RESEARCH

Open Access

Check for updates

Predicting Nottingham grade in breast cancer digital pathology using a foundation model

Jun Seo Kim^{1†}, Jeong Hoon Lee^{2†}, Yousung Yeon¹, Doyeon An¹, Seok Jun Kim¹, Myung-Giun Noh^{3*} and Suehyun Lee^{1*}

Abstract

Background The Nottingham histologic grade is crucial for assessing severity and predicting prognosis in breast cancer, a prevalent cancer worldwide. Traditional grading systems rely on subjective expert judgment and require extensive pathological expertise, are time-consuming, and often lead to inter-observer variability.

Methods To address these limitations, we develop an AI-based model to predict Nottingham grade from wholeslide images of hematoxylin and eosin (H&E)-stained breast cancer tissue using a pathology foundation model. From TCGA database, we trained and evaluated using 521 H&E breast cancer slide images with available Nottingham scores through internal split validation, and further validated its clinical utility using an additional set of 597 cases without Nottingham scores. The model leveraged deep features extracted from a pathology foundation model (UNI) and incorporated 14 distinct multiple instance learning (MIL) algorithms.

Results The best-performing model achieved an F1 score of 0.731 and a multiclass average AUC of 0.835. The top 300 genes correlated with model predictions were significantly enriched in pathways related to cell division and chromosome segregation, supporting the model's biological relevance. The predicted grades demonstrated statistically significant association with 5-year overall survival (p < 0.05).

Conclusion Our AI-based automated Nottingham grading system provides an efficient and reproducible tool for breast cancer assessment, offering potential for standardization of histologic grade in clinical practice.

Keywords Biological processes, Breast cancer, Gene expression data, Gene ontology, Multiple instance learning, Nottingham grade, TCGA

[†]Jun Seo Kim and Jeong Hoon Lee contributed equally to this work.

*Correspondence: Myung-Giun Noh md.mgnoh@gmail.com Suehyun Lee leesh@gachon.ac.kr ¹Department of Computer Engineering, Gachon University, Seongnam 13120, South Korea ²Department of Radiology, Stanford University School of Medicine, Stanford, CA 94305, USA ³Department of Pathology, School of Medicine, Ajou University, Suwon 16499, South Korea

Background

Breast cancer is now the most diagnosed cancer globally, accounting for one-eighth of all cancer cases [1]. According to GLOBOCAN 2022, the most recent global cancer statistics available, it remains the most commonly diagnosed cancer in women, with 2.3 million new cases and over 666,000 deaths worldwide, with projections indicating the overall cancer burden will increase by 77% by 2050 [2]. This emphasizes the importance of the early detection and effective treatment of breast cancer.

The Nottingham histologic grade (Elston-Ellis modification of the Scarff-Bloom-Richardson grading system)



© The Author(s) 2025, corrected publication 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://cre ativecommons.org/licenses/by-nc-nd/4.0/. is a key prognostic classification used to evaluate breast cancer [3]. It quantifies the severity of cancer based on the tumor's tubularity, nuclear pleomorphism, and mitotic count [4, 5]. It is a grading system recommended by various international professional organizations such as the World Health Organization (WHO), the American Joint Committee on Cancer (AJCC), the European Union (EU), and the Royal College of Pathologists (UK RCPath) [6]. Despite the common use of the Nottingham Grading System (NGS), the process requires expert knowledge, is time-consuming, and can be subject to inter- and intraobserver variability [7]. These issues highlight the need for approaches that are more automated and objective.

Advances in artificial intelligence (AI) and machine learning have greatly improved digital image analysis in pathology. Deep learning algorithms now match or exceed expert pathologists in analyzing histopathological slides and automating tasks like lymph node metastasis detection and Ki67 scoring in breast cancer, improving diagnostic reproducibility [8].

In particular, the field of histopathology, image analysis creates the possibility of improving the accuracy and efficiency of diagnosis owing to the emergence of foundational models, such as Lunit DINO and UNI [9–13]. Wang et al. first proposed the DeepGrade model to predict the Nottingham grade using deep learning, which was developed specifically to improve prognostic power by reclassifying patients in the intermediate-risk group (grade 2) [14]. Jaroensri et al. used a deep learning system to predict the pathological grade components of mitotic count, nuclear polymorphisms, and tubular body formation separately, with high agreement for each component [15]. Wetstein et al. focused on distinguishing low/ intermediate grades (grades 1 and 2) from high grades (grade 3) using a deep learning model based on multipleinstance learning (MIL) and ResNet-34 [16]. Sharma et al. focused on predicting Nottingham grades, specifically grades 1 and 3, using deep learning [17]. Although attempts have been made to predict Nottingham scores using AI-based tools, these studies were limited in their ability to effectively distinguish between grades 1 and 2. This indicates that there is still room for improvement in predicting Nottingham scores using existing approaches, particularly the lack of sophisticated predictions for clearly distinguishing between grades 1 and 2. Therefore, there remains a need to improve the prediction of Nottingham scores using AI-based tools.

In this study, we aimed to develop an AI model that can accurately predict the Nottingham grade from wholeslide images of hematoxylin and eosin (H&E)-stained breast cancer tissues. We introduce a novel methodology that combines MIL and self-supervised learning (SSL) to propose a model that can predict all Nottingham grades (grades 1, 2, and 3) in a unified manner. In particular, unlike previous studies, we attempted to clearly distinguish the boundaries between the grades and compared the UNI-foundation model with the ResNet-18 model to improve the prediction accuracy for each grade. Based on the predicted grades, we investigated the clinical utility of the automated cancer grading system through a review by an expert pathologist. We validated how AI-assisted histologic grade prediction correlates with patient survival outcomes. Furthermore, we analyzed the features of genomic data based on this histological prediction model using a multiomics approach.

Methods

Data source and patient selection

We used The Cancer Genome Atlas Program (TCGA) breast invasive carcinoma (BRCA) dataset of 1050 patients with invasive breast cancer from TCGA through the Genomic Data Commons Data Portal [18–20]. After quality review of digitized whole-slide images (WSIs) for female breast cancer patients, a total of 1,118 H&E-stained diagnostic slides from 1,050 patients were selected for the final analysis.

Male breast cancer cases (13 WSIs from 12 patients) and slides with unrecognizable tumor regions (2 WSIs from 2 patients) were excluded from the analysis. As one patient may have more than one WSI, the number of patients remained unchanged despite the exclusion of slides.

Figure 1 A illustrates the inclusion and exclusion criteria applied to the initial TCGA-BRCA dataset, as well as the composition of the final cohorts used for analysis. The clinical data for each patient (such as age at diagnosis, survival status, and follow-up duration) were obtained from Firehose Legacy, Pan-Cancer Atlas in Genomic Data Commons, and cBioPortal [21, 22]. There were 521 slides from patients with a recorded Nottingham histologic grade and 597 slides from patients without a recorded grade. Additionally, we selected 646 patients for whom RNA-seq expression data were available.

To assess the generalizability of our model beyond the TCGA-BRCA dataset, we employed an external validation cohort derived from the publicly available BReAst Carcinoma Subtyping (BRACS) dataset [23]. BRACS consists of a large collection of H&E-stained WSIs annotated for various breast lesion types and was developed through a collaborative effort among IRCCS Fondazione Pascale, Institute for High Performance Computing and Networking (ICAR)- National Research Council (CNR), and International Business Machines (IBM) Research Zurich. The dataset includes representative breast tissue samples from multiple diagnostic categories, such as benign lesions, atypical hyperplasia, ductal carcinoma in situ (DCIS), and invasive carcinoma (IC). For our study, we selected only the WSIs classified as IC, as



Fig. 1 Flowchart of patient and slide selection from the TCGA-BRCA and BRACS dataset. (A) TCGA-BRCA cohort selection for model development. (B) BRACS cohort selection for external validation

they are most relevant to the prediction of Nottingham histologic grade. From the IC-labeled subset, 132 WSIs were initially identified. After quality assessment by an expert pathologist, 4 WSIs were excluded due to poor slide quality, resulting in a final external validation set of 128 WSIs. These slides were used to evaluate the model's performance in predicting Nottingham grade in an independent cohort. Figure 1B illustrates the data selection process for the BRACS cohort.

Collection method for Nottingham histological grade in TCGA-BRCA and BRACS dataset

Asaoka et al. (2020) reviewed pathology reports from the TCGA-BRCA project using the Text Information Extraction System cancer research network [24]. Pathology reports of 1,046 H&E-stained tissue sections were analyzed to obtain Lymphovascular invasion (LVI) status and Nottingham histology scores [25]. Some histologic scores were obtained by utilizing histologic scores recorded in the literature, and those that were not included were obtained by manually reviewing pathology reports, resulting in a total of 521 histologic scores. A total of 521 slides with recorded Nottingham grades were used to build the deep learning model. Each image was processed as an independent sample and split into training, validation, and evaluation sets at a precise 6:2:2 ratio. Data partitioning was performed using stratified random sampling to maintain evenly distributed characteristics. In the BRACS dataset, Nottingham histologic grades were not originally provided. Therefore, all selected WSIs in the external validation cohort were independently reviewed and graded by an expert pathologist according to the standard Elston-Ellis modification of the Scarff-Bloom-Richardson system. Final histologic scores for both datasets are provided in Supplementary material, Tables S1 (TCGA-BRCA) and S2 (BRACS).

Histopathology image preprocessing and feature extraction

In the feature extraction phase, we used the UNI-foundation model, which is based on a ViT-L/16 architecture and trained by DINOv2 (Fig. 2B) [26, 27]. This model extracts 1,024 dimensional features from each patch. Additionally, for a performance comparison, we used the ImageNet-pretrained ResNet-18 model [28], which extracts 512 dimensional features for each patch.

For histopathological image segmentation, we used the Clustering-constrained Attention Multiple Instance Learning (CLAM) model, following parameter settings from previous studies, to extract tissue-containing tiles while excluding background and noise [29]. Tissue segmentation was performed using thresholding in the Hue, Saturation, Value (HSV) color space, and non-tissue or low-quality patches were removed, which also helps minimize potential batch effects by reducing slide-level variability. Extracted patches of a fixed size (224×224 pixels) were stored as HDF5 files for each patient to ensure efficient data management and processing for deep learning model training, as shown in Fig. 2A.

Development of a Nottingham score prediction model using multiple instance learning methods

We utilized the Attention-Challenging Multiple Instance Learning (ACMIL) method to train and build the model. As illustrated in Fig. 2C, this methodology integrates two main techniques, multi-branch learning and stochastic top-K instance masking, which focus on suppressing the concentration of attention and mitigating overfitting [30].



(a) Histopathology image preprocessing



(c) Complex analysis and result visualization using multiple instance learning

Fig. 2 Comprehensive Workflow for Histopathological Analysis. (A) Displays segmentation of histopathological images using the CLAM model, isolating tissue-only tiles and dividing them into 224×224 pixel patches. (B) The UNI model with DINO, pretrained using self-supervised learning (SSL), extracts 1,024-dimensional features at the patch level, which are subsequently aggregated within the MIL framework using attention-based selection. (C) The model predicts the Nottingham grade using a combination of multi-branch learning, stochastic top-K instance masking, and attention mechanisms. The results are visualized using survival analysis, heatmap visualization, and gene ontology analysis

These techniques are designed to address limitations in conventional MIL models, which often focus attention excessively on a small subset of instances, potentially leading to biased learning. The model was primarily trained using the standard multi-class cross-entropy loss function:

$$\mathcal{L}_b = -\sum_{i=1}^{C} y_i \log\left(p_i\right) \tag{1}$$

where y_i is the one-hot encoded ground truth label and p_i is the predicted probability for class *i*, obtained via softmax. To further enhance the robustness and generalization of the model, ACMIL incorporates two additional loss components derived from its multi-branch architecture. The first is the semantic regularization loss \mathcal{L}_p , which ensures consistent classification across multiple attention branches and is defined as:

$$\mathcal{L}_p = -\frac{1}{M} \sum_{i=1}^{M} y \log \widehat{y}_i + (1-y) \log (1-\widehat{y}_i)$$
 (2)

where M denotes the number of attention branches and \hat{y}_i is the prediction from the *i*-th branch. The second is the diversity regularization loss \mathcal{L}_d , which promotes feature diversity between attention branches by minimizing the cosine similarity between attention vectors. This loss is defined as:

$$\mathcal{L}_d = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=j+1}^{M} \cos\left(\alpha_i, \alpha_j\right)$$
(3)

where α_i denotes the attention vector of the *i*-th branch. These three loss components are combined into the final composite loss function as follows:

$$L = \mathcal{L}_b + \mathcal{L}_p + \mathcal{L}_d \tag{4}$$

In addition to loss regularization, ACMIL employs stochastic top-K instance masking to prevent the model from over-relying on a fixed set of highly attended instances. In this approach, instances with the top-K attention scores are selected, and a portion of them is randomly masked based on a predefined probability. The remaining attention scores are then normalized to maintain a valid probability distribution, thereby improving the generalization ability of the model. The model does not apply a hard attention threshold to define key instances; instead, predictions are computed by aggregating all instance features weighted by their attention scores, allowing key instances to emerge implicitly. To validate the performance of ACMIL, we performed a performance comparison analysis with the existing MIL models CLAM-SB, CLAM-MB, Transformer based Correlated multiple instance learning (Trans-MIL), Dual-stream multiple instance learning (DSMIL), Double-tier feature distillation multiple instance learning

(DTFD-MIL), and Attention-based multiple instance learning (ABMIL) [29, 31–34].

All models were implemented based on the UNI model and ResNet-18 architectures. Training was conducted for 100 epochs using the AdamW optimizer with a weight decay of 0.00001. The initial learning rates were set to 0.0001 for UNI model and 0.0002 for ResNet-18, and a cosine learning rate schedule was applied to progressively decrease the learning rate over the training epochs. Due to the variable number of instances per bag in the MIL framework, the batch size was set to 1. During training, we randomly selected 70–100% of patch features to achieve an augmentation effect. All hyperparameters used in this study were adopted from prior MIL-based digital pathology studies without additional tuning.

Statistical analysis

Model performance was evaluated using Area Under the Receiver Operating Characteristic Curve (AUROC) and F1 scores calculated on the validation and test datasets. AUROC measures the model's ability to distinguish between Nottingham histologic grades, while the F1 score assesses accuracy and sensitivity. For the validation and testing phases, the model in the epoch with the highest F1 score was selected and saved from the training results over 100 epochs.

To evaluate generalizability, external validation was performed using an independent cohort from the BRACS dataset, with AUROC and F1 scores similarly calculated. To further assess the agreement between model-predicted grades and pathologist-assigned grades, Cohen's kappa coefficients were calculated for each grade class in both TCGA-BRCA and BRACS datasets.

Survival analysis evaluated the impact of Nottingham grade on overall survival (OS) and the model's accuracy in predicting it, focusing on 5-year survival, a key metric in breast cancer prognosis [35]. We tested whether predicted survival differences for each grade were statistically significant. Survival time was defined as the time from diagnosis to death from any cause; patients with time from diagnosis to death recorded as zero or those with no follow-up time information (n = 4) were excluded from the analysis. Survival data were analyzed using the Kaplan-Meier method, and differences between survival curves were assessed using the log-rank test. Hazard ratios for each grade were estimated by univariate Cox proportional hazards regression analysis using the Nottingham histologic grade as a variable. Analyses were performed as survival analyses for patients with a Nottingham histologic grade recorded by the pathologist and survival analyses using the grade predicted by the model because the Nottingham histologic grade was not recorded.

For multivariate survival analysis, a Cox proportional hazards regression model was constructed incorporating key clinical covariates including tumor size (T1, T2, T3/T4), Estrogen receptor (ER), Progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) status, and age. Age was modeled both as a continuous and as a categorical variable using the median (\leq 58 years vs. >58 years). AI-predicted Nottingham grades were included as dummy variables with grade 1 as the reference group. Tumor stage was similarly encoded using T1 as the reference. Proportional hazard assumptions were tested, and L2 penalization was applied to stabilize coefficient estimates.

Heatmap visualization

Heatmap visualization was performed to highlight key areas of the tissue slide for each Nottingham grade, based on attention scores assigned to each patch. Attention scores were calculated using a neural network architecture with gating mechanisms, including tanh and sigmoid functions, to regulate the flow of information. The calculated attention scores were normalized using a softmax function to allow comparisons across all patches. The normalized attention values were visualized on a heat map, with areas of high attention represented by brighter colors.

Gene ontology analysis

We performed gene ontology (GO) analysis using RNAseq data from 646 TCGA patients, excluding the training cohort, including both patients with pathologist-assigned grades in the validation/test sets and those without such annotations, to explore gene expression patterns associated with model-predicted histologic grades. RNAseq data were normalized using the Trimmed Mean of M-values method to allow for comparisons between experiments, and log-transformed counts were estimated using the Voom method to make the data suitable for analysis. Genes with low expression were filtered using the count per million (CPM) method, which was applied only to genes with a CPM value >1 in more than 25% of the samples. This process resulted in 14,861 genes being selected for the final analysis. Additionally, for each Nottingham histologic grade 1-3 predicted by the model, we extracted the top 300 genes that were positively correlated and analyzed the genes associated with the predicted grade. We performed enrichment analysis on these gene lists using the DAVID web tool, focusing on Biological Processes ontology terms [36]. This assesses how well the model's predictions match the actual gene expression patterns.

Table 1	Overall	patient	characte	eristics	by I	Notting	ham	grad	е
presence	د								

Patient characteristics	With Nottingham Grade	Without Nottingham Grade	Total	
N	496	554	1050	
Age at biopsy				
Median (Min-Max)	58(27–90)	59(26–90)	58(26– 90)	
Tumor grade (n (%))				
grade 1(3–5 points)	67(14)	-	-	
grade 2(6–7 points)	229(46)	-	-	
grade 3(8–9 points)	200(40)	-	-	
Nuclear score (n (%))				
1	31(6)	-	-	
2	232(47)	-	-	
3	233(47)	-	-	
Tubular score (n (%))				
1	24(5)	-	-	
2	88(18)	-	-	
3	384(77)	-	-	
Mitotic score (n (%))				
1	218(44)	-	-	
2	130(26)	-	-	
3	148(30)	-	-	
Clinical stage (n (%))				
	106(21)	70(13)	176	
	288(58)	310(56)	598	
	96(19)	139(25)	235	
IV	5(1)	13(2)	18	
Unknown	1(1)	22(4)	23	
ER status (n (%))				
Positive	390(78)	381(69)	771	
Negative	103(21)	125(23)	228	
Indeterminate	1(0)	1(0)	2	
Unknown	2(1)	47(8)	49	
HER2 status (n (%))				
Positive	65(13)	92(17)	157	
Negative	263(53)	273(49)	536	
Equivocal	113(23)	60(11)	173	
Indeterminate	7(1)	4(1)	11	
Unknown	48(10)	125(22)	173	

The distribution of patient characteristics is stratified by the presence or absence of Nottingham grade. Included are demographic data (age at biopsy), tumor grade components (tumor grade, nuclear score, tubular score, and mitotic score), clinical staging, and biomarker status (ER and HER2). Age is expressed as median (range), while categorical variables are shown as counts and percentages

Results

Patient characteristics

The patient characteristics of the dataset of 521 slides with recorded Nottingham histologic.

grades (n = 496) and 597 slides without recorded Nottingham histologic grades (n = 554) are summarized in Table 1. The median age of patients with recorded Nottingham grades was 58 years (range 27–90), while those

Table 2	Distribution	of tumor	grades in	slides	with No	ottingham
grade ac	ross training,	validatio	n, and test	t sets		

Patient characteristics	Train	Validation	Test	<i>p</i> -value
N	314	103	104	
「umor grade (n (%))				1.00
grade 1(3–5 points)	42(13)	14(14)	14(14)	
grade 2(6–7 points)	145(47)	48(46)	48(46)	
grade 3(8–9 points)	127(40)	41(40)	42(40)	

The p-value, calculated using a chi-square test, shows no significant differences in tumor grade distributions across the datasets (p = 1.00)

without recorded grades had a median age of 59 years (range 26–90), with a significant age difference (p < 0.001). Regarding tumor grade distribution, among patients with recorded Nottingham grades, 14% had a low grade (grade 1), 46% had an intermediate grade (grade 2), and 40% had a high grade (grade 3). Regarding the observed clinical stage, when comparing the distribution of clinical stages between the groups of patients with and without documented Nottingham histologic grades, stage II was the most commonly observed stage in both groups, but there were some differences in the overall distribution. In the group of patients with recorded Nottingham stages, the intermediate stage (stage II) was dominant, with similar proportions of low (stage I) and high (stage III) stages. By contrast, in the unrecorded group, more stage III cases and fewer stage I cases were observed. ER and HER status also differed between the two groups, with ER positivity being 78% in the recorded group and 69% in the unrecorded group, and HER2 positivity being 13% in the recorded group and 17% in the unrecorded group. Detailed patient characteristics are presented in Table 1.

In total, 521 slides from 496 patients with recorded Nottingham histologic grades were used to construct the model. They showed the highest proportion of intermediate-grade tumors in the training set, with 145 cases (47%), whereas high-grade tumors represented 40% of the cases in each set. Similar patterns were observed in the validation and test sets, with no statistically significant differences observed across all the sets (p = 1.00). The detailed distribution of the tumor grades is shown in Table 2.

The distribution of tumor grade and its component scores in the BRACS external validation cohort is summarized in Table 3. Grade 2 tumors were the most common (44%), followed by grade 3 (36%) and grade 1 (20%). In terms of the component scores, the majority of cases had a tubular score of 3, a nuclear score of 2 or 3, and a mitotic score of 1.

Model performance comparison and selection

The performances of various MIL models were evaluated using TCGA-BRCA datasets. The evaluation, which mainly centered on the F1-score and AUC metrics,

 Table 3
 Patient characteristics in the BRACS external validation dataset

Patient characteristics	BRACS dataset	Total
Tumor grade (n (%))		128
grade 1(3–5 points)	26(20)	
grade 2(6–7 points)	56(44)	
grade 3(8–9 points)	46(36)	
Nuclear score (n (%))		128
1	14(10)	
2	57(45)	
3	57(45)	
Tubular score (n (%))		128
1	6(5)	
2	23(18)	
3	99(77)	
Mitotic score (n (%))		128
1	67(52)	
2	24(19)	
3	37(29)	

Summary of tumor grade and individual histologic component scores including nuclear, tubular, and mitotic scores for the 128 slides in the BRACS validation dataset

Table 4 Performance of MIL approaches using two pre-trained methods and metrics on TCGA-BRCA dataset

Pretrained	Performance	TCGA-BRCA				
Feature		Validation	n	Test		
Extractor		F1-score	AUC	F1-score	AUC	
ResNet18	CLAM-SB	0.666	0.772	0.647	0.768	
(ImageNet)	CLAM-MB	0.575	0.736	0.631	0.755	
	TransMIL	0.564	0.713	0.561	0.758	
	DSMIL	0.557	0.690	0.657	0.732	
	DTFD-MIL	0.647	0.757	0.607	0.755	
	ABMIL	0.624	0.759	0.637	0.783	
	ACMIL	0.633	0.773	0.734	0.797	
UNI	CLAM-SB	0.651	0.797	0.705	0.805	
(ViT-L/16, SSL)	CLAM-MB	0.674	0.808	0.702	0.817	
	TransMIL	0.700	0.826	0.702	0.817	
	DSMIL	0.710	0.767	0.709	0.755	
	DTFD-MIL	0.620	0.747	0.661	0.794	
	ABMIL	0.636	0.817	0.728	0.829	
	ACMIL	0.679	0.819	0.731	0.835	

The performance of MIL methods is compared using ResNet18 and UNI models, evaluated by F1-score and AUC on the TCGA-BRCA validation and test datasets

covered two main categories feature extractors: those pretrained with ImageNet and those pretrained with the proposed UNI model.

Among the models pretrained with ImageNet, CLAM-SB, CLAM-MB, TransMIL, DSMIL, DTFD-MIL, and ABMIL were evaluated using the F1 score and AUC on the test dataset. The ABMIL and ACMIL models performed particularly well. CLAM-SB performed well in validation, with an F1 score of 0.666 and an AUC of 0.772; however, in testing, the F1 score was slightly lower, with an F1 score of 0.647 and an AUC of 0.768. The

DSMIL and DTFD-MIL models exhibited consistent performance metrics in the tests. Particularly, ACMIL with the UNI-pretrained feature extractor performed the best on the test set, with an F1 score of 0.731 and an AUC of 0.835. The highest score for each class (grades 1, 2, and 3) was determined as the final prediction class. In this process, each score was multiplied by a randomized constant to obtain the final result. In the validation phase, it achieved an F1 score of 0.679 and an AUC of 0.819, whereas in the test set, it performed the best with an F1 score of 0.731 and an AUC of 0.835, demonstrating the effectiveness of the UNI-pretrained feature extractor in improving MIL model performance.

The detailed performance metrics and data are listed in Table 4. The performance evaluation results of the final selected ACMIL model (using UNI-pretrained features) were further analyzed using the ROC curves (Fig. 3A) and confusion matrix (Fig. 4A) derived for each Nottingham histological grade. The model showed different predictive capabilities across grades, achieving AUCs of 0.83 for grade 1, 0.77 for grade 2, and 0.90 for grade 3. Similarly, the confusion matrices demonstrated that the model achieved relatively high classification accuracy for grades 1 and 3, with slightly lower performance for grade 2. To evaluate the generalizability of the model, we additionally tested its performance on an independent external validation cohort from the BRACS dataset. The ROC curves and confusion matrices for this cohort are presented in Figs. 3B and 4B, respectively. The model achieved AUCs of 0.89 for grade 1, 0.70 for grade 2, and 0.83 for grade 3 in the external validation. While the classification performance for grade 2 remained moderate, the model maintained robust predictive ability for grades 1 and 3. These results indicate that the ACMIL model is capable of stratifying histologic grades in an independent cohort, supporting its applicability beyond the training dataset.

To further assess the agreement between the model predictions and pathologist-assigned grades, we additionally calculated Cohen's Kappa scores for each Nottingham grade class in both TCGA-BRCA dataset and BRACS dataset. The results are summarized in Table 5, demonstrating moderate to substantial agreement in the TCGA dataset, with lower agreement observed for grade 2 in the BRACS cohort. This highlights the model's consistency in predicting grades 1 and 3 across cohorts.

Survival analysis

We evaluated the OS of patients according to the NGS by comparing the graded survival based on pathologic classification and deep learning model prediction. Fiveyear survival rates were calculated for both the pathologist-classified grades and ACMIL model-predicted grades, which served as important indicators of tumor

A В ROC curves of model performance in the internal TCGA-BRCA test set ROC curves of model performance in the external BRACS validation set 1 1 0.8 0.8 Sensitivity Sensitivity 0.6 0.6 0.4 0.4 Nottingham Grade 1 (AUC = 0.83) Nottingham Grade 1 (AUC = 0.89) 0.2 0.2 Nottingham Grade 2 (AUC = 0.77) Nottingham Grade 2 (AUC = 0.70) Nottingham Grade 3 (AUC = 0.90) Nottingham Grade 3 (AUC = 0.83) 0 0 0.2 0.6 0.8 0.2 0.6 0 0.4 1 0 0.4 0.8 1 **False Positive Rate False Positive Rate**

Fig. 3 ROC Curves for Nottingham Grade Classification. (A) ROC curves of model performance in the internal TCGA-BRCA test set. (B) ROC curves of model performance in the external BRACS validation set



Classification performance on the TCGA test set





Classification performance on the BRACS external validation cohort



Fig. 4 Confusion Matrices for Nottingham Grade Classification in TCGA-BRCA and BRACS Datasets. (A) Confusion matrix from internal TCGA test set, showing the model's classification performance across Nottingham grades 1, 2, and 3. (B) Confusion matrix from external BRACS validation set, illustrating the model's generalizability in predicting Nottingham grades in an independent cohort

Table 5	Agreement of model versus pathologist: Cohen's ka	арра
scores in	ICGA-BRCA and BRACS datasets	

Dataset	Nottingham Grade	Cohen's Kappa (± SD)
TCGA-BRCA	Grade 1	0.41 ± 0.10
	Grade 2	0.51 ± 0.07
	Grade 3	0.65 ± 0.05
BRACS	Grade 1	0.46 ± 0.09
	Grade 2	0.25 ± 0.09
	Grade 3	0.43 ± 0.08

Cohen's Kappa coefficients with standard deviations are presented for each grade class, indicating the level of agreement between the model and pathologist classifications

progression and patient survival for each grade. The survival curves according to the tumor grade, as classified by the pathologist, showed relatively high survival rates for each grade. The ACMIL model predicted 5-year survival rates of 79.9% for grade 1, 75.7% for grade 2, and 59.1% for grade 3, with a trend toward lower survival rates for higher grades. These results were statistically significant (p < 0.05), indicating that the model predictions tended to match the actual clinical outcomes. This analysis aimed to show that model-predicted grades can stratify survival outcomes, highlighting their clinical relevance even without pathologist annotations. The log-rank test was used to assess whether the difference in survival between the graded patient populations was significant. This test was used to determine whether the OS curves for each of the three grades (1, 2, and 3) were significantly different. The survival curves based on pathologist classification (Fig. 5A) and those based on ACMIL model predictions

 Table 6
 Comparison of 5-year survival rates by tumor grade:

 pathologist vs. ACMIL model predictions

Outcome	5-year surviva	<i>p</i> -value		
	Grade 1	Grade 2	Grade 3	
os				
Pathologists	100.0 (100.0–100.0)	90.6 (77.1–96.3)	80.7 (66.6–89.3)	0.1232
ACMIL model	79.9 (74.3–84.4)	75.7 (57.4–86.9)	59.1 (38.0–75.1)	0.0173

The 5-year survival rates were estimated using the Kaplan-Meier method, while survival differences between tumor grades were tested using the log-rank test

(Fig. 5B) are shown. A comparison of the 5-year survival rates for each tumor grade is presented in Table 6.

To further assess the prognostic value of model-predicted Nottingham grades, we performed a multivariate Cox proportional hazards regression analysis, adjusting for tumor size, ER/HER2/PR status, and age. While the predicted grades did not retain statistical significance in the multivariate setting, the survival trends remained consistent with clinical expectations, with higher predicted grades associated with worse outcomes. These results indicate that model-derived grading captures prognostically relevant signals, although its independent predictive contribution may be influenced by other clinicopathological factors. The detailed results are presented in Supplementary material, Tables S3.



Fig. 5 Kaplan-Meier Survival Curves for Breast Cancer Patients by Nottingham Grades: Pathologist vs. Deep Learning. (A) Kaplan-Meier survival curves for pathologist-classified Nottingham grades show a clear trend, with grade 1 having the highest survival probability, followed by grades 2 and 3, reflecting the expected relationship between grades and survival outcomes. (B) Kaplan-Meier survival curves for Al-predicted Nottingham grades align with clinical expectations, showing grade 1 with the highest survival probability, grade 2 in the middle, and grade 3 with the lowest. Both graphs demonstrate consistent survival trends across Nottingham grades, with Al predictions closely matching clinical classifications

Visualization and focused analysis of Nottingham grades through heatmaps

The ACMIL model was used to generate heatmaps of breast cancer tissue slides according to Nottingham grades 1, 2, and 3, with attention scores visualized as color changes reflecting histological features. The original slide images were presented along with heat maps to show how the areas noted by the model corresponded to the actual pathological findings. For example, in Nottingham grade 1, a relatively uniform cellular structure and a small amount of microscopic cell proliferation were observed, which appeared as low-scoring areas in the heatmap of the model (Fig. 6A). By contrast, Nottingham Grade 2 showed a more irregular cell structure and an increase in microscopic cell proliferation, which was also reflected in the heatmap (Fig. 6B). In Nottingham grade 3, strong cell proliferation was observed along with a more irregular cell structure, which was reflected in the heatmap as high scoring areas (Fig. 6C). These results suggest that our prediction model can effectively distinguish the histological patterns associated with Nottingham grades. Furthermore, an expert pathologist reviewed the heatmaps and corresponding histologic slides, confirming that the attention regions identified by the model were consistent with clinically relevant histologic features. While these heatmaps qualitatively demonstrate the interpretability of the model, we also provided a quantitative evaluation of model-pathologist agreement



Fig. 6 Visualization of slides, heatmaps, and key regions for Nottingham grade. this figure illustrates the progression of histopathological features for Nottingham grades 1, 2, and 3 through original slides, corresponding heatmaps, and magnified grade-related regions: (A) Nottingham Grade 1: The original slide shows uniform cell structures. The heatmap highlights low-attention areas, reflecting minimal irregularities. The zoomed-in region confirms orderly tubular formations and low mitotic activity. (B) Nottingham Grade 2: The slide reveals moderately irregular structures. The heatmap shows mixed attention areas, indicating regions with moderate cellular pleomorphism and mitotic activity. The zoomed-in view corroborates an intermediate level of tubularity and nuclear atypia. (C) Nottingham Grade 3: The slide depicts highly irregular structures and significant cellular proliferation. The heatmap highlights intense attention areas, aligning with severe nuclear pleomorphism and high mitotic activity observed in the zoomed-in region. These visualizations demonstrate the AI model's ability to focus on histologically relevant regions that align with Nottingham grading criteria





Fig. 7 Key Biological Processes Associated with Nottingham Grade 3 Breast Cancer. The bar chart illustrates the key biological processes associated with Nottingham Grade 3 breast cancer, identified through gene ontology (GO) analysis. The most significant processes include cell division, chromosome segregation, and the mitotic cell cycle, reflecting the rapid and aggressive proliferation characteristic of Grade 3 tumors. Other notable processes, such as mitotic spindle organization, G2/M transition of the mitotic cell cycle, and mitotic cytokinesis, further highlight the enhanced mitotic activity observed in high-grade tumors. Additionally, processes like DNA replication and regulation of cyclin-dependent kinase activity underscore the genomic instability and dysregulated cell cycle mechanisms typical of advanced breast cancer. These results provide biological insights into the distinct and aggressive nature of Nottingham Grade 3

using Cohen's kappa statistics, as presented in Table 5, to support the reliability of the predicted grades.

Gene ontology analysis of biological processes related to Nottingham grade 3

GO analysis was performed to analyze the key biological processes associated with Nottingham Grade 3 breast cancer tissues. Based on the gene expression data according to the grade predicted using the ACMIL model, we identified statistically significant biological processes. The analysis showed that cell division, chromosome segregation, and mitotic cell

cycle pathways were prominently involved in grade 3. These processes reflect the aggressive and rapid growth properties of tumors and include important mechanisms involved in cancer progression. A detailed analysis is presented in Fig. 7. The gene expression analysis for grades 1 and 2 can be found in (supplementary material, Figure S1).

Discussion

We utilized the MIL methodology to predict the Nottingham histological grade from H&E-stained slide images of patients with breast cancer. The AI-based model developed in this study demonstrated that among the MIL models with different architectures and training methods, the model trained using the proposed UNI-pretrained features performed the best and could effectively predict the Nottingham histologic grade. The model outperformed other models in terms of the F1 score and AUC metrics, marking a significant advance in breast cancer histological analysis.

Although previous studies have applied AI models to predict the Nottingham histologic grade, they often showed limitations in clearly distinguishing between grades, particularly between grades 1 and 2. In contrast, our model demonstrated improved accuracy and consistency across all grades, addressing this key limitation in prior approaches.

Our model was able to predict cancer grades from Nottingham grades 1 to 3 with high accuracy and was successful in clearly distinguishing the boundaries between grades. Our model showed good consistency and predictive power at the boundary between grades 1 and 2, which previous studies failed to distinguish accurately.

An important feature of this study is that it analyzed the effectiveness of different approaches in comparing and evaluating state-of-the-art models. Different MIL models, including the ACMIL model, are based on different architectures, and a performance comparison among them allowed us to identify the most effective learning strategies. Particularly, the ACMIL model utilizes multibranch learning and stochastic top-K instance masking to achieve more sophisticated feature extraction and learning, which results in higher predictive accuracy and consistency compared with traditional MIL models.

Traditionally, the NGS relies heavily on the visual assessment of pathologists, which can lead to a lack of consistency in the assessment between observers or over time by the same observer. This study focused on minimizing the subjectivity between pathologists, especially in the assessment of histological grades. The AI-powered automated model developed in this study eliminated this subjectivity and provided an objective, quantified assessment based on histological characteristics, significantly improving consistency and reliability. By identifying meaningful patterns in complex medical data and providing deeper insight into clinical outcomes, these automated tools can improve medical diagnosis and treatment methods. These consistent and accurate predictions of the Nottingham histologic grades allowed us to explore how they relate to patient prognosis. The

results showed that patients with lower grades (grade 1) had higher survival rates, whereas those with higher grades (grade 3) had relatively lower survival rates. This confirms that pathological grade has a significant impact on patient prognosis and shows that AI models can support these important clinical decisions.

Additionally, we visualized the regions of interest in the tissue slide images according to the Nottingham grade using a heat map generated using the ACMIL model. The heatmap was generated based on the attention score assigned to each patch by the model and served as an important tool to visually confirm the histological characteristics. These visualizations showed that the AI model was in good agreement with the actual pathological findings, suggesting that it could be used more effectively by expert pathologists during the diagnostic process.

Furthermore, biological downstream analysis using RNA-seq data revealed that pathways related to cell division, chromosome segregation, and mitosis as biological processes that were particularly prominent in Nottingham grade 3. These pathways reflect the aggressive and rapid growth of cancer and include important mechanisms involved in disease progression.

Although this study predicted the Nottingham histologic grade of patients with breast cancer and used it to analyze survival, it had several limitations. First, because the study was conducted using publicly available data from TCGA, the follow-up duration, treatment information, and detailed data on tumor subtypes were limited. This may limit the depth of the analysis; in particular, the lack of inclusion of different tumor subtypes and treatment data may compromise the accuracy of the prognostic assessment. Future studies using larger cohorts with longer clinical follow-up periods, including data on different tumor subtypes and treatments, could overcome these limitations and clarify the clinical utility of this methodology.

Second, there was an imbalance in the data used in the analysis, with a relative lack of data from Nottingham histologic grade 1 compared to grades 2 and 3. This imbalance may have led to a lack of mortality information, particularly in survival analyses, thus affecting the reliability of the results. To resolve this data imbalance, it is necessary to obtain additional data corresponding to grade 1, which will play an important role in improving the reliability of the analysis.

Third, misclassification between certain Nottingham grades, particularly between grades 1 and 2, was observed in both the internal and external validation cohorts, as illustrated in the confusion matrices (Fig. 4A, B). This may be attributable to inherent histological ambiguities and overlapping morphological features between adjacent grades. Such limitations reflect challenges even in routine pathological assessments, where borderline cases often exist. Furthermore, as shown in Table 5, the external validation cohort demonstrated notably lower agreement between the model and pathologist for grade 2 (Cohen's kappa= 0.25 ± 0.09), compared to grades 1 and 3. This suggests the model's relatively weaker discriminative capacity for intermediate-grade tumors in an independent dataset. To improve classification robustness, future studies could explore model fusion techniques or integrate confidence estimation frameworks that assess the certainty of predictions. Such approaches could enable the model to flag borderline cases with low confidence, allowing for more cautious interpretation or further pathological review, especially in cases where histological ambiguity is high.

Fourth, this study partially utilized methodologies and models developed by existing researchers. This may make the performance of the models dependent on existing technical limitations and limit the development of original approaches or new models. Further studies using original model development and innovative methodologies could improve the accuracy and clinical applicability of breast cancer histologic grade prediction.

Fifth, Although the univariate survival analysis based on pathologist-assigned Nottingham grades demonstrated a clear stratification of overall survival, it did not reach statistical significance under the conventional threshold (p < 0.05), as shown in Fig. 5A. This limitation may reflect the intrinsic variability in patient prognosis that cannot be fully captured by histologic grading alone. To further assess the prognostic value of modelderived grading, we performed a multivariate Cox proportional hazards regression analysis incorporating key clinical covariates, including tumor size, ER/HER2/PR status, and age. In this multivariate context, neither the pathologist-assigned nor the model-predicted Nottingham grades retained statistical significance, indicating that their prognostic contribution may be attenuated by the influence of other clinicopathologic factors. This outcome may also be attributed to the overall cohort composition, limited availability of detailed clinical data, relatively low mortality rates in breast cancer populations, and the underrepresentation of grade 1 cases, all of which could reduce the statistical power of survival analyses and hinder robust stratification. Nonetheless, compared to pathologist-assigned grading, the modelpredicted grades demonstrated a relatively higher hazard ratio and more consistent stratification patterns across both univariate and multivariate analyses, particularly for higher-grade tumors. This suggests that the model may capture subtle histological features with prognostic relevance that are not easily discernible through visual inspection alone.

Conclusions

This study demonstrated the successful development of an AI model that can accurately predict Nottingham tissue grade from H&E-stained slides of breast cancer patients. The model effectively addresses a complex diagnostic task that traditionally requires expert-level knowledge and experience, enabling faster and more consistent assessments. By providing a standardized and automated diagnostic capability, it has the potential to significantly reduce inter-observer variability that is common in pathological assessments. Furthermore, prognostic and biological follow-up analyses validated the clinical utility of the model, highlighting the significance of our findings in the broader context of cancer research.

The model can predict tissue grade with high accuracy, which could be an important tool to support pathologists and improve the precision of diagnostic and treatment strategies. If integrated into clinical workflows, this AI model could contribute to improved patient outcomes by facilitating faster and more accurate diagnoses. The model also has the potential to be applied to other types of cancer and pathologic assessments, providing opportunities to further advance the automation and objectivity of pathology analysis. Ultimately, the integration of this AI model into the clinical environment could play an important role in enhancing personalized treatment approaches and optimizing cancer patient care.

Abbreviations

ABMIL	Attention-based Multiple instance learning
ACMIL	Attention-Challenging Multiple Instance Learning
AI	Artificial intelligence
AJCC	American Joint Committee on Cancer
AUROC	Area Under the Receiver Operating Characteristic Curve
BRACS	BReAst Carcinoma Subtyping
BRCA	Breast invasive carcinoma
CLAM	Clustering-constrained attention Multiple instance learning
CNR	National Research Council
CPM	count per million
DCIS	Ductal carcinoma in situ
DSMIL	Dual-stream Multiple instance learning
DTFD-MIL	Double-tier feature distillation Multiple instance learning
ER	Estrogen receptor
EU	European Union
GO	Gene ontology
H&E	Hematoxylin and eosin
HER2	Human epidermal growth factor receptor 2
HSV	Hue, Saturation, Value
IBM	International Business Machines
IC	Invasive carcinoma
ICAR	Institute for High Performance Computing and Networking
LVI	lymphovascular invasion
MIL	Multiple instance learning
NGS	Nottingham grading system
OS	Overall survival
PR	Progesterone receptor
SSL	Self-supervised learning
TCGA	The Cancer Genome Atlas
TransMIL	Transformer based Correlated Multiple instance learning
UK RCPath	United Kingdom Royal College of Pathologists
ViT-L/16	Vision Transformer-Large/16
WHO	World Health Organization
WSIs	Whole-slide images

Supplementary Information

The online version contains supplementary material available at https://doi.or g/10.1186/s13058-025-02019-4.

Supplementary Material 1: Supplementary material, Figure S1: Gene Ontology (GO) analysis results for GRADE 1(A) and GRADE 2(B).

Supplementary Material 2: Supplementary material, Table S1: Nottingham Histologic Score for TCGA-BRCA tumors examined in this study.

Supplementary Material 3: Supplementary material, Table S2: Nottingham Histologic Score for BRACS tumors analyzed in this study.

Supplementary Material 4: Supplementary material, Table S3: Multivariate Cox proportional hazards regression analysis comparing pathologistassigned and ACMIL model-predicted Nottingham grades in relation to 5-year overall survival.

Acknowledgements

Not applicable.

Author contributions

JK, JL, and MN designed the study. JK performed the analysis, prepared the tables and figures, and drafted the manuscript. JL contributed to the study design, prepared the figures, and revised the manuscript. YY interpreted the literature. DA interpreted the literature and drafted parts of the manuscript. SK contributed to the interpretation of the literature. MN prepared figures, visually identified key regions on tissue slide images according to each Nottingham grade, and revised the manuscript. SL provided overall guidance and critically reviewed the manuscript. All authors reviewed and approved the final version for submission.

Funding

This work was supported by the following research grants: the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health&Welfare, Republic of Korea (grant number: HI23C1494); and Clinical-Basic Collaborative Research Project at Ajou University Medical Center (M-2024-C0460-00073).

Data availability

TCGA image data used in this study are publicly available through the Genomic Data Commons Data Portal https://portal.gdc.cancer.gov). Survival data and other clinical data, including survival status and follow-up duration, can be accessed via Firehose Legacy, Pan-Cancer Atlas in Genomic Data Commons, and cBioPortal (https://www.cbioportal.org/). BRACS image data used in this study are publicly available through the BRACS dataset portal (https://www.bracs.icar.cnr.it). The Nottingham Histologic Score for TCGA-BRC A tumors collected in this study is provided in supplementary material Table S1. The Nottingham Histologic Score for BRACS tumors analyzed in this study is provided in Supplementary material, Table S2.

Declarations

Ethics approval and consent to participate

This study utilized publicly available and de-identified data from The Cancer Genome Atlas (TCGA) database and the BRACS dataset. Since the data are de-identified and publicly accessible, no additional ethical approval or informed consent was required for this study.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 17 December 2024 / Accepted: 9 April 2025 Published online: 19 April 2025

References

- Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49. https://doi.org/10.3322/caac.21660.
- Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229–63. https://doi.org/10.3322/caac.21834.
- 3. Bloom HJ, Richardson WW. Histological grading and prognosis in breast cancer. Br J Cancer. 1957;11(3):359–77. https://doi.org/10.1038/bjc.1957.43.
- Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology. 1991;19(5):403–10. https://doi.org /10.1111/j.1365-2559.1991.tb00229.x.
- Rakha EA, El-Sayed ME, Lee AH, et al. Prognostic significance of Nottingham histologic grade in invasive breast carcinoma. J Clin Oncol. 2008;26(19):3153– 8. https://doi.org/10.1200/JCO.2007.15.5986.
- Rakha EA, Reis-Filho JS, Baehner F, et al. Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Res. 2010;12(4):207. https://doi.org/10.1186/bcr2607.
- Ginter PS, Idress R, D'Alfonso TM, et al. Histologic grading of breast carcinoma: a multi-institution study of interobserver variation using virtual microscopy. Mod Pathol. 2021;34(4):701–9. https://doi.org/10.1038/s41379-020-00698-2.
- Acs B, Rantalainen M, Hartman J. Artificial intelligence as the next step towards precision pathology. J Intern Med. 2020;288(1):62–81. https://doi.org /10.1111/joim.13030.
- Kang M, Song H, Park S et al. Benchmarking self-supervised learning on diverse pathology datasets. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit. 2023:3344–54.
- Chen RJ, Ding T, Lu MY, et al. Towards a general-purpose foundation model for computational pathology. Nat Med. 2024;30(3):850–62. https://doi.org/10. 1038/s41591-024-02857-3.
- Cooper M, Ji Z, Krishnan RG, et al. Machine learning in computational histopathology: challenges and opportunities. Genes Chromosomes Cancer. 2023;62(9):540–56. https://doi.org/10.1002/gcc.23177.
- Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25(8):1301–9. https://doi.org/10.1038/s41591-019-0508-1.
- King H, Wright J, Treanor D, et al. What works where and how for uptake and impact of artificial intelligence in pathology: review of theories for a realist evaluation. J Med Internet Res. 2023;25:e38039. https://doi.org/10.2196/3803 9.
- Wang Y, Acs B, Robertson S, et al. Improved breast cancer histological grading using deep learning. Ann Oncol. 2022;33(1):89–98. https://doi.org/10.1016/j.a nnonc.2021.09.007.
- Jaroensri R, Wulczyn E, Hegde N, et al. Deep learning models for histologic grading of breast cancer and association with disease prognosis. NPJ Breast Cancer. 2022;8(1):113. https://doi.org/10.1038/s41523-022-00478-y.
- Wetstein SC, de Jong VMT, Stathonikos NM, et al. Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images. Sci Rep. 2022;12(1):15102. https://doi.org/10.1038/s41598-022-19112-9.
- Sharma A, Weitz P, Wang Y, et al. Development and prognostic validation of a three-level NHG-like deep learning-based model for histological grading of breast cancer. Breast Cancer Res. 2024;26(1):17. https://doi.org/10.1186/s1305 8-024-01770-4.
- Jensen MA, Ferretti V, Grossman RL, et al. The NCI genomic data commons as an engine for precision medicine. Blood. 2017;130(4):453–9. https://doi.org/1 0.1182/blood-2017-03-735654.
- Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, et al. The Cancer genome atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113–20. https://doi.org/10.1038/ng.2764.

- Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer genome atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19(1A):A68–77. https://doi.org/10.5114/wo.2014.47136.
- Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401–4. https://doi.org/10.1158/2159-8290.CD-12-0095.
- de Bruijn I, Kundra R, Mastrogiacomo B, et al. Analysis and visualization of longitudinal genomic and clinical data from the AACR project GENIE biopharma collaborative in cBioPortal. Cancer Res. 2023;83(23):3861–7. https://doi.org/10 .1158/0008-5472.CAN-23-0816.
- Dd Brancati N, Anniciello AM, Pati P, Riccio D, Scognamiglio G, Jaume G, et al. Bracs: A dataset for breast carcinoma subtyping in H&E histology images. Database (Oxford). 2022;2022:baac093. https://doi.org/10.1093/database/baa c093.
- Jacobson RS, Becich MJ, Bollag RJ, et al. A federated network for translational cancer research using clinical data and biospecimens. Cancer Res. 2015;75(24):5194–201. https://doi.org/10.1158/0008-5472.CAN-15-1973.
- Asaoka M, Patnaik SK, Zhang F, et al. Lymphovascular invasion in breast cancer is associated with gene expression signatures of cell proliferation but not lymphangiogenesis or immune response. Breast Cancer Res Treat. 2020;181(2):309–22. https://doi.org/10.1007/s10549-020-05630-5.
- 26. Caron M, Touvron H, Misra I et al. Emerging properties in self-supervised vision Transformers. In: Proc IEEE/CVF Int Conf Comput Vis. 2021:9650–60.
- 27. Dosovitskiy A. An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020.
- He K, Zhang X, Ren S et al. Deep residual learning for image recognition. In: Proc IEEE Conf Comput Vis Pattern Recognit. 2016:770–8.
- Lu MY, Williamson DF, Chen TY, et al. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng. 2021;5(6):555–70.
- 30. Zhang Y, Li H, Sun Y et al. Attention-challenging multiple instance learning for whole slide image classification. arXiv preprint arXiv:2311.07125. 2023.
- Shao Z, Bian H, Chen Y, et al. Transmil: transformer based correlated multiple instance learning for whole slide image classification. Adv Neural Inf Process Syst. 2021;34:2136–47.
- Li B, Li Y, Eliceiri KW et al. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit. 2021:14318–14328.
- Zhang H, Meng Y, Zhao Y et al. Dtfd-mil: double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit. 2022;18802–18812.
- 34. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In: Int Conf Mach Learn PMLR. 2018:2127–36.
- 35. Sherman BT, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. 2022;50(W1):W216–21.
- Vostakolaei FA, Karim-Kos HE, Janssen-Heijnen ML, et al. The validity of the mortality to incidence ratio as a proxy for site-specific cancer survival. Eur J Public Health. 2011;21(5):573–7. https://doi.org/10.1093/eurpub/ckq120.
- Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. N Engl J Med. 2016;375(12):1109–12.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.