CLASSIFYING CLINICALLY ACTIONABLE GENETIC MUTATIONS

CHEONG YU CHYE – DSI 13 – CAPSTONE PROJECT

OUR PROBLEM STATEMENT

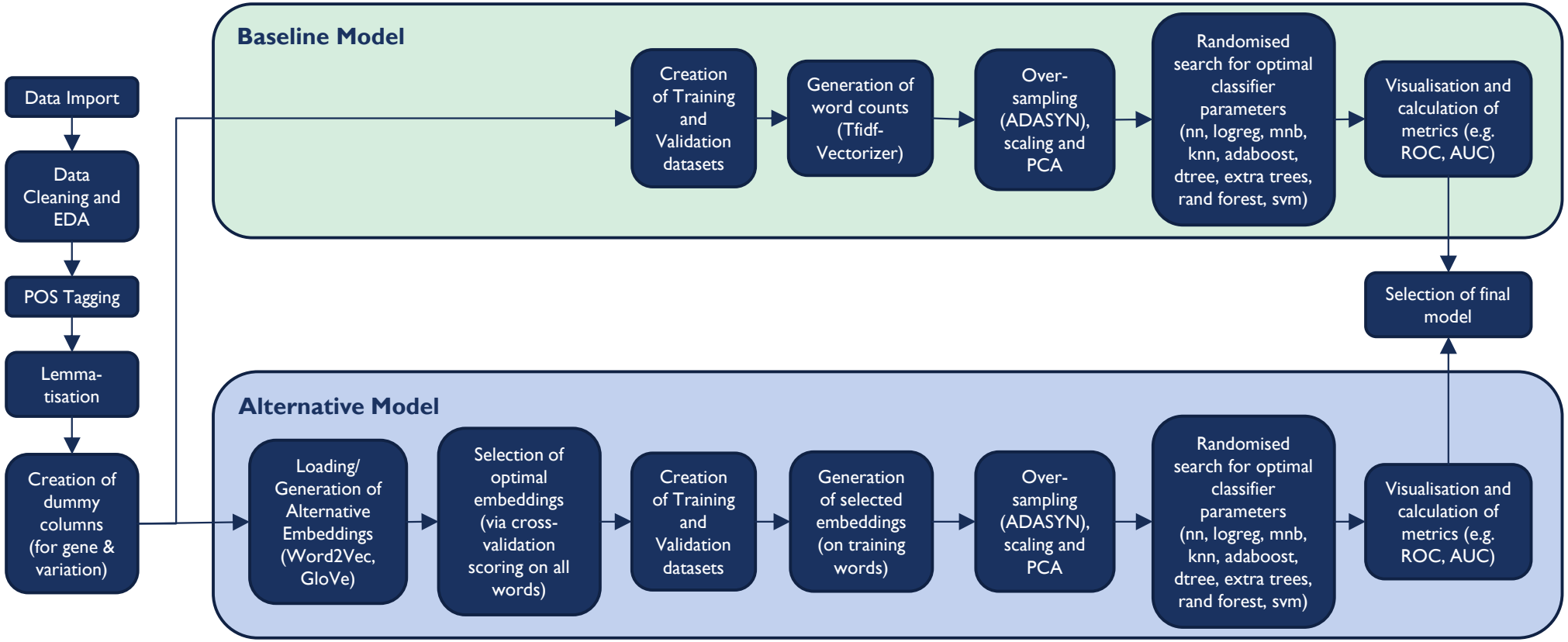
- Cancer is a top public health issue[1]:
  - Globally, about 1 in 6 deaths is due to cancer
  - One in every 4-5 people in Singapore may develop cancer in their lifetime
- Today, clinical pathologists can perform genomic sequencing on a patient's tumour sample to determine if it carries mutations that could aid in treatment, or clinical trials[2]
- Once these genes and mutations have been identified, clinical pathologists then have to manually review a growing corpus of related biomedical literature to classify the mutations – this process is tedious and time consuming
- Our problem statement: to build a classifier that can help to automate this classification
- Metrics: balanced (weighted) accuracy and F1 scores, micro-average AUC
- Success measure: beat baseline accuracy (0.287) by ≥ 10%

Sources:
[1] - https://www.nccs.com.sg/patient-care/cancer-types/cancer-statistics, https://www.who.int/news-room/fact-sheets/detail/cancers
[2] - https://www.mskcc.org/cancer-care/diagnosis-treatment/diagnosing/role-pathology

CHALLENGES

- A multi-class scenario with imbalanced classes
  - Our problem statement is a multi-class scenario involving 9 classes
  - Just two of the most frequent classes account for ~50% of all the classes
- Size of training and testing datasets
  - Only ~3,300 rows in the training dataset, but each row has a mean of ~63k words, and a maximum of ~526k words
  - After using one-hot encoding column creation and term frequency creation (TfidfVectorizer), we have >76k features
  - As-is, downstream model fitting is extremely slow (fitting for all models takes >17 hrs)
- Clinical text is difficult to classify effectively
  - Understanding the context of words is key but it is not easy to find related word embeddings
  - BioBERT word embeddings look promising but there are constraints to using them

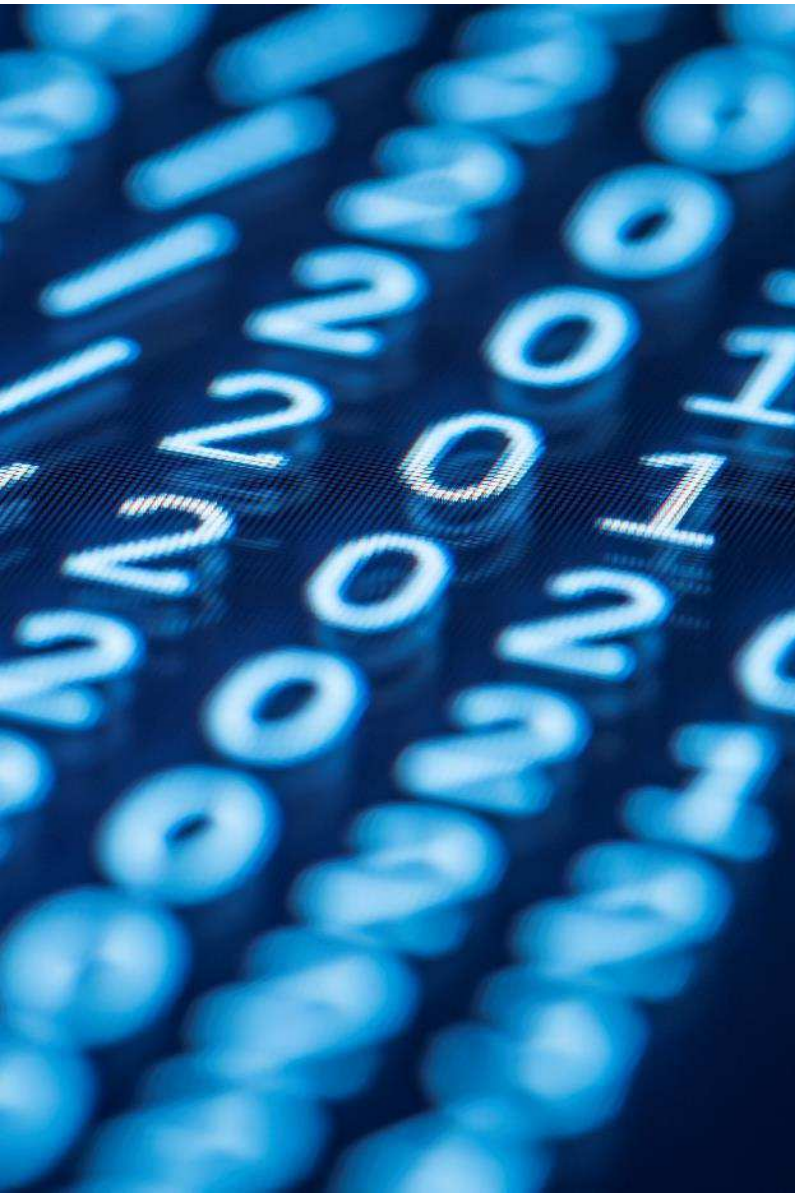
OUR APPROACH – AT A GLANCE

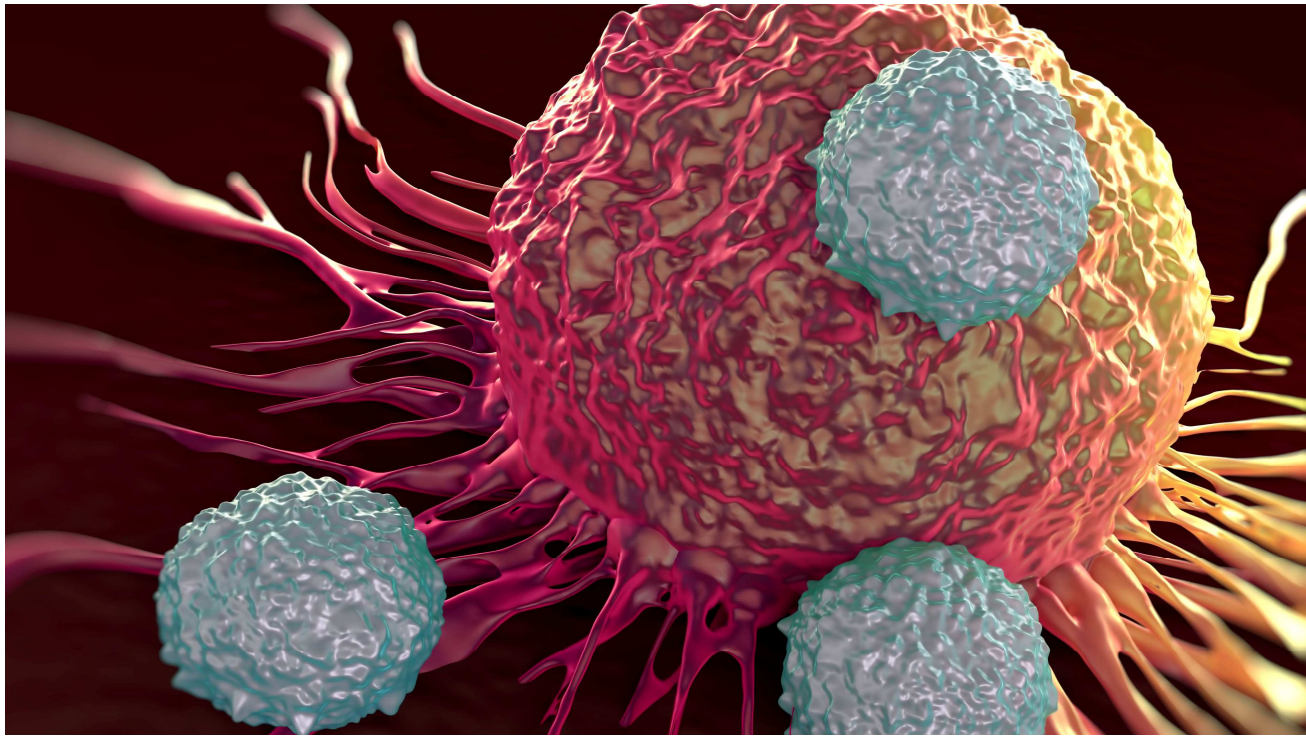


REPORT CARD (AS OF 15-APR)

	Goal	Status	Notes*
1	Obtain a baseline model	Completed	Logistic Regression Classifier based on Tfidf weighted word counts: • Balanced accuracy score: 0.540 • Balanced F1 score: 0.618 • Micro-average AUC: 0.760
2	Obtain an alternative model	Completed	Forward Neural Network based on mean Word2Vec word embeddings: • Balanced accuracy score: 0.397 • Balanced F1 score: 0.443 • Micro-average AUC: 0.663
3	Deal with imbalanced classes	Completed	Used partial ADASYN oversampling
4	Reduce overfitting (too many features)	Completed	Used PCA (no. of features dropped from ~76k to ~2k!)
5	Evaluate BERT or BioBERT word embeddings	Abandoned	Abandoned this approach as BERT has 1,024 word limitation
6	Evaluate ELMo word embeddings	Abandoned	Faced significant difficulty in creating the ELMo embeddings due to slow local processing speed and limited memory
7	Enhanced neural network with LSTM	In-progress	Attempt to introduce LSTM units into neural network

Legend: BERT = Bidirectional Encoder Representations from Transformers, ELMo = Embedding from Language Models, GloVe = Global Vectors (from Stanford University)
Note: * - all scores shown are based on validation dataset





THANK YOU

YUCHYE@GMAIL.COM