Capstone Project

Software Technology 1 (4483)

Year 1 Semester 2

**GitHub Repo**

Dylan Colwill

26/10/2023

Table of Contents

[Introduction 2](#_Toc149216111)

[Dataset Description 2](#_Toc149216112)

## Introduction

In this report, I will cover the design and structure of my Capstone Project for Software Technology 1. The project is to create software to detect different types of lung and colon cancer. This project works of the basis of a dataset filled with histopathological images of these two different cancer types.

The histopathological images provide valuable insights into cellular structures and abnormalities. The aim is to develop accurate predictive models by harnessing machine learning techniques, that would improve both the efficiency and accuracy of cancer diagnosis.

The project journey involved extensive data exploration, analysis, and modelling in order to achieve meaningful results. Through experimentation learning techniques such as Convolutional Neural Networks, promising outcomes are shown in classifying different types of cancerous tissues based on visual patterns identified within these images.

## Dataset Description

This dataset is sourced online from Kaggle with a free license <https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images>, with credit to (Borkowski AA, Bui MM, Thomas LB, Wilson CP, DeLand LA, Mastorides SM. Lung and Colon Cancer Histopathological Image Dataset (LC25000). arXiv:1912.12142v1 [eess.IV], 2019), can also be found originally on GitHub <https://github.com/tampapath/lung_colon_image_set>. The colon and lung cancer dataset consists of 25000 total images, comprised of five classes with 5000 images each. The dataset was created from 250 sample images from each class for a total of 1250, which was then expanded to 25000 using image augmentation methods, specifically the *Augmentor* Python package <https://github.com/mdbloice/Augmentor>, which is a technique to artificially generate images for the purpose of training a deep learning or machine learning model.123456

***Number of Classes:*** 5

***Classes:***

Colon Cancer:

* Colon Adenocarcinoma
* Colon Benign Tissue

Lung Cancer:

* Lung Adenocarcinoma
* Lung Benign Tissue
* Lung Squamous Cell Carcinoma

***Number of Images:*** 25000 total, 5000 per class

***File Type:*** jpeg

***Image Dimensions:*** 768 x 768 pixels

## Methodology

There are three major steps outlined to complete this project. Firstly, to perform Exploratory Data Analysis (EDA) on the dataset. Then creation of a prediction algorithm, Predictive Data Analytics (PDA) model. Finally to implement it into an application for ease of external users.

## Exploratory Data Analysis

EDA stage is crucial to the development of a strong and accurate predictive model. It helps to get a deep understanding of the data and its patterns and outliers, as well as identify any errors.