Ordinal deep learning for pulmonary nodule risk assessment: an external validation study on a novel screening cohort

Miriam Cobo Cano[1 [0000−0001−9208−9062]], Diego Serrano[2,3,4 [0000−0002−3729−9195]], Jennifer Barranco[3], Andrea Pasquier[3 [0000−0002−2581−433X]], Juan Pablo de-Torres[3,4,5 [0000−0003−1231−4894]], Ana Ezponda[3,4,6 [0000−0001−5502−2547]], Luis M Seijo[3,5 [0000−0001−9344−728X]], Gorka Bastarrika[3,4,6 [0000−0001−9493−6907]], Wilson Silva[7 [0000−0002−4080−9328]], Lara Lloret Iglesias[1 [0000−0002−0157−4765]⋆], Luis M Montuenga[2,3,4,8 [0000−0002−8739−1387]⋆]

[1] Advanced Computing and e-Science Group, Institute of Physics of Cantabria (IFCA), CSIC - UC, Santander, Spain, cobocano@ifca.unican.es, lloret@ifca.unican.es

[2] Department of Pathology, Anatomy and Physiology, Schools of Medicine and Sciences, University of Navarra, Navarra, Spain

[3] LUNGSEARCH Laboratory, Program in Solid Tumors, Cima Universidad de Navarra, Cancer Center Clínica Universidad de Navarra (CCUN), Navarra, Spain

[4] Instituto de Investigación Sanitaria de Navarra (IdISNA), Navarra, Spain

[5] Pulmonary Department, Clínica Universidad de Navarra, Navarra, Spain

[6] Radiology Department, Clínica Universidad de Navarra, Navarra, Spain

[7] AI Technology for Life, Department of Information and Computing Sciences, Department of Biology, Utrecht University, Utrecht, Netherlands

[8] Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), ISCIII, Spain

Abstract. This work investigates the generalizability of state-of-the-art deep learning models within a medically informed ordinal framework for lung nodule malignancy risk assessment. A three-class ordinal learning approach is used to evaluate the algorithms, reflecting the indeterminate nature of screening pulmonary nodules and mirroring radiological decision-making. Convolutional neural networks, vision transformers, and capsule networks were implemented in both 2D and 3D. These models were trained on the well-known LIDC-IDRI dataset and externally validated under identical conditions on the P-ELCAP dataset, a novel well-curated lung cancer screening cohort aligned with standard clinical guidelines for lung nodule management. The 2D vision transformer and 3D capsule network achieved the best and second-best external validation performance on P-ELCAP for distinguishing benign from malignant nodules, with the 3D capsule network producing the fewest indeterminate predictions, only one of which corresponded to a truly malignant. However, these results did not extend to nodules subsequently diagnosed as lung cancer at 1–3 years follow-up. The ordinal

⋆ Both authors share Senior Authorship.

framework can enhance indeterminate pulmonary nodule characterization in low-dose CT imaging, contributing to improved risk stratification. The findings reflect the challenges of generalization and the critical need for external assessment of deep learning models prior to clinical deployment. The dataset and the code are available at https://doi.org/10.5281/zenodo.21894313 and https://github.com/MiriamCobo/Deep_ordinal_lung_nodule_evaluation_framework, respectively.

Keywords: Generalizability · Lung cancer diagnosis · Ordinal learning.

1 Introduction

Reliable risk assessment of indeterminate pulmonary nodules (IPNs) detected in low dose computed tomography (LDCT) screening is essential to improve patient outcomes and avoid unnecessary biopsies [10,12,24]. IPNs are neither clearly benign nor highly suspicious for malignancy, making their management a major clinical challenge. The goal is to identify malignant IPNs at early stages while distinguishing them from a majority of benign nodules, which are typically deemed non-suspicious when stable in size on follow-up. DL has shown potential for IPN assessment in LDCT imaging [9,26], yet most methods in the literature lack external validation, limiting clinical translation and real-world adoption [14,3].

Many state-of-the-art (SOTA) DL approaches in lung nodule assessment [9,13] are trained on the publicly available LIDC-IDRI dataset [2], which includes nodule malignancy annotations (ranging from label 1, indicating low risk, to label 5, corresponding to high risk) provided by four independent radiologists. Since its collection in 2011, the criteria used in LIDC-IDRI have been updated by the Lung CT Screening Reporting and Data System (Lung-RADS), developed to standardize the reporting and management of screen-detected pulmonary nodules, and improve risk assessment [4]. In clinical practice, following Lung-RADS guidelines, nodules are broadly categorized into three stages: benign or indeterminate (requiring risk-dependent follow-up), and malignant (necessitating immediate intervention). This risk stratification motivates an ordinal learning formulation, enabling prior clinical knowledge to be incorporated through ordinal regularization [21,23,5] (illustrated in Fig. 1). Differing from prior work [1,9,13,26,28], which ignored the ordinal progression of malignancy risk and relied on binary or multiclass classification, our framework formulates lung cancer risk as an ordinal problem aligned with clinical reasoning. Notably, earlier methods excluded the indeterminate label (3) in LIDC-IDRI [1,9,13,23], overlooking its role as a transitional risk category. While Wu *et al.* [27] proposed a three-class ordinal regression approach using LIDC-IDRI, their 3D DenseNet architecture lacked comparison with other baselines and external validation, limiting conclusions on generalizability and clinical utility.

DL architectures in medical imaging are typically based on convolutional neural networks (CNNs), e.g. DenseNet [11] and EfficientNet [22], or attention

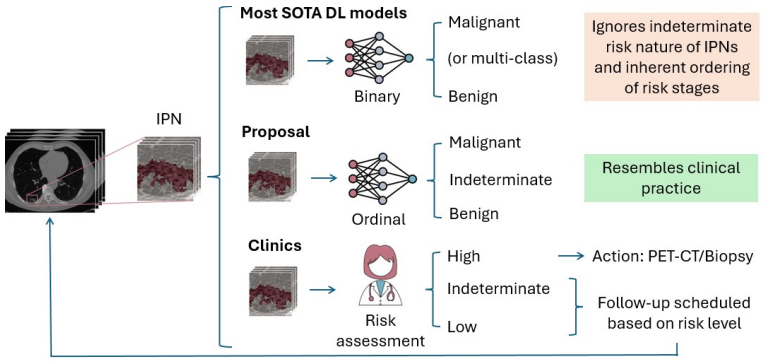

Fig. 1. Illustration of state-of-the-art methods, that disregard the ordinal nature of risk in IPNs, in contrast with our approach, aligned with radiological decision-making.

mechanisms, like Vision Transformers (ViT) [8]. As an alternative, capsule networks (CapsNets) were proposed to overcome CNNs' limitations in modeling spatial hierarchies and part–whole relationships [7]. In LIDC-IDRI, recent binary classification approaches used 2D CapsNets that learned privileged information on nodule attributes [9,13], which were also annotated in this dataset [2]. In 3D, Afshar *et al.* [1] proposed a multi-scale CapsNet with five nodule-centered patches. However, these approaches lacked external validation, which increases the risk of sequential overfitting, that occurs when the best-performing model has simply overfitted to the test data, rather than genuinely exhibiting better generalization [15]. In this context, Shao *et al.* [19] evaluated SOTA models trained on LIDC-IDRI on a new pathologically confirmed dataset (LIDP), concluding poor generalizability due to distribution shifts and the lack of pathological confirmation in LIDC-IDRI. Yet, we identified several limitations in this study: (1) the authors stated that the data originated from different distributions, but they did not thoroughly investigate how to better align radiological labels in LIDC-IDRI with pathological information in LIDP; (2) benign nodules confirmed with biopsy are biased to difficult samples and do not faithfully represent clinical practice; (3) label 3 in LIDC-IDRI was excluded, which represents the majority of IPNs and remains the primary unmet clinical need.

In this work, we validate medically informed DL ordinal approaches for lung nodule risk assessment on a novel well-curated screening cohort (P-ELCAP). Beyond LIDC-IDRI [2], publicly available high-quality datasets that follow standard clinical guidelines for lung nodule management are scarce, and the availability of corresponding annotations is even more limited. Although LIDP [19] was planned to be released after an embargo period, to the best of our knowledge, it is still not publicly accessible. Our contributions can be summarized as (1) exploring how to leverage LIDC-IDRI radiological labels in an ordinal approach that aligns with clinical practice and targets unmet needs in IPN characterization; (2) benchmarking popular architectures to assess generalizability.

2 Materials and methods

2.1 Data

LIDC-IDRI. All models were trained on the LIDC-IDRI [2] dataset. To minimize label uncertainty, only nodules annotated by 3 to 4 radiologists with a standard deviation below 1.0 were considered. In the training set, we repeated the same nodule, leveraging the different annotations available (with slightly different centroids) as a form of data augmentation, while in test and validation we took only the first annotation available. We considered the same labeling criteria as Wu *et al.* [27]: nodules with an average score above 3.5 were labeled as malignant (label 2), those within $[2.5, 3.5]$ were labeled as indeterminate (label 1), and those below 2.5 were labeled as benign (label 0), covering the full range of IPN classifications and considering the ordinal risk. Data were split into training (72%), validation (8%), and test (20%) using stratification based on the ordinal malignancy label, resulting in 3016 training, 91 validation and 223 test samples.

P-ELCAP. External validation was performed on the P-ELCAP dataset [6], retrospectively collected at Clínica Universidad de Navarra, Spain, with the approval of the University of Navarra Research Ethics Board (ref 2020.251, January 25th, 2021). P-ELCAP includes 67 lung cancer cases identified within 5 years prior to diagnosis, and 68 matched controls with benign nodules, confirmed by at least two years of stable follow-up LDCT imaging. There are 5 imaging false positive (FP) cases, who underwent surgery for a suspicious nodule that was conclusively excluded as lung cancer upon pathological examination. Among lung cancer participants, 31 cases were diagnosed at baseline. The remaining pre-lung cancer (Pre-LC) cases were distributed across subsequent years (8, 9, 10, 7, and 2 cases at years 1–5, respectively), with one case per year in years 1–3 in which the nodule was not visible at the corresponding acquisition time point. Validation analyses were restricted to years 1–3 due to the limited clinical relevance of more distant time points. P-ELCAP is biased toward challenging IPNs, reflecting the difficulty of distinguishing malignant from benign nodules in routine screening. LDCT scans were obtained before the administration of any relevant treatments and surgery, and revised to preserve the integrity and completeness of the dataset. Lung nodules were annotated by an experienced radiology technician, and subsequently revised by a thoracic radiologist.

2.2 Method

Each model is trained with a total loss $\mathcal{L}_{\text{total}}$, composed of multiple terms, on the LIDC-IDRI dataset. A prediction loss $\mathcal{L}_{\text{pred}}$ of nodule malignancy risk, representing the ordinal beta cross entropy (β_{CE}) loss from Vargas *et al.* [25], is optimized in all models. CapsNets include an additional attribute loss $\mathcal{L}_{\text{attr}}$. For the 2D CapsNet specifically, a reconstruction loss $\mathcal{L}_{\text{recon}}$ is incorporated, as in the original work by Gallée *et al.* [9]. The total loss is formally defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \lambda_{\text{attr}} \cdot \mathcal{L}_{\text{attr}} + \lambda_{\text{recon}} \cdot \mathcal{L}_{\text{recon}} \tag{1}$$

The nodule malignancy prediction loss $\mathcal{L}_{\text{pred}}$ is defined using the β_{CE} [25]:

$$\begin{aligned}\mathcal{L}_{\text{pred}} &= \beta_{\text{CE}}(y, \hat{y};\ \eta = 0.5,\ J = 3) \\ &= \sum_{i=1}^{J} q'(i) \cdot [-\log p(y = C_i \mid x)]\end{aligned} \tag{2}$$

where $J = 3$ is the number of ordinal classes, and $p(y = C_i \mid x)$ denotes the predicted probability for class C_i. The soft target distribution $q'(i)$ is defined as:

$$q'(i) = (1 - \eta)\delta_{i,y} + \eta f(x, a, b) \tag{3}$$

where $f(x, a, b)$ is the probability value sampled from a beta distribution centered in $x = \frac{2J-1}{2J}$, the parameters a and b are calculated as described by Vargas *et al.* [25], and $\delta_{i,y}$ is the Dirac delta function, which equals 1 if $i = y$ and 0 otherwise, and corresponds to the one-hot encoding of the ground-truth label. The parameter $\eta \in [0, 1]$ controls the linear combination between the one-hot ground-truth label and a unimodal prior derived from the beta distribution used to smooth the original one-hot label into a soft label [25]. We empirically set $\eta = 0.5$. The attribute loss $\mathcal{L}_{\text{attr}}$ is only used to train the 2D and 3D CapsNet models, with $\lambda_{\text{attr}} = 3$. This loss is defined as in Gallée *et al.* [9]:

$$\mathcal{L}_{\text{attr}} = \sum_{k=1}^{K} \text{MSE}\left(a_k^{\text{pred}}[i], a_k^{\text{gt}}[i]\right), \quad \forall i \in \mathcal{I}_{\text{attr}} \tag{4}$$

where $K = 8$ is the number of attributes, $a_k^{\text{pred}}[i]$ and $a_k^{\text{gt}}[i]$ denote the predicted and ground-truth values (mean attribute score by the radiologists), respectively, for the k-th attribute of sample i, and $\mathcal{I}_{\text{attr}}$ is the set of indices corresponding to the samples for which attribute annotations are available. The mean squared error (MSE) is computed independently for each attribute across the annotated samples. The reconstruction loss $\mathcal{L}_{\text{recon}}$ is only used in the 2D CapsNet to predict the segmentation mask of the nodule, with $\lambda_{\text{recon}} = 1$. This loss is defined as [9]:

$$\mathcal{L}_{\text{recon}} = \text{MSE}(x^{\text{recon}}, x) \tag{5}$$

where x is the segmentation mask and x^{recon} is the output produced by the decoder branch of the 2D CapsNet. Models were implemented in PyTorch 2.5.1 [18] on NVIDIA A30 GPUs (24 GB VRAM) using AdamW [16] (batch size 256).

Framework. Baseline models included DenseNet121 [11], EfficientNet-B0 [22], and ViT [8] in both 2D and 3D. We used the CapsNet-2D of Gallée *et al.* [9] without prototypes, since their ablation study showed no significant impact on performance. Their approach aimed to predict the distribution of radiologists' scores, excluding nodules with a mean score of 3, and considering predictions within ±1 of the reference label (within-1 accuracy) as correct [9,13]. However, as most LIDC-IDRI nodules have a score between 2–4, the model predominantly predicted the peak of the distribution at label 3, limiting its discriminative power.

We designed a 3D CapsNet (Fig.2) reducing feature maps (256→32) and routing iterations (3→2) in the original implementation [9] without compromising performance. The reconstruction module was excluded from the 3D version following suboptimal 2D evaluation performance, quantified by the Dice similarity coefficient used in prior work [9] (DSC: 0.66 in LIDC-IDRI, 0.55 in P-ELCAP).

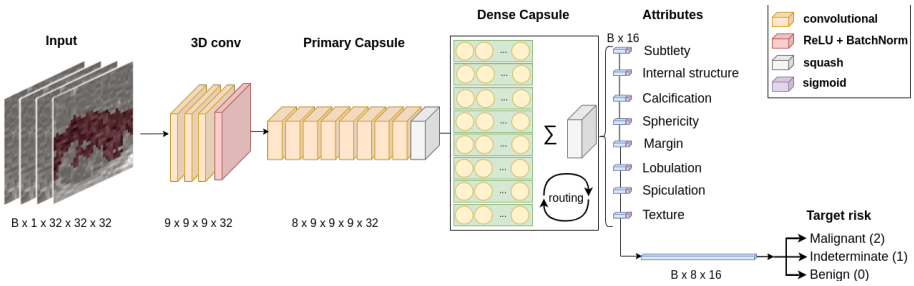

Fig. 2. Proposed 3D capsule network architecture incorporating attribute information to predict three-class ordinal malignancy risk in pulmonary nodules. B denotes the batch.

Preprocessing. A crop of size $[32, 32, 32]$ pixels centered on the nodule's centroid was used to capture 3D context. The minority of nodules exceeding this size in one or more dimensions were down-sampled along the corresponding axes. In 2D, only slices within this crop containing the annotated nodule were retained. In contrast, Afshar *et al.* [1] used only one central and four neighboring slices, and Gallée *et al.* [9] performed tight 2D crops based on segmentation masks that were resized to the input size, which limited the contextual information.

Evaluation metrics. Previous work on LIDC-IDRI reported a single metric, within-1 accuracy [9,13]. However, this metric alone fails to capture overall model performance and is insufficient for a multi-class classification problem with an inherent ordinal nature [17]. Instead, we used three-class accuracy (Acc.), balanced accuracy (B. acc.), mean absolute error (MAE), root mean squared error (RMSE), Uniform Ordinal Classification index (A_{UOC}) [20], and quadratic weighted Cohen's Kappa score (QWK) to evaluate on LIDC-IDRI test subset. For the 2D models, slice-level predictions were aggregated to obtain a nodule-level prediction using either the mean or the maximum predicted score across the retained slices. The aggregation strategy was selected based on the best performance on LIDC-IDRI validation subset. In the external validation on the P-ELCAP subset, malignant nodules diagnosed in the same year as the LDCT acquisition were assigned the maximum risk label (2), while benign nodules were assigned the minimum risk label (0). We report the number of nodules predicted as indeterminate (label 1), along with how many of these correspond to true benign or malignant cases. Binary classification metrics including Acc., B. acc., sensitivity (Sens.), precision (Prec.), specificity (Spec.), negative predictive value (NPV) and F1-score were calculated excluding indeterminate predictions. For all

the models, 95% confidence intervals estimated by bootstrapping were calculated generating 1000 bootstrap samples.

3 Results

Throughout all tables, subscripts denote the maximum half-width of the 95% confidence interval, best and second-best performances are marked **bold** and underlined, and DN, EN, CN refer to DenseNet, EfficientNet, and CapsNet.

Internal evaluation in LIDC-IDRI. Table 1 shows internal performance on the test set. To ensure correct implementation, the 2D CapsNet was reproduced as in Gallée et al. [9], achieving compatible within-1-accuracy (0.945).Table 2 shows binary classification performance excluding ground truth label 1 (indeterminates).

Table 1. Performance on internal LIDC-IDRI test subset with ordinal approach. Total true indeterminates are 53.4%. The aggregation method (Agg.) used in 2D models to compute the global score is indicated (Avg. = average, Max. = maximum).

Model	DN-2D	EN-2D	ViT-2D	CN-2D	DN-3D	EN-3D	ViT-3D	CN-3D
Agg.	Avg.	Avg.	Avg.	Max.	–	–	–	–
Acc.	$\mathbf{0.66}_{0.06}$	$0.58_{0.07}$	$\underline{0.64}_{0.07}$	$0.52_{0.07}$	$0.50_{0.07}$	$0.53_{0.07}$	$0.58_{0.06}$	$0.47_{0.07}$
B. acc.	$\mathbf{0.68}_{0.07}$	$0.62_{0.07}$	$\underline{0.65}_{0.07}$	$\underline{0.65}_{0.05}$	$0.55_{0.07}$	$0.56_{0.06}$	$0.59_{0.07}$	$0.63_{0.05}$
MAE	$\mathbf{0.38}_{0.07}$	$\underline{0.43}_{0.07}$	$\mathbf{0.38}_{0.07}$	$0.53_{0.08}$	$0.55_{0.08}$	$0.50_{0.07}$	$0.44_{0.07}$	$0.54_{0.07}$
RMSE	$\underline{0.68}_{0.09}$	$\underline{0.68}_{0.07}$	$\mathbf{0.64}_{0.07}$	$0.79_{0.08}$	$0.81_{0.08}$	$0.75_{0.07}$	$0.69_{0.06}$	$0.76_{0.06}$
$\boldsymbol{A_{UOC}}$	$0.46_{0.07}$	$0.50_{0.06}$	$0.48_{0.08}$	$0.48_{0.06}$	$\mathbf{0.58}_{0.06}$	$\underline{0.55}_{0.06}$	$0.52_{0.06}$	$0.47_{0.06}$
QWK	$0.52_{0.12}$	$0.52_{0.11}$	$\mathbf{0.55}_{0.11}$	$0.50_{0.10}$	$0.39_{0.11}$	$0.42_{0.10}$	$0.48_{0.10}$	$\underline{0.54}_{0.09}$
% Pred. indet.	51.1	48.4	51.1	19.3	39.5	52.9	53.4	12.6

Table 2. Performance on internal LIDC-IDRI test subset with ordinal approach, restricted to benign and lung cancer ground truth labels.

Model	DN-2D	EN-2D	ViT-2D	CN-2D	DN-3D	EN-3D	ViT-3D	CN-3D
Acc.	$0.88_{0.08}$	$\underline{0.94}_{0.08}$	$\underline{0.94}_{0.06}$	$0.88_{0.07}$	$0.84_{0.09}$	$0.88_{0.09}$	$0.93_{0.07}$	$\mathbf{0.96}_{0.04}$
B. acc.	$0.88_{0.07}$	$\underline{0.94}_{0.06}$	$\underline{0.94}_{0.06}$	$0.90_{0.05}$	$0.84_{0.09}$	$0.84_{0.10}$	$0.92_{0.08}$	$\mathbf{0.96}_{0.05}$
Sens.	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$0.94_{0.09}$	$\mathbf{1.00}_{0.00}$	$0.89_{0.11}$	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$\underline{0.97}_{0.06}$
Prec.	$0.80_{0.13}$	$0.90_{0.12}$	$\mathbf{0.94}_{0.09}$	$0.78_{0.12}$	$0.83_{0.13}$	$0.83_{0.12}$	$0.89_{0.11}$	$\underline{0.93}_{0.10}$
Spec.	$0.76_{0.14}$	$0.87_{0.13}$	$\mathbf{0.95}_{0.08}$	$0.80_{0.11}$	$0.78_{0.16}$	$0.69_{0.19}$	$0.83_{0.16}$	$\underline{0.94}_{0.07}$
NPV	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$0.95_{0.08}$	$\mathbf{1.00}_{0.00}$	$0.86_{0.13}$	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$\underline{0.98}_{0.04}$
F1-score	$0.88_{0.08}$	$\underline{0.94}_{0.07}$	$\underline{0.94}_{0.07}$	$0.87_{0.08}$	$0.86_{0.10}$	$0.91_{0.07}$	$\underline{0.94}_{0.06}$	$\mathbf{0.95}_{0.06}$
# Pred. indet.	31	39	36	10	35	48	49	11
# True benign	26	33	28	9	32	42	41	10
# True malignant	5	6	8	**1**	$\underline{3}$	6	8	**1**

External validation on P-ELCAP. Table 3 shows binary classification generalizability to P-ELCAP, and Table 4 contains predictions for pre-LC cases (years 1–3 before diagnosis) and imaging FP.

Table 3. Performance on external P-ELCAP validation, restricted to benign and lung cancer ground truth labels. [a]NPV undefined for resamples with no predicted negatives.

Model	DN-2D	EN-2D	ViT-2D	CN-2D	DN-3D	EN-3D	ViT-3D	CN-3D
Acc.	$0.54_{0.14}$	$0.57_{0.14}$	$\mathbf{0.84}_{0.14}$	$0.59_{0.12}$	$0.65_{0.12}$	$0.67_{0.12}$	$0.57_{0.13}$	$\underline{0.76}_{0.10}$
B. acc.	$0.59_{0.07}$	$0.63_{0.08}$	$\mathbf{0.85}_{0.11}$	$0.68_{0.07}$	$0.65_{0.09}$	$0.68_{0.10}$	$0.55_{0.10}$	$\underline{0.78}_{0.10}$
Sens.	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$\underline{0.97}_{0.08}$	$0.96_{0.08}$	$0.92_{0.11}$	$0.85_{0.14}$
Prec.	$0.50_{0.15}$	$0.50_{0.16}$	$\mathbf{0.75}_{0.19}$	$0.47_{0.13}$	$0.59_{0.14}$	$0.61_{0.14}$	$0.56_{0.14}$	$\underline{0.64}_{0.15}$
Spec.	$0.17_{0.15}$	$0.25_{0.17}$	$\underline{0.69}_{0.21}$	$0.36_{0.14}$	$0.33_{0.16}$	$0.39_{0.18}$	$0.17_{0.18}$	$\mathbf{0.71}_{0.13}$
NPV	–[a]	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$\underline{0.91}_{0.20}$	$\underline{0.91}_{0.20}$	–[a]	$0.88_{0.11}$
F1-score	$0.66_{0.14}$	$0.66_{0.15}$	$\mathbf{0.85}_{0.13}$	$0.64_{0.13}$	$0.73_{0.12}$	$\underline{0.74}_{0.11}$	$0.70_{0.13}$	$0.73_{0.13}$
# Pred. indet.	46	50	62	23	39	45	51	18
# True benign	39	40	49	20	37	40	45	17
# True malignant	7	10	13	3	$\underline{2}$	5	6	**1**

Table 4. Predictions in pre-LC nodules at years 1–3 prior to diagnosis and imaging false positives (FP). Indeterminate predictions are the complement of benign and malignant.

Model	Group	Pre-LC year 1	Pre-LC year 2	Pre-LC year 3	FP
DN-2D	Benign	0	0	0	0
	Malignant	6	5	7	5
EN-2D	Benign	0	0	1	0
	Malignant	2	4	4	3
ViT-2D	Benign	0	1	1	0
	Malignant	0	4	6	2
CN-2D	Benign	2	1	0	0
	Malignant	3	6	8	5
DN-3D	Benign	1	1	1	0
	Malignant	5	4	7	5
EN-3D	Benign	2	3	1	0
	Malignant	1	3	4	5
ViT-3D	Benign	1	0	0	0
	Malignant	2	3	5	4
CN-3D	Benign	3	2	3	0
	Malignant	4	3	5	4
Total	—	7	8	9	5

4 Discussion

The generalizability of SOTA DL models in lung nodule risk assessment was evaluated following a three class medically informed ordinal approach resembling clinical practice. The findings indicate that high performance on internal test data may not consistently translate to the external dataset, underscoring the need for robust validation before clinical translation. Results from previous work on LIDC-IDRI that lack external validation should be interpreted as upper bounds under idealized conditions rather than reliable estimates for clinical deployment [3]. The 3D CapsNet achieved the highest specificity and fewest indeterminate predictions on the external dataset, with only one corresponding to a true malignant (Table 3), consistent with its binary classification performance on LIDC-IDRI (Table 2). High specificity is critical in screening to minimize unnecessary procedures and patient anxiety. While the 2D ViT outperformed the 3D CapsNet across most binary metrics, it predicted 62 indeterminates, 13 of them truly malignant, compared to the 3D CapsNet's single malignant (Table 3), a key clinical distinction in screening, where missing malignant nodules poses significant patient risk. The remaining models exhibited very low specificity in the external validation, limiting clinical suitability. A further consideration for clinical translation is the dependence on full nodule segmentation. This limitation could be mitigated by integrating a pre-existing automated segmentation model from the literature into the pipeline, thereby eliminating the need for manual delineation. Performance on pre-LC cases (Table 4) was mixed across models, highlighting the absence of a clear superior approach and the need for further validation in larger independent cohorts.

We aligned LIDC-IDRI radiological labels with a clinical dataset, addressing previous limitations [19] and highlighting their utility despite the lack of pathological confirmation. This alignment enables adaptation of public datasets to smaller clinical cohorts, where expert annotations are costly and scarce. Our work also addresses the problem of sequential overfitting [15], demonstrating that solving real-world clinical needs requires robust validation, clinical workflow understanding, and awareness of dataset distribution shifts. The proposed ordinal framework shows potential for improving IPN risk stratification and supporting earlier lung cancer diagnosis. This work ultimately underscores the often over-looked importance of clinical needs and external validation. To support reproducible validation in lung cancer screening, we release P-ELCAP [6].

Acknowledgments. M. C. would like to acknowledge the support received by the Ministry of Education of Spain (FPU grant, reference FPU21-04458). M.C. And L.L.I. would like to acknowledge the support from the project AI4EOSC "Artificial Intelligence for the European Open Science Cloud" that has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement number 101058593. L. M. M. and G. B. received a Lung Ambition Alliance award grant for this project. L. M. M. was also supported by CIBERONC (CB16/12/00443), Spanish Ministry of Science and Innovation and Fondo de Investigación Sanitaria Fondo Europeo de Desarrollo Regional (PI22/00451). D. S. is a recipient of a Ramón y Cajal grant (# RYC2022-038084-I) funded by MICIU/AEI/10.13039/501100011033 and by

"ESF+". This work was supported by the Instituto de Salud Carlos III (ISCIII), project PI25/00123, co-funded by the European Union.

Disclosure of Interests. L. M. M.: Astra-Zeneca: Speaker bureau and Research Grant; AMADIX: Licenced patent co-holder on Complement in LC early detection and Research Grant; SIEMENS Healthineers: Research Grant; Pharmamar: Research Grant; Bristol-Myers Squibb´s: Research grant.

References

1. Afshar, P., Oikonomou, A., Naderkhani, F., Tyrrell, P.N., Plataniotis, K.N., Farahani, K., Mohammadi, A.: 3d-mcn: a 3d multi-scale capsule network for lung nodule malignancy prediction. Scientific reports **10**(1), 7948 (2020)
2. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2), 915–931 (2011)
3. Cabitza, F., Jurman, G., Molinari, F., Bellazzi, R.: Why almost all ml models for medicine are wrong-and what we need for evidence-based medical ai. International Journal of Medical Informatics p. 106538 (2026)
4. Christensen, J., Prosper, A.E., Wu, C.C., Chung, J., Lee, E., Elicker, B., Hunsaker, A.R., Petranovic, M., Sandler, K.L., Stiles, B., et al.: Acr lung-rads v2022: assessment categories and management recommendations. Journal of the American College of Radiology **21**(3), 473–488 (2024)
5. Cobo, M., Fontecha, D.C., Silva, W., Iglesias, L.L.: Physical foundations for trustworthy medical imaging: a survey for artificial intelligence researchers. Artificial Intelligence in Medicine p. 103251 (2025)
6. Cobo Cano, M., Serrano, D., Pasquier, A., Barranco, J., de Torres, J.P., Zulueta, J.J., Echeveste, J.I., Ezponda, A., Pueyo, J., Argueta Morales, A., Sanz-Ortega, J., Berto, J., Alcaide Ocaña, A.B., Di Frisco, M., Felgueroso Rodero, C., Campo, A., de la Fuente Añó, A., Escobar, A., Valencia, K., Montuenga, L.M.: Ordinal deep learning for pulmonary nodule risk assessment: an external validation study on a novel screening cohort. Zenodo Dataset (2026). https://doi.org/10.5281/zenodo.21894313
7. De Sousa Ribeiro, F., Duarte, K., Everett, M., Leontidis, G., Shah, M.: Object-centric learning with capsule networks: A survey. ACM Computing Surveys **56**(11), 1–291 (2024)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Gallée, L., Beer, M., Götz, M.: Interpretable medical image classification using prototype learning and privileged information. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 435–445. Springer (2023)
10. Henschke, C.I., Yip, R., Shaham, D., Markowitz, S., Cervera Deval, J., Zulueta, J.J., Seijo, L.M., Aylesworth, C., Klingler, K., Andaz, S., et al.: A 20-year follow-up of the international early lung cancer action program (i-elcap). Radiology **309**(2), e231988 (2023)

11. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
12. de Koning, H.J., van Der Aalst, C.M., de Jong, P.A., Scholten, E.T., Nackaerts, K., Heuvelmans, M.A., Lammers, J.W.J., Weenink, C., Yousaf-Khan, U., Horeweg, N., et al.: Reduced lung-cancer mortality with volume ct screening in a randomized trial. New England journal of medicine **382**(6), 503–513 (2020)
13. LaLonde, R., Torigian, D., Bagci, U.: Encoding visual attributes in capsules for explainable medical diagnoses. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. pp. 294–304. Springer (2020)
14. Lekadir, K., Frangi, A.F., Porras, A.R., Glocker, B., Cintas, C., Langlotz, C.P., Weicken, E., Asselbergs, F.W., Prior, F., Collins, G.S., et al.: Future-ai: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. bmj **388** (2025)
15. Lones, M.A.: Avoiding common machine learning pitfalls. Patterns **5**(10) (2024)
16. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
17. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., et al.: Metrics reloaded: recommendations for image analysis validation. Nature methods **21**(2), 195–212 (2024)
18. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32. pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
19. Shao, Y., Wang, M., Mai, J., Fu, X., Li, M., Zheng, J., Diao, Z., Yin, A., Chen, Y., Xiao, J., et al.: Lidp: A lung image dataset with pathological information for lung cancer screening. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 770–779. Springer (2022)
20. Silva, W., Pinto, J.R., Cardoso, J.S.: A uniform performance index for ordinal classification with imbalanced classes. In: 2018 international joint conference on neural networks (IJCNN). pp. 1–8. IEEE (2018)
21. Sirocchi, C., Bogliolo, A., Montagna, S.: Medical-informed machine learning: integrating prior knowledge into medical decision systems. BMC Medical Informatics and Decision Making **24**(Suppl 4), 186 (2024)
22. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
23. Tang, W., Yang, Z., Song, Y.: Disease-grading networks with ordinal regularization for medical imaging. Neurocomputing **545**, 126245 (2023)
24. Team, N.L.S.T.R.: Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine **365**(5), 395–409 (2011)
25. Vargas, V.M., Gutiérrez, P.A., Hervás-Martínez, C.: Unimodal regularisation based on beta distribution for deep ordinal regression. Pattern Recognition **122**, 108310 (2022)

26. Wang, C., Shao, J., He, Y., Wu, J., Liu, X., Yang, L., Wei, Y., Zhou, X.S., Zhan, Y., Shi, F., et al.: Data-driven risk stratification and precision management of pulmonary nodules detected on chest computed tomography. Nature Medicine pp. 1–12 (2024)
27. Wu, B., Sun, X., Hu, L., Wang, Y.: Learning with unsure data for medical image diagnosis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10590–10599 (2019)
28. Zhang, R., Wei, Y., Wang, D., Chen, B., Sun, H., Lei, Y., Zhou, Q., Luo, Z., Jiang, L., Qiu, R., et al.: Deep learning for malignancy risk estimation of incidental subcentimeter pulmonary nodules on ct images. European Radiology **34**(7), 4218–4229 (2024)