DCT Classification

Sarker Mohammed
Department of Electrical and Computer Engineering, Temple University
sarker.mohammed@temple.edu

Introduction: This study addresses pathological image classification using frequency-domain features from the TUH DPATH Breast Cancer dataset (v4.0.0). The input data consists of Discrete Cosine Transform (DCT) coefficients, which provide a compact representation of the original pathology images. Specifically, the data was generated by computing 256 × 256 DCT on annotated tissue patches, from which the most significant 32 × 32 coefficients were retained for the Red, Green, and Blue channels. This results in a feature space of 3,072 dimensions (3 × 32 × 32) per sample, capturing the essential spectral signatures of the breast tissue. The original dataset includes nine distinct classes but due to how the project is scored, the task was reformulated as a 6-way classification problem. Data corresponding to labels 1 ("artf"), 4 ("susp"), and 7 ("null") were filtered out, as they represent artifacts or unclassified segments that are ignored during evaluation. The remaining six relevant classes, representing normal tissue (norm), neoplastic tissue (nneo), inflammatory tissue (infl), ductal carcinoma in situ (dcis), invasive ductal carcinoma (indc), and background (bckg), were remapped to provide a focused diagnostic output.

The primary objective of this project is to minimize a custom weighted error metric. In this scoring system, tissue classification is weighed at 90% of the total score, while the background classification is weighed at 10%. This weighting scheme places extreme importance on the model's ability to differentiate between complex tissue pathologies. This report compares two distinct approaches including a statistical pipeline using Principal Component Analysis and Support Vector Machines as well as a deep learning approach utilizing a Convolutional Neural Network.

Algorithm No. 1 Description: The non-neural network approach utilized a classical statistical pipeline consisting of a dimensionality reduction followed by a kernel-based classifier. Since the input features consist of 3,072 DCT coefficients, representing high-dimensional frequency data, Principle Component Analysis (PCA) was employed to remove redundant spectral noise.

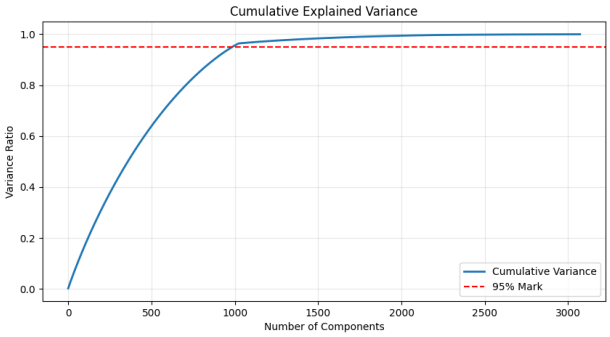

Figure 1: Cumulative Explained Variance

To determine the optimal feature space, a cumulative explained variance analysis was performed. As shown in the Scree Plot (Figure 1), the dataset experiences significant redundancy. It was found that 90% of the variance is captured by only 876 components, 95% is captured by 986 components, and 99% is captured by 1748 components. Experimental results justified the selection of the 95% variance mark. Reducing the feature space further to 300 resulted in a significant degradation of performance. We first tested various component values for PCA dimensionality reduction. In Table 1, we empirically tested the performance as we decreased dimensionality.

Components (N)	Train Error (%)	Dev Error (%)
3072	23.4708 / 22.8179	67.7371 / 63.8023
986	23.6487 / 22.8832	67.8551 / 63.7588
300	30.1532 / 28.8485	78.1173 / 72.8374

Table 2: PCA Dimensionality Reduction

For the classification portion, a Support Vector Machine (SVM) with a radial basis function (rbf) kernel was selected due to its ability to project the highly correlated features into a higher dimension space where they might become more separable.

The SVM algorithm seeks to define an optimal hyperplane that maximizes the margin between support vectors, which are the critical data points defining the boundary of each class region.

Since SVMs can be sensitive to outliers or mislabeled data, the prior dimensionality reduction via PCA served as a critical regularization step, preventing the model from overfitting to high-frequency noise. Extensive hyperparameter optimization was conducted via grid search across regularization parameters (C) and kernel coefficients (γ). The configuration of $C = 1$ and $\gamma = 'scale'$ was found to provide best dev accuracy, but configuration of $C = 0.5$ and $\gamma = 'scale'$ was selected to provide a balance between training accuracy and generalization.

Algorithm No. 2 Description: The second approach transitioned from a flat-vector statistical model to a Convolutional Neural Network (CNN). This shift was motivated by the spatial nature of 32 x 32 DCT coefficients grid. Online, there are various attempts at using DCT coefficients and a vast number of them utilize CNN of some sort, whether it be a custom CNN or ResNet-50. While the previous SVM approach treated the features as independent variables, the CNN leverages kernels to resolve spatial iterations.

One of the biggest challenges with this dataset is how similar each class appears to each other in the frequency domain. Neural networks are very good at memorizing training data rather than actually learning the difference between classes. In an attempt to prevent this from happening, we designed the feature extraction portion of the model to be very short. We opted for 2 small convolutional layers. By limiting this, we hoped for it to be forced to focus on the most important features.

To further help the model generalize, several "noise-fighting" strategies were added to the training process. A data augmentation step was used where the images were randomly flipped and small amounts of random noise were added to the values. This makes the training harder for the model and ensures that it doesn't rely on exact, specific numbers that might change in a new dataset. Additionally, a high Dropout rate (0.6) was used in the fully connected layer to randomly turn off parts of the network during training. Finally, the model used a weighted loss function. This essentially told the model to care more about the harder to find classes. By giving these difficult classes more weight, the model was forced to try harder to find differences between them.

Results: The performance of both algorithms were evaluated using the custom-weighed error metric provided by the client. Algorithm No. 1 showed stable error rates during development but reached an overall weighted error of 62.64% on the final evaluation set. Algorithm No. 2 performed slightly better, yielding a weighted error of 58.28%. Both models ultimately struggled to generalize and overfitted, as shown by the low Train/Dev results. While several strategies were used to prevent this, such as adding noise and using dropout, it is likely that the model size was reduced too much in an attempt to stop overfitting. This likely restricted the network from learning the complex features needed to generalize to the evaluation data.

	Data Set		
Algorithm	**Train**	**Dev Test**	**Eval**
PCA + SVM	23.23%	26.00%	62.65%
CNN	11.12%	11.17%	58.28%

Table 2: Final Results for each Algorithm

Conclusions: The classification of breast pathology from DCT coefficients remains a difficult problem due to the challenge of distinguishing between highly similar tissue signatures in the frequency domain. While both models struggled to maintain their performance on unseen data, Algorithm No. 2 achieved lower error rates by better capturing frequency relationships. The high evaluation errors suggest the models either overfitted or were too simplified to learn the features necessary for generalization. These results show that classifying these pathology samples remains a difficult challenge that requires a better balance between model capacity and the complexity of frequency-domain data.