Comparative Analysis of Machine Learning Algorithms for DCT-Based Digital Pathology Classification

Faezeh Bohlool
Department of Electrical and Computer Engineering, Temple University
Faezeh.bohlool@temple.edu

Introduction: Digital pathology has increasingly adopted computational methods to support diagnostic workflows, particularly in tasks involving large-scale image classification. In this project, we were provided with a dataset derived from the TUH DPATH Breast v4.0.0 collection. Instead of raw images, each pathology patch was transformed into a compressed frequency-domain representation using the Discrete Cosine Transform (DCT). For each RGB channel, the top-left 32×32 block of the 256×256 DCT was retained, resulting in 3072 features per sample (3 × 32 × 32). Each row in the dataset contains a numeric label (0–8) followed by these coefficients.
Although nine labels exist, only six are relevant for evaluation: 0 = norm, 2 = nneo, 3 = infl, 5 = dcis, 6 = indc, 8 = bckg. The remaining labels (1 = artf, 4 = susp, 7 = null) are explicitly ignored by the scoring program. The evaluation metric is a custom weighted. It computes the average error across the five tissue classes (0, 2, 3, 5, 6) and combines it with the background error (8) using a 90/10 weighting. The goal of the project was to implement two distinct machine-learning algorithms (one non-neural and one neural) and compare their performance on the training, development, and evaluation datasets.

Methods: The first challenge was understanding the structure of the DCT-based dataset and the implications of the custom scoring metric. Because the scoring script ignores labels 1, 4, and 7 entirely, we removed these samples from all datasets before training or evaluation. A custom scoring function was implemented to exactly replicate the instructor's script, including:

- A full 9×9 confusion matrix
- Per-class error computation
- Averaging over tissue classes
- Weighted combination with background error

This function became the central evaluation tool throughout the project. We then began by experimenting with several classical machine-learning models, including: Logistic Regression, Linear SVM, Random Forest, HistGradientBoosting, and MLP (simple neural network). These models were evaluated using a 5-fold cross-validation on the training set. Early results showed that:

- Linear SVM and Logistic Regression performed reasonably well but struggled with nonlinear class boundaries.
- Random Forest and HistGradientBoosting showed promise but were sensitive to hyperparameters.
- A simple MLP performed inconsistently due to the high dimensionality of the input.

This exploratory phase revealed that more sophisticated models were needed, particularly for capturing nonlinear relationships in the DCT feature space. After extensive experimentation, two algorithms emerged as the strongest candidates: Non-Neural Model: LightGBM, Neural Model: CNN

LightGBM: LGBM is well-suited for high-dimensional tabular data and can model nonlinear interactions between DCT coefficients. It also supports class weighting, which is essential due to the dataset's imbalance. Here we used:

- 5-fold stratified cross-validation
- Sample weights proportional to inverse class frequency
- Additional boosting for rare tissue classes (dcis, indc)
- Reduced weight for background to match the scoring emphasis

To obtain an unbiased estimate of training performance, we generated OOF predictions across all folds. This allowed us to compute a realistic training score without overfitting. For the final training, after cross-validation, the model was retrained on the full filtered training set using the same sample-weighting scheme.

CNN: At the beginning, we started with a simple CNN, which resulted in an error rate above 50. Then, because the dataset provides only DCT coefficients, we decided to reconstruct approximate spatial images using the inverse DCT (IDCT). These images preserve low-frequency structure, such as color gradients and coarse tissue morphology.

In image reconstruction, for each sample, the algorithm:

1. Reshape 3072 coefficients into (3, 32, 32)
2. Embed into a zero-padded (3, 256, 256) DCT grid
3. Apply 2D IDCT to each channel
4. Downsample to 64×64 for computational efficiency
5. Normalize using the training-set mean and standard deviation

For the CNN architecture, we designed a 5-block convolutional network:

- Conv → BatchNorm → ReLU → MaxPool
- Repeated with increasing channel depth (32 → 64 → 128 → 256)
- Adaptive average pooling
- Fully connected classifier with dropout
- Final output layer with 9 logits (matching original label space)

For training, we did:

- Cross-entropy loss with class weights
- AdamW optimizer, Learning-rate scheduler (ReduceLROnPlateau)
- Early stopping with patience of 12 epochs
- Batch size 64, up to 70 epochs

This setup stabilized training and prevented overfitting.

Results: The two models exhibited distinct performance profiles across classes and datasets. The CNN achieved substantially lower error rates for most tissue classes, with 32.74% for norm, 48.62% for nneo, 14.38% for infl, 37.50% for dcis, 36.09% for indc, and 8.09% for background. In contrast, LightGBM showed higher error on several key classes, with 43.55% for norm, 32.38% for nneo, 49.06% for infl, 98.31% for dcis, 96.03% for indc, and 10.04% for background. At the aggregate level, LightGBM obtained weighted errors of 58.11 (train), 58.48 (dev), and 57.72 (eval), whereas the CNN achieved markedly lower scores of 12.55, 31.29, and 37.44, respectively. These results confirm that the CNN, trained on IDCT-reconstructed images with class-weighted loss and regularization, outperformed LightGBM under the project's weighted error metric and provided the most effective overall model.

LightGBM Detailed Results

Class	Label	Error (%)
0	norm	43.55%
2	nneo	32.38%
3	infl	49.06%
5	dcis	98.31%
6	indc	96.03%
8	bckg	10.04%

CNN Detailed Results

Class	Label	Error (%)
0	norm	32.74%
2	nneo	48.62%
3	infl	14.38%
5	dcis	37.50%
6	indc	36.09%
8	bckg	8.09%

Final results

Algorithm	Train	Dev	Eval
GBM	58.11%	58.4828%	57.7249%
CNN	12.5465%	31.2880%	37.4418%