ICD-O code classification for pathological synoptic report diagnosis

Matthew Brennan, Haemin Lee, Hardi Rathod, Kevin Tsang

Pathology report

A pathology report is a document that contains the diagnosis determined by examining cells and tissues under a microscope.

Format: synoptic report and free text

Issue: Researchers cannot use just text to explore science, they would prefer medical codes

M8050	ConceptHierarchy	7	34	1	THYROID GLAND > TUMOR > Tumor Characteristics > Histologic Type > Papillary carcinoma, classic (usual, conventional)	8260
M8745	SmartDataValue	7	158	1	Desmoplastic melanoma	
M8430	ConceptHierarchy	6	47	1	MAJOR SALIVARY GLANDS > TUMOR > Histologic Type > Mucoepidermoid carcinoma, intermediate grade	
M8500	SmartDataValue	9	79	2	Ductal carcinoma in situ. Classified as Tis (DCIS) or Tis (Paget)	
M8742	SmartDataValue	9	8	1	Lentigo maligna melanoma	
M8260	ConceptHierarchy	6	45	1	KIDNEY: Nephrectomy > TUMOR > Histologic Type > Papillary renal cell carcinoma	
M8380	ConceptHierarchy	5	23	1	ENDOMETRIUM > TUMOR > Histologic Type > Endometrioid carcinoma, NOS	
M8490	ConceptHierarchy	5	216	1	STOMACH > TUMOR > Histologic Type > Adenocarcinoma > Alternative Optional Classification (based on WHO classification) > Poorly cohesive carcinoma (including signet-ring cell carcinoma and other variants)	
M8580	ConceptHierarchy	5	144	1	THYMUS > TUMOR > Histologic Type > Type B2 thymoma	8584
M8800	ConceptHierarchy	8	103	1	UTERUS (SARCOMA) > TUMOR > Histologic Type	
M8200	ConceptHierarchy	6	321	1	LIP AND ORAL CAVITY > TUMOR > Histologic Type > Adenoid cystic carcinoma, cribriform pattern	
M8441	SmartDataValue	6	31	1	Serous carcinoma	
M8743	SmartDataValue	6	6	1	Superficial spreading melanoma	
M8940	ConceptHierarchy	7	308	2	MAJOR SALIVARY GLANDS > TUMOR > Histologic Type > Preexisting Pleomorphic Adenoma Component > Carcinoma ex pleomorphic adenoma, invasive	8941
M8013	ConceptHierarchy	6	253	1	LUNG > TUMOR > Histologic Type > Combined large cell neuroendocrine carcinoma (LCNEC and other non-small cell component) > Type of Other Non-small Cell Carcinoma Component	
M8130	ConceptHierarchy	4	63	1	URINARY BLADDER: Cystectomy, Anterior Exenteration > TUMOR > Histologic Type > Papillary urothelial carcinoma, invasive	
M8312	SmartDataValue	5	3	1	Clear cell renal cell carcinoma	
M8340	ConceptHierarchy	4	106	1	THYROID GLAND > TUMOR > Tumor Characteristics > Histologic Type > Papillary carcinoma, follicular variant, encapsulated / well demarcated, non-invasive	
M9071	SmartDataValue	4	506	2	Yolk sac tumor (endodermal sinus tumor)	
M8041	ConceptHierarchy	4	240	1	URINARY BLADDER: Cystectomy, Anterior Exenteration > TUMOR > Histologic Type > Small cell neuroendocrine carcinoma	
M8575	SmartDataValue	4	66	1	Metaplastic carcinoma, mixed epithelial and mesenchymal type	
M8936	ConceptHierarchy	3	92	3	GASTROINTESTINAL STROMAL TUMOR (GIST): Resection > TUMOR > Histologic Type > Gastrointestinal stromal tumor, spindle cell type	
M8046	ConceptHierarchy	4	253	1	LUNG > TUMOR > Histologic Type > Combined large cell neuroendocrine carcinoma (LCNEC and other non-small cell component) > Type of Other Non-small Cell Carcinoma Component	8013
M8071	ConceptHierarchy	3	135	1	LUNG > TUMOR > Histologic Type > Invasive squamous cell carcinoma, keratinizing	
M8144	ConceptHierarchy	3	56	1	DISTAL EXTRAHEPATIC BILE DUCTS > TUMOR > Histologic Type > Adenocarcinoma, intestinal type	
M8510	ConceptHierarchy	3	328	1	THYROID GLAND > TUMOR > Tumor Characteristics > Histologic Type > Medullary carcinoma	
M8560	ConceptHierarchy	3	113	1	LUNG > TUMOR > Histologic Type > Adenosquamous carcinoma	
M8890	ConceptHierarchy	3	165	1	UTERUS (SARCOMA) > TUMOR > Histologic Type > Leiomyosarcoma, epithelioid type	
M9070	ConceptHierarchy	3	250	1	TESTIS: Radical Orchiectomy > TUMOR > Histologic Type > Intratubular embryonal carcinoma	
M9220	ConceptHierarchy	3	306	1	BONE: Resection > TUMOR > Histologic Type > Chondrosarcoma grade II	
M8160	ConceptHierarchy	3	349	1	INTRAHEPATIC BILE DUCTS > TUMOR > Histologic Type > Intrahepatic cholangiocarcinoma	
M8211	ConceptHierarchy	3	241	1	INVASIVE CARCINOMA OF THE BREAST: Resection > TUMOR > Histologic Type > Tubular carcinoma	
M8335	SmartDataValue	4	82	1	Follicular carcinoma, encapsulated angioinvasive	
M9085	ConceptHierarchy	6	340	1	TESTIS: Radical Orchiectomy > TUMOR > Histologic Type > Mixed germ cell tumor > Seminoma (percentage)	
M8310	SmartDataValue	3	188	1	Clear cell carcinoma	
M8330	SmartDataValue	3	76	1	Follicular carcinoma	
M8721	SmartDataValue	3	16	1	Nodular melanoma	
M9260	ConceptHierarchy	2	357	2	EWING SARCOMA: Resection > Histologic Type > Ewing Sarcoma	
M8072	ConceptHierarchy	2	51	1	LUNG > TUMOR > Histologic Type > Invasive squamous cell carcinoma, non-keratinizing	
M8083	ConceptHierarchy	2	61	1	PENIS > TUMOR > Histologic Type > Basaloid squamous cell carcinoma	
M8246	ConceptHierarchy	2	28	1	LIP AND ORAL CAVITY > TUMOR > Histologic Type > Moderately differentiated neuroendocrine carcinoma (atypical carcinoid tumor)	
M8249	ConceptHierarchy	2	28	1	LIP AND ORAL CAVITY > TUMOR > Histologic Type > Moderately differentiated neuroendocrine carcinoma (atypical carcinoid tumor)	
M8344	ConceptHierarchy	2	319	1	THYROID GLAND > TUMOR > Tumor Characteristics > Histologic Type > Papillary carcinoma, tall cell variant	

World Health Organization

International Classification of Diseases for Oncology (ICD-O)

Figure 3. ICD-O Coding of Lung Neoplasms

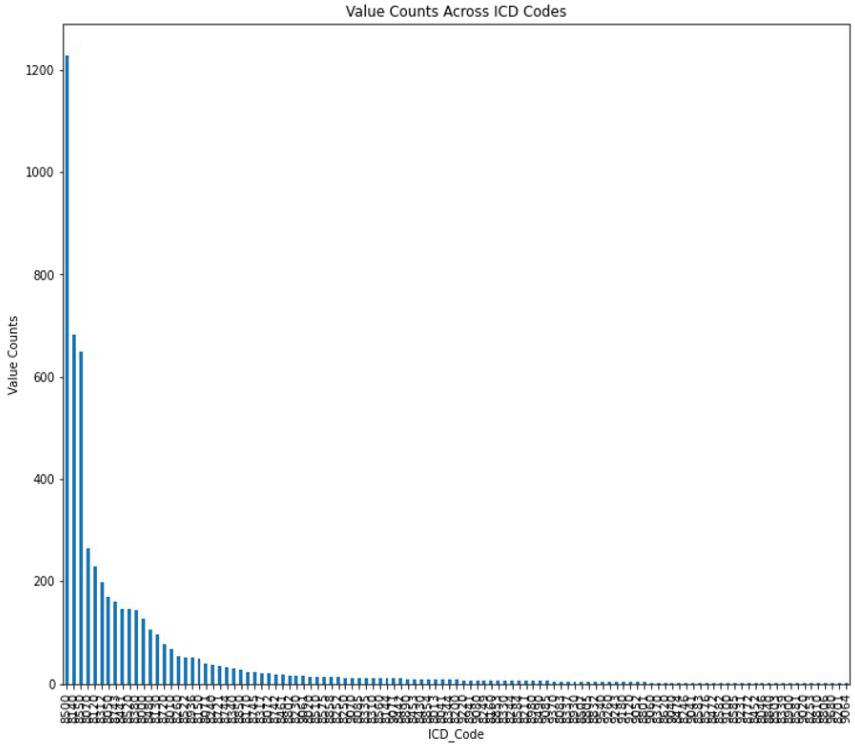
Lung Neoplasm	Topography Code	Behavior Code
Malignant neoplasm of the lung (such as carcinoma)	C34.9	M-8010/3
Metastatic neoplasm of the lung (such as metastatic seminoma from the testis)	C34.9	M-9061/6
In situ neoplasm of the lung (such as squamous carcinoma in situ)	C34.9	M-8070/2
Benign neoplasm of lung (such as adenoma)	C34.9	M-8140/0
Uncertain behavior of neoplasm of lung (such as carcinoid of uncertain behavior)	C34.9	M-8240/1

Figure 4. ICD-O Behavior Code and Corresponding Section of Chapter II, ICD-10

Behavior Code	Category	Term
/0	D10-D36	Benign neoplasms
/1	D37-D48	Neoplasms of uncertain and unknown behavior
/2	D00-D09	In situ neoplasms
/3	C00-C76, C80-C97	Malignant neoplasms stated or presumed to be primary
/6	C77-C79	Malignant neoplasms, stated or presumed to be secondary

Data example and distribution

Dx	Code
Acinar adenocarcinoma	8550
Acinar adenocarcinoma	8550
Acinar adenocarcinoma	8550
Acinar adenocarcinoma	8550
Acinar adenocarcinoma	8550
Acinar adenocarcinoma	8550
Acinar adenocarcinoma	8550
Acinar adenocarcinoma	8550
Acinar adenocarcinoma	8550
Acinar adenocarcinoma	8550
Acinar adenocarcinoma	8550
Acinar adenocarcinoma	8550
Acinar adenocarcinoma	8550
Acral melanoma	8744
Acral melanoma	8744
Acral-lentiginous melanoma	8744
Acral-lentiginous melanoma	8744
Acral-lentiginous melanoma	8744
Adenocarcinoma, intestinal type	8144
Adenoid cystic carcinoma, cribriform pattern#	8200
Carcinosarcoma (malignant mixed Müllerian tumor)	8000
Carcinosarcoma (malignant mixed Müllerian tumor)	8000
Clear cell renal cell carcinoma	8312
Clear cell renal cell carcinoma	8312
Clear cell renal cell carcinoma	8312
Clear cell renal cell carcinoma	8312
Clear cell renal cell carcinoma	8312
Clear cell renal cell carcinoma	8312
Clear cell renal cell carcinoma	8312
Clear cell renal cell carcinoma	8312
Clear cell renal cell carcinoma	8312
Clear cell papillary renal cell carcinoma	8260



6797 rows of diagnosis synoptic text and its classified ICD-O code

Data is imbalanced as 8500 was assigned 1250 times.

Objective

- Build an open-source classifier with the appropriate encoding method and run experiments on hyperparameters for upsampling, encoding and models.

Preprocessing and Data Embedding

Preprocessing Methods

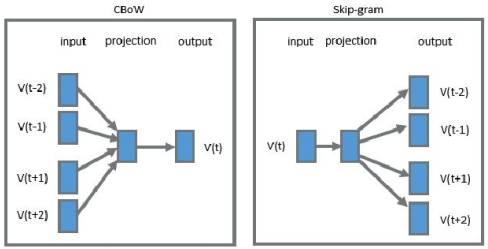
- Removed all non-alphanumeric values and other types of punctuation
- Casted all letters to lowercase

Example:

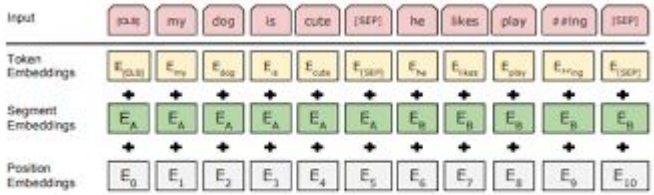
- Before: “Squamous cell carcinoma, conventional”
- After: “squamous cell carcinoma conventional”

Data Embedding Selection

- Word2Vec



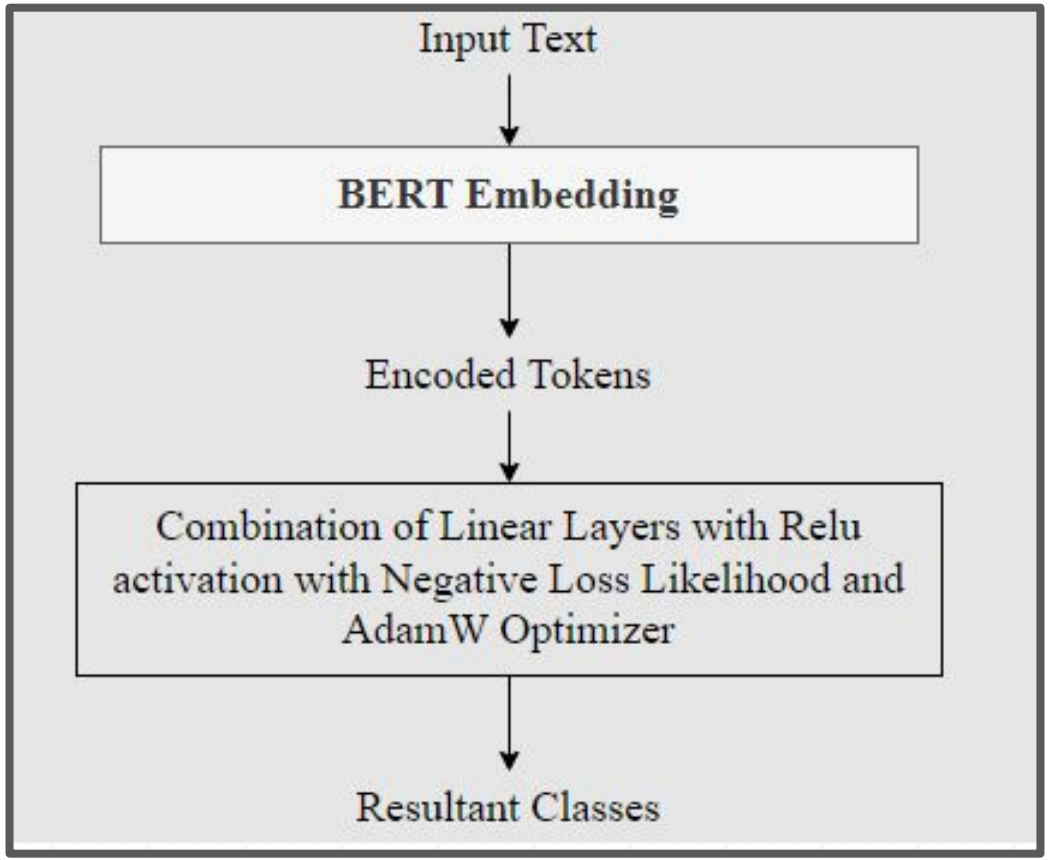
- BERT Encodings



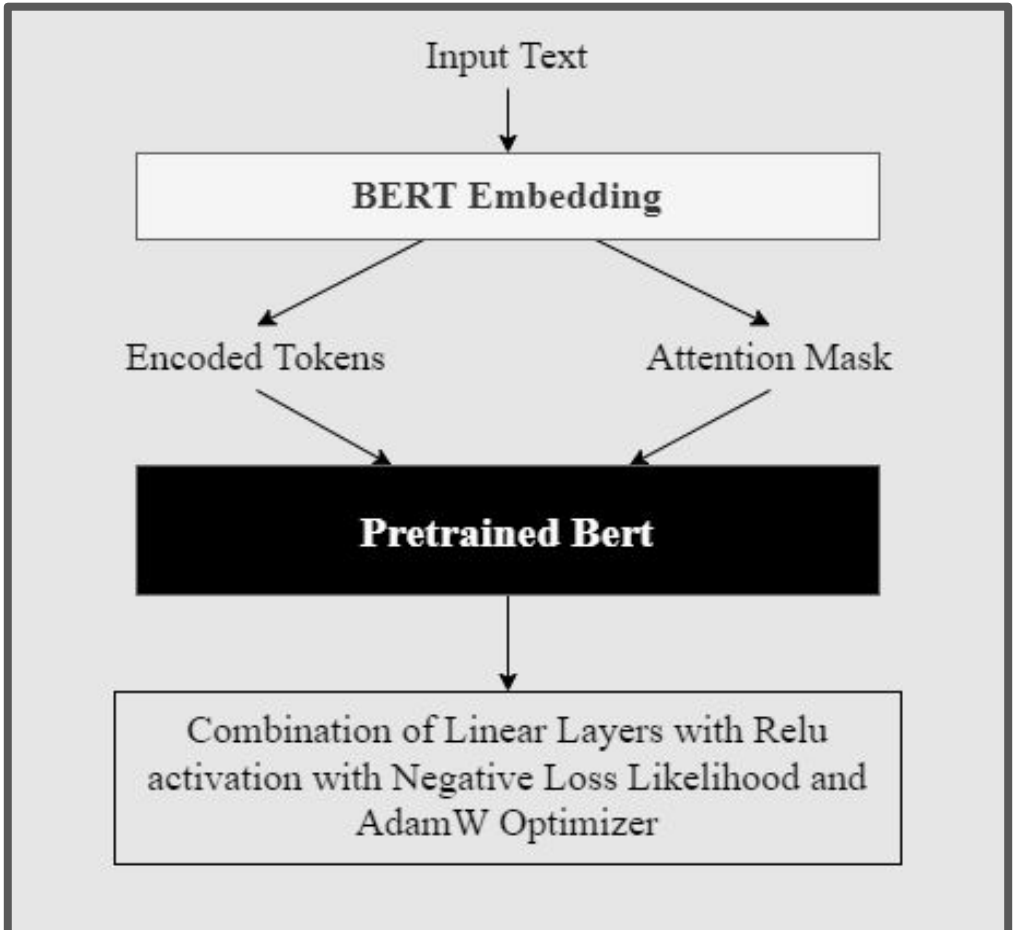
- Fine tuned BERT Encodings

Model and Algorithms

1. Simple Neural Network without Fine-tuning

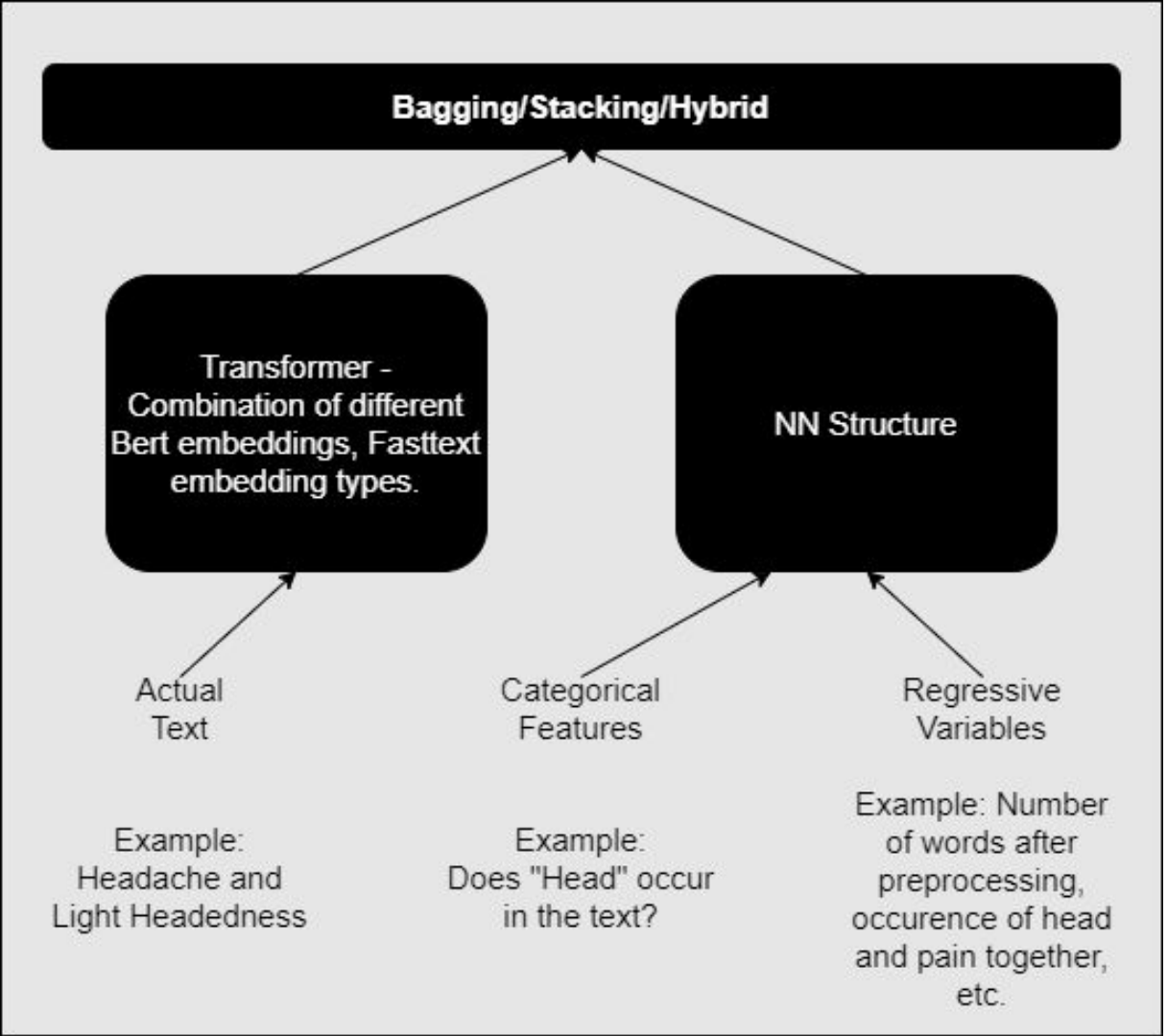


2. Neural Network with Bert Fine-Tuning



3. Tree Based Algorithms

- Using relations between various features of the data to determine the class
  - Example:
  - Mutation in BRCA1 and/or BRCA2 -> higher probability of Breast Cancer
- Pretty Slow
- Use of distributed/ scalable algorithm and pruning Techniques
- LightGBM Model, XGBoost, CatBoost



4. Multiple Encoding Based Multimodal Neural Network

Brief Results

Brief Results

Model	Embedding/Encoding/ Features	Accuracy	F1-score
Tree Based Algorithms			
LightGBM Model	BERT + CAT + NUM	98	99
CatBoost Model	BERT + CAT + NUM	96	98
XGBoost Model	BERT + CAT + NUM	98	95
Neural Network Based Algorithm			
Simple NN (no finetuning)	BERT	44	44
NN with BERT FineTuning	FineTuned BERT	72	72
M. Encoding Multimodal	BERT, FastText + CAT + NUM	99	99
BiLSTM	BERT	99.1	99.1

Best Model: BiLSTM

BLSTM - Model Structure

- Epochs-100, SGD Optimization, Cross Entropy Loss Fn.
  - Hyperparameter tuning on: batch_size, epochs, scheduler, lr, momentum.

```
class BLSTM(nn.Module):

    def __init__(self):
        super().__init__()

        self.lstm = nn.LSTM(input_size=34, hidden_size=256,
                            num_layers=1, batch_first=True, bidirectional=True)
        self.dropout = nn.Dropout(0.33)
        self.linear1 = nn.Linear(512, 128)
        self.elu = nn.ELU()
        self.linear2 = nn.Linear(128, 49)


    def forward(self, inputs):
        lstm_out, self.hidden = self.lstm(inputs.view(len(inputs), 1, -1))
        lstm_out_dropped = self.dropout(lstm_out)
        out = self.linear1(lstm_out_dropped.view(len(inputs), -1))
        elu_out = self.elu(out)
        l2_out = self.linear2(elu_out)
        log_probs = F.log_softmax(l2_out, dim=1)
        return log_probs
```

Encoder/Decoder Examples - BLSTM

```
Text: human papillomavirus hpvmediated positive squamous cell carcinoma oropharynx
Code: 8085
BERT Encoded: tensor([  101,  2529,  6643,  8197,  7174,  2863, 23350,  6522,  2615, 16969,
         3064,  3893,  5490,  6692, 27711,  3526,  2482, 21081,  2863, 20298,
        21890, 18143,  2595,   102,     0,     0,     0,     0,     0,     0,
            0,     0,     0,     0])
Pred: 8085

Text: invasive carcinoma of no special type ductal not otherwise specified
Code: 8500
BERT Encoded: tensor([  101, 17503,  2482, 21081,  2863,  1997,  2053,  2569,  2828, 23245,
         2389,  2025,  4728,  9675,   102,     0,     0,     0,     0,     0,
            0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
            0,     0,     0,     0])
Pred: 8500
```

Encoder/Decoder Examples - BLSTM

```
Text: squamous cell carcinoma conventional keratinizing
Code: 8071
BERT Encoded: tensor([  101,  5490,  6692, 27711,  3526,  2482, 21081,  2863,  7511, 17710,
         8609,  5498,  6774,   102,     0,     0,     0,     0,     0,     0,
            0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
            0,     0,     0,     0])
Pred: 8071


Text: pt tumor cm or less in greatest dimension
Code: 8000
BERT Encoded: tensor([  101, 13866, 13656,  4642,  2030,  2625,  1999,  4602,  9812,   102,
            0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
            0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
            0,     0,     0,     0])
Pred: 8000
```

Discussion

Our results confirmed that BiLSTM performed best in this type of text classification.

Upsampling

ICD-O cut off frequency	**Accuracy**
10	98%
5	97%

As we decreased the ICD-O cut off, the accuracy of the model goes down.

Future Scope

Negation

Our model is not robust enough to classify intentionally negated data. Further preprocessing and model adjustments are needed in future experiments.

Explainable AI

Explainability of the models allow researchers to learn how the models predict labels vs how physician diagnoses