Mechanosensitive TRPV4 immunohistochemistry improves deep learning-based grading of ductal carcinoma in situ beyond H&E morphology

Janghyun Yoo[1,†], Raghav Karthikeyan[2,3,†], Kashi Kamat[2,3], Christopher Chan[2], Shabnam Samankan[5], Elham Arbzadeh[5], Arnold Schwartz[5], Patricia S. Latham[5], Inhee Chung[2,4,*]

Affiliations:

1 Department of Physics and Astronomy, College of Natural Sciences, Seoul National University, Seoul, South Korea

2 Department of Anatomy and Cell Biology, School of Medicine and Health Sciences, George Washington University, Washington, DC, USA

3 Thomas Jefferson Science and Technology High School, Alexandria, VA, USA

4 Department of Biomedical Engineering, GW School of Engineering and Applied Science, George Washington University, Washington, DC, USA

5 Department of Pathology, George Washington Medical Faculty Associates, Washington, DC, USA

[†] These authors contributed equally to this work.

***** Correspondence: inheec@gwu.edu

Abstract

Ductal carcinoma in situ (DCIS) is a non-invasive breast cancer spanning a biologic continuum from atypical ductal hyperplasia (ADH) to high-grade lesions with variable risk of progression to invasive ductal carcinoma (IDC), yet diagnostic accuracy remains limited when based on morphologic assessment via hematoxylin and eosin (H&E) alone. TRPV4, a mechanosensitive ion channel we previously demonstrated to exhibit pathology-dependent spatial distribution patterns in DCIS, offers a biologically motivated immunohistochemical (IHC) marker that may refine classification beyond routine H&E assessment. We evaluated whether deep learning models trained on TRPV4 IHC outperform those trained on H&E for DCIS classification. We assembled a multi-institutional dataset of paired H&E and TRPV4 IHC whole-slide images from 108 patients (24,248 image tiles), with both stains available for most cases in an internal development cohort (n=69) and an external test cohort (n=39). Each cohort was digitized on different scanners to assess cross-platform robustness. Tiles from annotated regions were grouped into four ordered classes reflecting DCIS progression: normal/benign, ADH/low-grade DCIS, high-grade DCIS, and IDC. Xception and EfficientNet-B0 convolutional neural networks were trained with patient-level 3-fold cross-validation on the development cohort and evaluated as ensembles on the test cohort. On external testing at the patient level, H&E-based ensembles showed moderate performance (macro-F1=0.43–0.44, macro-AUC=0.73–0.80), whereas TRPV4 IHC-based models substantially improved classification (macro-F1=0.68–0.72, macro-AUC=0.91–0.92). Across tile-level predictions, 68–79% of errors were between adjacent grades, consistent with an ordinal DCIS spectrum. Per-class tile-level analyses on the external test cohort showed the greatest improvement with TRPV4 IHC over

H&E for ADH/low-grade DCIS (AUC 0.83–0.84 vs 0.70–0.81) and IDC (AUC 0.74–0.79 vs 0.65–0.66), supporting classification across the DCIS progression spectrum. These findings support TRPV4 IHC as a mechanistically grounded complement to H&E, improving deep learning-based DCIS classification in a pilot multi-institutional setting.

Keywords

ductal carcinoma in situ, breast pathology, immunohistochemistry, TRPV4, deep learning, digital pathology, whole-slide imaging, external validation, histopathologic grading, convolutional neural networks

Introduction

Ductal carcinoma in situ (DCIS) is a non-invasive breast cancer characterized by the proliferation of neoplastic epithelial cells confined within breast ducts and accounts for approximately 20% of newly diagnosed breast cancer cases (1, 2). Fewer than half of DCIS lesions progress to IDC if left untreated (3, 4), yet there is no reliable method to predict which cases will become invasive. Clinicopathologic factors and multigene assays have been developed to predict recurrence risk but show limited ability to predict IDC progression (5-14). As a result, the current standard of care frequently favors surgical intervention, raising concern for overtreatment (15, 16). Histopathologic evaluation of H&E-stained sections remains central to diagnosis and grading (15-20), but morphology-based assessment is inherently subjective and can limit both accuracy and precision, especially at biologically and clinically important boundaries (21-24).

In routine practice, breast lesions are categorized as benign, atypical ductal hyperplasia (ADH), DCIS, or IDC, with DCIS further stratified into low-, intermediate-, and high-grade based on nuclear atypia, architecture, and necrosis (15-20). However, these categories lie along a biologic continuum rather than representing discrete states. Small differences in nuclear features or necrosis can shift a lesion between ADH and low-grade DCIS or between DCIS grades, making some distinctions particularly fragile. The most diagnostically challenging and clinically consequential transitions include: (1) normal/benign epithelium versus ADH/low-grade DCIS (LG-DCIS), (2) ADH/LG-DCIS versus high-grade DCIS (HG-DCIS), and (3) the distinction of DCIS from IDC (25-27). Accordingly, in this study, we explicitly focus on four biologically ordered diagnostic groups along this DCIS-related progression axis to directly interrogate the spectrum most relevant to risk stratification and management decisions. Intermediate-grade DCIS cases, which are known to show particularly low interobserver agreement (25), were omitted from this initial deep learning analysis to reduce label noise in a modest-sized cohort.

To address limitations of morphology alone, deep learning approaches have been applied to breast pathology to computationally distinguish normal or benign tissue from DCIS, and between DCIS and IDC, typically using morphologic features from H&E (28-33). While these methods may enhance overall classification performance and reproducibility relative to purely manual assessment, they remain constrained by their reliance on morphology and therefore inherit the subjectivity and ambiguity of traditional grading. A critical unmet need is to integrate biologically grounded, mechanistically

meaningful markers with morphology within deep learning-enabled diagnostic frameworks to improve DCIS-focused classification across this progression spectrum.

We and others have characterized the mechanosensitive, calcium-permeable ion channel TRPV4 as a regulator of cellular responses to mechanical stress (34-37). In our recent study, we showed that in confined DCIS ducts, spatial crowding stress drives TRPV4 relocalization from intracellular compartments to the plasma membrane as part of a crowding-induced pro-invasive mechanotransduction program (37). Recent work from our group also indicates that plasma membrane-localized TRPV4 is enriched in high-risk DCIS and associated with subsequent progression to IDC (38), supporting the hypothesis that TRPV4 IHC captures mechanotransduction-related information not apparent on H&E alone. These mechanistic insights suggest that TRPV4 IHC images capture both morphologic features and biologically relevant protein distribution patterns that H&E cannot detect. This provides a strong rationale for comparing deep learning model performance on TRPV4 IHC versus H&E images for DCIS classification.

In the present study, we leverage TRPV4 IHC as a mechanotransduction-based stain with prior mechanistic justification (37, 38) and test whether deep learning models trained on TRPV4 IHC improve DCIS diagnostic performance compared with models trained on H&E. We assembled a multi-institutional dataset of 24,248 image tiles from pathology-specific regions of interest (ROIs) on formalin-fixed paraffin-embedded (FFPE) specimens from 108 DCIS-containing cases with mostly matched H&E and TRPV4 IHC whole-slide images, drawn from a UVA development cohort at the University of Virginia

(n=69) and an independent institutional test cohort at George Washington University (n=39). For each case, adjacent FFPE sections were stained with H&E and TRPV4 IHC to generate paired whole-slide images; slides and tiles failing pre-specified quality-control criteria (e.g., suboptimal staining, scanning artifacts, or incomplete annotations) were excluded. Tiles from pathologist-annotated regions were grouped into four ordered classes: normal/benign epithelium, ADH/LG-DCIS, HG-DCIS, and IDC, corresponding to the diagnostically critical groups described above.

We then trained and compared two convolutional neural network (CNN) architectures, Xception (39) and EfficientNet-B0 (40). Xception served as a higher-capacity backbone for capturing subtle spatial patterns, whereas EfficientNet-B0 provided a more parameter-efficient alternative to mitigate overfitting on this modest-sized cohort. Both architectures were trained separately on H&E and TRPV4 IHC image tiles using patient-level 3-fold cross-validation in the development cohort, and final performance was assessed by applying UVA-trained ensembles to the GWU cohort without retraining or fine-tuning. We quantified performance using macro-averaged F1-scores and areas under the receiver operating characteristic curve (ROC-AUC) at both tile and patient levels, along with per-class metrics and adjacency-weighted error maps to characterize the distribution of misclassifications. Together, this framework allows us to test whether deep learning models trained on TRPV4 IHC, capturing both tissue morphology and TRPV4 protein distribution, outperform models trained on H&E morphology alone, and whether any performance gains transfer to an independent institutional test cohort.

Materials and Methods

Study Design and Patient Cohorts

We conducted a multi-institutional retrospective study to develop and validate deep learning models for classifying TRPV4 immunohistochemistry (IHC) and matched hematoxylin and eosin (H&E) breast pathology. The study was approved by the institutional review boards at George Washington University (GWU) and the University of Virginia (UVA) with a waiver of informed consent for de-identified archival material. Specimens were obtained in part through the National Cancer Institute's Cooperative Human Tissue Network (CHTN; RRID: SCR_004446), coordinated at UVA.

After excluding intermediate-grade DCIS cases, the final dataset comprised 108 unique patients and 24,248 annotated tiles (299×299-pixel image patches extracted from whole-slide images). Patients were divided into a UVA development cohort (n=69) and an independent GWU external test cohort (n=39) (**Fig. 1A**; **Supplementary Table S1**). The GWU cohort substantially overlaps with cases previously analyzed in our mechanistic study of TRPV4 mechanotransduction in DCIS (37). Within UVA, evaluable H&E slides were available for 62 patients (9,623 tiles) and TRPV4 IHC for 51 patients (6,792 tiles). Within GWU, H&E slides were available for 31 patients (4,511 tiles) and TRPV4 IHC for 33 patients (3,322 tiles). Not all patients had both modalities available because of tissue exhaustion or staining/quality-control exclusions. Across cohorts and stain modalities, class frequencies were moderately imbalanced and differed between UVA and GWU (**Fig. 1A**; **Supplementary Table S1**).

UVA cases were used exclusively for model training, hyperparameter selection, and patient-level three-fold cross-validation. GWU cases were held out entirely for external testing and were not used at any stage of training or model selection. All data partitioning was performed at the patient level to prevent information leakage. Patient IDs were assigned using stratified 3-fold cross-validation to yield folds with comparable class distributions across the four ordered diagnostic groups.

Histologic Preparation, TRPV4 IHC, and Image Acquisition

Consecutive 4-µm FFPE sections from each case were stained with routine H&E and TRPV4 IHC; the GWU cohort largely overlaps with cases from our prior mechanistic study of TRPV4 mechanotransduction (37). TRPV4 immunostaining was performed on a Ventana Discovery Ultra automated platform using anti-TRPV4 antibody (Abcam ab39260, RRID:AB_1143677; rabbit polyclonal; 1:40) with CC1 (Tris-EDTA, pH 8.0) heat-induced epitope retrieval, HRP/DAB detection (OmniMap anti-rabbit HRP; ChromoMap DAB), and hematoxylin counterstaining, as previously described (37, 38). Each run included positive-control tissue and negative controls with primary antibody omitted.

Whole-slide images were digitized at 40× magnification. UVA slides were scanned using a Hamamatsu NanoZoomer (~0.25 µm/pixel), and GWU slides were scanned by VitroVivo Biotech (Rockville, MD) using an Olympus VS200 with comparable resolution. Three board-certified breast pathologists independently reviewed slides and annotated epithelial regions of interest (ROIs) including ducts and terminal duct-lobular units, while

excluding artifacts (e.g., tissue folds). Non-overlapping tiles at 299 × 299 pixels (matching standard deep learning input dimensions; example tiles in **Supplementary Fig. S1**) were extracted from annotated ROIs (examples shown in **Fig. 1B**) on both H&E and TRPV4 IHC slides; each tile was linked to patient ID, institution, modality, and diagnostic label.

Diagnostic Labels and Ordinal Grouping

For ROIs reviewed by multiple pathologists, the final diagnostic label was assigned by majority vote; equivocal or discordant ROIs without clear majority were excluded. Because pathologists reviewed different subsets of ROIs based on case availability, formal inter-rater reliability could not be calculated. Original labels included: normal, benign, ADH, LG-DCIS, HG-DCIS, IDC, and a small set of indeterminate ROIs annotated as intermediate-grade DCIS, which were excluded a priori.

Patient-Level Cross-Validation and Data Partitioning

To avoid leakage of patient-specific morphology across folds, we defined all splits at the level of patient ID. For each modality (TRPV4 IHC and H&E) within the UVA cohort, we assigned patients to 3 cross-validation folds as follows: for each patient, we computed the majority label group among that patient's image tiles. Patients were then stratified by this majority label, and within each label group, patients were ordered by decreasing tile count and distributed in a round-robin fashion across three folds (Fold 0, 1, and 2). This yielded three patient-disjoint folds with similar class distributions at the image tile level. For each fold in turn, the held-out fold served as internal validation, and the remaining

two folds were combined as the training set. All GWU patients were reserved as the external test cohort and were not included in any cross-validation procedure.

Image Preprocessing and Data Augmentation

Models were implemented in TensorFlow 2.10.1/Keras and trained on a GPU workstation (batch size=4). RGB tiles were resized to 299×299, normalized to [0,1], and processed with the backbone-specific ImageNet preprocessing function; augmented outputs were clipped to [0,1]. During training only, we applied on-the-fly geometric and photometric augmentations to both H&E and TRPV4 IHC tiles, including random flips, 90° rotations, random zoom/cropping, color jitter, and additive Gaussian noise. Validation and external test tiles were resized and preprocessed identically but were not augmented.

Network Architectures and Implementation

We trained two ImageNet-pretrained CNN backbones (Xception and EfficientNet-B0) for each imaging modality, replacing the original classification heads with a shared architecture: global average pooling, a 256-unit ReLU layer with L2 regularization (λ=0.001), dropout (0.6), and a 4-class output layer (**Fig. 1C**). Mixed-precision training (Keras mixed-float16 policy) was enabled to reduce memory use; the final layer was kept in float32 for numerical stability.

Training Procedure, Optimization, and Class Imbalance

For each stain–backbone combination and cross-validation fold, we used a two-stage transfer-learning protocol. In Stage 1, the ImageNet-pretrained backbone was frozen and only the classification head was trained; in Stage 2, the upper backbone layers were unfrozen for fine-tuning. Models were optimized with Adam using a class-weighted, label-smoothed categorical cross-entropy objective to address class imbalance, with class weights set inversely to training-set frequency and additional upweighting of IDC to penalize missed invasive disease. Training used validation macro-F1 for learning-rate scheduling and early stopping to limit overfitting. A class-weighted focal-loss variant was explored but was not used for primary experiments due to less stable optimization.

Ensemble Inference, Image- and Patient-Level Predictions

For each modality–architecture pair, three models (one per patient-level cross-validation fold) were trained on the UVA development cohort. For external testing, we applied the UVA-trained fold models to the GWU cohort without retraining or fine-tuning and averaged predicted class probabilities across folds to form an ensemble. Metrics were computed at both the tile level and the patient level. For patient-level inference, tile probability vectors were aggregated by mean probability pooling, and the class with the highest mean probability was assigned as the patient-level prediction. No GWU data were used for training, model selection, or hyperparameter tuning.

Evaluation Metrics and Statistical Analysis

Model performance was summarized by macro-averaged F1 score (macro-F1), computed across the four classes to weight each class equally regardless of frequency.

Macro-AUC (ROC-AUC) was computed in a one-vs-rest framework and macro-averaged across the four classes using the model's softmax probabilities. Tile-level macro-AUC was computed across all tiles, and patient-level macro-AUC was computed after aggregating tile probability vectors within each patient by mean-probability pooling. Sensitivity, precision, and per-class F1 score were calculated using standard definitions derived from the confusion matrix. For the external GWU test set, we estimated 95% confidence intervals (CI) for per-class F1 and AUC using non-parametric bootstrapping with 500 resamples. In each resample, we randomly selected patients (with replacement) and included all tiles from those patients, then recalculated both tile-level and patient-level metrics. We reported the 2.5th and 97.5th percentiles across resamples as the CI bounds.

Error Analysis and Adjacent-Grade Misclassification

Given the ordinal progression of the four labels, we distinguished adjacent-grade errors (normal/benign vs ADH/LG-DCIS, ADH/LG-DCIS vs HG-DCIS, HG-DCIS vs IDC) from non-adjacent errors (e.g., normal/benign vs IDC). From the multiclass tile-level confusion matrix, we quantified the total number of misclassified tiles, the subset differing by exactly one step in the ordinal scale, and the fraction of all errors that were adjacent-grade, providing a clinically interpretable measure of error severity.

Results

Internal Cross-Validation Performance on Development Cohort

Both Xception and EfficientNet-B0 architectures were trained separately on H&E and TRPV4 IHC tiles (**Fig. 1A,B**; representative input tiles shown in **Supplementary Fig. S1**) using patient-level 3-fold cross-validation in the UVA cohort (**Fig. 1C**; **Supplementary Table S2**). Training was monitored with early stopping to reduce overfitting. Unless otherwise specified, values are reported as mean±SD, where SD denotes the standard deviation across the three cross-validation folds.

At the tile level, H&E- and TRPV4 IHC-based models showed broadly comparable macro performance (**Supplementary Table S2**). For H&E, validation tile-level macro-F1 ranged from 0.36–0.74 across folds (0.56±0.19 for EfficientNet-B0; 0.59±0.09 for Xception), with macro-AUC of 0.71–0.93 (0.84±0.11 for EfficientNet-B0; 0.85±0.07 for Xception). For TRPV4 IHC, tile-level macro-F1 ranged from 0.49–0.76 (0.58±0.08 for EfficientNet-B0; 0.68±0.07 for Xception), with macro-AUC of 0.83–0.94 (0.87±0.03 for EfficientNet-B0; 0.90±0.04 for Xception).

At the patient level, aggregation increased separation between modalities (**Supplementary Table S2**). H&E models achieved patient-level macro-F1 of 0.60±0.09 (EfficientNet-B0) and 0.61±0.10 (Xception), with patient-level macro-AUC of 0.84±0.10 and 0.89±0.07, respectively. TRPV4 IHC models improved patient-level macro-F1 to 0.76±0.11 (EfficientNet-B0) and 0.75±0.09 (Xception), with patient-level macro-AUC increasing to 0.97±0.02 and 0.97±0.02, respectively. This pattern, modest differences at the tile level but markedly larger gains at the patient level with TRPV4 IHC, suggests

that TRPV4-related signal becomes more reliable when aggregated across multiple tiles within each patient.

Per-class performance on internal validation folds (**Supplementary Table S3**) showed improvements with TRPV4 IHC vs H&E. For ADH/LG-DCIS, mean tile-level F1 increased in both architectures (EfficientNet-B0: 0.67±0.17 vs 0.65±0.16; Xception: 0.71±0.10 vs 0.62±0.12), alongside increased sensitivity (EfficientNet-B0: 0.76±0.20 vs 0.70±0.10; Xception: 0.67±0.17 vs 0.61±0.01) and precision (increased for Xception: 0.80±0.13 vs 0.66±0.24; similar for EfficientNet-B0: 0.60±0.17 vs 0.62±0.21). For HG-DCIS, TRPV4 IHC improved F1 (EfficientNet-B0: 0.56±0.16 vs 0.51±0.33; Xception: 0.69±0.08 vs 0.58±0.19) with higher sensitivity (EfficientNet-B0: 0.51±0.12 vs 0.45±0.34; Xception: 0.69±0.18 vs 0.58±0.29) and precision (EfficientNet-B0: 0.65±0.26 vs 0.59±0.30; Xception: 0.74±0.19 vs 0.63±0.08). For IDC, mean tile-level F1 was unchanged for EfficientNet-B0 (0.50±0.27 vs 0.50±0.27) but increased for Xception (0.67±0.29 vs 0.57±0.24), with precision increasing in both architectures (EfficientNet-B0: 0.94±0.09 vs 0.86±0.13; Xception: 0.93±0.09 vs 0.76±0.07).

External Test Performance: TRPV4 IHC versus H&E

To assess generalizability, we evaluated ensembles formed by averaging predictions from the three UVA cross-validation fold models for each architecture and modality on the independent GWU cohort, without retraining (**Fig. 1A,B**; **Table 1**; **Fig. 2A,B**). At the tile level (**Fig. 2A**; **Table 1**), H&E ensembles achieved macro-F1 of 0.47 (EfficientNet-B0) and 0.51 (Xception), with macro-AUC of 0.78 and 0.80, respectively. TRPV4 IHC

ensembles improved to macro-F1 of 0.53 (EfficientNet-B0) and 0.60 (Xception), with macro-AUC of 0.84 and 0.85, corresponding to 12.8% and 17.6% relative improvements in tile-level macro-F1 (**Table 1**).

At the patient level (**Fig. 2B**; **Table 1**), H&E ensembles achieved macro-F1 of 0.43 (EfficientNet-B0) and 0.44 (Xception), with macro-AUC of 0.80 and 0.73, respectively. TRPV4 IHC ensembles improved to macro-F1 of 0.72 (EfficientNet-B0) and 0.68 (Xception), with macro-AUC of 0.91 and 0.92, corresponding to 67.4% and 54.5% relative improvements in patient-level macro-F1 (**Table 1**).

Class-wise discrimination at the tile level is visualized by ROC curves (**Fig. 2C**). In the Xception comparison shown, TRPV4 IHC improved per-class AUC relative to H&E, particularly for HG-DCIS (0.86 vs 0.82) and IDC (0.79 vs 0.66), while macro-AUC increased from 0.80 to 0.85 (**Table 1**; **Table 2**). ADH/low-grade DCIS also improved modestly (0.83 vs 0.81) (**Table 2**).

Tile-level per-class F1 scores similarly showed consistent TRPV4 IHC gains across architectures (**Fig. 2D**; **Table 2**), with the largest increases for ADH/low-grade DCIS (0.31–0.43 to 0.49–0.56) and IDC (0.20–0.26 to 0.33–0.46).

Patient-level per-class performance comparison

Fig. 3 summarizes patient-level sensitivity (**Fig. 3A**), precision (**Fig. 3B**), and AUC (**Fig. 3C**) by diagnostic category for all four modality-architecture combinations on the GWU

external test cohort (**Supplementary Table S4**). Sensitivity (**Fig. 3A**) showed architecture-dependent patterns, whereas AUC (**Fig. 3C**) consistently favored IHC over H&E, with the largest gains for ADH/low-grade DCIS (IHC 0.94–0.95 vs H&E 0.61–0.70) and IDC (IHC 0.77–0.85 vs H&E 0.61–0.69) (**Supplementary Table S4**). Sensitivity improvements were most prominent for EfficientNet-B0 in ADH/low-grade DCIS (IHC 1.00 vs H&E 0.25) and high-grade DCIS (IHC 0.86 vs H&E 0.40), while Xception showed the largest IDC sensitivity gain (IHC 0.63 vs H&E 0.14) (**Supplementary Table S4**). Precision (**Fig. 3B**) similarly favored IHC over H&E for ADH/low-grade (IHC 0.67–1.00 vs H&E 0.25–0.33), high-grade DCIS (IHC 0.75–0.86 vs H&E 0.46–0.50), and IDC (IHC 0.67–0.71 vs H&E 0.25–0.50), although H&E Xception achieved the highest normal/benign precision (0.88), reflecting a trade-off between specificity for normal tissue and sensitivity for abnormal classes (**Supplementary Table S4**). Together, these results indicate that TRPV4 IHC improves discrimination overall, with the greatest benefit for diagnostically challenging categories.

Error Pattern Analysis

To characterize model failure modes, we examined tile-level confusion matrices for all four modality–architecture combinations on the external GWU test cohort (**Supplementary Fig. S2**). Across all models, most errors reflected confusion between adjacent diagnostic categories along the DCIS progression axis. Adjacent-grade errors accounted for 68.3% (1359/1991) and 79.0% (1538/1947) of misclassified tiles for the H&E models (**Supplementary Fig. S2A,C**), and 72.8% (1031/1417) and 71.0% (875/1232) for the TRPV4 IHC models (**Supplementary Fig. S2B,D**). Overall, model

errors predominantly occurred between neighboring diagnostic groups, consistent with biologically plausible decision boundaries.

Discussion

In this multi-institutional deep learning study focused on the DCIS–IDC progression spectrum, we asked whether TRPV4 IHC, capturing both protein distribution patterns linked to mechanotransduction and tissue morphology, can provide diagnostically useful information beyond conventional H&E morphology alone. Using patient-level cross-validation on an internal development cohort and a fully held-out external test set from a different institution and scanner, we found that CNNs trained on TRPV4 IHC tiles achieved higher patient-level macro-AUC and macro-F1 than H&E models across both architectures. On the external GWU cohort, these gains were most evident at the patient level (**Table 1**; **Fig. 2B**), consistent with clinical workflows in which diagnoses integrate information across multiple regions rather than isolated fields. Per-class analyses further showed that TRPV4 IHC provided its largest discrimination improvements for ADH/low-grade DCIS (AUC +0.24–0.33) and for IDC (AUC +0.08–0.24) **(Fig. 3C**; **Supplementary Table S4**), two categories that are clinically consequential and often challenging at the microscope. While sensitivity and precision varied by architecture, the consistent AUC improvements suggest that RPV4 IHC encodes a robust, transferable signal that improves discrimination across architectures.

.

Our results extend prior work (37) in two directions. First, most DCIS-related deep learning studies have relied primarily on morphologic features from H&E to separate

benign from malignant lesions or to distinguish DCIS from IDC (28-33). These H&E-only approaches can perform well but inherit the ambiguity of morphology-based grading, particularly at clinically important boundaries. Here, we evaluate a mechanistically motivated stain, TRPV4 IHC, whose membrane relocalization has been linked to crowding-induced mechanotransduction in confined ducts (37, 38, 41). The consistent improvements in discrimination, particularly in ADH/low-grade DCIS and IDC, support the hypothesis that TRPV4 captures information that is not readily accessible from H&E morphology alone. Second, by emphasizing patient-level evaluation, we align more directly with diagnostic practice and show that tile-level information can translate into stronger case-level discrimination when aggregated per patient.

Architecture-dependent patterns provided additional insight into how different CNN backbones leverage TRPV4 signal. EfficientNet-B0 showed modest differences at the tile level but larger advantages after patient-level aggregation, suggesting that the TRPV4 signal may be most informative when integrated across multiple tissue foci. Xception demonstrated strong discrimination for specific classes in the external cohort, including improved IDC detection with IHC (**Fig. 3A–C**; **Supplementary Table S4**). Importantly, however, both architectures showed consistent patient-level AUC improvements with TRPV4 IHC, indicating that the mechanotransduction-encoded signal is not dependent on a single backbone choice and may generalize to other modern CNN families.

Our error analysis further supports the biological plausibility of the learned representations. Across all modality–architecture combinations, most misclassifications occurred between adjacent diagnostic categories along the DCIS progression axis (**Supplementary Fig. S2**), supporting biologically plausible decision boundaries. TRPV4 IHC models preserved this adjacency-weighted error structure while reducing overall misclassification, suggesting that improved performance reflects better separation along biologically plausible boundaries rather than spurious decision rules.

From a translational perspective, the observed performance patterns are consistent with potential applications in screening or triage. At the patient level, TRPV4 IHC improved discrimination in categories where under-calling has high clinical consequences, including ADH/low-grade DCIS and IDC (**Fig. 3C**; **Supplementary Table S4**). In practice, such models would be best positioned as decision-support tools that highlight suspicious foci or cases for closer review, rather than as stand-alone diagnostic systems. The architecture-dependent sensitivity/precision trade-offs also motivate ensembling or threshold tuning when optimizing for specific clinical objectives (e.g., prioritizing sensitivity for atypical or invasive disease).

This study has several limitations. The number of patients, particularly within ADH/LG-DCIS in the external cohort, was modest, leading to wide confidence intervals and limiting the precision of per-class estimates (**Supplementary Table S4**). Although we evaluated generalization across institutions and scanners, the external validation was performed at a single independent site, and additional multi-site testing will strengthen

confidence in broader deployment. Intermediate-grade DCIS and morphologic mimics were excluded to reduce label noise but remain clinically important and should be incorporated in future work. Finally, this study was not designed to assess prognostic value or treatment impact; prospective validation with outcome-linked endpoints will be essential before clinical translation.

Despite these limitations, our findings provide proof-of-concept that TRPV4 IHC paired with deep learning can enhance DCIS-focused classification in a multi-institutional setting. The largest gains in discrimination were observed in ADH/LG-DCIS and IDC, aligning with clinically difficult transitions and supporting the value of biologically grounded stains in the diagnostic “gray zones.” More broadly, these results suggest that mechanosensitive biomarkers such as TRPV4 may help bridge cellular mechanics and AI-assisted pathology, offering a path toward more interpretable and biologically informed diagnostic models.

Competing interests

I. Chung is an inventor on U.S. Patent No. 12,013,398 related to plasma-membrane relocalization based diagnostics. The patent is assigned to The George Washington University, is not licensed, and no commercial products are currently in development. No authors have received or will receive financial compensation related to this patent. The authors declare no other competing interests.

Funding

This work was supported by the Elsa Pardee Foundation Award, the George Washington University Technology Maturation Award, the George Washington Cancer Center, and the Katzen Cancer Research Pilot Award. The funders had no role in study design; data collection, analysis, or interpretation; or manuscript preparation.

Data availability

De-identified data and analysis code will be provided by the corresponding author upon reasonable request, subject to institutional policies and data use agreements.

Authors' contributions

J. Yoo: implemented the initial Xception model; wrote the initial prototype code to train the model on H&E versus IHC images; curated the image data. R. Karthikeyan: used Xception as a backbone and implemented EfficientNet-B0 in the pipeline; curated the image data. K. Kamat: data curation; coded the image input process; image summarization. C. Chan: data curation; image summarization. S. Samankan: ROI pathology annotation. E. Arbzadeh: ROI pathology annotation. A. Schwartz: ROI pathology annotation. P. Latham: identification of additional GWU cases. I. Chung: conceptualization; study design; methodology; final model development and optimization; formal analysis; supervision; project administration; funding acquisition; writing (original draft); writing (review and editing). All authors read and approved the final manuscript.

Acknowledgements

Tissue samples were provided by the Cooperative Human Tissue Network (CHTN), which is funded by the National Cancer Institute; other investigators may have received specimens from the same subjects. We thank Rebecca Blackwell (CHTN, University of Virginia) for case identification; Dr. Christopher Moskaluk (CHTN, University of Virginia) for histologic review; Dr. Patcharin Pramoonjago and Barushi Amarasinghe (Biorepository and Tissue Research Facility, University of Virginia) and Dr. Johnson Wang (VitroVivo Biotech, Inc, Rockville, MD) for tissue staining; and Ken Schill (DigitalScope) for assistance with IHC image data management.

References

1. Bocker W. Preneoplasia of the breast. Verh Dtsch Ges Pathol. 1997;81:502-13.

2. Gudjonsson T, Adriance MC, Sternlicht MD, et al. Myoepithelial cells: Their origin and function in breast morphogenesis and neoplasia. J Mammary Gland Biol. 2005;10(3):261-72.

3. Erbas B, Provenzano E, Armes J, et al. The natural history of ductal carcinoma in situ of the breast: a review. Breast Cancer Res Treat. 2006;97(2):135-44.

4. Cowell CF, Weigelt B, Sakr RA, et al. Progression from ductal carcinoma in situ to invasive breast cancer: revisited. Mol Oncol. 2013;7(5):859-69.

5. van Seijen M, Lips EH, Thompson AM, et al. Ductal carcinoma in situ: to treat or not to treat, that is the question. Br J Cancer. 2019;121(4):285-92.

6. Kumar AS, Bhatia V, Henderson IC. Overdiagnosis and overtreatment of breast cancer: rates of ductal carcinoma in situ: a US perspective. Breast Cancer Res. 2005;7(6):271-5.

7. Baxter NN, Virnig BA, Durham SB, et al. Trends in the treatment of ductal carcinoma in situ of the breast. J Natl Cancer Inst. 2004;96(6):443-8.

8. Srivastava S, Koay EJ, Borowsky AD, et al. Cancer overdiagnosis: a biological challenge and clinical dilemma. Nat Rev Cancer. 2019;19(6):349-58.

9. Ryser MD, Horton JK, Hwang ES. How Low Can We Go-and Should We? Risk Reduction for Minimal-Volume DCIS. Ann Surg Oncol. 2018;25(2):354-5.

10. Ryser MD, Weaver DL, Zhao F, et al. Cancer Outcomes in DCIS Patients Without Locoregional Treatment. J Natl Cancer Inst. 2019;111(9):952-60.

11. Hanna WM, Parra-Herran C, Lu FI, et al. Ductal carcinoma in situ of the breast: an update for the pathologist in the era of individualized risk assessment and tailored therapies. Mod Pathol. 2019;32(7):896-915.

12. Worni M, Akushevich I, Greenup R, et al. Trends in Treatment Patterns and Outcomes for Ductal Carcinoma In Situ. J Natl Cancer Inst. 2015;107(12):djv263.

13. Tevis SE, Bassett R, Bedrosian I, et al. OncotypeDX Recurrence Score Does Not Predict Nodal Burden in Clinically Node Negative Breast Cancer Patients. Ann Surg Oncol. 2019;26(3):815-20.

14. Dinan MA, Wilson LE, Reed SD, et al. Association of 21-Gene Assay (OncotypeDX) Testing and Receipt of Chemotherapy in the Medicare Breast Cancer Patient Population Following Initial Adoption. Clin Breast Cancer. 2020;20(6):487-94 e1.

15. Hwang S, Partridge A, Thompson A, et al. Comparing Two Treatment Approaches for Women with Ductal Carcinoma In Situ (DCIS) - The PORTAL Study. PCORI Final Research Reports. Washington (DC)2021.

16. Shehata M, Grimm L, Ballantyne N, et al. Ductal Carcinoma in Situ: Current Concepts in Biology, Imaging, and Treatment. J Breast Imaging. 2019;1(3):166-76.

17. Bomeisl P, Gilmore H. Spectrum of atypical ductal hyperplasia (ADH) and ductal carcinoma in-situ (DCIS): Diagnostic challenges. Semin Diagn Pathol. 2024;41(6):252-7.

18. Allred DC. Ductal carcinoma in situ: terminology, classification, and natural history. J Natl Cancer Inst Monogr. 2010;2010(41):134-8.

19. Bane A. Ductal carcinoma in situ: what the pathologist needs to know and why. Int J Breast Cancer. 2013;2013:914053.

20. Balleine RL, Webster LR, Davis S, et al. Molecular grading of ductal carcinoma in situ of the breast. Clin Cancer Res. 2008;14(24):8244-52.

21. Onega T, Weaver DL, Frederick PD, et al. The diagnostic challenge of low-grade ductal carcinoma in situ. Eur J Cancer. 2017;80:39-47.

22. Dano H, Altinay S, Arnould L, et al. Interobserver variability in upfront dichotomous histopathological assessment of ductal carcinoma in situ of the breast: the DCISion study. Mod Pathol. 2020;33(3):354-66.

23. Tsuda H, Yoshida M, Akiyama F, et al. Nuclear grade and comedo necrosis of ductal carcinoma in situ as histopathological eligible criteria for the Japan Clinical Oncology Group 1505 trial: an interobserver agreement study. Jpn J Clin Oncol. 2021;51(3):434-43.

24. van Seijen M, Jozwiak K, Pinder SE, et al. Variability in grading of ductal carcinoma in situ among an international group of pathologists. J Pathol Clin Res. 2021;7(3):233-42.

25. Elmore JG, Longton GM, Carney PA, et al. Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens. JAMA. 2015;313(11):1122-32.

26. Jain RK, Mehta R, Dimitrov R, et al. Atypical ductal hyperplasia: interobserver and intraobserver variability. Modern Pathol. 2011;24(7):917-23.

27. Van Bockstal MR, Berlière M, Duhoux FP, et al. Interobserver Variability in Ductal Carcinoma In Situ of the Breast. American Journal of Clinical Pathology. 2020;154(5):596-609.

28. Doyle S, Lips EH, Marcus E, et al. Deep learning for predicting invasive recurrence of ductal carcinoma in situ: leveraging histopathology images and clinical features. EBioMedicine. 2025;116:105750.

29. Wetstein SC, Stathonikos N, Pluim JPW, et al. Deep learning-based grading of ductal carcinoma in situ in breast histopathology images. Laboratory Investigation. 2021;101(4):525-33.

30. Huang CY, Chang RF, Lin CY, et al. Deep-learning model to improve histological grading and predict upstaging of atypical ductal hyperplasia / ductal carcinoma on breast biopsy. Histopathology. 2024;84(6):983-1002.

31. Hameed Z, Garcia-Zapirain B, Aguirre JJ, et al. Multiclass classification of breast cancer histopathology images using multilevel features of deep convolutional neural network. Sci Rep-Uk. 2022;12(1).

32. Tahir M, Hu Y, Kumar H, et al. A Comprehensive AI-Based Approach in Classifying Breast Lesions: Focusing on Improving Pathologists' Accuracy and Efficiency. Clinical Breast Cancer. 2025;25(6):e818-e25.

33. Hayward MK, Weaver VM. Improving DCIS diagnosis and predictive outcome by applying artificial intelligence. Biochim Biophys Acta Rev Cancer. 2021;1876(1):188555.

34. White JP, Cibelli M, Urban L, et al. TRPV4: Molecular Conductor of a Diverse Orchestra. Physiol Rev. 2016;96(3):911-73.

35. Rosenbaum T, Benitez-Angeles M, Sanchez-Hernandez R, et al. TRPV4: A Physio and Pathophysiologically Significant Ion Channel. Int J Mol Sci. 2020;21(11).

36. Michalick L, Kuebler WM. TRPV4-A Missing Link Between Mechanosensation and Immunity. Front Immunol. 2020;11:413.

37. Bu X, Ashby N, Vitali T, et al. Cell crowding activates pro-invasive mechanotransduction pathway in high-grade DCIS via TRPV4 inhibition and cell volume reduction. eLife. 2025:13:RP100490.

38. Chan C, Arbzadeh E, Samankan S, et al. Plasma Membrane TRPV4 Relocalization as a Mechanosensitive Prognostic Biomarker in Ductal Carcinoma In Situ. medRxiv. 2025:2025.11.04.25339245.

39. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. Proc Cvpr Ieee. 2017:1800-7.

40. Tan MX, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Pr Mach Learn Res. 2019;97.

41. Hua R, Jiang JX. How cell crowding causes cancer cells to spread. eLife. 2025;14:e106768.

Architecture	Modality	Tile-level Macro-F1	Tile-level Macro-AUC	Patient-level Macro-F1	Patient-level Macro-AUC
EfficientNet-B0	H&E	0.47	0.78	0.43	0.80
	TRPV4 IHC	0.53	0.84	0.72	0.91
	Absolute improvement	+0.06	+0.06	+0.29	+0.11
	Relative improvement	+12.8%	+7.7%	+67.4%	+13.8%
Xception	H&E	0.51	0.80	0.44	0.73
	TRPV4 IHC	0.60	0.85	0.68	0.92
	Absolute improvement	+0.09	+0.05	+0.24	+0.19
	Relative improvement	+17.6%	+6.3%	+54.5%	+26.0%

Table 1. External test performance on the GWU cohort.

External test performance comparing H&E and TRPV4 IHC ensemble models evaluated on the held-out GWU cohort without retraining. Each ensemble comprises 3- fold-specific models trained on the UVA development cohort (patient-level 3-fold cross-validation), with predictions averaged across folds. Tile-level macro-F1 and one-vs-rest

macro-AUC summarize tile-level discrimination, and patient-level metrics are computed after aggregating tile probability outputs within each patient. Absolute and relative improvements quantify the performance gain of TRPV4 IHC over H&E for each architecture, with the largest gains observed at the patient-level macro-F1 relative improvement 54.5–67.4%.

Model	**Class**	**Sensitivity**	**Specificity**	**Precision**	**F1-score [95% CI]**	**AUC [95% CI]**	**n tiles**
H&E Efficient Net-B0	Normal/ benign	0.88	0.78	0.66	0.75 [0.74–0.77]	0.92 [0.91–0.93]	1,454
	ADH/LG-DCIS	0.48	0.81	0.23	0.31 [0.28–0.34]	0.70 [0.68–0.74]	474
	HG-DCIS	0.50	0.89	0.75	0.60 [0.58–0.62]	0.85 [0.84–0.86]	1,786
	IDC	0.15	0.93	0.31	0.20 [0.17–0.23]	0.65 [0.63–0.67]	797
IHC Efficient Net-B0	Normal/ benign	0.75	0.86	0.61	0.67 [0.65–0.70]	0.91 [0.90–0.92]	770
	ADH/LG-DCIS	0.73	0.78	0.36	**0.49** [0.45–0.52]	**0.84** [0.82–0.86]	495
	HG-DCIS	0.55	0.87	0.77	**0.65** [0.62–0.67]	0.85 [0.84–0.87]	1,480
	IDC	0.25	0.94	0.46	**0.33** [0.29–0.37]	**0.74** [0.72–0.76]	577
H&E Xception	Normal/ benign	0.58	0.94	0.83	0.68 [0.66–	0.91 [0.90–	1,454

					0.71]	0.92]	
	ADH/LG-DCIS	0.58	0.88	0.35	0.43 [0.40–0.47]	0.81 [0.79–0.84]	474
	HG-DCIS	0.69	0.76	0.65	0.67 [0.65–0.69]	0.82 [0.81–0.84]	1,786
	IDC	0.27	0.84	0.26	0.26 [0.24–0.29]	0.66 [0.65–0.68]	797
IHC Xception	Normal/ benign	0.67	0.92	0.71	**0.69** [0.66–0.72]	**0.92** [0.91–0.93]	770
	ADH/LG-DCIS	0.65	0.88	0.49	**0.56** [0.52–0.59]	**0.83** [0.81–0.86]	495
	HG-DCIS	0.65	0.86	0.79	**0.71** [0.69–0.73]	**0.86** [0.85–0.88]	1,480
	IDC	0.52	0.84	0.41	**0.46** [0.42–0.49]	**0.79** [0.76–0.80]	577

Table 2. Tile-level class-wise performance on the GWU external test cohort

Per-class sensitivity, specificity, precision, F1 score, and one-vs-rest AUC are shown for ensemble models trained on the UVA cohort and evaluated on GWU without retraining. F1 score and AUC are reported with 95% confidence intervals estimated by bootstrap resampling (n=500). The number of evaluated tiles per class is shown for each modality.

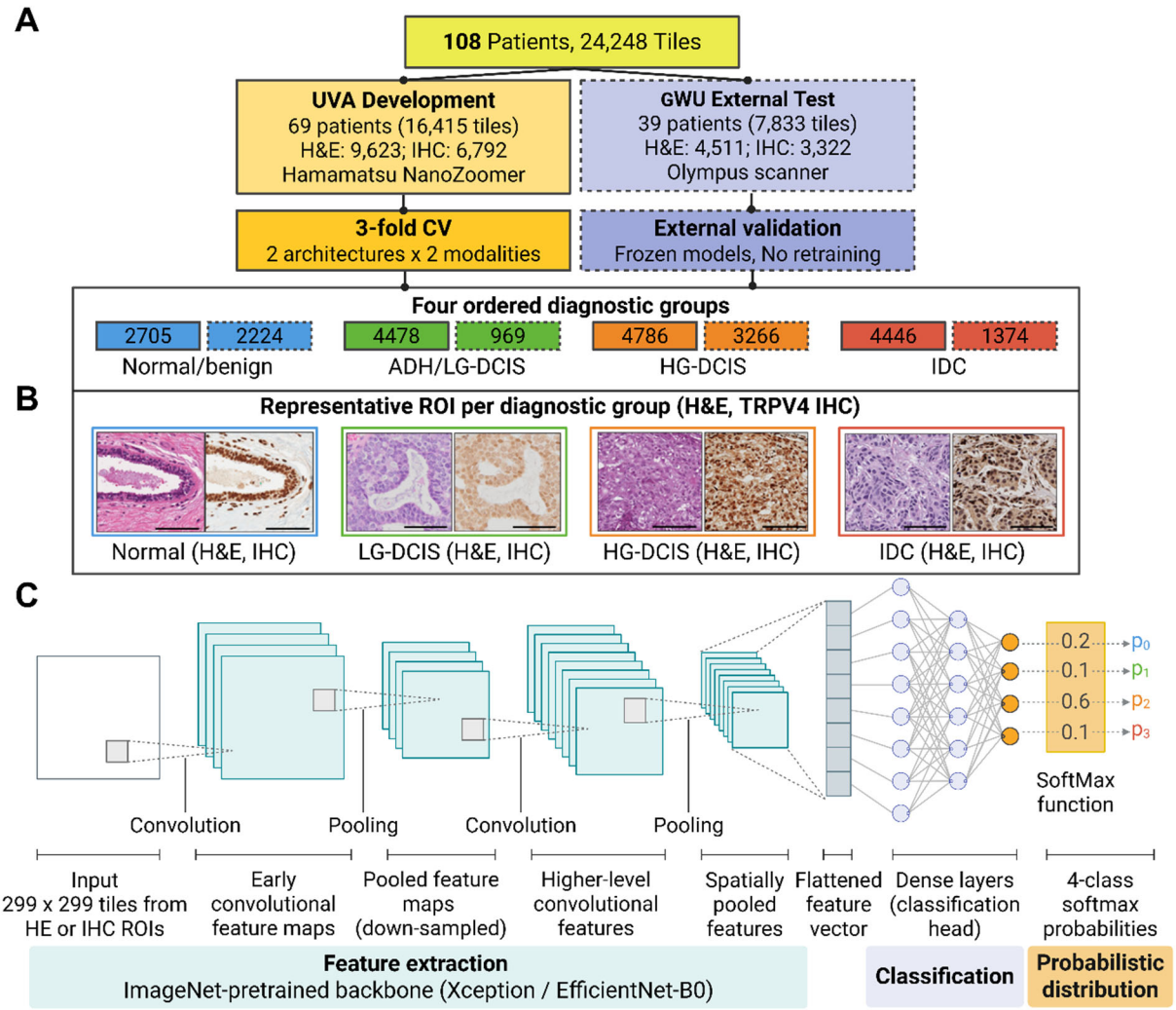

Figure 1. Study design and deep learning architecture.

Overview of the multi-institutional dataset and deep learning pipeline for TRPV4 IHC-enhanced DCIS classification. **(A)** Patient and tile distribution across UVA development (n=69 patients) and GWU external test (n=39 patients) cohorts, stratified by stain modality (H&E, TRPV4 IHC) and diagnostic category. The four ordered diagnostic groups reflect the DCIS progression spectrum: normal/benign (blue), ADH/low-grade DCIS (orange), high-grade DCIS (green), and invasive ductal carcinoma (IDC, red). **(B)** Representative ROIs for each diagnostic group in H&E and TRPV4 IHC, illustrating typical morphology and staining patterns from which tiles were sampled. Scale bars=100 μm. **(C)** The

schematic is adapted from “Convolutional Neural Networks (CNNs)” template by Martina Maritan, created with BioRender.com. We used two ImageNet-pretrained CNN architectures (Xception and EfficientNet-B0). UVA tiles were partitioned using patient-level 3-fold cross-validation to train separate models for each stain modality and backbone architecture. Each fold comprised tiles from distinct patients, ensuring no patient appeared in multiple folds. We evaluated ensemble models (averaging predictions from the three fold-specific models per modality and architecture) on an independent external test set scanned on a different whole-slide imaging platform (Olympus at GWU vs Hamamatsu at UVA), without retraining, to assess cross-platform generalizability.

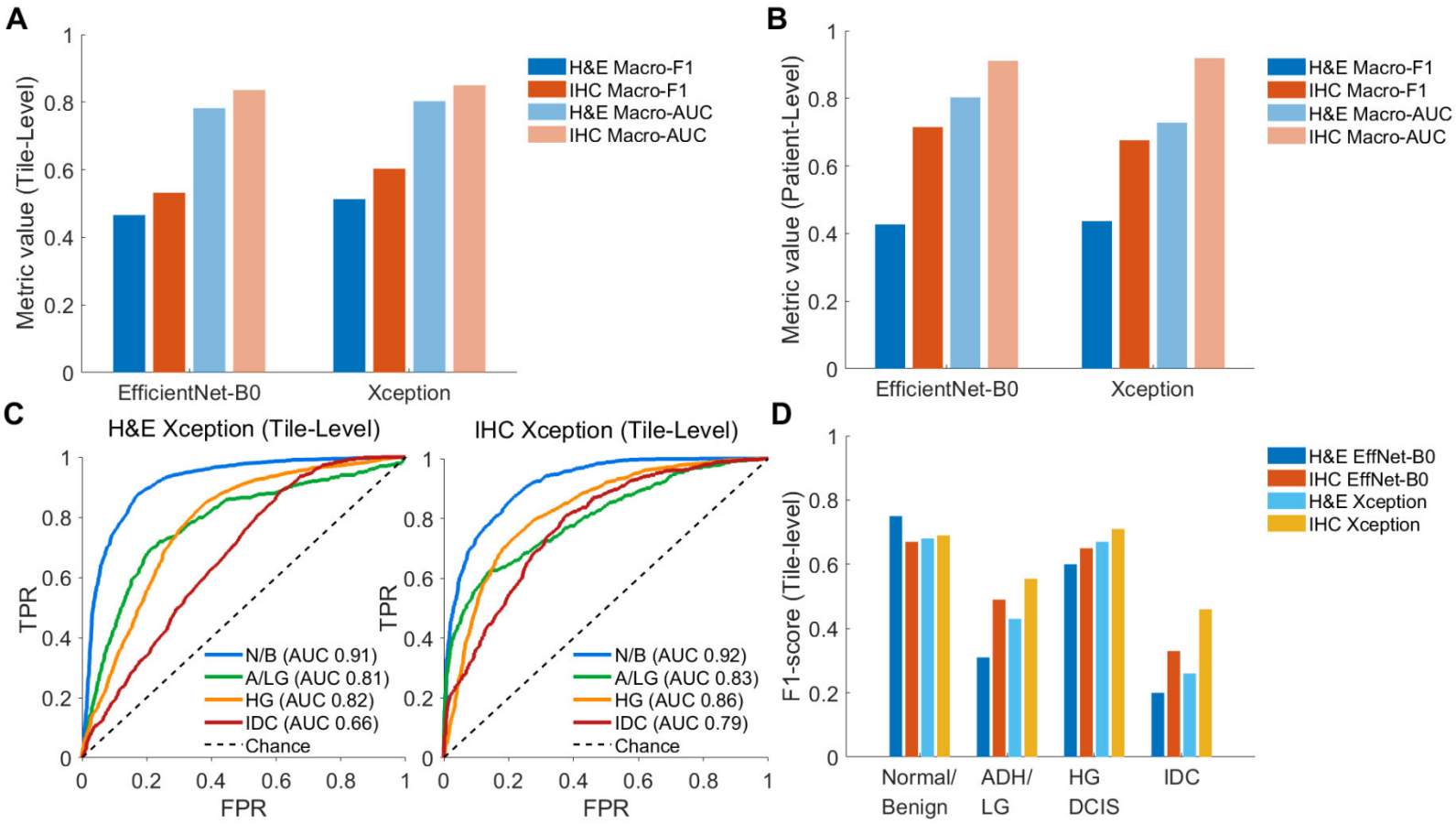

Figure 2. External test performance on the GWU cohort.

The GWU external test cohort comprised 39 unique patients; H&E was available for 31 patients (4,511 tiles) and TRPV4 IHC for 33 patients (3,322 tiles), totaling 7,833 tiles after excluding intermediate-grade cases. **(A)** Tile-level macro-F1 and macro-AUC for H&E and TRPV4 IHC ensemble models across EfficientNet-B0 and Xception architectures (**Table 1**). **(B)** Patient-level macro-F1 and macro-AUC after aggregating tile probabilities within each patient, showing larger performance gains for TRPV4 IHC than for H&E (**Table 1**). **(C)** Tile-level one-vs-rest ROC curves for the Xception ensembles (left: H&E; right: TRPV4 IHC), illustrating improved discrimination for HG-DCIS and IDC and a modest gain for ADH/low-grade DCIS (**Table 2**). Xception is shown because it achieved

marginally higher tile-level macro-F1 than EfficientNet-B0 on GWU (**Table 1**). **(D)** Tile-level per-class F1 scores for all model–modality combinations, highlighting the largest improvements for ADH/low-grade DCIS and IDC (**Table 2**).

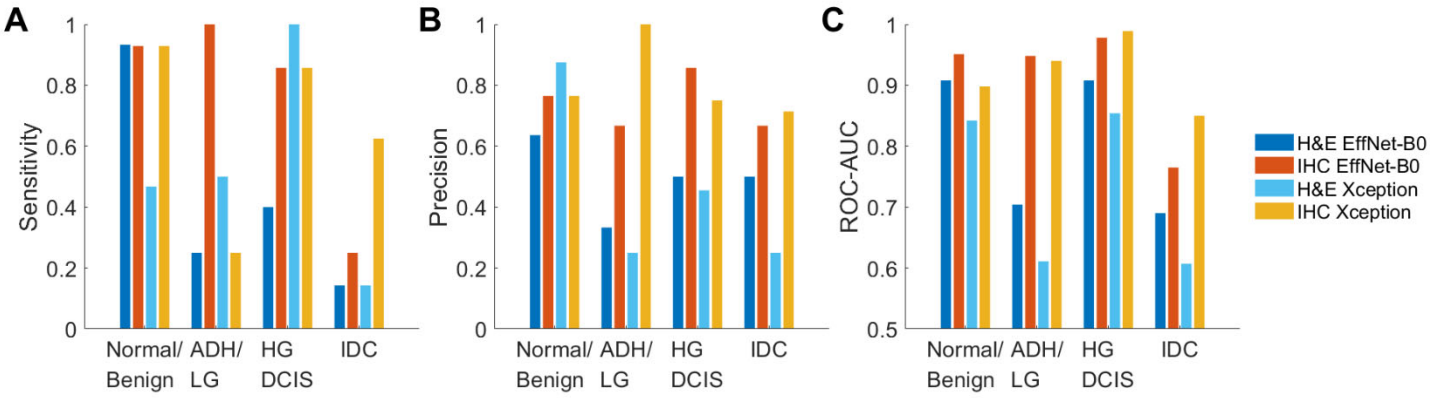

Figure 3. Patient-level per-class performance comparison on the GWU external test cohort.

Patient-level sensitivity **(A)**, precision **(B)**, and one-vs-rest ROC-AUC **(C)** bpay diagnostic category for all four model–modality combinations (H&E EfficientNet-B0, TRPV4 IHC EfficientNet-B0, H&E Xception, IHC Xception). The GWU cohort comprised 39 patients (H&E n=31; TRPV4 IHC n=33). IHC showed higher AUC than H&E across categories, with the largest gains for ADH/low-grade DCIS (IHC 0.94–0.95 vs H&E 0.61–0.70) and IDC (IHC 0.77–0.85 vs H&E 0.61–0.69). Exact values and 95% confidence intervals are provided in **Supplementary Table S4**.