Interpretable multimodal deep learning to predict breast cancer stage

Hannah Bibby, Tong Ding, Sharon Jiang, *Member, IEEE*, Charlie Lu, Chenwei Wu, Senthil Nachimuthu, MD

***Abstract*—Recently, more breast cancers have been found via biopsies, which has prompted more surgical procedures and chemotherapy. However, death rates from metastatic breast cancer have hardly changed. Thus, there remains a need to augment the determination of breast cancer staging, which will help predict the course the cancer will likely take, as well as determine the appropriate treatment. Deep learning algorithms have the potential to predict which cancers are high risk based on the biopsy image alone. However, there are few deep learning models that can integrate multiple clinical data types, which include histopathology imaging and structured patient information from electronic health records. Furthermore, interpretability issues remain a significant obstacle to the widespread deployment of deep learning models in clinical settings. Here, we introduce a novel multimodal method that combines attention-based multiple-instance learning on biopsy images and self-normalized networks on structured clinical metadata to predict breast cancer staging. We show the superior performance of the multimodal model over unimodal models on a large longitudinal dataset linking imaging and tabular data. Our results demonstrate that the multimodal model can be used to provide interpretable explanations by identifying clinically relevant features from both pathology images and metadata.**

***Index Terms*—computational pathology, interpretability, multimodal deep learning**

I. Introduction

The field of digital pathology has been rapidly expanding, and digital pathology workflows have become increasingly prevalent in hospitals [1]. Digital pathology not only aids collaboration between pathologists but also opens up a realm of possibilities for image analysis tools. Such tools can introduce efficiencies into pathology workflows by supporting pathologists in their current decision making. Additionally, they can be more sensitive to novel histopathology features, driving insights not achievable with manual interpretation [2].

The potential for artificial intelligence (AI) tools to be incorporated into the digital pathology workflow has been widely acknowledged. In 2021, a European consortium set up the BIGPICTURE project aiming to establish the largest database of pathology images to date in order to accelerate AI progress in the field. Breast cancer is one indication where significant progress in this space has been made. Breast cancer is an attractive indication for such a solution for several reasons: first, breast cancer has a high disease burden and affects an estimated 2.3 million new patients worldwide each year [3]. Second, the importance of histopathology to a patient's disease characterization for treatment and prognosis motivates advances in AI imaging tools. Finally, the complexity of interpretation drives the need for tools that improve reliability and reduce variability. [4]

There have been a select number of commercial releases of AI based digital pathology tools for breast cancer for use in the clinical setting along with others confined to research use only. For example, Owkin's RlapsRisk BC [5] predicts relapse risk in early stage breast cancer, Paige's tool detects breast cancer metastases in lymph nodes [6], and Roche's research uses AI tools to detect breast cancer markers (Ki-67, ER and PR) [7]. However, all these prediction solutions have to our knowledge relied solely on the image data. This raises the important question of whether the incorporation of clinical data into models alongside slide image data could improve the performance of predictive tasks.

To investigate the promise of multimodal models, this paper looks to compare a multimodal model against unimodal models in the prediction of breast cancer staging. Breast cancer staging is chosen as a clinically relevant prediction task because it is routinely performed across all patients and is a key determinant for patient management (i.e., dictating follow-up test requirements and treatment plans) [8].

Pathologists stage breast cancers on a scale of 0 (pre-invasive) to IV (metastatic) based on several features. Such features include overall tumor morphology (size and shape), cellular features (cell differentiation, degree of nuclear pleomorphisms, mitotic count), receptors present (ER, PR, HER2), and level of metastasis to lymph nodes and distant organs. Additional tests are required to determine receptor status, and confirming metastases requires other non-local imaging

All authors contributed equally and submitted this paper for review on 2/1/2023. This work was supported in part by Nightingale Open Science.

Hannah Bibby is with the Biomedical Informatics Department, Harvard Medical School, MA 02115 USA (e-mail: hbibby@hms.harvard.edu).

Tong Ding is with the Biostatistics Department, Harvard University, MA 02138 USA (e-mail: tong_ding@hsph.harvard.edu).

Sharon Jiang is with the Electrical Engineering and Computer Science Department, Massachusetts Institute of Technology, MA 02139 USA (e-mail: jiangs@mit.edu).

Charlie Lu is with the MIT Media Lab, Massachusetts Institute of Technology, MA 02139 USA (e-mail: luchar@mit.edu).

Chenwei Wu is with the Data Science Department, Harvard University, MA 02138 USA (e-mail: chenweiwu@g.harvard.edu).

Senthil Nachimuthu, MD is the Chief Medical Officer at Nightingale Open Science, Chicago, IL USA (e-mail: senthiln@nightingalescience.org).

techniques [8]. Although pathologists cannot stage a patient from a biopsy slide alone, they can tell with some level of confidence whether a cancer might be a low (0-II) or high stage (III-IV) from the biopsy slides. This presents an opportunity to examine whether AI tools can more confidently identify stage with incomplete information.

Additionally, there is scientific rationale for a multimodal model incorporating clinical data being superior to a unimodal model: metastatic risk varies between patients and may not be driven solely by tumor characteristics. Age related cellular changes in the tumor micro-environment have been demonstrated to play a role in tumor progression and metastatic risk [9]. Comorbidities have been proposed to have adverse effects on tumor biology [10]. Race has also been shown to be implicated in metastatic risk, with some studies suggesting black women are more likely to have aggressive forms of the disease [11].

Thus, there are many possible clinical benefits to developing an accurate staging prediction model. First, a model can reduce the number of man hours required to make a diagnosis. Second, a model can detect evidence of metastasis with a greater sensitivity than a pathologist is currently able to and can therefore promote more effective patient management.

II. Related Work

Deep learning models use artificial neural networks to extract non-linear, representative features and thus can learn complex relationships from an enormous amount of high-dimensional data. These methods have emerged as a promising avenue to transform cancer care due to the increasing availability of multiple data types from histopathology slides and electronic health records. Combining these data types via multimodal learning can enhance the predictive performance of deep learning models. Multimodal networks consist of two main components: creating representations that contain dense meaningful features of the input and a mathematical method to combine representations from modalities. A prevalent method used for the representation learning task uses auto-encoders (AE). The AE architecture contains an encoder and decoder. The encoder creates a representation vector of a lower dimension than the input, and the decoder reconstructs the original input using this low-dimensional representation [12]. With this framework, the encoder learns meaningful features that can generalize well, allowing deep learning models to integrate different modalities (e.g., medical images, clinical data) into a single end-to-end model. Few models have been developed in practice to integrate both imaging and clinical data. One study found that fusion models of CT images and electronic health records accurately classified Pulmonary Embolism cases [13]. This paper builds on these promising multimodal model results for a multiclass classification task.

TABLE I
Number of staged slides in the Nightingale dataset (unbalanced stages)

Stage	Train	Validation	Test
0	6685	962	973
I	18665	1755	2159
II	9112	839	895
III	2374	196	262
IV	485	67	171
Total	37321	3819	4460

In order for deep learning methods to be widely implemented in the clinical domain, they need to provide interpretable explanations to clinicians. Most deep learning models have limited interpretability since it is difficult to understand how millions of parameters interact in a neural network [12]. Specifically for medical images, interpretable explanations entail identifying salient localized regions used to make predictions [14]. We will introduce methods that incorporate interpretability to increase trust in the model and assist clinicians in decision-making when quantitative performance may be biased.

III. Methods

A. Dataset

A major barrier to deep learning applications in oncology is the need for large amounts of training data to develop generalizable models. For breast cancer, previous datasets linking biopsy images to patient outcomes have been far smaller than what is needed to apply modern deep learning approaches. Recently, Nightingale Open Science has filled this void by releasing a public dataset linking large amounts of histopathological and clinical structured data [15].

The Nightingale Open Science Breast Cancer Biopsy dataset contains 45,600 staged whole slide images (WSI) from biopsies and patient data from 4,200 cases collected between 2014 and 2020 from the Providence Cancer Institute in Portland, Oregon. The data was provided in two formats: WSI breast biopsy slides at 40x magnification and CSV files containing structured patient features and ground truth staging. Structured data features include: comorbidity conditions diagnosed up to two years prior to biopsy (e.g., dementia, pulmonary disease, diabetes), social determinants of health (e.g., BMI, tobacco use), and demographic information (e.g., age, race, ethnicity). The ground truth data for the slide stage is found via the Providence cancer registry for the patient during the year of the biopsy [16]. The train, validation, and test set splits are provided in Table I. Slides were split such that the same patient's slides were assigned to the same split.

The number of slides per stage is unbalanced, so we created five randomly split datasets with balanced classes for additional evaluation. To do this, we use a standard random undersampling approach (i.e., removing samples from majority groups) [17]. In this dataset, the least frequent stage in the total dataset is stage 4, which has 723 slides. Each dataset is split such that the number of training slides per stage is $80\% \pm 5\%$ of the total number of stage 4 slides, and the number of validation and test slides per stage are each $10\% \pm 2.5\%$ of the total number of stage 4 slides. Each patient can have multiple biopsies and a variable number of slides, so this modification results in a more balanced distribution while maintaining that one patient's slides are in the same split. The number of slides per stage after balancing across stages is shown in Table II.

TABLE II
NUMBER OF SLIDES PER STAGE IN RANDOMLY SPLIT DATASETS (BALANCED STAGES)

Dataset	Train	Validation	Test
1	565	72	86
2	578	72	73
3	555	89	79
4	577	78	68
5	586	64	73

B. Baseline: Supervised Learning with Structured Metadata

As an initial baseline, we developed an approach that only uses the associated metadata of the slide images but not the images themselves. Specifically, we used the following attributes as features: 'bmi', 'tobacco', 'dementia', 'peripheral_vascular_disease', 'pulmonary_disease', 'liver_disease', 'diabetes', 'cerebral_vascular_accident', 'congestive_heart_failure', 'diabetes_complications', 'peptic_ulcer', 'severe_liver_disease', 'connective_tissue_disorder', 'acute_myocardial_infarction', 'renal_disease', 'hiv', 'paraplegia', 'race', 'ethnicity', and 'birth_dt'. As a preprocessing step, we normalized numeric attributes such as BMI and one hot encoded categorical attributes such as ethnicity. To handle missing data, we drop rows with one or more missing attribute values. We also filtered out rows with missing stage labels.

For modeling approaches, we trained three types of classifiers: multi-layered perceptron (MLP), gradient-boosted trees (GBT), and support vector machine (SVM). We implemented the MLP in the Jax deep learning framework and used 3 layers with 100 hidden units, ReLU activations, and weighted cross entropy loss (stages 0-IV weighted accordingly `[0.9, 0.5, 0.7, 1.5, 1.9]`). We used the Adam optimizer with a learning rate of $1e^3$. For the GBT model, we use the XGBoost framework trained with multiclass softmax loss and the following hyperparameters: 'max_depth'=5, 'eta'=0.9, and 'num_round'=20. For the SVM classifier, we use the Sklearn framework with the default parameters such as RBF kernel and squared $l2$ regularization penalty.

C. WSI Processing

We used the public CLAM repository [14] for WSI processing. For segmentation, each WSI was read at a 64x downscaled resolution and converted to the HSV color space. The tissue regions were then identified in the image by thresholding the saturation channel after applying median blurring for smoothing. We used morphological closing to fill in any gaps or holes in the mask. Next, patches of size 256 x 256 from all tissue areas were identified in the image at the 20x equivalent pyramid level without overlap. We used a ResNet-50 model [18] pre-trained on ImageNet as an encoder to convert each patch into a 1024-dimensional feature vector by applying spatial average pooling after the third residual block. To make the process faster, we used multiple GPUs in parallel with a batch size of 256 per GPU.

D. Image Modality: Attention-based Multiple Instance Learning

To predict cancer staging from WSIs, we use an attention-based multiple instance learning (AMIL) algorithm, which was originally designed for weakly supervised classification [14]. Under the multiple instance learning framework, each gigapixel WSI is divided into smaller regions and treated as a collection (bag) of patches (instances) with a corresponding slide-level label used for training. After processing the WSIs, each WSI bag is represented by a $M_i \times C$ matrix tensor, where M_i is the number of patches (bag size) and varies between slides, and C is the feature dimension (which equals 1024 for the ResNet50 encoder we used).

The model has 3 components: the projection layer (f_p), the attention module (f_{attn}), and the prediction layer (f_{pred}). The extracted patch-level features of each bag $H \in \mathbb{R}^{M_i \times 1024}$ were first projected into more compact 512-dimensional embeddings by the projection layer f_p with weights $W_{proj} \in \mathbb{R}^{512 \times 1024}$ and bias $b_{proj} \in \mathbb{R}^{512}$. A gated attention module then scored each region for its relevance to slide-level staging prediction. We use attention-pooling [19] when aggregating patch features in the WSI so that regions with high attention scores contribute more to the patient-level representation compared to regions with low scores. Specifically, the gated attention module f_{attn} consists of three fully connected layers with weights $U_a \in \mathbb{R}^{256 \times 512}$, $V_a \in \mathbb{R}^{256 \times 512}$, and $W_a \in \mathbb{R}^{1 \times 512}$. For $patch_i$ in a slide, the features learned by the projection layer are $h_i \in \mathbb{R}^{512}$, and its attention score can be obtained by:

$$a_i = \frac{e^{W_a(\tanh(V_a h_i^T) \odot sigmoid(U_a h_i^T))}}{\sum_{m=1}^{M_i} e^{W_a(\tanh(V_a h_m^T) \odot sigmoid(U_a h_m^T))}}$$

Using the computed attention scores as weight coefficients, the gated attention pooling module aggregates the patch features by a weighted sum to obtain the slide-level representation $h_{slide} \in \mathbb{R}^{512}$ by:

$$h_{slide} = \sum_{m=1}^{M} a_m h_m$$

Lastly, the stage of the slide is decided by a prediction layer f_{pred} with weights $W_{pred} \in \mathbb{R}^{5 \times 512}$ and bias $b_{pred} \in \mathbb{R}^5$, where predicted stage $\hat{stage} = softmax(f_{pred}(h_{slide}))$.

E. Metadata Modality: Self-Normalized Networks

To learn the unimodal features from metadata, we consider two variants of deep neural networks (DNNs): self-normalized networks (SNNs) and traditional DNNs (for the sake of ablation studies). While DNNs utilize RELU activation, SNNs adopt scaled exponential linear units (SELU) as activation functions and automatically normalize the outputs of every layer to have a mean of zero and standard deviation of one [20]. Applied element-wise, the SELU function deals with an input x as follows:

$$SELU(x) = scale * (\max(0, x) + \min(0, \alpha * \exp(x) - 1))$$

We have a large number of metadata features and observe a significant drop in performance from train to test in our baseline supervised metadata-only model. This could be attributed to the traditional feed-forward DNN's proneness to overfitting on high-dimensional data and training instabilities from regularization techniques like stochastic gradient descent. Thus, we utilize a combination of SNNs and dropout to introduce more robustness into our metadata modality learning.

We design both our SNN and DNN model architecture to contain 2 hidden layers of 128 neurons each. We also add dropout with a probability of 0.1 at the end of every layer. The output of the fully connected layer learns a representation $h_{metadata} \in \mathbb{R}^{64}$.

F. Modality Fusion: Multimodal Learning

After the unimodal feature vectors are learned separately from our image component and metadata component, we apply a Kronecker product fusion controlled by gated attention and append a fully connected layer at the end to output staging classification probabilities. Before the fusion, we append one to $h_{metadata}$ and h_{slide} to preserve the unimodal information. Then, we apply the Kronecker product to multiply every neuron in $h_{metadata}$ with every neuron in h_{slide} to generate $h_{multimodal}$, which captures all cross-modality interaction between metadata and WSI's. A gated attention module is applied to both the slide and metadata features before fusion to control the expressivity of each modality and prevent noisy features. To prevent potential colinearity from one modality dominating the other, we compute the element-wise product of the $h_{unimodal}$ and $z_{attention}$ scores. The mathematical expression for this fusion process is as follows:

$$h_{multimodal} = \begin{bmatrix} h_{slides} * z \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_{metadata} * z \\ 1 \end{bmatrix}$$

The output $h_{multimodal} \in \mathbb{R}^{(512+1)\times(64+1)}$ goes through an additional fully connected fusion layer of $\mathbb{R}^{512}$ to generate a final $h_{pred} \in \mathbb{R}^{512}$.

Following our setup for image modality learning, we append a fully connected prediction layer f_{pred} with weights $W_{pred} \in \mathbb{R}^{5\times 512}$ and bias $b_{pred} \in \mathbb{R}^{5}$, where predicted stage $\hat{stage} = softmax(f_{pred}(h_{pred}))$ to generate predictions for cancer stages. The multimodal fusion architecture is shown in Figure 1.

IV. Results

A. Evaluation Metrics

To assess the performance of the task to predict breast cancer stage, we use quadratic weighted Cohen's Kappa coefficient $\kappa \in [-1, 1]$, which is commonly used to measure agreement between two raters. In Equation 1, P_o is the proportion of observed agreement and P_e is the agreement due to random chance.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

The Kappa coefficient is an appropriate metric because the Nightingale dataset is imbalanced. Furthermore, the prediction variable (cancer stage) is inherently an ordinal value with different costs associated with corresponding errors (e.g., predicting stage I is much worse than predicting stage III when the patient actually has stage IV cancer).

TABLE III

Structured Metadata Results

Model	Cohen's Kappa	Accuracy	F1
MLP	0.236	0.487	0.448
GBT	0.206	0.443	0.430
SVM	0.204	0.530	0.462

TABLE IV

Confusion matrix of GBT trained only on metadata (columns are predicted classes and rows are actual classes)

	Stage 0	Stage I	Stage II	Stage III	Stage IV
Stage 0	187	317	114	0	0
Stage I	85	1115	390	87	0
Stage II	33	447	44	27	0
Stage III	0	67	67	34	0
Stage IV	0	61	36	4	0

Since the dataset is imbalanced across stages, we also evaluate using macro-average AUC in a one vs. rest manner to gauge whether the model can properly distinguish classes.

B. Structured Metadata

Quantitative results for supervised learning of metadata features on unbalanced data are shown in Table III. There is similar performance in terms of Kappa scores across different models (MLP, GBT, and SVM). All models underperform on later cancer stages, likely due to a large class imbalance in data (refer to the confusion matrix in Table IV). Quantitative results on balanced datasets are different compared to those on unbalanced data but still, show poor performance. These metrics are expected and serve to highlight the superior performance of multimodal learning in later sections.

The distribution of mean Shapley values for stages 0, I, II, III, and IV are provided in Figures 2, 3, 4, 5, and 6, respectively. These distributions show that continuous attributes such as ‘birth_dt’ and ‘BMI’ have more impact on stage prediction than categorical attributes.

C. Image-Only

Quantitative results for all models (including the metadata only SVM, attention-based multiple instance learning, and multimodal fusion models) on balanced datasets are shown in Table VI. Compared to the metadata-only SVM model, the image-only model performs much better and achieves a Kappa score of 0.361 ± 0.07 from a 5-fold cross validation. Table V shows the AMIL confusion matrix. Similarly to the metadata models, the image-only AMIL model also underperforms on minority cancer stages. This is likely due to the difficulties of identifying cancer stages using histopathology images alone. While pathology images can provide valuable information about the size and location of a tumor, other factors (e.g., the presence of cancer cells in the lymph nodes and the spread of cancer to other parts of the body) are also important

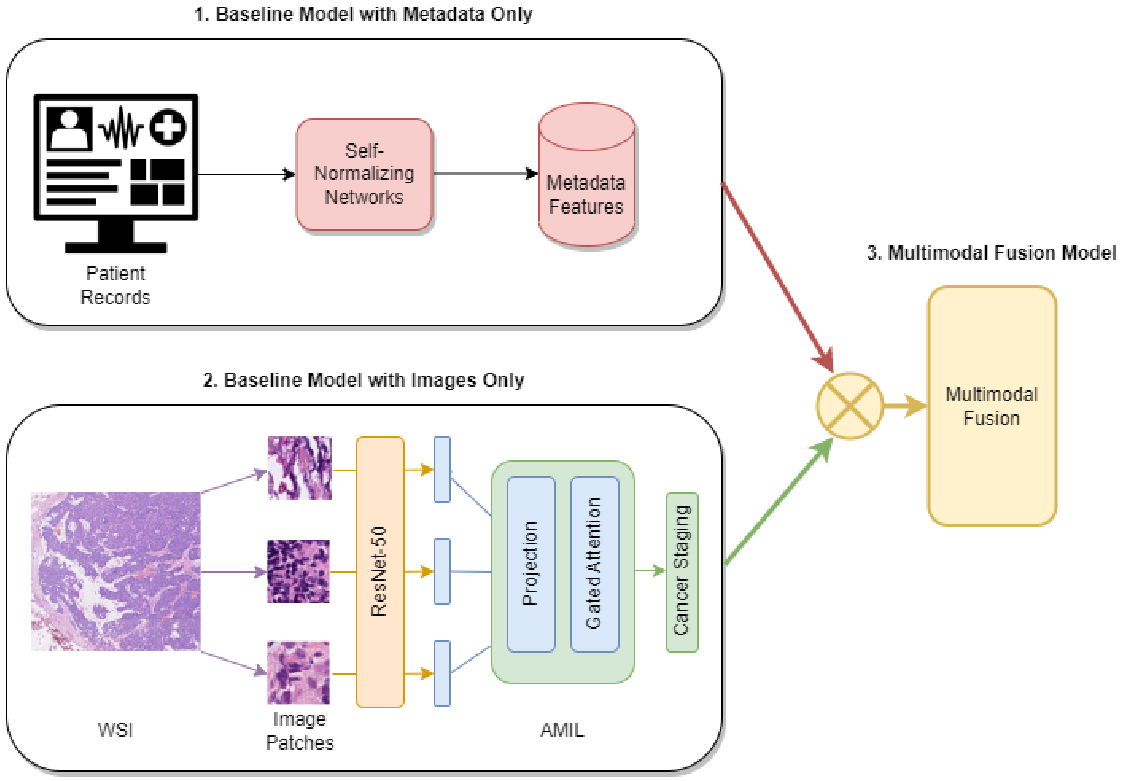
1. Baseline Model with Metadata Only
Patient Records
Self-Normalizing Networks
Metadata Features
2. Baseline Model with Images Only
WSI
Image Patches
ResNet-50
Projection
Gated Attention
AMIL
Cancer Staging
3. Multimodal Fusion Model
Multimodal Fusion

Fig. 1. Multimodal fusion architecture. Figure partially adapted from [21].

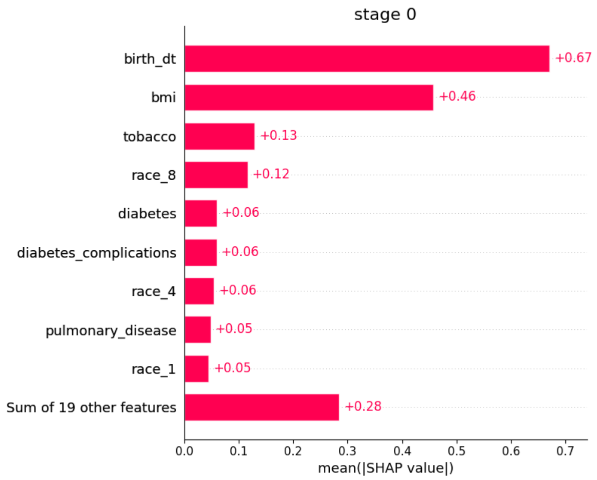

Fig. 2. Distribution of mean Shapley values for stage 0

in determining the stage, especially for late stages, which motivated us to integrate metadata for a more accurate staging prediction.

In terms of interpretation, we plot the highest attended patches via AMIL for all stages as shown in Figure 7. We consulted with a clinical expert to confirm that these patches highlighted relevant regions for staging. The clinical expert determined that the patches from stages II and III have a higher cellular grade than those from stage I. Cellular grading is one component of staging determination, and higher grades are more likely to metastasize compared to lower grades.

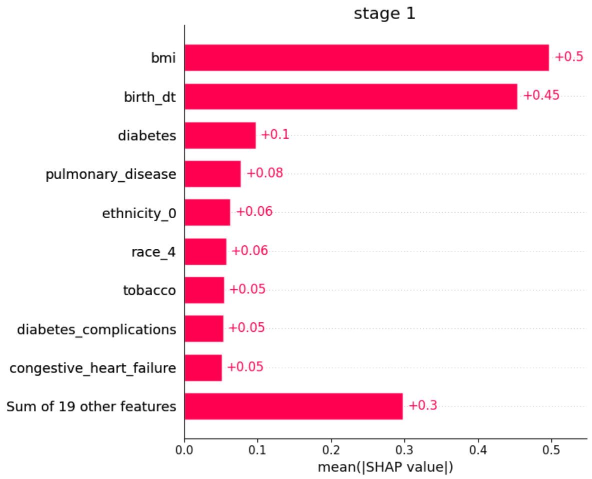

Fig. 3. Distribution of mean Shapley values for stage I

TABLE V

CONFUSION MATRIX OF THE IMAGE-ONLY AMIL MODEL (COLUMNS ARE PREDICTED CLASSES AND ROWS ARE ACTUAL CLASSES)

	Stage 0	Stage I	Stage II	Stage III	Stage IV
Stage 0	46	16	2	2	2
Stage I	7	41	12	8	0
Stage II	20	19	17	11	1
Stage III	17	10	1	32	8
Stage IV	12	8	4	15	29

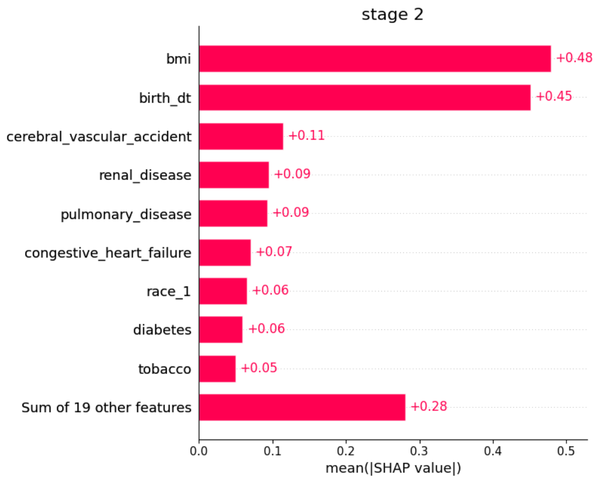

Fig. 4. Distribution of mean Shapley values for stage II

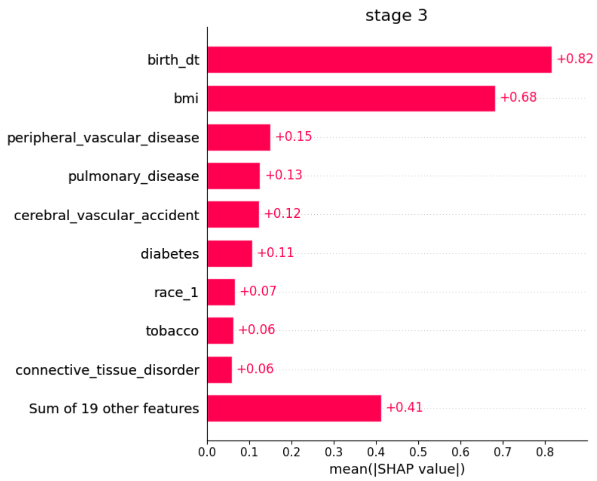

Fig. 5. Distribution of mean Shapley values for stage III

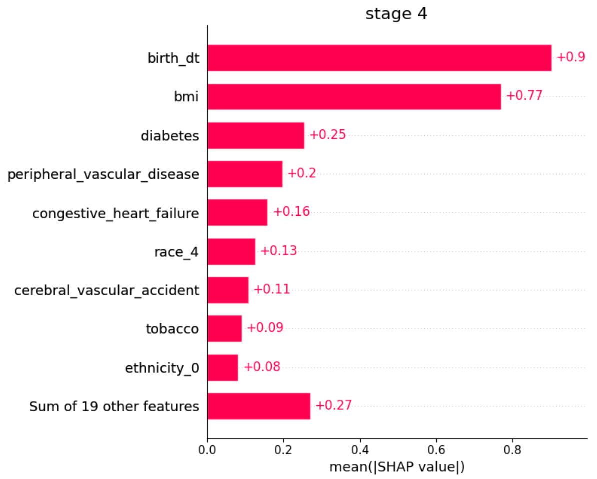

Fig. 6. Distribution of mean Shapley values for stage IV

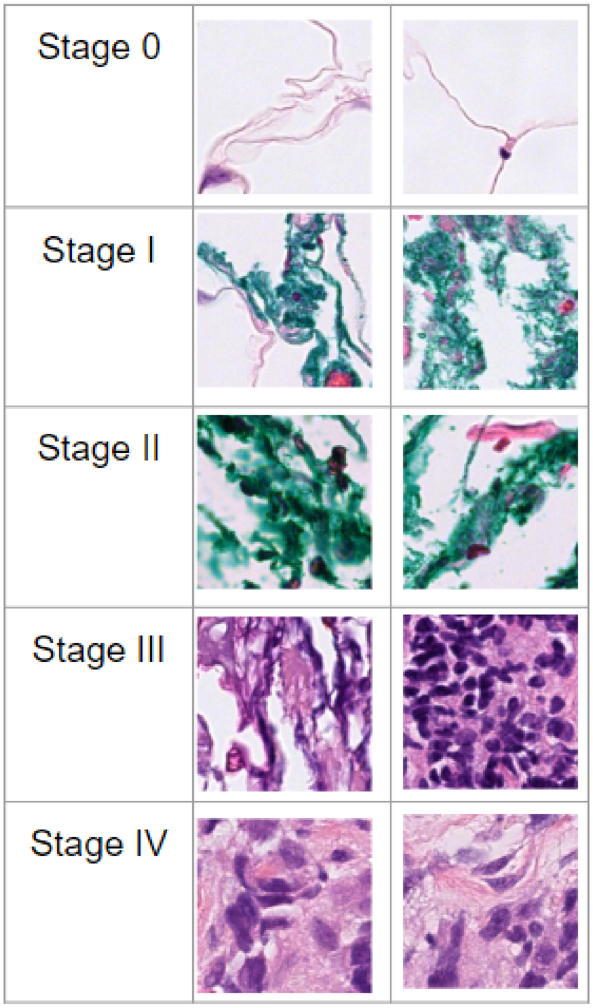

Fig. 7. Most important patches for all stages via AMIL. Each stage shows 2 patches with the highest attention score for different slides.

TABLE VI
QUANTITATIVE RESULTS FOR ALL MODELS

Model	Cohen's Kappa	AUC
Metadata only SVM	0.204	0.538
AMIL	0.361 ± 0.070	0.658 ± 0.100
Multimodal-DNN (Avg)	0.669 ± 0.113	0.767 ± 0.136
Multimodal-SNN (Avg)	0.706 ± 0.045	0.799 ± 0.087
Multimodal-DNN (Best)	0.842	0.897
Multimodal-SNN (Best)	0.838	0.911

D. Multimodal Learning

As described in Table VI, the two variants of the multimodal model outperform the unimodal models by a significant margin from a 5-fold cross validation. The Multimodal-DNN model has an average Cohen's Kappa of 0.669 compared to the image-only model's 0.361 and metadata-only model's 0.204. The Multimodal-DNN model has an average AUC of 0.767 compared to the image-only model's 0.658 and metadata-only model's 0.538. Multimodal-SNN is the best overall model, with a substantial increase in average Cohen Kappa of 0.706 and average AUC of 0.799. The self-normalizing feature helps achieve a smaller variance in model performance across 5 folds. The best fold model has Cohen's Kappa as high as

TABLE VII
CONFUSION MATRIX FOR MULTIMODAL-SNN MODEL (COLUMNS ARE PREDICTED CLASSES AND ROWS ARE ACTUAL CLASSES)

	Stage 0	Stage I	Stage II	Stage III	Stage IV
Stage 0	42	31	0	0	0
Stage I	0	73	0	0	0
Stage II	50	19	4	0	0
Stage III	13	0	0	55	5
Stage IV	0	0	0	2	72

0.842 from the DNN variant and AUC of 0.911 from the SNN variant.

Table VII shows the confusion matrix for the Multimodal-SNN model. This model correctly predicts stage I, III, and IV most of the time. However, it underperforms slightly on stage 0 and misses prediction of stage II entirely. Further analysis is needed to address this limitation and ensure robust prediction of the Multimodal-SNN model.

Figure 8 shows the patches with highest attention score for all stages. Similarly to the image-only model, we see that the multimodal SNN focuses on more higher cellular grades in the later stages (especially in stage III) compared to earlier stages like 0 and I. Thus, these patches demonstrate the interpretability of our multimodal learning approach.

V. DISCUSSION

The superior predictive performance of a model incorporating both clinical and imaging data demonstrates the potential multimodal models may hold for integration into the digital pathology workflow. Of particular relevance in this study was the incorporation of BMI and age data, as demonstrated by the higher Shapley values for these variables. This finding confirms previous scientific literature suggesting BMI and age are associated with a higher risk of metastasis [22], [23]. The Cohen's Kappa scores achieved by both the metadata DNN and metadata SNN models, 0.669 ± 0.113 and 0.706 ± 0.136 respectively, are promising for the clinical utility of such a model. The scores indicate that the model is able to successfully differentiate between stages. Given that pathologists will often only conclude from a slide biopsy if a tumor is low-stage (I-II) or high-stage (III-IV), the fact that the model can discern stages within these two classes is very positive. In addition, consultation with a clinician on the patches with the highest attention scores also confirmed, in a preliminary manner, that the model was using relevant clinical features to ground the prediction. Consulting with a larger pool of pathologists about the features used by the model would provide further confidence in the robustness of the methodology. In future consultations, we could potentially discuss merging stages into lower risk (I-II) and higher risk (III-IV) to simplify the multiclass classification task.

Further analytical steps required to confirm the model's ability to infer the presence of metastasis from a slide would be to perform sub-group predictive performance analyses. Possible sub-groups of interest may stratify on race, age, social determinants of health, ground truth staging, and presence of breast cancer receptors (ER, PR, HER2). Further hyperparameter tuning and using larger pathology images for pretraining models instead of ImageNet are expected to be able to improve these prediction metrics.

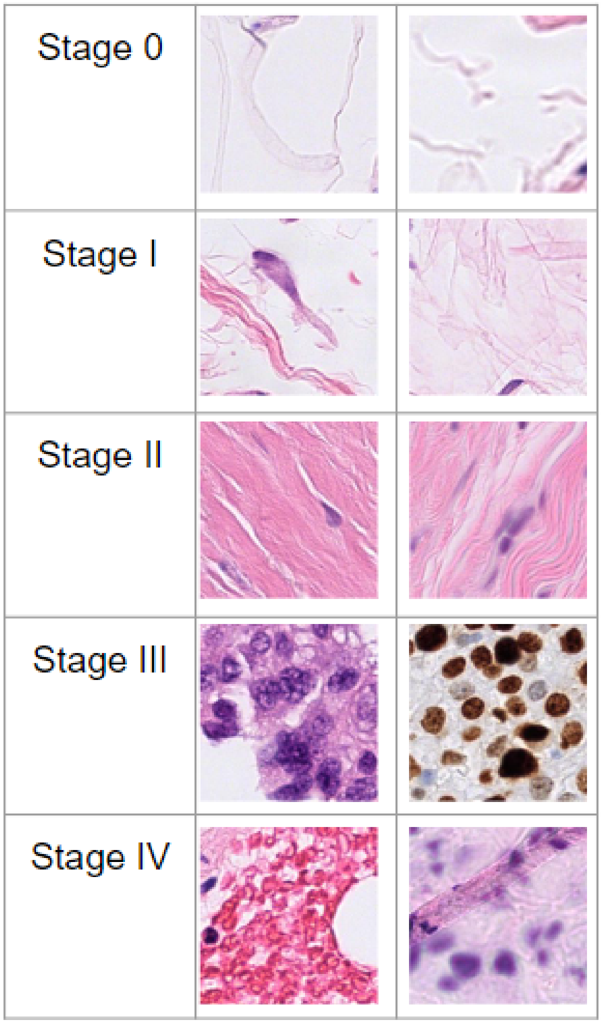

Fig. 8. Most important patches for all stages via Multimodal-SNN. Each stage shows 2 patches with the highest attention score for different slides.

In addition to improving performance, future work entails considering the real-world feasibility of model deployment. Digital pathology workflows are somewhat limited to academic centers, which limits their use in settings without high upfront investment in computational infrastructure. Additionally, at centers with a digital pathology workflow, the multimodal nature of this model may present its own challenges, as the image management system where biopsy images are stored may not be the same as the system for electronic health records.

Another potential hurdle to deployment involves generalizing models. The Nightingale dataset is currently restricted to one institution, and we look to further our collaboration to collect data from a variety of geographical settings and patient populations. With the increasing prevalence of digital pathology and the expansion of data-sharing networks in this space, we are confident that there is an enormous potential for this work to expand and be leveraged on other datasets.

Acknowledgment

The authors thank Nightingale Open Science for the curation of patient biopsy slides and financial support to use their computing platform in developing our deep learning pipeline. This manuscript was composed by participants in the HST.953 Clinical Data Learning, Visualization, and Deployments course at the Massachusetts Institute of Technology in Fall 2022.

References

[1] S. Betmouni, ”Diagnostic digital pathology implementation: Learning from the digital health experience,” (in eng), Digit Health, vol. 7, p. 20552076211020240, 2021 Jan-Dec 2021, doi: 10.1177/20552076211020240.

[2] V. Baxi, R. Edwards, M. Montalto, and S. Saha, ”Digital pathology and artificial intelligence in translational medicine and clinical practice,” (in eng), Mod Pathol, vol. 35, no. 1, pp. 23-32, 01 2022, doi: 10.1038/s41379-021-00919-2.

[3] H. Sung et al., ”Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” (in eng), CA Cancer J Clin, vol. 71, no. 3, pp. 209-249, May 2021, doi: 10.3322/caac.21660.

[4] A. Ibrahim et al., ”Artificial intelligence in digital breast pathology: Techniques and applications,” vol. Volume 49, ed. The Breast, 2020, pp. Pages 267-273.

[5] GArberis, I.J, Gaury, V., Drubay, D. et al. Blind validation of an AI-based tool for predicting distant relapse from breast cancer HES stained slides. Poster presented at: European Society for Medical Oncology (ESMO); May 9th - 13th 2022; Paris France.

[6] https://www.pathologynews.com/computational-pathology-ai/paige-earns-ce-ivd-and-ukca-marks-for-clinical-ai-application-to-detect-breast-cancer-metastases-in-lymph-nodes/

[7] https://diagnostics.roche.com/global/en/news-listing/2021/roche-announces-release-of-its-newest-ai-based-digital-pathology-algorithms-to-aid-pathologists-in-evaluation-breast-cancer-markers-ki67-er-pr.html

[8] https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-grades.html

[9] M. Fane and A. T. Weeraratna, ”How the ageing microenvironment influences tumour progression,” (in eng), Nat Rev Cancer, vol. 20, no. 2, pp. 89-106, Feb 2020, doi: 10.1038/s41568-019-0222-9.

[10] G. Panigrahi and S. Ambs, ”How Comorbidities Shape Cancer Biology and Survival,” (in eng), Trends Cancer, vol. 7, no. 6, pp. 488-495, Jun 2021, doi: 10.1016/j.trecan.2020.12.010.

[11] T. Akinyemiju, S. Sakhuja, J. Waterbor, M. Pisu, and S. F. Altekruse, ”Racial/ethnic disparities in de novo metastases sites and survival outcomes for patients with primary breast, colorectal, and prostate cancer,” (in eng), Cancer Med, vol. 7, no. 4, pp. 1183-1193, Apr 2018, doi: 10.1002/cam4.1322.

[12] Tran, K.A., Kondrashova, O., Bradley, A. et al. Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med 13, 152 (2021). https://doi.org/10.1186/s13073-021-00968-x.

[13] Huang, S.C., Pareek, A., Zamanian, R., Banerjee, I. and Lungren, M.P. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. Sci Rep 10, 22147 (2020). https://doi.org/10.1038/s41598-020-78888-w.

[14] Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., and Mahmood F. Data Efficient and Weakly Supervised Computational Pathology on Whole Slide Images. arXiv (2020). https://doi.org/10.48550/arXiv.2004.09666.

[15] Mullainathan, S., and Obermeyer, Z. (2022). Solving medicine’s data bottleneck: Nightingale Open Science. Nature Medicine, 28(5), 897–899. https://doi.org/10.1038/s41591-022-01804-4.

[16] Bifulco, C., Piening, B., Bower, T., Robicsek, A., Weerasinghe, R., Lee, S., Foster, N., Juergens, N., Risley, J., Haynes, K., and Obermeyer, Z. (2021). Identifying high-risk breast cancer using digital pathology images [Data set]. Nightingale Open Science. https://doi.org/10.48815/N5159B.

[17] Krawczyk, B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell 5, 221–232 (2016). https://doi.org/10.1007/s13748-016-0094-0.

[18] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[19] Ilse, M., Tomczak, J.M., and Welling, M. Attention-based Deep Multiple Instance Learning. ICML (2018). https://doi.org/10.48550/arXiv.1802.04712.

[20] Klambauer, G., Unterthiner, T., Mayr, A. and Hochreiter, S. Self-Normalizing Neural Networks. Advances in Neural Information Processing Systems 30 (NIPS 2017).

[21] Chen, R.J., Lu, M.Y., Williamson, D.F., Chen, T.Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B. and Mahmood, F., 2022. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell, 40(8), pp.865-878.

[22] White, M.C., Holman, D.M., Boehm, J.E., Peipins, L.A., Grossman, M., and Henley, S.J. Age and Cancer Risk, Am J Prev Med (2021). https://doi.org/10.1016/j.amepre.2013.10.029.

[23] Vucenik, I. and Stains, J.P. Obesity and cancer risk: evidence, mechanisms, and recommendations. Annals of the New York Academy of Sciences 1271, 37-43 (2012). https://doi.org/10.1111/j.1749-6632.2012.06750.x.