Informatics in Medicine Unlocked 65 (2026) 101799

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Bias in developing AI medical diagnosis: Impact of sequential CT scan slices on data leakage in lung cancer

Fredy Rojas, Samaneh Madanian *

Department of Computer and Information Sciences, Auckland University of Technology, Auckland, 1010, New Zealand
AUT Institute of Biomedical Technologies, Auckland University of Technology, Auckland, 1010, New Zealand

ARTICLE INFO

Dataset link: https://data.mendeley.com/datasets/bhmdr45bh2/4

Keywords:
Lung cancer
Digital health
Model evaluation protocol
Bias
Data leakage
Medical imaging
AI in healthcare
Reproducibility

ABSTRACT

Lung cancer is highly heterogeneous, making reliable evaluation of AI-based computer-aided diagnosis systems challenging. Public CT datasets such as IQ-OTH/NCCD are valuable for research, but the released dataset lacks subject identifiers, preventing verified subject-wise train–test partitioning and increasing the risk that slices from the same patient may appear in both training and test sets. This study re-examines IQ-OTH/NCCD to assess how partitioning affects performance and performance-overestimation risk.

We evaluated 14 deep-learning pipeline configurations that combined preprocessing, CNN architectures, partitioning, and training strategies. Performance was compared under two conditions: With Order, which preserved the released CT-slice sequence, and Without Order, which shuffled images before fold assignment. Under the null hypothesis, the two conditions were expected to produce comparable performance estimates. Systematically higher performance under Without Order was interpreted as evidence that this partitioning condition may produce more optimistic performance estimates. Performance was evaluated using accuracy, precision, recall, and F1-score

Results showed consistently higher performance estimates under the Without Order condition across the evaluated DL pipelines. In 100 repetitions of stratified 5KfoldCV, mean accuracy was 24.12 percentage points higher for the 2DCNN pipeline (95% CI: 23.52–24.74) and 12.42 percentage points higher for the ResNet pipeline (95% CI: 12.19–12.67). Both differences were statistically significant ($p < 0.001$). These findings suggest that the Without Order condition can produce more optimistic performance estimates when applied to IQ-OTH/NCCD.

This study raises awareness of the risk of performance overestimation without discouraging the use of IQ-OTH/NCCD. Rather, the findings show that partitioning choices can substantially influence performance estimates when subject identifiers are unavailable. We therefore recommend explicit reporting of data partitioning strategies and the provision of reproducible code within a containerised environment to support transparent evaluation.

1. Introduction

Lung cancer remains one of the leading causes of cancer-related mortality worldwide [1]. It is not a single disease but a spectrum of subtypes that can vary significantly in presentation across patients. Increasing evidence suggests that lung cancer exhibits considerable histological and molecular heterogeneity, even within the same histological subtype [2–4]. This complexity presents challenges for accurate and consistent diagnoses. Conversely, traditional imaging-based evaluations rely on visual inspection by specialists, introducing subjectivity and variability between observers [5]. For example, differentiating small-cell lung cancer (SCLC) from non-small-cell lung cancer (NSCLC) can be challenging, as subtle imaging differences are often difficult to discern visually with the naked eyes [6].

To address these limitations, artificial intelligence (AI)-based computer-aided diagnosis (CAD) systems have shown promising results in supporting radiologists and improving diagnostic accuracy [5]. However, the effectiveness of CAD outcomes depends on the diversity and volume of available training data. To illustrate the complexity of lung cancer detection, Table 1 presents the categories recognised by the World Health Organisation (WHO), with each ICD-O morphology code representing a distinct subtype of lung cancer. As more subtypes are targeted for detection, the need for larger datasets that capture this heterogeneity increases to enhance AI models and generalisation.

* Correspondence to: Department of Computer and Information Sciences, Auckland University of Technology, Auckland, New Zealand
E-mail address: sam.madanian@aut.ac.nz (S. Madanian).

https://doi.org/10.1016/j.imu.2026.101799
Received 27 April 2026; Received in revised form 13 July 2026; Accepted 29 July 2026
Available online 20 August 2026

Table 1
WHO 2021 classification of lung tumor categories and corresponding ICD-O morphology codes, illustrating the histological heterogeneity of lung cancer (adapted from [2]).

Categories	ICD-O morphology codes	Categories	ICD-O morphology codes
Epithelial tumors		*Mesenchymal tumors specific to the lung*	
Papillomas	8052/0, 8053/0, 8260/0, 8560/0	Pulmonary hamartoma	8992/0
Adenomas	8832/0, 8251/0, 8260/0, 8140/0, 8470/0, 8480/0, 8250/0, 8250/2, 8253/2, 8256/3, 8257/3, 8250/3, 8551/3, 8260/3, 8265/3, 8230/3, 8253/3, 8254/3, 8480/3, 8333/3, 8144/3, 8140/3	Chondroma	9220/0
Squamous precursor lesions	8070/2, 8077/0, 8077/2, 8077/2	Diffuse lymphangiomatosis	9170/3
Squamous cell carcinomas	8070/3, 8071/3, 8072/3, 8083/3, 8082/3	Pleuropulmonary blastoma	8973/3
Large cell carcinomas	8012/3	Intimal sarcoma	9137/3
Adenosquamous carcinoma	8560/3	Congenital peribronchial myofibroblastic tumor	8827/1
Sarcomatoid carcinomas	8022/3, 8031/3, 8032/3, 8972/3, 8980/3, 8980/3	Pulmonary myxoid sarcoma with EWSR1-CREB1 fusion	8842/3
Other epithelial tumors	8023/3, 8044/3	PEComatous tumors	9174/3, 8714/0, 8714/3
Salivary gland-type tumors	8940/0, 8200/3, 8562/3, 8430/3, 8310/3, 8982/0, 8982/3		
Lung neuroendocrine neoplasms		*Hematolymphoid tumors*	
Precursor lesion	8040/0	MALT lymphoma	9699/3
Neuroendocrine tumors	8240/3, 8249/3	Diffuse large B-cell lymphoma, NOS	9680/3
Neuroendocrine carcinomas	8041/3, 8045/3, 8013/3	Lymphomatoid granulomatosis, NOS	9766/1, 9766/3
		Intravascular large B-cell lymphoma	9712/3
		Langerhans cell histiocytosis	9751/1
Tumors of ectopic tissues			
Melanoma	8720/3	Erdheim–Chester disease	9749/3
Meningioma	9530/0		

Unfortunately, access to comprehensive medical imaging data remains limited, with publicly available datasets often derived from a small number of centres and lacking sufficient representation of broader patient populations [7,8].

Given the heterogeneity of lung cancer outlined in Table 1, leveraging diverse public datasets is critical for developing and validating generalisable AI solutions. Among these, the IQ-OTH/NCCD dataset [9] has been widely adopted in prior work, often reporting high performance, as summarised in Table 2. Although this dataset is a valuable public resource, particularly given the limited availability of open lung cancer CT datasets relative to the disease's clinical and biological heterogeneity, its released version does not include subject-level identifiers. Therefore, it is difficult to verify whether CT slices from the same subject are kept within the same training or testing partition. This limitation does not preclude the use of the dataset, but it raises an important evaluation concern: if related CT slices are distributed across training and test sets through randomised image-level splitting, model performance may be overestimated, as observed in other healthcare AI domains [10–12].

As subject-level identifiers are unavailable, it is not possible to directly construct or verify a true subject-wise split from the released metadata. However, the original IQ-OTH/NCCD scans were acquired in DICOM format, in which CT imaging data are commonly organised within a Patient-Study-Series-Instance hierarchy, and slices are generated as part of a scan-level volumetric acquisition [9,13]. Although the public release does not document whether this structure is fully preserved after slice extraction, it remains plausible that the released CT-slice sequence may retain some scan- or case-related ordering. Therefore, an important methodological question is whether disrupting this sequence through randomised image-level shuffling affects performance estimates. If the released order contains relevant scan- or case-level structure, preserving it during partitioning may reduce the random intermixing of related CT slices across training and test sets. Conversely, if the released order has no such structure, preserving or disrupting it should yield similar performance estimates.

Motivated by this concern, we designed a predefined set of comparative experiments to examine whether the data partitioning protocol affects performance estimates on the IQ-OTH/NCCD dataset. We evaluated 14 Deep Learning (DL) pipeline configurations combining different preprocessing procedures, CNN architectures, partitioning conditions, and training strategies. Within these configurations, we compared two partitioning conditions. The first condition, referred to as With Order, preserves the order of released CT slices when dividing the dataset. The second condition, referred to as Without Order, applies randomised image-level partitioning, in which images are shuffled before splitting.

This comparison was designed to examine the sensitivity of performance estimates to the data partitioning protocol, particularly whether randomised image-level splitting produces different estimates from partitioning that preserves the released CT-slice order. A consistent difference between the two partitioning conditions would suggest that image-level randomisation may influence evaluation outcomes and potentially inflate performance estimates, especially if related CT slices are distributed across the training and test sets. Based on these experiments, we discuss how the IQ-OTH/NCCD dataset can be used more reliably in future studies while maximising its value as a publicly available resource.

This study focuses exclusively on image-level classification while addressing the specific challenge posed by the IQ-OTH/NCCD dataset. We adopt an image-level classification approach as a scalable foundation for future dataset expansion. The outcomes of this work lay the groundwork for future phases, which will incorporate segmentation-based methods to enhance prediction localisation and interpretability.

Previous studies using IQ-OTH/NCCD have primarily focused on improving classification performance through model architecture, preprocessing, or training strategies, while the implications of the dataset's missing subject-level identifiers for data splitting and performance evaluation have received limited attention. To our knowledge, no previous study has directly examined whether shuffling IQ-OTH/NCCD images before fold construction systematically produces more optimistic performance estimates. This study addresses this gap by comparing the Without Order and With Order conditions across multiple deep-learning pipeline configurations and examining whether the observed differences remain consistent across repeated fold constructions.

The main contributions of this study are as follows:

- We provide systematic empirical evidence that the Without Order condition produces consistently more optimistic performance estimates on IQ-OTH/NCCD. This pattern was observed across 14

Table 2
Summary of prior work using the IQ-OTH/NCCD dataset.

Author	Year	Aim	ML/DL Algorithm	Model Evaluation	Reported Performance	Subject-Level Separation Consideration
Sazzad et al. [14]	2026	Authors explore modality-agnostic channel-attention-driven hybrid DL framework.	VGG16 and EfficientNetB0 with GRU	5KFoldCV	ACC: 99.55%	Acknowledged potential optimistic estimates
Kamala and Mohan [15]	2025	This study explores hybrid feature-fusion framework approaches, combining handcrafted descriptors, including GLCM and SIFT, with deep features extracted from VGG-16 and MobileNet.	GLCM + SIFT + MobileNet configuration	Single 70:20:10 split	ACC: 98.33%	Not explicitly reported
Shariff et al. [1]	2025	Integration of Differential Augmentation (DA) with CNNs to address the critical challenge of memory overfitting	CNN-based architectures	Single 80:20 train–test split	ACC: 98.78%	Not explicitly reported
Güraksın et al. [16]	2025	They propose using transfer learning with pretrained architectures, ensemble learning (EL) methods, and DA to improve cancer detection performance.	A majority-voting ensemble of pretrained GoogLeNet, EfficientNet, and DarkNet-19	Single 80:20 train–test split	ACC: 99.00%	Not explicitly reported
Kumar et al. [17]	2025	The study suggests that transfer learning and transformer-based architectures may improve lung cancer detection performance	pretrained model VGG19, Visual Transformer (ViT)	Single split 80:10:10	ACC: 99.06%	Not explicitly reported
Tabibian et al. [18]	2025	Authors propose two data augmentation strategies using cutting edge generative models to improve effect of class imbalance	DCGAN-based generation, Diffusion-based generation	10 times 80:20 split	ACC: 99.59%	Claimed (not verifiable)
Gupta et al. [19]	2024	This study explores a pipeline that combines multilevel Otsu thresholding for image preprocessing with a modified U-Net architecture optimised via differentiable architecture search (DARTS) for lung cancer detection	Modified U-Net architecture	Single 80:20 train–test split	ACC: 96.82%	Not explicitly reported
Rasa et al. [20]	2023	Novel approach using a transfer learning-based predictor called Lung-EffNet	EfficientNet-based architecture	Suffle images first, then used a train-test ration of 80:20	ACC: 99.10%	Not explicitly reported
Mohamed et al. [21]	2023	The Ebola Optimisation Search Algorithm (EOSA) is used to select the best combination of CNN weights and biases.	EOSA-CNN architecture	Single 80:20 train–test split	ACC: 93.21%	Not explicitly reported
V.R. et al. [22]	2023	The authors suggest that pretrained convolutional networks as feature extractor, combined with tree-based ensemble model as feature selector, may improve lung cancer detection.	pretrained VGG16, Extremely Randomised Tree Classifier, Multi-Layer Perceptron	Single 80:20 train–test split	ACC: 99.09%	Not explicitly reported

deep-learning pipeline configurations and remained evident under 100 repeated five-fold cross-validation runs for two representative pipelines. The magnitude and consistency of the effect were further quantified using mean accuracy differences, permutation testing, Cohen's d, and bootstrap intervals.

- We demonstrate that insufficient reporting of the partitioning procedure can confound the interpretation of performance improvements on IQ-OTH/NCCD. In particular, higher performance may be incorrectly attributed to preprocessing, architecture, or training choices when it may instead be influenced by the distribution of potentially related images across training and evaluation folds. Based on this evidence, we provide recommendations for reporting and interpretation of studies using medical imaging datasets without subject-level identifiers.

The remainder of this paper is organised as follows. Section 2 reviews previous studies that utilised IQ-OTH/NCCD for lung cancer classification and examines how their methodologies address the absence of subject-level identifiers in the released dataset. Section 3 describes the dataset, preprocessing procedures, evaluated pipeline configurations, partitioning conditions, and performance evaluation methods. Section 4 presents the experimental results, including the repeated cross-validation and statistical analyses. Section 5 discusses the methodological implications, limitations, and recommendations for the dataset's future use, and Section 6 concludes the paper.

2. Related work

Table 2 summarises prior studies that used the IQ-OTH/NCCD dataset. Although this dataset has been widely adopted, an important methodological concern remains underexamined: the absence of subject-level identifiers makes it difficult to ensure that images from the same scan or patient are not distributed across both the training and testing sets. This creates a potential risk of evaluation-protocol-induced bias, particularly when randomised image-level splitting is applied. Evidence from other medical imaging domains has shown that non-independent samples shared between training and testing partitions can

lead to overly optimistic performance estimates. The following discussion examines representative studies and their methodological choices, with particular attention to how dataset partitioning was handled.

Sazzad et al. [14] proposed a modality-agnostic channel-attention-driven hybrid DL framework evaluated independently on IQ-OTH/NCCD CT images and LC25000 histopathology images. The model combines VGG-16 and EfficientNetB0 feature-extraction backbones with feature fusion, channel attention, GRU-based feature refinement, and a softmax classification head. On IQ-OTH/NCCD, the model achieved 99.55% accuracy and 99.54% F1-score using 5-fold cross-validation. Although this evaluation is more rigorous than a single hold-out split, the authors explicitly acknowledged that the public datasets used do not provide patient- or slide-level identifiers, restricting strict patient-wise or scan-wise separation during cross-validation and potentially leading to optimistic performance estimates.

Kamala and Mohan [15] proposed a hybrid feature-fusion framework for IQ-OTH/NCCD lung CT classification, combining handcrafted descriptors, including GLCM and SIFT, with deep features extracted from VGG-16 and MobileNet. Six feature combinations were evaluated using a fixed 7:2:1 train–validation–test split, with the GLCM+SIFT+MobileNet configuration achieving the strongest reported performance. Although the study specifies the split ratio and reports that augmentation was applied only to the training set, the methodology does not make clear whether subject-wise or scan-aware safeguards were applied during data partitioning. In addition, using a single split with a 10% hold-out test set may limit the ability to assess the stability of the reported performance on a small dataset with potentially heterogeneous CT findings.

Shariff et al. [1] highlight the challenge of overfitting in AI models for lung cancer detection and propose image augmentation techniques such as random hue adjustments ($\pm 10°$), saturation changes (0.8–1.2), brightness scaling (0.9–1.1), and contrast modifications (0.85–1.15). These augmentations were applied randomly during training to increase data diversity and to mitigate class imbalance in the IQ-OTH/NCCD dataset. In their work, all images were resized to 256×256 pixels, and pixel values were normalised between 0 and 1. Their 2D-CNN architecture achieved an accuracy (ACC) of 98.78% using an 80:20 train–test split, outperforming DenseNet, EfficientNetB0, and ResNet in their comparison. However, the study does not clarify whether augmented variants of the same patient's scans were segregated across splits.

Similarly, Güraksın et al. [16] applied elastic transformations to augment the dataset by a factor of ten before performing the train–test split. When data augmentation is performed prior to partitioning, augmented variants derived from the same original image may be distributed across both training and test sets unless grouping constraints are explicitly applied. This can introduce a risk of data leakage and may lead to overly optimistic performance estimates [23]. The authors proposed a hybrid ensemble architecture, LECNN, which integrates GoogLeNet, EfficientNet, and DarkNet-19 and combines their predictions using a majority-voting mechanism. Their approach achieved 99.00% accuracy using an 80:20 split and was compared against 13 prior studies, indicating strong reported performance. However, the study did not explicitly discuss whether safeguards were implemented to prevent augmented variants of the original image or potentially related CT images from the same subject or scan from being distributed across both the training and test sets.

Tabibian et al. [18] propose two data augmentation mechanisms to address class imbalance, explicitly framing their study around data augmentation rather than classifier architecture optimisation. Although the public IQ-OTH/NCCD dataset does not provide explicit subject identifiers, the authors report using a subject-wise train–test split with an 80:20 ratio (88 patients, 877 images for training; 22 patients, 220 images for testing), repeated over 10 independent splits. However, no further methodological details are provided regarding how subject-level grouping was derived or how variability in the number of CT slices per individual was handled. While the authors state that their experiments are reproducible and provide a reference to a code repository, the corresponding project repository was not accessible at the time of writing. The best reported performance is achieved using diffusion-based augmentation, reaching an accuracy of 0.9959

Gupta et al. [19] explore two datasets, including IQ-OTH/NCCD. They propose an Otsu thresholding approach for image preprocessing and a modified U-Net architecture with differentiable architecture search (DARTS) to improve lung cancer detection. While their reported learning curves suggest minimal overfitting, the paper does not clarify whether subject-level leakage was addressed, since this information is absent in the IQ-OTH/NCCD dataset description. They perform a single 80:20 train–test split and report an ACC of 96.82%.

On the other hand, Rasa et al. [20] explored transfer learning with five pre-trained EfficientNet variations originally trained on ImageNet. The IQ-OTH/NCCD dataset was then used to fine-tune these models. Class imbalance was addressed using data augmentation techniques. The best model performance was 99.10% (ACC) with data augmentation, whereas without augmentation it was 98.64% (ACC). Similarly, Mohamed et al. [21] introduced the Ebola Optimisation Search Algorithm (EOSA), a biologically inspired method for optimising CNN weights and biases to classify normal, benign, and malignant lung conditions. The proposed CNN+EOSA pipeline achieved an accuracy (ACC) of 93.21% with an 80:20 train–test split, indicating state-of-the-art (SOTA) performance compared with existing approaches. Although the authors acknowledge limitations related to sample size and class imbalance, their reported evaluation protocols do not make it clear whether partitioning safeguards were considered to address the absence of subject-level metadata.

Another study has also explored hybrid architectures. In [17], a hybrid model is developed using a pretrained VGG19 for feature extraction and a Vision Transformer (ViT) to capture global contextual features. With an 80:10:10 train-validation-test split, the model achieved a test ACC of 99.06%. As with several prior studies using this dataset, the reported methodology does not make it clear whether subject-wise safeguards were applied during data partitioning. A separate effort by [22] combines VGG16 as a feature extractor, Extremely Randomised Trees as a feature selector, and a multilayer perceptron as the classifier. Using an 80:20 train-test split, this approach reached an ACC of 99.09%. However, the partitioning procedure was similarly reported at the image-split level, without further clarification regarding subject-level grouping.

Overall, recent studies using the IQ-OTH/NCCD dataset have reported high classification performance across a range of DL and hybrid modelling approaches. However, most of these studies focus primarily on model architecture, preprocessing, feature extraction, or optimisation strategies, while offering limited discussion of the data partitioning protocol with respect to subject-level separation. In particular, because the released IQ-OTH/NCCD dataset does not provide subject-level identifiers, it is often not possible to verify from the reported methodology whether CT slices originating from the same subject or scan were kept within the same training, validation, or testing partition.

This lack of discussion or methodological reporting on subject-level separation poses an important challenge for interpreting performance estimates in previous studies. It does not imply that prior studies necessarily suffered from data leakage; rather, it indicates that the possibility of leakage cannot always be assessed from the available information. Therefore, there is a need for methodological awareness when using the IQ-OTH/NCCD dataset, especially in studies that rely on randomised image-level splitting. To our knowledge, no previous study has directly examined whether shuffling IQ-OTH/NCCD images before fold construction (Without Order condition) systematically produces more optimistic performance estimates.

Beyond this concern, the dataset's limited size and the substantial variability in lung cancer manifestation make evaluation based on

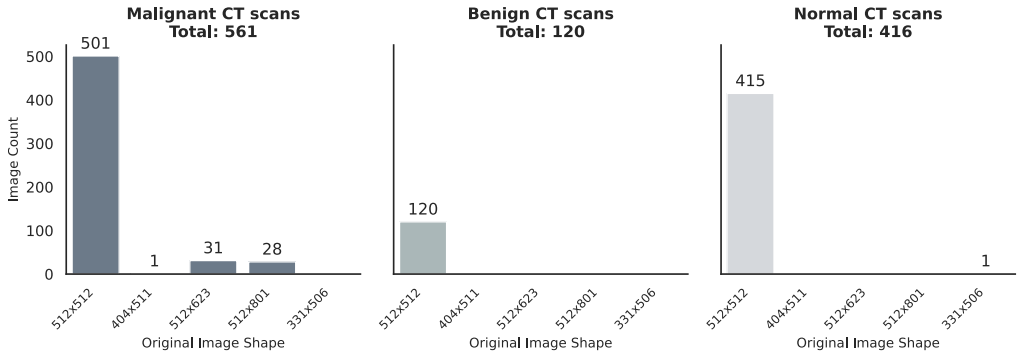

Fig. 1. Distribution of CT-scan images by diagnostic class and original image dimensions. The three subplots correspond to the Malignant, Benign, and Normal classes. Within each subplot, the horizontal axis shows the original image dimensions available in the dataset, and the bar height indicates the number of CT-scan images with each dimension.

a single 80:20 or 80:10:10 split potentially sensitive to the particular samples selected for training and testing. Such protocols may not fully capture the stability of performance estimates across alternative partitions. Moreover, incomplete reporting of implementation pipelines in several studies further limits reproducibility. Together, these methodological challenges motivate a more rigorous reassessment of the IQ-OTH/NCCD dataset, which we undertake in the following section.

3. Material and methods

3.1. IQ-OTH/NCCD dataset

The publicly available IQ-OTH/NCCD lung cancer dataset was collected over a three-month period in the fall of 2019 at the Iraq Oncology Teaching Hospital and the National Centre for Cancer Diseases. It comprises 1097 CT scan slices from 110 de-identified patients, including 40 malignant, 15 benign, and 55 normal subjects. All annotations were performed by oncologists and radiologists at the collection centres. While demographic details such as gender, occupation, and geographic origin (mostly from central Iraq) are noted in descriptive reports, structured metadata such as age or subject IDs are not included in the released dataset [9]. Fig. 1 shows the image count per class and per CT scan image size.

3.2. Experimental partitioning conditions

Each DL pipeline configuration was evaluated under two partitioning conditions. The two conditions differed only in whether the released CT-slice ordering was preserved before assigning images to the training and evaluation folds.

- **Condition 1 (With Order):** CT slices were partitioned while preserving the original sequential ordering provided in the dataset. Under this condition, images were assigned to folds in the order of their release.
- **Condition 2 (Without Order):** CT slices were randomly shuffled prior to partitioning. Under this condition, fold assignment was performed after the released image sequence was disrupted.

3.3. Rationale for with order partitioning condition

The IQ-OTH/NCCD dataset consists of CT scans originally acquired in DICOM format using a Siemens SOMATOM scanner [9], with each scan collected under breath-hold at full inspiration. In the original DICOM acquisition context, CT imaging data are commonly organised through a Patient–Study–Series–Instance hierarchy, where image instances belong to a specific imaging series and study [13]. For CT examinations, slices are generated as part of a scan-level volumetric acquisition and can be ordered using DICOM metadata associated with the image series. This structure supports the possibility that, if the public extraction and file organisation preserved aspects of the original scan-level arrangement, slices from the same patient or scan may appear in contiguous or near-contiguous groups in the released dataset.

However, the public IQ-OTH/NCCD release does not provide subject-level identifiers or the original DICOM metadata required to reconstruct Patient, Study, or Series groupings. Therefore, the released image order cannot be considered a verified patient-level grouping, and it is not possible to determine how accurately it captures true subject-wise separation. In this study, *With Order* partitioning condition is therefore treated as an assumption-based approximation rather than a confirmed subject-wise split. The purpose of this condition is to provide a controlled comparison against *Without Order* partitioning condition, where images are shuffled before splitting and potentially related CT slices may be distributed across training and testing partitions.

To illustrate the released CT-slice sequence used in the With Order partitioning condition, Fig. 2 shows examples of consecutive images from the Normal cases class. Visual changes across the sequence suggest that the released order may contain some local structure, although this does not constitute evidence of verified patient- or scan-level grouping in the absence of subject identifiers.

3.4. Hypothesis testing logic

The hypothesis testing framework was designed to determine whether performance estimates were affected by the partitioning strategy used to assign CT slices to training and evaluation folds. In this study, performance overestimation was operationally defined as a systematic increase in evaluation performance produced by one partitioning strategy relative to the alternative, under otherwise identical pipeline configurations.

The null hypothesis was that the With Order and Without Order conditions would produce comparable performance estimates. Under this hypothesis, disrupting the released CT-slice order before partitioning would not materially affect the estimated accuracy, precision, recall, or F1-score.

The alternative hypothesis was that the two partitioning conditions would produce different performance estimates across the evaluated DL pipeline configurations. In particular, higher performance under the Without Order partitioning condition would indicate that randomised image-level partitioning yields inflated performance estimates relative to the With Order partitioning condition. This would suggest that random shuffling before train–test partitioning can overestimate model performance, potentially because related CT slices may be distributed across both training and evaluation folds.

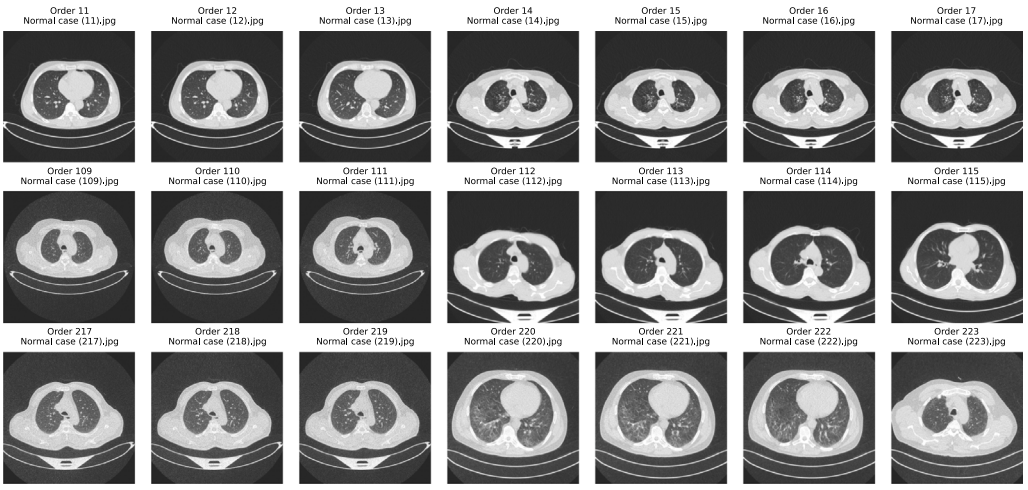

Fig. 2. Illustrative examples of consecutive CT slices from the Normal cases class in the released IQ-OTH/NCCD order. Each row shows a short segment of the original image sequence. The figure illustrates the type of ordering preserved by the With Order partitioning condition. Because subject-level identifiers and original DICOM metadata are unavailable in the public release, these visual segments should not be interpreted as confirmed patient- or scan-level groups.

3.5. Deep learning pipeline components

The proposed DL pipelines perform three-class classification at the individual CT-slice level and consist of four modular components: (1) raw-data preprocessing, (2) model architecture, (3) data split strategy, and (4) training strategy. Different combinations of these components define the DL pipelines evaluated in this study. The following subsections describe each component independently.

3.5.1. Raw data preprocessing component:

This section outlines the initial preprocessing of CT scan images. Four raw-data preprocessing configurations (PreProc001–PreProc004) were evaluated, as summarised in Table 3.

PreProc001 was designed as a minimal image transformation baseline. Because CT scan slices in the IQ-OTH/NCCD dataset exhibit heterogeneous spatial resolutions, all images were resized to 256 × 256 pixels to ensure architectural compatibility and consistent receptive fields across models. No further preprocessing was applied in PreProc001 in order to provide a neutral reference against which the effects of subsequent preprocessing steps could be assessed.

PreProc002–PreProc004 extend PreProc001 by applying Contrast Limited Adaptive Histogram Equalisation (CLAHE) with a fixed tile grid size and clip limits of 2, 5, and 10, respectively. CLAHE is widely used in medical imaging [24] to enhance local contrast and improve the visibility of structural details, such as texture and lesion boundaries, while limiting noise amplification [25]. Evaluating multiple clip limits enables assessment of model sensitivity to different levels of contrast enhancement and ensures that observed performance differences are not driven by a single preprocessing configuration. These pipelines also included pixel normalisation to a [0, 1] range. Fig. 3 presents an example of one CT scan image transformation across each data preprocessing strategy.

3.5.2. Architecture design component:

Three convolutional neural network (CNN) architectures were evaluated: (i) a custom Vanilla 2DCNN with two convolutional layers, (ii) an AlexNet-style architecture adapted from the PyTorch implementation, and (iii) an adopted ResNet-18-style residual network using a [2,2,2,2] residual-block configuration. All architectures were adapted to accept single-channel CT images of size 256 × 256 and to output raw logits for the three target classes. Table 4 summarises the main architectural configurations.

Table 3
Summary of the raw-data preprocessing components evaluated in the DL pipeline designs, including grayscale conversion, image resizing, and CLAHE parameter configurations.

Preprocessing ID	Grayscale method	Resize	ClipLimit	tileGridSize
PreProc001	grayscale_cv2	256px	–	–
PreProc002	grayscale_cv2	256px	2	(8,8)
PreProc003	grayscale_cv2	256px	5	(8,8)
PreProc004	grayscale_cv2	256px	10	(8,8)

3.5.3. Data split strategy component:

To obtain reliable performance estimates on the IQ-OTH/NCCD dataset, we developed a custom stratified K-fold procedure to generate 5-fold cross-validation (5-fold CV) splits. The proposed implementation preserves class stratification while allowing the ordering of CT scan slices to be either preserved or randomly shuffled through a partitioning parameter. Consequently, each split strategy was evaluated under the two partitioning conditions defined in Section 3.2, namely With Order and Without Order, as summarised in Table 5.

As predictive performance estimates from DL models on small datasets can be sensitive to the specific train–test split strategies, and because repeated DL training is computationally expensive, we designed two complementary split strategies to assess robustness to data partitions. Both strategies are described below.

Split strategy (a): A single five-fold partition was generated using a random seed of 0. Two partitioning options were evaluated.

For the With Order option, images were first separated by class and then sorted in ascending order by their original image-order index. To avoid always beginning the partition from the first image in the ordered sequence, a seed-dependent starting position was selected and the class-specific sequence was cyclically rotated. The rotated sequence was then divided into five contiguous subsets. This procedure preserved the relative cyclic ordering and local adjacency of the images within each class while varying the location of the fold boundaries. The corresponding subsets from all classes were combined to form the final stratified folds.

For the Without Order option, images for each class were randomly permuted with the same random seed before being divided into five approximately equal subsets. The corresponding class-specific subsets were then combined to construct the final folds. This option preserved the approximate class distribution across folds but discarded the original image ordering, allowing originally neighbouring images to be assigned to different training and test folds.

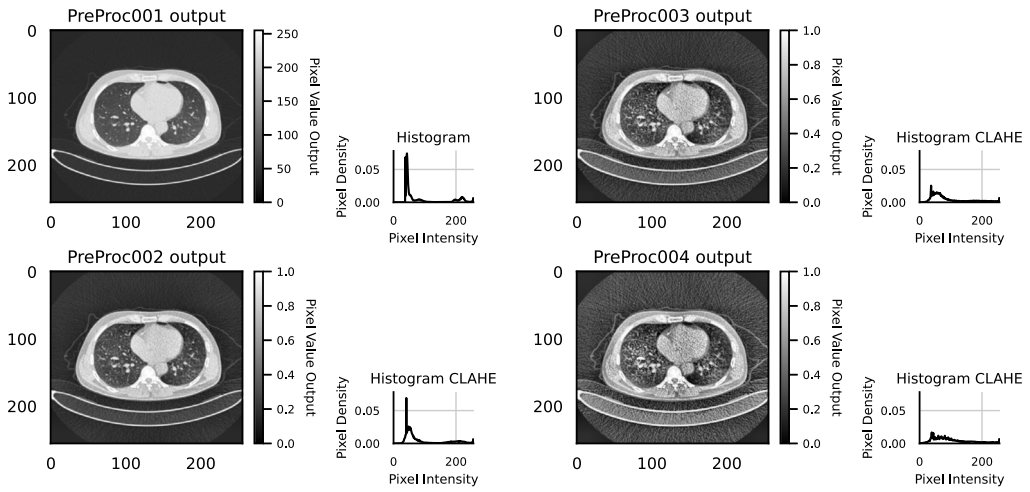

Fig. 3. Visual comparison of the outputs produced by preprocessing strategies PreProc001–PreProc004 for the same CT-scan slice. Each image is shown after grayscale conversion to a single channel, with the accompanying colour bar indicating pixel-intensity values. The corresponding histograms show the distribution of pixel intensities after each preprocessing strategy, highlighting the contrast changes introduced by the CLAHE-based transformations used in PreProc002–PreProc004.

Table 4
Summary of the model architecture components evaluated within the DL pipeline designs, including the input dimensions, main feature-extraction layers, and final classification layers for each evaluated architecture.

Arch. ID	Input size	Main layers	Final Layer
2DCNN	$1 \times 256 \times 256$	Conv2D(16) → MaxPool → Conv2D(32) → MaxPool	FC: 131072 → 3
AlexNet	$1 \times 256 \times 256$	Conv2D(64) → MaxPool → Conv2D(192) → MaxPool → Conv2D(384) → Conv2D(256) → Conv2D(256) → MaxPool → AdaptiveAvgPool2D	FC: 9216 → 4096 → 4096 → 3
RESNET	$1 \times 256 \times 256$	BatchNorm2D → Conv2D(64) → MaxPool → 2×RB(64) → 2×RB(128, stride 2) → 2×RB(256, stride 2) → 2×RB(512, stride 2) → AdaptiveAvgPool2D	FC: 512 → 256 → 125 → 3

RB denotes a residual block.

Thus, the main difference between the two options was how the images were assigned to the five folds. In the With Order option, images that were close together in the original sequence tended to remain together in the same fold. In the Without Order option, the images within each class were randomly shuffled and then divided across the five folds.

Split strategy (b): This split strategy repeated the partitioning procedure described in Split Strategy A 100 times. The random-state value was varied from 0 to 99, producing 100 seeded five-fold cross-validation runs for each partitioning option and, therefore, 500 train–test evaluations per option.

For the With Order option, changing the random-state value changed the cyclic starting position used before dividing each ordered class-specific sequence into five contiguous subsets. The relative ordering of the images was preserved within each repetition, but the locations of the fold boundaries varied across repetitions.

For the Without Order option, changing the random-state value generated a new random permutation of the images within each class before the five subsets were constructed. Therefore, the allocation of individual images to folds varied across repetitions while class stratification was maintained.

In this study, the term *split strategy* refers to the cross-validation protocol, whereas *partitioning condition* refers to how images are organised or assigned to each fold.

3.5.4. Training strategy component:

Hyperparameter settings influence how DL architectures learn patterns. Selecting optimal hyperparameters for DL pipelines is resource-intensive; thus, we adopted a pragmatic approach by conducting a preliminary phase of small-scale experiments. Based on the insights gained, we defined four training strategies tailored to our DL pipeline, preprocessed inputs, and architectures. Note that not all strategies were suitable for every model or input configuration; final strategy assignments are reported in the Results section.

Table 5
Summary of the data split strategy components evaluated in the DL pipeline design, including cross-validation protocols, class stratification, fold counts, and partitioning conditions.

Split Strategy ID	Protocol	Class stratification	Fold count	Partitioning Condition
Split a	5KfoldCV	yes	5	With Order
Split a	5KfoldCV	yes	5	Withour Order
Split b	100 Repeated 5KfoldCV	yes	100 × 5	With Order
Split b	100 Repeated 5KfoldCV	yes	100 × 5	Withour Order

All training strategies used the Adam optimiser [26], a computationally efficient adaptive learning-rate method with low memory requirements that is widely adopted in practice, together with Cross-Entropy Loss with a class weights balance strategy to address imbalance. We used PyTorch, a custom PyTorch Dataset, and DataLoader functions that do not introduce additional data transformations. Table 6 summarises the differences in batch size, number of epochs, and learning rate across the four strategies.

Each unique combination of preprocessing, architecture design, data splitting, and training configuration defines a DL pipeline variant. Each pipeline is trained to perform image-level classification, determining

Table 6
Parameter settings for the training-strategy components evaluated within the DL pipeline design.

Strategy ID	Batch Size	Epochs	Learning Rate
001	50	10	0.001
002	300	50	0.001
003	300	50	0.0001
004	300	75	0.0001

whether an input CT scan is most likely from a normal individual or indicative of benign or malignant lung cancer. During inference, the DL model predicts the class most likely associated with the input image.

3.6. Experimental setup

To ensure reproducibility, we provide the complete source code on GitHub alongside detailed setup instructions in the `README.md` file. An anonymous implementation of the code is available at: https://github.com/fredy-rojas/iq-oth-nccd-classification. A Docker image (built via `docker-compose.yml`) is included for seamless dependency management; alternatively, Python virtual environments can be configured using the provided `requirements.txt`. All experiments were conducted on an NVIDIA RTX A6000 GPU.

3.7. Performance evaluation metrics

Model performance for the three-class CT-slice classification task was evaluated using Accuracy (ACC) and macro-averaged Precision (PRE), Recall (REC), and F1-score (F1). These metrics were selected to quantify overall classification performance under the *With Order* and *Without Order* partitioning conditions. In line with the hypothesis-testing framework, they were used to assess whether performance estimates remained stable or changed systematically when the released CT-slice order was disrupted before fold assignment.

Let K denote the number of classes, N the total number of evaluated samples, and TP_k, FP_k, and FN_k the true positives, false positives, and false negatives for class k, respectively. Accuracy was computed as the proportion of correctly classified samples among all evaluated samples. Macro-averaged Precision, Recall, and F1-score were computed by first calculating the corresponding metric independently for each class and then taking the unweighted mean across classes [27,28]. The metrics were defined as follows:

$$ACC = \frac{\sum_{k=1}^{K} TP_k}{N}.$$

For each class k, precision, recall, and F1-score were computed as:

$$PRE_k = \frac{TP_k}{TP_k + FP_k},$$

$$REC_k = \frac{TP_k}{TP_k + FN_k},$$

$$F1_k = \frac{2 \cdot PRE_k \cdot REC_k}{PRE_k + REC_k}.$$

The macro-averaged metrics were then computed as the unweighted mean across classes:

$$PRE_{macro} = \frac{1}{K}\sum_{k=1}^{K} PRE_k,$$

$$REC_{macro} = \frac{1}{K}\sum_{k=1}^{K} REC_k,$$

$$F1_{macro} = \frac{1}{K}\sum_{k=1}^{K} F1_k.$$

3.8. Statistical comparison of partitioning procedures

A two-sided non-parametric permutation test was conducted separately for each pipeline to assess whether the mean accuracy differed between the With Order and Without Order partitioning options. For each repetition 5KfoldCV, the five fold-level accuracy values were first averaged, producing 100 repetition-level mean accuracies for each partitioning option. Although the same random-state values from 0 to 99 were used for both options, the observations were treated as unpaired because the random-state value controlled different fold-construction operations in each case. Under With Order, it determined the cyclic starting position of the ordered image sequence, whereas under Without Order, it determined the random permutation of the images.

Let $A_{p,i}^{\mathrm{WO}}$ denote the repetition-level mean accuracy for pipeline (p) under Without Order, and let $A_{p,j}^{\mathrm{W}}$ denote the corresponding accuracy under With Order. The observed difference was calculated as

$$\Delta_{p,\mathrm{obs}} = \overline{A}_p^{\mathrm{WO}} - \overline{A}_p^{\mathrm{W}},$$

where a positive value indicates higher accuracy under Without Order. To construct the null distribution, the 200 repetition-level accuracy values were randomly reassigned to the two partitioning conditions while preserving the original group sizes of 100 observations per condition. The difference between the two permuted group means was then calculated as

$$\Delta_{p,b}^{*} = \overline{A}_{p,b}^{*\mathrm{WO}} - \overline{A}_{p,b}^{*\mathrm{W}},$$

where (b) represents one permutation. This procedure was repeated (B=10,000) times for each pipeline. The two-sided permutation (p)-value was calculated as the proportion of permuted absolute mean differences that were equal to or greater than the absolute observed mean difference:

$$p_{\mathrm{perm},p} = \frac{1}{B}\sum_{b=1}^{B} I\left(\left|\Delta_{p,b}^{*}\right| \geq \left|\Delta_{p,\mathrm{obs}}\right|\right)$$

where $I(\cdot)$ is an indicator function that takes the value 1 when the stated condition is satisfied and 0 otherwise. The permutation test therefore assessed whether the observed difference between the two partitioning options was larger than would be expected if the repetition-level accuracy values were randomly assigned between the two conditions.

The magnitude of the difference between the partitioning options was additionally summarised using Cohen's d, calculated as the difference between the mean repetition-level accuracies divided by their pooled standard deviation:

$$d_p = \frac{\overline{A}_p^{\mathrm{WO}} - \overline{A}_p^{\mathrm{W}}}{s_{p,\mathrm{pooled}}}$$

where

$$s_{p,\mathrm{pooled}} = \sqrt{\frac{(n_{\mathrm{WO}}-1)s_{\mathrm{WO}}^2 + (n_{\mathrm{W}}-1)s_{\mathrm{W}}^2}{n_{\mathrm{WO}} + n_{\mathrm{W}} - 2}}$$

Because the repeated cross-validation runs reused the same dataset and produced overlapping training and evaluation sets, the repetition-level accuracy estimates were not statistically independent. Cohen's d was therefore interpreted as a descriptive standardised effect size that summarises the magnitude of the difference relative to the variability observed across the generated partition configurations, rather than as an effect size derived from independent samples.

The mean accuracy difference was retained as the primary measure of effect because it can be directly interpreted as percentage points in accuracy.

A 95% confidence interval for the unpaired mean difference was estimated using 10,000 bootstrap samples. In each bootstrap iteration, 100 repetition-level accuracies were sampled with replacement

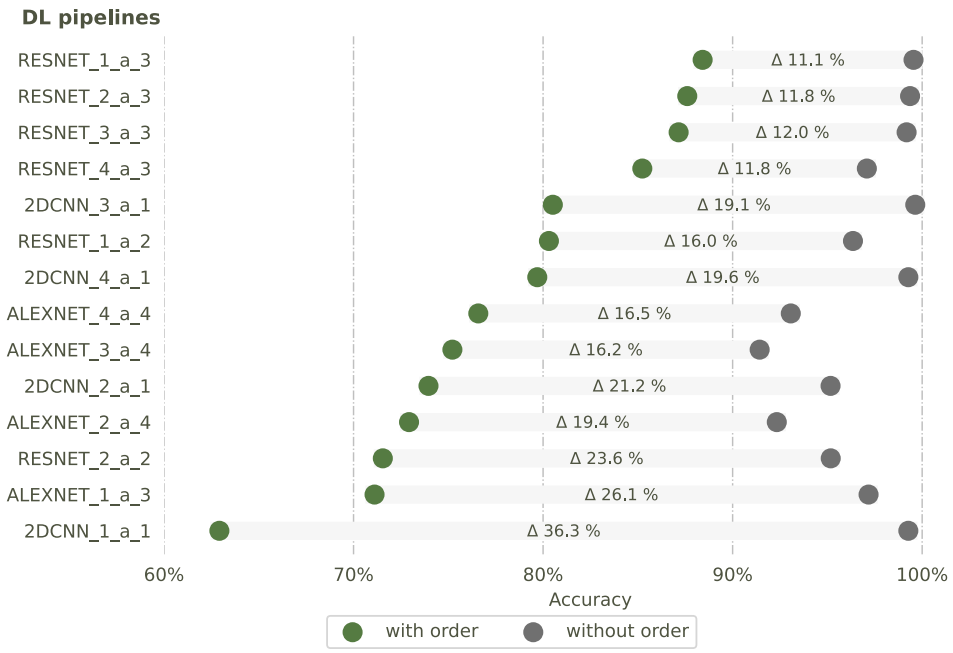

Fig. 4. Horizontal dumbbell plot comparing the mean accuracy obtained under the With Order and Without Order partitioning conditions for each DL pipeline using a single 5KfoldCV (split strategy (a)). The *y*-axis lists the DL pipeline configurations, while the *x*-axis shows accuracy. Green markers represent the With Order condition and gray markers represent the Without Order condition; the connecting lines highlight the paired difference between conditions. The labels report the absolute accuracy difference, defined as $\Delta ACC = ACC_{\text{Without Order}} - ACC_{\text{With Order}}$, in percentage points. Pipeline identifiers follow the convention `(Architecture)_(Preprocessing)_(Split Strategy)_(Training Strategy)`.

from the Without-Order condition and, independently, 100 repetition-level accuracies were sampled with replacement from the With-Order condition. The difference between the two bootstrap means was calculated, and the 2.5th and 97.5th percentiles of the resulting bootstrap distribution were used as the lower and upper confidence limits.

The permutation tests, effect sizes, and bootstrap confidence intervals were used to characterise the direction, magnitude, and consistency of the difference between the two partitioning procedures across repeated fold constructions. Because all repetitions reused the same underlying image dataset, these statistics characterise sensitivity to the partitioning procedure for the evaluated dataset and were not interpreted as estimates based on independent clinical samples or independent external datasets.

4. Results

This section presents the performance results for the evaluated DL pipeline configurations under the two experimental partitioning conditions described in Section 3.2. Results are reported using Accuracy (ACC), macro-averaged Precision (PRE), Recall (REC), and F1-score (F1). The With Order and Without Order conditions are compared to assess the sensitivity of performance estimates to the partitioning strategy. A cross-pipeline comparison is also presented to identify the configurations that produced the most consistent predictive performance across experimental conditions.

4.1. Single stratified 5KfoldCV (Split Strategy a)

Fig. 4 and Table 7 summarise the results of the 14 DL pipeline configurations evaluated using the split strategy (a), consisting of a single stratified 5KfoldCV. Accuracy was higher under the Without Order partitioning condition across all evaluated pipelines. Complete results for accuracy, F1-score, precision, and recall are provided in Appendix. The paired accuracy difference, defined as $\Delta ACC = ACC_{\text{Without Order}} - ACC_{\text{With Order}}$, ranged from 0.1112 to 0.3635. Across pipelines, accuracy ranged from 0.9143 to 0.9964 under Without Order, compared with 0.6292 to 0.8842 under With Order. This consistent pattern suggests that the Without Order condition produced more optimistic performance estimates than the With Order partitioning condition.

Table 7
Numerical summary of the single stratified 5-fold cross-validation (split strategy (a)) results for each DL pipeline under the With Order (WO) and Without Order (W/O) partitioning conditions. The table reports the mean accuracy obtained under each condition and the absolute difference ($\Delta ACC = ACC_{\text{Without Order}} - ACC_{\text{With Order}}$). Pipeline identifiers follow the naming convention `(Architecture)_(Preprocessing)_(Split Strategy)_(Training Strategy)`. Complete results for accuracy, F1-score, precision, and recall are provided in Appendix.

DL pipe id	arch id	preprocessing id	split id	train strategy id	ACC (WO)	ACC (W/O)	Δ ACC Δ ACC
RESNET_1_a_3	RESNET	PreProc001	splt_a	003	0.8842	0.9954	0.1112
RESNET_2_a_3	RESNET	PreProc002	splt_a	003	0.8761	0.9936	0.1175
RESNET_3_a_3	RESNET	PreProc003	splt_a	003	0.8715	0.9918	0.1203
RESNET_4_a_3	RESNET	PreProc004	splt_a	003	0.8524	0.9708	0.1184
2DCNN_3_a_1	2DCNN	PreProc003	splt_a	001	0.8051	0.9964	0.1913
RESNET_1_a_2	RESNET	PreProc001	splt_a	002	0.8031	0.9635	0.1604
2DCNN_4_a_1	2DCNN	PreProc004	splt_a	001	0.7969	0.9927	0.1958
ALEXNET_4_a_4	ALEXNET	PreProc004	splt_a	004	0.7658	0.9307	0.1649
ALEXNET_3_a_4	ALEXNET	PreProc003	splt_a	004	0.7522	0.9143	0.1621
2DCNN_2_a_1	2DCNN	PreProc002	splt_a	001	0.7395	0.9517	0.2122
ALEXNET_2_a_4	ALEXNET	PreProc002	splt_a	004	0.7293	0.9233	0.194
RESNET_2_a_2	RESNET	PreProc002	splt_a	002	0.7154	0.9518	0.2364
ALEXNET_1_a_3	ALEXNET	PreProc001	splt_a	003	0.7111	0.9717	0.2606
2DCNN_1_a_1	2DCNN	PreProc001	splt_a	001	0.6292	0.9927	0.3635

Under the With Order condition, the highest accuracies were generally obtained by the ResNet-based configurations. The best result came from RESNET_1_a_3 at 0.8842. Next was RESNET_2_a_3 with 0.8761, followed closely by RESNET_3_a_3 at 0.8715, and finally RESNET_4_a_3 at 0.8524. These four pipelines all used training strategy 003 and differed only in their preprocessing components

In contrast, under the Without Order condition, performance differences between architectures were considerably smaller. For example, the comparatively simple 2DCNN_3_a_1 pipeline achieved an accuracy

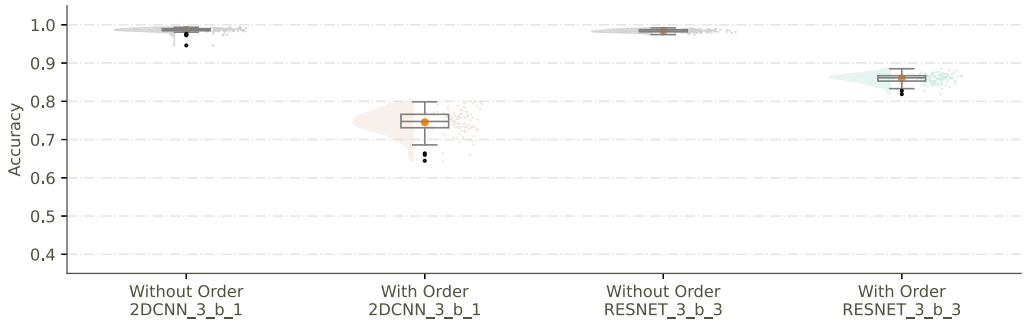

Fig. 5. Distribution of mean accuracy across 100 repetitions of stratified 5KfoldCV for pipelines 2DCNN_3_b_1 and RESNET_3_b_3 under the Without Order and With Order partitioning conditions. Each point represents the mean accuracy across the five folds from one repetition. The violin plots illustrate the distribution of the repeated estimates, while the embedded box plots show the median and interquartile range.

Table 8
Descriptive statistics of the mean accuracy estimates obtained across 100 repetitions of stratified 5KfoldCV for two DL pipelines under the With Order and Without Order conditions. For each repetition, accuracy was averaged across the five folds. The table reports the mean, standard deviation, range, quartiles, and empirical 95% interval across the 100 repetition-level estimates.

DL Pipeline	Partitioning Condition	Mean	Std	Min	25%	50%	75%	Max	Empirical 95% interval
2DCNN_3_b_1	Without Order	0.9865	0.0057	0.9461	0.9845	0.9872	0.99	0.9936	[0.97, 0.99]
2DCNN_3_b_1	With Order	0.7452	0.0307	0.6446	0.7309	0.7473	0.766	0.7986	[0.67, 0.80]
RESNET_3_b_3	Without Order	0.984	0.0034	0.9745	0.9818	0.9836	0.9872	0.9918	[0.98, 0.99]
RESNET_3_b_3	With Order	0.8598	0.0119	0.8187	0.8529	0.8615	0.867	0.8851	[0.84, 0.88]

of 0.9964, which was similar to RESNET_1_a_3 at 0.9954 and higher than all evaluated AlexNet configurations. Similarly, 2DCNN_1_a_1 increased from 0.6292 under With Order to 0.9927 under Without Order, producing the largest observed accuracy difference (ΔACC = 0.3635).

The magnitude of the accuracy difference also varied across architectures. The mean descriptive ΔACC was 0.1440 for the ResNet configurations, 0.1954 for AlexNet, and 0.2407 for the Vanilla 2DCNN configurations. The smallest difference was observed for RESNET_1_a_3 (ΔACC = 0.1112), although its accuracy still increased from 0.8842 under With Order to 0.9954 under Without Order. Overall, the consistent positive differences indicate that the estimated accuracy was sensitive to whether images were shuffled before fold construction.

4.2. Repeated 5KfoldCV (Split Strategy b)

We further evaluated two representative pipelines using 100 repeated stratified 5KfoldCV runs (split strategy (b)) to rule out the possibility that the results from strategy (a) were driven by a favourable train-test split. Fig. 5 presents the distributions of the 100 complete 5KfoldCV accuracy estimates, while Table 8 summarises their descriptive statistics and empirical 95% intervals. The selected pipelines were 2DCNN_3_b_1, representing a comparatively simple DL pipeline, and RESNET_3_b_3, representing a higher-complexity DL pipeline primarily due to its ResNet architecture.

For 2DCNN_3_b_1, the mean accuracy was 0.9865 under Without Order, compared with 0.7452 under With Order, corresponding to a descriptive mean difference of 0.2413. The empirical 95% intervals were [0.97, 0.99] and [0.67, 0.80], respectively, and did not overlap. This separation suggests that the Without Order condition consistently produced higher and more optimistic accuracy estimates across repeated partitions. Accuracy was also more variable under With Order ($SD = 0.0307$) than under Without Order ($SD = 0.0057$).

RESNET_3_b_3 exhibited a similar condition-dependent pattern. The mean accuracy was 0.9840 under Without Order and 0.8598 under With Order, yielding a descriptive mean difference of 0.1242. The corresponding empirical 95% intervals were [0.98, 0.99] and [0.84, 0.88], with standard deviations of 0.0034 and 0.0119, respectively. As with the Vanilla 2DCNN, the intervals did not overlap, and the Without Order estimates were both higher and less variable.

Although RESNET_3_b_3 outperformed 2DCNN_3_b_1 under With Order, their performance was nearly indistinguishable under Without Order. Their mean accuracies were 0.9840 and 0.9865, respectively, and their empirical 95% intervals substantially overlapped ([0.98, 0.99] and [0.97, 0.99]). This convergence towards near-ceiling performance across pipelines of different complexity further indicates that the Without Order condition produced consistently more optimistic performance estimates and reduced the observable performance differences between the evaluated pipelines. These empirical intervals describe the central 95% of the 100 repetition-level accuracy estimates and should be interpreted descriptively, rather than as confidence intervals for population-level predictive performance.

4.3. Permutation tests and effect sizes

To assess whether the differences between the partitioning conditions were larger than expected under the null hypothesis, non-parametric two-sample permutation tests were conducted using the 100 complete repeated 5-fold cross-validation accuracy estimates obtained under each condition. An independent comparison was conducted because the With Order and Without Order procedures yielded different fold memberships, even when the same numerical seeds were used.

For each pipeline, the test statistic was the difference in mean accuracy, defined as $\Delta\text{ACC} = \text{ACC}_{\text{Without Order}} - \text{ACC}_{\text{With Order}}$. The condition labels for the 200 accuracy estimates were permuted, while keeping 100 estimates per condition to generate the corresponding null distribution. Fig. 6 presents these null distributions, with the red vertical line indicating the observed mean difference. For both pipelines, the observed difference was located in the extreme positive tail of the null distribution, yielding permutation-test p-values below 0.001.

As summarised in Table 9, 2DCNN_3_b_1 produced an observed mean accuracy difference of 0.2412 (95% bootstrap CI: [0.2352, 0.2474]). This corresponds to an increase of approximately 24.12 percentage points under the Without Order condition. The descriptive standardised mean difference was also very large (Cohen's $d = 10.92$), indicating substantial separation between the accuracy distributions generated by the two partitioning procedures.

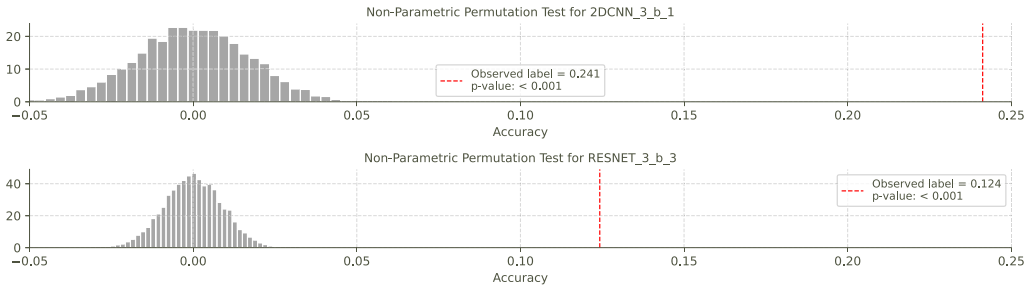

Fig. 6. Permutation-test null distributions for the difference in mean accuracy between the Without Order and With Order conditions. Each observation was the mean accuracy of one repetition of stratified 5-fold CV, and condition labels were randomly permuted to generate the grey null distributions. Red dashed lines indicate the observed differences for `2DCNN_3_b_1` (0.241) and `RESNET_3_b_3` (0.124). Both differences were significant ($p < 0.001$).

Table 9
Summary of the independent comparison between the With Order and Without Order conditions for the selected DL pipelines. The table reports the observed difference in mean accuracy, defined as $ACC_{Without\ Order} - ACC_{With\ Order}$, its 95% bootstrap confidence interval, the independent-groups Cohen's d, and the two-sample permutation-test p-value. Positive differences indicate higher accuracy estimates under the Without Order condition. Cohen's d is reported as a descriptive standardised effect size because the repeated cross-validation estimates were generated by reusing the same dataset.

DL pipe id	Metric	Observed Differences	95% CI	Cohen's d	p-value
2DCNN_3_b_1	ACC	0.2412	[0.2352, 0.2474]	10.92	<0.001
RESNET_1_b_3	ACC	0.1242	[0.1219, 0.1267]	14.26	<0.001

For RESNET_3_b_3, the observed mean accuracy difference was 0.1242 (95% bootstrap CI: [0.1219, 0.1267]), corresponding to an increase of approximately 12.42 percentage points under Without Order. The descriptive standardised mean difference was similarly very large (Cohen's $d = 14.26$). Although the absolute accuracy difference was smaller than that observed for the Vanilla 2DCNN pipeline, the lower variability of the ResNet accuracy estimates resulted in a larger standardised separation between the two partitioning-condition distributions.

For both pipelines, the bootstrap confidence intervals were entirely above zero, supporting the direction and consistency of the higher performance estimates under Without Order. Together with the permutation-test results and the substantial differences in accuracy, these findings support the interpretation that the Without Order procedure systematically produced more optimistic performance estimates for the evaluated dataset. These results also underscore the importance of explicitly reporting image-ordering and shuffling procedures when constructing training and test partitions, particularly when subject-level identifiers are unavailable and potentially related images may be distributed across folds.

5. Discussion

The main goal of this study was to assess whether applying the Without Order partitioning condition to the IQ-OTH/NCCD dataset can produce more optimistic performance estimates. We hypothesised that consistently higher performance under Partitioning Condition 2 (Without Order) than under Partitioning Condition 1 (With Order) would suggest that the estimated performance is sensitive to the assignment of CT slices to training and evaluation folds. Such a result would demonstrate that Without Order partitioning, the condition can produce more optimistic performance estimates when potentially related CT slices are distributed across training and evaluation folds. More broadly, this comparison highlights the importance of carefully selecting, justifying, and explicitly reporting the partitioning strategy, particularly for medical-imaging datasets in which multiple samples may originate from the same subject. Clear reporting of these methodological decisions is essential for allowing readers to appropriately interpret the reported performance and assess the validity, reproducibility, and generalisability of the findings.

The results presented in Fig. 4 show a consistent performance decrease across all 14 DL pipeline configurations when a single stratified 5KfoldCV (split strategy (a)) was applied under With Order condition. To further assess whether these differences were dependent on a specific train–test images arrangement, we repeated the stratified 5KfoldCV procedure 100 times (split strategy (b)). Across these repeated partitions, With Order condition yielded significantly lower performance than Without Order condition ($p < 0.001$), as shown in Fig. 5, Table 8, and Fig. 6. This consistency across multiple valid data partitioning configurations indicates that the performance gap observed under a single 5KfoldCV split is unlikely to be due to chance or a favourable split. Instead, it suggests that the higher performance observed under the Without Order condition reflects a systematic evaluation bias associated with how CT-scan images are assigned to training and evaluation folds rather than an improvement in generalisation.

Despite its simplicity, the vanilla 2DCNN pipeline (2DCNN_1_b_1) achieved a mean accuracy of approximately 0.99 under the Without Order partitioning condition, using only grayscale conversion, resizing to (256 × 256), and a minimal training strategy of 10 epochs. This performance was comparable to that obtained by pipelines incorporating more complex preprocessing methods, architectures, split strategies, and training strategies, as well as to the results reported by the studies presented in Table 7. Similar observations have been reported in medical-imaging studies, where substantially different CNN architectures achieved similarly high performance under image- or slice-level partitioning [10,11]. If the evaluation protocol had been limited to the Without Order condition, particularly without explicitly reporting how CT-scan images were assigned to the cross-validation folds, the near-ceiling performance achieved by pipelines with different levels of complexity could have been attributed to their individual DL components. However, the comparison with the With Order condition shows that performance is also strongly influenced by the partitioning procedure, making it difficult to determine from the Without Order results alone whether the high performance arises from pipeline design choices or from how the folds were constructed.

Although this was not observed in the present experiments, other combinations of preprocessing, architecture, and training strategies may yield similar performance under the With Order and Without Order conditions. However, comparable performance across partitioning conditions would not, by itself, demonstrate that the model had learned clinically meaningful patterns for lung-cancer classification. DL pipeline components are designed to constrain and direct the learning process towards task-relevant information, but the specific patterns captured by the model cannot be directly controlled. Consequently, the interpretation of model performance must also consider how the evaluation folds were constructed and whether potentially related CT-scan images could have been distributed across training and evaluation sets. Explicitly reporting the partitioning procedure is therefore essential for allowing

readers to assess whether the reported performance is more likely to reflect task-related generalisation rather than characteristics arising from the organisation of the dataset.

To reduce dependence on a single favourable or unfavourable partition, this study employed both a single stratified 5-fold cross-validation and 100 repetitions of stratified 5-fold cross-validation. The repeated procedure does not eliminate the dataset's limitations or guarantee performance in external populations; however, it provides a distribution of performance estimates across multiple valid fold configurations. This allows the stability of the reported results to be assessed and reduces the likelihood that the conclusions are driven by one particular train–test assignment. Therefore, when evaluating ML or DL models on small datasets, the choice of evaluation strategy should be explicitly considered and reported, as the construction of the evaluation folds can materially influence the estimated model performance.

An additional methodological consideration in studies using the IQ-OTH /NCCD dataset is the frequent reliance on a single train–test split (such as 80:20, 70:20:10). In small datasets, performance estimates obtained from a single holdout set can be highly dependent on the specific samples assigned to the test partition. Consequently, the selected test set may not adequately represent the variability and difficulty of the full dataset, potentially producing either optimistic or pessimistic performance estimates. This does not necessarily indicate an error in the model itself; rather, it reflects the inherent challenge of designing reliable evaluation protocols when the number and diversity of available samples are limited.

Slice-level and patient-level classification represent different predictive tasks and should not be treated as interchangeable. A slice-level pipeline may be designed to identify or prioritise suspicious CT images within an examination, whereas patient-level assessment requires combining information across multiple slices to produce a case-level decision. The methodological concern examined in this study does not arise from using individual CT slices as the prediction unit, but from how those slices are assigned to training and evaluation folds. Even when slice-level classification is the intended task, performance intended to represent independent cases should be evaluated using partitions that, as far as possible, avoid placing potentially related slices from the same subject or examination in both sets. Therefore, the interpretation of slice-level performance should account for the independence of the evaluation folds and the extent to which the partitioning procedure reflects the pipeline's intended use.

From a clinical-evaluation perspective, the relevance of these findings lies in how the partitioning procedure can alter the perceived usefulness and readiness of a DL pipeline, rather than in demonstrating the clinical utility of any individual configuration. In the repeated evaluation, the Without Order condition produced accuracy estimates approximately 12–24 percentage points higher than the With Order condition. Differences of this magnitude could lead to substantially different interpretations of whether a pipeline warrants further development or evaluation as a support tool for healthcare practitioners. Near-ceiling performance may also make it difficult to distinguish whether an apparent methodological improvement arises from preprocessing, architecture, or training design choices, or from how correlated CT-scan images were assigned to the evaluation folds. Consequently, insufficient reporting of the partitioning procedure may not only lead to inappropriate confidence in reported performance but also affect which pipeline design choices are prioritised for subsequent investigation or testing on additional datasets.

These implications should nevertheless be interpreted within the scope of the IQ-OTH/NCCD dataset and the image-level predictive task evaluated in this study. Repeated cross-validation provides evidence of the stability of performance across alternative partitions of the available data, but it does not establish generalisation to unseen patients, clinical centres, scanner configurations, demographic groups, or disease presentations that are not adequately represented in the dataset. Moreover, because subject identifiers are unavailable, the With Order condition represents a more conservative evaluation strategy rather than a confirmed subject-wise partition.

The development of ML and DL pipelines involves selecting preprocessing methods, model architectures, and training strategies to guide the learning process towards patterns relevant to the predictive task. Although the representations learned by a model cannot be fully controlled, the evaluation protocol can be designed to reduce the possibility that reported performance estimates are influenced by confounding or non-independent patterns. Defining individual CT slices as the prediction unit does not, by itself, address the possibility that potentially related or correlated images may be distributed across training and evaluation folds. In datasets where subject identifiers are unavailable, image-level partitioning may be a practical choice; however, the absence of such identifiers does not eliminate uncertainty about the independence of the resulting folds.

These findings therefore support more transparent reporting when IQ-OTH/NCCD or similar datasets without subject-level identifiers are used. Relevant details include whether images were shuffled before fold construction, the level at which splitting was performed (Ct-scan level, subject level, others), how the folds were generated, and whether subject-level separation could be verified. When these details or the corresponding code are unavailable, it may not be possible to determine whether related CT slices were distributed across training and evaluation folds. Clear reporting of the partitioning procedure and its limitations can therefore help readers determine whether observed improvements are more likely to reflect preprocessing, architecture, or training choices, or may also have been influenced by the distribution of potentially related images.

For relatively small datasets such as IQ-OTH/NCCD, the primary objective may be to investigate the potential of a particular pipeline design rather than to demonstrate clinical readiness. Nevertheless, the evaluation strategy should remain aligned with the intended interpretation and potential use of the designed pipeline. When image-level partitioning may place related CT-scan images, such as images potentially originating from the same patient, in both training and evaluation sets, it is advisable to clearly explain how this evaluation setting relates to the intended clinical use. In the absence of such justification, near-ceiling image-level performance should be interpreted cautiously.

5.1. Limitations

The absence of subject identifiers in IQ-OTH/NCCD prevents the implementation and verification of subject-level partitioning. Accordingly, the With Order condition used in this study should be interpreted as a more conservative partitioning strategy rather than as a confirmed subject-wise split. With Order partitioning, the condition may reduce the distribution of potentially related images across folds; however, exact subject boundaries cannot be identified, and independence between the training and evaluation folds cannot be guaranteed.

The unknown distribution of CT slices across subjects also limits the interpretation of the reported performance. In particular, it is not possible to determine the effective number of independent cases represented in each fold or whether some subjects contribute substantially more slices than others. Repeated 5-fold cross-validation reduces dependence on any single partition and provides a more stable distribution of performance estimates; however, it cannot quantify or eliminate the influence of unequal slice contributions because subject-level information is unavailable. This concern is compounded by the relatively small size of the dataset and its origin from a single geographic setting, which restricts the range of patient, acquisition, and disease characteristics represented. Furthermore, although the dataset includes normal, benign, and malignant categories, it provides no detailed information on lung-cancer subtypes or histological variation. These limitations do not alter the primary methodological objective of examining whether the Without Order condition yields more optimistic performance estimates; however, they restrict any broader interpretation of the findings to

the image patterns represented in IQ-OTH/NCCD rather than to wider clinical populations.

The class imbalance present in the dataset should also be considered when interpreting the results. Balancing techniques were not applied because the primary objective was to isolate the effect of the partitioning condition while keeping other pipeline configurations otherwise comparable. Introducing resampling or augmentation procedures would have added another source of variation to this comparison. Future studies focused on optimising classification performance should investigate class-aware training strategies and report class-specific metrics alongside aggregate performance measures. When data augmentation is used, it should be applied only to the training portion of each fold and should be selected based on the task's anatomical and predictive characteristics. This is important because related or augmented versions of an image should not be distributed across training and evaluation folds, while clinically implausible transformations may introduce patterns that are not representative of the intended predictive task.

Finally, this study evaluates classification performance at the CT-slice level. Slice-level performance does not directly translate to examination- or patient-level performance, as multiple predictions may need to be combined to produce a clinically interpretable decision for a complete CT examination. Although case-level prediction was outside the scope of this methodological investigation, future work would require an appropriate aggregation or decision-level strategy, together with independently identified cases, to assess performance at the level most directly relevant to healthcare practitioner support.

5.2. Future work

Future work should extend the present methodological investigation using newly collected or publicly available datasets that provide verified subject identifiers, scan-level metadata, and clearly defined examination boundaries. Such information would allow true subject-wise partitioning to be compared with the With Order and Without Order conditions evaluated in this study. It would also support investigation of whether preprocessing, architecture, and training design choices that perform well under image-level partitioning maintain their advantages when evaluated on independent cases. Where sufficient metadata are available, future studies could additionally examine scan- or patient-level prediction rather than relying only on individual CT-slice classification.

Alternative strategies for estimating possible scan boundaries were explored, including adjacent-image similarity and candidate change-point detection. However, the resulting pseudo-groups could not be verified as true subject or examination groups, and the number of detected groups exceeded the number of cases reported in the dataset description, suggesting possible over-segmentation. Incorporating these groups into the main evaluation would therefore have introduced additional assumptions beyond the primary objective of assessing whether the Without Order condition produces more optimistic performance estimates. In future work, pseudo-grouping methods could instead be used as sensitivity analyses to examine whether the main conclusions remain stable under different assumptions about image organisation. However, such analyses should not be interpreted as evidence of subject-level separation.

The findings of this study are not intended to discourage the use of IQ-OTH/NCCD. When the dataset is used on its own, With Order partition condition may provide a more conservative alternative to randomly shuffling the images before partitioning. However, this approach should not be described as subject-level partitioning because subject identifiers are unavailable. We therefore strongly recommend using the complete IQ-OTH/NCCD dataset as a single unit, either as part of the training data together with other datasets or as an external test dataset for a model trained on independent data. In both cases, the dataset limitations should be clearly reported when interpreting the results.

Finally, future extensions may investigate class-aware learning, clinically appropriate augmentation, and segmentation methods to identify image regions that contribute to model predictions. Segmentation and prediction-localisation methods may improve interpretation of slice-level outputs, although their clinical relevance would still require evaluation at the examination or patient level using independently identified cases.

6. Conclusion

Given the heterogeneity of lung cancer, it is important to make responsible use of all available public datasets, including IQ-OTH/NCCD. However, the absence of subject identifiers requires particular attention to the assignment of CT slices to training and evaluation folds. This study assessed whether the Without Order condition, in which images are randomly shuffled before partitioning, produces more optimistic performance estimates. Across the evaluated DL pipelines, the Without Order condition consistently outperformed the more conservative With Order condition. Repeated stratified 5KfoldCV showed accuracy differences of approximately 12 and 24 percentage points for the selected ResNet-1-b-3 based and 2DCNN-3-b-1 pipelines, respectively, with significant permutation-test results ($p < 0.001$). Differences of this magnitude can substantially affect how the contribution and potential usefulness of a DL pipeline design, including preprocessing, architecture, partitioning, and training strategies, are interpreted. Therefore, studies using IQ-OTH/NCCD should clearly report their partitioning procedures and acknowledge that related CT slices may be distributed across training and evaluation folds. Where possible, we recommend using IQ-OTH/NCCD as a complete unit within either the training or external evaluation data.

CRediT authorship contribution statement

Fredy Rojas: Writing – original draft, Investigation, Methodology, Data Curation, Formal analysis, Visualization, Software. **Samaneh Madanian:** Writing – review & editing, Validation, Supervision, Methodology, Funding acquisition, Resources, Project Administration, Conceptualization.

Declaration of generative AI and AI-assisted technologies

The authors declare that during the preparation of this, they used Microsoft Copilot (Enterprise Data Protection) and Grammarly for editing and proofreading purposes and fixing some LaTeX technical issues. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Samaneh Madanian reports financial support was provided by Auckland University of Technology. During the preparation of this work, the author(s) used GenAI (Grammarly and MS CoPilot - organisation data protection version) in order to edit and proofread. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is partially funded by AUT Faculty of Design and Creative Technology Summer Scholarship 2023–2024.

Table A.10
Detailed numerical summary of the single stratified five-fold cross-validation results (split strategy (a)) for the 14 deep-learning (DL) pipeline configurations under the With Order (WO) and Without Order (W/O) partitioning conditions. The table reports the mean accuracy (ACC), F1-score (F1), precision (PRE), and recall (REC) obtained across the five folds. For each metric, the difference was calculated as $\Delta = W/O - WO$; therefore, positive values indicate higher performance under the Without Order condition. Pipeline identifiers follow the naming convention `(Architecture)_(Preprocessing)_(Split Strategy)_(Training Strategy)`.

DL pipe id	arch id	preprocessing id	split (id)	train strategy id	ACC (WO)	ACC (W/O)	Δ ACC	F1 (WO)	F1 (W/O)	Δ F1	PRE (WO)	PRE (W/O)	Δ PRE	REC (WO)	REC (W/O)	Δ REC
RESNET_1_a_3	RESNET	PreProc001	a	3	0.8842	0.9954	0.1112	0.7966	0.9921	0.1955	0.8028	0.9965	0.1937	0.7989	0.9881	0.1892
RESNET_2_a_3	RESNET	PreProc002	a	3	0.8761	0.9936	0.1175	0.7566	0.9883	0.2317	0.7866	0.9949	0.2083	0.7579	0.9825	0.2246
RESNET_3_a_3	RESNET	PreProc003	a	3	0.8715	0.9918	0.1203	0.75	0.9845	0.2345	0.756	0.9934	0.2374	0.7521	0.977	0.2249
RESNET_4_a_3	RESNET	PreProc004	a	3	0.8524	0.9708	0.1184	0.7095	0.9437	0.2342	0.7084	0.966	0.2576	0.7153	0.9282	0.2129
2DCNN_3_a_1	2DCNN	PreProc003	a	1	0.8051	0.9964	0.1913	0.7064	0.9927	0.2863	0.7721	0.9947	0.2226	0.7332	0.9909	0.2577
RESNET_1_a_2	RESNET	PreProc001	a	2	0.8031	0.9635	0.1604	0.7	0.9395	0.2395	0.7033	0.9322	0.2289	0.7027	0.9549	0.2522
2DCNN_4_a_1	2DCNN	PreProc004	a	1	0.7969	0.9927	0.1958	0.6955	0.9877	0.2922	0.7065	0.9921	0.2856	0.7133	0.9839	0.2706
ALEXNET_4_a_4	ALEXNET	PreProc004	a	4	0.7658	0.9307	0.1649	0.6566	0.8642	0.2076	0.6548	0.8785	0.2237	0.6716	0.8583	0.1867
ALEXNET_3_a_4	ALEXNET	PreProc003	a	4	0.7522	0.9143	0.1621	0.6387	0.8665	0.2278	0.647	0.8677	0.2207	0.6623	0.8796	0.2173
2DCNN_2_a_1	2DCNN	PreProc002	a	1	0.7395	0.9517	0.2122	0.6558	0.9322	0.2764	0.7271	0.9477	0.2206	0.6666	0.9211	0.2545
ALEXNET_2_a_4	ALEXNET	PreProc002	a	4	0.7293	0.9233	0.194	0.6361	0.8961	0.26	0.6401	0.9225	0.2824	0.6484	0.9081	0.2597
RESNET_2_a_2	RESNET	PreProc002	a	2	0.7154	0.9518	0.2364	0.6109	0.9211	0.3102	0.6165	0.9076	0.2911	0.6227	0.9399	0.3172
ALEXNET_1_a_3	ALEXNET	PreProc001	a	3	0.7111	0.9717	0.2606	0.632	0.9631	0.3311	0.6624	0.964	0.3016	0.6338	0.9647	0.3309
2DCNN_1_a_1	2DCNN	PreProc001	a	1	0.6292	0.9927	0.3635	0.5653	0.9877	0.4224	0.6598	0.9899	0.3301	0.5602	0.9857	0.4255

Appendix. Complete performance metrics for the single 5kfoldcv

See Table A.10.

Data availability

The data that support the findings of this study are openly available in Mendeley Data at https://data.mendeley.com/datasets/bhmdr45bh2/4.

References

[1] Shariff V, Paritala C, Ankala KM. Optimizing non small cell lung cancer detection with convolutional neural networks and differential augmentation. Sci Rep 2025;15(1):15640. http://dx.doi.org/10.1038/s41598-025-98731-4.

[2] Nicholson AG, Tsao MS, Beasley MB, Borczuk AC, Brambilla E, Cooper WA, Dacic S, Jain D, Kerr KM, Lantuejoul S, et al. The 2021 WHO classification of lung tumors: Impact of advances since 2015. J Thorac Oncol 2022;17(3):362–87. http://dx.doi.org/10.1016/j.jtho.2021.11.003.

[3] Inamura K. Lung cancer: Understanding its molecular pathology and the 2015 WHO classification. Front Oncol 2017;7:193. http://dx.doi.org/10.3389/fonc.2017.00193.

[4] Ogawa K, Uruga H, Fujii T, Fujimori S, Kohno T, Kurosaki A, Kishi K, Abe S. Characteristics of non-small-cell lung cancer with interstitial pneumonia: Variation in cancer location, histopathology, and frequency of postoperative acute exacerbations in interstitial pneumonia. BMC Pulm Med 2020;20:1–10. http://dx.doi.org/10.1186/s12890-020-01347-9.

[5] Ueda D, Yamamoto A, Shimazaki A, Walston SL, Matsumoto T, Izumi N, Tsukioka T, Komatsu H, Inoue H, Kabata D, et al. Artificial Intelligence-supported lung cancer detection by multi-institutional readers with multi-vendor chest radiographs: A retrospective clinical validation study. BMC Cancer 2021;21:1–8. http://dx.doi.org/10.1186/s12885-021-08847-9.

[6] Chen BT, Chen Z, Ye N, Mambetsariev I, Fricke J, Daniel E, Wang G, Wong CW, Rockne RC, Colen RR, Nasser MW, Batra SK, Holodny AI, Sampath S, Salgia R. Differentiating peripherally-located small cell lung cancer from non-small cell lung cancer using a CT radiomic approach. Front Oncol 2020;Volume 10 - 2020. http://dx.doi.org/10.3389/fonc.2020.00593, URL https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2020.00593.

[7] Silva F, Pereira T, Neves I, Morgado J, Freitas C, Malafaia M, Sousa J, Fonseca J, Negrão E, Flor de Lima B, et al. Towards machine learning-aided lung cancer clinical routines: Approaches and open challenges. J Pers Med 2022;12(3):480. http://dx.doi.org/10.3390/jpm12030480.

[8] Gandhi Z, Gurram P, Amgai B, Lekkala SP, Lokhandwala A, Manne S, Mohammed A, Koshiya H, Dewaswala N, Desai R, et al. Artificial Intelligence and lung cancer: Impact on improving patient outcomes. Cancers 2023;15(21):5236. http://dx.doi.org/10.3390/cancers15215236.

[9] Alyasriy H, Al-Huseiny M. The IQ-OTH/NCCD lung cancer dataset. 2023, http://dx.doi.org/10.17632/bhmdr45bh2.4, URL https://data.mendeley.com/datasets/bhmdr45bh2/4.

[10] Tampu IE, Eklund A, Haj-Hosseini N. Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. Sci Data 2022;9(1):580. http://dx.doi.org/10.1038/s41597-022-01618-6.

[11] Yagis E, Atnafu SW, García Seco de Herrera A, Marzi C, Scheda R, Giannelli M, Tessa C, Citi L, Diciotti S. Effect of data leakage in brain MRI classification using 2D convolutional neural networks. Sci Rep 2021;11(1):22544. http://dx.doi.org/10.1038/s41598-021-01681-w.

[12] Rouzrokh P, Khosravi B, Faghani S, Moassefi M, Vera Garcia DV, Singh Y, Zhang K, Conte GM, Erickson BJ. Mitigating bias in radiology machine learning: 1. Data handling. Radiol: Artif Intell 2022;4(5):e210290. http://dx.doi.org/10.1148/ryai.210290.

[13] National Electrical Manufacturers Association (NEMA). Digital imaging and communications in medicine (DICOM) standard — Part 4: Service class specifications. 2024, URL https://dicom.nema.org/medical/dicom/2024e/output/chtml/part04/sect_C.3.html. [Accessed 14 January 2026].

[14] Sazzad SB, Islam I, Hossain MB. Interpretable lung cancer detection via channel attention driven hybrid deep learning model. SN Comput Sci 2026;7(5):556. http://dx.doi.org/10.1007/s42979-026-05184-1.

[15] Kamala L, Mohan KG. An efficient hybrid Artificial Intelligence framework for lung cancer classification using CT images. Sci Rep 2026;16:1777. http://dx.doi.org/10.1038/s41598-025-31432-0.

[16] Güraksın GE, Kayadibi I. A hybrid LECNN architecture: A computer-assisted early diagnosis system for lung cancer using CT images. Int J Comput Intell Syst 2025;18(1):35. http://dx.doi.org/10.1007/s44196-025-00761-3.

[17] Kumar A, Ansari MK, Singh KK. A combined approach of vision transformer and transfer learning based model for accurate lung cancer classification. SN Comput Sci 2025;6(5):1–11. http://dx.doi.org/10.1007/s42979-025-04050-w.

[18] Tabibian M, Razmpour T, Saha R. Diffusion models vs. DCGANs for class-imbalanced lung cancer CT classification: A comparative study. Intelligence-Based Med 2026;13:100336. http://dx.doi.org/10.1016/j.ibmed.2025.100336, URL https://www.sciencedirect.com/science/article/pii/S2666521225001413.

[19] Gupta A, Kumar A, Rautela K. UDCT: Lung cancer detection and classification using U-net and DARTS for medical CT images. Multimedia Tools Appl 2024;1–21. http://dx.doi.org/10.1007/s11042-024-19801-9.

[20] Raza R, Zulfiqar F, Khan MO, Arif M, Alvi A, Iftikhar MA, Alam T. Lung-EffNet: Lung cancer classification using EfficientNet from CT-scan images. Eng Appl Artif Intell 2023;126:106902. http://dx.doi.org/10.1016/j.engappai.2023.106902.

[21] Mohamed TI, Oyelade ON, Ezugwu AE. Automatic detection and classification of lung cancer CT scans based on deep learning and ebola optimization search algorithm. PLoS One 2023;18(8):e0285796. http://dx.doi.org/10.1371/journal.pone.0285796.

[22] VR N, Chandra SS V. ExtRanFS: An automated lung cancer malignancy detection system using extremely randomized feature selector. Diagnostics 2023;13(13):2206. http://dx.doi.org/10.3390/diagnostics13132206.

[23] Lones MA. Avoiding common machine learning pitfalls. Patterns 2024;5(10). http://dx.doi.org/10.1016/j.patter.2024.101046.

[24] Shin H, Kim T, Park J, Raj H, Jabbar MS, Abebaw ZD, Lee J, Van CC, Kim H, Shin D. Pulmonary abnormality screening on chest X-rays from different machine specifications: A generalized AI-based image manipulation pipeline. Eur Radiol Exp 2023;7(1):68. http://dx.doi.org/10.1186/s41747-023-00386-1.

[25] Jin S-H, Son D-M, Lee S-H, Go Y-H, Lee S-H. Enhancing local contrast in low-light images: A multiscale model with adaptive redistribution of histogram excess. Mathematics 2025;13(20). http://dx.doi.org/10.3390/math13203282, URL https://www.mdpi.com/2227-7390/13/20/3282.

[26] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proceedings of the international conference on learning representations. ICLR, 2015, p. 1–15, URL https://arxiv.org/abs/1412.6980.

[27] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manage 2009;45(4):427–37. http://dx.doi.org/10.1016/j.ipm.2009.03.002.

[28] Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, Parasa S. On evaluation metrics for medical applications of Artificial Intelligence. Sci Rep 2022;12(1):5979. http://dx.doi.org/10.1038/s41598-022-09954-8.