Artificial intelligence-based morphologic classification and molecular characterization of neuroblastic tumors from digital histopathology

Mark Applebaum

mapplebaum@bsd.uchicago.edu

University of Chicago

Siddhi Ramesh

University of Chicago

Emma Dyer

University of Chicago Medical Center

Monica Pomaville

Children's Hospital of Phildelphia https://orcid.org/0009-0008-0498-5544

Kristina Doytcheva

University of Chicago

James Dolezal

University of Chicago Medicine

Sara Kochanny

University of Chicago

Rachel Terhaar

University of Chicago

Casey Mehrhoff

University of Utah

Kritika Patel

University of Washington https://orcid.org/0009-0007-3569-294X

Jacob Brewer

University of Chicago

Benjamin Kusswurm

University of Chicago

Arlene Naranjo

University of Florida

Hiroyuki Shimada

Stanford University

Elizabeth Sokol

Ann & Robert H. Lurie Children's Hospital of Chicago https://orcid.org/0000-0002-1787-6901
Susan Cohn
University of Chicago https://orcid.org/0000-0001-5749-7650
Rani George
Harvard Medical School
Alexander Pearson
University of Chicago https://orcid.org/0000-0003-2801-7456

Brief Communication

Keywords:

Posted Date: June 4th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4396782/v1

Additional Declarations: There is a conflict of interest SR is the Chief Scientific Officer of Slideflow Labs. JD is Chief Executive Officer of Slideflow Labs. BK is a current employee of Youtube. JB is an employee of Milliman. SLC reports consulting fees from US WorldMeds. ATP reports consulting fees from Prelude Biotherapeutics, LLC, Ayala Pharmaceuticals, Elvar Therapeutics, Abbvie, and Privo, and contracted research with Kura Oncology, Abbvie, and EMD Serono. ATP is on the Scientific Advisory Board of Slideflow Labs. All other authors report no competing interests.

Abstract

A deep learning model using attention-based multiple instance learning (aMIL) and self-supervised learning (SSL) was developed to perform pathologic classification of neuroblastic tumors and assess *MYCN*-amplification status using H&E-stained whole slide digital images. The model demonstrated strong performance in identifying diagnostic category, grade, mitosis-karyorrhexis index (MKI), and *MYCN*-amplification on an external test dataset. This AI-based approach establishes a valuable tool for automating diagnosis and precise classification of neuroblastoma tumors.

Introduction

Neuroblastoma is a neuroblastic tumor (NT) and the most common extracranial pediatric solid tumor, affecting nearly 800 children in the United States annually.[1] To select optimal treatment strategies, patients are risk-stratified according to prognostic clinical, pathologic, and molecular variables including age, stage, histopathology, and *MYCN*-amplification.[2,3] Approximately 40% of patients with neuroblastoma are classified as high-risk, which carries a 60% overall three-year likelihood of event free survival.[4] *MYCN*-amplification is present in 20% of NTs and, when identified, places the patient in the high-risk category.[5]

The pathologic classification of NTs is a major contributor to risk stratification. The International Neuroblastoma Pathology Committee (INPC) uses combinations of four features—age, diagnostic category (neuroblastoma, ganglioneuroblastoma intermixed, ganglioneuroma, or ganglioneuroblastoma nodular), grade of differentiation, and mitosis-karyorrhexis index (MKI) — to classify tumors as favorable or unfavorable histology.[6] INPC classification has significant prognostic ability unto itself, as those with unfavorable histology have a four times higher likelihood of relapse compared to those with favorable histology.[2]

Histology from hematoxylin and eosin (H&E)-stained slides can also serve as a rich data source for deep learning models, which can be used to identify nuanced motifs in tumor morphology and produce precise risk stratification criteria.[7–9] Machine learning algorithms have been used to analyze NT digitized histology as early as 2009, with models that segmented cells and extracted texture features from histology images to predict tumor grade.[10] More recently, convolutional neural networks (CNNs) have been deployed on NT histology risk stratification.[11]

Using our open-source deep learning analysis pipeline, Slideflow (2.3.1), we developed an attention-based multiple instance learning (aMIL) model with features extracted by CTransPath, a pre-trained self-supervised learning (SSL) model.[12–14] In contrast to conventional CNNs, aMIL models rely on pre-trained features to begin model training (Fig. 1). These features are obtained by passing images through a feature extractor network that has been pre-trained on either domain-specific or non-specific images. CTransPath is a domain-specific model that has been trained on unlabeled H&E-stained slides from The

Cancer Genome Atlas (TCGA).[12] For limited datasets such as those obtainable in rare diseases, using domain-specific features to train an aMIL can offer significant performance advantages over non-specific models such as ImageNet.[15,16]

In this study, we leveraged the largest reported study cohort of digitized NTs analyzed with these state-of-the-art deep learning methods. We generated a training dataset of whole slide images (WSIs) from patients from the University of Chicago and the Children's Oncology Group. These WSIs were used to develop models for predicting diagnostic category, grade, MKI, and *MYCN*-amplification status. Model performance was validated on an external test dataset of WSIs from patients seen at Lurie Children's Hospital. We aimed to demonstrate the feasibility of using aMILs to aid in NT classification and risk stratification.

The median age of patients with digitalized NT in the training dataset (n = 172) was 2.63 years (SD = 4.37). Among patients with additional known clinical information, 84 of 138 (60.2%) had metastatic disease and 94 of 133 (70.7%) were high-risk. For diagnostic category, the dataset includes 24 ganglioneuroblastomas and 148 neuroblastomas which were confirmed by pathologists (KD, HS, PP). Of the 148 tumors with a diagnostic category of neuroblastoma, 93.2% were poorly differentiated and 25% had high MKI. Of the 135 tumors with known *MYCN* status, 40 were amplified (29.6%). The median age of the external test dataset (n = 25) was 3.33 years (SD = 2.90). All patients in the test dataset were high-risk and 23 of 25 (92%) had metastatic disease. Of the 23 tumors classified as neuroblastoma, all were poorly differentiated. Eleven of these 23 tumors (48%) had a high MKI. Eight of the 25 tumors (32%) were *MYCN*-amplified.

The final models demonstrated highly accurate performance across all outcomes in the training cohort (Fig. 2). Area Under the Receiver Operator Curve (AUROC) for diagnostic category, grade, MKI, and *MYCN* were 0.96, 0.85, 0.71, and 0.77, respectively, and (Area Under the Precision Recall Curve) AUPRC was 0.99, 0.99, 0.88, and 0.89, respectively. The model had the most success identifying diagnostic categories, with a sensitivity of 0.93 and specificity of 0.92. For *MYCN* status, a sensitivity of 0.75 and specificity of 0.73 was demonstrated in the analysis.

Using an independent cohort of clinically annotated NT tumors, the models demonstrated high accuracy across all analyzed outcomes, validating the findings in the training data set (Fig. 2). For diagnostic category, the AUROC was 0.85 [95% Confidence Interval (CI) 0.71–0.99], with an AUPRC of 0.99 (95% CI 0.94-1.0), sensitivity of 0.87 (95% CI 0.68–0.95), and specificity of 0.50 (95% CI 0.09–0.91). The AUROC for MKI was 0.74 (95% CI 0.56–0.92), with an AUPRC of 0.83 (95% CI 0.68–0.99), sensitivity of 0.50 (0.25–0.75), and specificity of 0.91 (95% CI 0.62–0.98). For MYCN status, the AUROC was 0.81 (95% CI 0.65–0.98), with an AUPRC of 0.77 (95% CI 0.64–0.97), sensitivity of 1.0 (95% CI 0.78-1.0), and specificity of 0.63 (95% CI 0.30–0.86). Grade could not be assessed in the external test cohort as all samples were poorly differentiating.

Expert pathologist (PP) review of the model's attention heatmaps, generated using GRAD-CAM, revealed that the models were primarily focusing on neoplastic areas of the tumor, rather than relying on non-tumor tissues such as fibrosis, fibrovascular stroma, or adrenal tissue. While in most cases the model accurately identified and focused on the relevant tumor regions, in some instances correlation was unevenly distributed across the relevant tumor area. This suggests that this variation in attention may correlate with less well characterized diffuse histopathological signatures that have unclear associations with standard pathologic descriptions. Further investigation into these attention patterns is necessary to elucidate novel morphological features or subtypes within neuroblastoma tumors.[17] Overall, the pathologist's analysis confirmed that the model was generally making predictions based on the most relevant areas within the neoplastic regions of each sample.

We show the feasibility of using small datasets of H&E-stained WSIs to develop models for morphologic classification of NTs and accurate assessment of *MYCN*-amplification status at diagnosis using an aMIL deep learning model. While prior deep learning models for NTs relied heavily on morphological feature extraction and labeled data, our method used unlabeled data in conjunction with SSL methods to improve model performance when working with a small dataset.[10,11] The model achieved notable performance in identifying diagnostic category and a strong ability to identify *MYCN*-amplification. The highly accurate automatic classification produced by the model could be refined with additional data to eventually streamline pathologist workflows.

The model’s ability to identify *MYCN*-amplification status from histology is an encouraging result, particularly given the limited data used to train the model. This suggests models could also be built to predict other relevant genomic features such as copy number variations and ploidy. As 50% of high-risk NTs do not harbor *MYCN*-amplification and typically have other findings such as 11q aberrations, a deep learning approach may also provide the ability to readily identify features that drive aggressive growth in non-*MYCN*-amplified high-risk tumors.[18] Unlike immunohistochemistry or fluorescence *in situ* hybridization where a single gene aberration is probed, deep learning models analyze the image at a global level and may be able to more readily identify morphological signatures produced by combinations of gene alterations that could further aid in stratifying NTs.

Limitations of this study arise largely from data availability. As NTs are rare, it remains difficult to collect sufficient samples to train a robust deep learning model. Our approach makes use of a network architecture that seeks to overcome this limitation. However, the model could further be improved with more data. Additionally, this study seeks to aid molecular pathology diagnostics and does not constitute a pathologist replacement. The model’s predictions act as a second pair of eyes and could alert a pathologist to review specific, notable aspects of the histology.

This work provides an important step forward in automating diagnosis and precise classification of NTs with the addition of deep learning-based image analysis. Ultimately, this can increase global access to molecular and pathological classification for tumors in regions without access to experts. We also demonstrate the ability of aMIL models to perform well on small datasets; this model architecture could

be extended to other rare cancers that suffer from low data availability. This artificial intelligence-based approach establishes another data modality in the pathologist's toolbox for NT classification.

Methods

Dataset description

H&E-stained slides from the time of initial diagnosis were obtained from the University of Chicago (n = 102), the Children's Oncology Group (n = 70), and Lurie Children's Hospital (n = 25). The images were reviewed by trained pathologists (HS, PP, KD) who annotated the tumor regions and defined the diagnostic category (ganglioneuroblastoma/neuroblastoma), grade (differentiating/poorly differentiating), and MKI (low/intermediate and high). *MYCN* status was abstracted from patient records (amplified/non-amplified). This study was approved by the University of Chicago (IRB20-0659) and Lurie Children's Hospital Internal Review Boards (IRB 2021–4498).

Image processing

WSIs were captured using an Aperio AT2 DX WSI Scanner. To remove normal background tissue and maximize cancer-specific training, image tiles were extracted from within pathologist-annotated regions of tumor. Image tiles were extracted from WSIs with a width of 302μ and 299 x 299 pixels using Slideflow version 2.3.1 and the libvips backend. Grayspace filtering, Otsu's thresholding, and gaussian blur filtering (sigma = 3, threshold = 0.02) were used to remove background.

Classifier training

Extracted tiles were converted into feature vectors using CTransPath with 'reinhard mask' normalization applied.[12] aMIL models were trained on extracted features in Slideflow with the FastAI API and Pytorch. The aMIL model parameters were: weight decay of $1e^{-5}$, bag size of 256, batch size of 32, and training for 10 epochs. aMIL models were evaluated with 5-fold cross validation and by calculating the average AUROC, AUPRC, sensitivity, specificity, and F-1 score. Patients were excluded from a given model if the measure of interest was unknown.

Model Validation

The aMIL model developed during training was used on the unseen external test dataset. Samples were evaluated in one run without any hyperparameter tuning on test data to ensure no validation leakage. Model performance was assessed as above.

Pathologist Explainability Assessment

Explainability heatmaps were generated using GradCAM.[19] PP reviewed the heatmaps to identify whether tumor regions that the model found important for outcome prediction had clinical correlation to the given outcome.

Declarations

Code Availability

This code relies extensively on the open-source software package Slideflow, version 2.3.1, which is available at https://github.com/jamesdolezal/slideflow. The code used for this experiment can be found at https://github.com/siddhir/NB_histology.

Acknowledgements

This work was supported in part by the Burroughs Wellcome Fund Early Scientific Training Program to Prepare for Research Excellence Post-Graduation (BEST-PREP; SR), University of Chicago Pritzker School of Medicine Summer Research Program (SR). Also supported by the National Institutes of Health P30CA014599 which provided funding for the analysis and supports the Human Tissue Research Core at the University of Chicago. Research reported in this publication was supported by the Children's Oncology Group, the National Cancer Institute of the National Institutes of Health under award numbers U10CA180886, U24CA196173, and U10CA180899. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Health. We'd like to thank Dr. Lynn Yee for facilitating inter-institutional exchange of H&E-stained slides.

Contributions

Siddhi Ramesh (SR): Conceptualization, Methodology, Investigation, Formal Analysis, and Writing - Reviewing & Editing.

Emma Chancellor Dyer (ECD): Writing - Original Draft, Visualization, Software

Monica Pomaville (MP): Writing – Data Curation, Reviewing & Editing

Kristina Doytcheva (KD): Writing - Histology Review, Reviewing & Editing

James Dolezal (JD): Writing - Reviewing & Editing

Sara Kochanny (SK): Writing – Data Curation, Reviewing & Editing

Rachel Terhaar (RT): Writing – Data Curation, Reviewing & Editing

Casey J Mehrhoff (CJM): Writing - Data Curation, Reviewing & Editing

Kritika Patel (KP): Writing - Data Curation, Reviewing & Editing

Jacob Brewer (JB): Writing - Formal Analysis, Reviewing & Editing

Benjamin Kusswurm (BK): Writing - Formal Analysis, Reviewing & Editing

Hiroyuki Shimada (HS): Writing - Data Curation, Reviewing & Editing

Arlene Naranjo (AN): Writing – Data Curation, Reviewing & Editing

Peter Pytel (PP): Writing - Histology Review, Reviewing & Editing

Elizabeth A Sokol (EAS): Writing – Data Curation, Reviewing & Editing

Susan L Cohn (SLC): Writing - Data Curation, Reviewing & Editing

Rani E George (REG): Writing - Data Curation, Reviewing & Editing

Alexander T Pearson (ATP): Writing – Methodology, Reviewing & Editing

Mark A Applebaum (MAA): Writing – Conceptualization, Methodology, Formal Analysis, Writing - Reviewing & Editing

Competing Interests

SR is the Chief Scientific Officer of Slideflow Labs. JD is Chief Executive Officer of Slideflow Labs. BK is a current employee of Youtube. JB is an employee of Milliman. SLC reports consulting fees from US WorldMeds. ATP reports consulting fees from Prelude Biotherapeutics, LLC, Ayala Pharmaceuticals, Elvar Therapeutics, Abbvie, and Privo, and contracted research with Kura Oncology, Abbvie, and EMD Serono. ATP is on the Scientific Advisory Board of Slideflow Labs. All other authors report no competing interests.

References

1. Campbell K, Siegel DA, Umaretiya PJ, et al. A comprehensive analysis of neuroblastoma incidence, survival, and racial and ethnic disparities from 2001 to 2019. Pediatric Blood & Cancer. 2024;71(1):e30732. doi:10.1002/pbc.30732
2. Irwin MS, Naranjo A, Zhang FF, et al. Revised Neuroblastoma Risk Classification System: A Report From the Children's Oncology Group. JCO. 2021;39(29):3229–3241. doi:10.1200/JCO.21.00278
3. Sokol E, Desai AV, Applebaum MA, et al. Age, Diagnostic Category, Tumor Grade, and Mitosis-Karyorrhexis Index Are Independently Prognostic in Neuroblastoma: An INRG Project. J Clin Oncol. 2020;38(17):1906–1918. doi:10.1200/JCO.19.03285
4. Pinto NR, Applebaum MA, Volchenboum SL, et al. Advances in Risk Classification and Treatment Strategies for Neuroblastoma. JCO. 2015;33(27):3008–3017. doi:10.1200/JCO.2014.59.4648
5. Thompson D, Vo KT, London WB, et al. Identification of patient subgroups with markedly disparate rates of MYCN amplification in neuroblastoma: A report from the International Neuroblastoma Risk Group project. Cancer. 2016;122(6):935–945. doi:10.1002/cncr.29848
6. Shimada H, Ambros IM, Dehner LP, et al. The International Neuroblastoma Pathology Classification (the Shimada system). Cancer. 1999;86(2):364–372. doi:10.1002/(SICI)1097-0142(19990715)86:2<364::AID-CNCR21>3.0.CO;2-7

7. Qaiser T, Lee CY, Vandenberghe M, et al. Usability of deep learning and H&E images predict disease outcome-emerging tool to optimize clinical trials. npj Precis Onc. 2022;6(1):1–12. doi:10.1038/s41698-022-00275-7
8. Hu J, Lv H, Zhao S, Lin CJ, Su GH, Shao ZM. Prediction of clinicopathological features, multi-omics events and prognosis based on digital pathology and deep learning in HR + /HER2 – breast cancer. Journal of Thoracic Disease. 2023;15(5). doi:10.21037/jtd-23-445
9. Liang J, Zhang W, Yang J, et al. Deep learning supported discovery of biomarkers for clinical prognosis of liver cancer. Nat Mach Intell. 2023;5(4):408–420. doi:10.1038/s42256-023-00635-3
10. Kong J, Sertel O, Shimada H, Boyer KL, Saltz JH, Gurcan MN. Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation. Pattern Recognit. 2009;42(6):1080–1092. doi:10.1016/j.patcog.2008.10.035
11. Gheisari S, Catchpoole DR, Charlton A, Kennedy PJ. Convolutional Deep Belief Network with Feature Encoding for Classification of Neuroblastoma Histological Images. Journal of Pathology Informatics. 2018;9(1):17. doi:10.4103/jpi.jpi_73_17
12. Wang X, Yang S, Zhang J, et al. Transformer-based unsupervised contrastive learning for histopathological image classification. Medical Image Analysis. 2022;81:102559. doi:10.1016/j.media.2022.102559
13. Ilse M, Tomczak JM, Welling M. Attention-based Deep Multiple Instance Learning. Published online June 28, 2018. doi:10.48550/arXiv.1802.04712
14. Dolezal JM, Kochanny S, Dyer E, et al. Slideflow: deep learning for digital histopathology with real-time whole-slide visualization. BMC Bioinformatics. 2024;25(1):134. doi:10.1186/s12859-024-05758-x
15. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition.*; 2009:248–255. doi:10.1109/CVPR.2009.5206848
16. Filiot A, Ghermi R, Olivier A, et al. Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling. Published online September 14, 2023:2023.07.21.23292757. doi:10.1101/2023.07.21.23292757
17. Dolezal JM, Wolk R, Hieromnimon HM, et al. Deep learning generates synthetic cancer histology for explainability and education. npj Precis Onc. 2023;7(1):1–13. doi:10.1038/s41698-023-00399-4
18. Luttikhuis MEMO, Powell JE, Rees SA, et al. Neuroblastomas with chromosome 11q loss and single copy MYCN comprise a biologically distinct group of tumours with adverse prognosis. Br J Cancer. 2001;85(4):531–537. doi:10.1054/bjoc.2001.1960
19. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Int J Comput Vis. 2020;128(2):336–359. doi:10.1007/s11263-019-01228-7

Figures

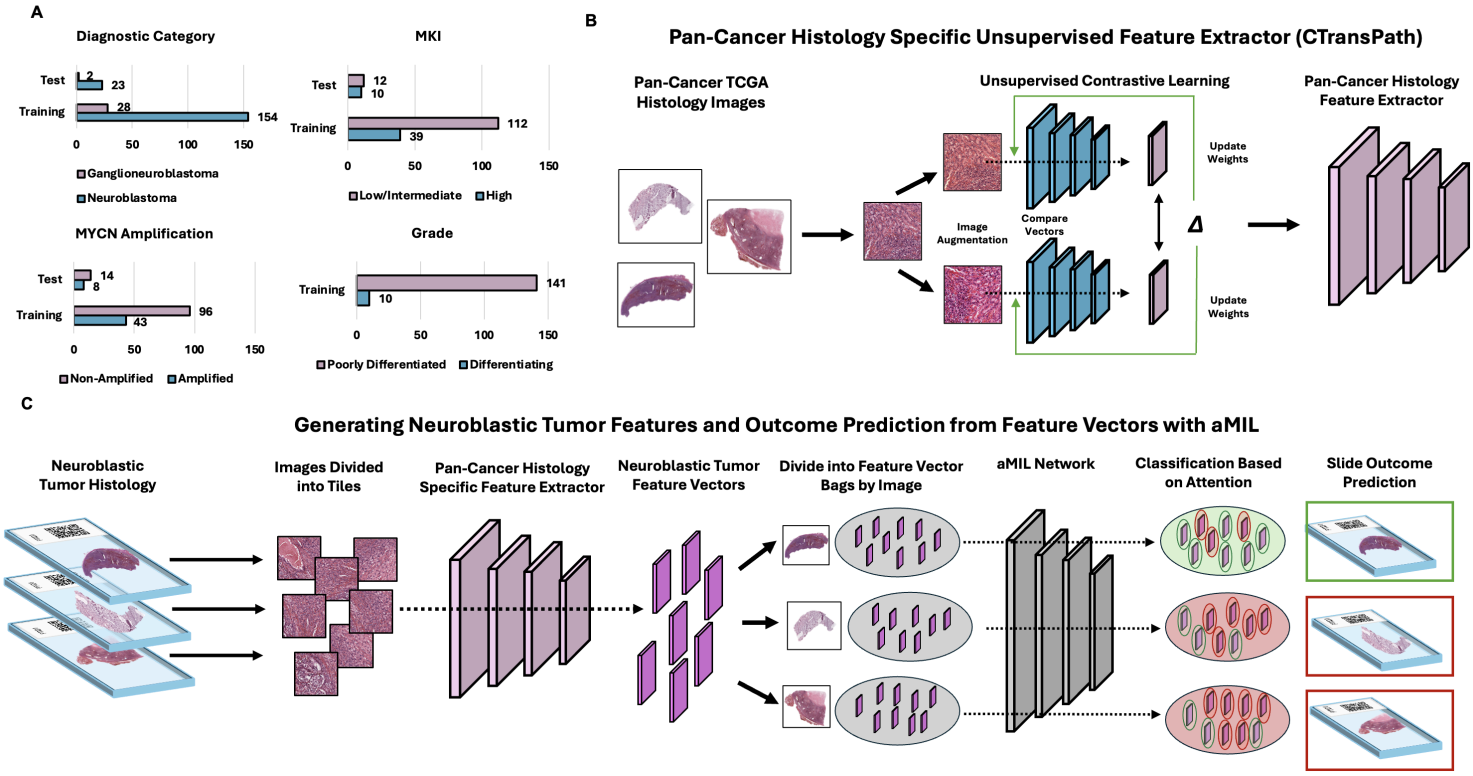


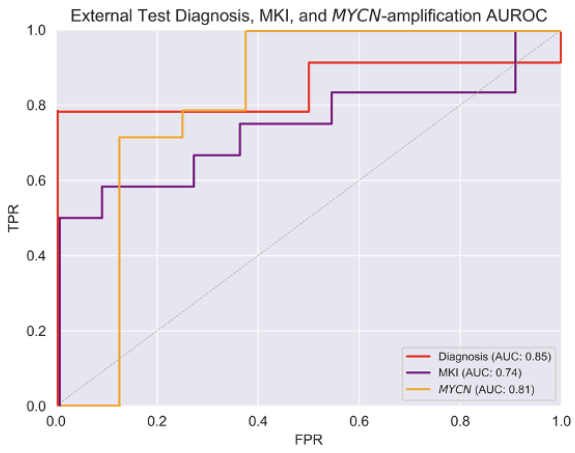
Figure 1

attention-based multiple instance learning (aMIL) models use feature vectors as inputs, grouped in bags, to make predictions aggregated from all vectors within a bag. (**A**) Number of slides in the training and test cohorts by pathologic category. (**B**) Models were pre-trained with histology-specific digital images using unsupervised domain-specific learning to extract features with CTransPath. (**C**) Whole slide images (WSI) were divided into tiles, passed through the fine-tuned network to generate neuroblastoma-specific feature vectors, which are divided into bags per WSI. The aMIL network assigns attention scores to vectors, and a slide-level prediction is determined based on the aggregated predictions weighted by attention scores.

A

	Training Dataset				External Test Dataset		
Metric	Diagnostic Category	Grade	MKI	*MYCN*	Diagnostic Category	MKI	*MYCN*
AUROC	0.96	0.85	0.71	0.77	0.85	0.74	0.82
AUPRC	0.99	0.99	0.88	0.89	0.99	0.83	0.84
Specificity	0.92	0.70	0.60	0.73	0.50	0.73	0.38
Sensitivity	0.93	0.80	0.77	0.75	0.87	0.67	1.00
Precision	0.99	0.97	0.85	0.87	0.95	0.73	0.74
F-1 Score	0.95	0.88	0.81	0.80	0.91	0.70	0.85

B



C

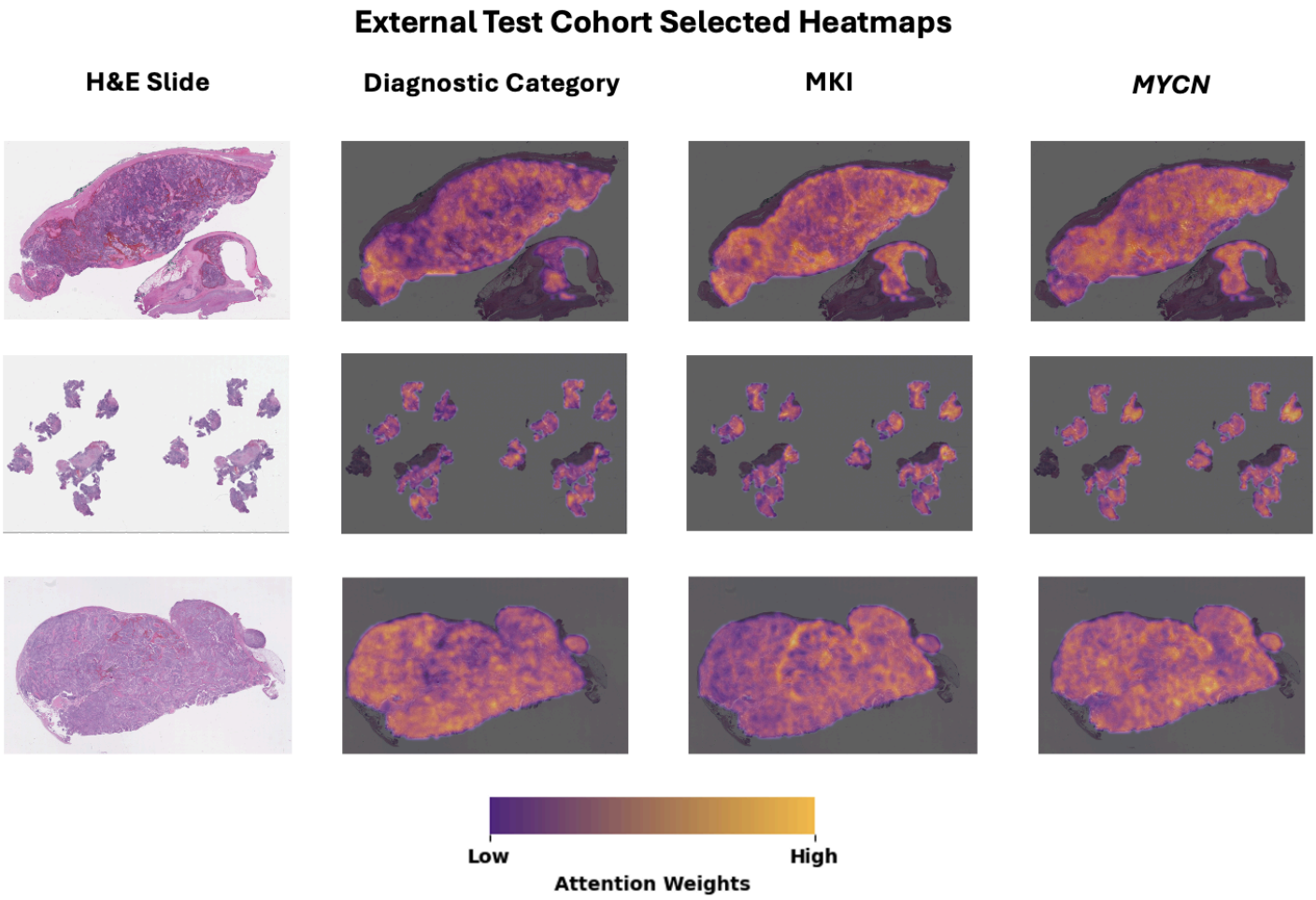


Figure 2

Model performance and explainability. (**A**) Performance metrics for training and external test models. (**B**) AUROC plots for the external test models. (**C**) Explainability heatmaps generated with GradCAM. Yellow regions were highly weighted and informative to the model while dark purple regions corresponded to low weights in generating predictions. Abbreviations: AUROC, Area Under the Receiver Operator Curve; AUPRC, Area Under the Precision Recall Curve.