

Quantifying the Impact of Evaluation Design on Deep Learning–Based CT Histologic Subtyping of Non–Small Cell Lung Cancer: A Leakage-Audited Cross-Cohort Multi-Architecture Study

Necati Vardar

necat.i.vardar@karatay.edu.tr

KTO Karatay University

Research Article

Keywords: Artificial intelligence, computed tomography, data leakage, patient-level validation, external validation, domain shift, calibration

Posted Date: October 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11223014/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Quantifying the Impact of Evaluation Design on Deep Learning–Based CT Histologic Subtyping of Non–Small Cell Lung Cancer: A Leakage-Audited Cross-Cohort Multi-Architecture Study

Necati Vardar^{1, *}

¹Department of Electrical and Electronic Engineering, Faculty of Engineering and Natural Sciences,
KTO Karatay University, Konya 42020, Turkey

*Corresponding author: necati.vardar@karatay.edu.tr

Abstract

Background: Computed tomography (CT)-based deep-learning performance may be strongly influenced by whether correlated images from the same patient cross evaluation partitions. We quantified the effect of evaluation design on histologic subtyping of non–small cell lung cancer (NSCLC) and assessed cross-cohort transportability.

Methods: A development cohort of 311 patients comprising 7,456 region-of-interest (ROI)-bearing CT slices was evaluated using ResNet50, ConvNeXt-Tiny, ViT-B/16, and Swin-T under two otherwise matched five-fold designs: an intentionally leakage-prone slice-random design (E1) and strict patient-grouped evaluation (E3). Final development-only models were applied without fine-tuning, threshold re-optimization, or recalibration to NSCLC-Radiomics (External A; 203 patients) and NSCLC-Radiogenomics (External B; 141 patients). Patient-level area under the receiver operating characteristic curve (AUROC) was the primary endpoint.

Results: Patient-level AUROC ranged from 0.9902–0.9997 under E1 and decreased to 0.7018–0.7342 under E3. Paired E1-minus-E3 AUROC differences ranged from 0.2559–0.2971, with 95% bootstrap confidence intervals excluding zero for all architectures. External AUROC decreased further to 0.5973–0.6298 in External A and 0.5040–0.6084 in External B. All External A AUROC confidence intervals remained above 0.50, whereas all External B intervals included 0.50. Calibration deteriorated after external transfer, and locked E3-derived thresholds produced cohort-dependent sensitivity–specificity tradeoffs. A five-epoch sensitivity analysis preserved the overall external-validation pattern.

Conclusions: Evaluation structure substantially influenced apparent CT-based deep-learning performance. Patient-level independence is necessary for credible internal assessment, but independent external testing remains essential for evaluating model transportability.

Keywords: Artificial intelligence; computed tomography; data leakage; patient-level validation; external validation; domain shift; calibration

1. Background

Non-small cell lung cancer (NSCLC) comprises a heterogeneous group of malignancies in which histologic classification remains central to diagnostic characterization and clinical management. Lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), the two major NSCLC histologies, differ in morphology, molecular profiles, patterns of genomic alteration, and therapeutic relevance. Contemporary thoracic tumor classification therefore integrates morphologic, immunohistochemical, and molecular information rather than treating NSCLC as a single pathologic entity [1]. Although tissue examination remains the reference standard for histologic diagnosis, biopsy samples represent only a limited portion of a spatially heterogeneous tumor. Computed tomography (CT), by contrast, depicts the entire lesion and surrounding tissue environment, motivating the use of quantitative imaging as a non-invasive approach to capture phenotypic information related to tumor biology.

Radiomics first formalized this concept by converting medical images into quantitative descriptors of intensity, morphology, texture, and heterogeneity [2,3]. Subsequent studies extended this framework to learned image representations generated by convolutional and transformer-based neural networks. In NSCLC, multiple investigations have shown that LUAD and LUSC are computationally distinguishable on CT, although reported performance varies substantially across datasets and methods. Radiomics-based studies have reported moderate discrimination, while deep-learning approaches have achieved AUROCs ranging from approximately 0.71 to 0.84

depending on cohort composition, dimensionality, representation strategy, and evaluation design [4–8]. More recent work combining radiomics, deep features, attention mechanisms, or multi-dataset training has reported variable but generally moderate-to-high discrimination across internal and external evaluations [9–11]. Collectively, these findings support the presence of histology-associated information in CT while also showing that apparent model performance is strongly influenced by how data are constructed and evaluated.

This issue is particularly important in two-dimensional CT modeling because a single patient may contribute many anatomically related tumor-bearing slices. Adjacent slices share lesion morphology, acquisition characteristics, reconstruction properties, and patient-specific anatomical context. If such slices are randomly allocated across training and testing folds, the resulting partitions are computationally separate but not biologically independent. Models may then exploit patient-specific similarities, acquisition signatures, or other recurring patterns rather than features that generalize to unseen individuals. Radiology machine-learning guidance therefore emphasizes grouping correlated observations from the same patient within a single partition, and quantitative-imaging standardization initiatives have similarly highlighted the effect of methodological choices on reproducibility [12,13]. This concern is consistent with the broader concept of shortcut learning, in which high-capacity models use statistical regularities that solve the training task without necessarily representing the intended biological target [14]. Medical-imaging studies have shown that neural networks can exploit institutional, scanner, patient, and healthcare-process signals, and methodological reviews have repeatedly identified leakage, biased dataset construction, inappropriate splitting, and insufficient external evaluation as threats to clinical credibility [15–17].

Patient-level independence is therefore necessary for credible internal evaluation but is not sufficient to establish transportability. Internal testing still occurs within the same overarching data-generating environment and may preserve acquisition protocols, reconstruction practices, disease prevalence, annotation conventions, and institutional characteristics. Independent external testing provides a stronger assessment by applying a fixed model to data generated outside the development framework. In a systematic review of externally tested radiologic deep-learning algorithms, 81% showed some reduction in performance outside their development setting [18]. Moreover, discrimination alone does not fully characterize transfer. AUROC measures ranking performance but does not determine whether predicted probabilities remain quantitatively reliable in a new population. Calibration measures such as the Brier score, calibration intercept, calibration slope, and expected calibration error provide complementary information on probability behavior [19,20]. Contemporary reporting frameworks, including CLAIM 2024 and TRIPOD+AI, accordingly emphasize transparent data partitioning, explicit separation of development and evaluation data, independent testing, and reporting beyond discrimination alone [21,22].

Despite these considerations, CT-based NSCLC histology studies have largely focused on improving representations, architectures, or classification performance rather than treating evaluation design itself as the experimental variable. Comparisons across separate studies cannot isolate the effect of partitioning because patient populations, preprocessing, segmentation, model families, optimization procedures, and outcome definitions often differ simultaneously. The present study therefore established a leakage-audited, cross-cohort framework in which the same development cohort, imaging inputs, preprocessing pipeline, task definition, and four model architectures—ResNet50, ConvNeXt-Tiny, ViT-B/16, and Swin-T—were evaluated under deliberately contrasting partition strategies. Performance under an intentionally leakage-prone

slice-random design was compared directly with strict patient-grouped evaluation, and the resulting difference was quantified at the patient level using paired resampling. Final models trained exclusively on the development cohort were then applied without fine-tuning, threshold re-optimization, or recalibration to two independent NSCLC cohorts with predefined confirmatory and secondary/exploratory roles. Discrimination, calibration, locked-threshold operating characteristics, and measurable acquisition-domain differences were evaluated jointly to determine how strongly apparent CT-based histologic classification performance depends on the independence and provenance of the data used for evaluation.

2. Methods

2.1. Study Design and Imaging Cohorts

This retrospective computational imaging study evaluated the effect of evaluation structure on computed tomography (CT)-based discrimination of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). Three publicly available, de-identified thoracic CT collections from The Cancer Imaging Archive (TCIA) were assigned predefined and non-interchangeable roles. Lung-PET-CT-Dx served exclusively as the development cohort, whereas NSCLC-Radiomics and NSCLC-Radiogenomics were reserved for independent external testing as External A (confirmatory) and External B (secondary/exploratory), respectively. No subject, image, label, or outcome information from either external cohort was used for model development, architecture selection, training-duration determination, threshold selection, or any other optimization step (Figure 1A).

After histology harmonization, CT-series selection, region-of-interest (ROI) eligibility assessment, and prespecified integrity checks, the development cohort comprised 311 patients and 7,456 ROI-bearing CT slices, including 250 patients with LUAD and 61 with LUSC [23]. External

A comprised 203 eligible patients and 3,687 slices (51 LUAD and 152 LUSC), whereas External B comprised 141 patients and 3,936 slices (112 LUAD and 29 LUSC) [2,24]. The locked analytical population therefore included 655 patients and 15,079 ROI-bearing CT slices.

Histopathologic diagnoses provided with the source collections were used as the reference standard. Labels were harmonized to a binary task with LUAD as the positive class and LUSC as the negative class. Cases without a harmonizable diagnosis, an eligible CT series, usable ROI-bearing slices, or compliance with the locked integrity criteria were excluded. Duplicate physical slice positions and exact-image duplicates were resolved before partitioning, and cross-cohort integrity checks were performed to prevent duplicated image content from propagating between development and external datasets. All subsequent preprocessing, model development, and statistical analyses were performed using the resulting locked cohort manifests.

Because the study used exclusively publicly available, de-identified imaging data and involved no interaction with participants, no additional institutional ethics approval or informed consent was required for this secondary analysis.

2.2. CT Preprocessing and ROI Standardization

A single preprocessing protocol was defined before model evaluation and applied identically across the development and external cohorts. Digital Imaging and Communications in Medicine (DICOM) pixel values were interpreted in their stored-value domain. Where pixel-padding tags were present, padding values and ranges were identified before Hounsfield unit (HU) conversion and assigned -1024 HU after conversion; untagged extreme stored values were not removed as presumed padding. Remaining pixels were converted to HU using the DICOM RescaleSlope and RescaleIntercept attributes.

For each eligible ROI-bearing CT slice, the locked lesion geometry defined a fixed 192×192 mm physical field of view, which was resampled to 224×224 pixels using bilinear interpolation, corresponding to an in-plane spacing of approximately 0.8571 mm/pixel. Regions extending beyond the image matrix were assigned -1024 HU. Use of a fixed physical rather than fixed pixel-space field of view was intended to reduce variation in apparent tumor scale attributable solely to native acquisition resolution. Each slice was then converted into a three-channel CT representation using prespecified lung (-1350 to 150 HU), mediastinal (-160 to 240 HU), and wide (-1200 to 800 HU) windows. Intensities were clipped to the respective ranges, linearly scaled to $[0,1]$, and subsequently normalized using ImageNet channel-wise statistics (Figure 1B).

No adaptive windowing, histogram equalization, contrast-limited adaptive histogram equalization, synthetic image generation, Synthetic Minority Over-sampling Technique (SMOTE), or image-level oversampling was used. Data augmentation was limited to a horizontal flip with probability 0.5 and was applied only during training. The same physical ROI dimensions, interpolation rule, CT windows, scaling, and normalization were used in all three cohorts.

Preprocessed tensors were stored in cohort-specific sharded caches together with slice and patient identifiers, source series information, Service-Object Pair (SOP) Instance Unique Identifier (UID), shard position, and per-slice tensor hashes. Cache files, model checkpoints, and prediction artifacts were verified using Secure Hash Algorithm 256-bit (SHA-256) hashing before downstream analyses.

2.3. Model Architectures and Training Protocol

Four ImageNet-pretrained architectures representing convolutional and transformer-based model families were evaluated: ResNet50, ConvNeXt-Tiny, ViT-B/16, and Swin-T [25–28]. The final classification layer of each network was configured for binary LUAD–LUSC classification.

All models were initialized using torchvision ImageNet-pretrained weights and subsequently fine-tuned for the study task. Architecture-specific hyperparameter searches were deliberately avoided to ensure that comparisons primarily reflected evaluation design rather than unequal optimization across model families.

A common training protocol was used for all architectures. Optimization employed AdamW [29] with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-4} . The learning rate was updated using cosine annealing with $T_{max} = 30$. Models were trained with a batch size of 16 for a maximum of 30 epochs using class-weighted cross-entropy loss. Class weights were calculated from the training portion of each fold as $w_c = N_{train}/(2n_c)$, where n_c denotes the number of training slices belonging to class c ; validation and test-set class distributions therefore did not contribute to loss weighting.

Training was monitored using fold-specific validation loss. Early stopping was triggered when validation loss failed to improve by at least 1×10^{-4} for five consecutive epochs, and the checkpoint with the minimum validation loss was restored. Mixed-precision computation was used during training and inference, and the random seed was fixed at 20260917. Except for the deliberately altered partitioning unit described below, preprocessing, optimization, augmentation, loss formulation, and patient-level aggregation were held constant across the two internal evaluation designs.

2.4. Leakage-Prone and Patient-Grouped Evaluation Designs

The central experimental variable was the unit of data partitioning. Two five-fold out-of-fold evaluation designs were constructed from the same 311-patient development cohort.

The first design (E1) used slice-random partitioning with StratifiedKFold, preserving histologic class stratification but not patient identity. Individual ROI-bearing CT slices were

therefore allocated independently, allowing slices from the same patient to occur in different folds. E1 was intentionally leakage-prone and served as an experimental reference rather than a recommended clinical evaluation strategy. Each slice received an out-of-fold prediction, and patient-level probabilities were calculated as the arithmetic mean of all slice-level probabilities for that patient. Because splitting occurred at the slice level, slices contributing to a single patient-level prediction could have been evaluated by different fold-specific models (Figure 1C).

The second design (E3) enforced strict patient-level separation using StratifiedGroupKFold, with patient identity as the grouping variable and histologic class used for stratification. All slices from a given patient remained within the same partition, and group-aware separation was maintained during fold-specific model optimization. Training, validation, and test patients were mutually exclusive. Each of the 311 patients entered an outer test fold exactly once, and all 7,456 development slices received one patient-independent out-of-fold prediction (Figure 1C).

Except for the partitioning unit, E1 and E3 used identical source images, labels, preprocessing, architectures, training hyperparameters, loss weighting, augmentation, and patient-level aggregation. Decision thresholds were not optimized for the primary E1–E3 comparison; the prespecified contrast was the change in patient-level AUROC produced by replacing slice-random partitioning with strict patient-grouped evaluation.

After completion of E3, the training duration for each final development model was determined using the median best epoch across the five E3 folds for that architecture. The median was prespecified as a robust fold-level summary and was determined exclusively from internal E3 validation. This procedure yielded a one-epoch training lock for all four architectures. A single

final model per architecture was then fitted using all 311 development patients and 7,456 slices. Fold-specific best epochs and final epoch locks are reported in Supplementary Table S1.

To assess sensitivity to the short primary training duration, a post hoc five-epoch analysis was performed. Five epochs were selected as a fixed alternative within the range of best epochs observed across the E3 folds, without reference to external-cohort performance. Each architecture was retrained on the complete development cohort using the same preprocessing, initialization, optimizer, learning-rate schedule, class-weighted loss, batch size, augmentation, and random seed. These models were applied unchanged to External A and External B without additional model selection, fine-tuning, threshold re-optimization, or recalibration. The one-epoch models remained the prespecified primary analysis. For each external cohort, one- and five-epoch patient-level predictions were aligned by patient identifier, and the change in AUROC was estimated using 10,000 paired patient-level bootstrap replicates with percentile-based 95% confidence intervals. Results are reported in Supplementary Table S2.

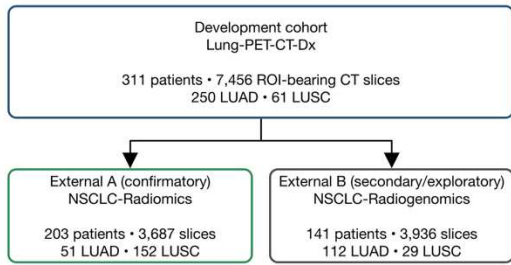
2.5. Independent External Testing and Locked Threshold Transfer

After final development-model fitting, all model parameters were frozen, and the four architectures were applied to External A and External B strictly in inference mode. No fine-tuning, domain adaptation, model selection, threshold re-optimization, or calibration fitting was permitted using external data. External A retained its predefined confirmatory role, whereas External B served as an independent secondary/exploratory test (Figure 1D).

For each external patient, LUAD probabilities were generated for all eligible ROI-bearing CT slices and averaged arithmetically to obtain a single patient-level probability, consistent with the aggregation rule used for internal evaluation. External A therefore contributed 203 patient-level predictions from 3,687 slices, and External B contributed 141 predictions from 3,936 slices.

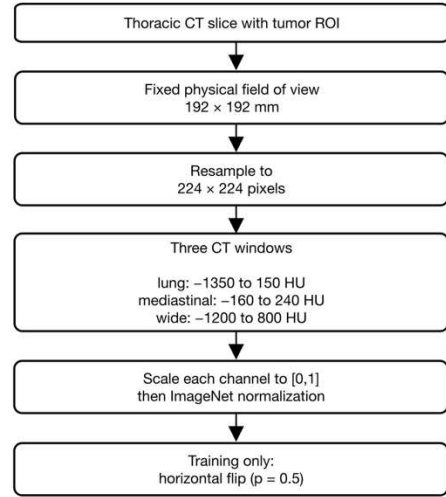
Model-specific classification thresholds were derived exclusively from pooled patient-level out-of-fold predictions generated under strict E3 evaluation. For each architecture, the threshold maximizing Youden's J statistic was selected [30], locked, and transferred unchanged to both external cohorts (Figure 1D). Sensitivity, specificity, balanced accuracy, precision, macro-F1 score, and Matthews correlation coefficient were therefore calculated using internally determined decision rules rather than thresholds optimized on external data. No external recalibration or temperature scaling was performed.

A Cohort allocation

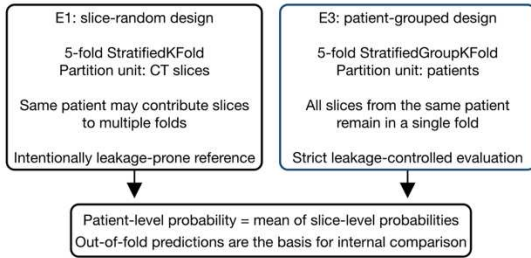


External A and External B were reserved for independent testing only.

B CT preprocessing and ROI standardization



C Internal evaluation designs



D Final model fitting and external testing

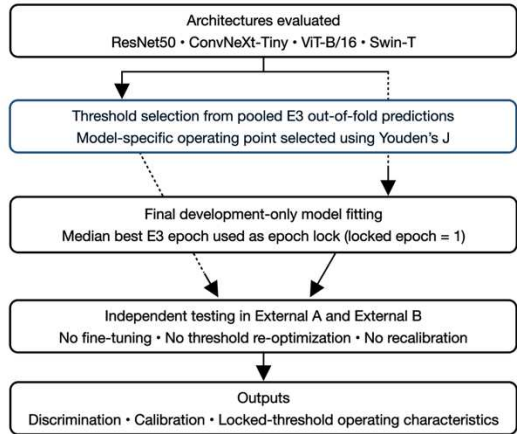


Figure 1. Study design and leakage-audited evaluation framework. Schematic overview of cohort allocation, CT preprocessing, internal evaluation, final model fitting, and independent external testing. E1 used intentionally leakage-prone slice-random partitioning, whereas E3 enforced strict patient-grouped separation. Final models were trained exclusively on the development cohort and applied unchanged to External A (NSCLC-Radiomics; confirmatory) and External B (NSCLC-Radiogenomics; secondary/exploratory). Model-specific thresholds were derived from pooled E3 out-of-fold patient predictions and transferred unchanged to both external cohorts without fine-tuning, threshold re-optimization, or recalibration.

2.6. Performance, Calibration, and Cross-Cohort Statistical Analysis

Patient-level AUROC was prespecified as the primary performance endpoint. Secondary metrics included precision–recall area under the curve (PR-AUC), balanced accuracy, macro-F1 score, Matthews correlation coefficient (MCC), sensitivity, specificity, and precision. Because PR-AUC is prevalence dependent, it was interpreted within cohorts rather than used as a prevalence-invariant basis for direct cross-cohort comparison.

The E1–E3 evaluation-design effect was quantified separately for each architecture using 10,000 paired patient-level bootstrap replicates. E1 and E3 predictions from the same patient were resampled together, and the 2.5th and 97.5th percentiles of the resulting Δ AUROC distribution defined the 95% confidence interval (CI). This paired procedure preserved within-patient correspondence and based uncertainty on patient-level sampling variation. External performance uncertainty was estimated using 10,000 patient-level bootstrap samples per model and cohort; samples containing only one histologic class were discarded because AUROC is undefined. Percentile-based 95% CIs were calculated for AUROC and principal secondary metrics.

Calibration was assessed using the Brier score, expected calibration error (ECE), calibration intercept, and calibration slope [19,20]. ECE was calculated using ten equal-width probability bins. Calibration intercept and slope were estimated with a binomial generalized linear

model relating the observed binary histology label to the logit of the predicted patient-level probability. Calibration was evaluated for E3 and both external cohorts without external recalibration.

Cross-cohort differences in available acquisition characteristics were analyzed at the patient level. Continuous variables were screened using the Kruskal–Wallis test, followed by pairwise Mann–Whitney comparisons between the development cohort and each external cohort. Multiplicity was controlled using the Holm procedure [31], and Cliff’s delta was used to quantify ordinal effect size [32]. Categorical acquisition variables were evaluated using contingency-table analysis and Cramér’s V; rare levels were combined where necessary, and significance testing was interpreted cautiously when expected cell counts remained small. Differences in LUAD:LUSC composition were assessed using two-sided Fisher exact tests with Holm correction.

All acquisition-domain analyses were considered descriptive or associative and were not interpreted as causal explanations of external performance degradation. Cohort manifests, fold assignments, preprocessing caches, model checkpoints, prediction files, and principal statistical outputs were locked and integrity-verified before downstream synthesis.

2.7. Use of AI-assisted tools in manuscript preparation

3. Results

3.1. Cohort Composition and Evaluation Integrity

The final analytical population comprised 655 patients and 15,079 ROI-bearing CT slices across one development and two independent external cohorts. The development cohort included 311 patients (250 LUAD, 61 LUSC), External A included 203 patients (51 LUAD, 152 LUSC),

and External B included 141 patients (112 LUAD, 29 LUSC). LUAD prevalence was therefore 80.4% in the development cohort, 25.1% in External A, and 79.4% in External B (Table 1).

E1 and E3 used the same 311 development patients and 7,456 CT slices but differed in partitioning unit. E1 allowed slices from the same patient to occur across different folds, whereas E3 enforced strict patient-level separation. Under E3, each patient entered an outer test fold exactly once, all 7,456 slices received one patient-independent out-of-fold prediction, and no patient overlap occurred between training, validation, and test subsets. External cohorts were introduced only after development-model fitting and threshold locking.

Table 1. Cohort composition and predefined analytical roles

Cohort	Analytical role	Patients, n	CT slices, n	LUAD, n	LUSC, n	LUAD prevalence	Mean slices/patient
Development	Model development and internal evaluation	311	7456	250	61	0.804	23.97
External A	Confirmatory external testing	203	3687	51	152	0.251	18.16
External B	Secondary/exploratory external testing	141	3936	112	29	0.794	27.91

Note. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

3.2. Effect of Evaluation Design on Internal Performance

Changing only the partitioning strategy produced a marked reduction in discrimination. Under slice-random E1 evaluation, patient-level AUROC ranged from 0.9902 to 0.9997 across the four architectures, whereas strict patient-grouped E3 evaluation reduced AUROC to 0.7018–0.7342 (Figure 2A).

The corresponding E1-minus-E3 Δ AUROC values ranged from 0.2559 to 0.2971. In paired patient-level bootstrap analysis, the 95% CIs excluded zero for all architectures, confirming a consistent evaluation-design effect across both convolutional and transformer-based models (Table 2; Figure 2B).

Table 2. Patient-level discrimination across evaluation designs and independent external cohorts

Model	E1	E3	E1-E3	External A	External B
	AUROC	AUROC	Δ AUROC (95% CI)	AUROC (95% CI)	AUROC (95% CI)
ResNet50	0.9997	0.7191	0.2806 (0.2176–0.3490)	0.6246 (0.5339–0.7142)	0.5199 (0.3959–0.6420)
ConvNeXt-Tiny	0.9997	0.7216	0.2781 (0.2094–0.3518)	0.5973 (0.5096–0.6836)	0.5579 (0.4414–0.6733)
ViT-B/16	0.9902	0.7342	0.2559 (0.1890–0.3282)	0.6298 (0.5436–0.7123)	0.5040 (0.3856–0.6217)
Swin-T	0.9990	0.7018	0.2971 (0.2244–0.3753)	0.6267 (0.5359–0.7155)	0.6084 (0.4925–0.7201)

Note. Δ AUROC represents the difference in AUROC between E1 and E3 (E1 minus E3). Internal differences were estimated using 10000 paired patient-level bootstrap replicates. External confidence intervals were obtained from 10000 patient-level bootstrap samples per model and cohort. AUROC, area under the receiver operating characteristic curve; CI, confidence interval.

3.3. Independent External Testing

Discrimination decreased further during independent external testing. In External A, patient-level AUROC ranged from 0.5973 to 0.6298, corresponding to decreases of approximately 0.075–0.124 relative to E3. Nevertheless, the lower bounds of the 95% bootstrap CIs remained above 0.50 for all four architectures. External B showed weaker and more heterogeneous discrimination, with AUROCs of 0.5040–0.6084 and reductions of approximately 0.093–0.230 relative to E3; all 95% CIs included 0.50 (Figure 2C; Table 2).

Locked-threshold performance was similarly cohort dependent. Balanced accuracy ranged from 0.552 to 0.593 in External A and approximately 0.512 to 0.569 in External B, with corresponding MCC ranges of 0.107–0.164 and approximately 0.020–0.126, respectively (Table 3). These values were obtained using E3-derived thresholds without external re-optimization.

PR-AUC ranged from 0.3272 to 0.3818 in External A and from 0.7875 to 0.8596 in External B. Because LUAD prevalence differed markedly between cohorts (25.1% vs. 79.4%), PR-AUC was interpreted only as a within-cohort secondary metric rather than as a prevalence-invariant basis for direct comparison.

The five-epoch sensitivity analysis preserved the overall external-validation pattern. In External A, five-epoch AUROCs ranged from 0.5955 to 0.6353, with paired Δ AUROC of -0.0343 to $+0.0107$. In External B, five-epoch AUROCs ranged from 0.5068 to 0.6071, with paired Δ AUROC of -0.0480 to $+0.0676$. Paired Δ AUROC confidence intervals included zero for all models except ResNet50 in External B, which increased by 0.0676 (95% CI, 0.0022–0.1381). The qualitative pattern was unchanged: all External A AUROC confidence intervals remained above 0.50, whereas all External B intervals included 0.50 (Supplementary Table S2).

Table 3. Locked-threshold external operating-point performance

Cohort	Model	AUROC	PR-AUC	Balanced accuracy	MCC
External A	ResNet50	0.6246	0.3818	0.558	0.107
External A	ConvNeXt-Tiny	0.5973	0.3272	0.571	0.123
External A	ViT-B/16	0.6298	0.3693	0.552	0.118
External A	Swin-T	0.6267	0.3465	0.593	0.164
External B	ResNet50	0.5199	0.7875	0.550	0.084
External B	ConvNeXt-Tiny	0.5579	0.8210	0.512	0.020
External B	ViT-B/16	0.5040	0.8024	0.525	0.041

Cohort	Model	AUROC	PR-AUC	Balanced accuracy	MCC
External B	Swin-T	0.6084	0.8596	0.569	0.126

Note. External A was predefined as confirmatory and External B as secondary/exploratory. AUROC and PR-AUC were calculated from continuous patient-level probabilities. Balanced accuracy and MCC were calculated using model-specific thresholds derived exclusively from pooled E3 out-of-fold patient predictions and transferred unchanged to both external cohorts. PR-AUC is prevalence sensitive and is therefore reported as a secondary within-cohort metric. AUROC, area under the receiver operating characteristic curve; PR-AUC, precision–recall area under the curve; MCC, Matthews correlation coefficient.

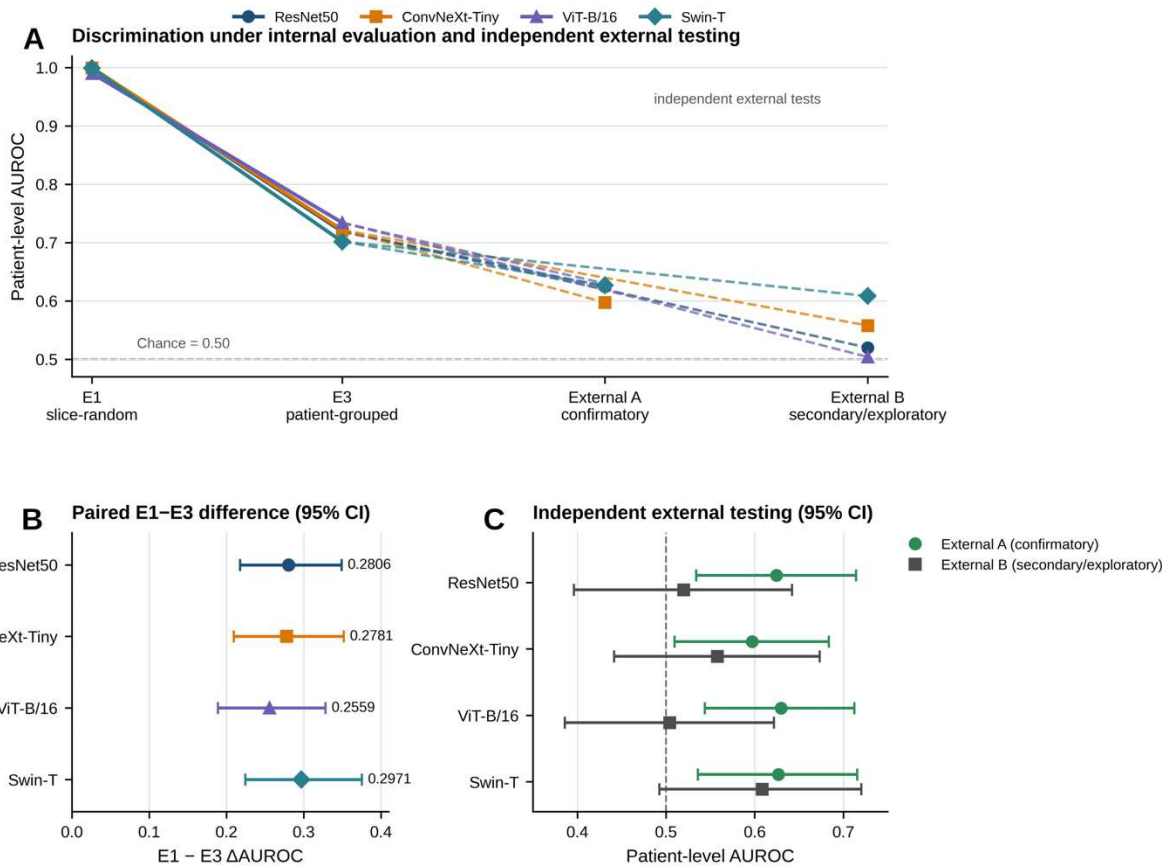


Figure 2. Effect of evaluation design and progressive loss of discrimination during independent testing. (A) Patient-level AUROC trajectories across leakage-prone E1, patient-grouped E3, External A, and External B; the horizontal line denotes AUROC = 0.50. (B) Paired E1-minus-E3 Δ AUROC values with 95% CIs from 10,000 patient-level bootstrap replicates. (C) Patient-level AUROC with 95% bootstrap CIs in the two independent external cohorts. No external model adaptation was performed.

3.4. Calibration and Transfer of Locked Operating Points

Calibration deteriorated after transfer from strict internal evaluation to the external cohorts. Under E3, ECE ranged from 0.070 to 0.148, compared with 0.234–0.413 in External A and approximately 0.175–0.226 in External B (Figure 3A). Brier scores likewise increased from 0.1457–0.1845 under E3 to 0.2559–0.4118 in External A and 0.2142–0.2487 in External B, indicating poorer probability calibration after transfer (Figure 3B).

Model-specific thresholds were derived exclusively from pooled E3 out-of-fold patient probabilities using Youden’s J statistic and fixed before external testing. The locked thresholds ranged from 0.6112 to 0.8341, with corresponding E3 sensitivity/specificity pairs reported in Table 4.

Direct transfer of these operating points produced architecture- and cohort-specific shifts in sensitivity and specificity. In External A, sensitivity/specificity pairs were 0.412/0.704 for ResNet50, 0.549/0.592 for ConvNeXt-Tiny, 0.255/0.849 for ViT-B/16, and 0.725/0.461 for Swin-T. In External B, the corresponding pairs were 0.652/0.448, 0.679/0.345, 0.464/0.586, and 0.759/0.379, respectively (Figure 3C). No threshold re-optimization or probability recalibration was performed using external data.

Table 4. E3 calibration characteristics and locked model-specific thresholds

Model	E3 Brier score	E3 ECE	Calibration intercept	Calibration slope	Locked threshold	Sensitivity	Specificity
ResNet50	0.1778	0.1204	0.8858	0.3355	0.7964	0.632	0.738
ConvNeXt-Tiny	0.1762	0.1369	0.9082	0.2292	0.8341	0.684	0.705
ViT-B/16	0.1457	0.0700	0.7582	0.4552	0.7242	0.720	0.705
Swin-T	0.1845	0.1476	1.0640	0.2997	0.6112	0.768	0.607

Note. ECE, expected calibration error. Thresholds were selected from pooled E3 out-of-fold patient predictions using Youden’s J statistic and were subsequently frozen for both external cohorts.

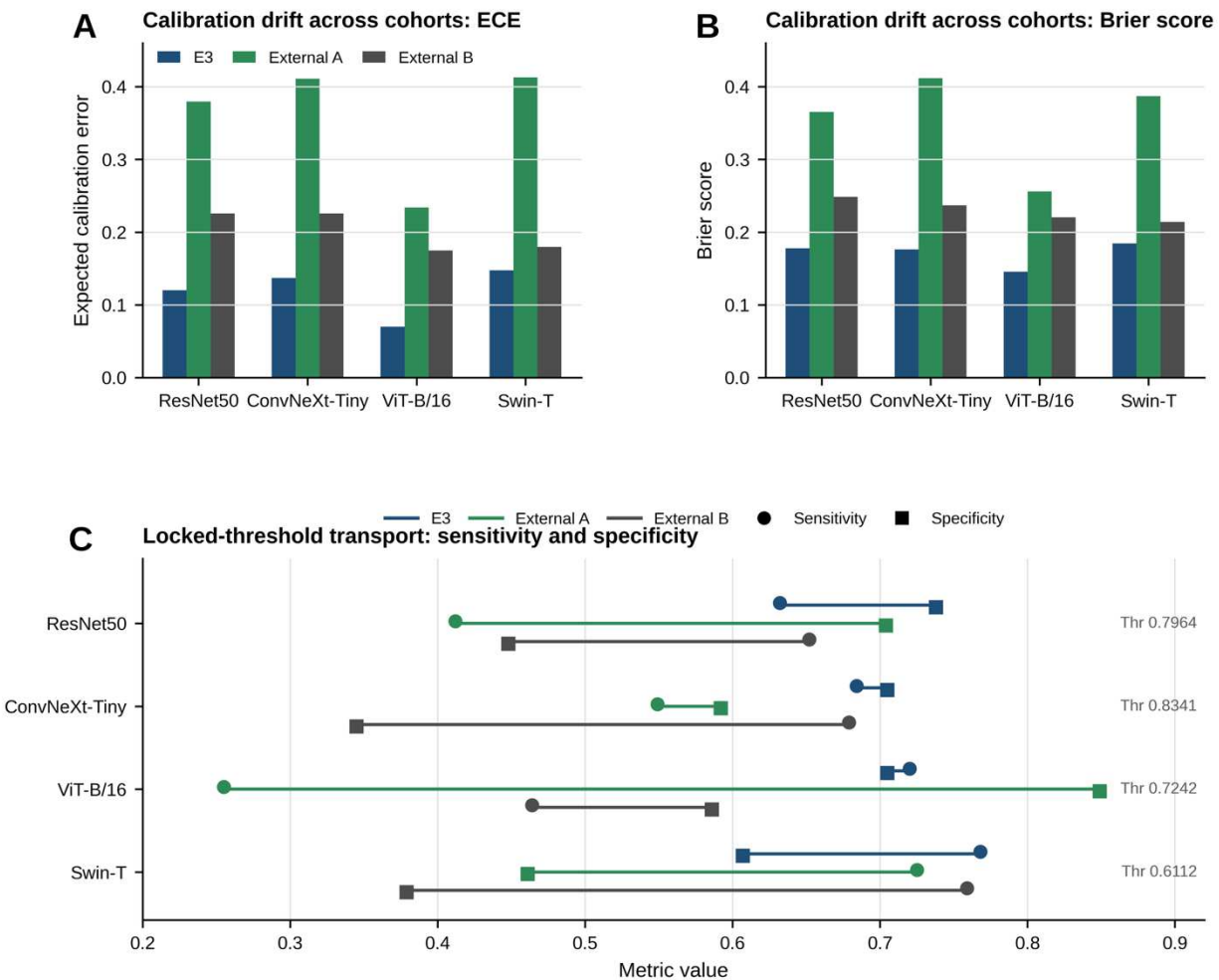


Figure 3. Calibration and locked operating-point transport across internal and external evaluation. (A) Expected calibration error (ECE) across E3, External A, and External B. (B) Brier scores across the same cohorts. (C) Sensitivity and specificity obtained using model-specific thresholds derived from pooled E3 out-of-fold predictions and transferred unchanged to both external cohorts. No external threshold re-optimization or probability recalibration was performed.

3.5. Cross-Cohort Histologic and Acquisition-Domain Shift

Histologic composition differed markedly across cohorts. LUAD prevalence was 80.4% in the development cohort and 79.4% in External B, but only 25.1% in External A (Figure 4A). The LUAD:LUSC distribution differed strongly between the development cohort and External A (OR

12.21, 95% CI 8.00–18.65; Holm-adjusted Fisher exact $p < 0.001$), whereas no corresponding difference was detected between the development cohort and External B (OR 1.06, 95% CI 0.65–1.74; Holm-adjusted $p = 0.801$) (Table 5).

No continuous acquisition variable simultaneously met both the Holm-adjusted $p < 0.05$ criterion and the prespecified moderate-or-larger effect-size threshold of $|\delta| \geq 0.33$ (Figure 4B; Table 5). The largest categorical shift involved convolution-kernel distribution, with Cramér’s $V = 0.842$. Because the corresponding contingency table was sparse, with a minimum expected cell count of approximately 1.29, categorical inference was treated as exploratory, and the effect-size estimate was emphasized (Figure 4C).

Acquisition-domain analysis was limited by available metadata; several potentially relevant scanner descriptors, including native pixel spacing, slice thickness, manufacturer, scanner model, and reconstruction diameter, were unavailable. Accordingly, these analyses characterize measurable cohort differences rather than providing an exhaustive or causal explanation of external performance degradation.

Table 5. Principal cross-cohort distribution-shift findings

Comparison / feature	Development	External cohort	Effect estimate	Adjusted significance
LUAD prevalence: Development vs External A	80.4%	25.1%	OR = 12.21 (95% CI 8.00–18.65)	$p < 0.001$
LUAD prevalence: Development vs External B	80.4%	79.4%	OR = 1.06 (95% CI 0.65–1.74)	$p = 0.801$
Continuous acquisition variables	—	—	No variable met both Holm-adjusted $p < 0.05$ and $ \delta \geq 0.33$.	—

Comparison / feature	Development	External cohort	Effect estimate	Adjusted significance
Convolution-kernel distribution	Development	External cohorts	Cramér's $V = 0.842$	Exploratory due to sparse expected cells

Note. OR, odds ratio for LUAD relative to LUSC; δ , Cliff's delta.

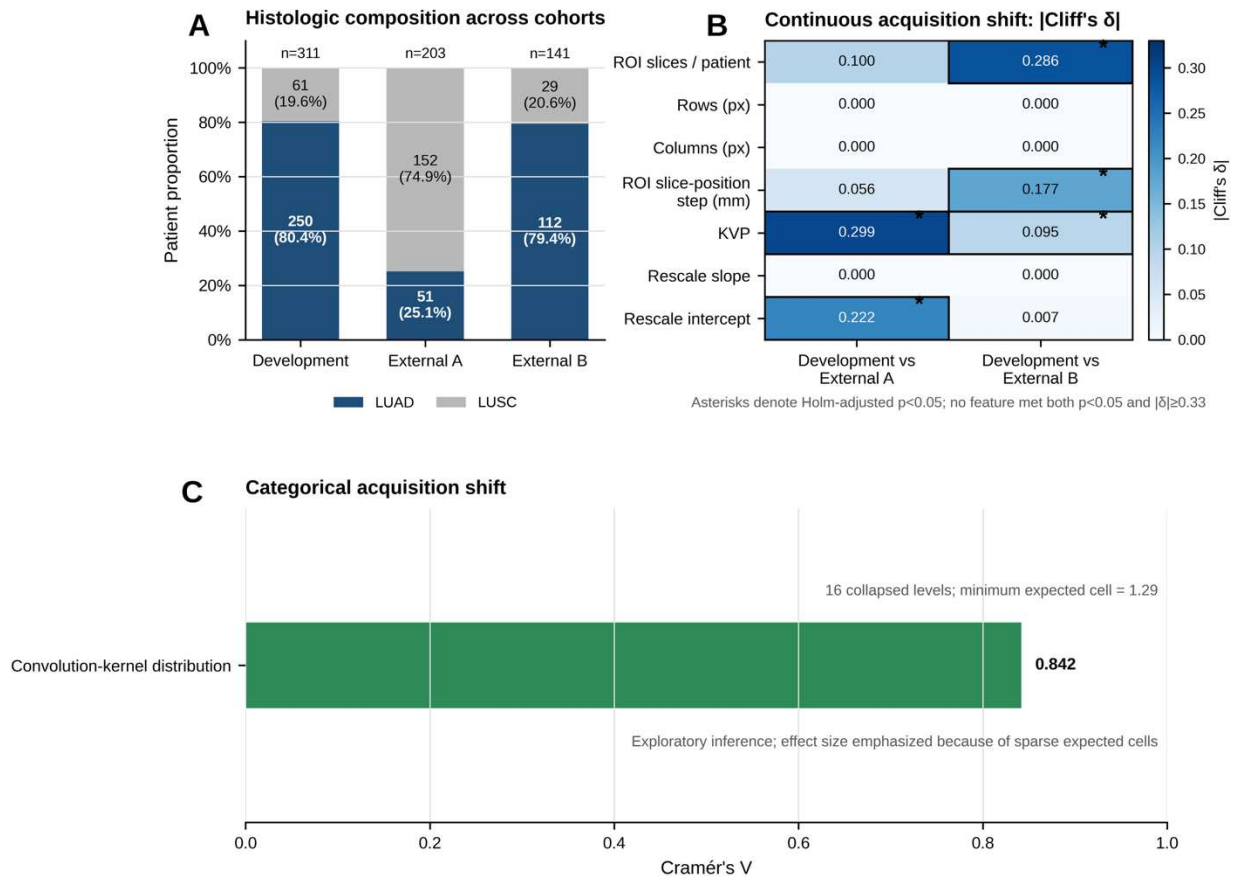


Figure 4. Histologic composition and measurable acquisition-domain differences across cohorts. (A) Patient-level LUAD and LUSC composition across the development, External A, and External B cohorts. (B) Absolute Cliff's delta effect sizes for continuous acquisition differences; asterisks denote Holm-adjusted $p < 0.05$. No feature met both $p < 0.05$ and $|\delta| \geq 0.33$. (C) Convolution-kernel distribution shift summarized by Cramér's $V = 0.842$; inference was considered exploratory because of sparse expected cell counts. Acquisition-domain findings were interpreted as associative rather than causal explanations of external performance degradation.

4. Discussion

The present study examined evaluation structure as an experimental variable in CT-based deep learning for NSCLC histologic subtyping. Across four architectures, patient-level AUROC decreased from approximately 0.99–1.00 under intentionally leakage-prone slice-random evaluation to approximately 0.70–0.73 under strict patient-grouped testing, with paired E1-minus-E3 differences of 0.256–0.297. Because the development cohort, preprocessing, task definition, architectures, optimization framework, and patient-level aggregation were held constant, this contrast isolates the influence of partitioning more directly than comparisons across independently conducted studies. The findings indicate that apparent CT classification performance can be substantially inflated when correlated slices from the same patient cross evaluation boundaries.

This effect is particularly relevant to two-dimensional CT modeling, in which adjacent tumor-bearing slices share lesion morphology, anatomy, reconstruction characteristics, and patient-specific imaging context. Under slice-random partitioning, test images may therefore remain statistically related to images encountered during training even when no exact image is duplicated. Models can consequently exploit patient- or acquisition-specific signals in addition to histology-related features. This interpretation is consistent with shortcut learning [14] and with evidence that medical-imaging models can exploit institutional, scanner, and healthcare-process characteristics that are predictive within a dataset but unstable across environments [15,16]. The present findings are consistent with this broader concern in CT histologic subtyping, showing that apparent performance can change substantially when only the partitioning unit is altered.

Strict patient grouping reduced but did not eliminate discrimination. Under E3, all four architectures retained AUROCs of approximately 0.70–0.73, broadly consistent with previous CT-based studies of adenocarcinoma–squamous cell discrimination [4–8]. This suggests that the development data contained genuine histology-associated information, but at substantially lower

strength than implied by slice-random evaluation. The similarity of E3 performance across ResNet50, ConvNeXt-Tiny, ViT-B/16, and Swin-T further indicates that the dominant experimental contrast was evaluation structure rather than architectural family.

Independent external testing produced a further reduction in discrimination, with AUROCs of approximately 0.60–0.63 in External A and 0.50–0.61 in External B. This progression from near-perfect E1 performance to moderate E3 performance and weaker external performance illustrates that patient-level independence and cross-cohort transportability address different questions. E3 prevents direct patient overlap, whereas external testing additionally changes the population and imaging environment. The observed degradation is consistent with previous evidence that radiologic deep-learning models frequently perform less well outside their development setting [18], reinforcing that internal patient-level testing should not substitute for independent external validation.

The two external cohorts also showed that prevalence shift alone did not explain transportability. External A retained AUROC confidence intervals above 0.50 for all four architectures despite a marked change in LUAD prevalence, whereas External B had a histologic composition similar to development but all AUROC confidence intervals included 0.50. Acquisition metadata likewise demonstrated measurable cross-cohort heterogeneity, including a large difference in convolution-kernel distribution, although no continuous acquisition variable satisfied both the prespecified statistical and moderate-effect-size criteria. Because several potentially relevant scanner descriptors were unavailable, these analyses should be interpreted as evidence of domain heterogeneity rather than as a causal explanation of external performance loss.

Calibration provided complementary evidence of transfer degradation. ECE and Brier scores were less favorable in both external cohorts than under E3, demonstrating that

transportability cannot be characterized by discrimination alone. Internally derived thresholds were transferred unchanged without recalibration, producing cohort-dependent shifts in sensitivity and specificity. These findings distinguish preservation of rank discrimination from preservation of probability calibration and operating-point behavior, both of which are important when prediction models are applied outside their development environment.

The post hoc training-duration analysis reduced concern that the external findings were driven solely by the one-epoch final-training lock. One epoch resulted from the prespecified median best-epoch rule across E3 folds rather than external performance optimization. Extending training to five epochs did not produce a consistent improvement across architectures or cohorts. External A showed only small paired AUROC changes with confidence intervals including zero, whereas in External B only ResNet50 showed a positive paired difference whose confidence interval excluded zero. Nevertheless, the overall pattern was preserved: all five-epoch AUROC confidence intervals remained above 0.50 in External A and included 0.50 in External B. Training duration therefore influenced some model-specific estimates but did not account for the overall cross-cohort pattern.

Several limitations should be acknowledged. First, the study used retrospective public CT datasets and was restricted to binary LUAD–LUSC classification, limiting direct generalization to other histologies or prospective clinical populations. External precision was also constrained by relatively small minority-class counts, particularly 51 LUAD cases in External A and 29 LUSC cases in External B. Second, the models used two-dimensional ROI-bearing slices with patient-level averaging; three-dimensional or explicitly spatial approaches may behave differently. Third, acquisition-domain analysis was limited by available metadata and could not comprehensively evaluate variables such as manufacturer, scanner model, native pixel spacing, slice thickness, or

reconstruction diameter. Fourth, no external fine-tuning, recalibration, or domain adaptation was performed because the objective was to measure unmodified transportability. Accordingly, the study does not determine how much external performance might be recovered through such adaptation. Finally, the four architectures were selected to represent distinct model families rather than to provide an exhaustive architecture or hyperparameter benchmark.

These limitations are balanced by several design strengths. Evaluation leakage was introduced as a controlled experimental condition and compared directly with patient-grouped testing using the same patients and imaging inputs. Internal differences were quantified with paired patient-level resampling, external evaluation used two independent cohorts with predefined analytical roles, and thresholds were derived exclusively from E3 predictions and transferred without external optimization. Together with joint assessment of discrimination, calibration, operating-point behavior, and measurable domain differences, the findings support a central methodological conclusion: in repeated-measure medical imaging datasets, evaluation design can influence apparent performance at least as strongly as architectural choice. Credible assessment of CT-based deep-learning systems therefore requires explicit control of patient-level dependence and independent external testing before claims of generalizability are made.

5. Conclusions

This study demonstrates that evaluation design substantially influences the apparent performance of deep learning models for CT-based NSCLC histologic subtyping. Slice-random evaluation produced near-perfect discrimination across four architectures, whereas strict patient-grouped testing reduced AUROC and independent external testing further decreased discrimination and calibration. The five-epoch sensitivity analysis preserved the overall external-validation pattern. These findings indicate that patient-level independence is necessary but

493 insufficient to establish transportability. Reliable assessment of imaging models should therefore
494 incorporate patient-level partitioning, independent external validation, locked decision thresholds,
495 and transparent reporting of calibration and cohort differences. In repeated-measure imaging
496 datasets, evaluation structure should be treated as a central component of experimental design
497 rather than as a secondary implementation detail.

498 **List of abbreviations**

Abbreviation	Definition
AUROC	Area under the receiver operating characteristic curve
CI	Confidence interval
CT	Computed tomography
DICOM	Digital Imaging and Communications in Medicine
E1	Slice-random evaluation design
E3	Patient-grouped evaluation design
ECE	Expected calibration error
HU	Hounsfield unit
KVP	Peak kilovoltage
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MCC	Matthews correlation coefficient
NSCLC	Non–small cell lung cancer
OR	Odds ratio

Abbreviation	Definition
PR-AUC	Precision–recall area under the curve
ROI	Region of interest
SHA-256	Secure Hash Algorithm 256-bit
SMOTE	Synthetic Minority Over-sampling Technique
SOP Instance UID	Service-Object Pair Instance Unique Identifier
TCIA	The Cancer Imaging Archive

499 **Declarations**

500 **Ethics approval and consent to participate**

501 This study was a secondary analysis of publicly available, de-identified imaging data obtained
502 from The Cancer Imaging Archive (TCIA). No new participants were recruited, no intervention
503 was performed, and no identifiable private information was accessed. Because the study involved
504 exclusively publicly available, de-identified data and no direct interaction with participants, ethics
505 approval and informed consent to participate were not required for this secondary analysis.

506 **Consent for publication**

507 Not applicable.

508 **Availability of data and materials**

509 The imaging datasets analyzed during the current study are publicly available through The Cancer
510 Imaging Archive (TCIA): Lung-PET-CT-Dx (<https://doi.org/10.7937/TCIA.2020.NNC2-0461>),
511 NSCLC-Radiomics (<https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI>) [33], and NSCLC-
512 Radiogenomics (<https://doi.org/10.7937/K9/TCIA.2017.7hs46erv>) [34]. Derived cohort

manifests, model prediction outputs, calibration outputs, and analysis scripts are available from the corresponding author upon reasonable request.

Competing interests

The author declares that there are no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

NV conceptualized the study, developed the methodology and software, performed validation and formal analysis, curated and investigated the data, prepared the visualizations, and drafted and revised the manuscript. NV also administered the project and approved the final manuscript.

Acknowledgements

The author gratefully acknowledges KTO Karatay University for its institutional support.

Generative AI statement

During the preparation of this work, the author used Grammarly (Grammarly Inc.) solely for language editing to improve grammar and readability. The tool was not used to generate scientific content, analyses, results, or interpretations. After using this tool, the author reviewed and edited the text as needed and takes full responsibility for the content of the publication.

References

[1] WHO Classification of Tumours Editorial Board. Thoracic Tumours. 5th ed. Vol. 5. Lyon: International Agency for Research on Cancer; 2021.

534 [2] Aerts HJWL, Rios Velazquez E, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al.
535 Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat*
536 *Commun.* 2014;5:4006. <https://doi.org/10.1038/ncomms5006>.

537 [3] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images are more than pictures, they are data.
538 *Radiology.* 2016;278(2):563–577. <https://doi.org/10.1148/radiol.2015151169>.

539 [4] Linning E, Lu L, Li L, Yang H, Schwartz LH, Zhao B. Radiomics for classification of lung
540 cancer histological subtypes based on nonenhanced computed tomography. *Acad Radiol.*
541 2019;26(9):1245–1252. <https://doi.org/10.1016/j.acra.2018.10.013>.

542 [5] Ferreira-Junior JR, Koenigkam-Santos M, Magalhães Tenório AP, Faleiros MC, Garcia
543 Cipriano FE, Fabro AT, et al. CT-based radiomics for prediction of histologic subtype and
544 metastatic disease in primary malignant lung neoplasms. *Int J Comput Assist Radiol Surg.*
545 2020;15(1):163–172. <https://doi.org/10.1007/s11548-019-02093-y>.

546 [6] Chaunzwa TL, Hosny A, Xu Y, Shafer A, Diao N, Lanuti M, et al. Deep learning classification
547 of lung cancer histology using CT images. *Sci Rep.* 2021;11(1):5471.
548 <https://doi.org/10.1038/s41598-021-84630-x>.

549 [7] Marentakis P, Karaiskos P, Kouloulas V, Kelekis N, Argentos S, Oikonomopoulos N, Loukas
550 C. Lung cancer histology classification from CT images based on radiomics and deep learning
551 models. *Med Biol Eng Comput.* 2021;59(1):215–226.
552 <https://doi.org/10.1007/s11517-020-02302-w>.

553 [8] Guo Y, Song Q, Jiang M, Guo Y, Xu P, Zhang Y, et al. Histological subtypes classification of
554 lung cancers on CT images using 3D deep learning and radiomics. *Acad Radiol.* 2021;28(9):e258–
555 e266. <https://doi.org/10.1016/j.acra.2020.06.010>.

- [9] Song F, Song X, Feng Y, Fan G, Sun Y, Zhang P, et al. Radiomics feature analysis and model research for predicting histopathological subtypes of non-small cell lung cancer on CT images: A multi-dataset study. *Med Phys*. 2023;50(7):4351–4365. <https://doi.org/10.1002/mp.16233>.
- [10] Liang B, Tong C, Nong J, Zhang Y. Histological subtype classification of non-small cell lung cancer with radiomics and 3D convolutional neural networks. *J Imaging Inform Med*. 2024;37(6):2895–2909. <https://doi.org/10.1007/s10278-024-01152-4>.
- [11] Oh G, Gi Y, Lee J, Kim H, Wu HG, Park JM, et al. Hybrid approach to classifying histological subtypes of non-small cell lung cancer (NSCLC): Combining radiomics and deep learning features from CT images. *J Imaging Inform Med*. 2025;38(6):3535–3547. <https://doi.org/10.1007/s10278-025-01442-5>.
- [12] Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing medical imaging data for machine learning. *Radiology*. 2020;295(1):4–15. <https://doi.org/10.1148/radiol.2020192224>.
- [13] Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295(2):328–338. <https://doi.org/10.1148/radiol.2020191145>.
- [14] Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, Wichmann FA. Shortcut learning in deep neural networks. *Nat Mach Intell*. 2020;2:665–673. <https://doi.org/10.1038/s42256-020-00257-z>.
- [15] Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med*. 2018;15(11):e1002683. <https://doi.org/10.1371/journal.pmed.1002683>.

- [16] Badgeley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digit Med*. 2019;2:31. <https://doi.org/10.1038/s41746-019-0105-1>.
- [17] Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell*. 2021;3:199–217. <https://doi.org/10.1038/s42256-021-00307-0>.
- [18] Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: A systematic review. *Radiol Artif Intell*. 2022;4(3):e210064. <https://doi.org/10.1148/ryai.210064>.
- [19] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group ‘Evaluating diagnostic tests and prediction models’ of the STRATOS initiative. Calibration: The Achilles heel of predictive analytics. *BMC Med*. 2019;17:230. <https://doi.org/10.1186/s12916-019-1466-7>.
- [20] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: *Proceedings of the 34th International Conference on Machine Learning*. Proc Mach Learn Res. 2017;70:1321–1330.
- [21] Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, Kahn CE Jr; CLAIM 2024 Update Panel. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiol Artif Intell*. 2024;6(4):e240300. <https://doi.org/10.1148/ryai.240300>.
- [22] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. <https://doi.org/10.1136/bmj-2023-078378>.

602 [23] Li P, Wang S, Li T, Lu J, HuangFu Y, Wang D. A Large-Scale CT and PET/CT Dataset for
603 Lung Cancer Diagnosis (Lung-PET-CT-Dx) [dataset]. The Cancer Imaging Archive; 2020.
604 <https://doi.org/10.7937/TCIA.2020.NNC2-0461>.

605 [24] Bakr S, Gevaert O, Echegaray S, Ayers K, Zhou M, Shafiq M, et al. A radiogenomic dataset
606 of non-small cell lung cancer. *Sci Data*. 2018;5:180202. <https://doi.org/10.1038/sdata.2018.202>.

607 [25] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings*
608 *of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016:770–778.
609 <https://doi.org/10.1109/CVPR.2016.90>.

610 [26] Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. In:
611 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
612 2022:11966–11976. <https://doi.org/10.1109/CVPR52688.2022.01167>.

613 [27] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An
614 image is worth 16×16 words: Transformers for image recognition at scale. In: *International*
615 *Conference on Learning Representations*; 2021.

616 [28] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: Hierarchical vision
617 transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on*
618 *Computer Vision*. 2021:9992–10002. <https://doi.org/10.1109/ICCV48922.2021.00986>.

619 [29] Loshchilov I, Hutter F. Decoupled weight decay regularization. In: *International Conference*
620 *on Learning Representations*; 2019.

621 [30] Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32–35.
622 [https://doi.org/10.1002/1097-0142\(1950\)3:1<32::AID-CNCR2820030106>3.0.CO;2-3](https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3).

623 [31] Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 1979;6(2):65–
624 70. <http://www.jstor.org/stable/4615733>

625 [32] Cliff N. Dominance statistics: Ordinal analyses to answer ordinal questions. Psychol Bull.
626 1993;114(3):494–509. <https://doi.org/10.1037/0033-2909.114.3.494>.

627 [33] Aerts HJWL, Wee L, Rios Velazquez E, Leijenaar RTH, Parmar C, Grossmann P, et al. Data
628 From NSCLC-Radiomics (Version 4) [dataset]. The Cancer Imaging Archive. 2014.
629 <https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI>.

630 [34] Bakr S, Gevaert O, Echegaray S, Ayers K, Zhou M, Shafiq M, et al. Data for NSCLC
631 Radiogenomics (Version 4) [dataset]. The Cancer Imaging Archive. 2017.
632 <https://doi.org/10.7937/K9/TCIA.2017.7hs46erv>.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [04SupplementaryInformation30092026.docx](#)