Graphs HELP Predict Cancer: Multimodal GNNs for Cancer Recurrence Risk

Akhil Arularasu
Emory University
akhil.arularasu@emory.edu

Alexander Liu
Emory University
aliu266@emory.edu

Thomas Ma
Emory University
thomas.ma@emory.edu

Abstract

Predicting if a patient's cancer will recur is an important task that prevents unnecessary chemotherapy for patients who would not benefit from treatment and helps others receive the time-sensitive treatment they need. Baseline models such as a ViT-CNN and logistic regression hybrid by Goyal et al. (2024) offer impressive performance. However, unlike graph-based models, convolutional neural networks (CNNs) struggle to capture non-local context and do not carry inductive bias on structural representations, while vision transformers (ViTs) struggle to capture the finer details of cell interactions. In this paper, we present Histologic Extremely Lossy Predictor (HELP), a gated fusion hybrid model that combines a graph transformer and a hypergraph neural network classifier to exploit the strengths of graph-based models. We use a graph transformer as an image model to better match the data structure of whole slide images over CNNs and ViTs. Our experiments are conducted on the Dartmouth BMIRDS Breast Cancer Dataset (Goyal et al., 2024). Our graph models achieved strong performance relative to our baselines on whole slide image data and clinicopathologic metadata. To the best of our knowledge, this is the first work to incorporate graph-based models into cancer recurrence risk predictions on H&E stained WSIs.

1 Introduction

Cancer is the second leading cause of death globally. Of the 10 million cancer-related deaths in 2020, 70%, or nearly one in six deaths, came from low- and middle-income countries (ACS, 2022) (WHO, 2025). An important tool to prevent cancer deaths is medical imaging, which involves creating visual representations of the body for physicians to provide accurate diagnoses and explain critical medical decisions to patients (Dao and Ly, 2024). To diagnose some cancers, pathologists receive stained patient biopsies on glass slides that they can view to examine aberrances. Whole-slide imaging (WSI) uses special scanners and cameras to capture many small high-resolution frames of a glass slide to be assembled into a complete gigapixel-sized digital image of the entire slide (Kumar et al., 2020). This allows pathologists to use specialized software for more complicated analysis, remote access and sharing of samples, and automated image analysis tools.

Automated cancer recurrence risk analysis has recently gained significant attention due to its clinical importance in treatment planning (Zuo et al., 2023) and early detection and diagnosis of cancer improve the probability that a patient survives (WHO, 2025). Building a generalizable and reliable automated model can simplify the process of making accurate treatment decisions, monitoring disease progression, and planning patient aftercare. Furthermore, it can improve access to healthcare in developing countries where the traditional Oncotype DX test for risk of cancer recurrence is not available.

Although there has been significant development of automated tools, most existing systems struggle with scalability, generalization (across hospitals or scanners), and integration with existing patient information (e.g. radiology text reports, clinical records, images). Current models are optimized for narrow tasks and require substantial compute resources that limit their deployment in real world settings.

Several studies have proposed convolutional neural networks (CNNs) and vision transformers (ViTs) to overcome these issues. However, CNNs and ViTs are still limited because they do not fit the behavior of WSI data (Kwadkar, 2025) (Dosovitskiy et al., 2021) (Touvron et al., 2021). CNNs struggle with long-range dependencies and the complexity of cellular arrangements, while ViTs are better at interpreting global context but struggle

to capture the finer details of cell interactions.

In this paper, we present a novel framework, Histologic Extremely Lossy Predictor (HELP), that takes advantage of whole slide image data and its corresponding tabular clinicopathological metadata through gated fusion graph neural networks to improve risk classification accuracy (Section 3). Our approach uses a GPS graph transformer (Rampášek et al., 2023) on the whole-slide image data to resolve the limited preservation of similarity relationships between image patches found in CNNs and ViTs due to their lack of graphical representations that encode relational information and the promising track record of GNN-based methods in related fields such as survival analysis (Wang et al., 2024), cancer subtyping (Li and Nabavi, 2024), and biomarker prediction (Lu et al., 2022). Based on the graph-based architecture of GNNs, we hypothesize that they are a better model of the cellular relationships that are important in the interpretation of WSI data. We also use a hypergraph neural network (HGNN) classifier for our metadata to preserve richer structural information after promising performance over other models on clinicopathologic information (Xu et al., 2023).

Our approach is trained and evaluated on the Dartmouth Breast Cancer Recurrence Risk Dataset, which is made up of 990 hematoxylin and eosin (H&E)-stained formalin-fixed paraffin-embedded (FFPE) whole-slide images and clinicopathologic metadata. The metadata includes the Oncotype DX Breast Recurrence Score, patient age, tumor size, tumor grade, histologic type, estrogen receptor (ER) status, progesterone receptor (PR) status, and HER2 protein status. We will evaluate our model against other competitive models including CNN and ViT.

This work makes contributions as follows:

1. **Dataset innovation.** We construct graphs from the whole slide image data, creating nodes from patches and defining edges to describe similarity through weights and adjacency through nearest neighbors between patches.

2. **Novel framework.** We introduce a novel framework based on GNNs that preserves relational and spatial structure in WSIs and handles the multi-modality of corresponding WSIs and patient metadata.

3. **Comprehensive evaluation.** We benchmark our framework against strong baselines (CNNs and ViTs) and demonstrate competitive predictive accuracy.

2 Related Work

Although there are only a few studies that apply deep learning models to the specific problem of cancer recurrence risk, we also examine the application of deep learning models to the general and related problem of cancer prognosis based on whole slide images for a more comprehensive picture.

2.1 Deep Learning on Whole Slide Images

Studies on deep learning models for medical image data began as soon as the first architectures were designed, even back to artificial neural networks (ANNs). However, more concrete work began with CNNs, which became one of the first deep learning architectures to be applied to cancer prognosis (Lee, 2023), and Silveira et al. (2025) found that CNNs were the most popular for predicting recurrence solely from image data.

CNNs and transformer architectures such as the ViT have found significant success in several medical imaging tasks, including COVID-19 detection from chest X-rays (Krishnan and Krishnan, 2021) and eye disease detection from retinal fundus images (Djoumessi et al., 2025). However, this does not generalize to the specific nature of H&E-stained whole slide images, which require a more nuanced understanding that can also capture cellular relationships and behavior. As a result, deep learning on whole slide images was only possible after widespread adoption of scanners that could produce digital whole slide images for computers to use (FDA, 2017).

Qu et al. (2022) found that CNNs could produce an accurate histological scoring system that predicts short-term and long-term recurrence of early-stage hepatocellular carcinoma, with consistent high performance across most subtypes when validated on an external cohort. The authors used a weakly-supervised approach with each manual annotations for each slide in six categories (tumor region, normal liver tissue, portal area, fibrosis, hemorrhage/necrotic area, and lymphocyte area).

More recently, Goyal et al. (2024) introduced a CNN-ViT and logistic regression hybrid model trained on the same Dartmouth dataset we use to train and evaluate our models. The authors

achieved an AUC of 0.91 on the Dartmouth training dataset and 0.84 on an external University of Chicago cohort.

Kwadkar (2025) conducted a comparative analysis that found CNNs outperformed transformers on tasks with distinct anomalies. An image classifier CNN, EfficientNet-B0, even outperformed ViTs on average. However, transformers made strong use of attention on more subtle tasks like brain tumor classification which CNNs struggled on. Although CNNs are known to be equivariant to translation and invariant to small translation via pooling, as well as parameter sharing leading to a lower parameter count, they can struggle with more significant transformations, nonlocal relationships, identifying objects, and preserving certain spatial information after pooling. On the other hand, ViTs may perform well when the global context is important but loses information on nuanced data since the core attention mechanism calculates attention across image patches while diluting local information contained within each patch. This demonstrates a fundamental weakness that hinders its performance on the crucial details in whole slide images.

This analysis illustrates the strengths of both architectures, but also displays their weaknesses with specific spatial information that can be represented via graph structures. Hybrid-GNNs such as the GAT can preserve the attention mechanism, while incorporating better spatial locality. Because the patch samples are independent, it would be beneficial to generate spatial dependencies and reasoning through GNNs to preserve the cellular behavior and relationships.

2.2 Graph Neural Networks

Introduced in 2005, the GNN was designed to preserve graph data relationships when the convention for deep learning at the time was to flatten data into vectors, losing valuable insight on data relationships (Gori et al., 2005).

To our knowledge, GNNs have not been applied to the specific task of cancer recurrence risk prediction, offering potential for novel contributions to an underexplored task within medical imaging. However, GNNs have found some use in related medical imaging tasks.

(Xiao et al., 2025) found that a U-Shaped variant of GNNs surpassed state-of-the-art transformer-based models in tumor medical image segmentation. This shows that GNNs have the potential to incorporate spatial information better than transformers.

GNNs were also combined with an attention mechanism for a multi-stacked-layered graph attention network (MSL-GAT) that predicted critical driver genes associated with bladder cancer progression through multi-omics (genomics, transcriptomics, proteomics, metabolomics) data (Ibrahim et al., 2025). This approach outperformed all competitive models tested against and highlights a strong potential for a combined approach based on a GNN and attention mechanism.

Another state-of-the-art graph-based model in cancer survival prediction, the DM-GNN, incorporates both convolution and attention-based mechanisms in a dual-stream architecture, utilizing graphs to better represent similarities between different WSI samples based on structure similarity and dependencies in biological units (Wang et al., 2024).

A series of proposed models for preprocessing tumor microenvironment gastric cancer image data into a graph of cellular information yielded better results than the globally adopted "TNM Classification of Malignant Tumors" (TNM) standard (Wang et al., 2022). GNNs also demonstrated high performance in lung cancer survival analysis (Lian et al., 2022), an important area of cancer treatment related to cancer recurrence. The image-based graph convolution network (GCN) outperformed classical ML models, a tumor-based CNN (Tumor-CNN), and the traditional and universal TNM (tumor, nodes, metastasis) used by hospitals on the same lung CT images.

Another paper trained a novel GNN on whole slide images (SlideGraph+) to predict HER2 status, a protein where over-expression leads to tumor growth and cancer development (Lu et al., 2022). The model demonstrated improved computational efficiency compared to classic patch-based WSI models, since the patch-level features only needed to be extracted once before the graph representation modeled the global context of the WSI. The model also outperformed the state-of-the-art and all other competitive models it was evaluated against. The researchers also propose that the architecture of SlideGraph+ could also be applied to cancer survival analysis and recurrence prediction

Given the success of GNNs in cancer analysis fields, we believe a GNN can capture the cellular details of whole slide images and outperform both CNNs and ViTs.

3 Approach

We frame cancer recurrence prediction as a **binned binary classification** problem. Based on standard practice in the medical field (Sparano et al., 2018) and to remain consistent with previous work with the same dataset (Goyal et al., 2024), each slide is labeled *high risk* if its Oncotype DX recurrence score is 26-100 and *low risk* if the score is 0-25.

As seen in Figure 2, our proposed model consists of three components: a Graph Transformer classifies based on graphs constructed from whole slide images, a Hypergraph Neural Network classifies based on clinicopathologic metadata, and a gated fusion module combines classification results from image and metadata block to output a final prediction label based on dynamic weighting.

3.0.1 Graph Feature Construction

Our graph feature engineering pipeline can be visualized in Figure 1. Due to the extreme spatial resolution and gigapixel scale of WSIs, directly processing the entire image is computationally infeasible. The images also vary in size. To address these issues, we create 224 x 224 patches based on precedent set by (Krizhevsky et al., 2012) and (He et al., 2016) for each whole slide image and feed them to a Vision Transformer to generate high-dimensional visual embeddings as feature space based on past success with DINOv2 (Oquab et al., 2024). The visual embeddings of the patches and coordinates for the patch centers are fed into a graph building algorithm (Algorithm 1) to construct our features for the Graph Transformer.

Algorithm 1 Graph Construction

Input: Z: list of patch lists P for each WSI.

function GRAPH-BUILDER(Z)
  $G \leftarrow [\,]$
  for all P in Z **do**
    $X \leftarrow [\,]$
    $C \leftarrow [\,]$
    $y \leftarrow label$
    for all p in P **do**
      $x \leftarrow visual(p)$ ▷ Get visual embedding
      $append(X, x)$
      $c \leftarrow center(p)$
      $append(C, c)$
    end for
    $k = number\_neighbors$
    $E \leftarrow build\_edges(C, n)$
    $W \leftarrow dist(E)$ ▷ Calculate L2 dist. b/w edges
    $g \leftarrow graph(X, y, C, E, W)$ ▷ Any structure
    $append(G, g)$
  end for
  return G
end function

To construct a sparse and semantically meaningful graph structure, we employ a nearest neighbors strategy based on the Euclidean distance between patch center coordinates. Each patch corresponds to a node, and edges are established between nodes whose centers lie among each other's k nearest neighbors. The edge weights are proportional to the inverse Euclidean distance between patches, promoting stronger interactions between spatially proximal regions.

The resulting graphs thus consist of:

- Node features: visual embeddings concatenated with positional encodings
- Edge weights: distance-based scalars with nearest neighbors adjacency between patch centers

A specific and detailed description of the preprocessing pipeline, selection of Vision Transformer, and hyperparameter choices is provided in Section 4.

3.0.2 WSI Graph Transformer Block

The WSI graph representation is processed using a General, Powerful, Scalable (GPS) Transformer (Rampášek et al., 2023), which combines local message passing with global attention to capture both fine-grained and long-range dependencies across image patches.

Each GPS layer consists of two primary components:

- A local MLP-GINEConv module that propagates information through edges and aggregates neighborhood-level features
- A global attention mechanism implemented using either a Multi-Head Attention (MHA) or Performer-based efficient attention module for scalable global context modeling.

The outputs of these modules are combined through residual connections and layer normalization, ensuring stable training and improved representational capacity. A final multilayer perceptron (MLP) projects the aggregated node embeddings into a fixed-dimensional latent space for downstream fusion.

3.0.3 Metadata Hypergraph Block

While graph edges face limitations on each edge connecting exactly two nodes, hypergraphs can

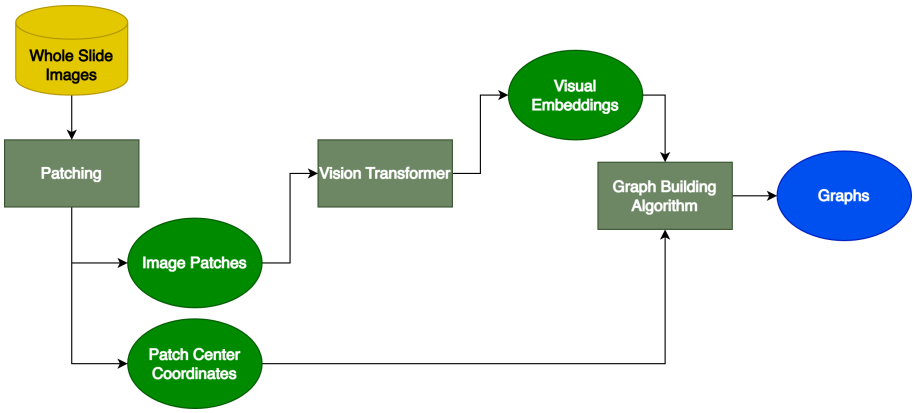


Figure 1: Graph construction from whole slide images.

offer more complex features and semantic information by removing the constraint. As a result, it can capture similarities and differences between different samples with more nuance.

While graph-based architectures constrain relationships to pairwise interactions, hypergraphs provide a natural way to model higher-order relations among multiple entities simultaneously. In this work, we leverage a Hypergraph Neural Network to capture complex dependencies among WSI-level metadata features (e.g., clinical or demographic information).

Each sample is represented as a hypergraph in which hyperedges connect multiple metadata nodes that share semantically related attributes. The constructed hypergraph is then processed through one or more HypergraphConv layers (Feng et al., 2019), which propagate messages along these high-order edges. This allows the network to learn abstract representations that capture the joint influence of multiple metadata factors.

The resulting latent representation is passed through a lightweight MLP to align its embedding space with that of the WSI Graph Transformer Block.

3.0.4 Gated Fusion Block

To integrate the complementary information captured by the two modalities, we employ a Gated Fusion Mechanism. This module adaptively weighs the contributions of the WSI-based and metadata-based embeddings by learning a gating coefficient $a \in [0, 1]$ that balances their relative importance:

$$z_{\text{fusion}} = \alpha z_{\text{graph}} + (1 - \alpha) z_{\text{hypergraph}}$$

where z_{graph} and $z_{\text{hypergraph}}$ denote the respective latent representations.

The fused representation z_{fusion} is then processed through a final MLP classifier that outputs the final binary prediction for cancer recurrence. This mechanism allows the model to adaptively integrate multimodal signals while mitigating potential dominance of one modality over the other.

3.1 Baseline Approaches

We evaluate our model against the state-of-the-art multi-model approach that uses logistic regression and OncoDHNet (Goyal et al., 2024), which happens to be trained and evaluated on the same Darthmouth dataset. The model performs binary classification by binning patients on the Oncotype DX breast recurrence score with the same intervals: 0-25 for low risk and 26-100 for high risk patients.

The OncoDHNet architecture randomly selects 50 regions from a WSI, creating class tokens for each region. It breaks down the randomly selected regions into smaller patches, which are passed into a ResNet-18 to extract feature vectors. These vectors are then passed into a pre-trained transformer, MaskHIT (Jiang et al., 2024), specifically built for histology image analysis. This provides a classification for each of the 50 regions, so they can be averaged to provide the final label for the entire slide.

The clinicopathologic metadata for each WSI were used as features for a simple logistic regression model optimized with grid search to maximize AUC score. The logistic regression model gave a score classification independent of the OncoDHNet model.

The outputs of the two models were combined to provide a single prediction. The prediction was optimized by evaluating different thresholds for confidence scores between the two models to find

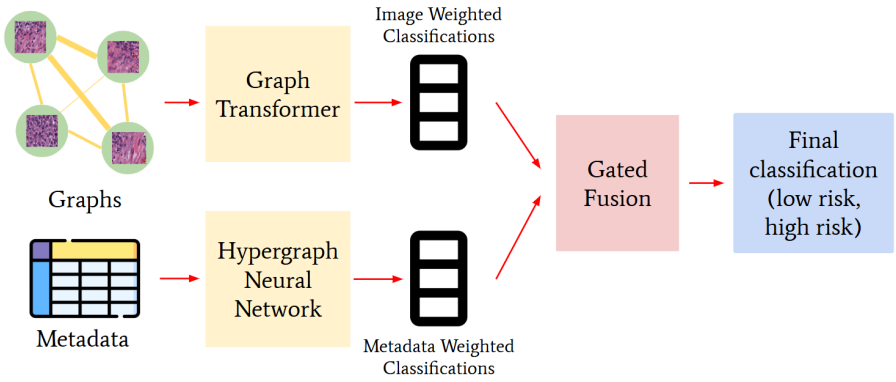


Figure 2: Graph Transformer and Hypergraph pipeline takes graph data (originally image data) and cliniocpathologic metadata to output a single classification label.

the one that best discriminated between the low- and high-risk categories.

3.2 Novelty

Our approach improves upon the state-of-the-art by keeping the same pipeline idea between the whole slide images and clinicopathologic metadata, but integrating a Graph Transformer for the image data that will capture both similarity and spatial relations between patches and a Hypergraph Neural Network for the metadata that will improve upon the simple logistic regression model currently used. For justification on why these models will perform better, read more in Section 2.

4 Experiments

4.1 Data Processing Pipeline

Using the BMIRDS Dartmouth Breast Cancer Recurrent Risk Dataset (Goyal et al., 2024), we prepare data under two modalities for our model. The dataset consists of 990 whole-slide images, which complement image metadata information including age, tumor size and grade, histologic type, etc. From this original dataset, we find that 849 WSIs are accompanied with metadata. We then sent the WSIs and metadata through our WSI and metadata data processing pipeline.

4.1.1 Whole-Slide Images

Whole-slide images can be incredibly intensive in storage and computation due to the detail and gigapixel resolution at which it stores information. We outline our image processing pipeline in Figure 3.

We perform sequential patching with a pretrained DINOv2 ViT-S/14 model of resolution 16×16 on the WSI with a downsample factor of 20 in recognition of precedent on patching size (Krizhevsky et al., 2012) (He et al., 2016). We selected our factor of 20 due to quadratic gains in space and computational efficiency via downsampling without sacrificing too much information.

We then perform graph construction on the patches. Each node represents the information of a single image patch. We build direct eight-neighbor adjacency with edge weights represented by a softmax on cosine similarity for normalization and to preserve positive edge weights. This intends to highlight adjacency between patches that carry similar features, utilizing the inductive bias of graph-based networks.

4.2 Data Split & Model Training

To match our paper baseline (Goyal et al., 2024), we train our model under K-Fold ($K = 5$) with an Adam optimizer (learning rate of 10^{-3} and weight decay of 10^{-5} for 30 epochs per fold due to computation constraints. To address class imbalance and small dataset size, we utilize weighted BCE Loss and Weighted Resampling.

The dataset consists of two strongly imbalanced classes which can deeply skew model predictions towards negative predictions. Under serious class imbalance, AUROC can be very misleading, potentially obfuscating underfitting and a failure of models to generalize (Movahedi et al., 2023). To counteract, we pursued two approaches to mitigate the effects of class imbalance. First, our weighted Binary-Cross-Entropy loss weighs minority class predictions more to ensure generalizability. Second, we impose a weighted random sampler to resample minority classes due to the small size of

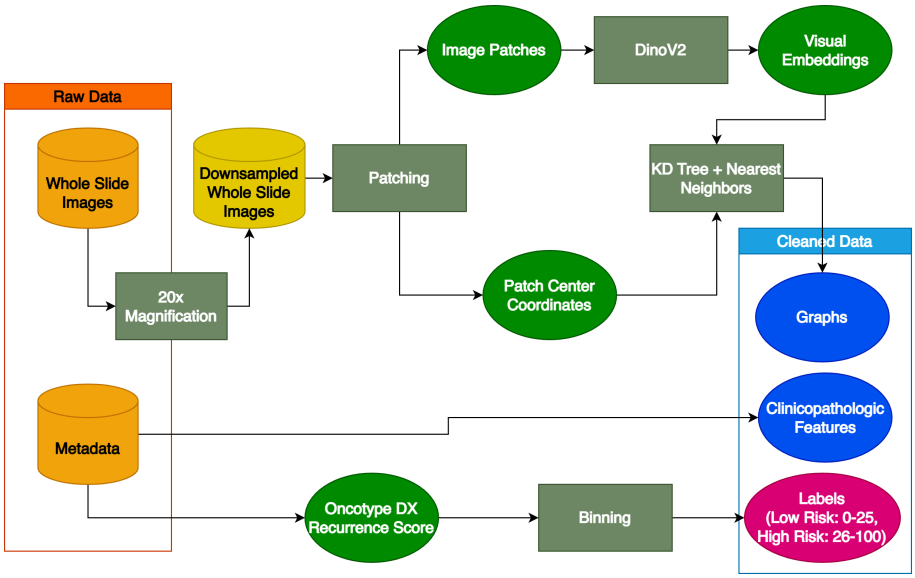


Figure 3: Complete whole slide image and clinicopathologic metadata preprocessing pipeline.

our dataset and class imbalance.

4.3 Metadata Training

Our metadata consists of the following features: age, tumor size, tumor grade, histologic type, ER, PR, and HER markers on breast cancer cells. For categorical features of N, we one-hot encode into $N - 1$ features to prevent multicollinearity and ensure ease of training. For continuous data, we perform quantile-binning. We then train on hold-out with a 80-0-20 split for 500 epochs under a weighted BCE loss function and Adam optimizer.

For our baselines, we use popular approaches including a simple MLP, RandomForest, and XGBoost.

4.3.1 Evaluation Metrics

For our metrics, we include AUROC due to its prevalence, F1-Score as an overall representation of model performance, and Precision and Recall for a deconstructed representation of model's performance in terms of both positive and negative predictions due to the potential of the model underfitting on an imbalanced dataset. While we believe AUROC is a misleading metric for our particular dataset, it is still important as an overall evaluation of the model performance. F1-Score is a well-rounded representation of model performance since it represents the harmonic mean of both precision and recall. Recall helps us capture the Type-II errors we want to avoid due to the consequences of failing to detect a high-risk case. However, while we do want to emphasize recall performance and penalize false-negative predictions over false-positive predictions as precision does, we must consider that class imbalance prevent a robust capture of the model's susceptibility to making false-negative predictions.

4.4 Experimental Settings

For our WSI models, we trained with the following hyperparameters: k folds, epochs, batch size, learning rate, weight decay, dropout, and attention type for our Transformer model.

We chose our K folds value based on our baseline paper. We set the number of epochs based on the number of folds and computational demands of training locally. We set batch size to 8 to reduce the noise of training updates. We set the initial learning rate and weight decay for Adam optimizer based on common starting values without evaluating too-large of a portion of our small dataset at a time. For attention type, GPS convolution supports both multi-headed attention (MHA) and Performer attention. We set dropout to 0.3 to counteract overfitting and promote generalization.

For our metadata models, we plan on optimizing our hyperparameters when their output is fused with our WSI model output to help produce a classification. We plan on experimenting with the same hyperparameters as listed above, including L_p regularization, number of layers, etc.

We plan on using GridSearch to optimize all of the listed hyperparameters for our WSI models and their joint fusion with our metadata models.

All models were trained locally on the following

machines: 14-core Apple M4 CPU Sequoia 15.6.1 MAC laptop, Intel Core Ultra 7 155U CPU and integrated GPU 16 GB RAM laptop, RTX 4070 GPU and Intel i9-13879HX CPU 16 GB DDR5 RAM laptop, and RTX 2070 GPU and Intel i9-9900k CPU 32 GB DDR4 RAM desktop PC [1].

4.5 Results

With the previously listed hyperparameters, we have achieved the results in Table 1.

It is important to note that the baseline study by Goyal et al. (2024) did not report F1-Score, Precision, or Recall, limiting direct comparison beyond AUROC. The lack of balance sensitive metrics makes it difficult to compare the performance differences between our models.

We consider the F1-Score an important and robust performance measure for this task, as AUROC alone can be misleading in the presence of significant class imbalance. Before we applied the methods to handle the imbalance we outlined in Section 3, our AUROC was comparable to the external baseline despite our model being practically unable to detect high risk cases.

After implementing methods to account for imbalance, however, our observed F1-Scores remain notably low across all tested graph-based architectures, which raises concerns regarding the effectiveness of our current data processing and graph construction pipeline. This suggests that further experimentation is warranted—particularly with respect to the patching mechanism and graph connectivity. Future work may explore optimizing patch size, balancing resolution trade-offs via more aggressive downsampling, replacing random patch sampling with spatially sequential patching, and improving message-passing strategies to enhance feature communication between nodes.

While relatively high Recall scores were achieved, this result is likely influenced by the dataset imbalance which means these values may not reflect meaningful sensitivity to minority class instances (Section 5).

5 Analysis

5.1 Performance Analysis

Multi-head attention resulted in a higher precision value, while Performer attention resulted in a higher recall and F1 score as seen in Table 1. Based on these scores, the Performer attention resulted in better overall model performance and minimized false negatives. This may be due to the Performer Attention's linear attention mechanism demonstrating an ability to potentially handle longer sequences, meaning that some predictions may rely on non-local relationships.

The discrepancy between the state-of-the-art AUC of 0.910 (Goyal et al., 2024) and our HELP model AUC of 0.745 can be explained by the authors' use of the OncoDHNet model for the whole slide image component of their pipeline, which was pre-trained on 2244 whole slide images for about 20 million image patches. We base this on the low AUC, precision, and F1 we found for the ResNet-50 and ViT-B/16, the same models used in the whole-slide image OncoDHNet component of their architecture. Since we cannot produce comparable results on our models without access to the same data (or even a similar amount and type of data), we cannot make a sound comparison between the two architectures or even reproduce their results.

However, despite not being able to do the same pre-training to compare models, the ResNet-50 and ViT-B/16 still serve as components in their whole-slide image OncoDHNet model that we can compare to our graph-based approaches. Our graph-based models outperform the generic CNN ResNet-50 baseline on all metrics except for recall. ResNet-50 achieves an AUC of 0.6216 and F1 of 0.2968 with perfect recall (1.0) but very low precision (0.1743), while the frozen ViT baseline performs even worse in terms of AUC (0.5373) and F1 (0.3029), as shown in Table 1. This shows that, in the context of training solely on the Dartmouth dataset without external training, Graph-Transformer and GAT-based models appear to offer competitive performance over CNNs and vision transformers.

Our metadata models perform noticeably better across all metrics compared to our model on the image modality. Random Forest performed the best on AUC, which is attributable to incredibly high recall. However, it had underwhelming precision, indicating a very high concentration of false positive predictions and a high possibility of underfitting its predictions to a singular class. XGBoost had the highest F1-Score and Precision, with competitive AUC and Recall metrics, offering more balanced performance. Our stronger performance overall across all metadata models compared to the

[1] Our source code can be found on github.com/Fobertree/CancerRecurrencePrediction

Model (Fused)	AUC	F1	Precision	Recall
GraphTransformer + HGNN + Gated Fusion	0.745	**0.437**	**0.341**	**0.674**
GraphTransformer + MLP + Gated Fusion	0.659	0.359	0.318	0.412
External Benchmark (Fused)				
ViT + CNN + Metadata Logistic Regression (Goyal et al., 2024)	**0.910**	-	-	-

Table 1: Performance comparison of complete models on the entire dataset including image data and metadata. All metrics are averaged over k-fold cross-validation.

Model (Image)	AUC	F1	Precision	Recall
GAT	**0.760**	0.422	0.369	0.593
GraphTransformer (MHA)	0.730	0.421	**0.381**	0.572
GraphTransformer (Performer)	0.718	**0.439**	0.347	0.703
ResNet-50 (frozen)	0.622	0.297	0.174	**1.000**
ViT-B/16 (frozen)	0.537	0.303	0.174	**1.000**

Table 2: Performance comparison of models on the whole slide image data. All metrics are averaged over k-fold cross-validation.

Model (Metadata)	AUC	F1	Precision	Recall
MLP	0.801	0.539	0.429	0.727
RandomForest	**0.966**	0.308	0.182	**1.000**
XGBoost	0.889	**0.659**	**0.528**	0.875
Hypergraph Classifier	0.750	0.529	0.514	0.545

Table 3: Performance comparison of models on the metadata. All metrics are averaged over 3 runs.

whole slide image models indicates a strong relationship between clinicopathologic metadata and breast cancer recurrence outcome.

5.2 Error Analysis

Between the two attention variants, Multi-Head Attention (MHA) yielded higher precision, whereas Performer attention achieved higher recall and F1 scores, as shown in Table 1. This indicates that Performer attention provided better overall performance, particularly by reducing false negatives, while MHA tended to reduce false positives. The superior generalization of Performer attention is likely due to its comparatively lower computational complexity. MHA's higher representational capacity may have caused it to overfit, especially given the limited number of whole-slide image samples available for training.[2]

Label	Samples
Low Risk	825
High Risk	168

Table 4: Breakdown of Dartmouth Breast Cancer Recurrence Risk Dataset by label.

We set a weighted BCE loss for our models to account for the data imbalance shown in Table 4. Although this improved our metrics, the image models and the metadata models still have higher recall than precision, showing the models struggle to accurately detect positive cases. Without this loss, the overall performance would be even worse.

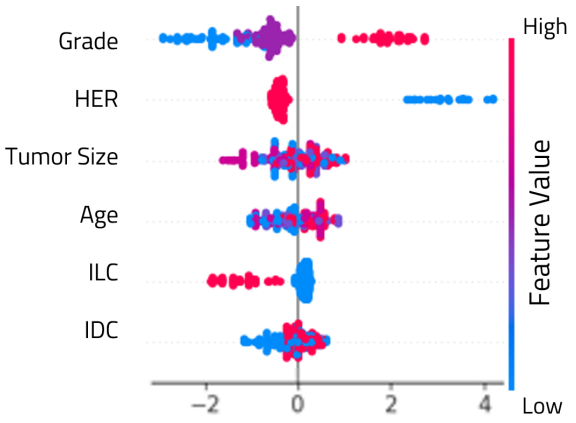


Figure 4: SHAP summary for metadata models.

Furthermore, because we avoided image data augmentation to maintain comparability with our baseline model, the model may not have been exposed to sufficient visual variability during training. Introducing controlled image augmentations that preserve class ratios could help counter overfitting to oversampled minority examples, thereby reducing false positives while retaining improved recall.

[2]Unfortunately, we are not able to include specific examples of image classification errors due to the privacy nature of the data agreement

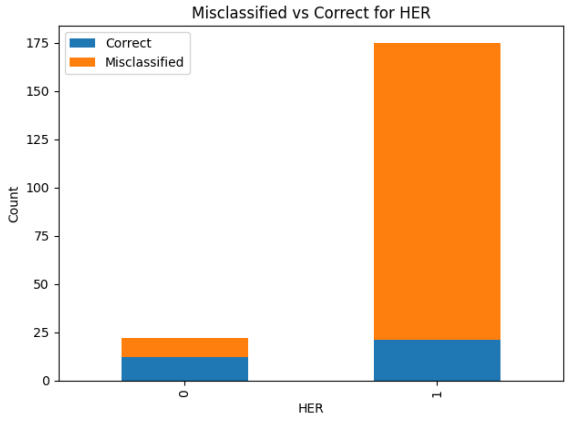


Figure 5: HER2 Receptor misclassification distribution.

To analyze our most important features, the SHAP summary in Figure 4 shows that tumor grades 1 and 2 (blue and purple, respectively) caused the models to predict a patient is low-risk while grade 3 (red) pushed the model to high-risk prediction. This is expected behavior, as an increase in tumor grade means the cells appear more abnormal. The tumor size doesn't seem to offer as meaningful of a discrimination, and age follows an expected distribution where older ages push the model to high risk prediction.

However, negative HER2 status appears to push the model hard into predicting high-risk while positive HER2 status pushes the model toward low-risk. This is contradictory; a positive HER2 status would mean aggressive cell growth linked to cancer, while negative HER2 status would be normal. Upon further analysis in Figure 5, we can see a large portion of the positive HER2 status cases were misclassified. Breast cancers involving positive HER2 protein receptors in general are more difficult to detect and make conclusions about than those with negative HER2, so the model appears to be struggling with the potential unreliability of HER2 status in risk prediction.

5.3 Discussions

Our study is limited by our lack of compute resources to handle the scale of the whole slide images and our shortage of data due to the patient privacy concerns that limit public availability.

False positives remain a major challenge for all our tested models. We attribute this to a combination of class imbalance, oversampling during training, and limited variation in the positive class. The mismatch between the artificially balanced training distribution and the more imbalanced validation distribution amplified the model's bias toward predicting positive recurrence. Furthermore, the absence of image augmentation—preserved to maintain consistency with the baseline model—may have contributed to overfitting on repeated minority-class patches, further elevating false-positive rates. Although false positive errors are less consequential than false negative errors in the medical field, our model still makes too many false positive predictions to see actual deployment and practical use in the medical field.

We identify the following research directions for our work:

- **Larger and more diverse WSI datasets**: For image-based models, increasing dataset size through augmentation would alleviate issues of class imbalance, reduce overfitting, and improve generalizability. Furthermore, we have not considered training with segmentation masks, which may help isolate certain details we want to emphasize. We found that our model performed as well as foundational ResNet CNN and ViT models without pre-training, and hypothesize our model requires a larger training corpus for semantic understanding.

- **Alternative Approaches with Positional Encodings and Representation Learning**: Experiments can be performed with fine-tuned contrastive learning approaches such as CLIP with a MedVLM can offer the potential to improve semantic understanding of WSIs and improve model performance. Our pre-trained DINOv2 model is not fine-tuned for medical imaging tasks, and was trained on the LVD-142M dataset, a non-public natural image dataset.

- **Reducing aggressive downsampling and upsampling**: Due to the computational memory and processing efficiency costs of high resolution images, we downsampled the WSIs by $20\times$, processed patches in sizes of 16×16, then aggressively upsampled the images to fit a 384×384 shape for our DINOv2 input dimensions. The strong upsampling is likely to lose heavy amounts of important image information. To address this, larger image patches can be taken with reduced downsampling.

- **Rotation Equivariance**: Imposing rotation equivariance through the addition of EGNN (Satorras et al., 2022) blocks can potentially improve model performance in our image block. Intuitively, tissue in slides are not manually corrected to a specific rotation, which may hurt our model's ability to analyze semantic information on tissue.

- **Related Image Data Sources**: Patients usually undergo initial imaging procedures such as a mammogram, ultrasound, or MRI that could offer additional patient history for better context.

6 Conclusion

In this work, we present a hybrid Graph Transformer and Hypergraph Neural Network Classifier model that outperforms a ResNet-50 and ViT-B/16 in predicting breast cancer recurrence risk.

Although the results of our hybrid approach cannot be compared to the results presented in Goyal et al. (2024), our approach shows better performance than ResNet and ViT baseline models for predicting whether a patient is low-risk or high-risk for breast cancer recurrence based on a small, imbalanced dataset with whole slide images and image metadata.

In particular, our Graph Transformer models offer better performance in F1-Score and Precision. However, we note the weaker performance of our image-based models compared to traditional machine-learning methods on the accompanying metadata, which implies the metadata is a stronger predictor of breast cancer recurrence compared to the WSIs.

Despite better performance than the baselines, our model lacks in AUC and Recall performance. We plan to tackle these challenges by introducing a new PE approach with a medical-VLM or domain-specific input through CLIP to fine-tune our positional encoding with semantic context of human tissue in WSIs. We also struggled to properly counteract challenges with class imbalance and dataset size, as all models seem to struggle on our binary classification task. To address this issue, we plan on pre-training our model on a significantly larger dataset of WSI images for a stronger understanding of medical histology, such as the data corpus from the TAILORx trials (Sparano et al., 2018). Furthermore, we plan on exploring graph and image data-augmentation techniques to improve training and generalizability without generating data leakage in training.

Acknowledgments

Special thanks to the professors who contributed their input and expertise to this paper: Jinho Choi, Carl Yang, Shengpu Tang, and Judy Gichoya; and to Grace Byun for being an excellent source for suggestions about our paper and for her willingness to put in extra work to answer questions and re-evaluate our writing several times.

References

ACS. 2022. The global cancer burden.

Loan Dao and Ngoc Quoc Ly. 2024. Recent advances in medical image classification. *International Journal of Advanced Computer Science and Applications*, 15(7).

Kero Djoumessi, Samuel Ofosu Mensah, and Philipp Berens. 2025. A hybrid fully convolutional cnn-transformer model for inherently interpretable disease detection from retinal fundus images. *Preprint*, arXiv:2504.08481.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

FDA. 2017. Fda allows marketing of first whole slide imaging system for digital pathology. *FDA News Release*.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks.

Marco Gori, Gabriele Monfardini, and Franco Scarselli. 2005. A new model for learning in graph domains. In *IEEE International Joint Conference on Neural Networks*.

Manu Goyal, Jonathan D. Marotti, Adrienne A. Workman, Seth K. Tooker, Graham M. andRamin, Elaine P. Kuhn, Mary D. Chamberlin, Roberta M. diFlorio Alexander, and Saeed Hassanpour. 2024. A multimodel approach integrating whole-slide imaging and clinicopathologic features to predict breast cancer recurrence risk. *npj breast cancer*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition.

Taghreed S. Ibrahim, M. S. Saraya, Ahmed I. Saleh, and Asmaa H. Rabie. 2025. An efficient graph attention framework enhances bladder cancer prediction. *Scientific Reports*.

Shuai Jiang, Liesbeth Hondelink, Arief A. Suriawinata, and Saeed Hassanpour. 2024. Masked pre-training of transformers for histology image analysis. *Journal of Pathology Informatics*.

Koushik Sivarama Krishnan and Karthik Sivarama Krishnan. 2021. Vision transformer based covid-19 detection using chest x-rays. In *IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. *neurIPs*.

Neeta Kumar, Ruchika Gupta, and Sanjay Gupta. 2020. Whole slide imaging (wsi) in pathology: Current perspectives and future directions. *J Digit Imaging*.

Kunal Kwadkar. 2025. Comparative analysis of vision transformers and convolution neural networks for medical image classification. *Preprint*, arXiv:2507.21156.

Minhyeok Lee. 2023. Recent advancements in deep learning using whole slide imaging for cancer prognosis. *Bioengineering*, 10(8).

Bingjun Li and Sheida Nabavi. 2024. A multimodal graph neural network framework for cancer molecular subtype classification.

Jie Lian, Yonghao Long, Fan Huang, Kei Shing Ng, Faith M Y Lee, David C L Lam, Benjamin X L Fang, Qi Dou, and Varut Vardhanabhuti. 2022. Imaging-based deep graph neural networks for survival analysis in early stage lung cancer using ct: A multicenter study. *Radiology: Artificial Intelligence*.

Wenqi Lu, Michael Toss, Muhammad Dawood, Emad Rakha, Nasir Rajpoot, and Fayyaz Minhas. 2022. Slidegraph+: Whole slide image level graphs to predict her2 status in breast cancer. *Medical Image Analysis*, 80:102486.

Faezeh Movahedi, Rema Padman, and James F. Antaki. 2023. Limitations of receiver operating characteristic curve on imbalanced data: Assist device mortality risk scores.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, arXiv:2304.07193.

Wei-Feng Qu, Meng-Xin Tian, Jing-Tao Qiu, Yu-Cheng Guo, Chen-Yang Tao, Wei-Ren Liu, Zheng Tang, Kun Qian, Zhi-Xun Wang, Xiao-Yu Li, Wei-An Hu, Jian Zhou, Jia Fan, Hao Zou, Ying-Yong Hou, and Ying-Hong Shi. 2022. Exploring pathological signatures for predicting the recurrence of early-stage hepatocellular carcinoma based on deep learning. *Frontiers in Oncology*, Volume 12 - 2022.

Ladislav Rampášek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. 2023. Recipe for a general, powerful, scalable graph transformer. *Transactions on Machine Learning Research*, arXiv:2205.12454.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. 2022. E(n) equivariant graph neural networks. *Preprint*, arXiv:2102.09844.

Jaqueline Alvarenga Silveira, Alexandre Ray da Silva, and Mariana Zuliani Theodoro de Lima Lima. 2025. Harnessing artificial intelligence for predicting breast cancer recurrence: a systematic review of clinical and imaging data. *Discover Oncology*.

Joseph A. Sparano, Robert J. Gray, Della F. Makower, Kathleen I. Pritchard, Kathy S. Albain, Daniel F. Hayes, Charles E. Geyer, and George W. Sledge. 2018. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, pages 10347–10357.

Yanan Wang, Yu Guang Wang, Changyuan Hu, Ming Li, Yanan Fan, Nina Otter, Ikuan Sam, Hongquan Gou, Yiqun Hu, Terry Kwok, John Zalcberg, Boussioutas, Roger J Daly, Guido Montufar, Pietro Lio, Dakang Xu, Geoffrey I Webb, and Jiangning Song. 2022. Cell graph neural networks enable the precise prediction of patient survival in gastric cancer.

Zhikang Wang, Jiani Ma, Qian Gao, Chris Bain, Seiya Imoto, Pietro Liò, Hongmin Cai, Hao Chen, and Jiangning Song. 2024. Dual-stream multi-dependency graph neural network enables precise cancer survival analysis. *Medical Image Analysis*.

WHO. 2025. Cancer.

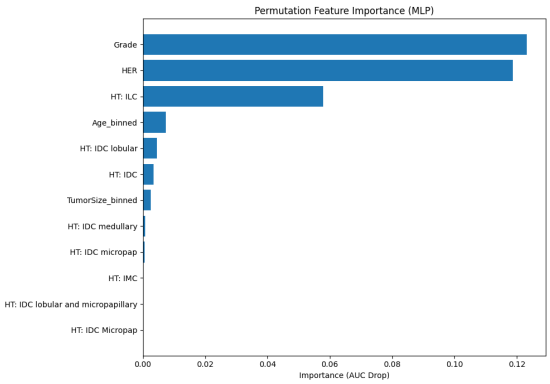
Huimin Xiao, Guanghua Yang, Zhuocheng Li, and Changhua Yi. 2025. Gnns surpass transformers in tumor medical image segmentation. *Scientific Reports*, 15(1).

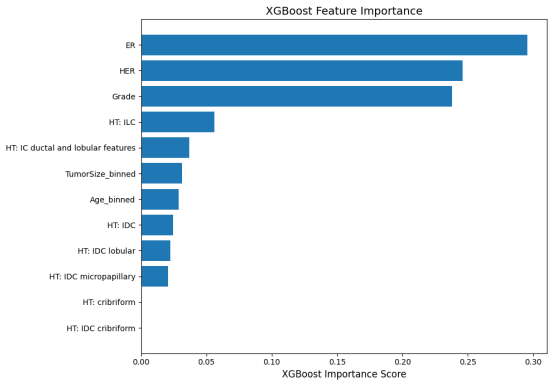
Ran Xu, Mohammed K Ali, Joyce Ho, and Carl Yang. 2023. Hypergraph transformers for ehr-based clinical predictions. In *AMIA Jt Summits Transl Sci Proc.*

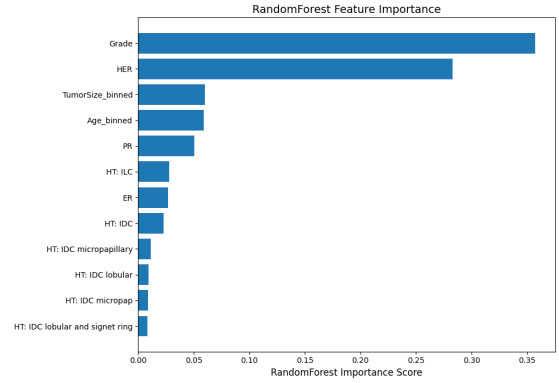
D. Zuo et al. 2023. Machine learning-based models for the prediction of breast cancer recurrence. *BMC Medical Informatics and Decision Making*, 23(377).

A Appendix

A.1 Feature Importance







Feature importance for each model. Note that the "Histologic Type" feature was one-hot encoded, and we display these individually to visualize how each type influences the model.

Tumor grade and HER2 protein status were consistently among the important features for all metadata models. Estrogen receptor (ER) status was the most important feature for XGBoost, which is important to note given the significant performance advantage with XGBoost compared to MLP and RandomForest.