Histopathology Patch Classification Using Gradient Boosting and a Multilayer Perceptron

Jackson Walker
Department of Electrical and Computer Engineering, Temple University
Jackson.walker@temple.edu

Introduction: This paper presents two machine learning systems for the histopathology patch classification task of microscopy images of breast tissue. After a 2D Discrete Cosine Transform (DCT) is applied to a 256x256 RGB patch, each patch is a 3,072-dimensional vector of DCT coefficients representing three 32×32 image channels. The DCT converts the image into a frequency-domain representation, and dimensionality is reduced by retaining only the low-frequency 32×32 coefficients per channel, which preserve the most informative structural content. The goal of this project was to classify patches into six tissue categories: normal (norm), non-neoplastic (nneo), inflammatory (infl), ductal carcinoma in situ (dcis), invasive ductal carcinoma (indc), and background (bckg). Labels 1, 4, and 7 (artifact, suspicious, null) were ignored during scoring. The official scoring metric is Score = 0.90 × avg_error(norm, nneo, infl, dcis, indc) + 0.10 × error(bckg), where lower is better and all five primary classes are weighted equally regardless of the sample frequency. The training set contains 9,340 usable samples with significant class imbalance: we found that nneo accounts for 33.5% of samples while dcis accounts for only 5.5%; however, both contribute equally to the score. This mismatch in the abundance of samples is the central challenge of the task. Feature engineering transforms each raw DCT vector into a 6,262-dimensional representation. So, for each channel, we reconstructed the DCT block into the spatial domain with a 2D Inverse DCT (IDCT). The engineered features include DCT zone energy ratios across a total of five frequency bands, spatial statistics (mean, std, four percentiles), higher-order moments (skewness and kurtosis), gradient magnitude statistics, and an 8-bin direction histogram, LBP-lite texture variance, radial DCT energy bands, center-versus-border spatial contrast, and cross-channel correlation and intensity ratio features. All these features help describe tissue uniformity, irregular boundaries, tissue density, and color relationships. Something important we found through cross-validation was that dcis and nneo were nearly inseparable in this feature space: a binary dcis-vs-nneo classifier achieved only 41% recall for dcis, with 59.7% of dcis samples falling into nneo. This told us that these features were greatly overlapping each other in the feature space. This suggests that both classes must share similar glandular structures, and classification depends on the larger-scale spatial context. This finding motivated the use of aggressive class weight boosting and post-hoc probability rescaling as the primary strategy for improving minority class recall.

Algorithm No.1 Description: Hierarchical Gradient Boosting Classifier (HGB). Gradient boosting builds a collection of decision trees sequentially, where each tree is trained to approximate the gradient of the loss function, effectively focusing on correcting the model's current errors.

$$F_M(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

Where $h_m(x)$ are the individual decision trees and γ_m are their corresponding weights.

$$r_i^{(m)} = -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}$$

At each iteration, a new tree is trained to fit the negative gradient of the loss function, which represents the direction of greatest error reduction. Gradient boosting handles nonlinear feature interactions well and is robust to mixed feature types. Scikit-learn's HistGradientBoostingClassifier uses histogram-based feature binning for memory-efficient training on high-dimensional data. We

designed a three-stage hierarchical architecture to match the biological structure of the classes. This seemed to be the most logical approach. Stage 1 is a 3-way classifier separating bckg, non-cancer (norm, nneo), and cancer (infl, dcis, indc). Isolating background at Stage 1 was important because bckg patches were found to severely mix in with the norm/nneo boundary when included in the same classifier. Stage 2a classifies norm vs. nneo within the non-cancer branch. Stage 2b classifies the three cancer subtypes (infl, dcis, indc). A parallel flat 6-class model is trained simultaneously. Its probability outputs are mixed into the hierarchical predictions for the hard classes using weights of 0.5 for dcis, 0.3 for infl, and 0.3 for indc, allowing the model to leverage global boundaries that individual branches cannot see. The flat model, which is not constrained by the hierarchy, can capture alternative decision boundaries and provide corrective probability mass to ambiguous classes. Post-hoc probability rescaling is applied before argmax to align decision boundaries with the equal-weight scoring formula: dcis ×11.5, indc ×4.5, infl ×8.0, norm ×1.3. These multipliers were tuned systematically on the development set. A lot of time was spent systematically evaluating different combinations of class weights. This was the heuristic part of this model. Training used a max_depth=10–12, max_iter=300–400, learning_rate=0.05, min_samples_leaf=12, and l2_regularization=0.1. The final model was retrained on the combined train and development sets before generating the final eval predictions.

Algorithm No. 2 Description: Multilayer Perceptron (MLP). A multilayer perceptron is a feedforward neural network in which input features are passed through successive layers of learned linear transformations followed by non-linear activation functions.

$$h^{(l)} = \sigma(W^{(l)}h^{(l-1)} + b^{(l)})$$

Where $W^{(l)}$ and $b^{(l)}$ are the weights and biases of the layer l and $\sigma(\cdot)$ is a nonlinear activation function. Each layer learns increasingly complex nonlinear combinations of the input features. A key design decision was to restrict the MLP input to the compact engineered feature subset (approximately 118 statistical features) rather than the full 6,262-dimensional vector. We found that with only 9,340 training samples, feeding a neural network 6,262 features causes severe overfitting. This was observed experimentally as the model achieved low training loss but failed to generalize to the development set. The 118-feature subset provides a sample-to-feature ratio that permits good generalization. The architecture is three layers: Linear(118→256) + BatchNorm1d + ReLU + Dropout(0.3), Linear(256→64) + BatchNorm1d + ReLU + Dropout(0.2), Linear(64→9). Batch normalization normalizes each layer's inputs to accelerate training and reduce sensitivity to initialization. Dropout randomly zeroes a fraction of activations during training to prevent neuron co-adaptation. Gaussian noise (σ=0.03) was added to inputs during training as data augmentation. Training used AdamW (lr=1e-3, weight_decay=1e-2) with cosine learning rate adjustment over 25 epochs and batch size 128. Class weights and SCORE_MULTIPLIERS were tuned independently from the HGB model since neural networks are more sensitive to extreme weight ratios: dcis ×1.5, indc ×2.5, infl ×5.0, norm ×1.0. The final model was retrained on the train and dev datasets combined.

Results: Table 1 shows the error scores for both models across all three data splits. Lower is better.

	Data Set		
Algorithm	**Train**	**Dev Test**	**Eval**
HGB	26.33%	26.62%	49.88%
MLP	20.70%	21.24%	37.17%

Table 1. Final error scores for each model and dataset

Both models improved substantially over the course of development, beginning from a baseline error above 68%. The HGB model reached 26.62% on dev and 49.88% on eval. The MLP achieved 21.24% on dev and 37.17% on eval, outperforming HGB on both splits. The gap between dev and eval scores, particularly for HGB, suggests possible overfitting on the dev set through multiplier tuning. The MLP generalized more consistently, likely because the compact 118-feature input prevented memorization. The most obvious source of error in both models was the dcis/nneo boundary, which we determined is a fundamental limitation of single-patch DCT features rather than a modeling failure.

Conclusions: Two independent classifiers were developed for the histopathology patch classification task. The HGB model used a three-stage hierarchical structure with a parallel flat model and probability blending to handle the overlapping class boundaries. The MLP used compact engineered features, batch normalization, dropout, and moderate class weights to prevent overfitting on the small training set. The MLP achieved a better result: 37.17% eval error versus 49.88% for HGB. Both models relied on the key information that the scoring formula weights all five primary classes equally, regardless of frequency. This required deliberate upweighting of rare cancer subtypes and post-hoc probability rescaling aligned to the metric. Future improvements could explore patch context beyond single tiles or multi-scale DCT representations.