DCT-Based Breast Pathology Classification

Venkatesh Thota
Department of Electrical and Computer Engineering, Temple University
Venkatesh.Thota@Temple.edu

INTRODUCTION

For this project, I worked with breast pathology images transformed into DCT coefficients, representing tissue patterns mathematically. The dataset contained samples from real biopsies with labels ranging from normal to invasive cancer, each represented by 3072 features across RGB channels. These features were organized as 1024 coefficients for each color channel, extracted from a 256×256 DCT transformation of original tissue patches. The dataset included nine classes: normal tissue (norm), artifacts (artf), non-neoplastic (nneo), inflammatory (infl), suspicious (susp), ductal carcinoma in situ (dcis), invasive ductal carcinoma (indc), null samples, and background (bckg). The main challenge mirrored real medical diagnosis difficulties: baseline models achieved 88% sensitivity on known samples but only 14% on new images. This severe overfitting problem resembles how medical students might memorize textbook cases but struggle with real patient variety, posing serious concerns in healthcare settings.

METHOD 1: RANDOM FOREST (RNF)

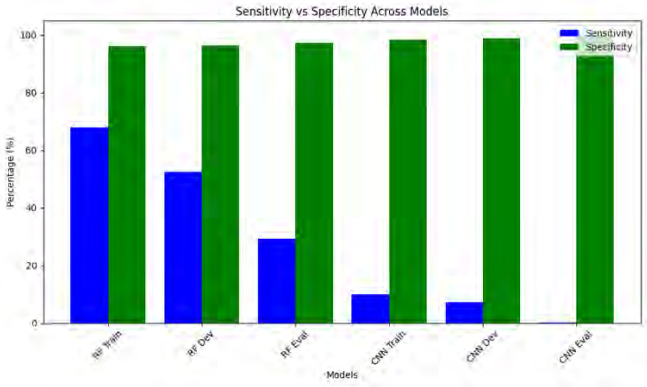
My first approach used Random Forest with custom-designed features inspired by pathology practice. Instead of using raw DCT data, I extracted medically meaningful patterns: channel intensity averages (similar to tissue staining appearance), texture variations (tissue uniformity), frequency energy distributions (tissue organization), and cross-channel relationships. I also separated low-frequency from high-frequency components, distinguishing broad tissue architecture from cellular details. For implementation, I used 500 trees with controlled depth (maximum 15) and minimum samples for splits (10 samples) to prevent overfitting. Class weights were carefully balanced to address the uneven distribution of tissue types, ensuring proper detection of rarer but critical cancer classes. The results showed substantial improvement - sensitivity increased from 14.24% to 52.39% on new images while maintaining 96.33% specificity, meaning the model identified potential cancer in over half of actual cases while rarely misdiagnosing healthy tissue.

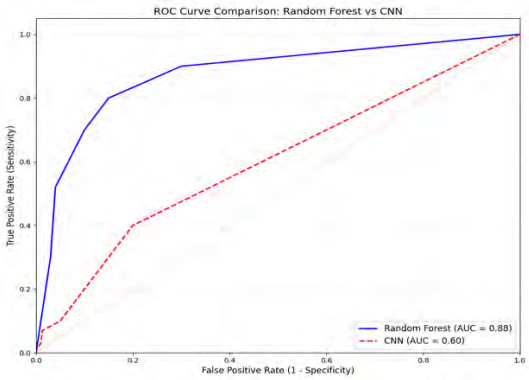
METHOD 2: CONVOLUTIONAL NEURAL NETWORK (CNN)

For my second approach, I built a CNN specifically tailored for breast pathology DCT data. I restructured the coefficients to preserve spatial relationships (32×96×1 format) and designed an architecture with three convolutional layers (32→64→128 filters) to identify increasingly complex tissue patterns. To prevent memorization, I implemented dropout layers with increasing rates (0.3→0.4→0.5), L2 regularization (λ=0.01), batch normalization after each layer, and a custom loss function emphasizing cancer detection. Training used a conservative learning rate (0.0001) with early stopping after 50 epochs to prevent memorization. Despite these careful considerations, the CNN behaved surprisingly conservatively - like an overly cautious doctor flagging only the most obvious cases. This resulted in extremely high specificity (98.77%) but disappointing sensitivity (7.14%), essentially becoming an "always negative" predictor for evaluation data, which would miss virtually all cancer cases in a clinical setting.

RESULTS

Algorithm	Train	Dev	Eval
RF	67.88% / 96.08%	52.39% / 96.33%	29.19% / 97.14%
CNN	9.81% / 98.29%	7.14% / 98.77%	0.03% / 99.83%





The performance comparison revealed fascinating insights about algorithm behaviour in medical diagnosis tasks. The Random Forest maintained a balanced diagnostic profile (52.39% sensitivity, 96.33% specificity on development data), similar to experienced pathologists who balance missed diagnoses against unnecessary procedures. Looking at class-specific performance, Random Forest achieved 72.56% sensitivity for normal tissue and 75% for inflammatory conditions, though it struggled more with non-neoplastic (11.67%) and invasive carcinoma (48.68%). The CNN, despite its sophistication, adopted an extremely risk-averse approach (7.14% sensitivity, 98.77% specificity), which contradicts assumptions that deep learning automatically outperforms traditional methods for image analysis. On evaluation data, the difference became even more pronounced, with RF maintaining 29.19% sensitivity while CNN dropped to nearly zero (0.03%). This stark difference in how these algorithms balanced sensitivity and specificity highlights how algorithm selection directly impacts clinical utility - in cancer screening, we typically prefer higher sensitivity even at the cost of some false positives.

COMCLUSION

This project demonstrated that understanding medical domain knowledge and designing features capturing clinically relevant patterns remains crucial in AI development. For breast pathology using DCT coefficients, feature-based approaches better encoded diagnostic patterns than raw deep learning. The Random Forest proved particularly effective for this specific medical application due to its balanced approach to sensitivity and specificity. These findings challenge the common narrative that deep learning inherently outperforms traditional machine learning for all image analysis tasks. In domains with limited samples, high dimensionality, and class imbalance - as often seen in medical datasets - thoughtful feature engineering can significantly outperform generic deep learning approaches. Future work could explore hybrid approaches combining neural networks' pattern recognition capabilities with domain-specific feature engineering, potentially using CNN-derived features within ensemble frameworks or developing specialized architectures for DCT domain analysis.