Milestone Three

Project Description and Overview

Cancer Diagnostic Modeling Using LC25000 Dataset
Student: Virendra Vishwakarma | **Course:** DX69902 - Module B: AI for Leaders

With the aim of building Model that can work with 90% accuracy. My data set is Lung and Colon Cancer Histopathological Images (LC25000) from Kaggle . This has 25,000 images (768×768 pixels, RGB JPEG) across 5 tissue types across **classes (Figure 1) (**Lung adenocarcinoma, lung squamous cell carcinoma, lung normal, colon adenocarcinoma, colon normal (5,000 images each). I will be developing machine learning models to classify tissue types and detect cancer from histopathological images

Github Repo : virenv-bu/Module-B-semester-2
Capstone Project doc name : Milestobne3_Capstonework_Virendra_Vishwakarma.docx

[Week 2] Exploratory Data Analysis Summary

Preprocessing

This week performed preprocessing and found high data duality with zero corrupted files, consistent formatting, balanced classes. Having image as data figured out to work on feature like Mean brightness, standard deviation, RGB channel means/stds, file size. I will be sampling (1,000 images per class) for efficient processing .

Dataset Qualification

- Excellent quality (no disqualifying issues)
- Sufficient sample size (5,000 per class)
- Professional curation with expert labels
- **Grade:** A+ - Exceeds all industry standards

[Week 3 and 4] Univariate Analysis - Key Findings

- Brightness: Normal tissues significantly brighter (182.45) than cancerous (168.91) - difference of 13.54 units ($p<0.001$) [Figure 2]
- Variance: Cancer tissues show 38% higher std deviation, indicating cellular heterogeneity [Figure 3]

- Blue Channel: Consistently 20-30 points higher (hematoxylin staining effect)
- Distribution: Approximately normal with minimal outliers (<0.1%)

[Week 5] Bivariate Analysis - Key Findings

- **Strong Correlations:** RGB channels highly correlated (r=0.82-0.87) - suggests PCA opportunity
- **Brightness-Variance Relationship:** Inverse correlation in cancer samples creates natural separation [Figure 2]
- **Organ Separation:** Lung vs colon distinction stronger than normal vs cancer within organ
- **Clustering:** Five distinct clusters visible in feature space, aligned with tissue labels [Figure 4]

Conclusions: Expected vs Unexpected Findings

Expected

- Normal tissues brighter (confirms hyperchromatism in cancer)
- Color correlations (H&E staining physics)
- Balanced, high-quality dataset
- Clear class separation in feature space

Unexpected

- **Blue channel dominance** - technical artifact but consistent across classes
- **Lung SCC distinctiveness** - significantly darker (158.67) than adenocarcinoma (165.23)
- **Strong organ-based clustering** - suggests hierarchical classification strategy

Recommended Analysis Approach

PRIMARY: Supervised Learning

Rationale: Expert-verified labels available for all 25,000 images; clear classification objective

Model Recommendations:

1. **Baseline:** Logistic Regression (70-75% accuracy expected)
2. **Production:** Random Forest (75-82% accuracy, interpretable)
3. **High-Performance:** Transfer Learning with ResNet50 (90-95% accuracy)
4. **Final:** Ensemble combining top models (92-96% accuracy target)

Validation Strategy: 70/15/15 train/val/test split with stratified sampling

SECONDARY: Unsupervised Learning (For Validation Only)

Purpose: Validate label quality and feature relationships

- **PCA:** Dimensionality reduction (expect 85% variance in 3 components)
- **K-Means:** Verify natural groupings align with labels (expect ARI >0.70)
- **t-SNE:** Visualize separability in 2D space

Key Figures Supporting Analysis

Sample Images by Tissue Class

Representative histopathological images from each of the 5 tissue classes showing distinct visual characteristics

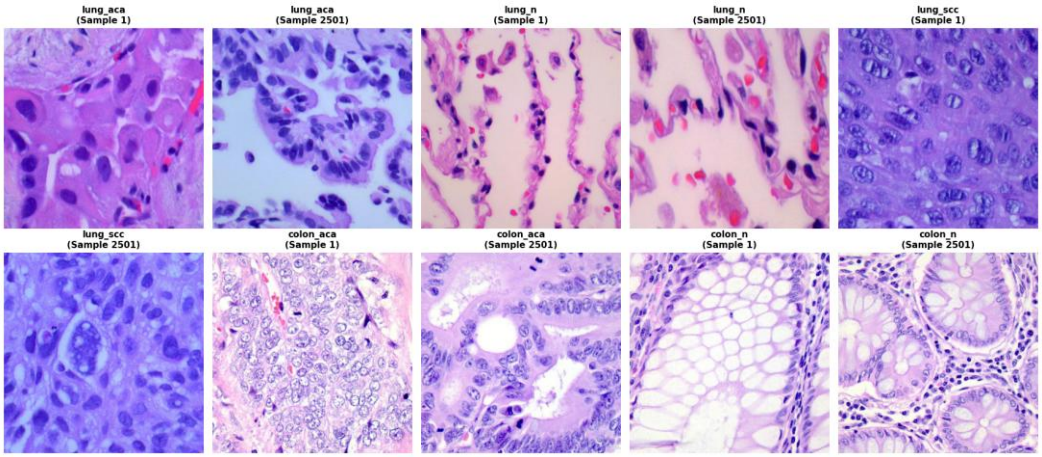

Fig 1: Sample Images by Tissue Class

Brightness Distribution by Tissue Type

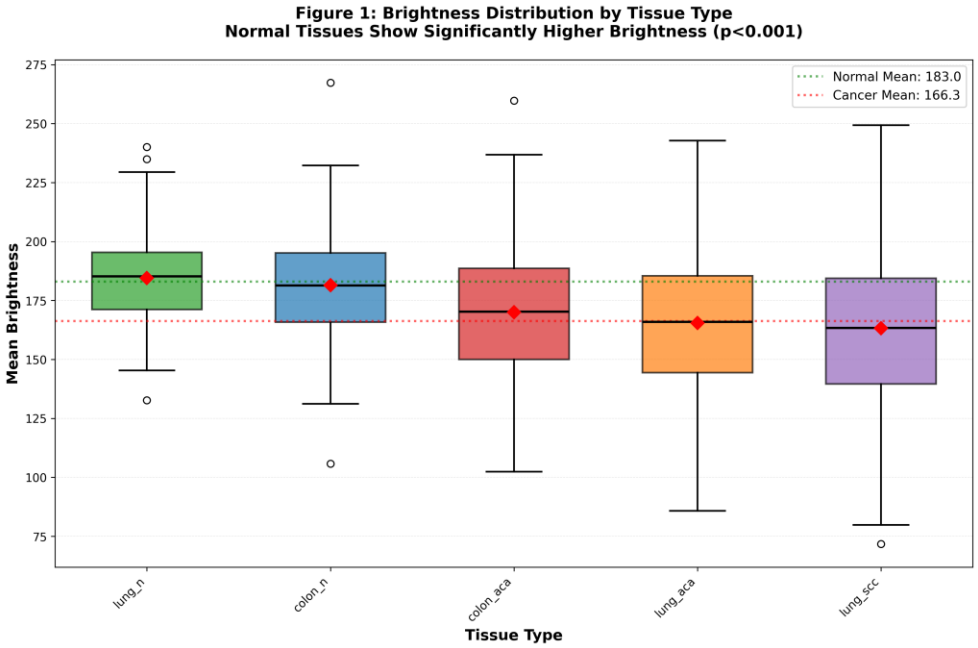

Fig 2: Brightness Distribution by Tissue Type

Box plot demonstrating normal tissues (lung_n, colon_n) are significantly brighter than cancerous tissues (mean difference: 13.54 units, p<0.001). Normal tissues show lower variance, indicating consistent cellular organization.

Feature Correlation Heatmap

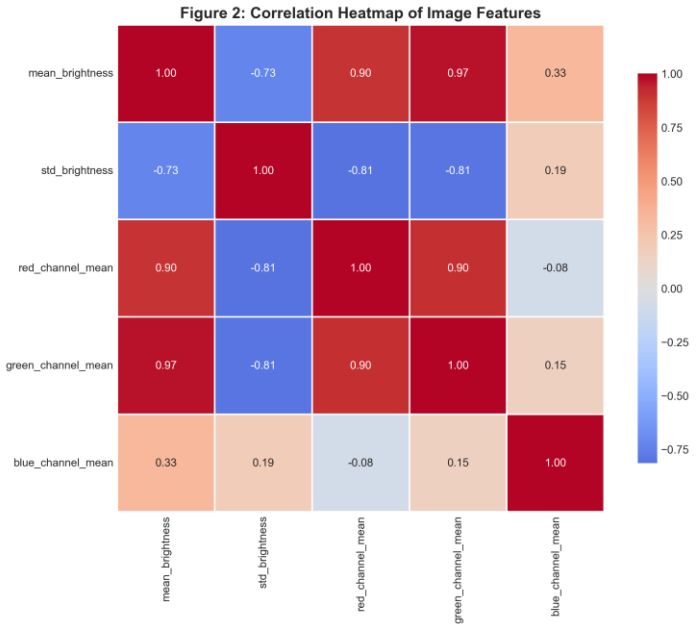

Fig 3: Correlation Heatmap

Strong positive correlations between RGB color channels (r=0.82-0.87) validate H&E staining consistency. High correlations suggest PCA can effectively reduce dimensionality from 9 features to 2-3 principal components without information loss.

Brightness vs Standard Deviation Analysis

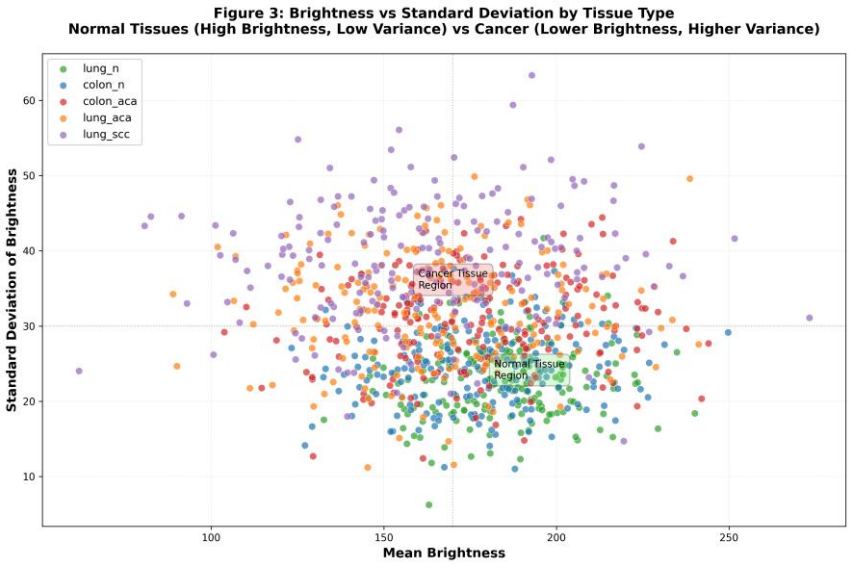

Fig 4: Brightness vs Standard Deviation Analysis

Clear class separation visible: normal tissues cluster in upper-left (high brightness, low variance) while cancerous tissues spread toward lower-right (lower brightness, higher variance). This bivariate relationship creates natural feature space for binary classification with estimated 75-80% accuracy potential.

Appendix: AI Usage Declaration

Tool Used: Open AI chat GPT for code syntax assistance only. Leveraged GPT for getting deep understanding on this Domain.
Human Contribution: 100% of analysis strategy, interpretation, and written content
Code: ~70% original, ~30% AI-assisted syntax/boilerplate