Deep Feature Extraction for Class-Separable t-SNE Visualization of MedMNIST Datasets: Comparing Generic and Biomedical Pretraining

Sarra Kharbech
Ph.D Student, Computer Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

Dhuha Abouemter
Hamad Bin Khalifa University, Doha, Qatar

Sara Hamed
Hamad Bin Khalifa University, Doha, Qatar

Abdul-Rahman Abdel-Fattah
Hamad Bin Khalifa University, Doha, Qatar

Adnan Khan
Hamad Bin Khalifa University, Doha, Qatar

Samir Belhaouari
Hamad Bin Khalifa University, Doha, Qatar

Abstract:
t-Distributed Stochastic Neighbor Embedding (t-SNE) is widely used to visualize high-dimensional data in two dimensions, but its output quality is fundamentally limited by the feature space on which it operates. When applied directly to raw pixel intensities of medical images, t-SNE tends to produce poorly separated clusters because Euclidean distances are dominated by superficial intensity and positional variations rather than clinically meaningful morphological differences. In this work, we examine whether replacing raw pixels with deep representations extracted from pretrained neural networks yields more class-separable t-SNE embeddings on four MedMNIST datasets: BloodMNIST, PathMNIST, OrganAMNIST, and OrganCMNIST. We compare three feature spaces: (i) raw pixels with PCA preprocessing, (ii) penultimate-layer activations of ResNet-50 pretrained on ImageNet, and (iii) embeddings from BiomedCLIP, a vision-language model pretrained on 15 million biomedical image-text pairs from PubMed Central. Both deep backbones substantially improve class separability over raw pixels across all four datasets, with the largest gains observed on high-contrast modalities such as CT (OrganAMNIST, OrganCMNIST) and microscopy (BloodMNIST, PathMNIST). BiomedCLIP outperforms ResNet-50 on some datasets and is comparable on others, suggesting that domain-specific pretraining provides a measurable but dataset-dependent advantage for unsupervised visual exploration of medical image collections.

Keywords:
t-SNE, dimensionality reduction, medical image analysis, MedMNIST, ResNet-50, BiomedCLIP, transfer learning, feature visualization.