Cancer Classification with Deep Learning using Genomic Data

Supervisor
Tareque Mohmud Chowdhury
Assistant Professor, Islamic University of Technology

Co-supervisor
Tasnim Ahmed
Lecturer, Islamic University of Technology

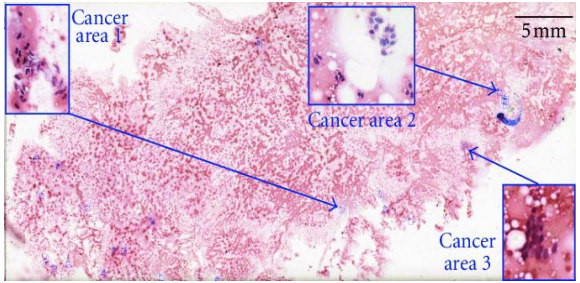
Ahmad Omar Ahsan	160041001
Azmaeen Bin Ansar	160041030
Minhajul Islam Minhaz	160041061

Outline

- Introduction
- Problem Statement
- Literature Review
- Research Challenges
- Implementation
- Result Analysis
- References

Introduction

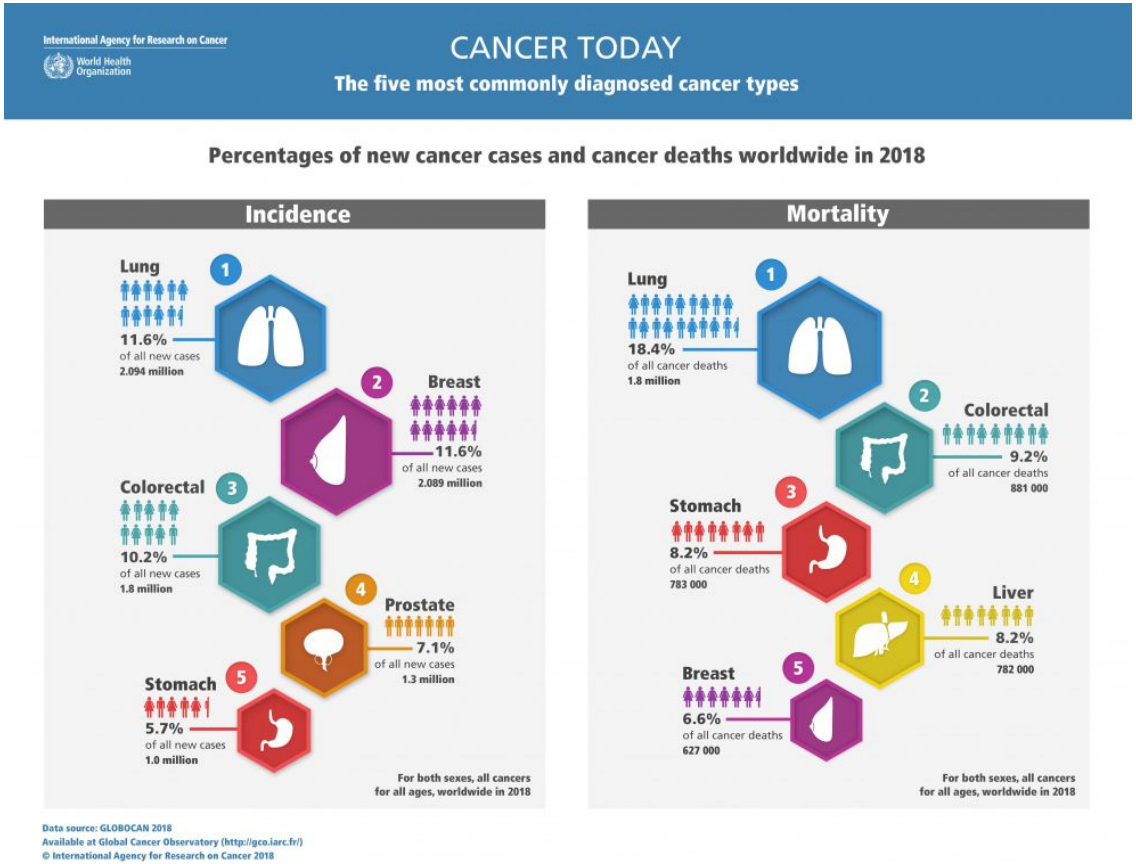
Traditional Cytological Identification



Radiological diagnosis

Histological diagnosis

Motivation

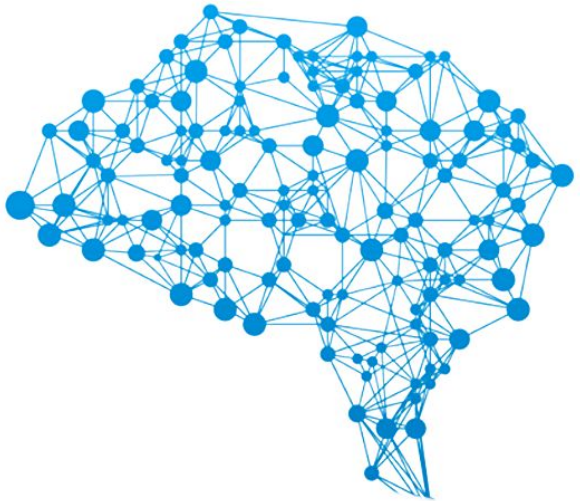


Source : WHO

Problem Statement

A cancer diagnosis system needs to be built that is

- Less expensive
- Stage independent
- Efficient



Literature Review

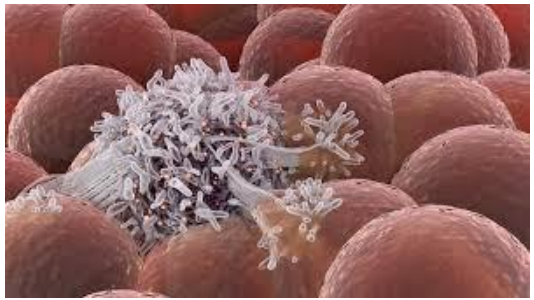
- The role of MicroRNAs in human cancer
  - Authors: Yong Peng, Carlo M Croce
  - Journal : Nature, Signal Transduction and Targeted Therapy (2016)

- A Machine Learning Approach for the Classification of Kidney Cancer Subtypes Using miRNA Genome Data
  - Authors: Ali Muhamed Ali, Hanqi Zhuang, Ali Ibrahim, Oneeb Rehman , Michelle Huang and Andrew Wu
  - Journal : MDPI, Applied Biosciences and Bioengineering (2018)

The role of miRNA in human cancer

miRNA dysregulation in cancer

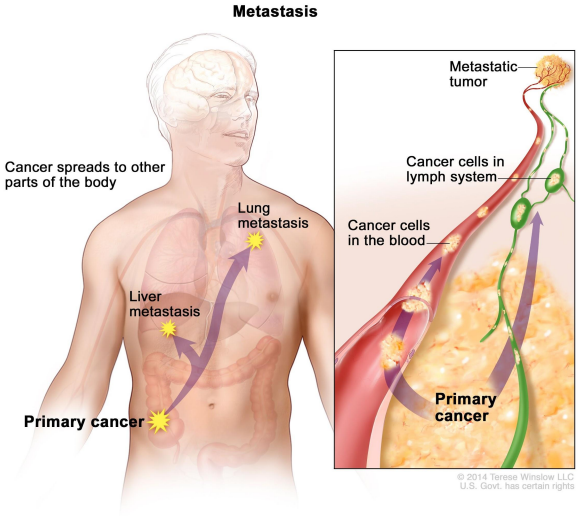
- miRNA are non coding nucleotides
- Compelling evidences have demonstrated that miRNA expression is dysregulated in human cancer
- Amplification or deletion of miRNA genes
- Transcriptional control of miRNAs
- Defects in miRNA biogenesis machinery
- miRNA drastically affects tumors



Contd : 2/3

Significance of altered miRNA in tumors

- Evading growth suppressors and sustaining proliferative signaling
- Resisting cell death
- Activating invasion and metastasis



Contd : 3/3

Paper 2

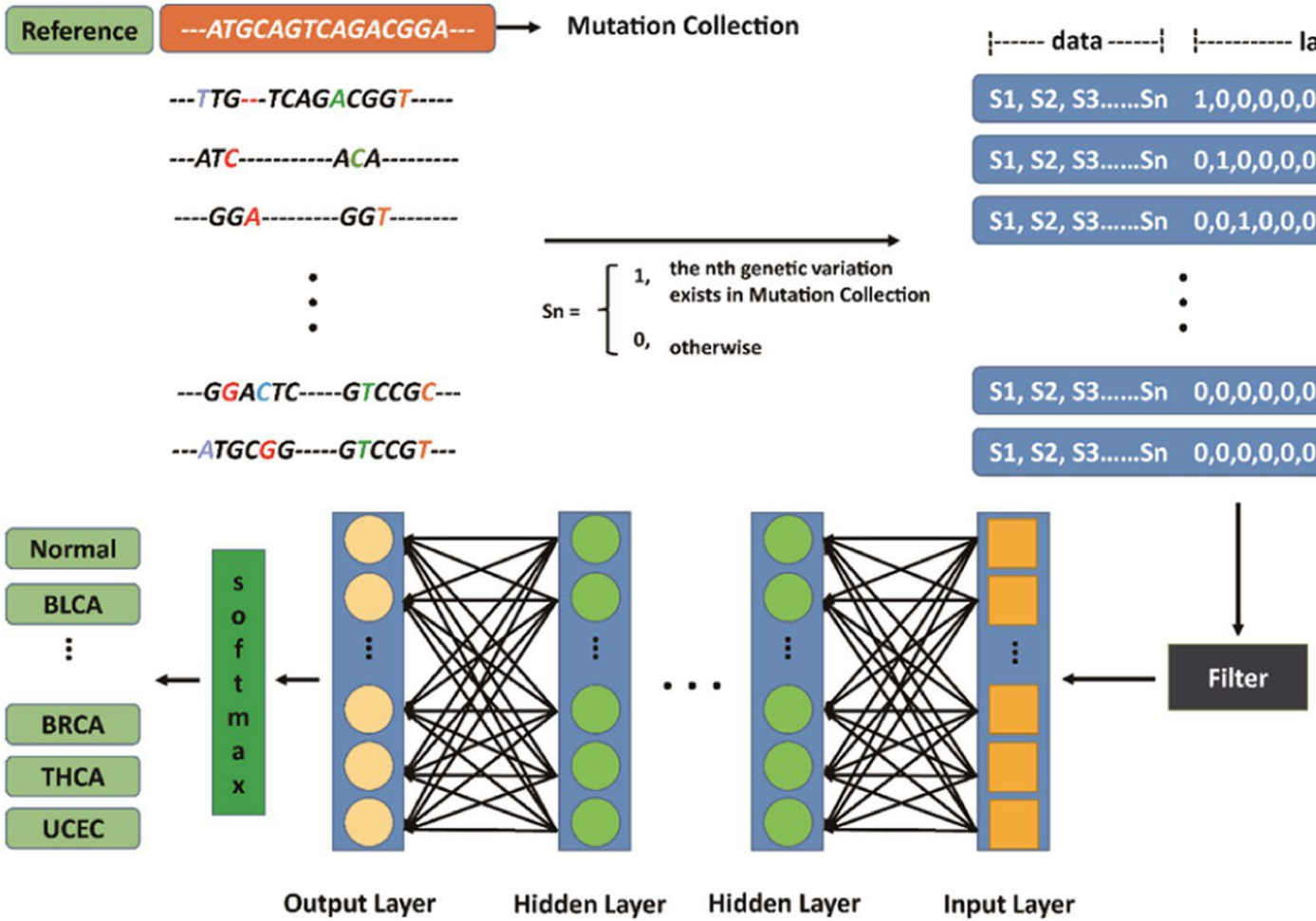
Identification of 12 cancer types through genome deep learning

Model description

Model	Accuracy
Specific(12)	97.47%
Mixture	70.08%
Total Specific	94.70%

- 14 models for 12 types of cancer
- 10,000 input features
- 4 layers

Contd : 2/3


Reference
---ATGCAGTCAGACGGA---
Mutation Collection
---TTG---TCAGACGGT-----
---ATC----------ACA---------
----GGA---------GGT--------
---GGACTC-----GTCCGC---
---ATGCGG-----GTCCGT---
Sn =
1, the nth genetic variation exists in Mutation Collection
0, otherwise
data
label
S1, S2, S3.....Sn 1,0,0,0,0,0,0,0,0,0,0,0
S1, S2, S3.....Sn 0,1,0,0,0,0,0,0,0,0,0,0
S1, S2, S3.....Sn 0,0,1,0,0,0,0,0,0,0,0,0
S1, S2, S3.....Sn 0,0,0,0,0,0,0,0,0,0,1,0
S1, S2, S3.....Sn 0,0,0,0,0,0,0,0,0,0,0,1
Filter
Normal
BLCA
BRCA
THCA
UCEC
softmax
Output Layer
Hidden Layer
Hidden Layer
Input Layer

Contd : 3/3

A Machine Learning Approach for the Classification of Kidney Cancer Subtypes Using miRNA Genome Data

Data Selection

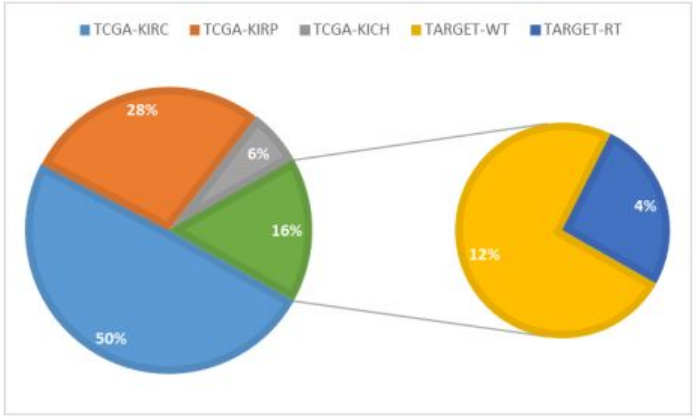


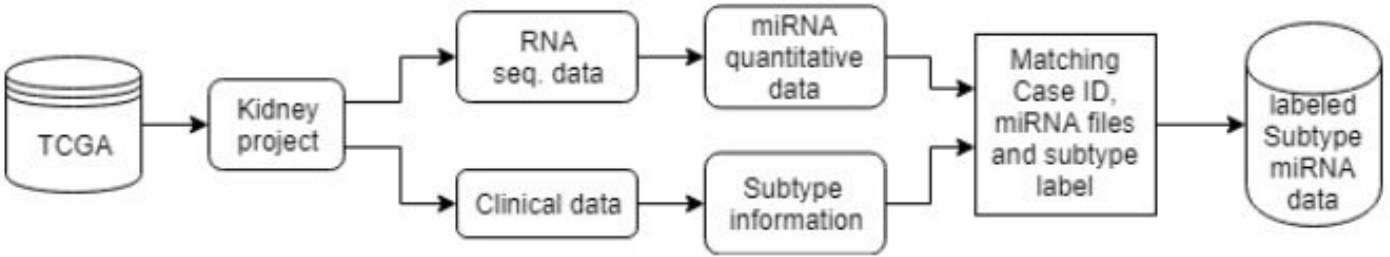
Figure 2. Kidney diseases projects available in the The Cancer Genome Atlas (TCGA) data repository.

Table 1. Table Type Styles.

Disease Type	Project Name	No. of Cases	No. of Files
Kidney Renal Clear Cell Carcinoma	TCGA-KIRC	516	616
Kidney Renal Papillary Cell Carcinoma	TCGA-KIRP	291	323
High-Risk Wilms Tumor	TARGET-WT	127	138
Kidney Chromophobe	TCGA-KICH	66	91
Rhabdoid Tumor	TARGET-RT	44	50

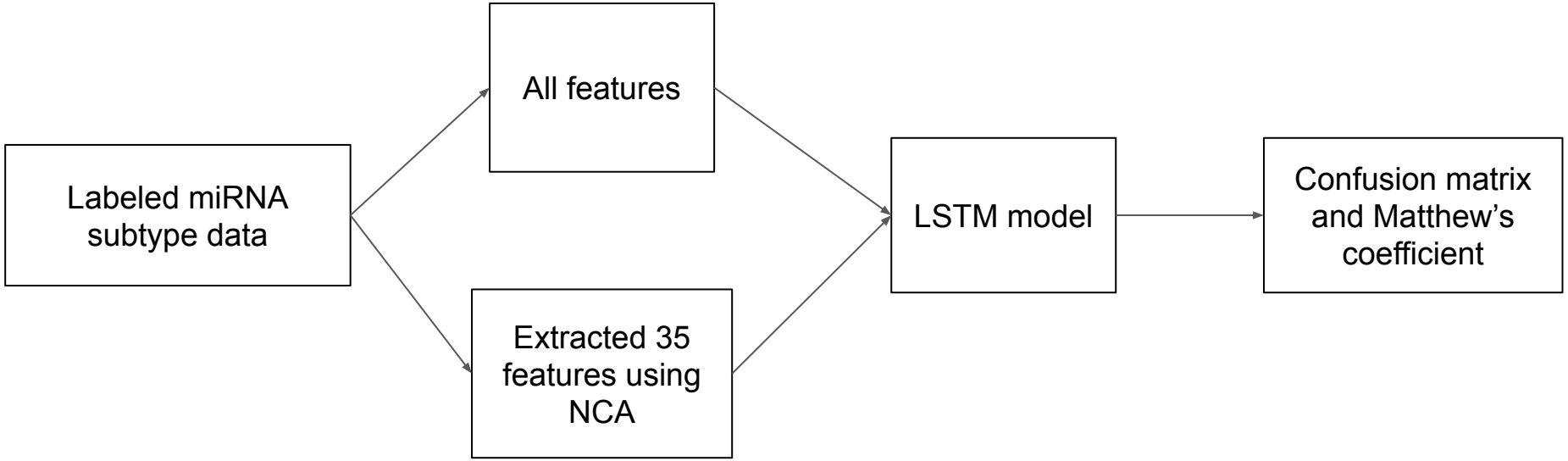
Contd : 2/5

Data Preprocessing Pipeline



Contd : 3/5

Model training pipeline



Contd : 4/5

Result Analysis

Table 5. Classification performance in terms of Matthews Correlation Coefficients.

Class	Selected 35 miRNA for Balanced Classes	1627 miRNA for Balanced Classes	Selected 35 miRNA for Unbalanced Classes	1627 miRNA for Unbalanced Classes
WT	0.953	0.968	0.963	0.943
KICH	0.898	0.938	0.880	0.817
KIRC	0.949	0.964	0.950	0.902
KIRP	0.917	0.957	0.911	0.877
RT	0.884	0.918	0.914	0.8965
Over all	0.920	0.949	0.924	0.887

Contd : 5/5

Matthew’s Correlation Coefficient (MCC)

A machine learning metric used to measure the quality of classification.

- Used widely in the field of Bioinformatics.
- Values range from -1 to 1.
- Basically a measure of the correlation between true and predicted values.

One of the most complete way to describe a confusion matrix.

- The coefficient takes into account true and false positive and negatives.
- Gives a balance measure even though classes are of very different sizes.

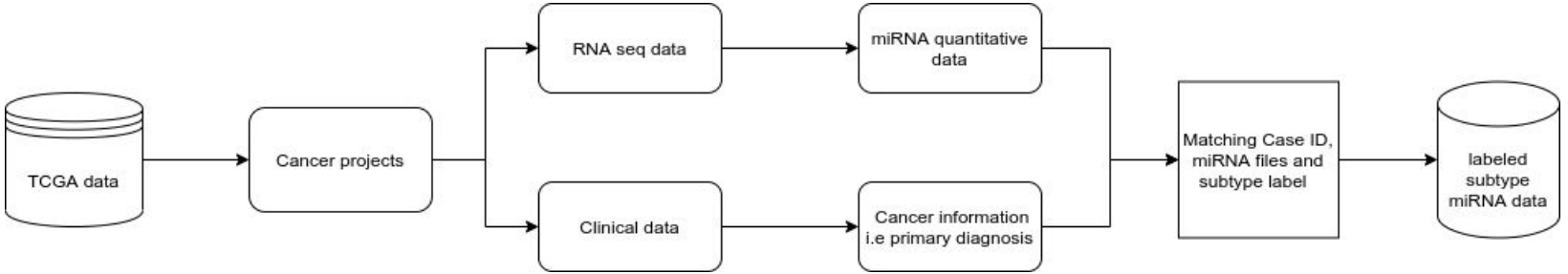
$$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Research Challenges

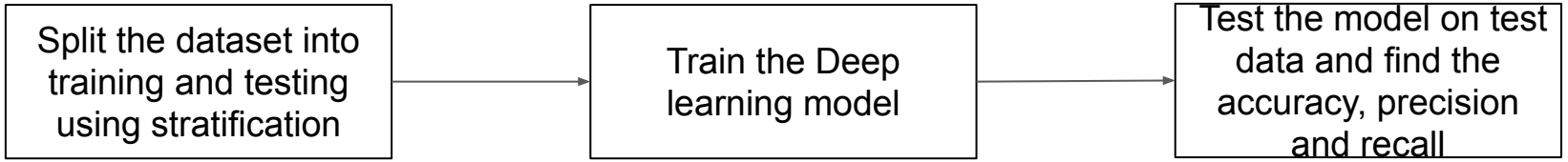
- Restricted to subtypes of an organ
- LSTM may not capture the underlying patterns for classification of different cancer types
- Hand picked features

Implementation

Data Pipeline



Model Training pipeline



Contd : 1/10

Data Exploration

Contd : 2/10

MiRNA Quantification file

miRNA_ID	read_count	reads_per_million_miRNA_mapped	cross-mapped
hsa-let-7a-1	120721	17656.97199	N
hsa-let-7a-2	120517	17627.13441	N
hsa-let-7a-3	120674	17650.09765	N
hsa-let-7b	311736	45595.32989	N
hsa-let-7c	6153	899.954015	N
hsa-let-7d	2808	410.705489	N
hsa-let-7e	15507	2268.094736	N
hsa-let-7f-1	45136	6601.710454	N
hsa-let-7f-2	46155	6750.752082	N
hsa-let-7g	3314	484.714384	N
hsa-let-7i	5154	753.837639	N
hsa-mir-1-1	8	1.170101	N
hsa-mir-1-2	8	1.170101	N
hsa-mir-100	46034	6733.054303	Y
hsa-mir-101-1	32044	4686.83999	N
hsa-mir-101-2	32730	4787.17616	N
hsa-mir-103a-1	28548	4175.505805	Y
hsa-mir-103a-2	28598	4182.818937	Y
hsa-mir-103b-1	0	0	N
hsa-mir-103b-2	0	0	N
hsa-mir-105-1	0	0	N
hsa-mir-105-2	0	0	N
hsa-mir-106a	20	2.925253	Y
hsa-mir-106b	3336	487.932162	N
hsa-mir-107	472	69.035965	Y
hsa-mir-10a	150118	21956.65478	N
hsa-mir-10b	12052	1762.75732	N
hsa-mir-1178	0	0	N
hsa-mir-1179	0	0	N
hsa-mir-1180	36	5.265455	N
hsa-mir-1181	5	0.731313	N

Contd : 3/10

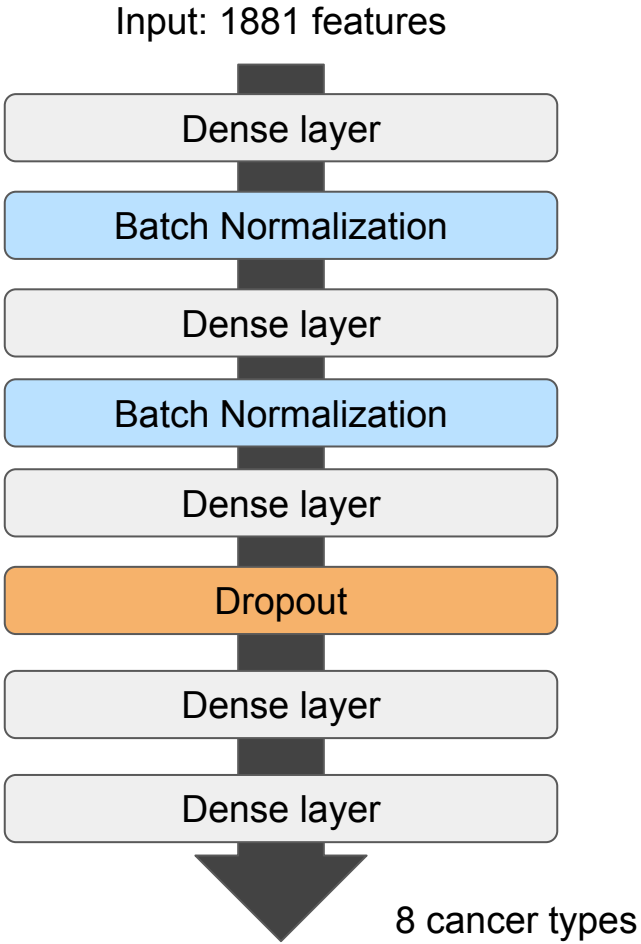
Summary of Data

Train samples	1749
Test samples	583

Cancer Types	Number of samples
Acute myeloid leukemia	301
Infiltrating duct carcinoma	301
Clear cell adenocarcinoma	301
Serous cystadenocarcinoma	301
Malignant melanoma	301
Papillary adenocarcinoma,Thyroid	301
Papillary adenocarcinoma, Kidney	291
Squamous cell carcinoma, NOS,Upper lobe, lung	235
Total	2332

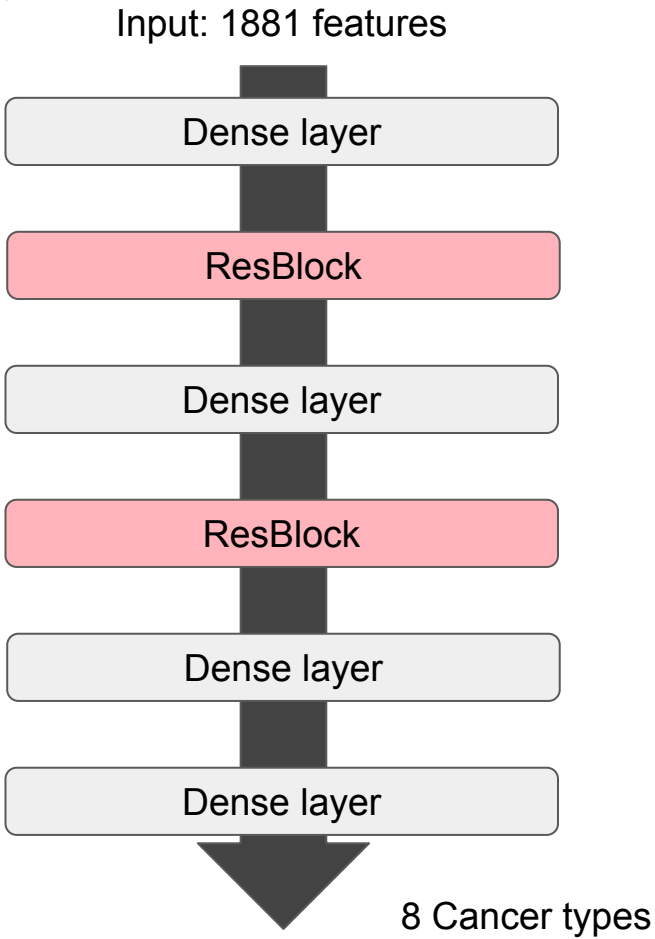
Contd : 4/10

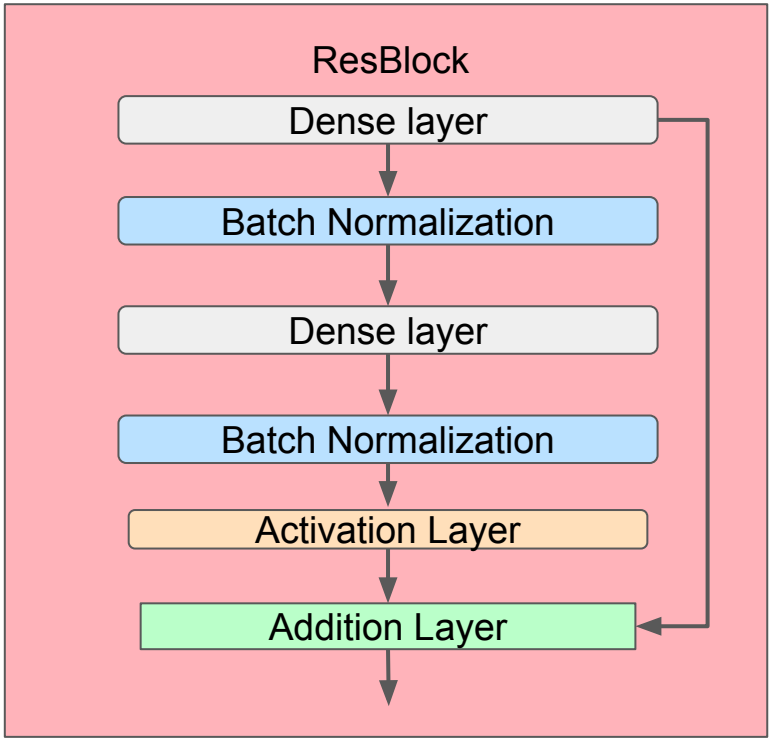
Artificial Neural Network



Contd : 5/10

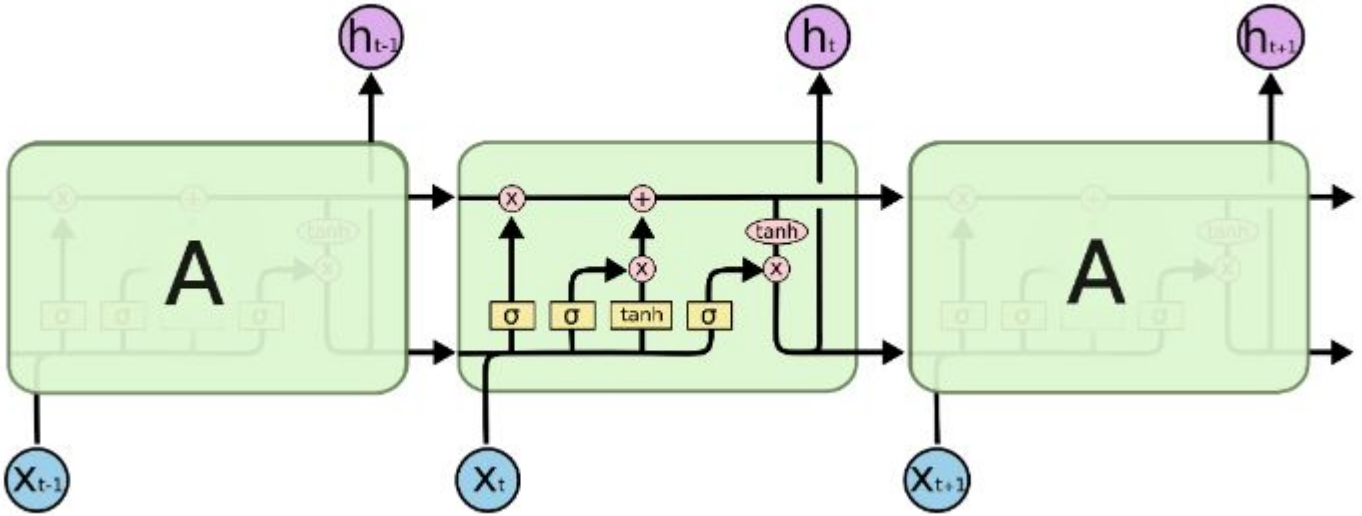
CResNet





Contd : 6/10

LSTM

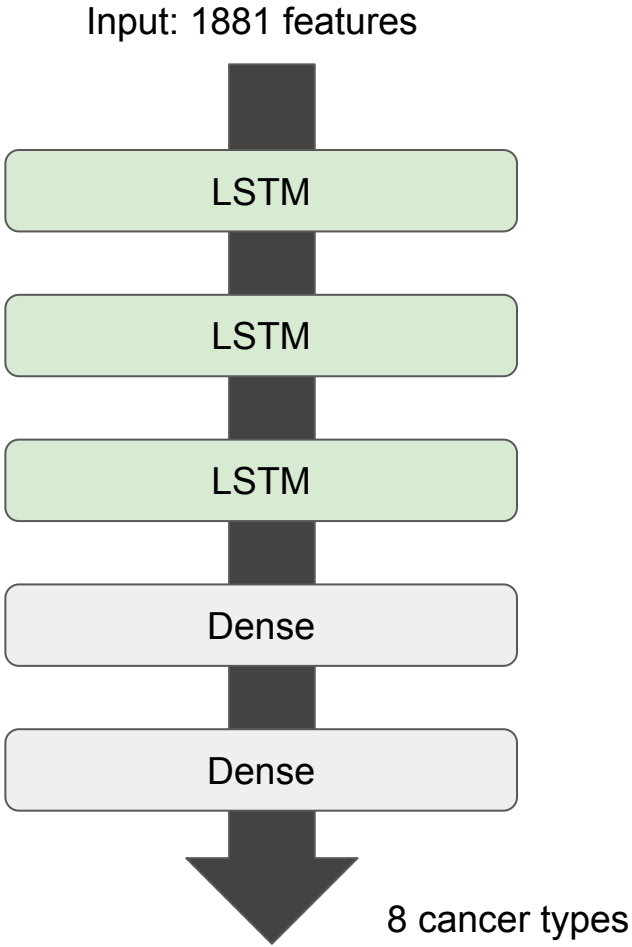


1) Forget 2) Store 3) Update 4) Output

Contd : 7/10

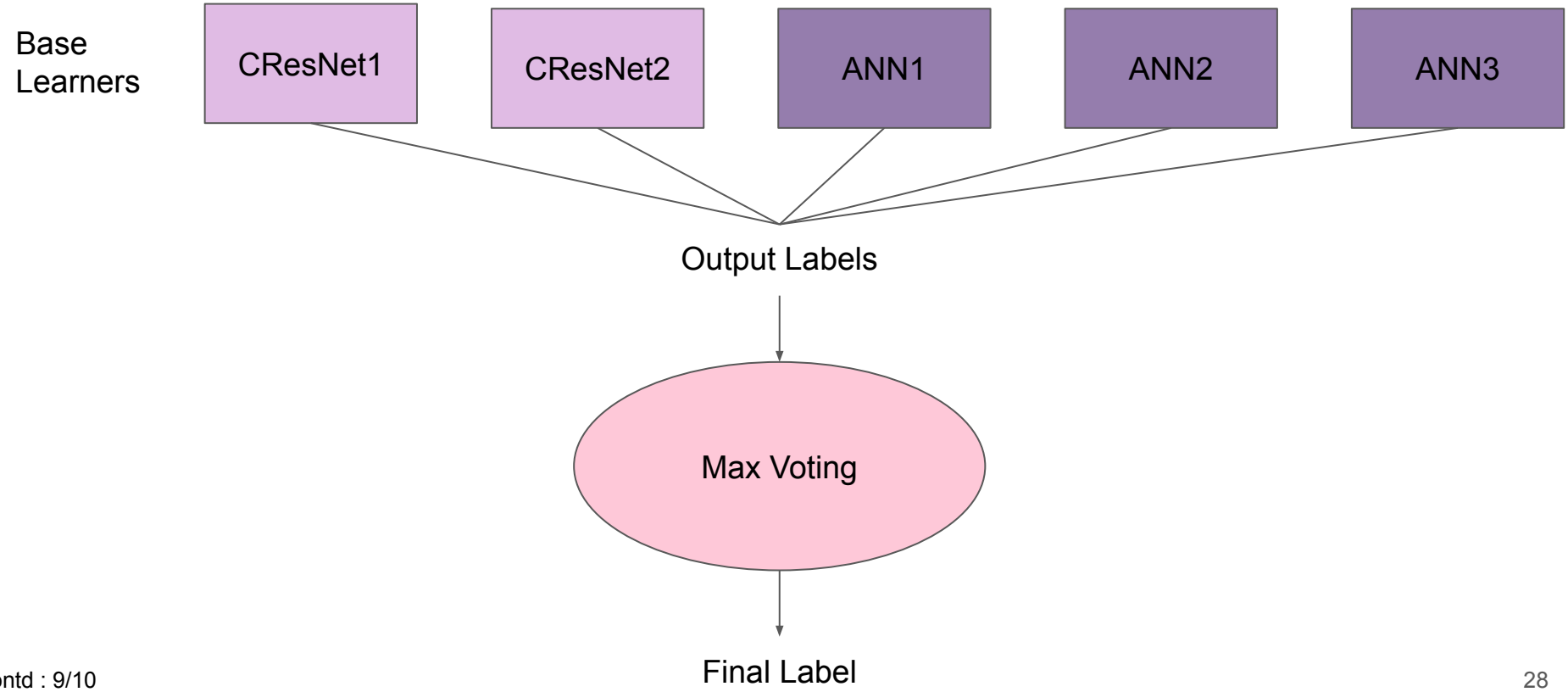
Source: colah's blog

LSTM Model architecture



Contd : 8/10

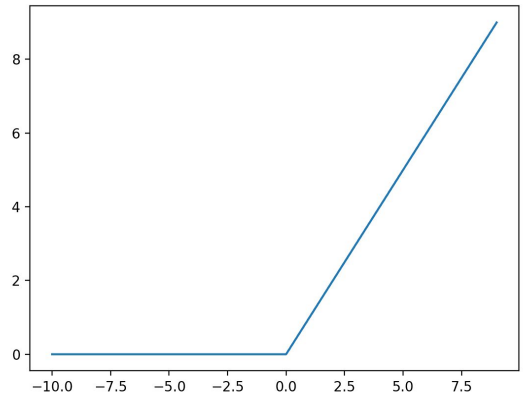
Ensemble Method



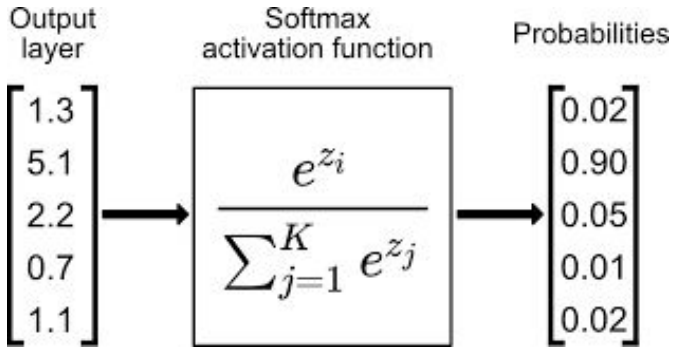
Contd : 9/10

Activation functions and Optimization

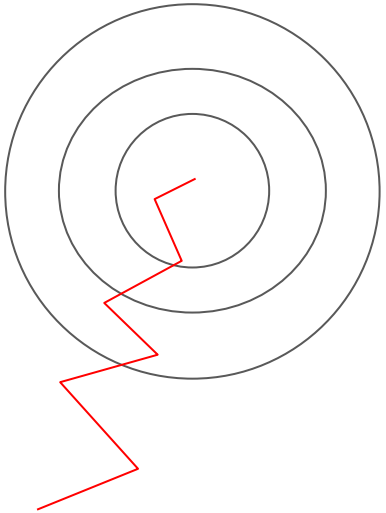
ReLu



Softmax



Adam

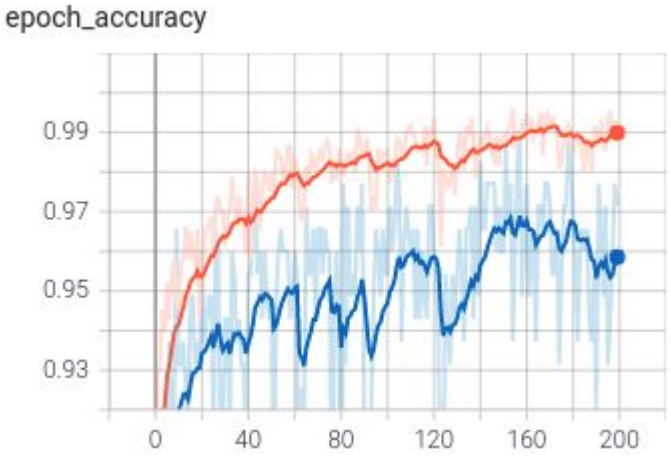


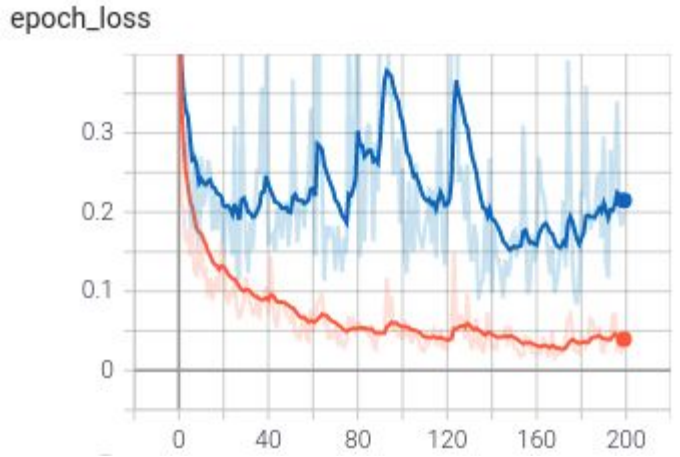
Contd : 10//10

Technical Specification

- Framework: TensorFlow
- GPU : Tesla K80 (Provided by Google Colab)
- RAM : 13 GB
- Epochs : 200
- Learning rate : 0.01
- Tensorboard : Accuracy and Loss

Accuracy graph and Loss graph of ANN

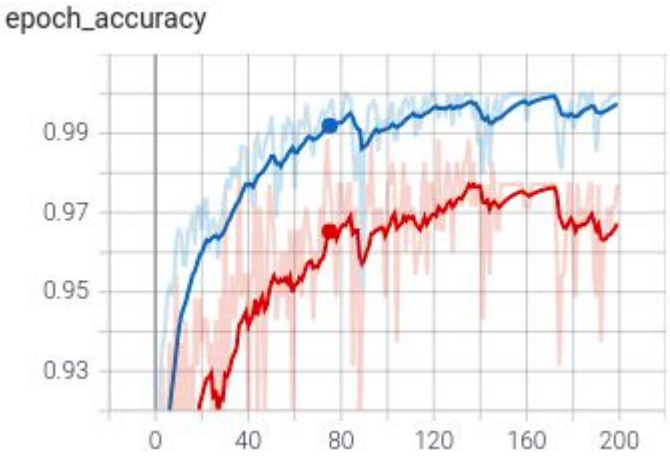


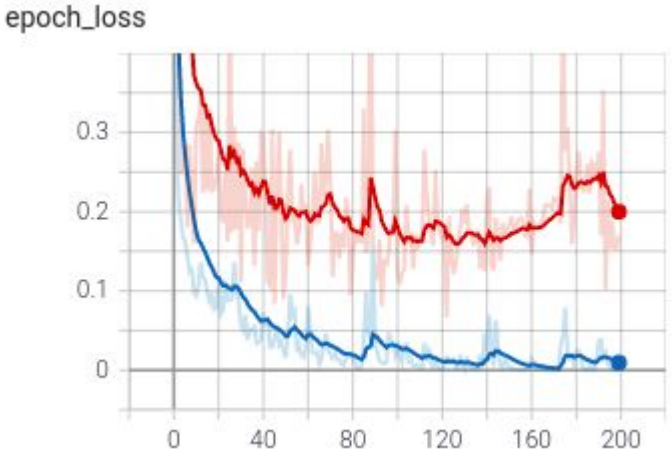


Training

Validation

Accuracy and Loss graph of CResNet



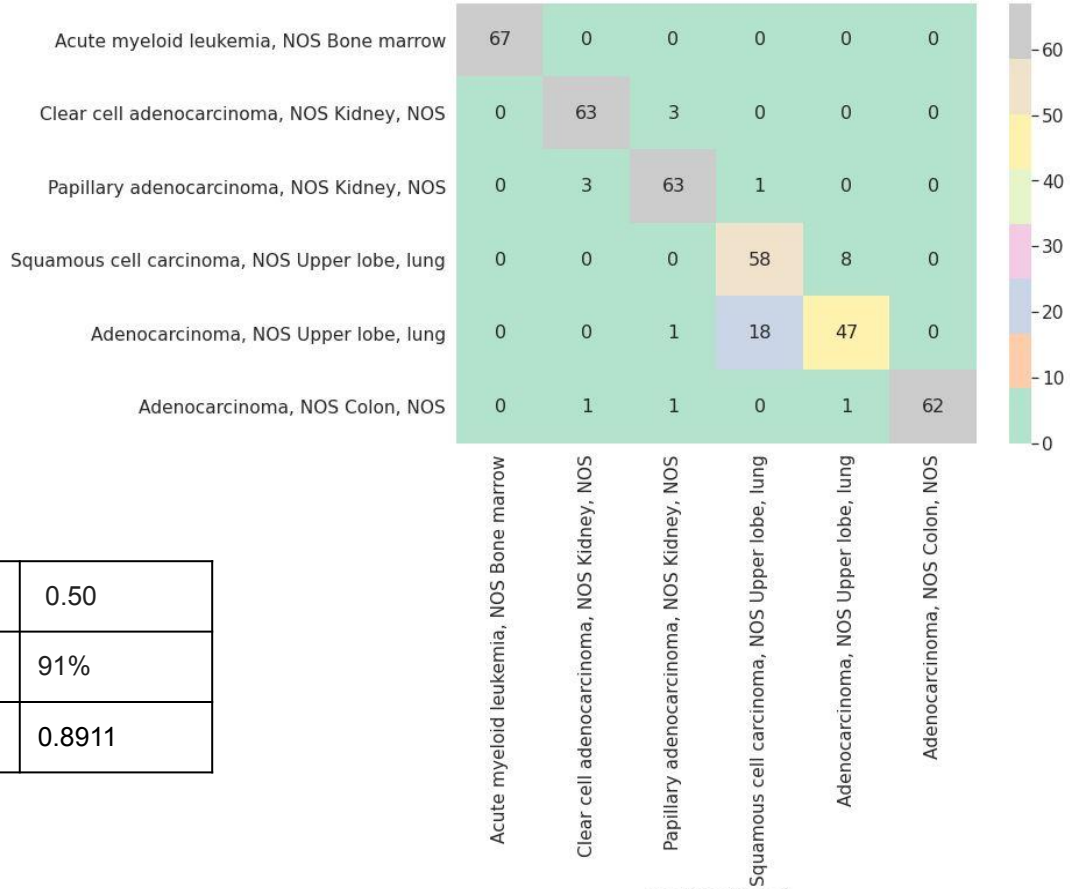


Training

Validation

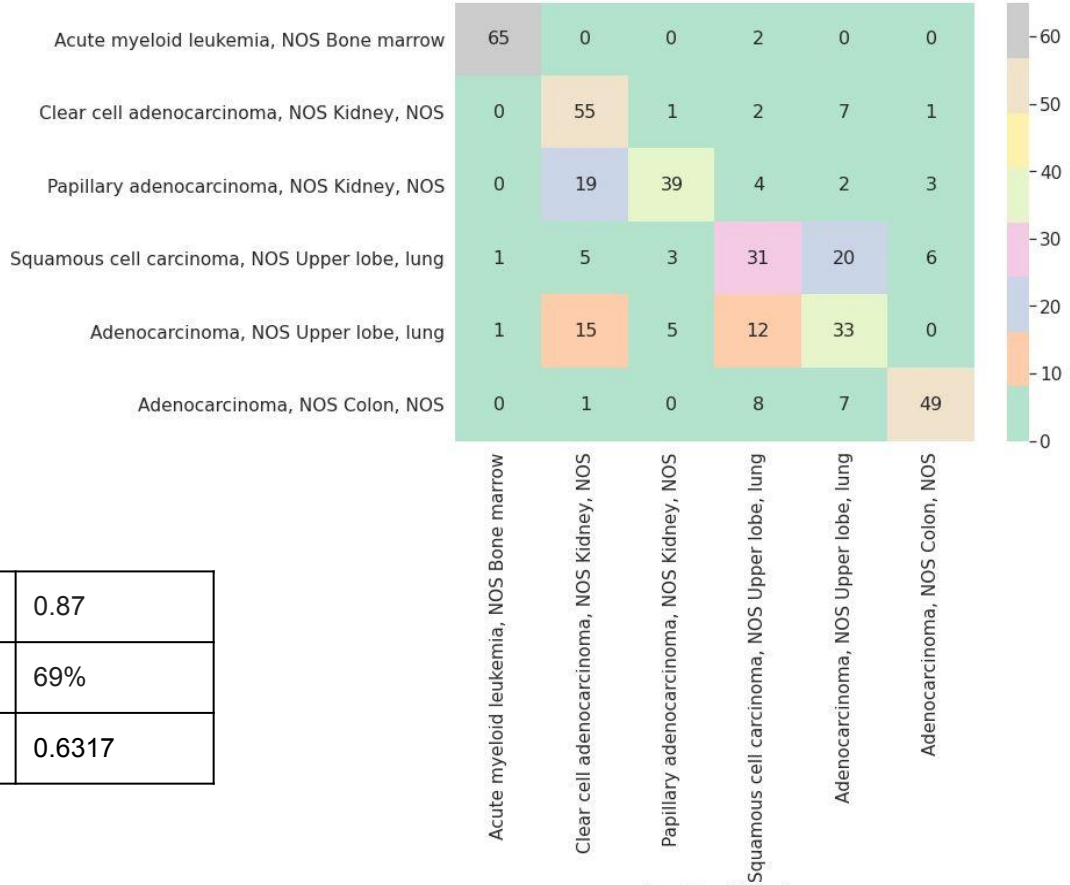
Previously trained models

ANN



Loss	0.50
Accuracy:	91%
MCC	0.8911

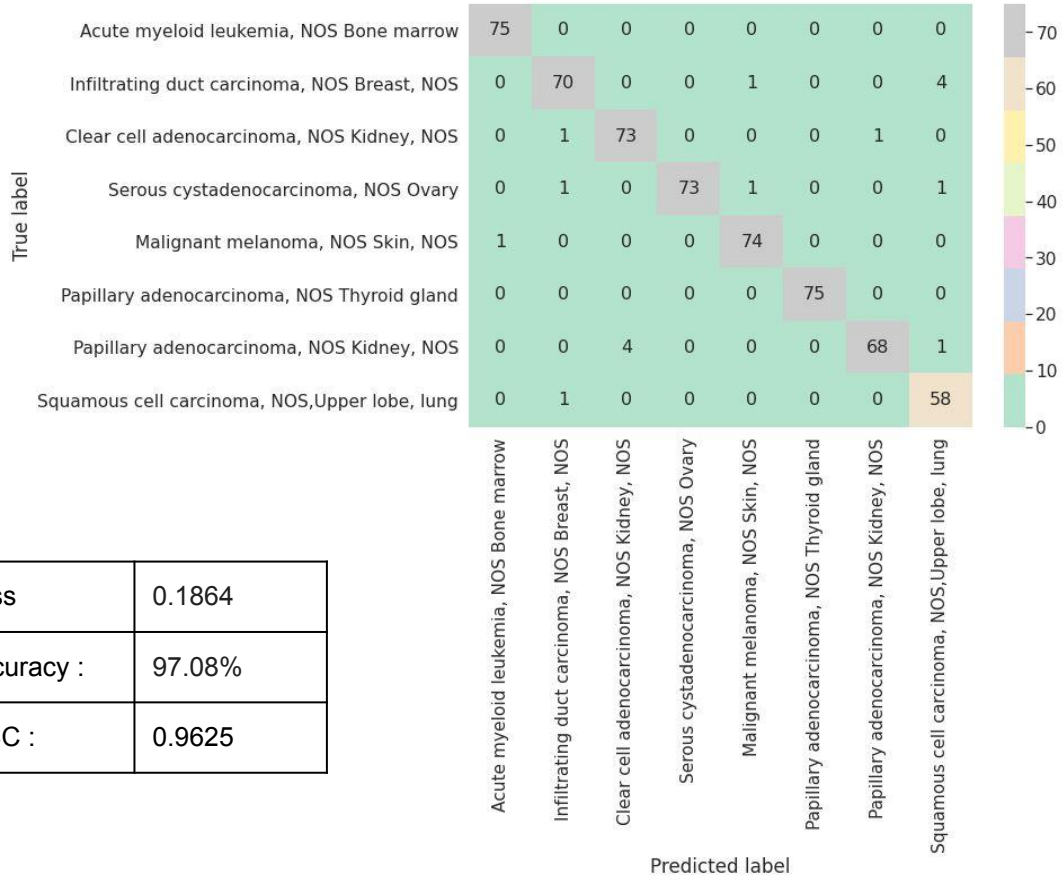
LSTM



Loss	0.87
Accuracy	69%
MCC	0.6317

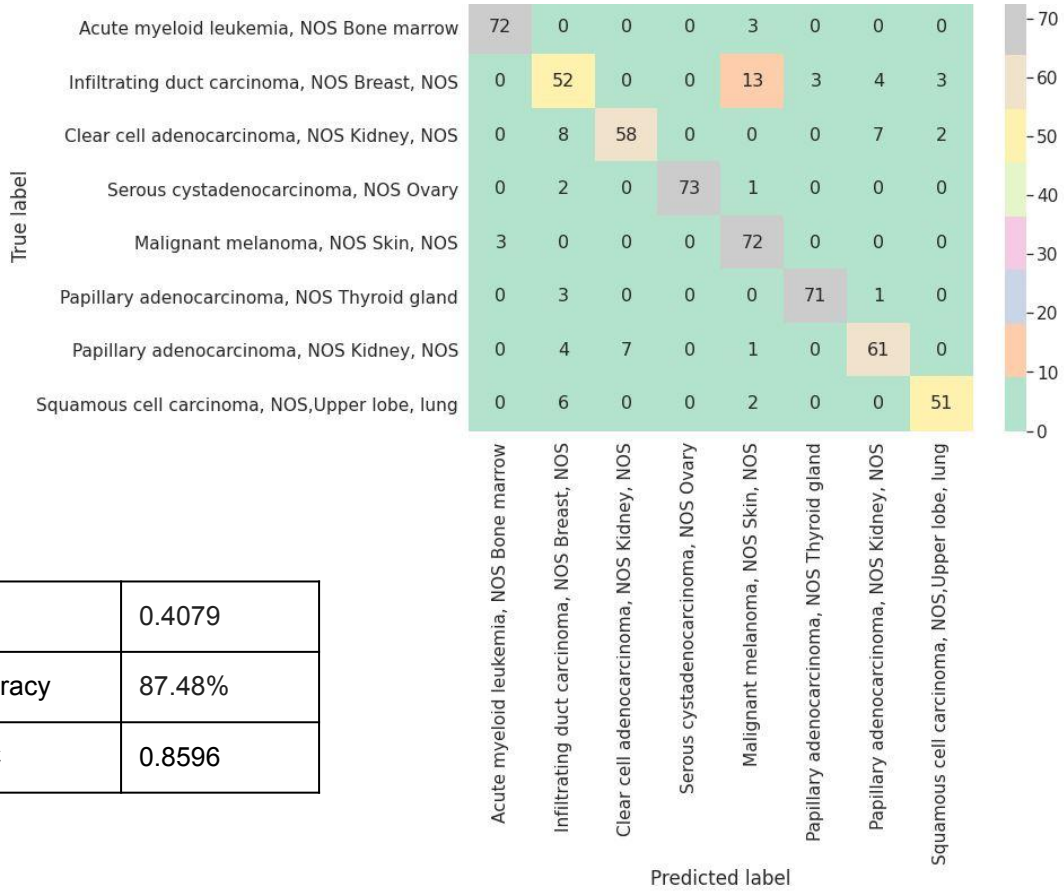
Result Analysis

ANN



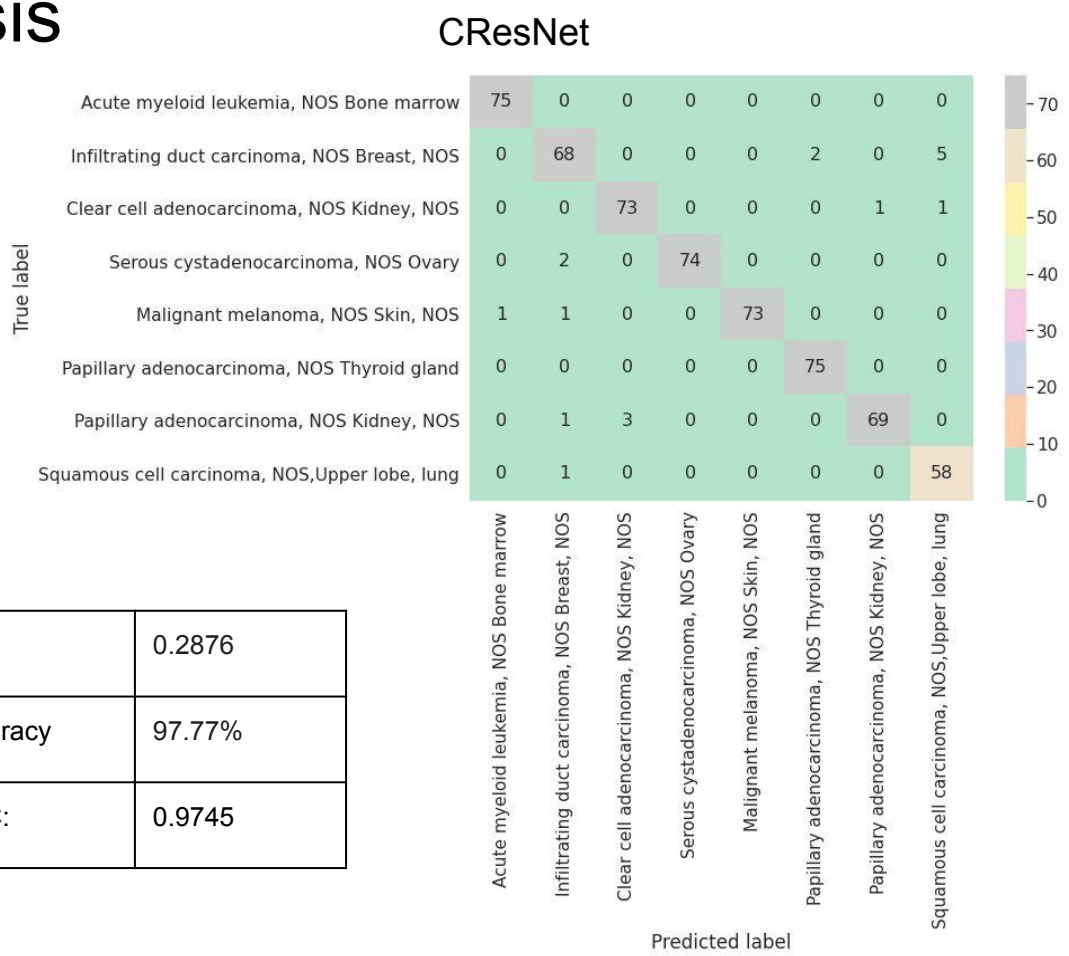
Loss	0.1864
Accuracy :	97.08%
MCC :	0.9625

LSTM



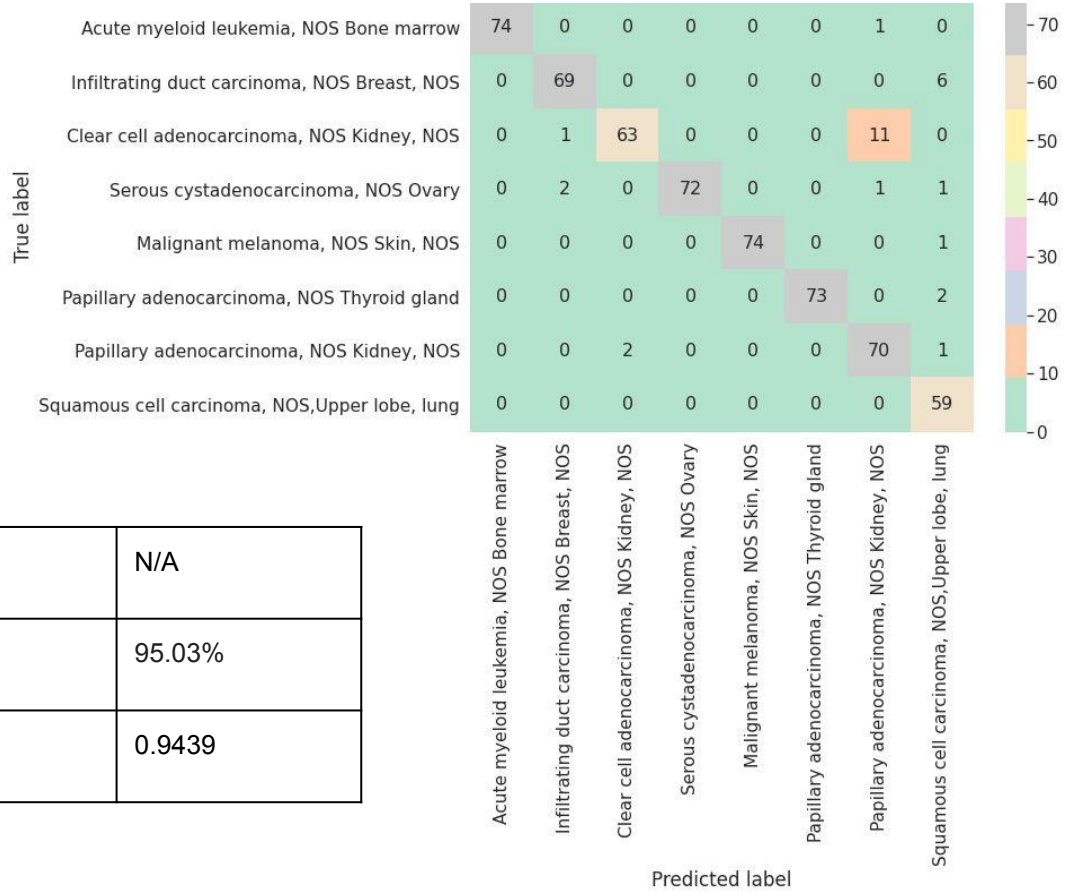
Loss	0.4079
Accuracy	87.48%
MCC	0.8596

Result Analysis contd.



Loss	0.2876
Accuracy	97.77%
MCC:	0.9745

Ensemble Model



Loss	N/A
Accuracy	95.03%
MCC	0.9439

MCC for the individual models

	ANN				
	True Positive	True Negative	False Positive	False Negative	MCC (Matthew's correlation coefficient)
Acute myeloid leukemia, NOS Bone marrow	75	507	1	0	0.9924
Infiltrating duct carcinoma, NOS Breast, NOS	70	505	3	5	0.9084
Clear cell adenocarcinoma, NOS Kidney, NOS	73	505	4	2	0.9547
Serous cystadenocarcinoma, NOS Ovary	73	508	0	3	0.9772
Malignant melanoma, NOS Skin, NOS	74	509	2	1	0.9772
Papillary adenocarcinoma, NOS Thyroid gland	75	508	0	0	1
Papillary adenocarcinoma, NOS Kidney, NOS	68	514	1	5	0.9524
Squamous cell carcinoma, NOS Upper lobe, lung	58	525	6	1	0.9374
				Average	0.9624625

	LSTM				
	True Positive	True Negative	False Positive	False Negative	MCC (Matthew's correlation coefficient)
Acute myeloid leukemia, NOS Bone marrow	72	505	3	3	0.9541
Infiltrating duct carcinoma, NOS Breast, NOS	52	485	23	23	0.6481
Clear cell adenocarcinoma, NOS Kidney, NOS	58	509	7	17	0.8083
Serous cystadenocarcinoma, NOS Ovary	73	509	0	3	0.9772
Malignant melanoma, NOS Skin, NOS	72	508	20	3	0.8463
Papillary adenocarcinoma, NOS Thyroid gland	71	511	3	4	0.9462
Papillary adenocarcinoma, NOS Kidney, NOS	61	522	12	12	0.8131
Squamous cell carcinoma, NOS Upper lobe, lung	51	532	5	7	0.8837
				Average	0.859625

	Ensemble				
	True Positive	True Negative	False Positive	False Negative	MCC (Matthew's correlation coefficient)
Acute myeloid leukemia, NOS Bone marrow	74	508	0	1	0.9923
Infiltrating duct carcinoma, NOS Breast, NOS	69	506	3	6	0.9302
Clear cell adenocarcinoma, NOS Kidney, NOS	63	508	2	12	0.8894
Serous cystadenocarcinoma, NOS Ovary	72	510	0	4	0.9695
Malignant melanoma, NOS Skin, NOS	74	509	0	1	0.9923
Papillary adenocarcinoma, NOS Thyroid gland	73	509	0	2	0.9846
Papillary adenocarcinoma, NOS Kidney, NOS	70	513	13	3	0.8846
Squamous cell carcinoma, NOS Upper lobe, lung	59	525	11	0	0.9086
				Average	0.9439375

	CResNet				
	True Positive	True Negative	False Positive	False Negative	MCC (Matthew's correlation coefficient)
Acute myeloid leukemia, NOS Bone marrow	75	507	1	0	0.9924
Infiltrating duct carcinoma, NOS Breast, NOS	71	505	3	4	0.9462
Clear cell adenocarcinoma, NOS Kidney, NOS	73	506	2	2	0.9694
Serous cystadenocarcinoma, NOS Ovary	74	506	1	2	0.9772
Malignant melanoma, NOS Skin, NOS	73	507	1	2	0.9769
Papillary adenocarcinoma, NOS Thyroid gland	75	507	1	0	0.9924
Papillary adenocarcinoma, NOS Kidney, NOS	70	510	1	3	0.9684
Squamous cell carcinoma, NOS Upper lobe, lung	59	521	3	0	0.9727
				Average	0.97445

Conclusion

- Our CResNet model achieved MCC of 0.9745 which is greater than the SOTA LSTM 0.949
- Possible improvements explainability between this 1881 features and the cancer type and subtypes

Thank you

References

1. Peng, Y., Croce, C. The role of MicroRNAs in human cancer. *Sig Transduct Target Ther* 1, 15004 (2016). https://doi.org/10.1038/sigtrans.2015.4
2. Shapcott Mary, Hewitt Katherine J., Rajpoot Nasir Deep Learning With Sampling in Colon Cancer Histology Frontiers in Bioengineering and Biotechnology (2019) doi: 10.3389/fbioe.2019.00052 \
3. Lopez-Rincon, Alejandro et al. “Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection.” *BMC bioinformatics* vol. 20,1 480. 18 Sep. 2019, doi:10.1186/s12859-019-3050-8
4. Schetter, Aaron J et al. “The role of microRNAs in colorectal cancer.” *Cancer journal (Sudbury, Mass.)* vol. 18,3 (2012): 244-52. doi:10.1097/PPO.0b013e318258b78f
5. Muhamed Ali, A.; Zhuang, H.; Ibrahim, A.; Rehman, O.; Huang, M.; Wu, A. A Machine Learning Approach for the Classification of Kidney Cancer Subtypes Using miRNA Genome Data. *Appl. Sci.* **2018**, *8*, 2422.
6. Tang Binhua, Pan Zixiang, Yin Kang, Khateeb Asif; Recent Advances of Deep Learning in Bioinformatics and Computational Biology; Frontiers in Genetics (2019) doi: 10.3389/fgene.2019.00214
7. Sun, Y., Zhu, S., Ma, K. *et al.* Identification of 12 cancer types through genome deep learning. *Sci Rep* 9, 17256 (2019). https://doi.org/10.1038/s41598-019-53989-3