Comparative DCT-Feature Classification of TUH Breast Images

Y. Qwareeq
Temple University, Philadelphia, Pennsylvania, USA
qwareeq@temple.edu

Introduction: This study addresses a pathological image classification problem using features derived from the Discrete Cosine Transform (DCT). The input data consists of 3-channel, 32×32 DCT coefficients obtained from patches of the TUH DPATH Breast dataset (`v4.0.0`). The initial task involved nine distinct classes, but based on project scoring rules which prioritize specific diagnostic categories, the task was reformulated as a 6-way classification problem. Data corresponding to original classes 1 (`'artf'`), 4 (`'susp'`), and 7 (`'null'`) were filtered out. The remaining relevant classes (original labels `0='norm'`, `2='nneo'`, `3='infl'`, `5='dcis'`, `6='indc'`, `8='bckg'`) were remapped to labels 0 through 5, respectively, for model training.

The primary evaluation metric for this task is a custom weighted error (WE) designed to heavily penalize misclassifications among the diagnostically relevant non-background classes. Specifically, the score is calculated as Score $= 0.9 \times$ `avg_lbls_err` $+ 0.1 \times$ `avg_bckg_err`, where `avg_lbls_err` is the mean error rate across the remapped classes 0-4, and `avg_bckg_err` is the error rate for the remapped background class 5. Minimizing this score was the central objective. This paper compares two distinct approaches: a classical machine learning algorithm, `LightGBM`, chosen for its effectiveness on tabular data; and a deep learning approach using a Vision Transformer (ViT), selected for its state-of-the-art performance on image classification tasks. We detail the feature engineering, model tuning, and final prediction generation process for both methods. All experiments utilized the `train.csv` and `dev.csv` files provided. For feature normalization, `sklearn.preprocessing.StandardScaler` was fit only on the filtered training data features and subsequently used to transform both the training and development sets. This resulted in 9340 training samples and 5410 development samples after filtering and remapping.

Algorithm No. 1 Description: The non-neural network approach utilized `LightGBM` with the `DART` booster. Feature engineering focused on low-frequency DCT coefficients. The final feature set was derived by selecting the top-left DCT blocks of size $k = (3, 4, 2)$ for each of the 3 channels, applying Principal Component Analysis (PCA) retaining $\approx$99.95% variance, and concatenating the three DC coefficients (one per channel), resulting in 32 features. Extensive hyperparameter optimization for the 6-class task was performed using `Optuna` with 3-fold cross-validation over three phases. This optimized the feature parameters (k, PCA variance), model structural hyperparameters (e.g., `num_leaves`=24), and class weights ($w^* \approx [4.1, 2.6, 36.9, 36.5, 19.8, 1.8]$). This process yielded a final cross-validated weighted error of $\approx 42.44\%$.

Algorithm No. 2 Description: Given the image-based nature of the DCT features, a deep learning approach using a ViT was explored. Preliminary attempts with MLPs (Multilayer Perceptrons), basic CNNs (Convolutional Neural Networks), smaller ViT variants (`Tiny`, `Small`), ResNets (Residual Networks) on inverse DCT (IDCT) images, and CNNs on raw coefficients were unsuccessful (failed convergence or weighted error $>\approx 50\%$). The successful approach involved fine-tuning a pre-trained `vit_base_patch16_224` model from the `timm` library. Data preparation for the ViT involved taking the filtered six-class DCT coefficients (size $3 \times 32 \times 32$), transforming them back to spatial images using the inverse DCT (`scipy.fftpack.idctn`), scaling the resulting pixel values to the 0–255 range, resizing the images to 224×224 using bicubic interpolation, and finally applying standard ImageNet mean and standard-deviation normalization.

For training, the model's classification head was replaced with a new one for six output classes. Optimization was performed using `AdamW` with `nn.CrossEntropyLoss` weighted by class frequencies in the training set. A cosine annealing learning rate schedule was employed, along with layer-wise learning rate decay (`create_param_groups_lrd` function) and `DropPath` regularization to mitigate overfitting. Early stopping based on the 90/10 weighted error metric on the development set was used over a maximum of 40 training epochs. Hyperparameter tuning via `Optuna` involved two 30-trial sweeps (coarse then fine) to optimize `learning_rate`, `weight_decay`, `layer_decay`, and `drop_path`, minimizing the cross-validated weighted error (using an internal stop patience of 8 and the `MedianPruner`). The best hyperparameters found were `learning_rate` $\approx 1.98 \times 10^{-4}$, `weight_decay` ≈ 0.0203, `layer_decay` ≈ 0.859, and `drop_path` ≈ 0.190, which achieved a best cross-validated weighted error of 27.56%.

Results: To assess the generalization performance of the final optimized models, a repeated testing procedure was conducted using 10 independent runs. In each run, the combined `train + dev` data was split into 70% for training, 15% for validation, and 15% for testing (stratified by class), with the test set held out completely. Across these 10 runs, the optimized `LightGBM` (`DART`) model achieved a mean 90/10 weighted error of $41.87\% \pm 1.56\%$ (standard deviation) on the test sets. The optimized `ViT-B/16` model achieved a significantly lower mean weighted error of $31.85\% \pm 1.45\%$ on the test sets, demonstrating superior accuracy and stability.

For the final submission, models were trained using all samples from the `train` set, with the `dev` set used for early stopping (`patience=50` boosting rounds for `LightGBM`, `patience=2` epochs for ViT). The ViT model underwent an additional fine-tuning epoch on the combined `train + dev` data. The `LightGBM` model achieved 45.06% weighted error on the development set, stopping around boosting round 352. The `ViT-B/16` model achieved 28.80% weighted error on the development set, converging by epoch 12 before the final fine-tuning step. The 90/10 weighted error achieved on the three datasets using these final models is reported in the adjacent table.

	Data Set		
Algorithm	`train`	`dev`	`eval`
DRT (`LightGBM`)	17.2135%	45.5868%	51.6819%
ViT	14.8427%	28.8026%	35.2692%

Table 1: Final 90/10 weighted error (%) on `train`, `dev`, and `eval` datasets for both algorithms.

Conclusions: This study compared classical machine learning (`LightGBM`) and deep learning (`ViT-B/16`) approaches for classifying pathological tissue types from 3-channel, 32×32 DCT coefficients, optimizing for a custom 6-class, 90/10 weighted-error metric. Extensive feature engineering involving selection of low-frequency DCT coefficients, PCA, and inclusion of DC components, combined with `Optuna`-based hyperparameter tuning (model structure, regularization, class weights) resulted in a strong `LightGBM` (`DART` booster) baseline achieving a mean weighted error of approximately 41.9% over 10 independent test sets.

The Vision Transformer approach, which involved reconstructing spatial images via inverse DCT, leveraging a pre-trained `ViT-B/16` model, and applying fine-tuning techniques such as layer-wise learning-rate decay and `Optuna` optimization, significantly outperformed the classical method, achieving a mean weighted error of approximately 31.9% across the same 10 test sets. These findings underscore the benefit of leveraging powerful pre-trained vision architectures, even when starting from frequency-domain data, by reconstructing spatial representations. While the optimized `LightGBM` provided a robust non-neural network benchmark, the fine-tuned Vision Transformer demonstrated superior accuracy and stability for this specific image classification task under the defined evaluation criteria.