**ECE 8527: Digital Pathology Final Project**

*Claudia-Anca Dumitrescu*

Department of Electrical and Computer Engineering, Temple University

claudia-anca.dumitrescu@temple.edu

**Introduction:** Digital pathology datasets derived from discrete cosine transform (DCT) coefficients carry a unique statistical signature: they are intrinsically sparse and exhibit non‐Gaussian marginal distributions. By transforming image patches into the DCT domain, many coefficients approach zero or cluster tightly around low‐frequency components, while a handful capture the critical high‐frequency edges and texture variations that distinguish pathological structures. These characteristics, namely sparsity, heavy tails, and pronounced skew, pose distinct challenges for model selection, feature engineering, and regularization.

In this work, we first harness an AutoML framework, to traverse a broad algorithmic landscape: from linear and probabilistic classifiers to tree‐based ensembles, each combined with class‐imbalance strategies tailored to our sparse DCT features. We then validate our top performer with a five‐fold cross‐validation, observing stable F1 performance, which attests to its robustness across data splits. Finally, we probe deep‐learning architectures trained on images re-reconstructed from our DCT data at an alternative resolution (a deliberate variation introduced to assess model sensitivity to preprocessing). This experiment sheds light on how convolutional inductive biases interact with the reconstructed signal and integrating coefficient-based and pixel-based representations.

**LightGBM:** The DCT coefficients in our dataset exhibit pronounced sparsity and heavy-tailed distributions - traits that violate the Gaussian assumptions. AutoML’s evaluation across 32 classifiers revealed that probabilistic models underperformed because their assumptions treat the dense, near-zero regions uniformly, obscuring the contribution of rare, high-amplitude coefficients. In contrast, boosting methods such as LightGBM and XGBoost ranked among the top performers. These models excelled not because of any inherent superiority, but because their iterative error reweighting and decision-tree structures adapt naturally to skewed, nonparametric, and high-dimensional spaces - precisely our DCT features.

The No Free Lunch theorem reminds us that, without strong priors on the true data‐generating process, which in our DCT feature space is too complex to characterize, no single algorithm can be expected to excel across every task. Here, both the sparsity pattern and heavy‐tail behavior of coefficients invalidate common assumptions such as smooth, low‐dimensional decision boundaries or Gaussian noise. By deploying AutoML, we avoid anchoring on one inductive bias: the framework evaluates linear, kernel, tree‐based, and ensemble methods under a unified pipeline, tuning each to our data’s idiosyncrasies. This breadth ensures our choice rests on empirical performance, not unverifiable assumptions about smoothness or separability.

LightGBM’s strong performance on our DCT feature space stems from its ability to natively accommodate sparse and skewed data distributions through histogram-based binning and leaf-wise gradient boosting. Its tree growth strategy prioritizes high-gain splits, enabling deeper exploration of the few high-frequency coefficients that carry the most discriminative signal, while regularization controls (e.g., reg\_alpha and num\_leaves) mitigate overfitting to noise. The model’s handling of missing values and categorical features further aligns with the zero-dominated structure of DCT vectors. Hyperparameter tuning via nested cross-validation (mean F1 ≈ 0.5717 ± 0.0128) confirmed its consistent generalization. Class imbalance was addressed through inverse-frequency weighting, and a low learning rate (0.05) alongside L1 regularization effectively pruned uninformative, near-zero splits—highlighting LightGBM’s architectural synergy with our feature space, where models with incompatible inductive biases faltered.

**Deep Learning:** Our deep‐learning experiments relied on two state‐of‐the‐art, ImageNet‐pretrained backbones, each fine‐tuned in a two‐stage protocol (freeze base, train head; then unfreeze the last ~20 layers and continue at a lower learning rate). This strategy preserves broad, general features learned on ImageNet and avoids catastrophic forgetting, while later enabling targeted adaptation of high-level filters to the fine-grained, high-frequency textures in our DCT-reconstructed pathology images.

**EfficientNetB0** combines mobile inverted bottleneck convolution blocks (MBConv) with squeeze-and-excitation (SE) attention to achieve high parameter efficiency. We froze the pre-trained backbone, added a global pooling and dropout head, and optimized using a custom, differentiable macro-F1 loss that directly maximizes our target metric while masking ignored classes. After 10 epochs (LR: 10-3), we unfroze the top 20 layers and fine-tuned (LR: 10-4) with early stopping and learning-rate reduction.

**ResNet50**, by contrast, stacks identity-mapped residual blocks that facilitate very deep feature hierarchies and preserve low-level details. We trained a dropout head with sparse-categorical-crossentropy. On the dev, ResNet achieved 43.63% error, indicating its residual structure more readily modeled the reconstructed textures - but its error ballooned to 86.06% on the final eval split, revealing sensitivity to the training distribution and a risk of overfitting when loss and evaluation metrics diverge.

**Results:** Across our classical‐ML pipelines, the baseline LightGBM classifier began with a dev‐set error of 61.65%, which fell to 55.99% after tuning it. On the original eval split, error similarly dropped from 60.59% to 56.15%, with virtually no train–test discrepancy - strong evidence that balanced‐class weighting plus LightGBM’s built-in shrinkage effectively controlled overfitting. Introducing PCA (200 components) further reduced dev error to 55.09%, and fine‐tuning that pipeline reached 54.92%. The final SelectKBest (k=150) stage yielded the best classical performance, driving dev error to 53.92% and eval error to 56.62%, underscoring how pruning uninformative DCT coefficients improves the model’s focus. ResNet incurred an 86.09% error, while our best LGBM pipeline saw an 85.06%. These results remind us that even when hyperparameters and data partitions are correctly applied, both gradient‐boosted trees and deep convolutional networks can struggle on sparse, DCT‐derived inputs, and that comparative evaluation remains critical to identifying the best AI strategy, which was briefly addressed using xAI (Figure 1).

|  |  |
| --- | --- |
|  | **Data Set** |
| **Algorithm** | **Train** | **Dev** | **Eval init** | **Eval** |
| LGBMClassifier | 0.01% | 61.65% | 60.59% | - |
| LGBMClassifier tuned | 0.00% | 55.99% | 56.15% | - |
| LGBMClassifier + PCA | 0.00% | 55.09% | 58.08% | - |
| LGBMClassifier + PCA (tuned) | 0.00% | 54.92% | 57.54% | - |
| PCA + SelectKBest + LGBMClassifier | 0.00% | 54.66% | 57.45% | - |
| PCA + SelectKBest + LGBMClassifier (tuned) | 0.00% | 53.92% | 56.62% | 85.06% |
| EfficientNet | 87.84% | 89.80% | 88.29% | - |
| ResNet | 29.89% | 43. 64% | 49.67% | 86.06% |

Table 1. Results obtained from all the experiments

**Conclusions:** Our investigation into modeling DCT-derived features in digital pathology underscores the effectiveness of tree-based methods, particularly LightGBM, in handling the inherent sparsity and heavy-tailed distributions of such data. AutoML-guided tuning produced a robust LightGBM pipeline, while deep models showed mixed outcomes. ResNet50 exhibited moderate adaptability to the reconstructed images, whereas EfficientNetB0 significantly underperformed, with a development error of approximately 89.8%. This suggests that EfficientNetB0's compact architecture may struggle to capture the fine-grained textures present in DCT-reconstructed images. Overall, performance hinges on aligning model inductive biases with the statistical nature of the transformed data.



Figure 1. Results of xAI for EfficientNet (top) and ResNet (bottom).

**References:**

1. Abramova, Victoriya, Vladimir Lukin, Sergey Abramov, Sergii Kryvenko, Piotr Lech, and Krzysztof Okarma. 2023. "A Fast and Accurate Prediction of Distortions in DCT-Based Lossy Image Compression" *Electronics* 12, no. 11: 2347. <https://doi.org/10.3390/electronics12112347>
2. Tan, Mingxing & Le, Quoc. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. 10.48550/arXiv.1905.11946.
3. [https://medium.com/@aimldl1984/explainable-ai-grad-cam-8acd04dd2dd3](https://medium.com/%40aimldl1984/explainable-ai-grad-cam-8acd04dd2dd3)
4. <https://pypi.org/project/lazypredict/>