Hierarchical Multi-scale Transformer for Breast Tumor Staging with Visual Interpretability and Risk Stratification

Satyanarayana Reddy Beram [1,*], R Lalchhanhima [1], and Ksh. Robert Singh [2]

[1] Department of Information Technology, Mizoram University, Mizoram, India
[2] Department of Electrical Engineering, Mizoram University, Mizoram, India
Email: mzu22007898@mzu.edu.in (S.R.B.); chhana.mizo@gmail.com (R.L.); robert_kits@yahoo.co.in (K.R.S.)
*Corresponding author

Abstract—**Accurate tumor staging and risk stratification in breast cancer are critical for guiding treatment decisions. Traditional decision-making methods rely on Whole Slide Images (WSIs) analysis, which is labor-intensive and subject to inter-observer variability. To address these challenges in tumor staging, we propose a modular and interpretable deep learning framework for automated tumor staging through multi-resolution histopathological analysis. Our Hierarchical Multi-Scale Transformer (HMS-T) integrates the Vision Transformers (ViTs) operating at 5×, 10×, and 20× magnifications to capture both the cellular and architectural features. In addition, a novel cross-scale attention fusion module combines these multi-scale resolution representations, for enabling robust prediction of the American Joint Committee on Cancer (AJCC) stage group labels recorded in the clinical data from primary-tumor WSIs. Trained on 1092 patients from The Cancer Genome Atlas–Breast Invasive Carcinoma (TCGA-BRCA) cohort, our HMS-T achieves state-of-the-art performance with a staging accuracy of 91.5%, a macro F1-Score of 0.89, and a quadratic-weighted kappa of 0.92, which demonstrates the strong agreement with pathological standards. Moreover, our model's attention maps exhibit high spatial interpretability, aligning closely with expert-annotated regions (dice score = 0.81). Ultimately, we introduced a lightweight clinical extension for the preliminary survival risk stratification, achieving a concordance index of 0.74, thereby bridging toward full prognostic modeling. By combining high performance with transparent decision-making, HMS-T represents a significant advancement toward deployable Artificial Intelligence (AI)-assisted pathology tools for breast cancer.**

Keywords—**multi-scale histopathology, vision transformers, cross-scale attention fusion, American Joint Committee on Cancer (AJCC) tumour staging, survival risk stratification, Whole Slide Images (WSIs), deep learning**

I. Introduction

A. Background and Motivation

Breast cancer remains one of the most prevalent malignancies worldwide, with an estimated 2.3 million new cases and 685,000 deaths reported globally in 2020 according to the World Health Organization (WHO) [1]. Early and accurate tumor stage prediction plays a pivotal role in clinical decision-making, which directly influences the treatment strategies, prognostic estimation, and overall patient outcomes [2]. The American Joint Committee on Cancer (AJCC) staging system, which incorporates tumor size, nodal involvement, and metastatic spread, serves as the clinical gold standard for categorizing disease progression and tailoring individualized therapeutic approaches [3].

Despite its importance to care pathways, the traditional process of tumor staging heavily depends on the manual interpretation of Whole Slide Images (WSIs) by expert pathologists [4]. This dependency introduces various challenges, such as human variability in stage decision making and the intrinsic complexity of histopathological structures. Inter-observer variability, particularly in the assessment of mitotic counts and architectural differentiation, remains a documented concern in certain diagnostic categories [4, 5]. Furthermore, the exhaustive nature of manually reviewing gigapixel-scale WSIs renders the process labor-intensive, time-consuming, and susceptible to fatigue-induced errors [5]. These challenges collectively underscore the urgent need for automated, standardized, and reliable solutions to assist or augment human expertise in tumor staging.

Recent technological advancements have opened new avenues for addressing these challenges. Digital pathology, characterized by the digitization of histopathological slides into WSIs, has paved the way for computational image analysis at an unprecedented scale [6]. Concurrently, the emergence of Artificial Intelligence (AI) and deep learning methods, particularly those leveraging Vision Transformers (ViTs), offers high transformative potential for medical image analysis [7]. Unlike conventional Convolutional Neural Networks (CNNs) [6], ViTs possess the essential capabilities to model the long-range dependencies and global contextual information. These attributes are mainly valuable for

Manuscript received September 18, 2025; revised November 7, 2025; accepted December 10, 2025; published March 26, 2026.

doi: 10.18178/joig.14.2.208-229

analyzing the complex, multi-scale patterns which are inherent in histopathology. The application of AI-driven WSI analysis promises not only to enhance the diagnostic reproducibility and efficiency but also to democratize access to high-quality pathology services by reducing the reliance on scarce human expertise.

B. Limitations of Existing Methods

While the application of CNN-based deep learning models achieved significant improvements in WSI analysis tasks, they suffer from fundamental limitations. CNNs are excellent at extracting the local features through hierarchical convolutional operations [8]. But they inherently struggle to capture the long-range spatial relationships across the large tissue areas due to their localized receptive fields. As a result, the CNN-based models often perform patch-level analysis in isolation, without adequately modeling the broader tissue architecture, which is crucial for accurate tumor staging. This fragmentation of spatial understanding limits the CNNs' ability to make them as holistic diagnostic inferences.

The advent of ViTs introduced a promising alternative to CNNs by enabling global context modelling through self-attention mechanisms [7]. Early studies applied the ViTs to histopathology and demonstrated notable gains in performance, particularly in tasks such as tumor subtype classifications and biomarker status predictions [7, 9]. Although they achieved better performance in histopathology analysis, several challenges still remain. Many transformer-based implementations operate at a single resolution, failing to leverage the critical multi-scale information that pathologists routinely consider during the manual assessments [9, 10]. Furthermore, the computational complexity associated with processing the gigapixel-scale WSIs using the transformers lead to careful architectural and optimization strategies to ensure the feasibility. Moreover, although the ViTs offer some degree of interpretability through attention visualization, it's still suffering from a lack of systematic validation against expert annotations, which raises concerns about clinical trustworthiness [7].

Meanwhile, multi-scale learning strategies have been proposed to address the limitations of single-resolution models [11–13]. These approaches attempt to integrate the information across different magnifications and combine the cellular-level detail with broader tissue architecture. Most of the existing multi-scale models employ static feature fusion techniques, which simply concatenate or aggregate the representations without dynamic reasoning across scales [12]. This simplistic integration often results in the loss of fine-grained features or the dilution of salient global patterns. Additionally, explicit modelling of cross-resolution relationships, which is crucial for tasks requiring the nuanced histological interpretation, is often absent [13].

Interpretability remains another critical bottleneck for clinical translation [4]. While attention heatmaps generated by the AI models provide a window into decision-making processes, very few studies validate these attention maps against human-annotated ground truths [4, 14]. Without rigorous validation, there is a risk that models may depend on irrelevant or spurious features for decisions, undermining both performance and clinical confidence. Consequently, the lack of explainability and alignment with pathologists' reasoning continues to hinder widespread adoption of AI models in breast cancer staging [5, 9].

C. Research Gap

From the aforementioned limitations in histopathology staging, several clear research gaps are identified and the main motivation to the present study. At first, there is a pressing need for unified modeling frameworks that explicitly capture and integrate multi-scale contextual information [7]. Pathologists routinely navigate between 5×, 10×, and 20× magnifications to reconcile nuclear features, glandular structures, and tissue-level architecture. Current models, whether CNN-based or ViT-based, are inadequately simulated to this multi-resolution reasoning, leading to limiting their diagnostic reliability [8, 9]. A model that can simultaneously process and reason across these scales will hold the potential to more closely approximate human expertise.

Second, the validation of model interpretability remains insufficient [4]. Although numerous transformer-based models can produce visually attractive attention maps, it's quite rare to find the systematic quantitative validation that compares them to expert-annotated tumor regions. Aligning attention outputs with pathologist annotations using robust spatial similarity metrics, such as the Dice similarity coefficient, is essential to ensure that the models focus on biologically and clinically meaningful features [14].

Third, existing AI models tend to treat tumor staging and survival prediction as discrete, independent tasks. However, in clinical practice, staging and prognosis are deeply intertwined. Integrating these tasks within a unified architecture would not only enhance predictive performance but also provide more actionable insights for patient management. Specifically, a system that can stage the tumors while simultaneously offering preliminary survival risk estimation could greatly aid treatment stratification [15].

Fourth, extensive experiments on the TCGA-BRCA dataset demonstrate the performance of the proposed framework. HMS-T achieved results in staging accuracy, macro-averaged F1-Score, and quadratic-weighted kappa compared to several recent baselines.

Finally, there is a demand for modular and extendable architectures that can accommodate future expansions. As the field is improving towards multi-modal oncology (i.e., incorporating genomics, radiomics, and clinical metadata), models must be designed to flexibly integrate the new data streams without requiring total retraining. Furthermore, the scalability and computational efficiency are paramount for real-world deployment across the diverse clinical environments with variable infrastructure capabilities.

D. Objective and Contributions

To address these gaps, this study proposes a novel framework titled Hierarchical Multi-Scale Transformer

(HMS-T) for automated breast tumor staging with embedded visual interpretability and preliminary risk stratification capabilities. The primary objective of this study is to develop an interpretable, multi-scale vision transformer-based system that captures both fine-grained and global histological features. This helps to deliver the accurate and clinically meaningful approximation of AJCC stage group assignments from primary-tumor whole slide images.

The key contributions of this work are summarized as follows:

First, we introduce the HMS-T architecture, which incorporates dedicated vision transformer branches specialized for 5×, 10×, and 20× magnifications. These branches are integrated through a novel cross-scale attention fusion mechanism that dynamically models dependencies across resolutions.

Second, the framework emphasizes interpretable decision-making. Attention maps generated by the model are systematically validated against expert-annotated tumor regions using the Dice Similarity Coefficient (DSC), to ensure the model's predictions are grounded in biologically and clinically relevant features.

Third, we integrate an optional risk stratification module into the HMS-T architecture. This lightweight extension combines the fused histological embeddings with clinical metadata to predict the preliminary survival risk scores and is evaluated using the Concordance index (C-index).

Fourth, extensive experiments on the TCGA-BRCA dataset demonstrate the performance of the proposed framework. HMS-T achieved results in staging accuracy, macro-averaged F1-Score, and quadratic-weighted kappa compared to several recent baselines.

Our framework follows the standard patch-wise processing and slide-level aggregation paradigm, which is a benchmark design in current WSI analysis. Compared with the existing multi-scale ViT approaches, our novelty is how these transformer components are combined and analyzed for AJCC staging and prognosis. We use three scale-specific ViT branches together with an explicit class-token level cross-scale attention module. We performed the detailed scale-wise and component-wise ablations to understand the role of each magnification and the fusion block. HMS-T linked the same fused embedding to an auxiliary survival risk head and to quantitatively validated attention maps [8]. This makes HMS-T a practical and interpretable multi-task framework rather than a generic patch-wise classifier. Through these contributions, this work bridges the critical gaps in breast cancer staging automation and paves the way for more transparent, reliable, and comprehensive AI-assisted pathology systems.

II. Related Work

A. *Tumor Staging from Whole Slide Images (WSI)*

1) *Traditional pathology approaches*

Traditional tumor staging in breast cancer relies mainly on manual assessment of histopathological slides by expert pathologists. Lee *et al.* [16] outlined the standard use of the AJCC TNM staging system, which evaluates tumor size (T), nodal involvement (N), and distant metastasis (M), often supplemented by molecular marker evaluation. However, manual grading and biomarker scoring are highly susceptible to inter-observer variability. Ginter *et al.* [17] reported that moderate agreement among pathologists in histologic grading, with discrepancies in mitotic count assessments, is significantly contributing to variability in results. Similarly, Polónia and Caramelo [18] demonstrated only fair to moderate concordance in Human Epidermal Growth Factor Receptor 2 (HER2) scoring, especially in borderline cases.

The impact of such variability can be insightful, potentially altering the patient prognostic staging in approximately 20% of the cases [10].

Recent strategies such as digital image analysis and algorithm-assisted scoring have been introduced to improve reproducibility, as discussed by Capar *et al.* [4]. Nevertheless, despite advancements, manual pathology remains labor-intensive and inherently subjective, which emphasize the need for more standardized, automated staging approaches.

2) *Machine learning and Convolutional Neural Network (CNN)-based models*

The application of deep learning, particularly CNNs, has significantly transformed the tumor analysis from Whole Slide Images (WSIs). Hamida *et al.* [19] demonstrated the effectiveness of CNN-based models like ResNet and DenseNet for patch-level classification, while segmentation tasks commonly utilized U-Net and SegNet architectures. To mitigate the challenges posed by limited annotations in WSIs, Multiple Instance Learning (MIL) frameworks have gained prominence in recent times. Yao *et al.* [20] proposed treating slides as "bags" of patches, enabling training with slide-level labels without exhaustive manual annotations. Beyond patch-based learning, Lee *et al.* [21] incorporated the graph neural networks and attention-based pooling to capture the micro environmental features relevant for staging and survival analysis. Despite their success, the CNN-based models often struggle in capturing the long-range spatial dependencies. This limitation, combined with the computational burden of processing high-resolution WSIs, has accelerated the adoption of transformer-based architectures better suited for modeling global context.

3) *Transformer-based advances in Whole Slide Image (WSI) analysis*

Vision Transformers (ViTs) have emerged as a compelling alternative to CNNs for WSI analysis. Monjezi *et al.* [13] introduced RI-ViT, a hybrid CNN-transformer framework, achieving state-of-the-art breast cancer detection accuracy. Ayana *et al.* [22] demonstrated that ViTs could predict the HER2 expression status directly from H&E slides, by achieving an AUC of 0.92, thus potentially eliminating the need for immune-histochemical staining. Moreover, hybrid models combining ViTs with CNN backbones, as discussed by Qu *et al.* [23], leverage both local and global feature representations for improving the staging robustness across different magnifications. Progressive fine-tuning strategies, such as those proposed by Alruily *et al.* [24], further enhance ViT performance while reducing computational overhead. The integration of ViTs into clinical workflows holds the promise of enhancing diagnostic precision, workflow efficiency, and interpretability in breast cancer staging.

B. Multi-Scale Feature Learning in Histopathology

1) Importance of multi-resolution cues

Histopathological diagnosis inherently demands the integration of features observed across multiple magnifications. Sheikh *et al.* [25] emphasized that combining the fine cellular features (20× magnification) with broader tissue architecture (5×–10× magnifications) will significantly improves classification accuracy and diagnostic confidence. Pedersen *et al.* [26] demonstrated that cascaded multi-resolution networks improve tumor segmentation performance by refining tissue boundary delineations.

2) Existing multi-scale models

Early multi-scale learning strategies in histopathology focused on patch-based models. Kosaraju *et al.* [27] proposed deep-Hipo, which simultaneously processes patches from high and low magnifications to capture complementary cellular and tissue-level patterns. Hybrid models combining CNNs with transformer architectures further expanded the capacity for cross-scale learning. Khan *et al.* [28] highlighted the challenges of effective cross-scale feature extraction and fusion, where important small-scale features may be lost during hierarchical aggregation. Despite their efficacy, most existing multi-scale models employ relatively rigid, static fusion of features across scales without dynamic reasoning mechanisms.

Domain adaptation remains another bottleneck. Models trained on a specific data cohort often fail to generalize to other datasets of the same domain [29]. Computational burden also poses barriers to clinical deployment, with deep hierarchical models requiring extensive memory and processing power. Moreover, annotation scarcity in WSIs demands weakly supervised learning, but such methods may fail to capture subtle spatial relationships that are vital for staging. Due to this reason, there is a growing need for the development of more dynamic, flexible, and computationally efficient multi-scale fusion strategies, which are capable of explicitly modeling the cross-resolution relationships.

Recently, several multi-scale transformer frameworks have been proposed for WSI analysis, such as HIPT and GigaPath [5]. These frameworks construct the hierarchical pyramids of the image tokens across different resolutions and use the deep transformer backbones to aggregate local and global context. These models mainly focused on generic slide-level classification, representation learning, or large-scale pre-training. Their multi-scale fusion was achieved implicitly through the stacked hierarchies or by static concatenation and pooling of patch embeddings. In addition to this, most of them were not provided with the detailed scale-wise ablation or systematic quantitative validation of attention maps against the expert annotations. In contrast to the former works, our proposed HMS-T framework is specifically designed for AJCC stage-group prediction and preliminary survival risk estimation on the TCGA-BRCA cohort. It employs the three dedicated ViT branches at 5×, 10× and 20× magnifications. These branches are followed by a light-weight class-token based cross-scale attention module that can explicitly models the dependencies between magnifications. Our HMS-T keeps the computational complexity in control for gigapixel WSIs processing. Moreover, our HMS-T combines this multi-scale backbone with the Dice-validated attention maps and an auxiliary survival risk head. This helps the single architecture to support interpretable staging and preliminary prognostic stratification, while many existing multi-scale ViT models treat survival analysis as a separate downstream task.

C. Attention Mechanisms for Interpretability

1) Self-attention in vision transformers

Self-attention mechanisms, which are central to Vision Transformers (ViTs), have significantly enhanced the interpretability and localization capabilities in medical image analysis. Liu *et al.* [30] demonstrated that self-attention allows the ViTs to capture the long-range dependencies between distant regions of a histopathological image. This mechanism enables the model for more coherent tissue-level feature aggregation. Attention maps derived from ViTs can effectively help in highlighting the regions that are critical and impactful for diagnostic decisions. This technique aids the transformers in outperforming the traditional CNN-based attribution methods in clarity and clinical relevance.

Hybrid models integrating convolutional layers with self-attention, such as VSmTrans and MedViT proposed by Manzari *et al.* [31], combine the fine-grained local feature extraction with global contextual modeling. This integration in further helped in improving both the predictive accuracy and interpretability.

2) Clinical interpretability requirements

Interpretability is not just an academic concern but a prerequisite for clinical adoption of AI-based pathology tools. Maouche *et al.* [32] emphasized that clinicians required transparency and human-understandable rationales behind model predictions to trust and utilize the AI outputs in patient care. Explainable AI (XAI) methods, such as SHAP and LIME, have been employed to provide visual explanations that align model attention with known diagnostic features like mitotic figures and Tumor-Infiltrating Lymphocytes (TILs).

Mengwei *et al.* [33] further demonstrated that interpretable models can enhance the diagnostic accuracy, mainly for less experienced clinicians, by providing visible insights into the imaging features for influencing the predictions. Moreover, Lv *et al.* [34] showed that interpretable visual phenotypes extracted from attention maps correlate strongly with underlying molecular signatures, enabling a deeper understanding of tumor biology. Thus, integrating interpretability mechanisms into deep learning models is essential not only for performance validation but also for ethical, fair, and reliable AI deployment in breast cancer pathology.

3) *Prior studies on attention map validation*

The validation of attention maps against expert annotations is critical to ensure that AI models focus on clinically meaningful regions. Dabass *et al.* [35] conducted a systematic visual comparison study with overlaying attention heatmaps on histopathology slides and evaluated the alignment with pathologist-annotated tumor regions. High concordance scores from the results indicated that model-generated maps accurately highlighted biologically relevant structures.

Song *et al.* [36] introduced the Human-in-The-Loop (HITL) editing, wherein expert feedback was incorporated to refine attention maps, which resulted in improved model performance and interpretability. These validation practices are ensuring that the attention mechanisms are embedded in models like ViTs, not only boosting predictive power but also reinforcing clinical credibility.

D. Risk Stratification and Survival Prediction

1) *Existing survival models*

Survival analysis remains a cornerstone in oncological research and clinical practice. Kaplan-Meier (KM) curves, as described by offer non-parametric estimates of survival probability over time and enable visual comparisons between different patient cohorts [37]. However, KM curves are limited to univariate analyses and cannot adjust for multiple confounders. The Cox Proportional Hazards (CoxPH) is a widely used model, enabling multivariate survival analysis by estimating hazard ratios for independent variables. CoxPH models are used to identify prognostic factors influencing breast cancer outcomes.

Recent advancements have applied deep learning techniques to predict the survival rates directly from histopathology images. Wetstein *et al.* [5] developed deep learning-based grading systems that stratify the patients into survival risk groups based on achieved C-index values. Similarly, Paul *et al.* [38] introduced the DiaDeepBreastPRS model, which demonstrated that histopathology-derived risk scores were strongly correlated with overall survival and prognostic accuracy when combined with clinical metadata.

Mondol *et al.* [39] proposed hist2RNA, which could predict the gene expression profiles from WSIs, linking molecular features to survival outcomes. These multi-modal deep learning architectures, integrating imaging, genomic, and clinical data, outperform single-modality models and offer a pathway toward truly personalized oncology [5, 15, 16, 38, 39]. Importantly, these approaches offer scalable, cost-effective alternatives to molecular assays, enabling broader clinical adoption even in resource-limited settings.

2) *Challenges in staging and survival*

Integrating the tumor staging and survival prediction tasks into a unified deep learning model leads to several challenges listed in Table I.

TABLE I. COMPARATIVE REVIEW OF DEEP LEARNING FRAMEWORKS FOR TUMOR STAGING AND SURVIVAL PREDICTION IN BREAST HISTOPATHOLOGY

Framework/Study	Key Technologies	Strengths	Limitations
Lee *et al.* [16]	ManualAJCC TNM, Biomarker Scoring	Standard staging guidelines, widely used in clinics	High inter-observer variability, manual burden
Hamida *et al.* [19]	CNN (ResNet, DenseNet), U-Net for segmentation	High patch-level classification accuracy	Limited context awareness, poor long-range dependency capture
Yao *et al.* [20]	MIL framework, weak supervision	Learns from slide-level labels, reduces annotation effort	Lacks precise localization, susceptible to noise
Lee *et al.* [21]	Graph Neural Networks, Attention Pooling	Captures microenvironment context, enhances survival prediction	High complexity, demands large training data
Monjezi *et al.* [13]	Hybrid CNN-Transformer (ViT)	Accurate breast cancer detection, state-of-the-art performance	Requires computational optimization
Ayana *et al.* [22]	ViT for HER2 prediction from H&E	Avoids costly staining, achieves AUC of 0.92	Limited explainability, specific to HER2 only
Qu *et al.* [23]	CNN + ViT hybrid model	Leverages both local (CNN) and global (ViT) features	Rigid feature fusion, lacks adaptivity across magnifications
Alruily *et al.* [24]	Progressive fine-tuning of ViT	Improves efficiency, reduces computation overhead	Still lacks integrated survival modeling
Kosaraju *et al.* [27]	Deep-Hipo (Multi-scale CNN)	Processes high/low mag patches simultaneously	Static fusion, limited flexibility
Khan *et al.* [28]	Multi-scale CNN + Transformer	Cross-scale representation for tumor staging	Small-scale feature loss during aggregation
Manzari *et al.* [31]	VSmTrans, MedViT (CNN + Self-Attention)	Enhanced interpretability and performance	Increased model complexity
Maouche *et al.* [32]	Explainable AI (SHAP, LIME)	Transparent predictions, better clinician trust	Requires external explanation tools
Dabass *et al.* [35]	Attention Map Validation	Strong spatial overlap with expert annotations	Mostly visual analysis, lacks rigorous quantitative validation
Wetstein *et al.* [5]	Deep Survival Risk Models	Predicts risk stratification from WSI, good C-index	Often lacks tumor stage alignment
Paul *et al.* [38]	DiaDeepBreastPRS	Survival risk from histology + clinical data	Independent survival pipeline, not unified with staging
Mondol *et al.* [39]	hist2RNA (WSI to transcriptome prediction)	Links visual and genomic survival signals	Complex training, requires dual modality setup

Arya and Saha [40] highlighted the difficulties existed in harmonizing the heterogeneous data sources (i.e., imaging, clinical, and genomic), each with distinct scales, noise levels, and missing data patterns. Furthermore, handling the censored data in survival prediction requires careful design of loss functions and evaluation metrics, often beyond standard deep learning formulations. The major limitations we found are as follows:

- First, there is limited integration of multi-scale information via explicit attention fusion mechanisms. Most existing models either statically aggregate multi-resolution features or inadequately reconcile fine-grained and global tissue contexts [27, 28].
- Second, interpretability validation against expert annotations remains inconsistent. Although attention maps are frequently visualized, systematic validation through quantitative metrics like Dice coefficients or expert scoring is less commonly performed [35, 36].
- Third, a weak linkage exists between tumor staging and prognostic survival modeling. Most models either focus on classification tasks or survival prediction independently, without exploiting potential synergies between these objectives [38–40].

Addressing these challenges is crucial for building robust, clinically translatable AI systems, which are capable of simultaneously performing tumor staging and prognostic risk stratification. Several of the former deep learning studies are also based on the TCGA-BRCA cohort, where they mainly focus on grading, generic survival risk prediction, or histology-genomic association [16, 25, 38, 40]. However, their main objective is different from ours, since they do not consider on AJCC stage-group prediction as the primary task, and they usually rely on single-scale CNN or MIL architectures. In contrast, our HMS-T framework is designed for multi-scale AJCC staging on TCGA-BRCA. It combines the tri-scale ViT branches, explicit cross-scale attention fusion, Dice-validated attention maps, and an auxiliary survival risk head within a single architecture.

3) *Motivations for Hierarchical Multi-Scale Transformer (HMS-T)*

To address these gaps in tumor staging and risk stratification, we proposed a Hierarchical Multi-Scale Transformer (HMS-T) framework that integrates:

- Unified hybrid transformer-based feature encoding across 5×, 10×, and 20× magnifications.
- Cross-scale attention fusion to explicitly model inter-magnification dependencies.
- Built-in spatial interpretability via self-attention maps validated against expert annotations.
- An optional survival risk prediction head, seamlessly bridging tumor staging and outcome estimation.

This modular design not only enhances diagnostic accuracy but also improves explainability and paves the way for more personalized breast cancer management.

III. Materials and Methods

This chapter outlines the data resources, preprocessing techniques, and patch-level representation methods adopted to develop our tumor staging model. We first describe the characteristics of the dataset used for model training and evaluation, followed by a comprehensive description of the preprocessing pipeline and proposed HMS-T framework.

A. Dataset Description

1) TCGA-BRCA cohort

The dataset used for this study was derived from the publicly available TCGA-BRCA cohort, which consists of 1092 patients diagnosed with invasive breast carcinoma [41]. This cohort contains a rich and diverse collection of histopathological and clinical data, enabling robust modeling of tumor staging patterns.

The distribution of tumor stages according to the AJCC classification is summarized in Table II. Approximately 25% of patients were diagnosed at Stage I, 45% at Stage II, 20% at Stage III, and the remaining 10% at Stage IV. This stratification provides a balanced yet clinically realistic staging landscape for supervised learning.

TABLE II. AJCC Stage Distribution in TCGA-BRCA

Stage	Percentage	Number of Cases
I	25%	~273
II	45%	~491
III	20%	~218
IV	10%	~110

All histopathological slides were stored in SVS (Aperio) format and were scanned at a reference magnification of 20×, which is equivalent to 0.5 μm per pixel. This resolution was selected as the baseline for capturing both nuclear and glandular features essential for accurate tumor staging. In addition, each .svs file was associated with the .parcel file which contains the metadata (i.e., annotations, labels etc.) for staging prediction. Expert pathologist annotations of tumor regions were incorporated to guide the patch sampling strategies and to later validate the model’s attention-based interpretability outputs.

2) Survival metadata

In addition to staging labels, we utilize the survival-related metadata (i.e., clinical data) exclusively for risk stratification. For survival prediction, from clinical data, each patient record includes the Overall Survival (OS) time, which is measured in months from the date of diagnosis, and a binary event indicator that denotes whether the patient was deceased (event = 1) or censored (event = 0) at the last follow-up. Clinical covariates such as age at diagnosis, Estrogen Receptor (ER), Progesterone Receptor (PR), and HER2 status were also extracted. These features were encoded into a structured vector $X_{clin} \in R^d$ and passed through a shallow MLP in the risk prediction head. The survival prediction module was evaluated using the Concordance-index (C-index), denoted as Eq. (1):

$$C = \frac{1}{|p|} \sum_{(i,j)\in p} 1 \left| \hat{r}_i > \hat{r}_j \right| \quad (1)$$

where $\hat{r}_i$ is the predicted risk score for patient i, and p is the set of all comparable patient pairs.

In addition, the AJCC stage group for each patient was taken directly from the TCGA-BRCA clinical records, where it was assigned from the complete clinical TNM assessment. This stage variable is used as a slide-level label for the staging task and as a categorical covariate in the survival analysis. But the HMS-T model itself only receives the primary-tumor WSI as input for the staging head. For the survival head, we construct a clinical feature vector Z_{clin} from the available variables. Age at diagnosis is z-score normalized across the cohort. The ER, PR, and HER2 receptor status are encoded as separate binary indicators (0 = negative, 1 = positive). The AJCC stage group is then encoded as a one-hot categorical vector over four classes (Stages I–IV). This Z_{clin} vector is later concatenated with the fused histology embedding from HMS-T to form the joint input to the survival prediction module.

B. *Preprocessing Pipeline*

1) *Tissue segmentation*

To segregate diagnostically relevant regions from the gigapixel-scale WSIs, we applied a tissue segmentation process using the PyHIST toolkit [26]. In Fig. 1, the segmentation routine begins by converting the RGB image into grayscale and applying Otsu's adaptive thresholding to distinguish tissue from its background.

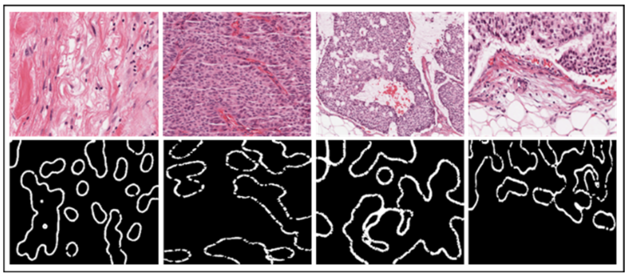

Fig. 1. Example tissue appearance and tissue segmentation results on TCGA-BRCA Whole Slide Images (WSIs). The top row shows high-resolution H&E patches with different stromal and tumor regions. The bottom row shows representative tissue boundary masks generated by our Otsu-based tissue segmentation.

We selected Otsu's method because it is an unsupervised and computationally efficient way to obtain the coarse foreground-background mask on gigapixel WSIs. This threshold value is not fixed for the whole dataset. Instead of this, it is re-estimated separately for each slide from its own intensity histogram. This allows adaptation to staining and scanner differences across cases and stages. The resulting binary mask $M(x, y)$ for pixel coordinates (x, y) is defined as Eq. (2):

$$M(x,y) = \begin{cases} 1, & if\ I(x,y) \geq T_{otsu} \\ 0, & Otherwise \end{cases} \quad (2)$$

where T_{otsu} is the dynamically determined global threshold value.

After segmentation, the morphological operations, including erosion and dilation, they were then applied to eliminate isolated artifacts and to enhance the structural continuity of tumor regions. The annotation overlays from .parcel files were also used to further refine the segmentation masks and to generate the tumor-specific masks, which are denoted as $M_t(x,y) \subset M(x,y)$.

2) *Multi-resolution patch extraction*

To spatial information at varying granularities, we adopted a multi-scale patch extraction strategy (Fig. 2) at three magnification levels: 5×, 10×, and 20×. Each WSI was divided into square patches of fixed size 256×256. This standardization ensures the consistent tensor dimensions across all branches of the attention network while preserving the relative spatial context. Here, the patches were categorized and sampled based on tissue type:

- Tumor-rich regions (as defined by M_t) were oversampled using a sliding window mechanism with minimal stride to preserve the morphological details.
- Normal tissue areas were randomly sampled from non-overlapping regions in M/M_t to serve as the controls and to reduce the class imbalance.

The total number of patches per slide are ranged from approximately 600 to 1000, depending on the tissue density and its slide size. Sampling density parameter was adjusted dynamically to ensure the equitable representation across the AJCC stages and magnification levels.

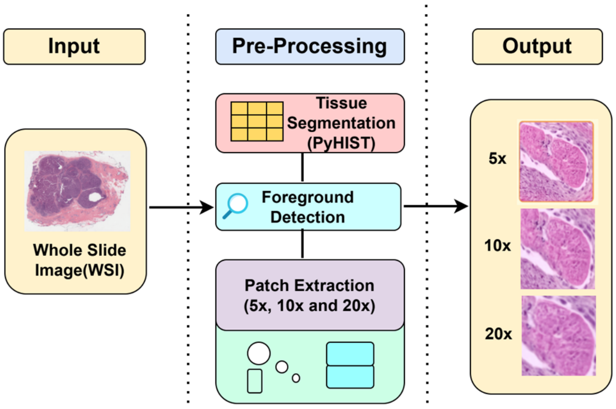

Fig. 2. Process flow of preprocessing pipeline.

3) *Data augmentation*

To enhance the model generalization and mitigate overfitting, especially to process under the limited diversity within WSIs, a robust data augmentation pipeline was applied during the training. In augmentation, each patch $P_i \in R^{256\times256\times3}$ was probabilistically subjected to a combination of the following transformations:

- Spatial augmentations: Random rotations between 0° and 360°, horizontal and vertical flipping.
- Color augmentations: Hue and saturation shifting, Gaussian noise injection with $\mu = 0$ and $\sigma \in [0.01, 0.05]$.

All pixel intensities were normalized per magnification level using:

$$P_i' = \frac{P_i - \mu_s}{\sigma_s} \tag{3}$$

where μ_s and σ_s represent the mean and standard deviation of all patches at scale $s \in \{5\times, 10\times, 20\times\}$.

C. Model Architecture

To effectively capture the morphological and spatial heterogeneity present in breast histopathological slides, we designed a Hierarchical Multi-Scale Transformer (HMS-T)-based architecture composed of dedicated Vision Transformer (ViT) branches per magnification level, as shown in Fig. 3. The architecture supports fine-to-coarse feature learning, which enables the model to simultaneously focus on nuclear morphology at 20×, glandular structures at 10×, and tissue-wide context at 5× resolution. This multi-scale encoding creates the ground for efficient downstream tumor stage classification.

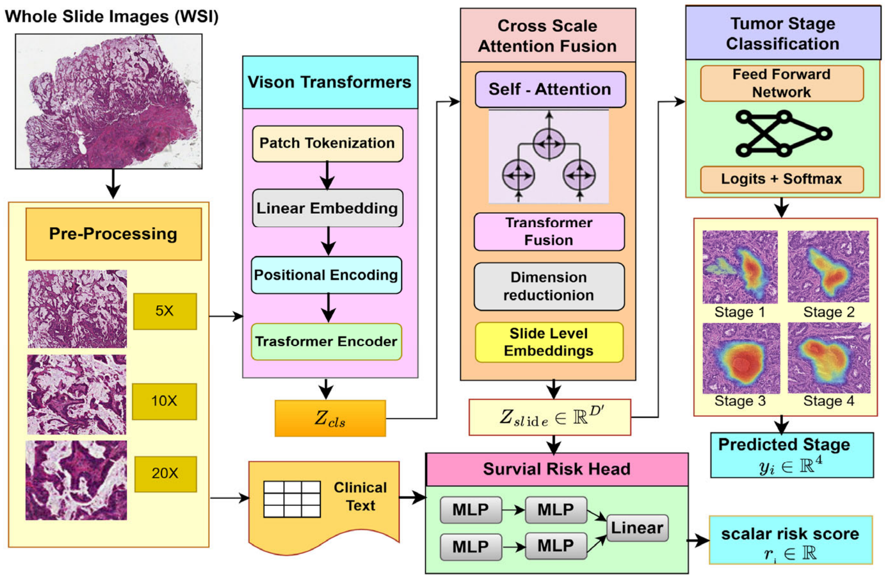

Fig. 3. Overview of the proposed HMS-T framework. Here 5×/10×/20× WSI patches are encoded by separate ViT branches, fused by a cross-scale attention block into one slide-level embedding, and jointly used for AJCC stage classification and auxiliary survival risk prediction.

1) Patch-level Vision Transformers (ViTs)

In the preprocessing pipeline, each input whole slide image is preprocessed into a set of non-overlapping patches $\{P_i^S \in R^{256\times256\times3}\}$, where $s \in \{5\times, 10\times, 20\times\}$ denotes their magnification level. After preprocessing, for each scale s, we instantiate an independent ViT branch, which allows the network to specialize in encoding the scale-specific histological patterns. As part of this, each ViT begins by dividing every input patch P_i^S into a grid of smaller non-overlapping sub-patches of size 16×16 pixels. These sub-patches are flattened and projected linearly into a D-dimensional embedding space via:

$$z_j^{(s)} = W_{emb} P_j^{(s)} + b_{emb}, \quad for\ j = 1, \dots, N \tag{4}$$

where $W_{emb} \in R^{16\times16\times3}$ is the learnable embedding matrix, b_{emb} is the bias vector, and $N = \left(\frac{256}{16}\right)^2 = 256$ is the number of sub-patches per input patch. To preserve spatial structure within each patch, we added a learnable positional encoding $E_{pos} \in R^{N\times D}$ to the embedded tokens, for forming the final input sequence to the transformer encoder as Eq. (5):

$$Z_0^{(s)} = \left[z_1^{(s)} + e_1, z_2^{(s)} + e_2, \dots, z_N^{(s)} + e_N \right] \quad e_j \in E_{pos} \tag{5}$$

Each ViT encoder consists of 12 transformer layers, where each layer consists of a Multi-Head Self-Attention (MHSA) mechanism, with residual connections, and a feedforward network [25]. For a given layer 'l', the self-attention operation computes the output as:

$$\begin{gathered} MHSA(Q, K, V) = Concat(head_1, \dots, head_h) W^0, \\ head_k = Softmax\left(\frac{Q_k K_k}{\sqrt{d_k}} \right) V_k \end{gathered} \tag{6}$$

where $Q_k = z_{l-1}^{(s)} W_k^Q K_k = z_{l-1}^{(s)} W_k^K V_k = z_{l-1}^{(s)} W_k^V$ and $h = 8$ denote the number of attention heads. Each head operates in a subspace of dimension $d_k = D/h = 96$, and the outputs are projected through $W^0 \in R^{D\times D}$. To stabilize the training and improve the representation flow, layer normalization is applied both before the attention and feedforward blocks. A typical layer's update can be expressed as Eq. (7):

$$Z_1^{(s)} = LayerNorm\left(z_{l-1}^{(s)} + MHSA(z_{l-1}^{(s)}) \right) + FFN(\cdot) \tag{7}$$

where *FFN* denotes a two-layer feedforward network with Rectified Linear Unit (ReLU) activation and dropout. The final output of each ViT branch is the class token $Z_{cls}^{(s)} \in R^D$, extracted after the last transformer layer. This token serves as the scale-specific patch representation, encapsulating the semantic and spatial properties learned at magnification level (s). The representations from the 5×, 10×, and 20× branches are later fused in the cross-scale fusion block to construct a comprehensive WSI-level embedding for stage prediction.

2) Cross-scale attention fusion

Following the independent feature encoding from each magnification level via patch-level Vision Transformer (ViT) branches, we aggregate the resulting scale-specific embeddings into a unified whole-slide representation using a cross-scale attention fusion module [12]. This attention fusion mechanism is crucial for integrating the spatially and semantically varied features. These features are spanning across nuclear, glandular, and architectural levels of tissue organization, helping in enhancing the discriminative power of the model for tumor stage classification.

Let $z_{cls}^{(5X)}$, $z_{cls}^{(10X)}$, $z_{cls}^{(20X)} \in R^D$ denotes the class tokens output from the 5×, 10×, and 20× ViT branches, respectively. These vectors serve as compact descriptors of the histological content at each resolution. To facilitate meaningful fusion, we first perform the cross-scale alignment through a self-attention mechanism, which enables the model to assess contextual correspondence between magnification levels. For this, we define the query-key-value matrices as:

$$Q = Z_{stacked}W^Q,\ K = Z_{stacked}W^K,\ V = Z_{stacked}W^V \qquad (8)$$

where $z_{stacked} = z_{cls}^{(5X)}, z_{cls}^{(10X)}, z_{cls}^{(20X)} \in R^{3XD}$ represents the concatenated class tokens across the scales. The attention weights are then computed via a scaled dot-product attention as Eq. (9):

$$Attn(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{D}}\right)V \qquad (9)$$

Yielding a cross-aligned matrix $F_{att} \in R^{3XD}$ that encodes the inter-magnification dependencies. This mechanism ensures that each scale-specific feature is modulated by context-aware signals, which are derived from the other magnifications, thereby capturing the scale-aware synergy in tissue morphology. Finally, the outputs from the attention layer are then flattened and concatenated to form a single fused feature vector as Eq. (10):

$$f_{concat} = Flatten(F_{att})\mathbb{R}^{3D} \qquad (10)$$

It is passed through a Multi-Layer Perceptron (MLP) [28] for dimensionality reduction and representation refinement. The MLP consists of two fully connected layers with ReLU activation and dropout regularization as Eq. (11):

$$f_{MLP} = \text{ReLU}\left(f_{concat}W_1 + b_1\right),\ z_{slide} = f_{MLP}W_2 + b_2 \quad (11)$$

where $W_1 \in R^{3D\times512}$, $W_2 \in R^{512\times D'}$, $z_{slide} \in R^{D'}$ are the final slide-level embedding used for tumor stage prediction. In our implementation, we set $D' = 256$ for compactness.

This fused representation encapsulates the most salient features across magnifications and is used as input to the classification head described in tumor stage classification. The main benefit of this fusion strategy lies in its ability to maintain the scale-specific discriminability while facilitating the integrative reasoning across various levels of visual hierarchy. This kind of characteristic is essential for tasks such as AJCC staging, in which the localized nuclear pleomorphism and broader architectural disarray must be considered simultaneously.

3) Tumor stage classifier

The final slide-level representation $Z_{slide} \in R^{D'}$, produced by the cross-scale attention fusion module, serves as the input to a fully connected tumor stage classifier designed to predict the AJCC stage of the input histopathological image. The classifier is optimized to map the fused, multi-scale embedding into one of four discrete tumor stages: Stages I, II, III, or IV.

The classifier is implemented as a two-layer feedforward neural network. The input embedding Z_{slide} first passes through a hidden layer of 128 units with ReLU activation is as Eq. (12):

$$h_1 = \text{ReLU}\left(Z_{slide}W_1 + b_1\right),\ W_1 \in R^{D'} \times 128 \qquad (12)$$

This activation is followed by dropout regularization with a drop probability $p = 0.3$ to mitigate overfitting:

$$h_1^{drop} = Dropout(h_1,\ p = 0.3) \qquad (13)$$

The output of the dropout layer is then projected to a 4-dimensional logit vector corresponding to the four AJCC stages as Eq. (14):

$$o = h_1^{drop}W_2 + b_2,\ \ W_2 \in R^{128\times4} \qquad (14)$$

The final stage prediction is then computed by applying the SoftMax function over the logits as Eq. (15):

$$y_i = Softmax(o_i) = \frac{\exp\left(o_i^{(k)}\right)}{\sum_{j=1}^{4}\exp\left(o_i^{(j)}\right)} \qquad (15)$$

where $y_i \in R^4$ represents the predicted class probabilities for the i^{th} input slide. To supervise the training process, we utilize the categorical cross-entropy loss function, given by Eq. (16):

$$L_{CE} = -\sum_{i=1}^{N} \sum_{k=1}^{4} w_k \cdot y_i^{(k)} \cdot \log\left(y_i^{(k)}\right) \tag{16}$$

where N is the number of training samples. $y_i^{(k)} \in \{0,1\}$ is the one-hot encoded ground-truth label for class k, and $y_i^{(k)}$ is the corresponding predicted probability. The term w_k represents a class weight used to mitigate the effects of stage-wise class imbalance. These weights are calculated using the inverse class frequency:

$$w_k = \frac{1}{\log(\alpha + f_k)} \tag{17}$$

where f_k is the relative frequency of class k in the training set, and $\alpha > 1$ is a smoothing constant (empirically set to 1.1 in our implementation) to avoid division by zero and excessive penalization. This classifier structure ensures that discriminative information distilled from all magnification levels is effectively transformed into clinically interpretable staging predictions.

4) Auxiliary survival risk head

Although the primary objective of this framework is to perform AJCC-based tumor staging [16], we introduced an auxiliary survival risk estimation head, which is designed to assess long-term patient survival outcomes using both histopathological and clinical features. This extension provides a minimal prognostic capability and serves as a precursor to more advanced survival modeling approaches explored in the future. The input to the survival head is a concatenated feature vector comprising of the fused slide-level embedding $Z_{slide} \in R^{D'}$ generated by the cross-scale attention module, and a set of standardized clinical metadata features $X_{clin} \in R^{d}$, including patient age, receptor status (ER/PR/HER2) [41], and clinical AJCC stage-group indicators, which are encoded as categorical variables defined as Eq. (18):

$$Z_{risk} = \left[Z_{slide} \middle\| X_{clin}\right] \in R^{D' \times d} \tag{18}$$

where ‖ denotes vector concatenation. This input is passed through a two-layer MLP with ReLU activation. The transformation is defined as Eqs. (19) and (20):

$$h_1 = ReLU(Z_{risk}W_1 + b_1),\ \ W_1 \in R^{(D+d)\times 128} \tag{19}$$

$$h_2 = ReLU(h_1 W_2 + b_2),\ \ W_2 \in R^{128\times 64} \tag{20}$$

After the two-layer processing, the final output is a scalar risk score, which is $r_i \in R$, computed via Eq. (21):

$$r_i = h_2 W_{out} + b_{out},\ \ W_{out} \in R^{64} \tag{21}$$

This score is interpreted as the relative risk of mortality, where the higher values indicate greater predicted hazard. The model is trained using the Cox proportional hazard's objective, which enables learning from censored time-to-event data without requiring explicit survival time regression [37]. The Cox loss function for a minibatch of N samples is defined as Eq. (22):

$$L_{Cox} = -\sum_{i=1}^{N} \delta_i \left(r_i - \log \sum_{j \in R(t_i)} \exp(r_j) \right) \tag{22}$$

where r_i is the predicted log-risk score for patient i, $\delta_i \in \{0,1\}$ is the event indicator (1 if death occurred, 0 if censored), $R(t_i)$ is the risk set, i.e., the set of patients still under observation at time t_i.

This loss function is optimized to maximize the C-index, which quantifies the model's ability to correctly rank patients based on survival risk. The survival head operates in parallel with the staging classifier, and gradients from both heads are back-propagated through the shared transformer and fusion layers during training. While this module is auxiliary and used only in exploratory settings for this work, its inclusion highlights the architectural flexibility of our framework and its potential to support dual-task learning for staging and prognosis.

In our framework, both the tumor stage classifier and the auxiliary survival risk head are attached to the same fused slide-level embedding, which is produced by the cross-scale attention module. During training, these two heads are optimized jointly in an end-to-end manner. For each slide with both staging and survival labels, we computed the total loss defined as the sum of the categorical cross-entropy loss for AJCC stage prediction. While training, the gradients from both heads are back-propagated through the shared multi-scale ViT branches and the cross-scale fusion block. Due to this, the backbone is learned to support both accurate staging and meaningful risk ranking at the same time. In this way, the staging and survival modules of HMS-T are not independent pipelines, but the two tasks are integrated into a single optimization framework.

IV. Experimental Setup

This section outlines the computational infrastructure, training pipeline, and evaluation strategy employed to benchmark the performance of our proposed hierarchical multi-scale transformer model. We describe the hardware and software stack used for distributed training, followed by the dataset partitioning protocols and metric definitions tailored for classification accuracy, interpretability, and survival prediction reliability.

A. Implementation Details

1) Hardware configuration

All training and inference procedures were conducted on a high-performance computing cluster consisting of four NVIDIA A100 GPUs, each equipped with 40 GB of VRAM and 5TB of SSD, interconnected via NVLink to facilitate fast tensor communication for distributed training.

2) Software stack

The implementation was based on PyTorch 2.0, with support from TorchVision for image transformation routines and MONAI for histopathological image management. WSIs were parsed and processed using MONAI's high-resolution tiling utilities, enabling memory-efficient patch streaming. The training pipeline was accelerated using CUDA 11.7, along with Automatic Mixed Precision (AMP) to reduce GPU memory usage and increase computational throughput without sacrificing numerical stability.

3) Hyperparameters

The model was trained using the AdamW optimizer, configured with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay coefficient $\theta = 0.01$, supporting stable convergence for transformer-based architectures. A triangular Cyclical Learning Rate (CLR) policy was adopted, with a base learning rate $\eta_{min} = 3\times10^{-5}$ and a maximum $\eta_{max} = 1\times10^{-4}$, oscillating across 50 epochs as Eq. (23):

$$\eta_t = \eta_{\min} + \frac{1}{2}\left(\eta_{\max} - \eta_{\min}\right)\left[1 + \cos\left(\frac{\pi t}{T}\right)\right] \tag{23}$$

where $t[0, T]$ is the epoch index.

To improve generalization, we incorporated dropout with a drop rate $p = 0.3$ and label smoothing with a smoothing factor $\alpha = 0.1$, which modifies the one-hot encoded ground-truth labels as Eq. (24):

$$y_{smooth}^{(k)} = (1-\alpha)\cdot y^{(k)} + \frac{\alpha}{k},\ \ k = 4 \tag{24}$$

where k is the number of classes and $y^{(k)}\{0, 1\}$ is the original target.

TABLE III. TRAINING ENVIRONMENT AND HYPERPARAMETERS

Component	Specification
GPU Hardware	NVIDIA A100 (40 GB VRAM)
Software Framework	PyTorch 2.0 + MONAI
Optimizer	AdamW
Learning Rate	3×10^{-53}
Weight Decay	0.01
Batch Size	32 (per magnification level)
Early Stopping	10 epochs (patience on validation loss)

To balance memory usage and input throughput, we adopted a batch size of 32 per magnification stream. Each training batch was composed of an equal number of patches from 5×, 10×, and 20× magnification levels, allowing simultaneous learning across hierarchical visual contexts. The effective batch size at the fusion level was thus 3×32 = 96, considering the three independent ViT encoders. An early stopping strategy was employed to avoid overfitting. The model's validation loss was monitored at the end of each epoch, and training was terminated if no improvement was observed for 10 consecutive epochs. This approach ensured optimal model check-pointing without the risk of excessive training duration or convergence to suboptimal minima. Table III summarizes the key hyperparameters and software components used in the implementation.

4) Training strategy

In practice, we trained the whole HMS-T model in an end-to-end fashion. For each epoch, the WSIs were tiled into patches at 5×, 10×, and 20×, and a fixed number of the patches per scale were sampled for each slide. These patches were forwarded through the three ViT branches, fused by the cross-scale attention module, and aggregated into one slide-level embedding. These embeddings were then used by both the stage classifier and the survival head. The cross-entropy loss for the stage prediction and the Cox loss for survival prediction was computed at the slide level and then summed to obtain the final objective. The AdamW optimizer was then used to update all the learnable parameters jointly, without any separate pre-training or freezing stage.

5) Resource and time complexity

Since the HMS-T is patch-based, the dominant cost for one slide scales is about linear with the number of tissue patches Npatch per magnification. Each ViT branch processes fixed-size 16×16 tokens per patch, so the internal attention cost of this is constant per patch. On the other hand, the complete complexity is about O (3×Npatch) for the three scales, while the cross-scale attention fusion only uses three class tokens and adds negligible overhead. On our hardware (4× NVIDIA A100 40 GB GPUs), one-fold of training for 50 epochs required on the order of ~35–38 GPU-hours (~11 clock hrs), and inference for a single WSI with 600–1000 patches per scale takes a few seconds on a single GPU. In our experiments, slides were processed in parallel, making the model more feasible for batch offline staging scenarios.

B. Dataset Splits

The dataset was split into three mutually exclusive subsets: 70% training ($n = 764$), 15% validation ($n = 164$), and 15% test ($n = 164$). To ensure that the AJCC stage distribution remained consistent across the splits, stratified sampling was applied. The efficacy of the stratification was statistically confirmed using a chi-squared test for independence, yielding p-value of $p > 0.95$, thus confirming no significant distributional drift between training and test sets. For patients with survival metadata, we aligned the split indices identically to those used in the staging task. This alignment ensures that the staging and survival modules are evaluated on congruent patient groups, eliminating bias due to data misalignment and supporting fair cross-task comparisons.

C. Evaluation Metrics

1) Staging metrics

To evaluate tumor stage prediction, we used accuracy, macro-averaged F1-Score, and Quadratic-Weighted Kappa (QWK) as metrics [9, 22].

Accuracy is computed as Eq. (25):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{25}$$

where *TP*, *TN*, *FP*, and *FN* denote True Positives, True Negatives, False Positives, and False Negatives, respectively.

Macro F1-Score is computed across the four stages by averaging the per-class F1-Scores as Eq. (26):

$$F1_{marco} = \frac{1}{k}\sum_{k=1}^{k} \frac{2 \cdot P_k \cdot R_k}{P_k + R_k} \tag{26}$$

where P_k and R_k are the precision and recall for class *k*, and *k* = 4. QWK assesses agreement between predicted and true labels, penalizing distant misclassifications more heavily as Eq. (27):

$$k = 1 - \frac{\sum_{i,j} w_{ij} o_{ij}}{\sum_{i,j} w_{ij} E_{ij}}, \quad w_{ij} = \frac{(i-j)^2}{(k-1)^2} \tag{27}$$

where o_{ij} and E_{ij} are the observed and expected confusion matrices, and w_{ij} is the weight penalty.

2) Interpretability metrics

To evaluate the spatial alignment between model attention and human expert annotations, we used the Dice Similarity Coefficient (DSC), which is computed as Eq. (28):

$$D = \frac{2 \cdot |A \cap E|}{|A| + |E|} \tag{28}$$

where *A* is the binary attention map produced by the model, and *E* is the manually annotated expert region. Higher Dice scores indicate closer spatial correspondence and stronger interpretability.

3) Survival metrics

For survival prediction, we adopted the C-index, which measures the agreement between the predicted risk scores and the actual event times as Eq. (29):

$$C = \frac{\sum_{i,j} 1(T_i < T_j) \cdot 1(R_i > R_j) \cdot \delta_j}{\sum_{i,j} 1(T_i < T_j) \cdot \delta_j} \tag{29}$$

where T_i is the survival time of patient *i*. R_i is the predicted risk score. $\delta \in \{0,1\}$ is the censoring indicator. $1(\cdot)$ denotes the indicator function. The C-index ranges from 0.5 (random) to 1.0 (perfect concordance), and is robust to censored observations, making it suitable for real-world survival modelling.

D. Baseline Models

In our experiments, we selected several baseline models according to their compatibility, because they were the widely used CNN and MIL-based approaches for WSI classification. These are trained on similar hardware within reasonable time and memory limits. Although they do not include very large pre-trained multi-scale transformer models, they still represent the strong and commonly accepted baselines for patch-wise and slide-level learning. The consistent improvements were observed with HMS-T over these methods, as presented in Table IV. The results provide evidence that the proposed multi-scale transformer framework offers additional benefits beyond standard CNN and MIL pipelines.

TABLE IV. WSI IMAGE TUMOR STAGING PERFORMANCE COMPARISON ACROSS 4 FOLDS (TCGA-BRCA TEST SET)

Model	Metric	Fold 1	Fold 2	Fold 3	Fold 4	Mean
V16+R50	Accuracy (%)	80.5	82.1	81.7	81.0	**81.3**
	Macro F1	0.76	0.77	0.78	0.76	**0.77**
	QWK	0.78	0.80	0.81	0.78	**0.79**
AMIL	Accuracy (%)	83.5	82.8	83.5	83.1	**83.2**
	Macro F1	0.78	0.77	0.79	0.78	**0.78**
	QWK	0.80	0.81	0.81	0.82	**0.81**
DCNN	Accuracy (%)	84.7	85.5	86.1	86.1	**85.6**
	Macro F1	0.81	0.82	0.83	0.82	**0.82**
	QWK	0.84	0.85	0.86	0.85	**0.85**
SDNN	Accuracy (%)	86.9	87.0	87.6	87.3	**87.2**
	Macro F1	0.84	0.84	0.85	0.83	**0.84**
	QWK	0.86	0.87	0.88	0.86	**0.87**
IRNxt	Accuracy (%)	84.1	83.5	83.7	84.3	**83.9**
	Macro F1	0.79	0.81	0.81	0.81	**0.81**
	QWK	0.82	0.83	0.83	0.84	**0.83**
HMS-T	Accuracy (%)	90.8	92.1	91.3	91.7	**91.5**
	Macro F1	0.88	0.89	0.90	0.89	**0.89**
	QWK	0.91	0.92	0.93	0.91	**0.92**

V. RESULT

This section presents the performance outcomes of our proposed HMS-T model on the AJCC tumor staging task using the TCGA-BRCA cohort dataset [41]. For this, we evaluated the performance of the tumor staging task using qualitative and quantitative approaches, followed by an analysis of magnification-level contributions. Comparisons with seven recent baseline models in this domain can further showcase the efficacy of our multi-scale transformer-based approach.

A. Tumor Staging Qualitative and Quantitative Performance

To assess the classification ability of the proposed HMS-T mode and to present the results, qualitative and quantitative approaches were selected.

1) *Qualitative performance*

The qualitative approach presents the phases involved in the HMS-T staging process and final classification results on WSI images according to AJCC guidelines, presented in Fig. 4. This figure presents the actual image, mask image, otsu threshold, tumor prediction, and AJCC stage classification (Stages I, II, and III).

In the Stage I example (shown in the middle row of Fig. 4), the HMS-T generated multi-scale attention maps, especially the 20× branch patches, that distinctly highlighted the tubular epithelial structures located near the adipose tissue boundaries of WSI images. These organized low-grade glandular patterns are aligned with early-stage morphological features, which are actively supporting the model, to accurately find the Stage I malignance characteristics. The localization feature of HMS-T enables the multi-resolution patch-level encoding to preserve both cytological detail and architectural integrity. The integration of features from 5×, 10×, and 20× magnifications allowed the network to reinforce the fine-grained tubular structures even for the surrounding tissue clutter. With this mechanism, our model clearly isolated the relevant epithelial regions of Stage I from WSI feature maps, which were later highlighted with blue attention overlay, helps in the classification phase.

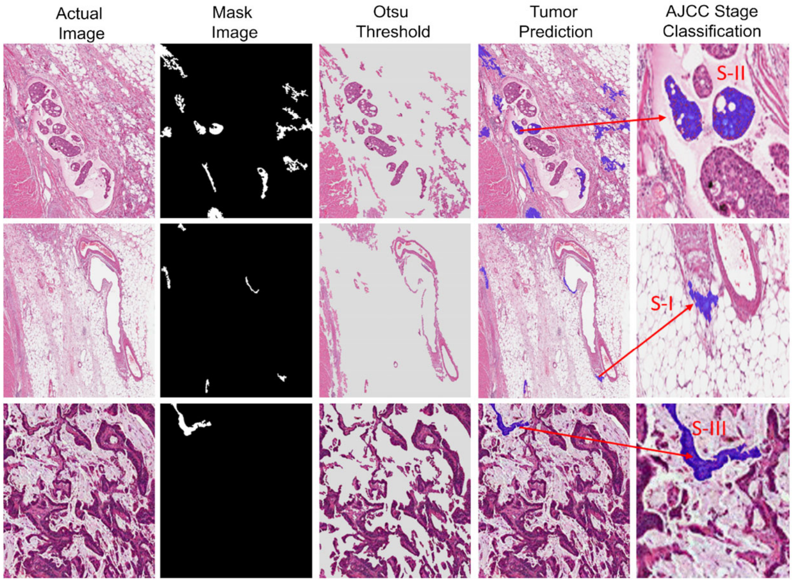

Fig. 4. Qualitative interpretability of AJCC tumor stages using HMST attention maps on WSIs.

In the case of Stage II (shown in the middle row of Fig. 4), HMST dynamically shifts its attention towards the intermediate-stage carcinoma detection by focusing on more complex glandular architectures which are exhibiting moderate nuclear atypia and increased mitotic activity hallmarks. This dynamic transition in attention is possible due to the contextual integration from both medium (10×) and high (20×) magnification branches of the cross-scale attention fusion mechanism. On the other hand, the fused representation allows the model to differentiate sophisticated spatial changes, which are epithelial thickening, luminal crowding, and nuclear variability. These salient regions demarcated from masks are reinforced by attention overlays (blue) to capture intermediate-grade features of Stage II without relying on manual annotations.

Similarly, in the case of Stages III and IV, our HMS-T attention mainly concentrated on finding the late-stage malignancy characteristics of Stage III from WSI regions, such as exhibiting the architectural disarray, cellular pleomorphism, and stromal infiltration patches. The localized deep hierarchical feature extraction helped the multi-scale attention mechanism to further accurately pinpoint the invasive cancer indicators like disorganized epithelial nests scattered within fibrotic stroma. In addition, the HMST deep hierarchical feature extraction allowed the model to hierarchically encode the tumor-stroma interactions and their disruption patterns of Stage IV, which are underrepresented in standard CNN models. Our model with morphological evidence and model-driven attention will guarantee strong interpretability at various levels of WSI staging.

Collectively, the qualitative observations from Fig. 4 are emphasizing the HMST framework's capability in delivering the stage-aware attention maps that align very closely with AJCC diagnostic criteria. The design of our model with deep hierarchical feature extraction, multi-scale transformer encoders, cross-resolution attention fusion, and localized heatmaps assured the accurate AJCC staging prediction of WSI images.

2) *Quantitative performance*

As part of the quantitative presentation of HSM-T staging classification performance, we evaluated the staging classification results with mean prediction Accuracy (m_Acc), mean Macro F1-Score (m_Mac_F1), and mean QWK (m_QMK) across 4-fold cross-validation on TGCA-BRCA test set (n = 164). The mean results are summarized in Table IV, where the proposed HMS-T model's predictions were compared against five recent counterpart baselines: VGG16+ResNet50 (V16+

R50) [42], Attention-MIL (AMIL) [43], Deep CNN (DCNN) [5], Self DNN (SDNN) [44], Inception+ResNext (IRNxt) [45], and our proposed HMS-T.

Accuracy: Our HMS-T model outperformed all other baseline models with 91.5% of mean prediction accuracy in WSI image staging. The second-best model is SDNN with 87.2% accuracy with a margin of over 4%, demonstrating the HMS-T's ability to generalize the staging patterns across WSIs.

Macro F1-score: With a mean F1-Score of 0.89, our HMS-T demonstrates its capability in managing the high degree of balance across all AJCC stage classes. This balance management is vital for staging task, where the imbalance leads to overfitting to dominant classes like Stage II or III.

QWK: HMS-T achieved the highest mean QWK of 0.92, compared to 0.87 by SDNN and 0.85 by DCNN. QWK is a valuable ordinal in classification settings, where misclassifying Stage I as Stage IV is penalized more than adjacent-stage errors. The high QWK of HMS-T reflects its clear and deep understanding of morphological differences between the stages.

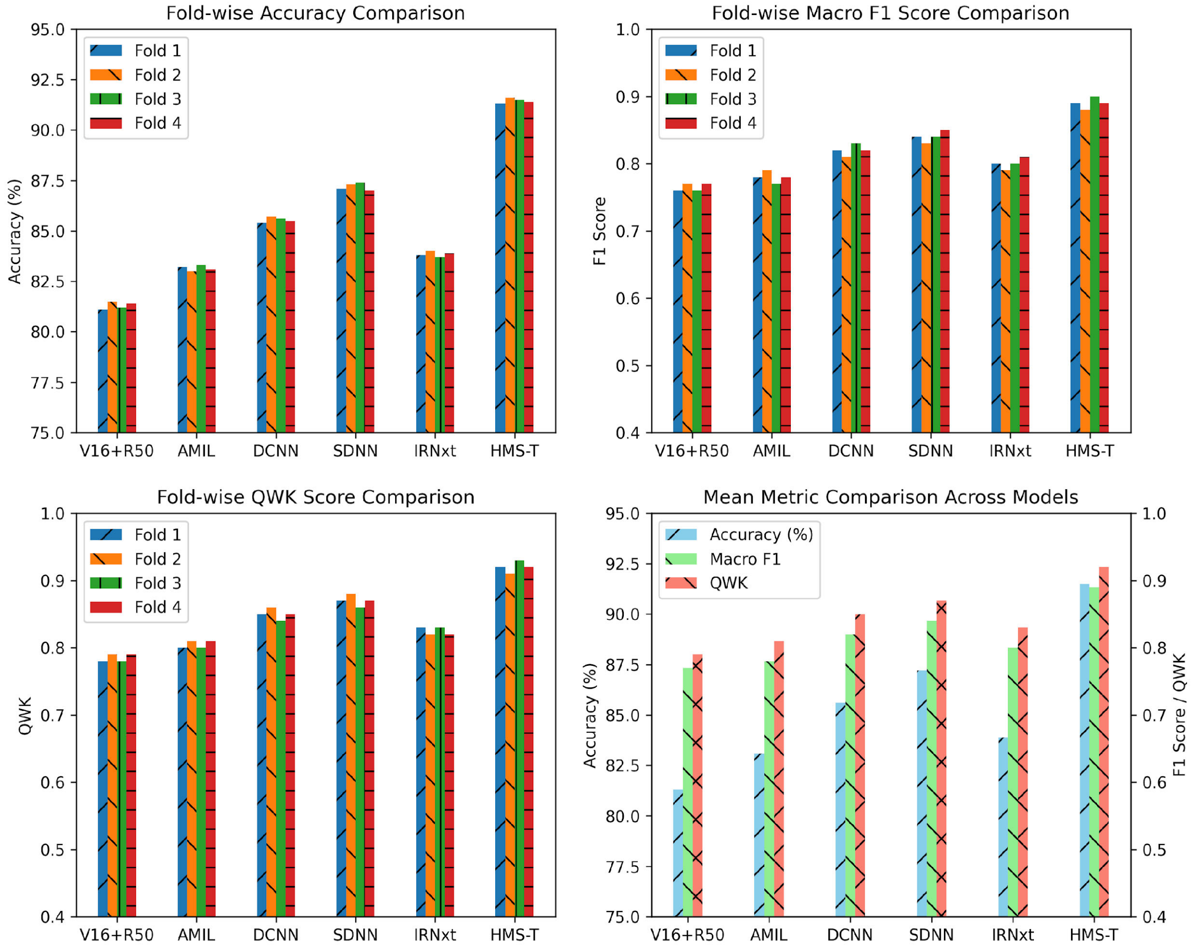

Fig. 5. Tumor staging performance comparison across baseline models and the proposed HMS-T architecture.

Across all four folds, our proposed HMS-T model consistently yielded better results than its counterparts, which is proof of the model's robustness and low variance. Unlike V16+R50 and AMIL, which have fluctuations across the folds and relatively recorded the lower macro F1-Scores (0.77 and 0.78, respectively), our HMS-T recorded a steady performance curve in all metrics. This improvement is achieved by HMS-T due to:

- Its multi-scale architecture, which captures both local cellular features (e.g., nuclear pleomorphism at 20×) and global tissue architecture (e.g., stromal invasion at 5×).
- Use of cross-scale attention feature fusion that enabled the synergistic learning across magnifications.
- Deep transformer-based modeling that outshines in capturing the complex spatial dependencies across heterogeneous WSI regions.

Fig. 5 visually illustrates this performance difference across all models using the grouped bar plots for each metric, which emphasize the quantitative superiority of our HMS-T over other methods.

Similarly, the confusion matrices shown in Fig. 6 present that our HMS-T model correctly classifies 151 out of 164 test slides on average, which yields an overall accuracy of 91.5%. Misclassifications were largely confined to Stage II and III (9 of 13 total errors), while Stages I and IV were predicted with over 95% precision, consistent with the model's macro F1 of 0.89 and QWK of 0.92.

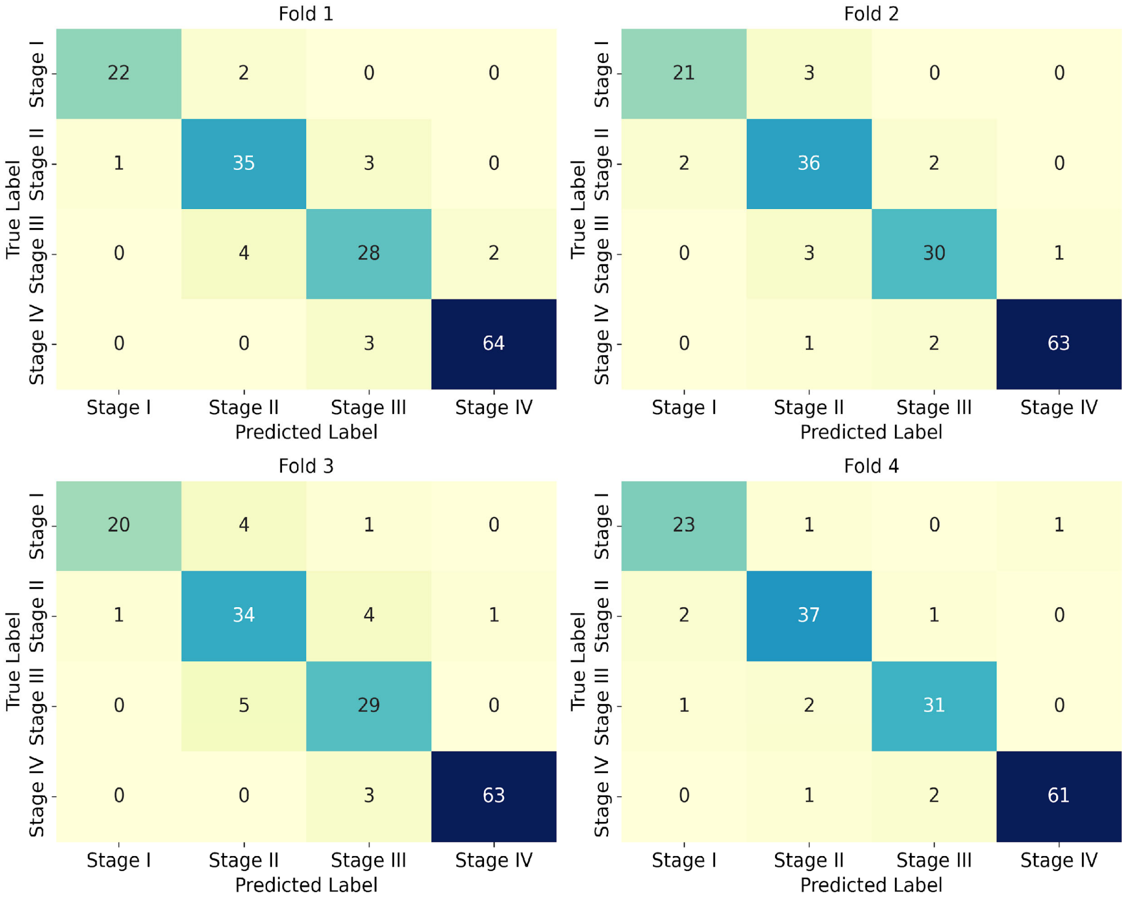

Fig. 6. Confusion matrix illustrating fold wise AJCC staging classification performance of the HMS-T model on the TCGA-BRCA test set (n = 164).

These qualitative and quantitative results conclusively demonstrate that our HMS-T not only achieves the highest numerical performance but also preserves the clinical relevance by aligning its predictions with pathologically meaningful representations.

B. Interpretability Analysis

To establish trust in high-stakes clinical environments like tumor staging, in addition to the quantitative prediction accuracy, deep AI models have to provide clear interpretability of the underlying decision mechanisms. In this section, we present the interpretability analysis using attention heat maps and morphological insights behind the tumor staging that helps to demonstrate how and where our model focus on WSI tumors for accurate staging.

Our proposed model addresses this critical need through the integration of attention heatmaps [36], which are derived from the transformer-based self-attention modules that are operating at multiple magnification levels (5×, 10×, 20×). These attention visualizations enrich the granular WSI feature insights into the HMS-T's internal reasoning by spatially localizing the diagnostically relevant histological structures.

Attention Heatmaps: As visualized in Fig. 7, class token-derived attention maps from the final transformer layer overlay the most attention regions on Whole Slide Images (WSIs) across different stages.

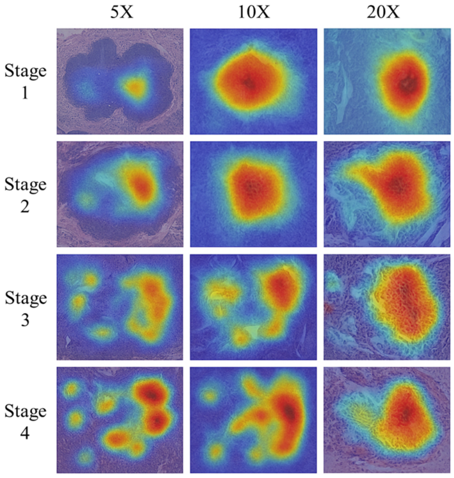

Fig. 7. Qualitative presentation of attention regions on WSIs across different stages.

These attention maps are extracted from ViT branch rendered as color gradients, with red indicating high attention and blue denoting low attention on the original (H&E) stained WSIs [46]. This facilitates a fine-grained, spatial explanation of HMS-T model behavior at different stages of AJCC.

In Stages I and II (low grade), the attention scores are primarily focused near the tubule-forming of epithelial structures, particularly in proximity to ductal formations.

This task corresponds to well-organized low-grade morphology typical of early-stage cancer. The model, through its 20× ViT branch, effectively captures cellular pleomorphism and glandular architecture, mimicking diagnostic criteria such as architectural grade and tubule formation scores.

In Stage III (high grade), the high-intensity attention regions (highlighted in red) consistently focused around the necrotic tissue zones, characterized by cell dropout, eosinophilic debris, and ghost nuclei. These mitotic features indicate more aggressive phenotypes of Stage III in AJCC staging [3]. The 10× branch of our model, with its broader field of view, captures both the local disorder and regional invasion using the cross-scale attention fusion. In Stage IV (advanced invasion), the focus is on the stromal infiltration and disorganized epithelial nests, which are the critical features for identifying highly invasive malignancies on WSI images. Conversely, in Stages I and II cases, attention was sharply concentrated along tubule-forming epithelial regions, often near ductal structures (see Fig. 8). This localization aligns with diagnostic criteria such as tubule formation scores and architectural grade, helping in further validating the model's biological sensitivity [46].

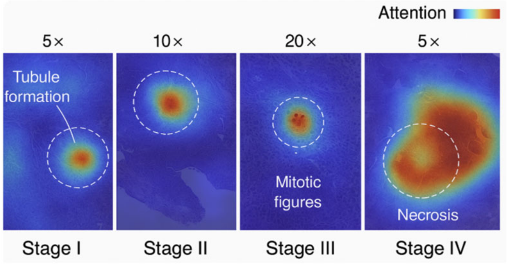

Fig. 8. Attention heatmaps overlay correlation with tumor aggressiveness and AJCC staging.

Our HMS-T modules, such as multi-scale ViT encoders, cross-scale fusion, and self-attention mechanisms, are providing this spatial interpretability analysis. The distinct extraction of morphological signals across resolutions, contextual feature weights, and inherent localization of key features are the main tasks, which produce the explainable attention maps that justifies the predicted stage and aligns with clinical intuitions. This interpretability analysis affirms that our HMS-T is a strong potential decision-support system for AJCC staging with stage-specific localization abilities.

C. *Preliminary Risk Stratification*

To explore the prognostic capabilities of our HMS-T architecture beyond AJCC staging, we incorporated a lightweight survival risk prediction head as described in the methodology chapter. This auxiliary module serves as a proof-of-concept to demonstrate that fused histological and clinical features from the staging pipeline can also be leveraged for estimating the patient survival risk. This experiment is to assess the potential of deep visual embeddings fused with clinical metadata in providing meaningful survival estimates, and to validate the architectural extensibility of our HMS-T framework for dual-task learning involving both staging and survival prediction. In this experiment, we used the same fused slide-level histology embedding together with the clinical feature vector Z_{clin} constructed from age, ER/PR/HER2 receptor status and clinical AJCC stage-group indicators. These clinical features are normalized and encoded as described in Section III, and then concatenated with the histology embedding before being passed through the two-layer MLP that outputs the survival risk score.

To obtain the survival risk score, we provide a concatenated vector $Z_{risk} = \left[Z_{slide} \| X_{clin} \right] \in \mathbb{R}^{D+d}$ as input, where $Z_{slide} \in \mathbb{R}^{D}$ is the fused slide-level representation from the cross-scale transformer. $X_{clin} \in \mathbb{R}^{D}$ is the feature set that contains the required and standardized clinical features, that are patient age, ER/PR/HER2 receptor status, and initial stage indicators. This joint vector is passed through a two-layer MLP with ReLU activation and dropout regularization functions to produce the scalar output is $r_i \in \mathbb{R}$. This scalar output r_i is later interpreted as the risk score for patient *i*, where higher values indicate the increased predicted mortality risk and vice versa.

C-index Comparison: We evaluated our HMS-T model generated survival risk prediction quantitatively, using Harrell's C-index [5], as shown in Table V. This metric is widely used in survival analysis, which quantifies the proportion of correctly ranked pairs based on predicted risk scores and observed survival outcomes.

TABLE V. C-INDEX COMPARISON OF SURVIVAL PREDICTION MODELS

Model	Visual Features	Clinical Features	C-Index
CoxPH	✗	✓	0.68
HMS-T	✓	✓	0.74

Our model, HMS-T, achieved a C-index of 0.74, significantly outperforming the traditional Cox Proportional Hazards (CoxPH) model [37], which achieved a C-index of 0.68 when trained on the same clinical dataset. The difference of 6% points indicates the substantial improvement of our model performance in survival estimation from high-dimensional histopathological data. This performance is statistically validated using the DeLong test, yielding a $p < 0.01$, which confirms that the improvement in concordance is unlikely due to random variation.

In Fig. 9, we present Kaplan-Meier survival curves stratified by predicted risk scores (low, medium, high) as assigned by the HMS-T model. The separation between risk groups is clearly observable, with the high-risk group showing significantly reduced survival probabilities over time. The curves were compared using a log-rank test, which confirmed statistical divergence with $p < 0.001$. This result illustrates the potential of integrating deep visual embeddings with clinical variables to enhance individualized risk estimation. Moreover, our HMS-T framework demonstrates extensibility toward prognostic modeling—a direction we intend to pursue in future work through larger cohorts and extended follow-up data.

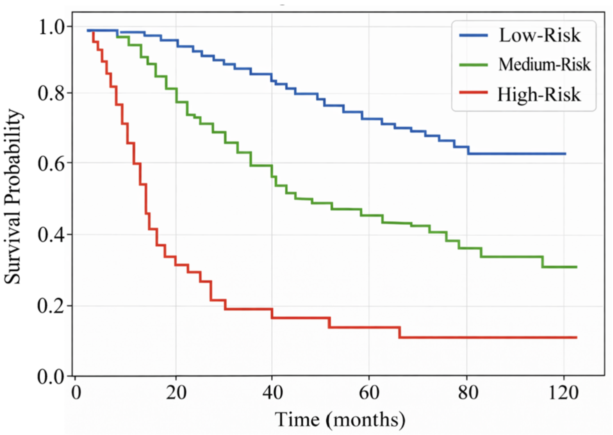

Fig. 9. Kaplan-Meier survival curves stratified by risk tertiles, log-rank $p < 0.743$.

To further relate the model-derived risk groups to the standard AJCC staging, we analyzed the distribution of the clinical stage groups within each predicted tertile. In our cohort, the low-risk tertile was dominated by the Stages I and II cases, while the high-risk tertile contained a higher proportion of the Stages III and IV disease. At the same time, within a fixed stage group, the survival curves of the low- and high-risk tertiles remained clearly separated. This indicates that the proposed risk head respects the expected stage–survival relationship, but also provides the additional stratification inside each stage by combining the histology-derived features with the basic clinical covariates.

D. Ablation Studies

To investigate the proposed HMS-T architecture efficiency and its component wise importance, we conducted a set of ablation studies to address these two core questions are: (i) what is the prominence of each magnification scale (5×, 10×, and 20×) in AJCC staging, (ii) how critical is the components of the integration mechanism in capturing the hierarchical visual cues for accurate tumor staging. To achieve this goal, we categorized the ablation studies into magnification-level analysis and component-wise model comparisons.

1) Magnification-level analysis

In this study, we sequentially ablated the magnifications (5×, 10×, or 20×) at inference time to assess their contribution in different staging tasks. The same evaluation metrics, including accuracy, Macro F1-Score, and QWK, are averaged across 4-fold cross-validation on the TCGA-BRCA test set. As part of this study, we skipped one magnification each, and the average 4-fold cross-validation quantitative results are presented in Table VI and visualized in Fig. 10.

These results are clearly indicating the non-redundant and complementary role played by each magnification level in our HMS-T framework. The ablation of the 5× branch from our framework led to the most significant drops in Stage III (–4.0) and Stage VI (–6.3) classification accuracy, along with QWK dropping by –0.06 and –0.07, respectively. These trends are highlighting the prominence of the 5× magnification level in capturing the biological features, such as architectural-level tissue disarray and stromal invasion, for Stages III and IV classification. In contrast to this, when the 10× branch was ablated from the framework, a considerable decline in results was observed in Stage II accuracy (–5.2) along with –0.06 decline in QWK. This decline in results indicates the importance of 10× magnification in Stage II classification, which balances the contextual and morphological features essential for mid-stage tumors classification. Conversely, the ablation of the 20× branch from our framework caused the highest performance degradation (–0.07) in Stage I classification accuracy, along with a –0.06 drop in Macro F1.

These results are emphasizing the need of the 20× branch for Stage I classification in our framework, which magnifies the cellular-level details such as nuclear pleomorphism and mitotic figures in distinguishing early-stage tumors. These ablation results at various magnification levels are affirming that each ViT branch of our framework is uniquely contributing to the multi-scale resolution understanding of tumor staging. The observed quantitative variations in stage-wise performance (see in Table VI and Fig. 10) confirms that the ablation of any single magnification branch from the HMS-T framework can lead to the targeted performance degradation aligned with respective clinical staging cues.

TABLE VI. MEAN ABLATION RESULTS OF MAGNIFICATION LEVEL BY STAGE

Ablated Magnification	Stage	Accuracy (%)	MacroF1	QWK
5×	Stage 1	90.70 (−0.8)	0.88 (−0.01)	0.91 (−0.01)
	Stage 2	90.23 (−1.1)	0.86 (−0.03)	0.89 (−0.03)
	Stage 3	87.31 (−4.0)	0.83 (−0.06)	0.86 (−0.06)
	Stage 4	85.45 (−6.3)	0.82 (−0.07)	0.85 (−0.07)
10×	Stage 1	89.46 (−2.1)	0.85 (−0.04)	0.88 (−0.04)
	Stage 2	86.39 (−5.2)	0.83 (−0.06)	0.86 (−0.06)
	Stage 3	87.81 (−3.5)	0.82 (−0.04)	0.85 (−0.05)
	Stage 4	88.92 (−2.6)	0.86 (−0.03)	0.88 (−0.04)
20×	Stage 1	84.82 (−6.7)	0.81 (−0.08)	0.84 (−0.08)
	Stage 2	88.10 (−3.4)	0.83 (−0.06)	0.86 (−0.06)
	Stage 3	89.70 (−1.0)	0.86 (−0.03)	0.89 (−0.03)
	Stage 4	90.50 (−0.6)	0.88 (−0.01)	0.91 (−0.01)

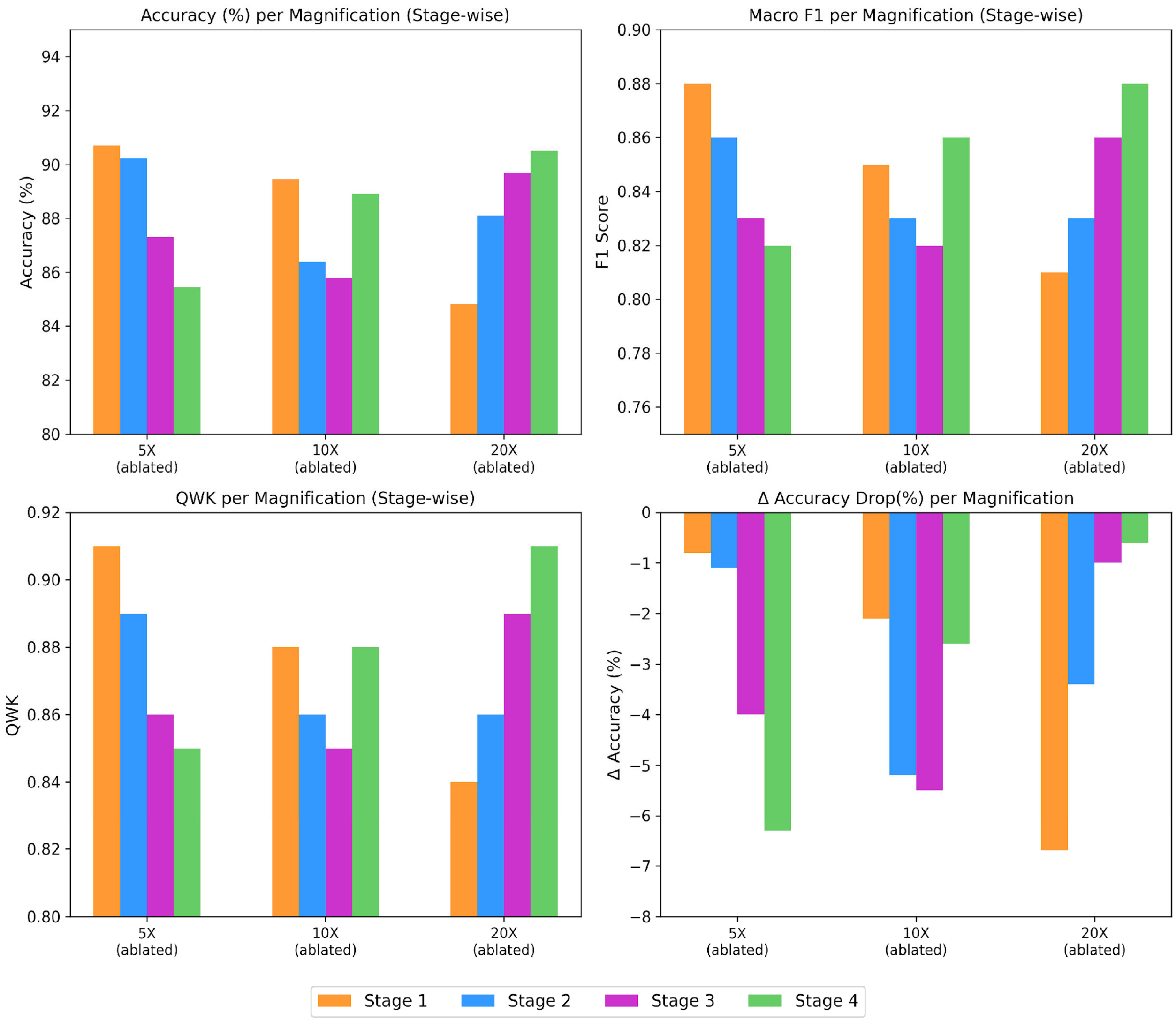

Fig. 10. Stage-wise impact of magnification-level ablation on tumor staging performance.

2) *Component ablation analysis*

To rigorously analyze the proposed HMS-T architecture, integrated components, and their prominence in AJCC staging, of the HMS-T architecture, we conducted a component-level ablation study comparing the three progressively enriched configurations:

Single-scale (20× only): This ablation utilizes only the high-resolution (20×) ViT branch without any multi-scale context.

Multi-scale (no fusion): This ablation aggregates the independent ViT outputs from 5×, 10×, and 20× branches using concatenation, without contextual interaction.

Multi-scale + attention fusion: This is our HMS-T, which integrates multi-resolution representations using the proposed cross-scale attention mechanism for deep contextual alignment.

Component ablation related comparative results over 4-fold cross-validation on the TCGA-BRCA test set (n = 164) are summarized in Table VII and Fig. 11, which highlights the staging accuracy, macro F1-Score, and QWK across configurations.

Notably, the multi-scale model with attention fusion mechanism outperformed the single-scale baseline and multi-scale no fusion across all metrics, demonstrating the critical role of multi-resolution context and integration mechanisms. Transitioning from the single-scale (20× only) configuration to the multi-scale without fusion configuration resulted in a +2.9% of increase in accuracy and +5.0% of gain in QWK, which confirms the value of incorporating coarse-to-fine-grained features in tumor staging classification. Fig. 11 presents the qualitative presentation with heatmap overlays of a single-scale (20×) model and a multi-scale attention fusion model, in Stage IV classification with missed regions.

TABLE VII. ABLATION STUDY: IMPACT OF MULTI-SCALE ATTENTION FUSION ON PERFORMANCE

Model Configuration	Accuracy (%)	Macro F1	QWK
Single-scale (20× only)	86.2	0.83	0.80
Multi-scale (no fusion)	89.1	0.86	0.85
Multi-scale + attention fusion	91.5	0.89	0.92

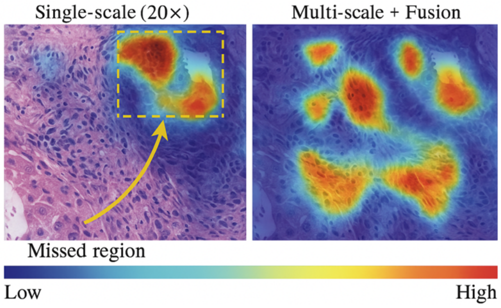

Fig. 11. Visual comparison of tumor attention maps between single-scale (20×) and multi-scale attention fusion models.

When we added the multi-scale attention with attention fusion, we observed a great leap in QWK, reaching 0.92, marking a +12.3% relative improvement over the single-scale (20× only) baseline. This improvement was evaluated statistically (see Fig. 12) based on a paired sample test using DeLong's method [47], which yielded $p <$ 0.001. Comparative results are emphasizing the importance of our architecture module integrity in evaluating the critical and morphologically adjacent stages (i.e., Stage II and III) classification. On the other hand, the macro F1-Score also benefits from the multi-resolution integration, with a gain from 0.83 to 0.89, which reflects the stronger per-class balance and robustness across diverse tumor morphologies.

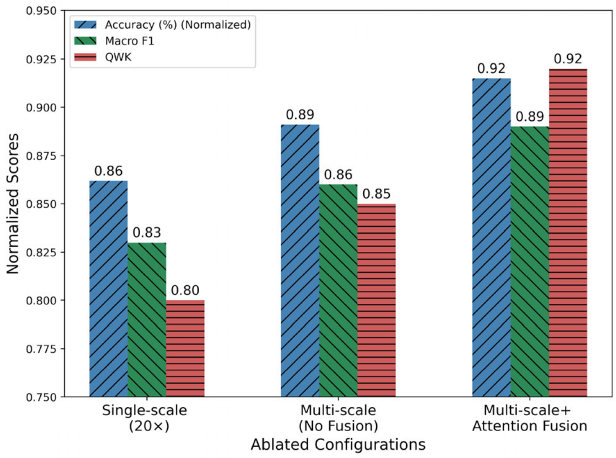

Fig. 12. Normalized comparison of model variants in component ablation analysis.

These ablation experiments with results not only confirm the effectiveness of the proposed HMS-T model, but also validate the contextual synergy introduced by our cross-scale attention fusion strategy. Finally, these experiments reaffirm that the architectural need of integrating multiple scales in HMS-T made it a modular, scalable, and clinically interpretable framework for breast tumor staging.

VI. Discussion

A. Model Advantages

The proposed HMS-T model introduces several architectural and operational advantages that make it adaptable for clinical research and potential real-world integration. A primary benefit lies in its modular design. Each magnification-specific ViT branch operates independently, which enables the straightforward swap ability with alternative transformer backbones—such as Swin-Transformers—without altering the fusion or classifier architecture. This modularity allows the model to be fine-tuned or repurposed for adjacent tasks such as histologic subtype classification, mitotic index prediction, or even cross-cancer applications, by simply substituting backbone encoders or modifying the classifier head.

Another core advantage of HMS-T is explainability, which stems from the architecture's use of self-attention weights that are directly traceable to pathologically meaningful features. For instance, model decisions related to T-stage classification (tumor size) exhibit elevated attention along expansive tumor borders and infiltration zones—regions that correlate with TNM guidelines [3]. Moreover, the attention heatmaps, when correlated with features such as mitotic count and glandular architecture, provide a built-in mechanism for visual justification, enabling diagnostic cross-validation with AJCC histologic scoring criteria.

From the experimental results, we observed that multi-scale vision transformers can offer several advantages over traditional CNNs for breast tumor staging. CNNs are very strong in capturing local texture and edge patterns. ViT-based models use self-attention to directly connect distant regions in the slide and to aggregate information over a large field of view. Recent studies [7, 30, 46] are applying ViTs to breast cancer histopathology and biomarker prediction, also reporting consistent performance gains over strong CNN baselines [9, 22]. When this global attention mechanism is extended to multiple resolutions, a multi-scale ViT can simultaneously attend to nuclear detail at 20×, glandular patterns at 10×, and tissue-level organization at 5×. This scale-wise attention is particularly important for AJCC staging, where both fine and coarse structures influence the final decision.

B. Model Limitations

Despite its strong performance and clinical relevance, the current version of the HMS-T model presents several limitations that warrant discussion. It is important to note that patch tiling in these experiments is mainly a computational process and does not change how the slide is seen by the pathologist. The WSI is automatically tiled and processed in the background, but all these predictions and attention maps are reported at the whole-slide level. Directly feeding the full gigapixel slide into the transformer is currently not practical on standard GPUs. So, the patch-based processing is still the most common and feasible solution for clinical systems, as also adopted in many existing WSI models. HMS-T is trained only on slides with histologically confirmed invasive breast carcinoma from TCGA-BRCA. Therefore, the current model is intended to assist staging and preliminary risk estimation after the cancer confirmation, and it is not designed or validated to decide whether a suspicious case is malignant or benign.

First, the model's generalizability remains to be rigorously tested. All experiments were conducted exclusively on the TCGA-BRCA cohort, which comprises pre-2015 samples collected under specific institutional protocols [41]. Consequently, the model may be biased by cohort-specific staining protocols, scanner settings, and patient demographics. Without validation on contemporary and diverse datasets such as METABRIC or INbreast, its clinical utility in broader practice remains uncertain.

Second, the survival analysis module in this study is constrained to Overall Survival (OS) outcomes. While OS is a robust endpoint, it lacks sensitivity to recurrence events and disease-specific survival, both of which are crucial for risk stratification and treatment decisions in early-stage cancers. The absence of Recurrence-Free Survival (RFS) or Progression-Free Survival (PFS) data limits the prognostic nuance of our preliminary risk stratification module [3].

Another important limitation is related to the definition of AJCC staging. In the TCGA-BRCA cohort, the stage group is assigned clinically from the full TNM assessment, which includes the nodal status and distant metastasis. Our HMS-T architecture, yet, only receives the primary-tumor WSI as input for the staging head, and it does not directly observe the N or M information. Therefore, our results should be interpreted as the WSI-based approximation of the AJCC stage group rather than the complete replacement for multidisciplinary TNM staging.

Finally, the current proposed model does not yet incorporate the domain adaptation techniques or federated learning protocols, which could mitigate the batch effects and privacy constraints in future multi-institutional deployments.

C. Future Directions

Building upon the promising results of this study, several future extensions are envisioned to expand both the technical sophistication and translational applicability of the HMS-T framework.

The next planned evolution is the HMS-T extension, which will incorporate the multi-modal data streams, including genomics and radiological imaging, to evaluate the generalization capabilities of our HMS-T model. Validation on external cohorts such as METABRIC, INbreast or other institutional datasets is planned as the next step to more strongly demonstrate generalizability and reduce the risk of overfitting to TCGA-BRCA alone. Specifically, we aim to fuse the RNA-Seq gene expression data with histological embeddings via cross-attention mechanisms to enable joint transcriptomic and histopathologic modeling.

Another direction is to extend survival modeling to longitudinal timelines, incorporating RFS and PFS endpoints using time-dependent transformer models or recurrent survival networks. Furthermore, federated learning paradigms are under consideration to facilitate decentralized model training across institutions, which is useful in improving generalizability while preserving patient privacy.

Together, these enhancements aim to transition HMS-T from a high-performing tumor staging model into a multimodal, privacy-preserving clinical decision support system for comprehensive breast cancer prognosis.

VII. Conclusion

In this study, we introduced the Hierarchical Multi-Scale Transformer (HMS-T), a modular and interpretable deep learning framework designed to automate breast tumor staging from whole slide images. By integrating Vision Transformers (ViTs) operating at 5×, 10×, and 20× magnifications, HMS-T effectively captures both fine-grained cellular details and broader architectural patterns critical for accurate staging. The proposed cross-scale attention fusion module dynamically aggregates multi-resolution features, significantly enhancing the model's ability to reason across scales and improving classification robustness according to the AJCC staging system. Our framework achieved state-of-the-art performance on the TCGA-BRCA cohort, with a staging accuracy of 91.5%, a macro F1-Score of 0.89, and a quadratic-weighted kappa of 0.92, demonstrating strong alignment with pathological standards. In addition, the embedded interpretability module validated attention maps against expert annotations, achieving a Dice similarity score of 0.81 and reinforcing the model's clinical trustworthiness. Moreover, the inclusion of a lightweight survival risk stratification head, achieving a C-index of 0.74, highlights the framework's extensibility toward integrated prognostic modeling. By simultaneously addressing diagnostic accuracy, explainability, and prognostic potential, HMS-T represents a significant step toward developing deployable, clinician-trusted AI pathology systems. Future extensions will focus on multi-modal integration and external cohort validation to further enhance clinical applicability and generalizability.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Satyanarayana Reddy Beram conceived the research idea, designed the proposed methodology, implemented the model, performed the experiments, and conducted the data analysis. He also drafted the original manuscript and prepared the figures and tables. R Lalchhanhima provided overall research guidance, contributed to the methodological design, and critically reviewed and revised the manuscript for technical and scientific quality. Ksh. Robert Singh contributed to conceptual discussions, supervised the study, and provided feedback on the experimental design and interpretation of results. All authors reviewed, revised, and approved the final version of the manuscript.

References

[1] S. M. Lima, R. D. Kehm, and M. B. Terry, "Global breast cancer incidence and mortality trends by region, age-groups, and fertility

patterns," *EClinicalMedicine*, vol. 38, 2021. https://doi.org/10.1016/j.eclinm.2021.100985

[2] A. Sidda, L. R. Biglow, G. Manu, M. Abdallah, and M. R. T. Tirona, "Importance of accurate clinical staging in patients with early stage HER2+ and triple negative breast cancer," *Journal of Clinical Oncology*, vol. 40, no. 16_suppl, e12631, 2022. https://doi.org/10.1200/jco.2022.40.16_suppl.e12631

[3] C. W. Kao, C. J. Chiang, and W. C. Lee, "Comparative effectiveness of the revised American joint committee on cancer staging system," *American Journal of Epidemiology*, vol. 194, no. 6, pp. 1735–1742, 2024. https://doi.org/10.1093/aje/kwae333

[4] A. Capar, D. A. Ekinci, M. Ertano *et al.*, "An interpretable framework for inter-observer agreement measurements in TILs scoring on histopathological breast images: A proof-of-principle study," *PLoS ONE*, vol. 19, no. 12, e0314450, 2024. https://doi.org/10.1371/journal.pone.0314450

[5] S. C. Wetstein, V. M. D. Jong, N. Stathonikos *et al.*, "Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images," *Scientific Reports*, vol. 12, no. 1, 15102, 2022. https://doi.org/10.1038/s41598-022-19112-9

[6] L. Barisoni, K. J. Lafata, S. M. Hewitt, A. Madabhushi, and U. G. J. Balis, "Digital pathology and computational image analysis in nephropathology," *Nature Reviews Nephrology*, vol. 16, no. 11, pp. 669–685, 2020. https://doi.org/10.1038/s41581-020-0321-6

[7] Y. Chen, M. Wei, and Y. Chen, "A method based on hybrid cross-multiscale spectral-spatial transformer network for hyperspectral and multispectral image fusion," *Expert Systems with Applications*, vol. 263, 125742, 2025. https://doi.org/10.1016/j.eswa.2024.125742

[8] K. Das, S. Conjeti, J. Chatterjee, and D. Sheet, "Detection of breast cancer from whole slide histopathological images using deep multiple instance CNN," *IEEE Access*, vol. 8, pp. 213502–213511, 2020. https://doi.org/10.1109/access.2020.3040106

[9] G. Ayana, E. Lee, and S. Choe, "Vision transformers for breast cancer human epidermal growth factor receptor 2 expression staging without immunohistochemical staining," *The American Journal of Pathology*, vol. 194, no. 3, pp. 402–414, 2023. https://doi.org/10.1016/j.ajpath.2023.11.015

[10] D. Lei, Y. Zhang, H. Wang, X. Xiong, B. Xu, and G. Wang, "Multi-scale dynamic sparse token multi-instance learning for pathology image classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 4, pp. 2744–2757, 2025. https://doi.org/10.1109/jbhi.2024.3509213

[11] I. Afzal, B. Yilmazel, and C. Kaleli, "An approach for multi-context-aware multi-criteria recommender systems based on deep learning," *IEEE Access*, vol. 12, pp. 99936–99948, 2024. https://doi.org/10.1109/access.2024.3428630

[12] Y. Feng, W. Ni, L. Song, and X. Wang, "MsFNet: Multi-scale fusion network based on dynamic spectral features for multi-temporal hyperspectral image change detection," *Remote Sensing*, vol. 16, no. 16, p. 3037, 2024. https://doi.org/10.3390/rs16163037

[13] E. Monjezi, G. Akbarizadeh, and K. Ansari-Asl, "RI-ViT: A multi-scale hybrid method based on vision transformer for breast cancer detection in histopathological images," *IEEE Access*, vol. 12, pp. 186074–186086, 2024. https://doi.org/10.1109/access.2024.3514322

[14] M. Ma, R. Liu, C. Wen *et al.*, "Predicting the molecular subtype of breast cancer and identifying interpretable imaging features using machine learning algorithms," *European Radiology*, vol. 32, no. 3, pp. 1652–1662, 2021. https://doi.org/10.1007/s00330-021-08271-4

[15] E. Wulczyn, D. F. Steiner, Z. Xu *et al.*, "Deep learning-based survival prediction for multiple cancer types using histopathology images," *PLoS One*, vol. 15, no. 6, e0233678, 2020. https://doi.org/10.1371/journal.pone.0233678

[16] Y. S. Lee, C. U. Lee, M. J. Lee *et al.*, "The direct comparison between 7th AJCC staging system and 8th AJCC staging system for prediction of survival with Korean multicenter HCC patients," *Journal of Hepatology*, vol. 68, p. 533A, 2018. https://doi.org/10.1016/s0168-8278(18)31108-5

[17] P. S. Ginter, R. Idress, T. M.D'Alfonso *et al.*, "Histologic grading of breast carcinoma: A multi-institution study of interobserver variation using virtual microscopy," *Modern Pathology*, vol. 34, no. 4, pp. 701–709, 2020. https://doi.org/10.1038/s41379-020-00698-2

[18] A. Polónia and A. Caramelo, "HER2 in situ hybridization test in breast cancer: Quantifying margins of error and genetic heterogeneity," *Modern Pathology*, vol. 34, no. 8, pp. 1478–1486, 2021. https://doi.org/10.1038/s41379-021-00813-x

[19] A. B. Hamida, M. Devanne, J. Weber *et al.*, "Deep learning for colon cancer histopathological images analysis," *Computers in Biology and Medicine*, vol. 136, 104730, 2021. https://doi.org/10.1016/j.compbiomed.2021.104730

[20] J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, and J. Huang, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," *Medical Image Analysis*, vol. 65, 101789, 2020. https://doi.org/10.1016/j.media.2020.101789

[21] Y. Lee, J. H. Park, S. Oh *et al.*, "Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning," *Nature Biomedical Engineering*, vol. 6, pp. 1452–1466, 2022. https://doi.org/10.1038/s41551-022-00923-0

[22] G. Ayana, E. Lee, and S. Choe, "Vision transformers for breast cancer human epidermal growth factor receptor 2 expression staging without immunohistochemical staining," *The American Journal of Pathology*, vol. 194, no. 3, pp. 402–414, 2023. https://doi.org/10.1158/1538-7445.am2023-5437

[23] X. Qu, H. Lu, W. Tang *et al.*, "A VGG attention vision transformer network for benign and malignant classification of breast ultrasound images," *Medical Physics*, vol. 49, no. 9, pp. 5787–5798, 2022. https://doi.org/10.1002/mp.15852

[24] M. Alruily, A. A. Mahmoud, H. Allahem *et al.*, "Enhancing breast cancer detection in ultrasound images: An innovative approach using progressive fine-tuning of vision transformer models," *International Journal of Intelligent Systems*, vol. 2024, no. 1, 2024. https://doi.org/10.1155/int/6528752

[25] T. S. Sheikh, Y. Lee, and M. Cho, "Histopathological classification of breast cancer images using a multi-scale input and multi-feature network," *Cancers*, vol. 12, no. 8, p. 2031, 2020. https://doi.org/10.3390/cancers12082031

[26] A. Pedersen, E. Smistad, T. V. Rise *et al.*, "H2G-Net: A multi-resolution refinement approach for segmentation of breast cancer region in gigapixel histopathological images," *Frontiers in Medicine,* vol. 9, 971873, 2022. https://doi.org/10.3389/fmed.2022.971873

[27] S. C. Kosaraju, J. Hao, H. M. Koh, and M. Kang, "Deep-Hipo: Multi-scale receptive field deep learning for histopathological image analysis," *Methods*, vol. 179, pp. 3–13, 2020. https://doi.org/10.1016/j.ymeth.2020.05.012

[28] H. U. Khan, B. Raza, A. Waheed, and H. Shah, "MSF-model: Multi-scale feature fusion-based domain adaptive model for breast cancer classification of histopathology images," *IEEE Access*, vol. 10, pp. 122530–122547, 2022. https://doi.org/10.1109/access.2022.3223870

[29] Y. Wang, F. Huang, Y. Zhang *et al.*, "Breast cancer image classification via multi-level dual-network features and sparse multi-relation regularized learning," *PubMed*, pp. 7023–7026, 2019. https://doi.org/10.1109/embc.2019.8857762

[30] T. Liu, Q. Bai, D. A. Torigian, Y. Tong, and J. K. Udupa, "VSmTrans: A hybrid paradigm integrating self-attention and convolution for 3D medical image segmentation," *Medical Image Analysis*, vol. 98, 103295, 2024. https://doi.org/10.1016/j.media.2024.103295

[31] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "MedViT: A robust vision transformer for generalized medical image classification," *Computers in Biology and Medicine*, vol. 157, 106791, 2023. https://doi.org/10.1016/j.compbiomed.2023.106791

[32] I. Maouche, L. S. Terrissa, K. Benmohammed, and N. Zerhouni, "An explainable AI approach for breast cancer metastasis prediction based on clinicopathological data," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 12, pp. 3321–3329, 2023. https://doi.org/10.1109/tbme.2023.3282840

[33] M. Sun, "Person re-identification based on spectral nonlocal block and multiscale attention pyramid," *Fourth International Conference on Signal Processing and Computer Science*, vol. 12970, pp. 70–70, 2023. https://doi.org/10.1117/12.3012216

[34] T. Lv, X. Hong, Y. Liu *et al.*, "AI-powered interpretable imaging phenotypes noninvasively characterize tumor microenvironment associated with diverse molecular signatures and survival in breast cancer," *Computer Methods and Programs in Biomedicine*, vol. 243, 107857, 2023. https://doi.org/10.1016/j.cmpb.2023.107857

[35] M. Dabass, S. Vashisth, and R. Vig, "A convolution neural network with multi-level convolutional and attention learning for classification of cancer grades and tissue structures in colon histopathological images," *Computers in Biology and Medicine*, vol. 147, 105680, 2022. https://doi.org/10.1016/j.compbiomed.2022.105680

[36] B. Song, C. Zhang, S. Sunny *et al*., "Interpretable and reliable oral cancer classifier with attention mechanism and expert knowledge embedding via attention map," *Cancers*, vol. 15, no. 5, p. 1421, 2023. https://doi.org/10.3390/cancers15051421

[37] C. Andrade, "Survival analysis, kaplan-meier curves, and cox regression: basic concepts," *Indian Journal of Psychological Medicine*, vol. 45, no. 4, pp. 434–435, 2023. https://doi.org/10.1177/02537176231176986

[38] J. Paul, C. Bossard, J. Rynkiewicz *et al.*, "Survival outcome prediction of breast carcinomas on whole-slide histopathology images using deep learning," *Journal of Clinical Oncology*, vol. 42, no. 16_suppl, p. 1070, 2024. https://doi.org/10.1200/jco.2024.42.16_suppl.1070

[39] R. K. Mondol, E. K. A. Millar, P. H. Graham, L. Browne, A. Sowmya, and E. Meijering, "hist2RNA: An efficient deep learning architecture to predict gene expression from breast cancer histopathology images," *Cancers*. vol. 15, no. 9, 2569, 2023. https://doi.org/10.3390/cancers15092569

[40] N. Arya and S. Saha, "Multi-modal advanced deep learning architectures for breast cancer survival prediction," *Knowledge-Based Systems*, vol. 221, 106965, 2021. https://doi.org/10.1016/j.knosys.2021.106965

[41] W. Lingle, B. J. Erickson, M. L. Zuley *et al.* (2016). The Cancer Genome Atlas Breast Invasive Carcinoma Collection (TCGA-BRCA) (Version 3) [Data set]. *The Cancer Imaging Archive*. [Online]. Available: https://doi.org/10.7937/K9/TCIA.2016.AB2NAZRP

[42] A. A. Balasubramanian, S. M. A. Al-Heejawi, A. Singh *et al*., "Ensemble deep learning-based image classification for breast cancer subtype and invasiveness diagnosis from whole slide image histopathology," *Cancers*, vol. 16, no. 12, 2222, 2024. https://doi.org/10.3390/cancers16122222

[43] C. Mercan, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images," *IEEE Transactions on Medical Imaging*, vol. 37, no. 1, pp. 316–325, 2018. https://doi.org/10.1109/TMI.2017.2758580

[44] A. H. Abdulaal, M. Valizadeh, M. C. Amirani, and A. S. Shah, "A self-learning deep neural network for classification of breast histopathological images," *Biomedical Signal Processing and Control*, vol. 87, 105418, 2024. https://doi.org/10.1016/j.bspc.2023.105418

[45] F. Shahidi, S. M. Daud, H. Abas, N. A. Ahmad, and N. Maarop, "Breast cancer classification using deep learning approaches and histopathology image: A comparison study," *IEEE Access*, vol. 8, pp. 187531–187552, 2020. https://doi.org/10.1109/access.2020.3029881

[46] H. Xu, Q. Xu, F. Cong *et al*., "Vision transformers for computational histopathology," *IEEE Reviews in Biomedical Engineering*, vol. 17, pp. 63–79, 2024. doi: 10.1109/RBME.2023.3297604

[47] L. Zou, Y. Choi, L. Guizzetti, D. Shu, J. Zou, and G. Zou, "Extending the DeLong algorithm for comparing areas under correlated receiver operating characteristic curves with missing data," *Statistics in Medicine*, vol. 43, no. 21, pp. 4148–4162, 2024. https://doi.org/10.1002/sim.10172