Preprocessing of Skin Cancer Whole Slide Images to Predict Five-Year Survival

Ewan Ward
Emerging Diagnostic and Investigative Technologies, Department of Pathology, Dartmouth Hitchcock Medical Center

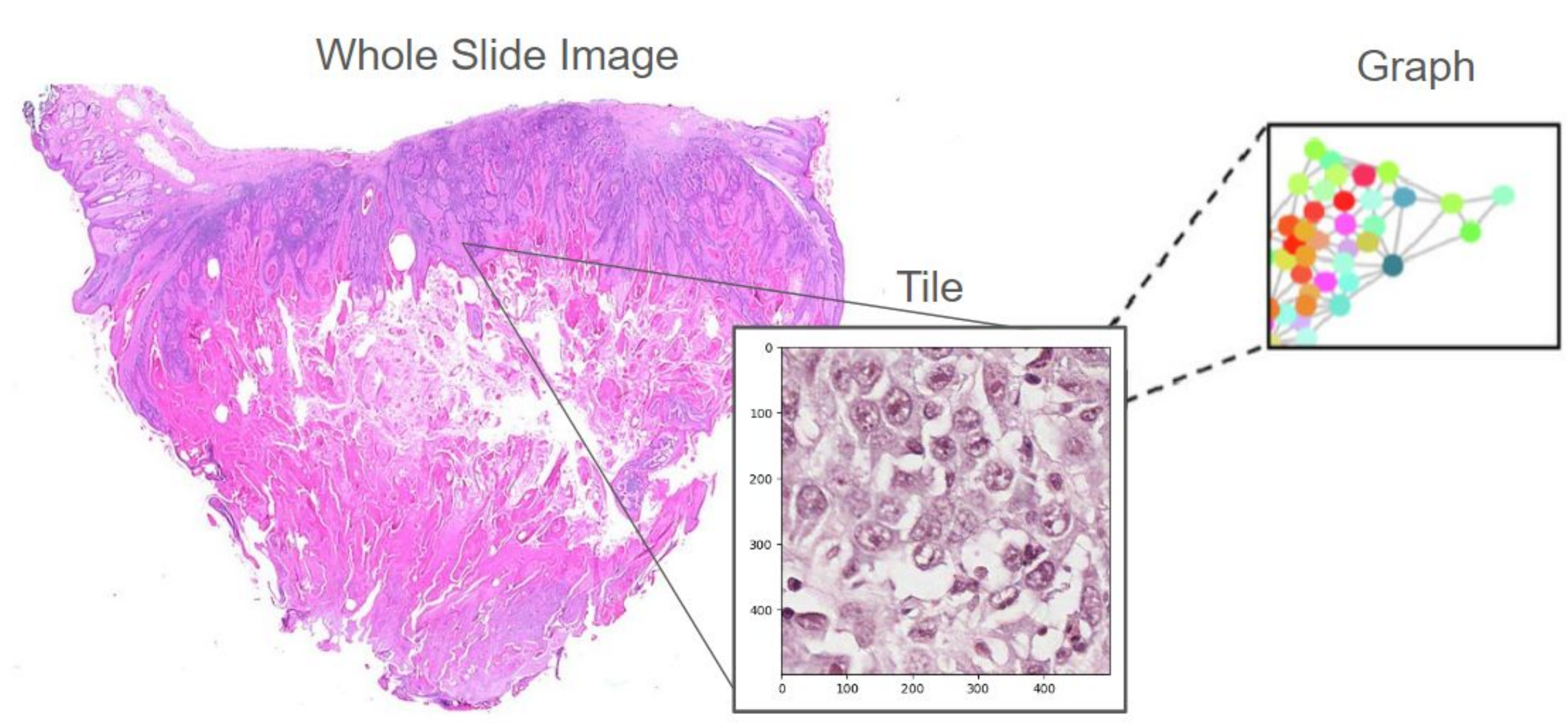
Dartmouth Health

ABSTRACT

- **Data Acquisition and Labeling:** WSIs and associated patient data were obtained from The Cancer Genome Atlas (TCGA). The data was used to assign binary survival labels ("survived" or "not survived") to each slide.
- **Tile Generation and Filtering:** Each WSI was divided into smaller tiles. Tiles containing mostly white space were discarded.
- **Tile Labeling and Splitting:** The remaining tiles were labeled with their corresponding survival status and randomly assigned to training, validation, and testing sets.
- **Feature Extraction:** Key features, including Haralick features (contrast and energy) and RGB color information, were extracted from each tile.
- **Graph Construction:** Graphs were built where each tile represents a node, and edges connect spatially adjacent tiles.
- **Next Step:** The project is ready for the subsequent phase of building and training a GCN using the generated graphs and features.

INTRODUCTION

- **Skin cancer** is a significant public health issue with millions of cases and thousands of deaths annually in the US.
- **Early detection of melanoma**, the deadliest form of skin cancer, is crucial due to its high survival rate.
- **Advancements in digital pathology** allow for rapid and consistent analysis of skin cancer tissue samples.
- **Graph Convolutional Networks (GCNs)** are a promising machine learning technique for analyzing whole slide images (WSIs) of skin cancer tissue.
- **Preprocessing WSIs involves** converting images into graphs through tiling, feature extraction, and graph construction.



Cancer Slide Image Credit: Brollo, A. https://commons.wikimedia.org/wiki/File:Skin_keratoacanthoma_whole_slide.jpg
Tile Image Credit: Ward, E. (2024). Preprocessing of Skin Cancer Whole Slide Images to Predict Five-Year Survival
Graph Image Credit: Wenqi L., et al. (2022) SlideGraph+: Whole slide image level graphs to predict HER2 status in breast cancer, Medical Image Analysis, Volume 80

Figure 1: Example workflow for converting Whole Slide Images to Graphs

METHODS

- **Goal**: Develop a preprocessing method for skin cancer whole slide images (WSIs) to prepare data for a graph convolutional network (GCN) that predicts 5-year survival.
- **Data Acquisition**: Obtain WSIs and survival data from TCGA.
- **Data Preparation**: Divide WSIs into smaller tiles, filter out low-quality tiles, and assign survival labels.
- **Feature Extraction**: Extract Haralick features (contrast, energy) and color information from each tile.
- **Graph Construction**: Create a graph where tiles are nodes and their spatial relationships are edges.

RESULTS

- **WSI Tiling:** Large WSIs are divided into smaller tiles for efficient processing.
- **Data Selection:** 40 WSIs were chosen from TCGA and processed into approximately 24,000 tiles each.
- **Tile Filtering:** Tiles consisting primarily of white space were removed based on pixel color thresholding.
- **Data Labeling:** Each tile was labeled as "survived" or "not survived" based on patient outcome.
- **Data Splitting:** Tiles were randomly divided into training, validation, and testing sets for model development.
- **Feature Extraction:** Image features were extracted for analysis.
- **Feature Types:** Haralick features (contrast and energy) and color histograms were used.
- **Haralick Features:** Measure image texture based on pixel intensity differences.
- **Color Histogram:** Measures the distribution of colors within an image.
- **Feature Combination:** Haralick features and color histograms were combined to represent each tile.
- **Graph Representation:** Tiles are represented as nodes in a graph.
- **Node Relationships:** Edges connect neighboring tiles in the graph.
- **GCN Analysis:** A GCN analyzes the graph to identify patterns and relationships.
- **Prediction Goal:** The GCN aims to predict 5-year survival based on the graph representation.
- **Graph Creation:** The graph is constructed using extracted tile features and spatial information.

CONCLUSION

- **Initial Goal:** Build a GCN to predict skin cancer recurrence using whole slide images.
- **Project Adjustment:** Due to data accessibility, shifted focus to predicting 5-year survival.
- **Challenges Faced:** Limited coding experience, data acquisition difficulties, and preprocessing hurdles.
- **Project Outcome:** Successfully preprocessed 40 whole slide images for future GCN development.
- **Skill Development:** Learned Python, Jupyter Notebook, and high-performance computing.
- **Code Acquisition:** Gained experience in code searching, adaptation, and troubleshooting.

edit.