Capstone Final Project

Berat Bozkurt

Video Link:
https://drive.google.com/file/d/1RwdzStHDpD09nMRa4IcC60BzX6nvlX3O/view

Background

Classification of articles on specific types of cancer by looking at the abstracts

- **Leukaemia**
- **Non-Hodgkin Lymphoma**
- **Bladder Cancer**
- **Thyroid Cancer**

2 Past Model

- CatBoostClassifier
- Convolutional Neural Network (CNN)

2 More Model

- RandomForestClassifier
- Long-Short Term Memory (LSTM)

Cat Boost

```
0:      learn: 0.9644511        total: 5.49s    remaining: 2m 39s
1:      learn: 0.7580971        total: 11.3s    remaining: 2m 38s
2:      learn: 0.6224037        total: 15.8s    remaining: 2m 22s
3:      learn: 0.5305688        total: 21.5s    remaining: 2m 19s
4:      learn: 0.4598252        total: 27s      remaining: 2m 14s
5:      learn: 0.4076880        total: 32.8s    remaining: 2m 11s
6:      learn: 0.3645890        total: 37.3s    remaining: 2m 2s
7:      learn: 0.3293235        total: 42.7s    remaining: 1m 57s
8:      learn: 0.3036073        total: 48s      remaining: 1m 52s
9:      learn: 0.2788366        total: 52.4s    remaining: 1m 44s
10:     learn: 0.2620795        total: 56.9s    remaining: 1m 38s
11:     learn: 0.2465066        total: 1m 2s    remaining: 1m 33s
12:     learn: 0.2322181        total: 1m 6s    remaining: 1m 26s
13:     learn: 0.2200928        total: 1m 10s   remaining: 1m 20s
14:     learn: 0.2105664        total: 1m 15s   remaining: 1m 15s
15:     learn: 0.2026067        total: 1m 20s   remaining: 1m 10s
16:     learn: 0.1983615        total: 1m 24s   remaining: 1m 4s
17:     learn: 0.1913409        total: 1m 29s   remaining: 59.9s
18:     learn: 0.1866337        total: 1m 34s   remaining: 54.5s
19:     learn: 0.1816165        total: 1m 38s   remaining: 49.2s
20:     learn: 0.1765551        total: 1m 43s   remaining: 44.4s
21:     learn: 0.1742230        total: 1m 47s   remaining: 39.1s
22:     learn: 0.1695757        total: 1m 52s   remaining: 34.1s
23:     learn: 0.1667856        total: 1m 57s   remaining: 29.3s
24:     learn: 0.1643047        total: 2m 1s    remaining: 24.4s
25:     learn: 0.1611411        total: 2m 5s    remaining: 19.3s
26:     learn: 0.1593874        total: 2m 10s   remaining: 14.5s
27:     learn: 0.1578821        total: 2m 14s   remaining: 9.61s
28:     learn: 0.1544903        total: 2m 21s   remaining: 4.86s
29:     learn: 0.1534716        total: 2m 25s   remaining: 0us
```

	precision	recall	f1-score	support
0	1.00	0.94	0.97	13853
1	0.95	0.99	0.97	67966
2	0.96	0.90	0.93	18432
3	1.00	0.94	0.97	14924
accuracy			0.96	115175
macro avg	0.98	0.94	0.96	115175
weighted avg	0.96	0.96	0.96	115175

Convolutional Neural Network

```
Epoch 1/10
2100/2100 - 600s - loss: 0.1037 - accuracy: 0.9645 - val_loss: 0.0665 - val_accuracy: 0.9786 - 600s/epoch - 286ms/step
Epoch 2/10
2100/2100 - 754s - loss: 0.0423 - accuracy: 0.9863 - val_loss: 0.0690 - val_accuracy: 0.9778 - 754s/epoch - 359ms/step
Epoch 3/10
2100/2100 - 774s - loss: 0.0167 - accuracy: 0.9949 - val_loss: 0.0819 - val_accuracy: 0.9780 - 774s/epoch - 369ms/step
Epoch 4/10
2100/2100 - 778s - loss: 0.0068 - accuracy: 0.9978 - val_loss: 0.1130 - val_accuracy: 0.9758 - 778s/epoch - 371ms/step
Epoch 5/10
2100/2100 - 733s - loss: 0.0043 - accuracy: 0.9986 - val_loss: 0.1409 - val_accuracy: 0.9729 - 733s/epoch - 349ms/step
Epoch 6/10
2100/2100 - 723s - loss: 0.0040 - accuracy: 0.9986 - val_loss: 0.1441 - val_accuracy: 0.9759 - 723s/epoch - 344ms/step
Epoch 7/10
2100/2100 - 680s - loss: 0.0029 - accuracy: 0.9991 - val_loss: 0.1665 - val_accuracy: 0.9742 - 680s/epoch - 324ms/step
Epoch 8/10
2100/2100 - 734s - loss: 0.0027 - accuracy: 0.9991 - val_loss: 0.1466 - val_accuracy: 0.9758 - 734s/epoch - 350ms/step
Epoch 9/10
2100/2100 - 754s - loss: 0.0025 - accuracy: 0.9993 - val_loss: 0.1487 - val_accuracy: 0.9755 - 754s/epoch - 359ms/step
Epoch 10/10
2100/2100 - 720s - loss: 0.0022 - accuracy: 0.9993 - val_loss: 0.1604 - val_accuracy: 0.9761 - 720s/epoch - 343ms/step
```

	precision	recall	f1-score	support
0	0.99	0.98	0.98	13853
1	0.98	0.98	0.98	67966
2	0.94	0.95	0.95	18432
3	0.98	0.98	0.98	14924
accuracy			0.98	115175
macro avg	0.97	0.97	0.97	115175
weighted avg	0.98	0.98	0.98	115175

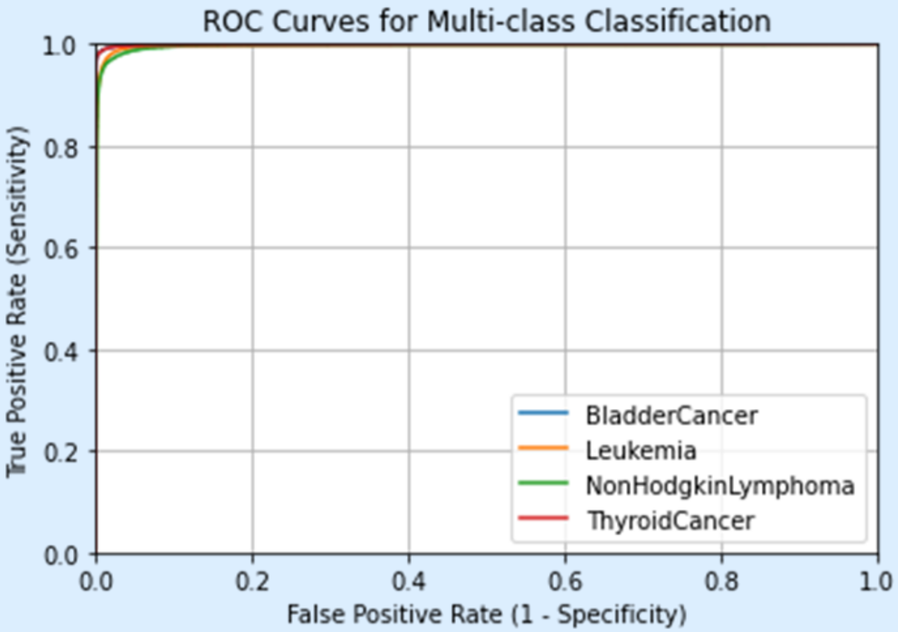
Random Forest

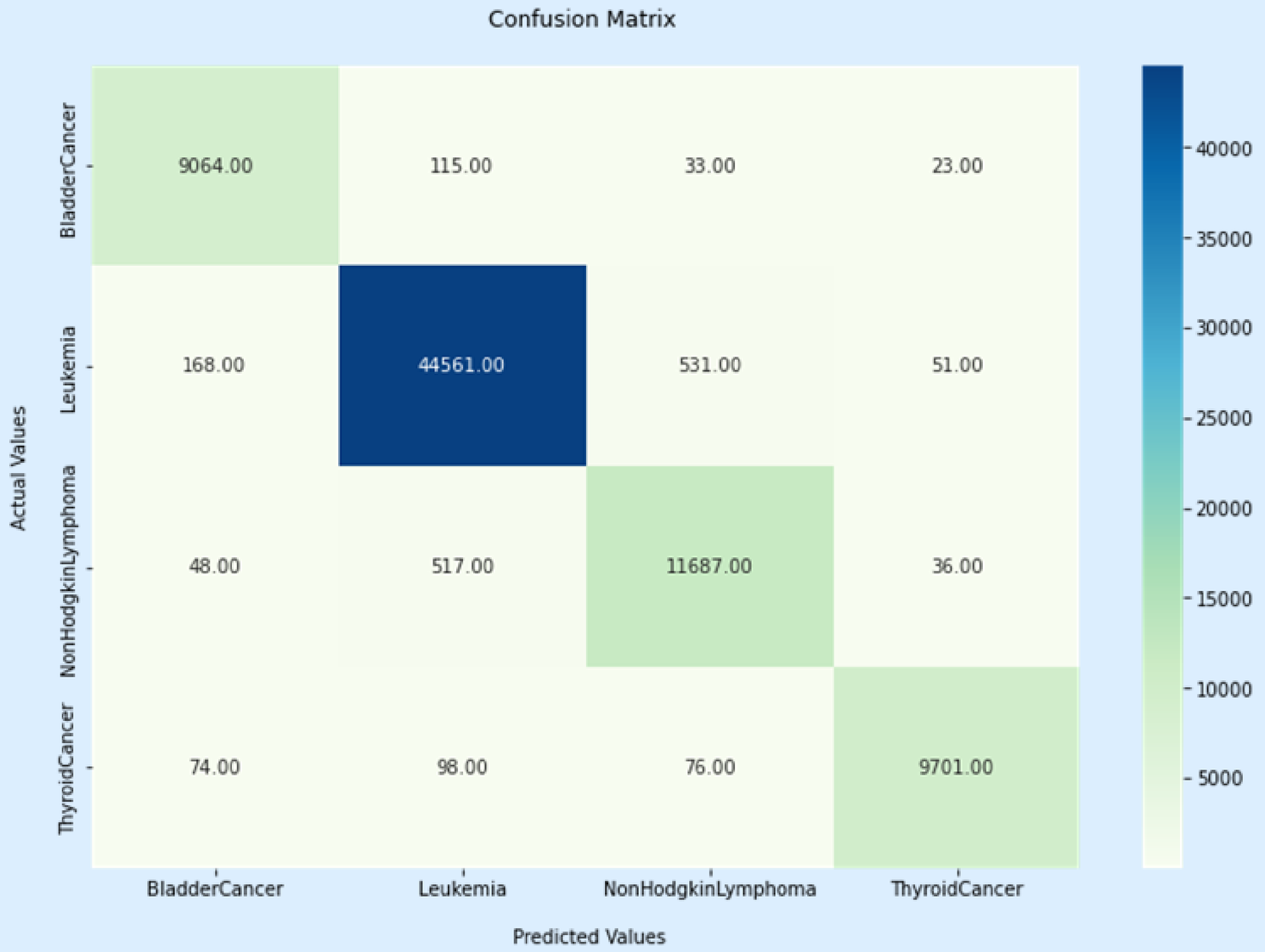
	precision	recall	f1-score	support
BladderCancer	0.99	0.94	0.96	9235
Leukemia	0.93	0.99	0.96	45311
NonHodgkinLymphoma	0.97	0.81	0.89	12288
ThyroidCancer	0.99	0.93	0.96	9949
accuracy			0.95	76783
macro avg	0.97	0.92	0.94	76783
weighted avg	0.95	0.95	0.95	76783

LSTM

```
Epoch 1/5
2100/2100 - 1291s - loss: 0.7985 - accuracy: 0.6764 - val_loss: 0.3330 - val_accuracy: 0.8314 - 1291s/epoch - 615ms/step
Epoch 2/5
2100/2100 - 1318s - loss: 0.1682 - accuracy: 0.9260 - val_loss: 0.0848 - val_accuracy: 0.9737 - 1318s/epoch - 628ms/step
Epoch 3/5
2100/2100 - 1298s - loss: 0.0647 - accuracy: 0.9801 - val_loss: 0.0706 - val_accuracy: 0.9779 - 1298s/epoch - 618ms/step
Epoch 4/5
2100/2100 - 1217s - loss: 0.0503 - accuracy: 0.9842 - val_loss: 0.0750 - val_accuracy: 0.9763 - 1217s/epoch - 580ms/step
Epoch 5/5
2100/2100 - 1173s - loss: 0.0387 - accuracy: 0.9881 - val_loss: 0.0816 - val_accuracy: 0.9759 - 1173s/epoch - 558ms/step
Accuracy: 97.69%
```

	precision	recall	f1-score	support
0	0.97	0.98	0.98	9235
1	0.98	0.98	0.98	45311
2	0.95	0.95	0.95	12288
3	0.99	0.98	0.98	9949
accuracy			0.98	76783
macro avg	0.97	0.97	0.97	76783
weighted avg	0.98	0.98	0.98	76783


ROC Curves for Multi-class Classification
True Positive Rate (Sensitivity)
False Positive Rate (1 - Specificity)
0.0
0.2
0.4
0.6
0.8
1.0
BladderCancer
Leukemia
NonHodgkinLymphoma
ThyroidCancer


Confusion Matrix
BladderCancer
Leukemia
NonHodgkinLymphoma
ThyroidCancer
Actual Values
9064.00
115.00
33.00
23.00
168.00
44561.00
531.00
51.00
48.00
517.00
11687.00
36.00
74.00
98.00
76.00
9701.00
BladderCancer
Leukemia
NonHodgkinLymphoma
ThyroidCancer
Predicted Values
40000
35000
30000
25000
20000
15000
10000
5000

Model Tuning

1. For the traditional machine learning algorithms (CatBoostClassifier and Random Forest Classifier), use grid search to tune the hyperparameters.

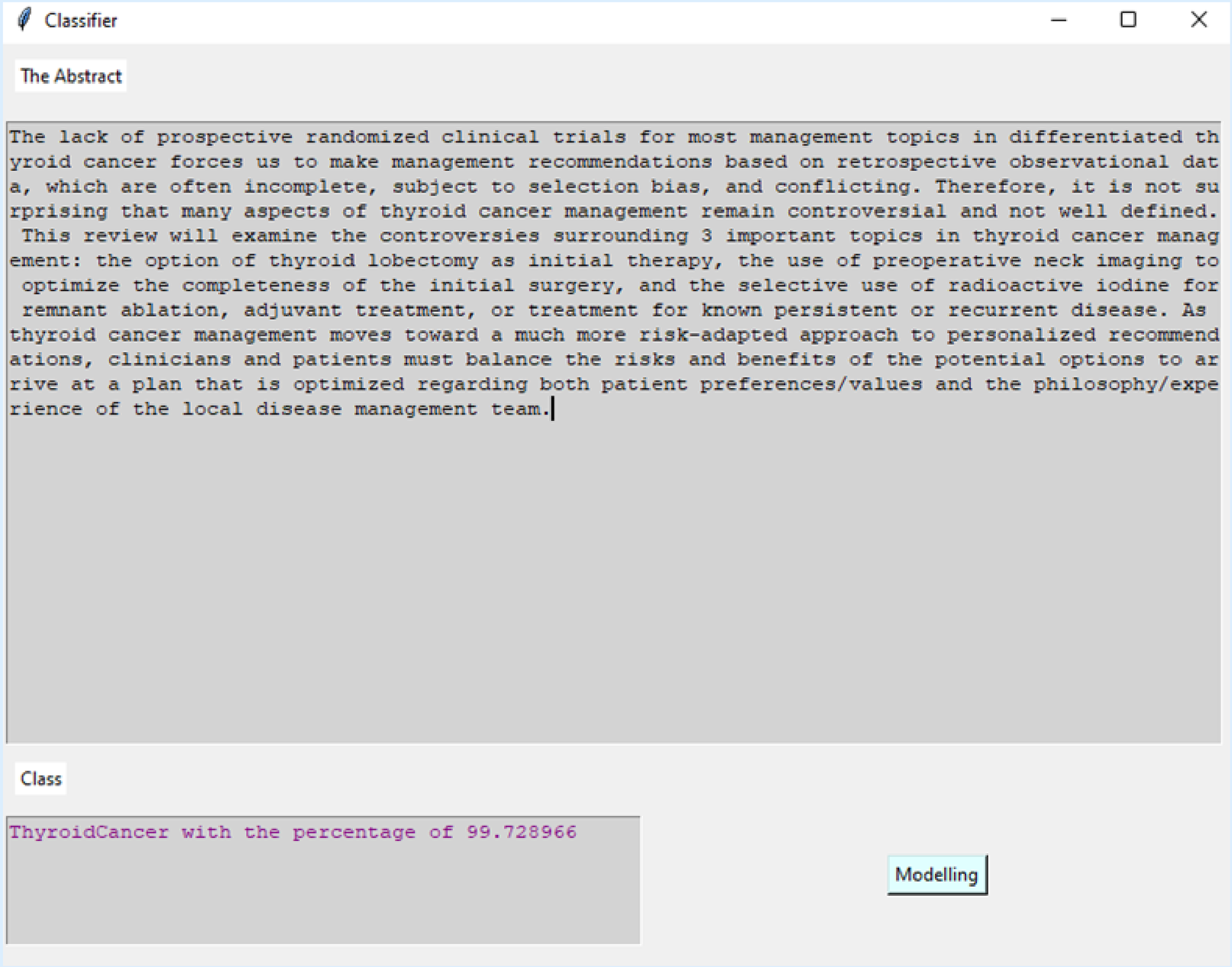
2. For the deep learning algorithms (CNN and LSTM), manually tune the hyperparameters by experimenting with different values and evaluating the models' performance on the validation set.

Traditional Machine Learning vs Deep Learning

- Deep learning is better for text classification because it can automatically learn features from raw text data

- Artificial neural networks can extract increasingly complex features as data passes through layer

DEMO


Classifier
The Abstract
The lack of prospective randomized clinical trials for most management topics in differentiated th
yroid cancer forces us to make management recommendations based on retrospective observational dat
a, which are often incomplete, subject to selection bias, and conflicting. Therefore, it is not su
rprising that many aspects of thyroid cancer management remain controversial and not well defined.
This review will examine the controversies surrounding 3 important topics in thyroid cancer manag
ement: the option of thyroid lobectomy as initial therapy, the use of preoperative neck imaging to
optimize the completeness of the initial surgery, and the selective use of radioactive iodine for
remnant ablation, adjuvant treatment, or treatment for known persistent or recurrent disease. As
thyroid cancer management moves toward a much more risk-adapted approach to personalized recommend
ations, clinicians and patients must balance the risks and benefits of the potential options to ar
rive at a plan that is optimized regarding both patient preferences/values and the philosophy/expe
rience of the local disease management team.
Class
ThyroidCancer with the percentage of 99.728966
Modelling

Thanks