models based on the American dataset to the Asian dataset may have reduced performance, and vice versa. This limitation of international applications is not unique to lung cancer; similar issues have been reported in DL applications for other diseases, including chronic hepatitis B, coronary arterial calcifications, and mammographic lesions [17–19].

One fundamental barrier to improving international application is the difficulty of exchanging medical imaging data across countries due to privacy concerns and regulatory restrictions. Consequently, many DL models are developed using datasets from a single country or ethnic group, which limits their robustness when deployed in different clinical settings characterized by variations in imaging protocols and population genetics [20]. While pooling data from multiple regions could theoretically improve model generalizability, it also introduces heterogeneity that complicates training and may degrade performance [21]. Techniques such as intensity normalization and spatial smoothing have been employed to mitigate scanner-related variability [22], but these methods cannot fully account for population-level differences in genetics, diet, and lifestyle factors that influence disease presentation. To address these challenges, transfer learning (TL) has emerged as a promising strategy. TL involves reusing feature extraction layers from a model pretrained on one population and fine-tuning the classifier layers on data from a target population, thereby adapting the model to local characteristics without requiring extensive retraining from scratch [23]. We anticipated that TL could facilitate the international application of SLN malignancy prediction by enhancing the model performance across different populations.

In this study, we investigated the impact of population variations on the SLN malignancy prediction between American and Asian datasets and further proposed the TL technique to overcome this influence. This study was designed to achieve the following aims: first, to identify the differences in SLN characteristics between Asian and American datasets; second, to propose a malignancy prediction model for SLNs based on LDCT using a DL approach; third, to assess the reduction of prediction performance when applying models across two datasets; and finally, to evaluate the efficacy of TL in improving prediction models for SLN malignancy.

2. Materials and Methods

2.1. Study Cohort

This study built an Asian dataset of SLNs by retrospective collection of patients with lung nodules in Taipei, Hsinchu, and Sijhih branches of the Cathay General Hospital (CGH) system from 2006 to 2022. Afterward, we excluded patients with unavailable histology for malignancy, insufficient quality of LDCT images for diagnosis, and lung nodules larger than

10 mm. In total, 628 patients with SLNs (530 from Taipei, 42 from Hsinchu, and 56 from Sijhih CGH) were included in the CGH dataset (Figure S1). We also collected a publicly available American dataset (National Lung Screening Trial, NLST) containing LDCT data from The Cancer Imaging Archive (https://www.cancerimagingarchive.net/collection/nlst/, available on 4 April 2025) [24,25]. Among these patients, we excluded patients with the absence of detectable lesions or lung nodules larger than 10 mm, resulting in 7913 patients with SLN. To balance the sample size between the Asian and American datasets, we selected 600 patients from the 7913 SLN patients and ensured that the distributions of age, gender, and malignancy rates of the selected patients were consistent with those of the entire database of the NLST dataset (Figure S2). The inclusion criteria for both datasets were as follows: (a) the diameter of the lung nodule was less than 10 mm; (b) the malignancy of the nodule was confirmed by histology or longitudinal imaging follow-up; (c) the image quality of pre-treatment LDCT was sufficient. Age, gender, SLN volume, and SLN type were also recorded. The institutional review board of CGH approved this study (CGH IRB: CGH-P111079), and informed consent was waived because of the retrospective data collection. Data collection and all research methods in this study were performed in accordance with the Declaration of Helsinki and the regulations of CGH IRB.

2.2. Imaging Parameters of Low-Dose Computed Tomography

The LDCT data in the CGH dataset were obtained using 8 different CT scanners, including Aquilion 640 (Canon Medical Systems, Japan); Aquilion ONE (Canon Medical Systems, Japan); Aquilion Prime SP (Canon Medical Systems, Japan); Brilliance (Philips, Netherlands); Brilliance 64 (Philips, Netherlands); Ingenuity CT (Philips, Netherlands); Sensation 16 (Siemens, German); Vereos PET/CT (Philips, Netherlands). The scanning coverages are similar in all CT scanners, ranging from the thoracic inlet to the upper abdomen. CT images were reconstructed with a slice thickness of 1 to 5 mm. Pixel sizes ranged from 0.483×0.483 mm^2 to 0.912×0.912 mm^2. The matrix size for each slice was 512×512 with a 16-bit gray depth in Hounsfield units (HU). The peak tube voltage was 120 kVp, and the tube current ranged from 20 to 429 mA.

The LDCT data in the NLST dataset were acquired from 8 different CT scanners, including Aquilion (Canon Medical Systems, Japan); HiSpeed QX/i (GE HealthCare, Chicago, Illinois, USA); LightSpeed Plus (GE HealthCare, Chicago, Illinois, USA); LightSpeed QX/i (GE HealthCare, Chicago, Illinois, USA); LightSpeed 16 (GE HealthCare, Chicago, Illinois, USA); Mx8000 (Philips, Netherlands); Sensation 16 (Siemens, German); Volume Zoom (Siemens, German). The scanning coverage, matrix size, and gray depth for the NLST dataset were similar to those of the CGH dataset. CT images were reconstructed with a slice thickness of 2 to 5 mm. Pixel sizes ranged from 0.488×0.488 mm^2 to 0.957×0.957 mm^2. The peak tube voltage was 120 or 140 kVp, and the tube current ranged from 40 to 320 mA. Detailed information on CT scanners and imaging parameters for both datasets is listed in Table S1.

2.3. Segmentation of Small Lung Nodules

The SLNs were delineated by two well-trained radiological technologists and verified by an experienced radiologist. The soft tissue window (window width: 350 HU, window level: 50 HU) and lung window (window width: 1500 HU, window level: −600 HU) on LDCT images were applied for lesion delineation. The soft tissue window was used to distinguish between SLNs and fluid components, such as pleural and pericardial effusions, and the lung window was applied to determine the border of SLNs. Several image preprocessing steps were employed to reduce the scanning variations. The spatial resolution of LDCT was adjusted to a voxel size of $1 \times 1 \times 3$ mm^3. The gray values of LDCT were

adjusted to the lung window, followed by intensity normalization to a range from 0 to 255. The volume of SLNs was cropped into a matrix size of 40 × 40 × 13. Afterward, we separated the CGH and NLST datasets into the training (70% of SLNs) and test (30% of SLNs) sets, respectively.

2.4. Data Augmentation of the Training Sets

The images in the training set from the CGH dataset were augmented thrice by left-right flipping and rotating by 5 and 10 degrees. Contrarily, due to the imbalanced proportion of malignant and benign SLNs in the training set of the NLST dataset, the benign SLN images were augmented twice using left-right flipping and 5-degree rotation. The malignant SLN images in the NLST dataset were augmented six times by upside-down flipping, left-right flipping, and rotating by −10, −5, 5, and 10 degrees. All processes were conducted within the MATLAB R2023a environment (MathWorks Inc., Natick, MA, USA).

2.5. Network Architecture and Model Building

The two-channel U-Net was employed for the malignancy prediction in this study [26], which comprised an input layer, encoder blocks, a bottleneck block, decoder blocks, and a classifier using cross entropy as the loss function (Figure 1). In the first two encoder blocks, feature maps were processed through two pathways with different kernel sizes of convolutional layers: one with the kernel size of 3 × 3 × 1 and the other with the kernel size of 1 × 1 × 3. These maps were normalized via batch normalization, activated using a leaky ReLU, and downsampled with a maximum pooling layer (kernel: 5 × 5 × 5). The feature maps from the two pathways were concatenated in the bottleneck block, followed by a convolutional layer (kernel: 3 × 3 × 3). In the decoder blocks, transposed convolutional layers (kernel: 3 × 3 × 3) were applied, concatenating the two-pathway maps. Finally, fully connected layers, a softmax activation, and a final classifier were applied. The DL models were trained with the following hyperparameters: the stochastic gradient descent with momentum optimizer, an initial learning rate of 0.001, a drop factor of 0.5 for the learning rate, a batch size of 16, an L2 regularization of 0.01, and a momentum of 0.9. For TL, the model was initially pre-trained based on the source dataset (either the CGH or NLST dataset) and then fine-tuned by adjusting the weights of the final 8 layers in the two-channel U-Net based on the target dataset (the other dataset).

Overall, 5 models were built based on the same architecture in this study (Figure 1). **Models 1 (CGH model)** and **Model 2 (NLST model)** were trained solely based on the training set of the CGH and NLST datasets, respectively. **Model 3 (pooling model)** was built by pooling the training sets of the CGH and NLST datasets. **Model 4 (NLST model with TL)** was pre-trained based on the training set of the NLST dataset, followed by the TL on the training set of the CGH dataset. **Model 5 (CGH model with TL)** was pre-trained based on the training set of the CGH dataset, followed by the TL on the training set of the NLST dataset.

2.6. Assessment of Model Performance and Statistical Analysis

Differences in continuous and categorical variables between the CGH and NLST datasets were identified by two-sample t-tests and chi-square tests, respectively. The assessments of model performance were conducted based on the test sets. The sensitivity, specificity, and accuracy derived from the confusion matrix, as well as the area under the receiver operating characteristic curve (AUC), were used to assess the prediction performance of the models. To compare the performance among the five proposed models, the bootstrap method with 100 times resampling followed by a two-sample t-test with a Bonferroni correction was applied. To further explain how the networks predict malignancy probability for SLNs, we employed the analysis of gradient-weighted class activation

mapping (Grad-CAM) in the prediction models [27]. Grad-CAM computed the gradient of the malignancy probability with respect to the feature maps of the final convolutional layer. We averaged the gradients to obtain importance weights, which were then used to produce a heatmap by performing a weighted combination of the feature maps. The resulting activation map highlighted the spatial regions that strongly influenced the model's decision, allowing us to assess whether the model focused on clinically relevant regions/features.

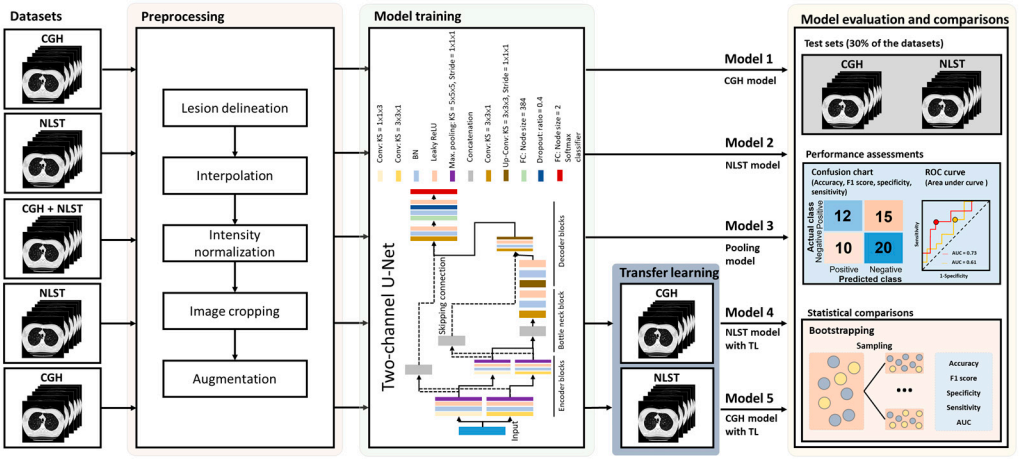

Figure 1. Diagram of study workflow. The LDCT images were preprocessed by resolution adjustment, lesion delineation, intensity normalization, image cropping, and data augmentation. The DL models were built based on processed training-set images using a two-channel U-Net. **Model 1** and **Model 2** were trained based on the training set of the CGH and NLST datasets, respectively. **Model 3** was built by pooling training sets of the CGH and NLST datasets. **Model 4** was trained based on the training set of the NLST dataset, followed by the transfer learning to the training set of the CGH dataset (pale blue block). **Model 5** was trained based on the training set of the CGH dataset, followed by the transfer learning to the training set of the NLST dataset (pale blue block). Assessments of the models were conducted by calculating the accuracy, sensitivity, specificity, and area under the receiver operating characteristics curve (AUC) based on the test sets of the CGH and NLST datasets, respectively. The statistical comparison between models was performed using the bootstrap method followed by a two-sample t-test. Conv: convolution layer, KS: kernel size, BN: batch normalization layer, Leaky ReLU: leaky Rectified Linear Unit, Concat: concatenate layer, up-Conv: transpose convolution layer, FC: fully connected layer.

3. Results

3.1. Demography of Study Cohorts

The study collected 669 lesions from 628 Asian patients at Cathay General Hospital (CGH), referred to as the CGH dataset. Additionally, we also collected an American dataset by downloading 600 lesions from 600 patients in the National Lung Screening Trial (NLST), designated as the NLST dataset (Table 1). The age at diagnosis in the CGH da-taset (59.45 ± 13.07) was significantly younger ($p < 0.001$) than in the NLST dataset (61.48 ± 4.71). Females comprised a higher proportion ($p < 0.001$) in the CGH dataset (58.76%) compared to the NLST dataset (39.17%). No significant difference in the equivalent diameter and volume of SLNs was found between the two datasets. A higher proportion ($p < 0.001$) of malignancy was observed in the CGH dataset (47.09%) compared to the NLST dataset (28.83%). The distribution of SLN types (solid/partial solid/ground-glass opacity (GGO)) was significantly different ($p < 0.001$) between the CGH dataset (40.66%/30.79%/28.55%) and the NLST dataset (71.17%/10.00%/18.83%).

Table 1. Demography of study cohorts.

Characteristics	CGH Dataset	NLST Dataset	*p* Value
Patient number	628	600	
Age at diagnosis	59.45 ± 13.07	61.48 ± 4.71	<0.001 *
Gender (M/F)	259 (41.24%)/369 (58.76%)	365 (60.83%)/235 (39.17%)	<0.001 *
SLN number	669	600	
Pathology			<0.001 *
Benignness	354 (52.91%)	427 (71.17%)	
Malignancy	315 (47.09%)	173 (28.83%)	
Eq. Diameter (mm)	3.48 ± 1.86	4.37 ± 1.58	0.792
Volume (mm^3)	557.96 ± 764.60	503.25 ± 619.38	0.453
SLN types			<0.001 *
Solid	272 (40.66%)	427 (71.17%)	
Partial solid	206 (30.79%)	60 (10.00%)	
GGO	191 (28.55%)	113 (18.83%)	

* $p < 0.001$.

Additionally, we found that the malignancy probability of partial solid SLNs in the CGH dataset (49.03%) was significantly higher ($p = 0.005$) compared to the NLST dataset (28.33%, Table S2). A similar finding was observed for GGO SLNs, with the malignancy probability being significantly higher ($p < 0.001$) in the CGH dataset (64.40%) compared to the NLST dataset (24.78%, Table S2).

3.2. Reduced Model Performance in Different Datasets

Given these differences in SLN type distributions between the CGH and NLST datasets (Table 1), we further assessed whether the different distribution of SLN types between the CGH and NLST datasets leads to a reduction in model performance across populations. We separately developed prediction models for SLN malignancy based on the CGH and NLST datasets, referred to as **Model 1 (CGH model)** and **Model 2 (NLST model)**, and further applied these models to another dataset. Our results showed significant reductions in model performance when employing prediction models to different test populations (Figure 2). For instance, **Model 1 (CGH model)** showed significantly reduced performance ($p < 0.001$) in the test set of the NLST dataset (reductions in accuracy, sensitivity, and AUC of 18.4%, 97.9%, and 52.6%, respectively) compared to the test set of the CGH dataset (Figure 2A). **Model 2 (NLST model)** showed significantly reduced performance ($p < 0.001$) in the test set of the CGH dataset (reductions in accuracy, specificity, sensitivity, and AUC of 34.1%, 15.2%, 57.0%, and 43.6%, respectively) compared to the test set of the NLST dataset (Figure 2B).

3.3. Impact of Population Variation on Model Performance

A potential strategy to develop a prediction model with international application is training models using pooled datasets from different populations. In this study, we further investigated whether a prediction model (**Model 3**, **pooling model**) using both populations for training would improve the generalization of SLN malignancy prediction (Figure 2C). **Model 3** showed smaller differences in model performance (absolute differences in accuracy, sensitivity, and AUC of 0.16, 0.07, and 0.20, respectively) between the CGH and NLST dataset test sets compared to **Models 1** and **2**.

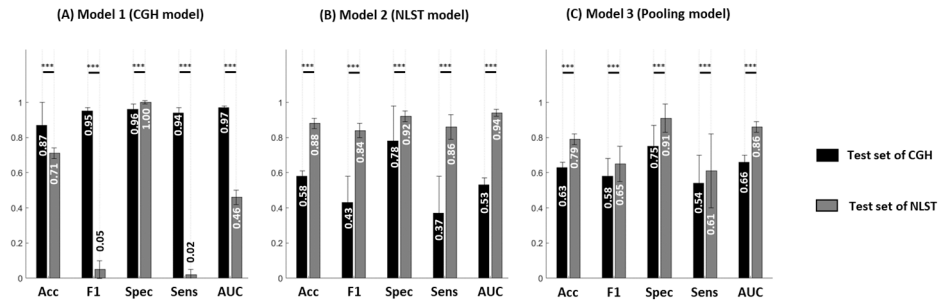

Figure 2. Comparisons of performance between the CGH and NLST datasets for two-channel U-Net models. (**A**) **Model 1** is the prediction model based on the CGH dataset. (**B**) **Model 2** is based on the NLST dataset. (**C**) **Model 3** is based on the pooling dataset. Each bar chart contains accuracy, F1 score, specificity, sensitivity, and AUC. Black and gray bars indicate the model performance in the test set of the CGH and NLST datasets, respectively. *** $p < 0.001$.

Figure S3 shows the performance comparisons between **Models 1**, **2**, and **3** in the test sets of the CGH and NLST datasets. **Model 3** showed significantly poorer ($p < 0.001$) performance in the test set of the CGH dataset (reductions in accuracy, specificity, sensitivity, and AUC of 27.6%, 21.9%, 42.6%, and 32.0%, respectively) than in **Model 1**. The performance of **Model 3** is significantly lower ($p < 0.001$) in the test set of the NLST dataset (reductions in accuracy, sensitivity, and AUC of 10.2%, 29.1%, and 8.5%, respectively) than in **Model 2**. Accordingly, training a model by pooling the SLNs from different datasets reduced the performance for SLN malignancy prediction in both test sets.

3.4. Improvement of Model Performance Through TL

To overcome the challenge of the low performance induced by population variation and to facilitate international application, TL was applied to **Model 2 (NLST model)** to create **Model 4 (TL from NLST to CGH dataset)**, where the model parameters were fine-tuned in the final layers using the CGH dataset. Similarly, TL was applied to **Model 1 (CGH model)** to create **Model 5 (TL from CGH to NLST dataset)**, where the model parameters were fine-tuned in the final layers using the NLST dataset. Figure 3 shows the performance comparisons between models with and without TL. **Model 4 (TL from NLST to CGH dataset)** significantly enhanced ($p < 0.001$) the performance (increases in accuracy, F1 score, specificity, sensitivity, and AUC of 56.9%, 116.3%, 14.1%, 159.5%, and 83.0%, respectively) in the test set of CGH dataset compared to **Model 2** (Figure 3A). **Model 5 (TL from CGH to NLST dataset)** significantly enhanced ($p < 0.001$) the performance (increases in accuracy, F1 score, sensitivity, and AUC of 21.1%, 1520.0%, 3950.0%, and 95.7%, respectively) in the test set of the NLST dataset compared to **Model 1** (Figure 3B).

In the comparisons between the pooling model (**Model 3**) and models with TL, **Model 4** showed significantly better ($p < 0.001$) performance (increases in accuracy, F1 score, specificity, sensitivity, and AUC of 44.4%, 60.3%, 18.7%, 77.8%, and 47.0%, respectively) in the test set of CGH dataset compared to **Model 3** (Figure 3A). **Model 5** showed significantly better ($p < 0.001$) performance (increases in accuracy, F1 score, sensitivity, and AUC of 8.9%, 24.6%, 32.8%, and 4.7%, respectively) in the test set of the NLST dataset compared to **Model 3** (Figure 3B). Accordingly, TL could overcome the cross-population variation by adapting the parameters of the pre-trained model (built based on the source population) to the target population.

3.5. Training Time for the Proposed Models

The training time for the base models was 66.5 min for **Model 1**, 58.9 min for **Model 2**, and 130.5 min for **Model 3**. The training time for the TL was 21.4 min in **Model 4** and 13.8 min in **Model 5**.

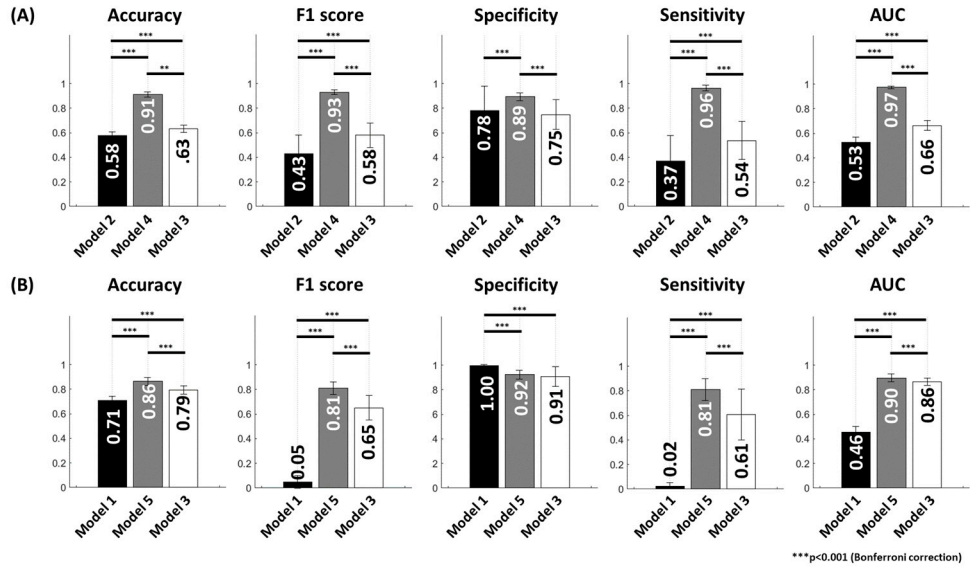

Figure 3. Performance comparisons in the target dataset among specialized models and TL models. (**A**) Accuracy, F1 scores, specificity, sensitivity, and AUC of **Model 2**, **Model 4**, and **Model 3** in the test set of the CGH dataset, respectively. (**B**) Accuracy, F1 scores, specificity, sensitivity, and AUC of **Model 1**, **Model 5**, and **Model 3** in the test set of the NLST dataset, respectively. ** $p < 0.01$, *** $p < 0.001$.

3.6. Demonstrative Cases of Malignancy Prediction

To explore the interpretability of **Model 2 (NLST model)** and **Model 4 (TL from NLST to CGH dataset)**, the Grad-CAM was analyzed on two representative SLN cases from the CGH dataset (Figure 4). Case #1 was diagnosed with a benign solid SLN measuring 6.87 mm in diameter. In contrast to **Model 2**, which considered a more nonspecific area, including the ribs, **Model 4** focused on the lesion, associated vessels, and pleural retraction after TL. Case #2 was diagnosed with a malignant GGO SLN 1.56 mm in diameter. In contrast to **Model 2**, **Model 4** focused on the lesion, the vascular distribution, and the gap between SLN and pleura after TL. In summary, the model with TL could pay more attention to lesion-related areas and make a more accurate malignancy prediction.

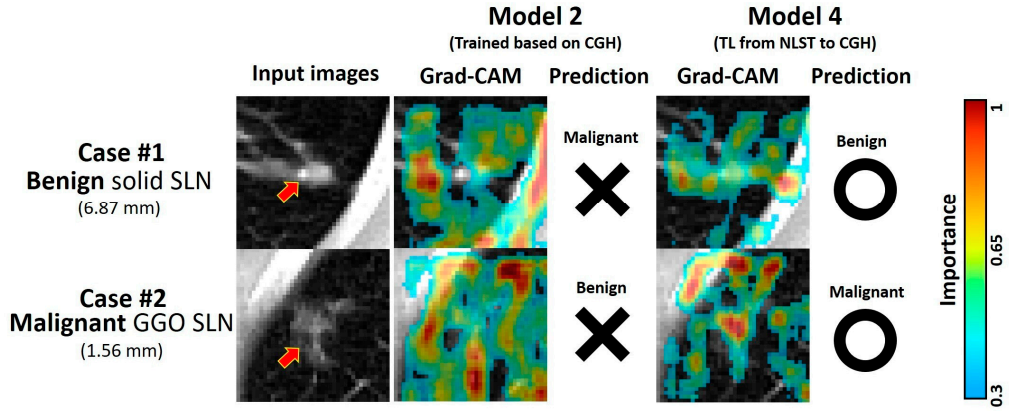

Figure 4. Grad-CAM and prediction for two representative SLN cases, which were mispredicted by **Model 2** but correctly predicted by **Model 4**. The Grad-CAM represents the importance of areas for malignancy prediction in the model, with red and blue colors representing high and low importance, respectively. The red arrows indicate the location of the SLNs. The circle and cross symbols indicate the correctness of the model prediction.

4. Discussion

In this study, the CGH (Asian) and NLST (American) datasets showed differences in age at diagnosis, gender, and SLN types. We identified a reduced performance of specialized models (**Models 1** and **2**) in these distinct datasets (Figure 5A), potentially due to the population variations between datasets (Table S2). Additionally, cross-population variations after pooling different datasets may also reduce the modeling performance in both datasets (Figure 5B). This study demonstrated that the models with TL showed enhanced performance compared to the initial models, hence improving the international application of prediction models for SLN malignancy across different populations (Figure 5C).

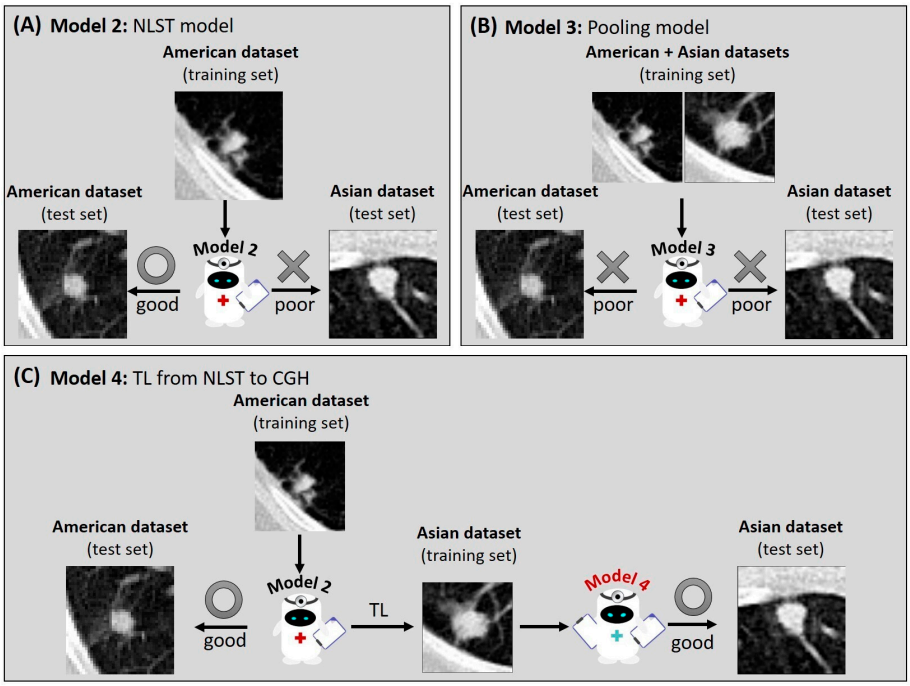

Figure 5. A diagram summarizing the benefits of TL. (**A**) **Model 2**, built based on the American dataset, performed poorly in the Asian dataset. (**B**) **Model 3**, trained by pooling American and Asian training sets, showed low performance in both datasets. (**C**) **Model 4**, built with TL from American to Asian datasets, overcame the cross-population variation and enhanced model performance to facilitate international application. The circle and cross symbols represent good and poor model performance in the test sets, respectively.

In this study, over 50% of cases in the CGH dataset were female. The CGH dataset also presented a younger age at diagnosis than the NLST dataset. The literature reports a young age and high proportion of female cases in the Asian cohort due to exposure to cooking oil fumes [11,28,29]. Compared to Americans, partial and GGO SLNs in Asians showed a distinctly higher malignancy risk, which has been revealed in the previous comparative study [15,16]. The variations in individual and SLN imaging characteristics across populations may mislead malignancy prediction when employing a model developed in different countries.

One potential solution to achieve satisfactory model performance across populations is to construct a generalized model by pooling datasets from multiple countries. However, our results showed that **Model 3**, based on a pooling dataset, could not effectively improve model performance (Figure 2), even though the imaging parameters (including image intensity and voxel size) had been adjusted to eliminate scanning variations [22]. Pooling different datasets with population inconsistencies can generate a high variance within the training set and interfere with the learning process of imaging patterns. Therefore,

we considered that the poor performance of **Model 3** might be attributed to population variations. Instead of data pooling, an alternative approach to accommodate these inherent variations is required to facilitate the international application of the AI model.

This study identified the benefits of TL in enhancing the model performance in different datasets (Figure 3). First, the training times for **Models 1**, **2**, **3**, **4**, and **5** are 66.5, 58.9, 130.5, 21.4, and 13.8 min, respectively. The shorter training times observed in the TL models suggest that TL could adapt model parameters to population variations more efficiently than retraining the whole model. Second, compared to the specialized model, the TL models focused on SLN-associated features, including fibrosis, vascular distribution, and pleural retraction (Figure 4), all commonly used to evaluate nodule malignancy by radiologists and pulmonologists [30–33]. Furthermore, due to various image traits for SLNs, we also investigated the improvement of model performance by TL in different SLN types. Our sub-analysis showed that TL significantly improved performance across all three types in test sets of both the CGH and NLST datasets (Figure S4). We suggested that fine-tuning model parameters in TL could improve model performance, enabling the model to recognize the characteristics of different SLN types specific to the target population. Accordingly, TL could overcome population variations and facilitate the international application of malignancy prediction in SLN. Additionally, TL avoided data transfer across institutes and reduced the potential risk of data leakage. The TL technique offered a solution for developing international applications while ensuring compliance with privacy regulations in a short training time.

Limitations in this study were addressed as follows. First, because of the lack of clinical factors in the NLST dataset, we could only build the prediction model of SLN malignancy based on images. Although the proposed TL models could achieve an AUC above 0.90, additional benefits from including clinical risk factors, such as tuberculosis and chronic obstructive pulmonary disease [34], require further studies to confirm. Second, we reported superior performance of the employed two-channel U-Net compared to the ResNet and traditional U-Net for predicting malignancy of SLNs (Table S3 and Figure S5). Using advanced deep-learning architectures that are suitable for small lesions may further improve the malignancy prediction. Third, future studies, including more populations in addition to Asian and American datasets, might fully explore TL's efficacy in facilitating the international application of SLN malignancy prediction.

This study observed the disparities in age at diagnosis, gender, SLN types, and malignancy rates between Asian and American datasets. Even though the images underwent harmonization, this study observed reduced performance when applying the prediction model to the independent external dataset. Finally, we suggest that TL could improve model performance and facilitate the international application of malignancy prediction models for SLNs.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics15121460/s1, Table S1: CT scanners and imaging parameters in the CGH and NLST datasets; Table S2: The malignancy of three SLN types in the CGH and NLST dataset; Table S3: The network architecture for the DL models; Figure S1: The patient enrollment flowchart for the CGH dataset; Figure S2: The patient enrollment flowchart for the NLST dataset; Figure S3: Performance comparisons in the target dataset among specialized models and pooling model; Figure S4: Comparisons of accuracy, F1 score, specificity, sensitivity, and AUC between models with and without TL in the target datasets in different SLN types; Figure S5: Performance comparisons in the target dataset among ResNet-18, U-Net, and 2C-U-Net. Refs [35,36].

Author Contributions: Conception and design, J.-R.C., K.-Y.H., and C.-F.L.; acquisition of data, J.-R.C., K.-Y.H., Y.-C.W., S.-P.L., Y.-H.M., and S.-C.P.; development of methodology, J.-R.C., K.-Y.H., and C.-F.L.; funding acquisition, K.-Y.H. and C.-F.L.; supervision, Y.-C.W., S.-P.L., Y.-H.M., and

C.-F.L.; visualization, J.-R.C., K.-Y.H., and C.-F.L.; writing—original draft, J.-R.C., K.-Y.H., and C.-F.L.; writing—review and editing, J.-R.C., K.-Y.H., Y.-C.W., S.-P.L., Y.-H.M., S.-C.P., and C.-F.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Veterans General Hospitals and University System of Taiwan Joint Research Program: VGHUST114-G1-2-3; Cathay General Hospital: CGH-MR-A111136; National Science and Technology Council: NSTC 113-2314-B-A49-048-MY3.

Institutional Review Board Statement: The Institutional Review Board of Cathay General Hospital approved this study (CGH IRB: CGH-P111079) on 8th February 2023, and the informed consent was waived because of the retrospective data collection. Data collection and all research methods in this study were performed in accordance with the Declaration of Helsinki and the regulations of CGH IRB.

Informed Consent Statement: Written informed consent was waived by the institutional review board of CGH due to the retrospective data collection.

Data Availability Statement: The raw data for the CGH dataset cannot be publicly available for ethical and legal reasons. However, researchers can submit inquiries for analyzed data to the corresponding authors by reasonable request. The NLST dataset can be accessed through the data portal of the Cancer Imaging Archive (https://www.cancerimagingarchive.net/collection/nlst/ available till 22 April 2025).

Acknowledgments: This work was supported by the Veterans General Hospitals and University System of Taiwan Joint Research Program (VGHUST114-G1-2-3), Cathay General Hospital (CGH-MR-A111136), and the National Science and Technology Council (NSTC 113-2314-B-A49-048-MY3). The funding bodies had no role in the design of the study, the analysis and interpretation of data, or in the manuscript crafting. This work is part of the doctoral dissertation of Jyun-Ru Chen at the department of Biomedical Imaging and Radiological Sciences, National Yang Ming Chiao Tung University.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SLN	Small lung nodule
DL	Deep-learning
LDCT	Low-dose computed tomography
TL	Transfer learning
CGH	Cathay General Hospital
NLST	National Lung Screening Trial
IRB	Institutional review board
HU	Hounsfield unit
ReLU	Rectified linear unit
AUC	Area under the receiver operating characteristic curve
Grad-CAM	Gradient-weighted class activation map

References

1. Yang, C.Y.; Lin, Y.T.; Lin, L.J.; Chang, Y.H.; Chen, H.Y.; Wang, Y.P.; Shih, J.Y.; Yu, C.J.; Yang, P.C. Stage Shift Improves Lung Cancer Survival: Real-World Evidence. *J. Thorac. Oncol.* **2023**, *18*, 47–56. [CrossRef] [PubMed]
2. He, S.; Li, H.; Cao, M.; Sun, D.; Yang, F.; Yan, X.; Zhang, S.; He, Y.; Du, L.; Sun, X.; et al. Survival of 7,311 lung cancer patients by pathological stage and histological classification: A multicenter hospital-based study in China. *Transl. Lung Cancer Res.* **2022**, *11*, 1591–1605. [CrossRef] [PubMed]
3. Huang, M.-D.; Weng, H.-H.; Hsu, S.-L.; Hsu, L.-S.; Lin, W.-M.; Chen, C.-W.; Tsai, Y.-H. Accuracy and complications of CT-guided pulmonary core biopsy in small nodules: A single-center experience. *Cancer Imaging* **2019**, *19*, 51. [CrossRef]

4. Uzun, Ç.; Akkaya, Z.; Düşünceli Atman, E.; Üstüner, E.; Peker, E.; Gülpınar, B.; Elhan, A.H.; Ceyhan, K.; Atasoy, K. Diagnostic accuracy and safety of CT-guided fine needle aspiration biopsy of pulmonary lesions with non-coaxial technique: A single center experience with 442 biopsies. *Diagn. Interv. Radiol.* **2017**, *23*, 137–143. [CrossRef]
5. Cai, J.; Vonder, M.; Du, Y.; Pelgrim, G.J.; Rook, M.; Kramer, G.; Groen, H.J.M.; Vliegenthart, R.; de Bock, G.H. Who is at risk of lung nodules on low-dose CT in a Western country? A population-based approach. *Eur. Respir. J.* **2024**, *63*, 2301736. [CrossRef]
6. McWilliams, A.; Tammemagi, M.C.; Mayo, J.R.; Roberts, H.; Liu, G.; Soghrati, K.; Yasufuku, K.; Martel, S.; Laberge, F.; Gingras, M.; et al. Probability of cancer in pulmonary nodules detected on first screening CT. *N. Engl. J. Med.* **2013**, *369*, 910–919. [CrossRef] [PubMed]
7. Zhang, R.; Wei, Y.; Wang, D.; Chen, B.; Sun, H.; Lei, Y.; Zhou, Q.; Luo, Z.; Jiang, L.; Qiu, R.; et al. Deep learning for malignancy risk estimation of incidental sub-centimeter pulmonary nodules on CT images. *Eur. Radiol.* **2024**, *34*, 4218–4229. [CrossRef]
8. Cui, S.L.; Qi, L.L.; Liu, J.N.; Li, F.L.; Chen, J.Q.; Cheng, S.N.; Xu, Q.; Wang, J.W. A prediction model based on computed tomography characteristics for identifying malignant from benign sub-centimeter solid pulmonary nodules. *J. Thorac. Dis.* **2024**, *16*, 4238–4249. [CrossRef]
9. Yu, A.C.; Mohajer, B.; Eng, J. External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. *Radiol. Artif. Intell.* **2022**, *4*, e210064. [CrossRef]
10. Papalampidou, A.; Papoutsi, E.; Katsaounou, P.A. Pulmonary nodule malignancy probability: A diagnostic accuracy meta-analysis of the Mayo model. *Clin. Radiol.* **2022**, *77*, 443–450. [CrossRef]
11. Lam, D.C.; Liam, C.K.; Andarini, S.; Park, S.; Tan, D.S.W.; Singh, N.; Jang, S.H.; Vardhanabhuti, V.; Ramos, A.B.; Nakayama, T.; et al. Lung Cancer Screening in Asia: An Expert Consensus Report. *J. Thorac. Oncol.* **2023**, *18*, 1303–1322. [CrossRef] [PubMed]
12. Balzer, B.W.R.; Loo, C.; Lewis, C.R.; Trahair, T.N.; Anazodo, A.C. Adenocarcinoma of the Lung in Childhood and Adolescence: A Systematic Review. *J. Thorac. Oncol.* **2018**, *13*, 1832–1841. [CrossRef] [PubMed]
13. Jung, K.J.; Jeon, C.; Jee, S.H. The effect of smoking on lung cancer: Ethnic differences and the smoking paradox. *Epidemiol. Health* **2016**, *38*, e2016060. [CrossRef]
14. Brenner, D.R.; Hung, R.J.; Tsao, M.S.; Shepherd, F.A.; Johnston, M.R.; Narod, S.; Rubenstein, W.; McLaughlin, J.R. Lung cancer risk in never-smokers: A population-based case-control study of epidemiologic risk factors. *BMC Cancer* **2010**, *10*, 285. [CrossRef] [PubMed]
15. Lui, N.S.; Benson, J.; He, H.; Imielski, B.R.; Kunder, C.A.; Liou, D.Z.; Backhus, L.M.; Berry, M.F.; Shrager, J.B. Sub-solid lung adenocarcinoma in Asian versus Caucasian patients: Different biology but similar outcomes. *J. Thorac. Dis.* **2020**, *12*, 2161–2171. [CrossRef]
16. Qin, Y.; Xu, Y.; Ma, D.; Tian, Z.; Huang, C.; Zhou, X.; He, J.; Liu, L.; Guo, C.; Wang, G.; et al. Clinical characteristics of resected solitary ground-glass opacities: Comparison between benign and malignant nodules. *Thorac. Cancer* **2020**, *11*, 2767–2774. [CrossRef]
17. Kim, H.Y.; Lampertico, P.; Nam, J.Y.; Lee, H.C.; Kim, S.U.; Sinn, D.H.; Seo, Y.S.; Lee, H.A.; Park, S.Y.; Lim, Y.S.; et al. An artificial intelligence model to predict hepatocellular carcinoma risk in Korean and Caucasian patients with chronic hepatitis B. *J. Hepatol.* **2022**, *76*, 311–318. [CrossRef]
18. Gernaat, S.A.M.; van Velzen, S.G.M.; Koh, V.; Emaus, M.J.; Isgum, I.; Lessmann, N.; Moes, S.; Jacobson, A.; Tan, P.W.; Grobbee, D.E.; et al. Automatic quantification of calcifications in the coronary arteries and thoracic aorta on radiotherapy planning CT scans of Western and Asian breast cancer patients. *Radiother. Oncol.* **2018**, *127*, 487–492. [CrossRef]
19. Yamaguchi, T.; Inoue, K.; Tsunoda, H.; Uematsu, T.; Shinohara, N.; Mukai, H. A deep learning-based automated diagnostic system for classifying mammographic lesions. *Medicine* **2020**, *99*, e20977. [CrossRef]
20. Yan, K.; Cai, J.; Zheng, Y.; Harrison, A.; Jin, D.; Tang, Y.-B.; Tang, Y.-X.; Huang, L.; Xiao, J.; Lu, L. Learning From Multiple Datasets With Heterogeneous and Partial Labels for Universal Lesion Detection in CT. *IEEE Trans. Med. Imaging* **2020**, *40*, 2759–2770. [CrossRef]
21. Baltagi, B.H.; Griffin, J.M.; Xiong, W. To Pool or Not to Pool: Homogeneous Versus Heterogeneous Estimators Applied to Cigarette Demand. *Rev. Econ. Stat.* **2000**, *82*, 117–126. [CrossRef]
22. Seoni, S.; Shahini, A.; Meiburger, K.M.; Marzola, F.; Rotunno, G.; Acharya, U.R.; Molinari, F.; Salvi, M. All you need is data preparation: A systematic review of image harmonization techniques in Multi-center/device studies for medical support systems. *Comput. Methods Programs Biomed.* **2024**, *250*, 108200. [CrossRef]
23. Iman, M.; Arabnia, H.R.; Rasheed, K. A Review of Deep Transfer Learning and Recent Advancements. *Technologies* **2023**, *11*, 40. [CrossRef]
24. Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J. Digit. Imaging* **2013**, *26*, 1045–1057. [CrossRef] [PubMed]

25. National Lung Screening Trial Research, T.; Aberle, D.R.; Berg, C.D.; Black, W.C.; Church, T.R.; Fagerstrom, R.M.; Galen, B.; Gareen, I.F.; Gatsonis, C.; Goldin, J.; et al. The National Lung Screening Trial: Overview and study design. *Radiology* **2011**, *258*, 243–253. [CrossRef]
26. Lee, W.K.; Wu, C.C.; Lee, C.C.; Lu, C.F.; Yang, H.C.; Huang, T.H.; Lin, C.Y.; Chung, W.Y.; Wang, P.S.; Wu, H.M.; et al. Combining analysis of multi-parametric MR images into a convolutional neural network: Precise target delineation for vestibular schwannoma treatment planning. *Artif. Intell. Med.* **2020**, *107*, 101911. [CrossRef]
27. Quach, L.D.; Quoc, K.N.; Quynh, A.N.; Thai-Nghe, N.; Nguyen, T.G. Explainable Deep Learning Models With Gradient-Weighted Class Activation Mapping for Smart Agriculture. *IEEE Access* **2023**, *11*, 83752–83762. [CrossRef]
28. Zhang, X.; Rao, L.; Liu, Q.; Yang, Q. Meta-analysis of associations between cooking oil fumes exposure and lung cancer risk. *Indoor Built Environ.* **2021**, *31*, 820–837. [CrossRef]
29. Xue, Y.; Jiang, Y.; Jin, S.; Li, Y. Association between cooking oil fume exposure and lung cancer among Chinese nonsmoking women: A meta-analysis. *Oncol. Targets Ther.* **2016**, *9*, 2987–2992. [CrossRef]
30. Manos, D.; Seely, J.M.; Taylor, J.; Borgaonkar, J.; Roberts, H.C.; Mayo, J.R. The Lung Reporting and Data System (LU-RADS): A Proposal for Computed Tomography Screening. *Can. Assoc. Radiol. J.* **2014**, *65*, 121–134. [CrossRef]
31. Abu Qubo, A.; Numan, J.; Snijder, J.; Padilla, M.; Austin, J.H.M.; Capaccione, K.M.; Pernia, M.; Bustamante, J.; O'Connor, T.; Salvatore, M.M. Idiopathic pulmonary fibrosis and lung cancer: Future directions and challenges. *Breathe* **2022**, *18*, 220147. [CrossRef] [PubMed]
32. Wang, X.; Leader, J.K.; Wang, R.; Wilson, D.; Herman, J.; Yuan, J.M.; Pu, J. Vasculature surrounding a nodule: A novel lung cancer biomarker. *Lung Cancer* **2017**, *114*, 38–43. [CrossRef] [PubMed]
33. Zhao, H.C.; Xu, Q.S.; Shi, Y.B.; Ma, X.J. Clinical-radiological predictive model in differential diagnosis of small ($\leq$ 20 mm) solitary pulmonary nodules. *BMC Pulm. Med.* **2021**, *21*, 281. [CrossRef] [PubMed]
34. Ang, L.; Ghosh, P.; Seow, W.J. Association between previous lung diseases and lung cancer risk: A systematic review and meta-analysis. *Carcinogenesis* **2021**, *42*, 1461–1474. [CrossRef]
35. Ebrahimi, A.; Luo, S.; Chiong, R. Introducing Transfer Learning to 3D ResNet-18 for Alzheimer's Disease Detection on MRI Images. In Proceedings of the 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ), Wellington, New Zealand, 25–27 November 2020; pp. 1–6.
36. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016, Athens, Greece, 17–21 October 2016; pp. 424–432.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.