Team members: Qiuyang Wang, Xu Cai, Yukun Zhang

Contributions:

- Qiuyang Wang:
  - chose BioBERT as the text classification model;
  - wrote the major finetuning and prediction code for this project;
  - kept trying different hyperparameters for better performance.

- Xu Cai:
  - Preprocessing medical notes. For patients who have disease, split their medical note into several notes to address original data imbalance issue. Remove unimportant text including characters, repetitive key words. Fine-tuned and trained llama2 model, but finally turned out not good enough and searched for other models.

- Yukun Zhang:
  - Preprocessing medical notes by splitting the notes and removing uppercase words. Adjust the parameters of different models and add some functions to better generate the results. Train the models to get the best result.

Approach:

1. **Split the "True" medical notes into many sub-notes for that patient in order to balance the data.**
2. **Tokenized the notes and then encoded them by using BERT tokenizer.**
3. **Employed dataloader to put data into batches (size = 16)**
4. **Trained the model**