**Capstone Project**

**Document Skeleton**

# Process overview

The following diagram shows the overall end-to-end process for defining, designing and delivering the Capstone project.



Note: The following are the candidate sections of the document. They are presented here for guidance. Questions in each section could be used as possible aspects to cover. Some questions may not be applied to each project. On the other hand, additional information may be needed.

# Problem statement

* What is the problem or the opportunity that the project is investigating?
  + Pathologists often face a high volume of histopathological slide, which leads to stress, fatigues, and potential diagnostic errors.
* Why is this problem valuable to address?
  + The issue is critical in cancer detection, extended waiting periods have become a significant challenge, potentially delaying timely treatment for patients
* What is the current state (e.g. unsatisfied customers, lost revenue)?

Current state: busy laboratory, delayed diagnosis.

* What is the desired state?

Automating image classification can reduce stress, save time and improve accuracy

* Has this problem been addressed by other research projects? What were the outcomes?

Past research has used deep learning for medical images, but histopathology needs models that capture dine texture and details

# Industry/ domain

* What is the industry/ domain?

Medical laboratories / pathology

* What is the current state of this industry? (e.g. challenges from startups)

Current state: all the cases are still manually examined by pathologists, which is time-consuming and labour-intensive

Shortage of trained pathologists, high risk of burnout and human error

* What is the overall industry value-chain?

Sample collection -> sample fixation-> slide preparation -> manual diagnosis -> report generation

* What are the key concepts in the industry?
  + Histopathological imaging
  + Cancer subtype classification
  + Diagnostic automation
* Is the project relevant to other industries?
  + Radiology
  + Telemedicine and remote diagnosis

# Stakeholders

* Who are the stakeholders? (be as specific as possible)
  + Histology laboratory manager and pathologists
* Why do they care about this problem?
  + Pathologists want to reduce workload and avoid errors
  + Hospitals aim to improve efficiency and diagnostic speed
  + Researchers seek real-world applications for AI in medicine
* What are the stakeholders’ expectations?
  + A reliable AI tool that supports pathologists
  + Improve accuracy in Cancer diagnosis
  + Fater turnaround times for pathology reports

# Business question

* What is the main business question that needs to be answered?
  + Can an AI-based image classification system reduce diagnostic workload and improve accuracy in detecting lung and colon cancers in histology labs?
* What is the business value of answering this question? (quantify value and make necessary assumptions)
  + Reducing diagnostic workload can save time and lower operational costs in pathology labs
  + Faster turnaround time may lead to earlier treatment, improving patient outcomes and hospital efficiency.
  + Reducing pathologist burnout can improve job satisfaction and retention
  + Even a 10%-15% improvement in diagnostic speed could results in significant savings.
* What is the required accuracy? What are the implications of false positives or false negatives?
  + Required accuracy should aim for at least 99%accuracy. False negatives are more critical, recall must be 100% otherwise missed diagnosis can lead to serious consequences
  + False positives may lead to unnecessary stress, additional tests and healthcare costs.

# Data question

* What is the data question that needs to be answered?

Can a deep learning model accurately classify histopathological images to detect Lung and Colon cancer?

* What is the data required to answer the question?

A labelled balanced dataset of histopathological images representing different tissue classes for training and evaluating a classification model

# Data

* Where was the data sourced?

From Kaggle, with images originally collected from James, A Haley Veteran’s Hospital in Florida

* What is the volume and attributes of the data?
  + 25000 images in total, divided into 5 classes. 5000 each classes.
  + All images are 768 x 768 pixels in JPEG format
* How reliable is the data?

Reliable because its from actual medical laboratory and standard H&E stain

* What is the quality of the raw data?

High quality but most were augmented from smaller set of original images

* How was this data generated?

Original images were collected from pathology slides, then segmented to increase volume

* Is this data available on an ongoing basis?

The dataset us static and not updated, not available on an ongoing basis

# Data science process

## Data analysis

* What data pipeline was to wrangle the raw data?

Images resized to 180 X 180 pixels, normalized to [0,1], split into train(70%), validation(15%) and test (15%) processed in batches of 16

* What are the highlights of the Exploratory Data Analysis (EDA)?

Check class balance

Confirmed clear differences between benign and malignant samples

* Is the pipeline reusable? (for example, to process future data?)

Yes

* What are the intermediary data structures used (if any)?

Used Numpy arrays and TensorFLow datasets for batching and processing

## Modelling

* What are the main features used?

Image pixel values

* Did you find any interesting interactions between features?

Not applicable.. but increasing pixel values can also increase accuracy in pre-trained models

* Is there a subset of features that would get a significant portion of your final performance? Which features?

Not manually selected, models use full image input

* How did you select features?

Not manually selected

* What feature engineering techniques are used?

Rescaling and data augmentation

* What are the models used?

Convolutional Neural Networks and pre-trained models

* How long does it take to train your model?

Approximately 30-60 mins for one model on GPU for customised CNN

* What are the tools used? (cloud platform, for example)

Google colabs, TensorFlow/Keras, Numpy, sklearn,

* What are the model performance metrics?

Accuracy, precision, Recall, F1 score and confusion matrix

* Which model was selected?

Customised CNN model

## Outcomes

* What are the main findings and conclusions of the data science process?
  + The custom CNN models successfully classififed histopathological images with high accuracy
  + Preprocessing steps like resizeing and normalization improved model performances
  + Model is effective at filtering benign from malignant cases. Making it useful as a diagnostic support tool
  + Pre-trained model tested, one of them can reach up to 100% accuracy for certain types of cancer and benign conditions

## Implementation

* What are the considerations for implementing the model in production?
  + Ensure model is integrated into a secure, user-friendly interface
  + Model must be explainable to gain trust from clinicians
  + Regular train with new data is needed to keep the model updated and unbiased
  + Consider performances across different devices
  + Address privacy and compliance concerns with medical image data

# Data answer

* Was the data question answered satisfactorily?

Yes, the dataset provided enough labelled, high quality histopathlogical images to train and evaluate the model effectively

* What is the confidence level in the data answer?

High

# Business answer

* Was the business question answered satisfactorily?

Yes – the project showed that AI can support pathologists by reducing workload and improving diagnostic efficiency

* What is the confidence level in the business answer?

High- model accuracy and clear application in clinical workflows support the business values

# Response to stakeholders

* What are the overall messages and recommendations to the stakeholders?

The model can help reduce diagnostic workload and improve consistency

It’s not a replacement but a supportive tool for pathologists

With further training and regular updates, it could be deployed in clinical settings

Clear visual explanations and performances metrics can improve trust in system

# End-to-end solution

* What is the overall end-to-end solution to use the model developed in the project?

The solution starts with images preprocessing, customised CNN model is trained on the dataset then deployed using streamlit app, users able to upload images and receive predictions.

This end-to-end system supports real-time, interpretable decision -making for image classification

# References

* Where are the data and code used in the project? (show a simplified list of main items: notebooks, datasets, exported models)

Notebooks: <https://colab.research.google.com/drive/1bCnVJUeWPpwMKL_En2zwjXo_k5aaRHo0>

* + Dataset: LC25000 : <https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images>
  + Exported models: pretrained models: ResNet50 and DenseNet121 , custom CNN
* What are the resources used in the project? (libraries, algorithms, etc)
  + Libraries: TensorFlow, Keras, Numpy, Pandas, Matplotlib, OpenCV
  + Tools: google Colab, StreamLit
  + Algorithms: CNN, Transfer learning with pre-trained architectures