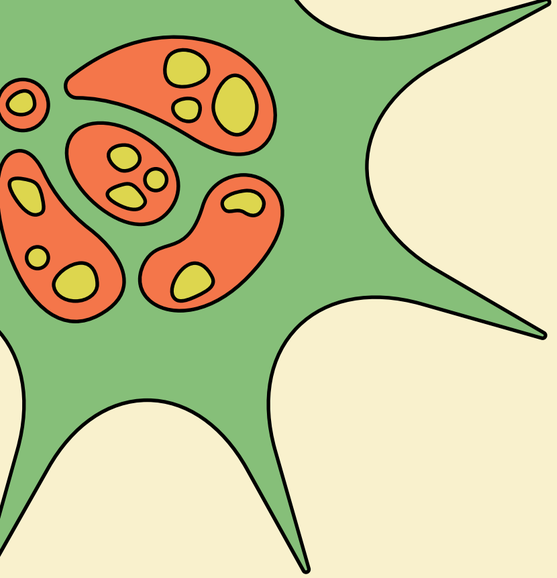
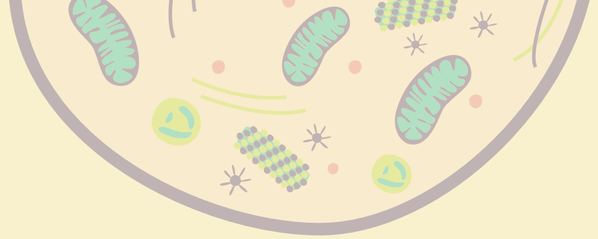
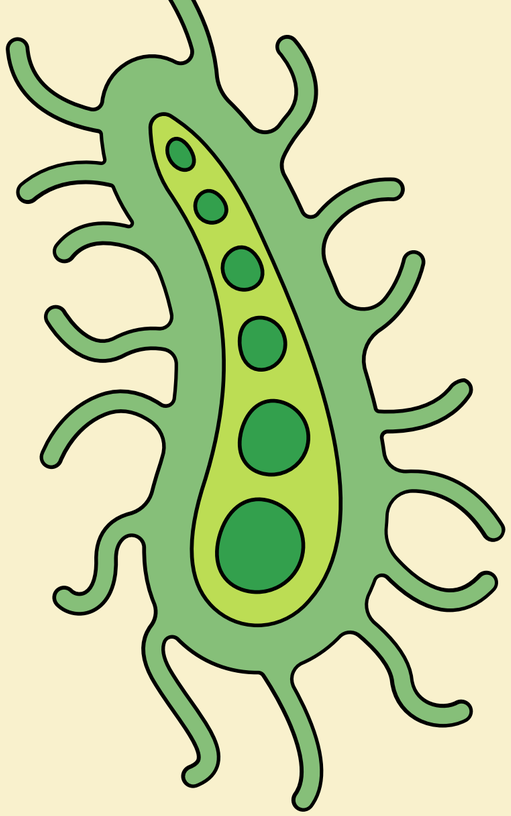


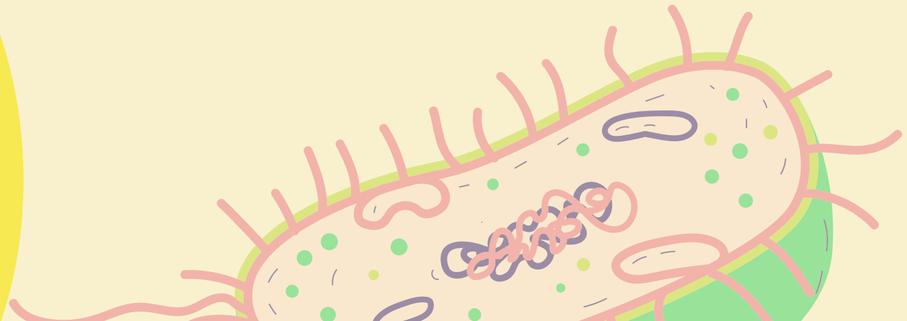




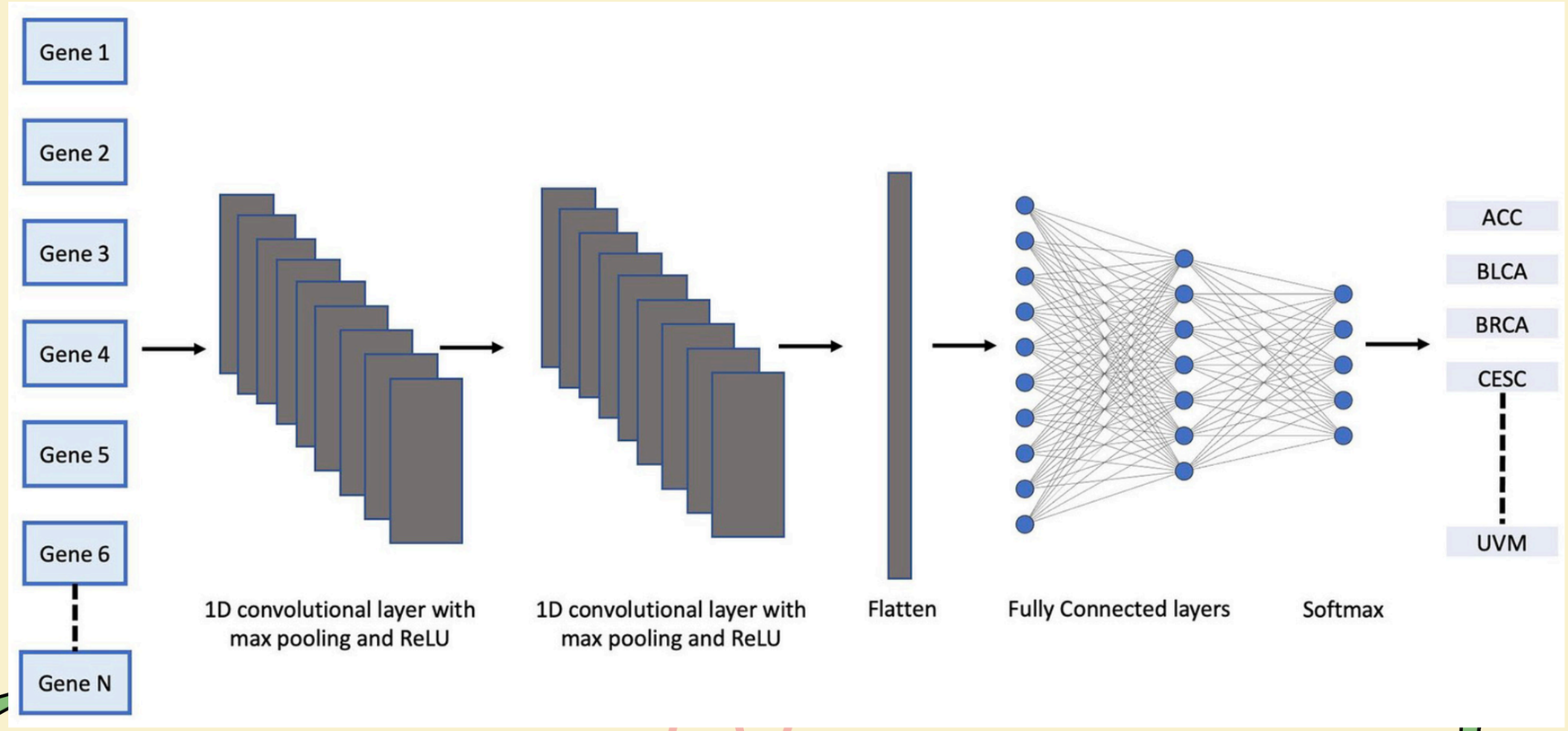
TULIP

Giorgi Kuchava

Giorgi Bendianishvili

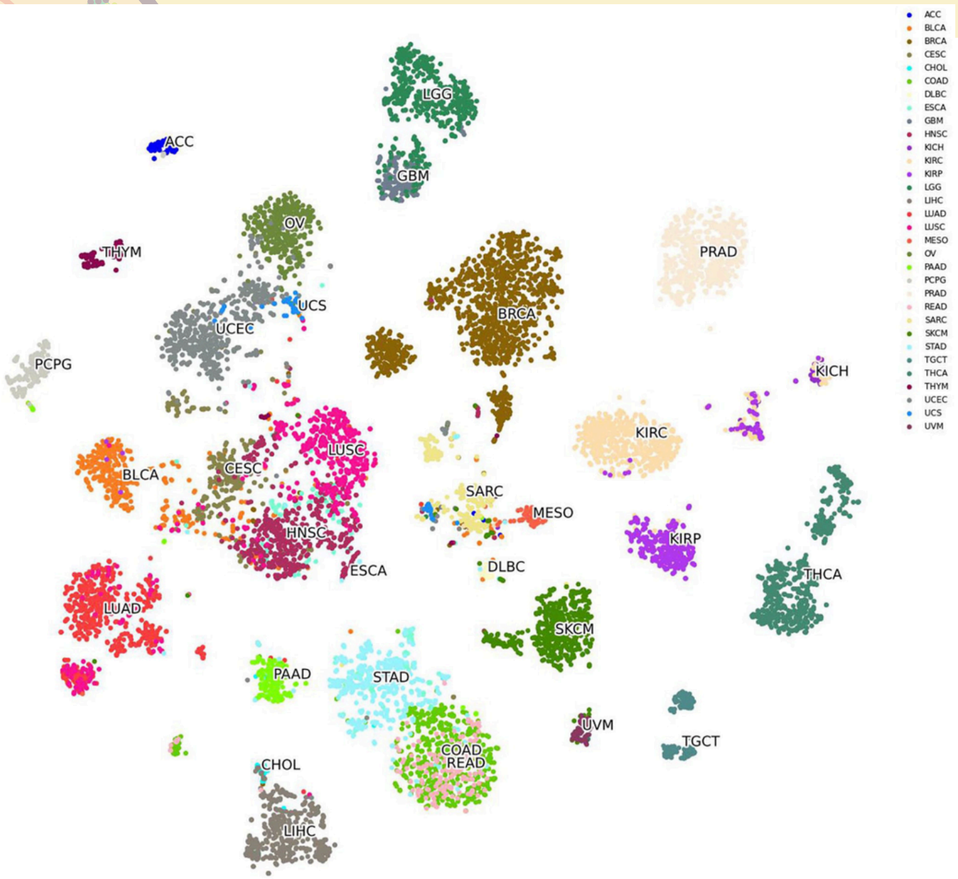


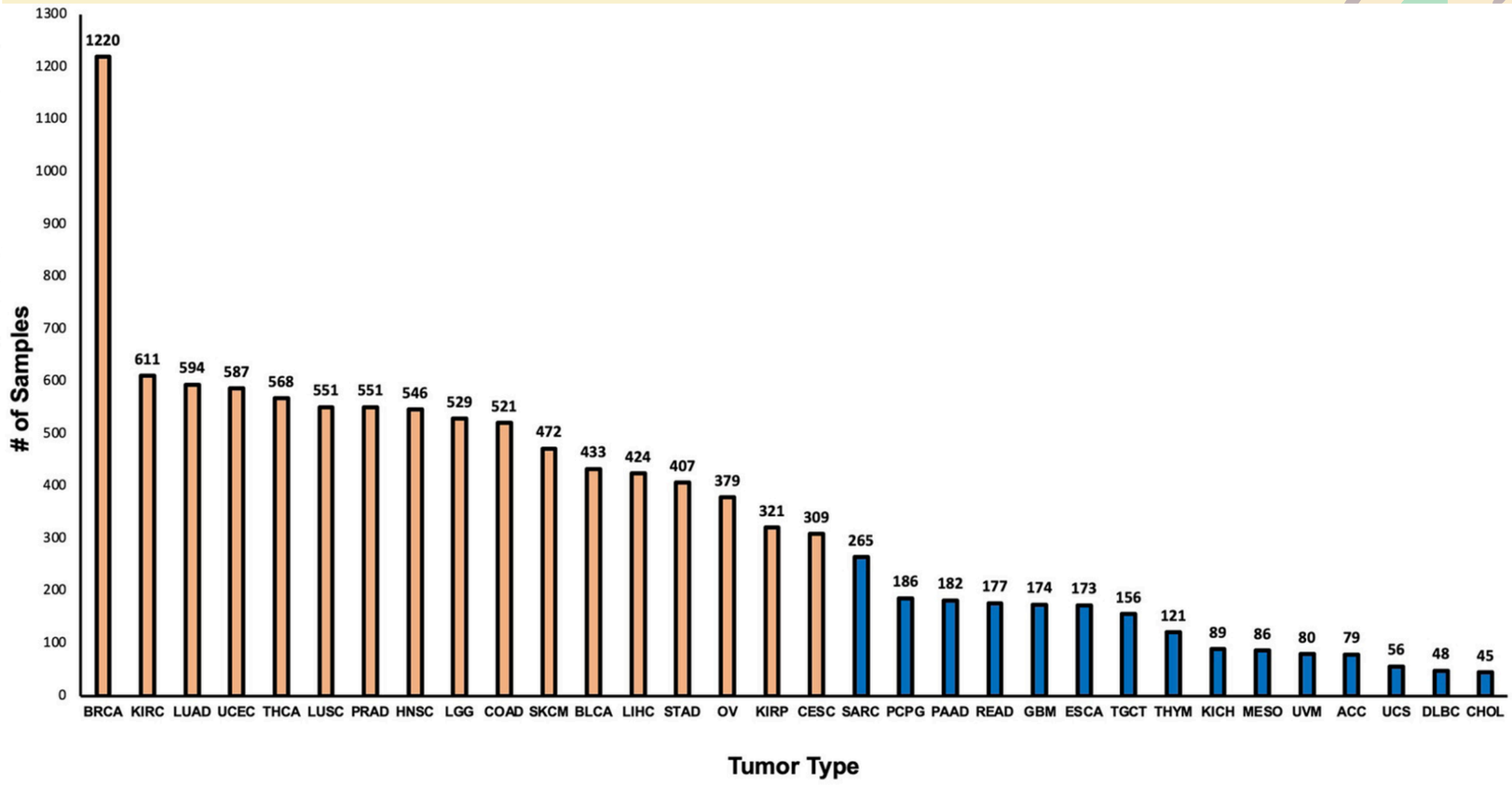
TULIP: An RNA-seq-based Primary Tumor Type Prediction Tool Using Convolutional Neural Networks



Tumor Types

19,758 protein coding genes as input and 32 primary tumor types

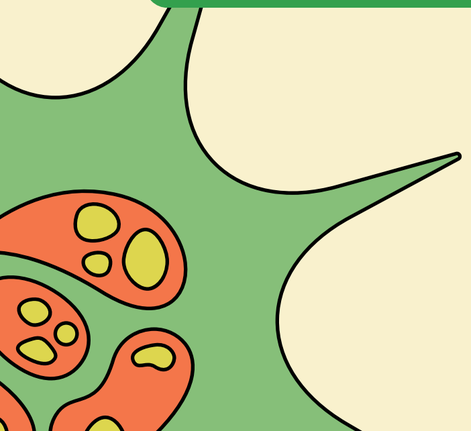


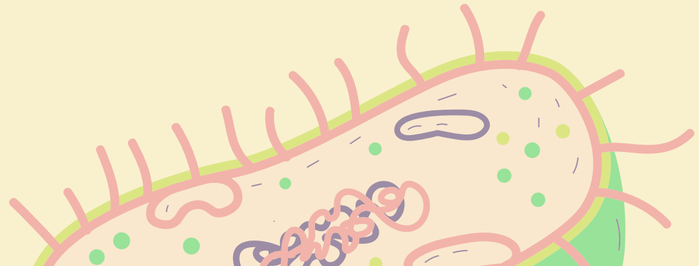


Preprocessing

Protein-coding gene list → Phenotype filtering → Expression matrix loading → Sample alignment

Tumor type filtering (32 classes) → Normalization (TPM + log) → Padding → Train / Val / Test split



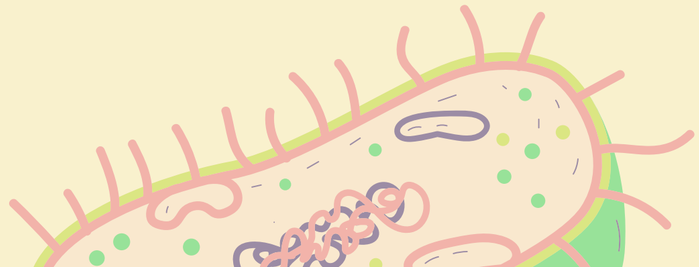


Raw Data Structure

Gene ID	Sample1	Sample2	Sample3	Sample4	Sample5
ENSG1	32	51	0	73	11
ENSG2	104	82	63	22	5
ENSG3	0	0	44	55	61
ENSG4	91	13	22	0	33
ENSG5	77	88	99	11	22

Primary Tumor Filtering

Sample	Type
S1	Primary Tumor
S2	Normal
S3	Metastatic
S4	Primary Tumor
S5	Normal



Sample	Type
S1	Primary Tumor
S4	Primary Tumor

Tumor-Type Filtering

Sample	Cancer Type
S1	Breast
S2	Lung
S3	Rare Sarcoma
S4	Thyroid
S5	Unknown

Sample	Cancer Type
S1	Breast
S2	Lung
S4	Thyroid

Protein-Coding Gene Filtering(chunking)

Gene	Sample1	Sample2
ENSG1 (coding)	32	51
ENSG2 (non-coding)	104	82
ENSG3 (coding)	0	44
ENSG4 (lncRNA)	91	13
ENSG5 (coding)	77	88

Gene	Sample1	Sample2
ENSG1	32	51
ENSG3	0	44
ENSG5	77	88

FPKM → TPM → log10

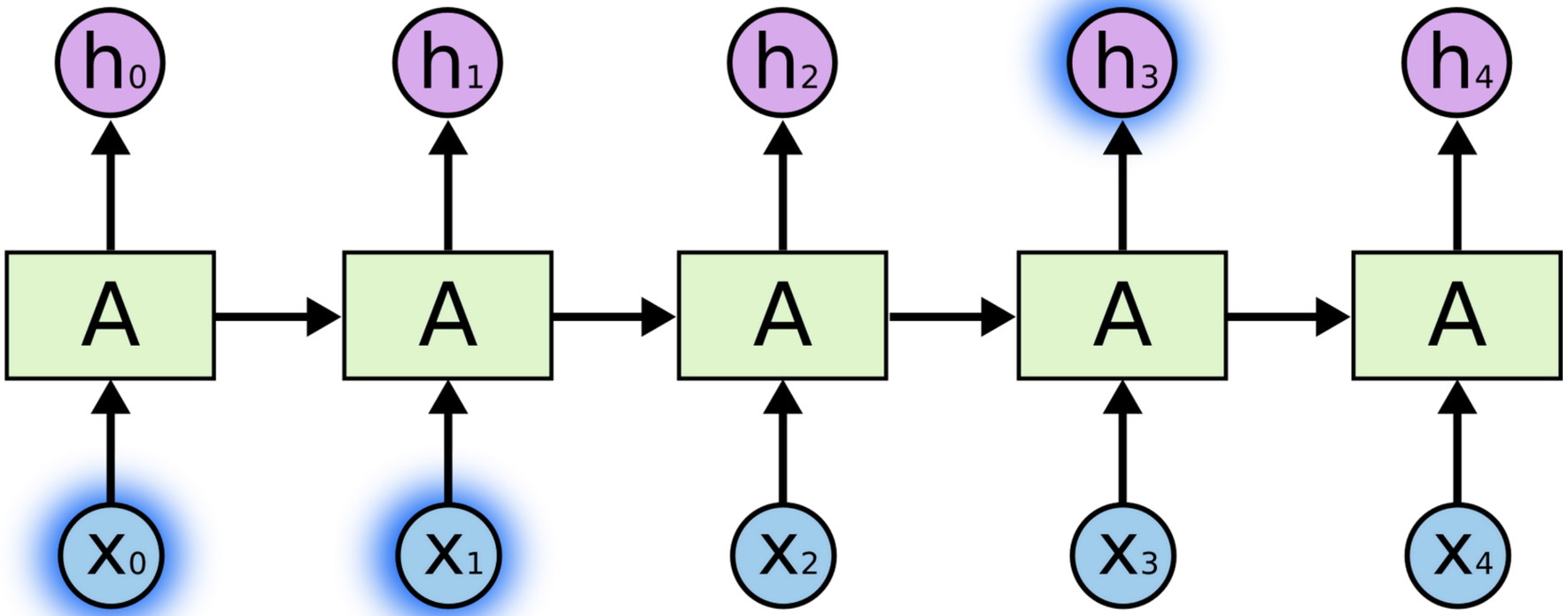
Gene	Value	Gene	TPM	Gene	log10(TPM)
G1	50	G1	500 000	G1	5,7
G2	30	G2	300 000	G2	5,48
G3	20	G3	200 000	G3	5,3

Train/Val/Test split

80% Train
10% Validation
10% Test

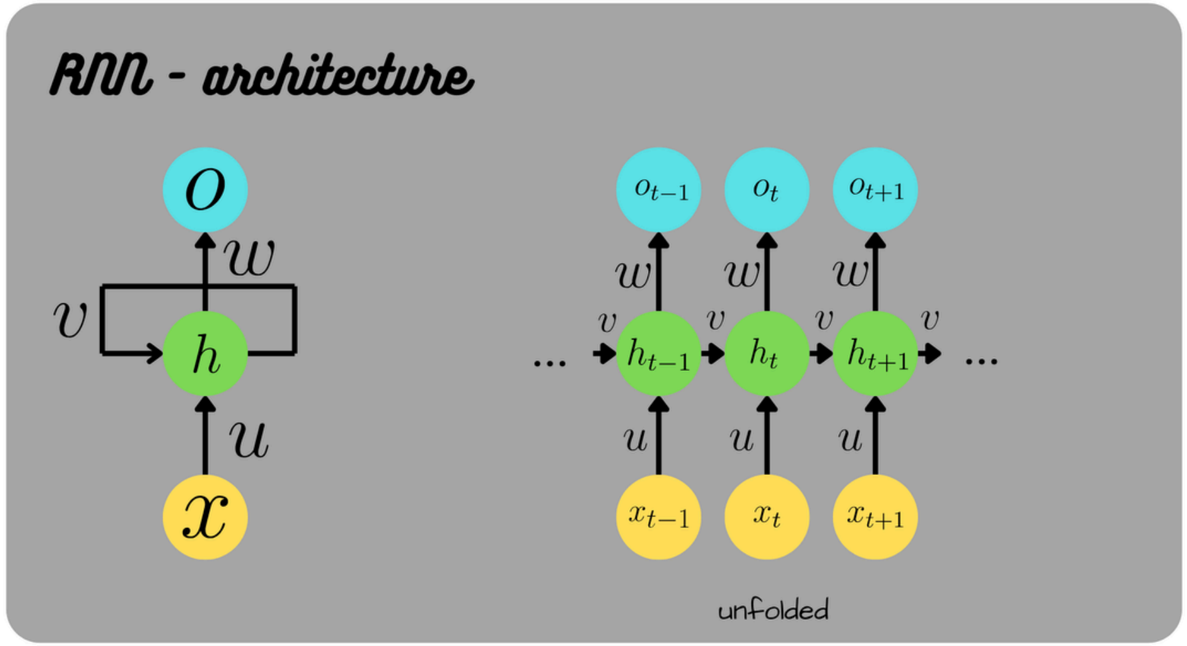
LSTM

Main Problem with RNN



For example: “The movie that I watched yesterday was amazing”

Problems with RNN

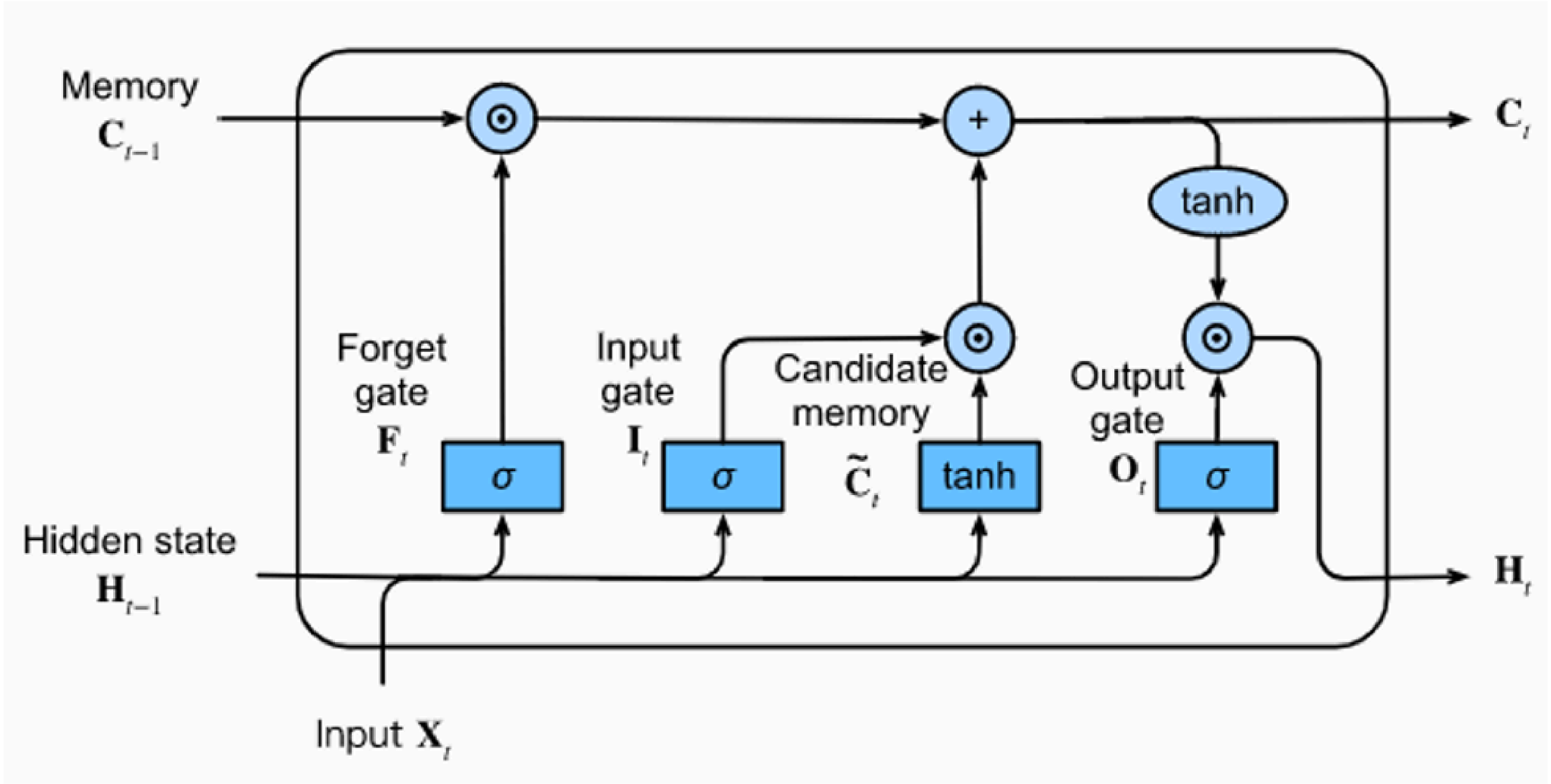


The h gets overwritten every time

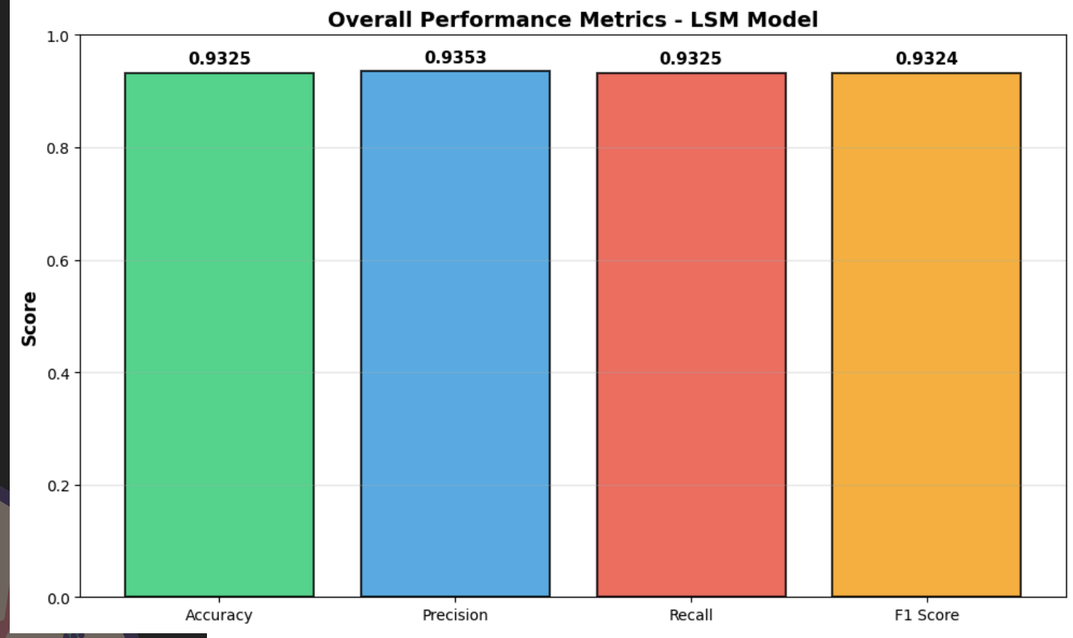
Exploding Gradients if weights > 1

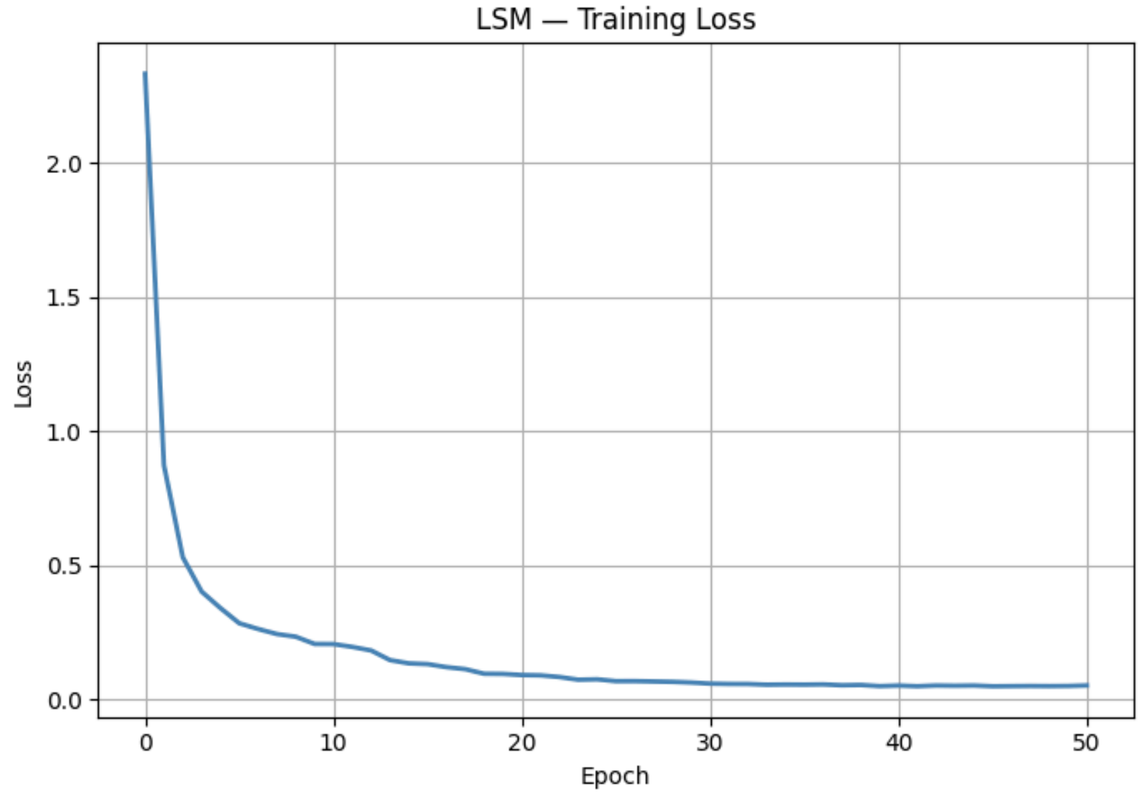
Vanishing Gradients if weights < 1

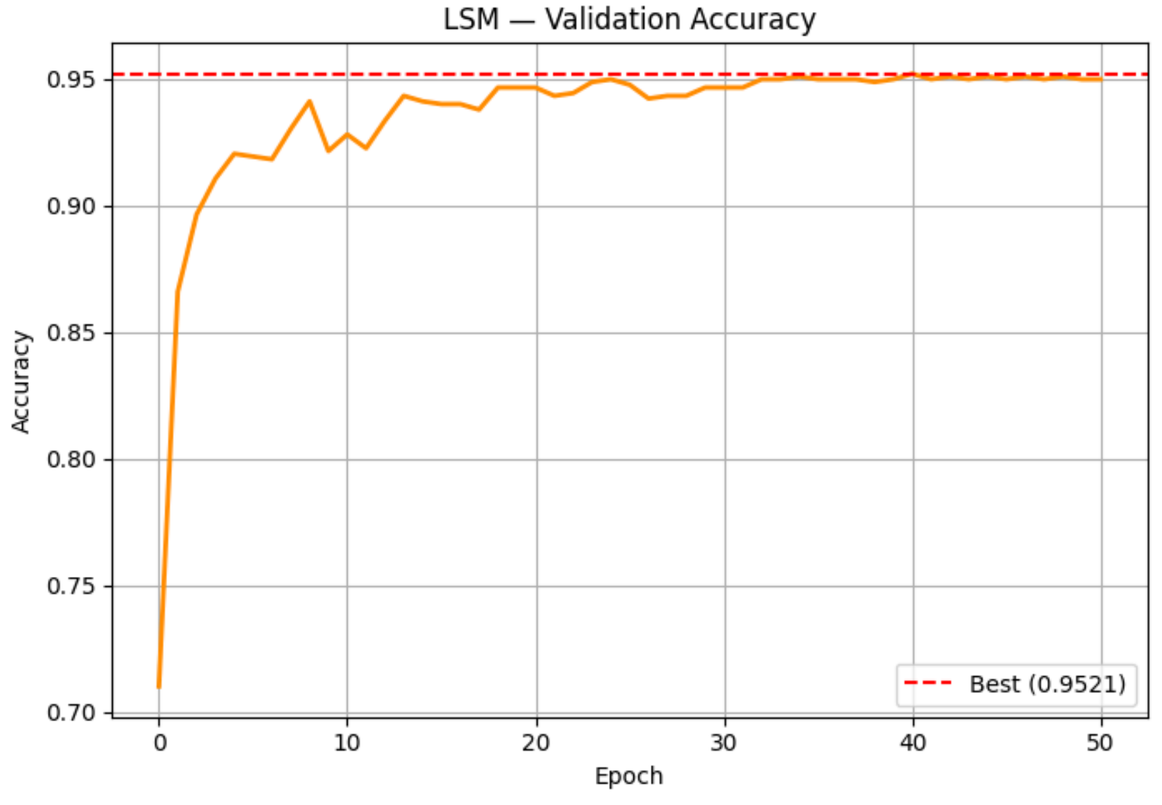
LSTM


Memory
$\mathbf{C}_{t-1}$
$\mathbf{C}_t$
tanh
Forget gate
$\mathbf{F}_t$
Input gate
$\mathbf{I}_t$
Candidate memory
$\tilde{\mathbf{C}}_t$
Output gate
$\mathbf{O}_t$
σ
σ
tanh
σ
Hidden state
$\mathbf{H}_{t-1}$
$\mathbf{H}_t$
Input $\mathbf{X}_t$

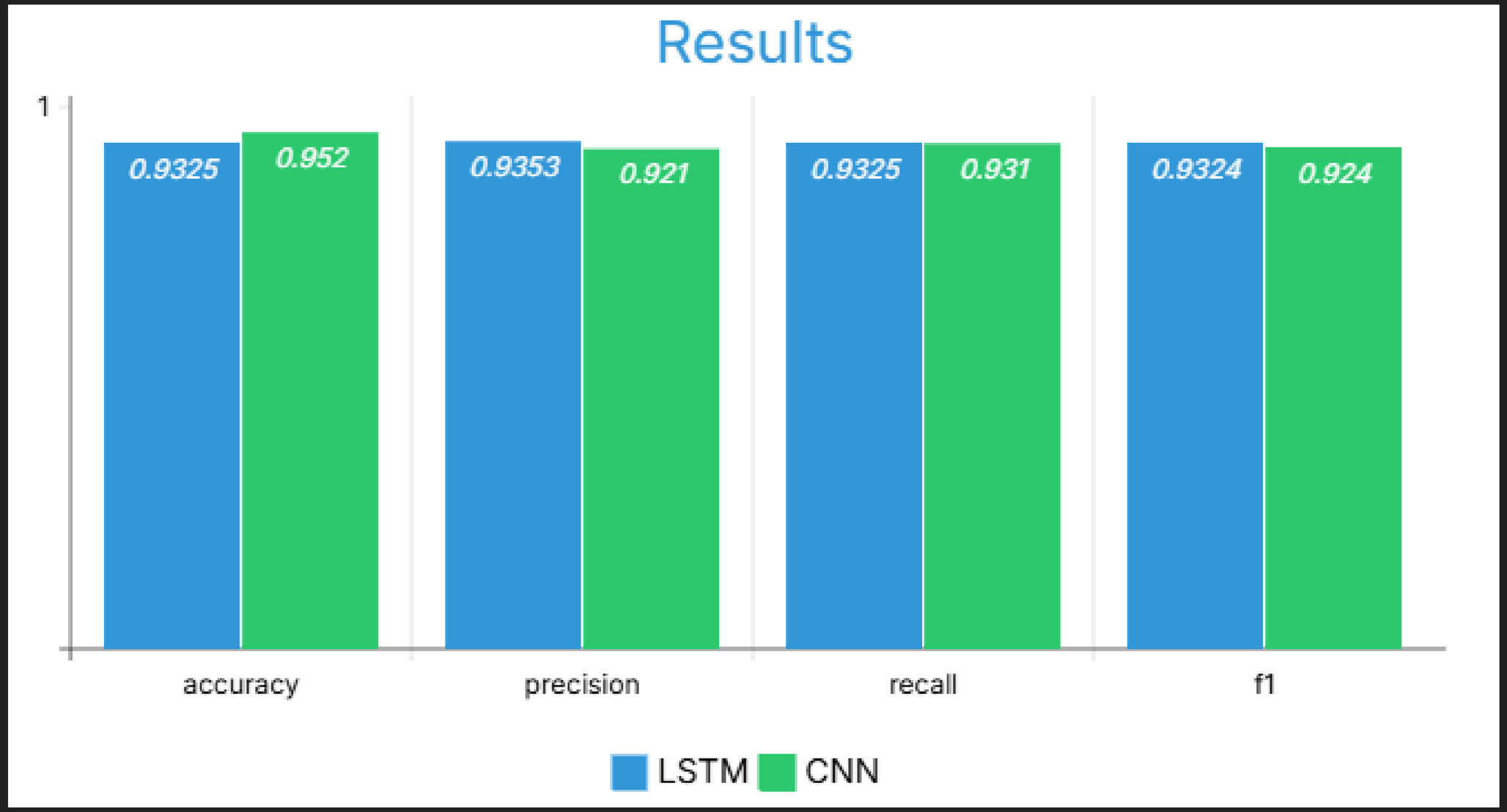
Training


Overall Performance Metrics - LSM Model
Score
1.0
0.8
0.6
0.4
0.2
0.0
0.9325
0.9353
0.9325
0.9324
Accuracy
Precision
Recall
F1 Score

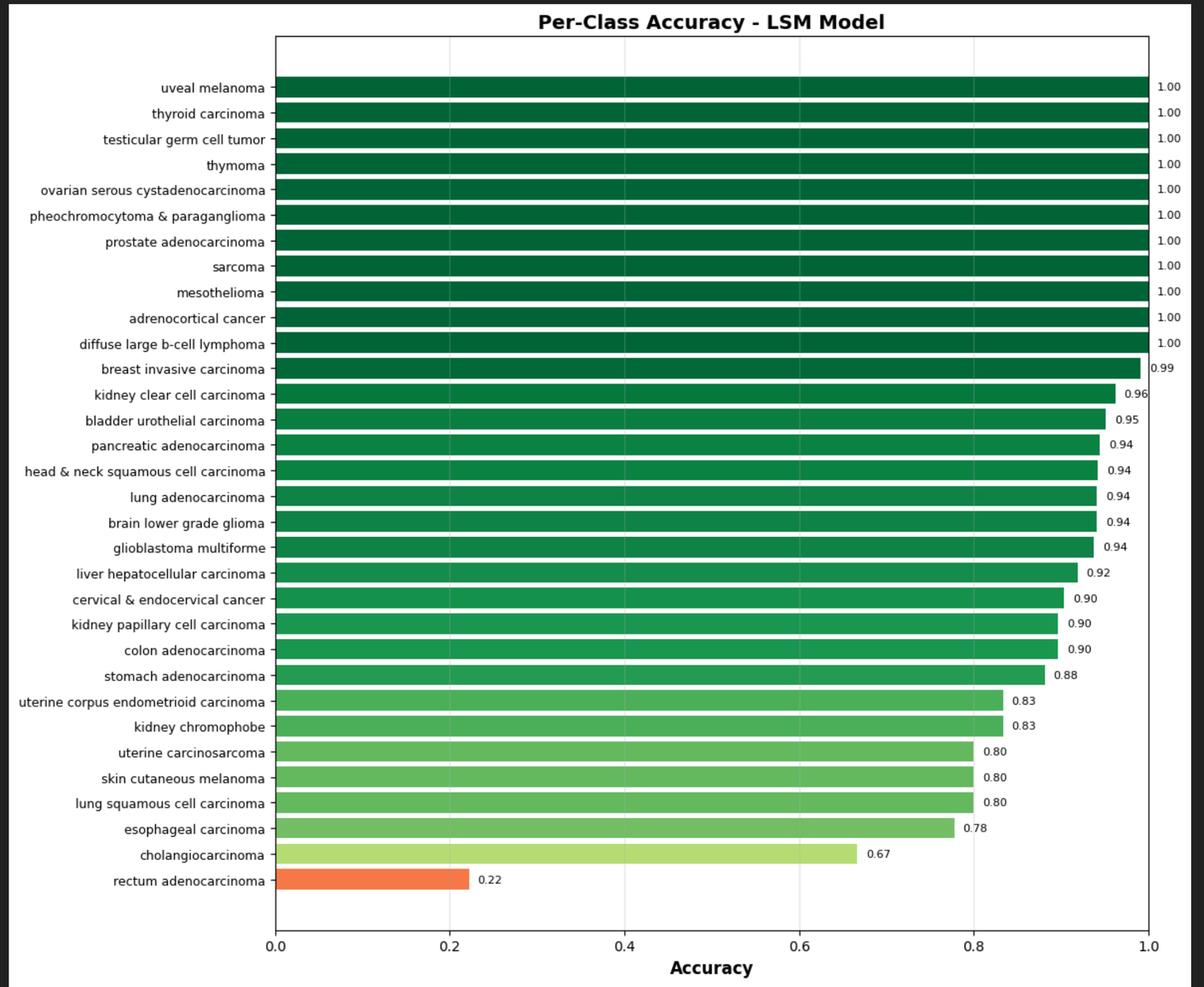

LSM — Training Loss
Loss
2.0
1.5
1.0
0.5
0.0
0
10
20
30
40
50
Epoch


LSM — Validation Accuracy
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
Best (0.9521)
0
10
20
30
40
50
Epoch

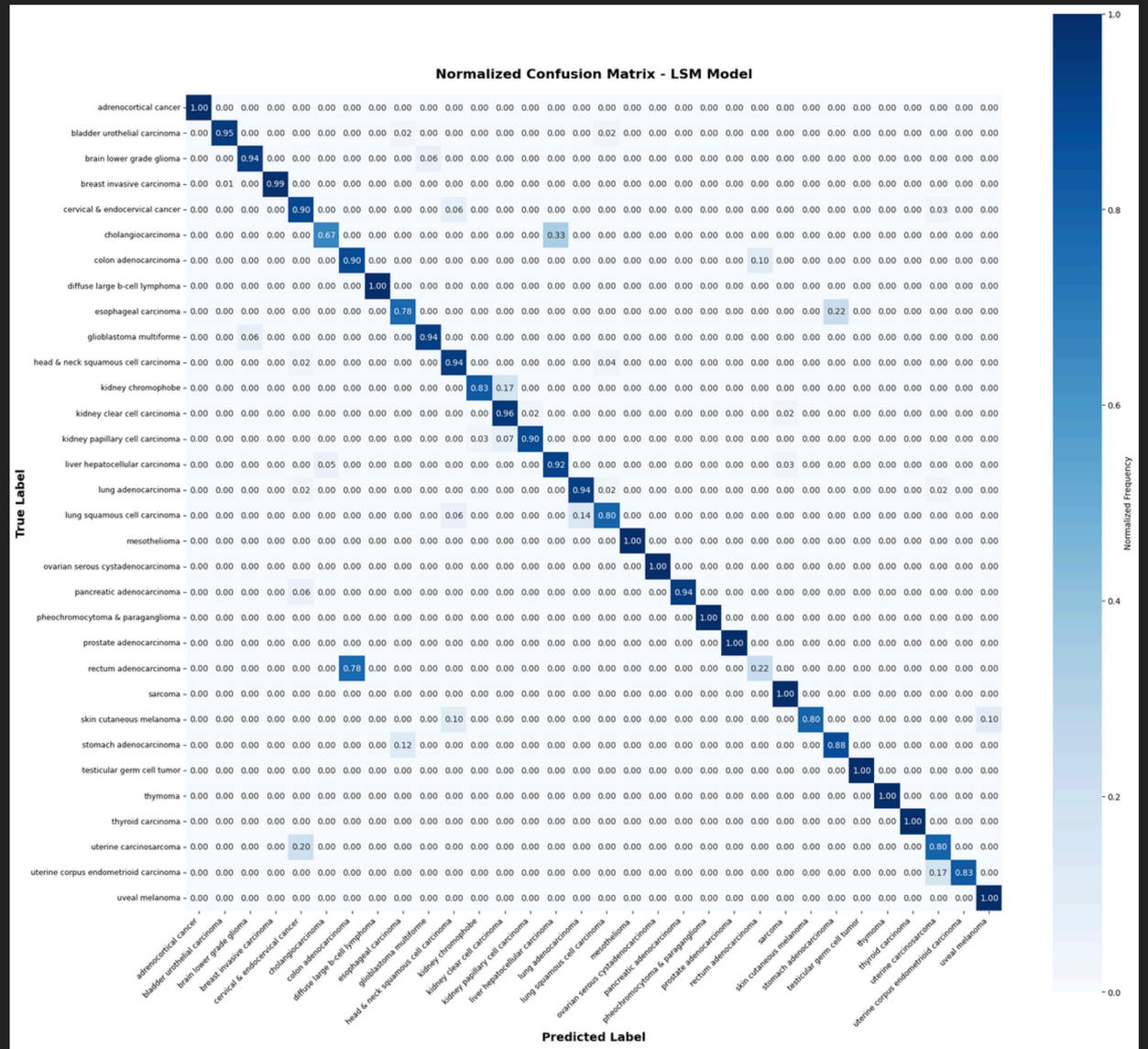
Comparison


Results
1
0.9325
0.952
0.9353
0.921
0.9325
0.931
0.9324
0.924
accuracy
precision
recall
f1
LSTM
CNN

Types


Per-Class Accuracy - LSM Model
uveal melanoma 1.00
thyroid carcinoma 1.00
testicular germ cell tumor 1.00
thymoma 1.00
ovarian serous cystadenocarcinoma 1.00
pheochromocytoma & paraganglioma 1.00
prostate adenocarcinoma 1.00
sarcoma 1.00
mesothelioma 1.00
adrenocortical cancer 1.00
diffuse large b-cell lymphoma 1.00
breast invasive carcinoma 0.99
kidney clear cell carcinoma 0.96
bladder urothelial carcinoma 0.95
pancreatic adenocarcinoma 0.94
head & neck squamous cell carcinoma 0.94
lung adenocarcinoma 0.94
brain lower grade glioma 0.94
glioblastoma multiforme 0.94
liver hepatocellular carcinoma 0.92
cervical & endocervical cancer 0.90
kidney papillary cell carcinoma 0.90
colon adenocarcinoma 0.90
stomach adenocarcinoma 0.88
uterine corpus endometrioid carcinoma 0.83
kidney chromophobe 0.83
uterine carcinosarcoma 0.80
skin cutaneous melanoma 0.80
lung squamous cell carcinoma 0.80
esophageal carcinoma 0.78
cholangiocarcinoma 0.67
rectum adenocarcinoma 0.22
0.0
0.2
0.4
0.6
0.8
1.0
Accuracy

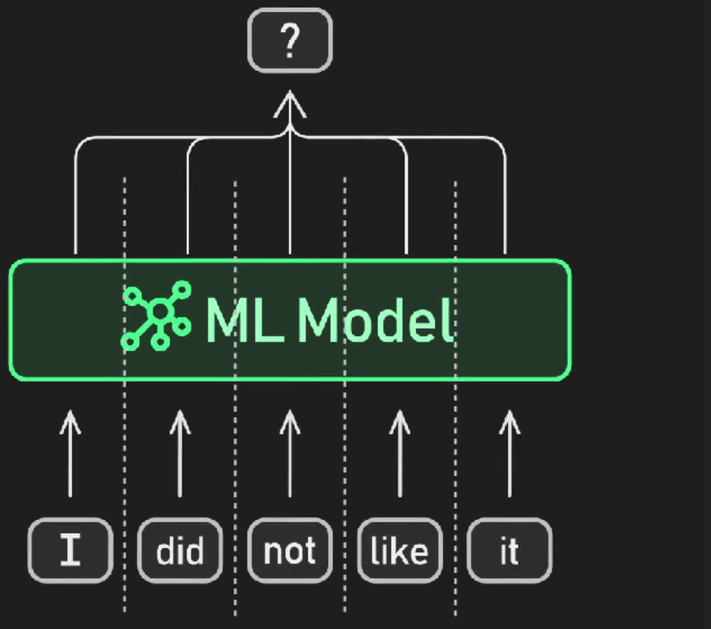
Confusion Matrix



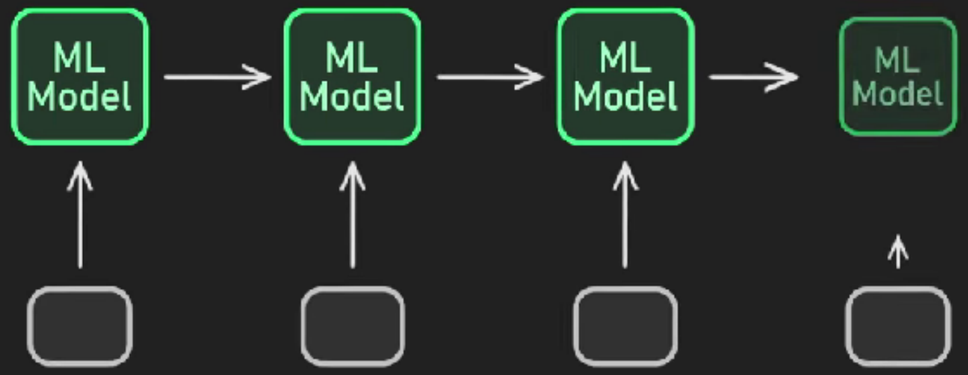
Transformers

Before Transformer Problems

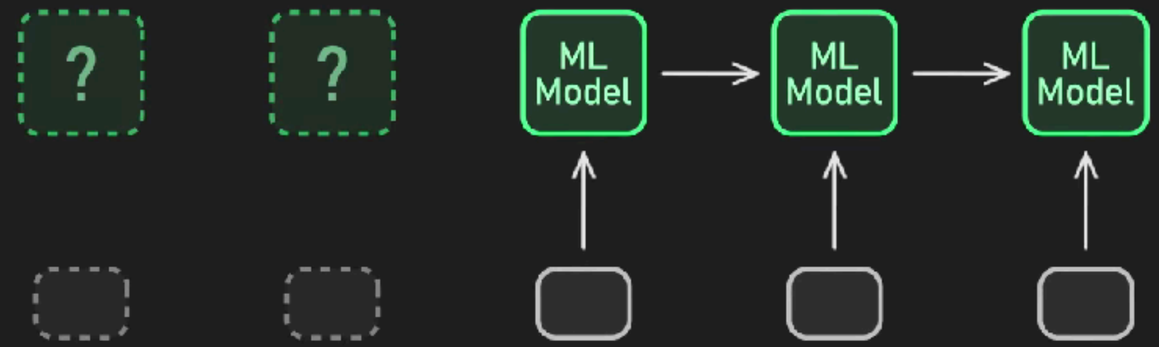
Tokens independent
Lost context

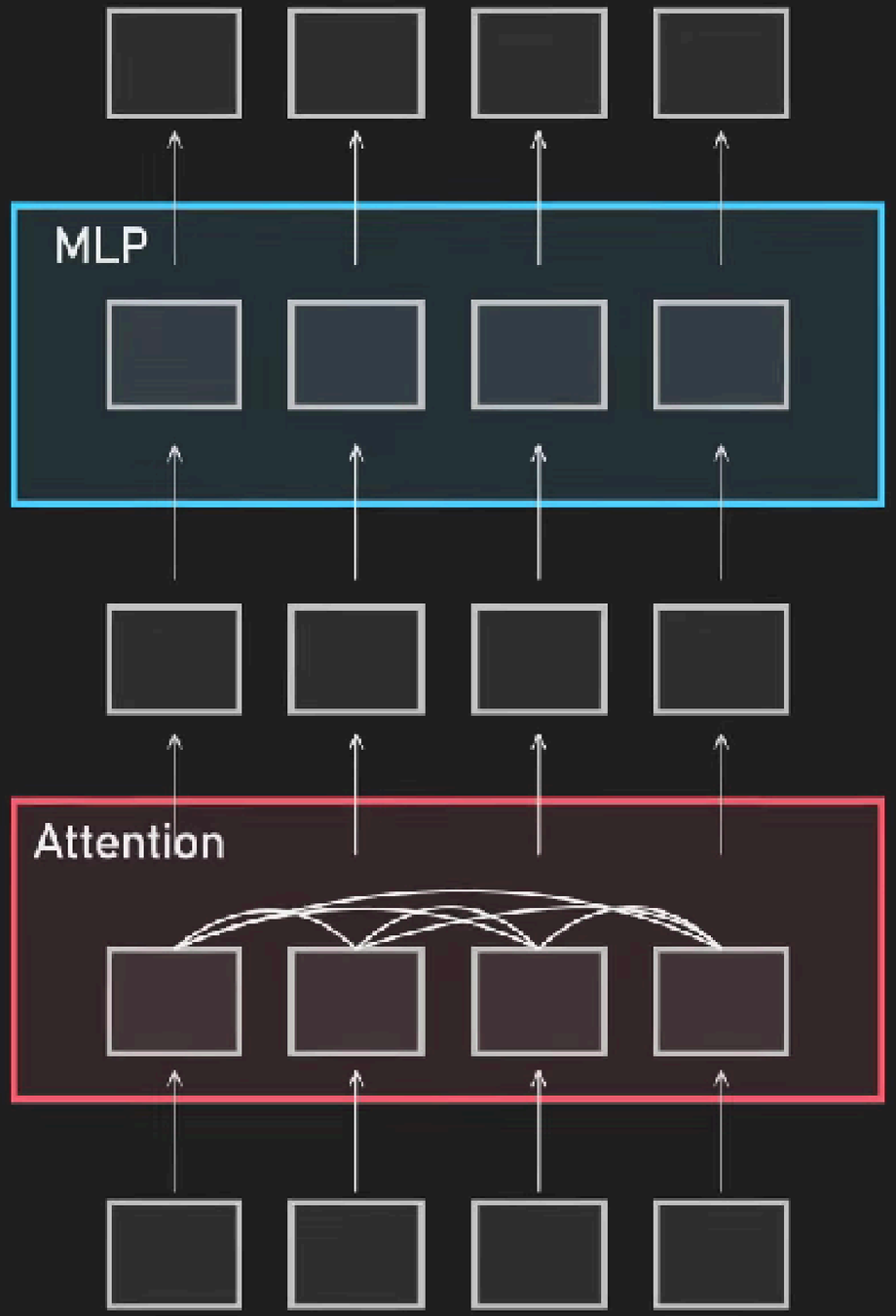


Sequential

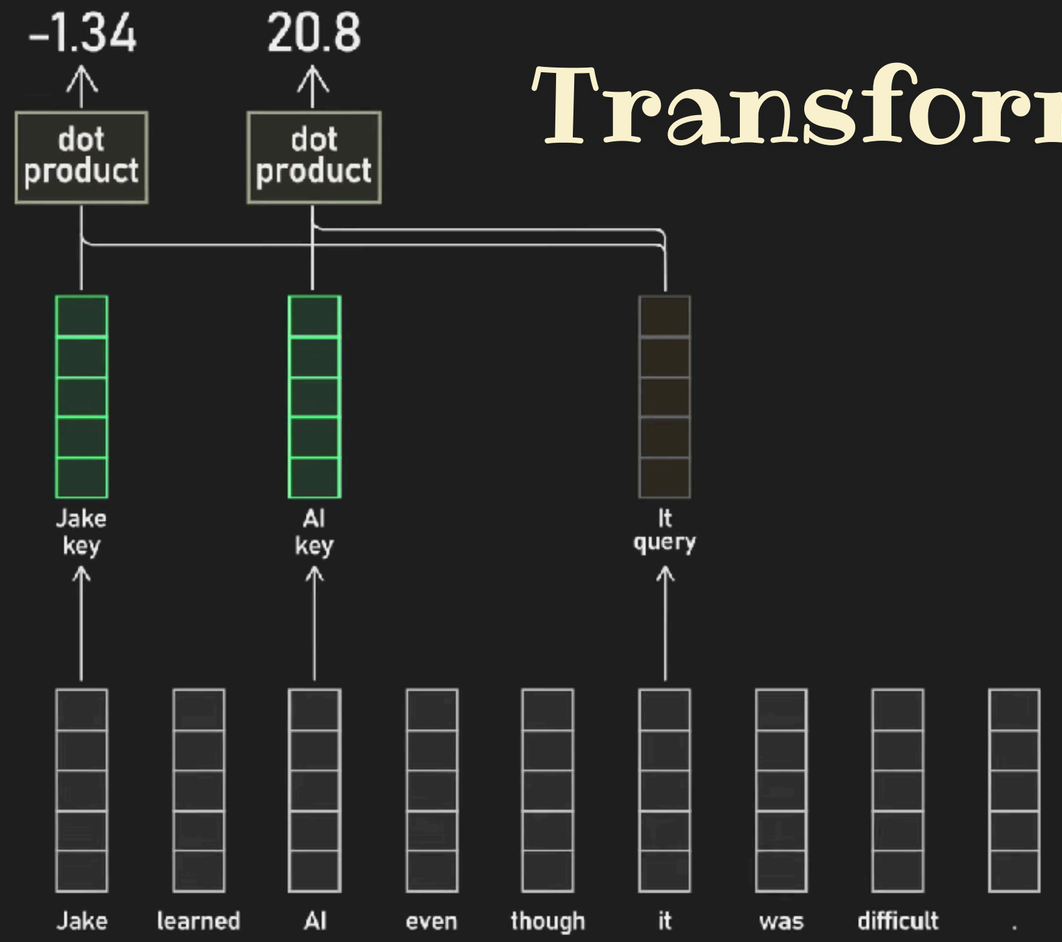


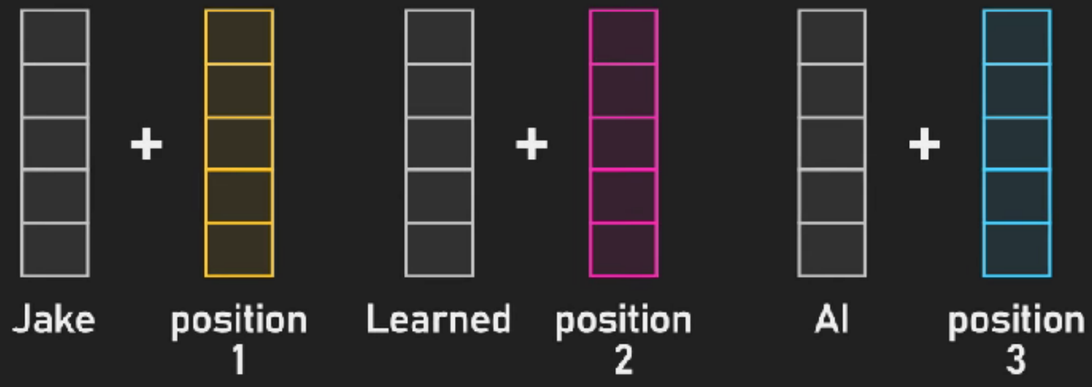
No Long Term Dependencies

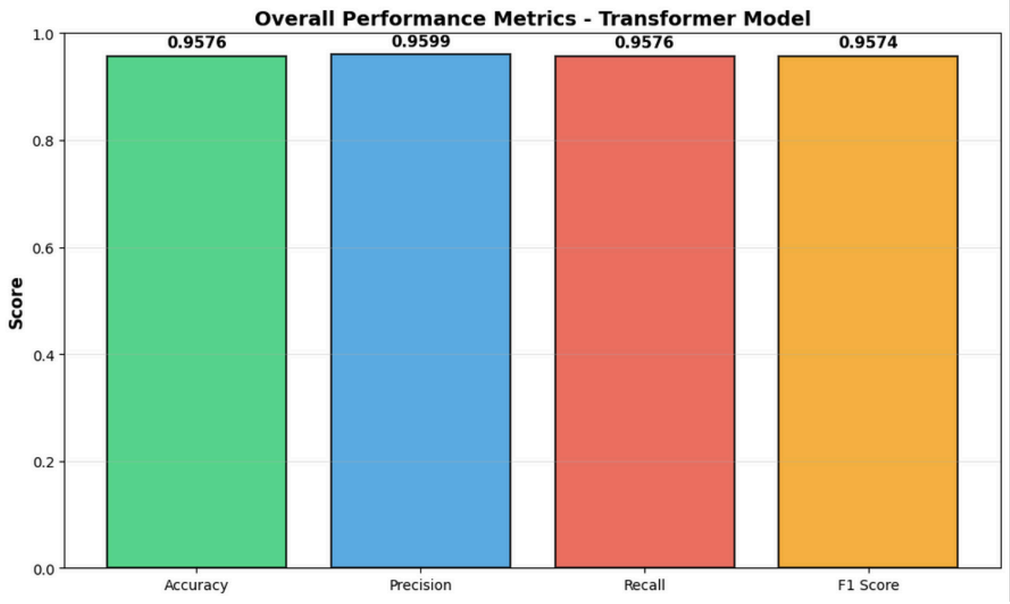




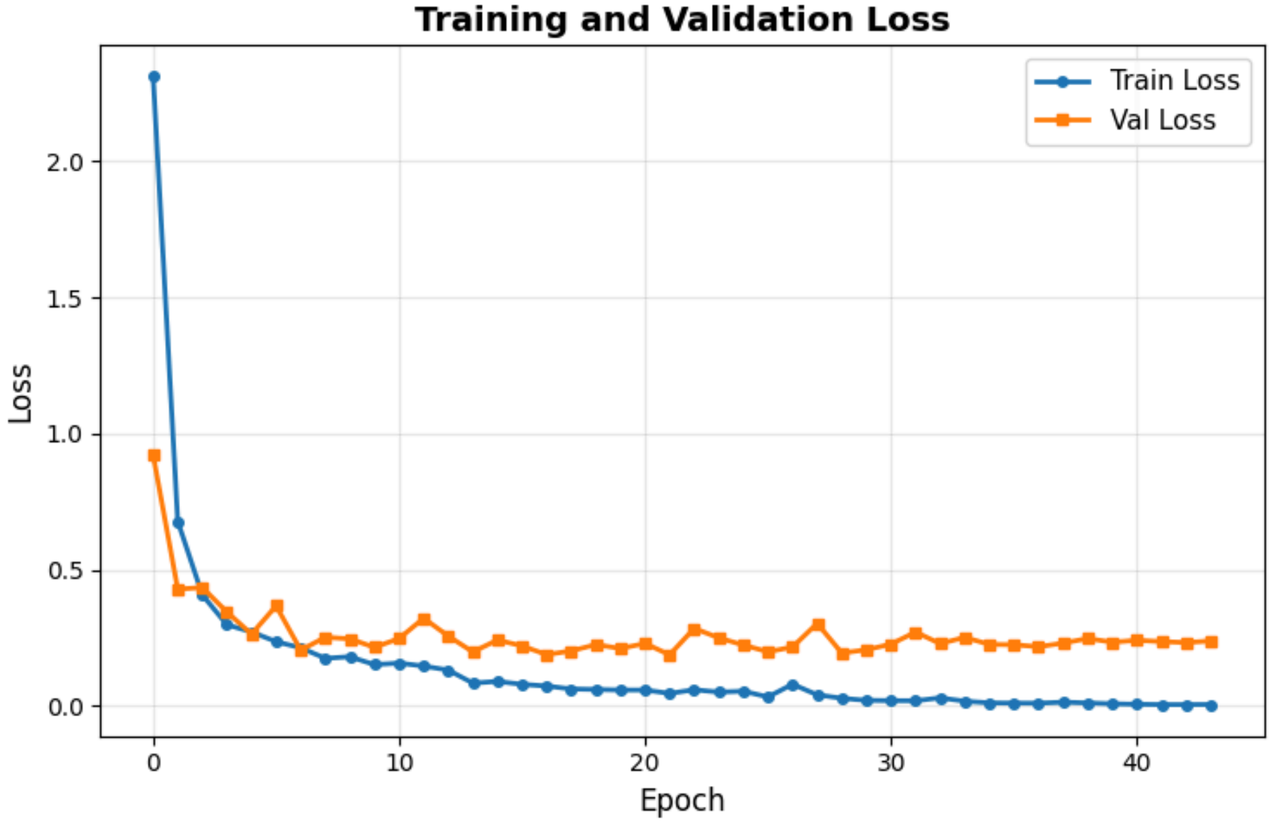
Transformer

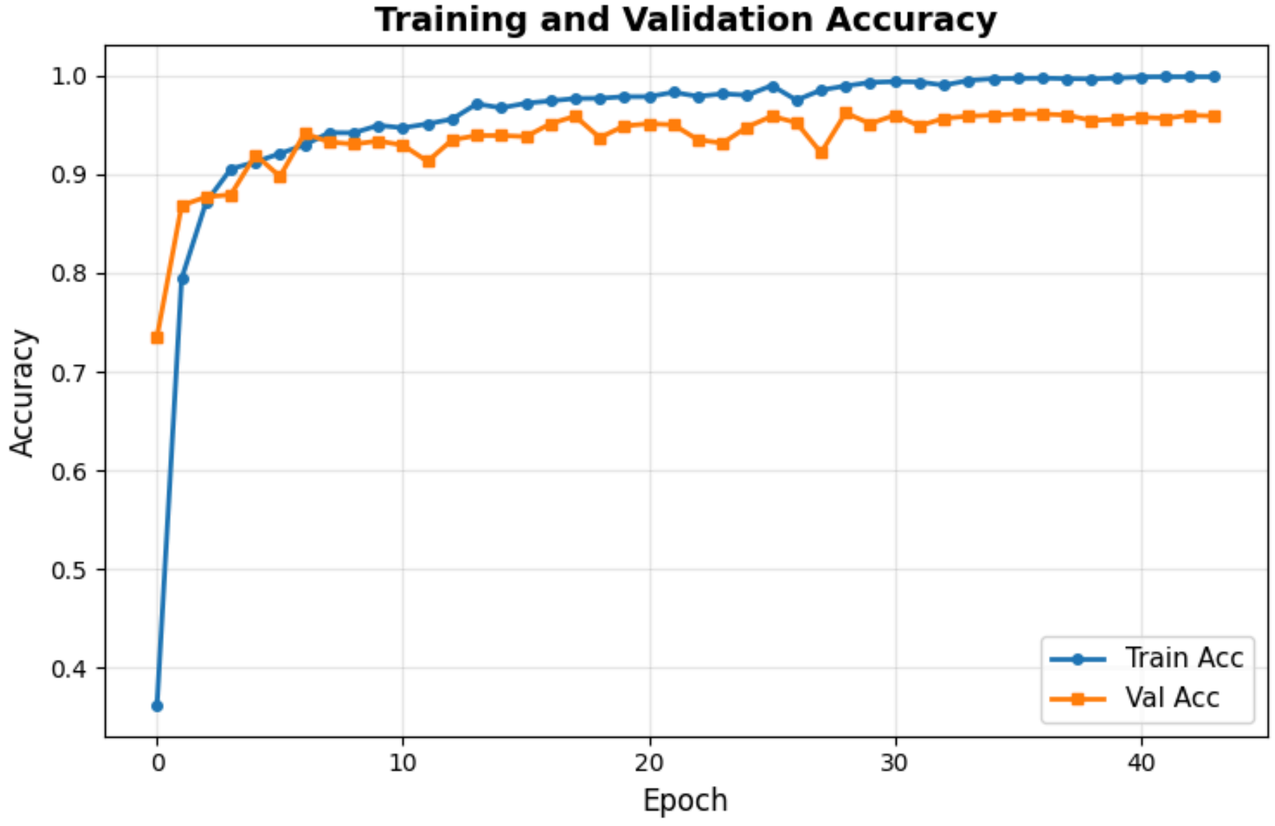




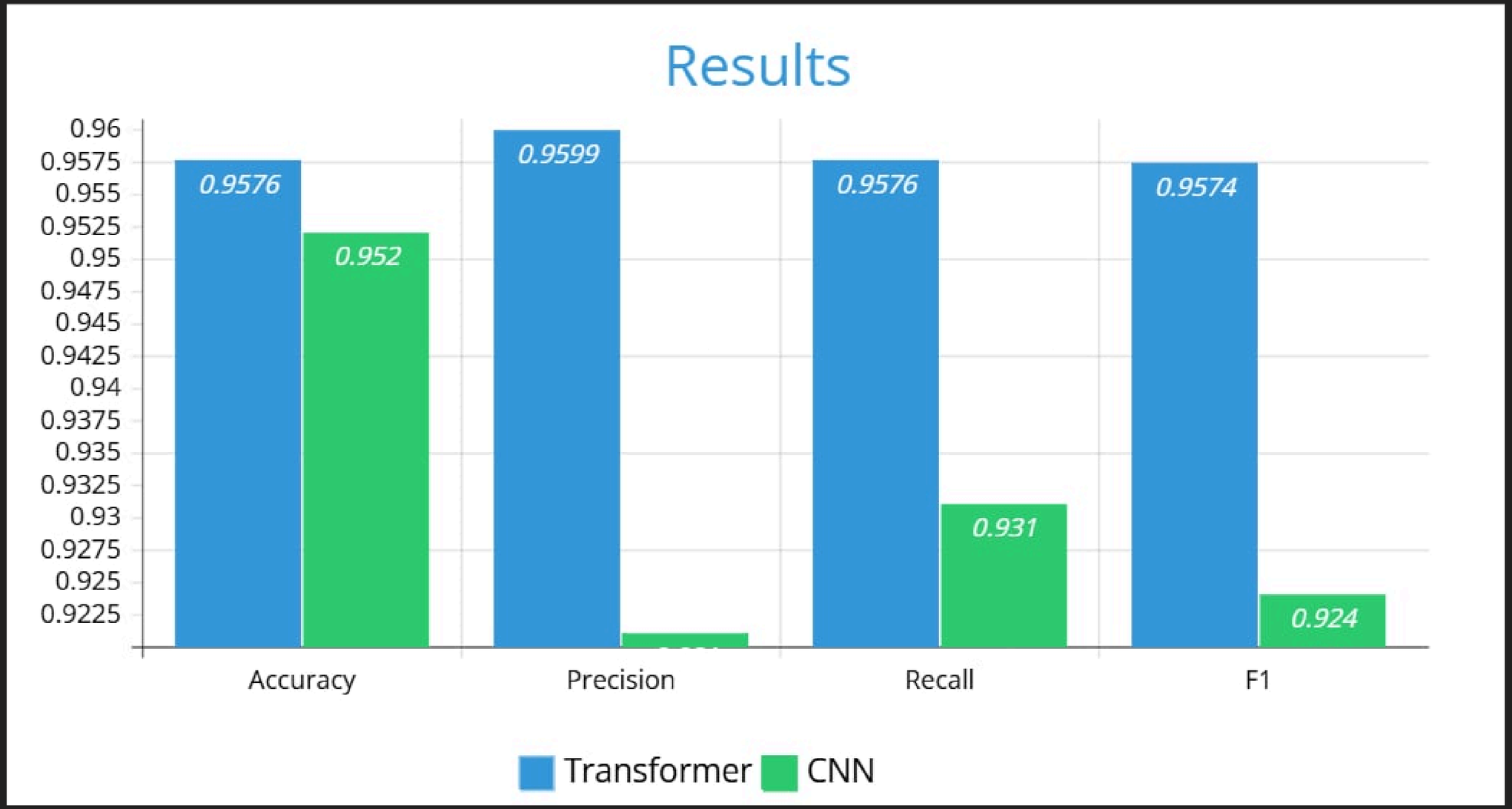

Overall Performance Metrics - Transformer Model
Score
0.9576
0.9599
0.9576
0.9574
Accuracy
Precision
Recall
F1 Score

Training

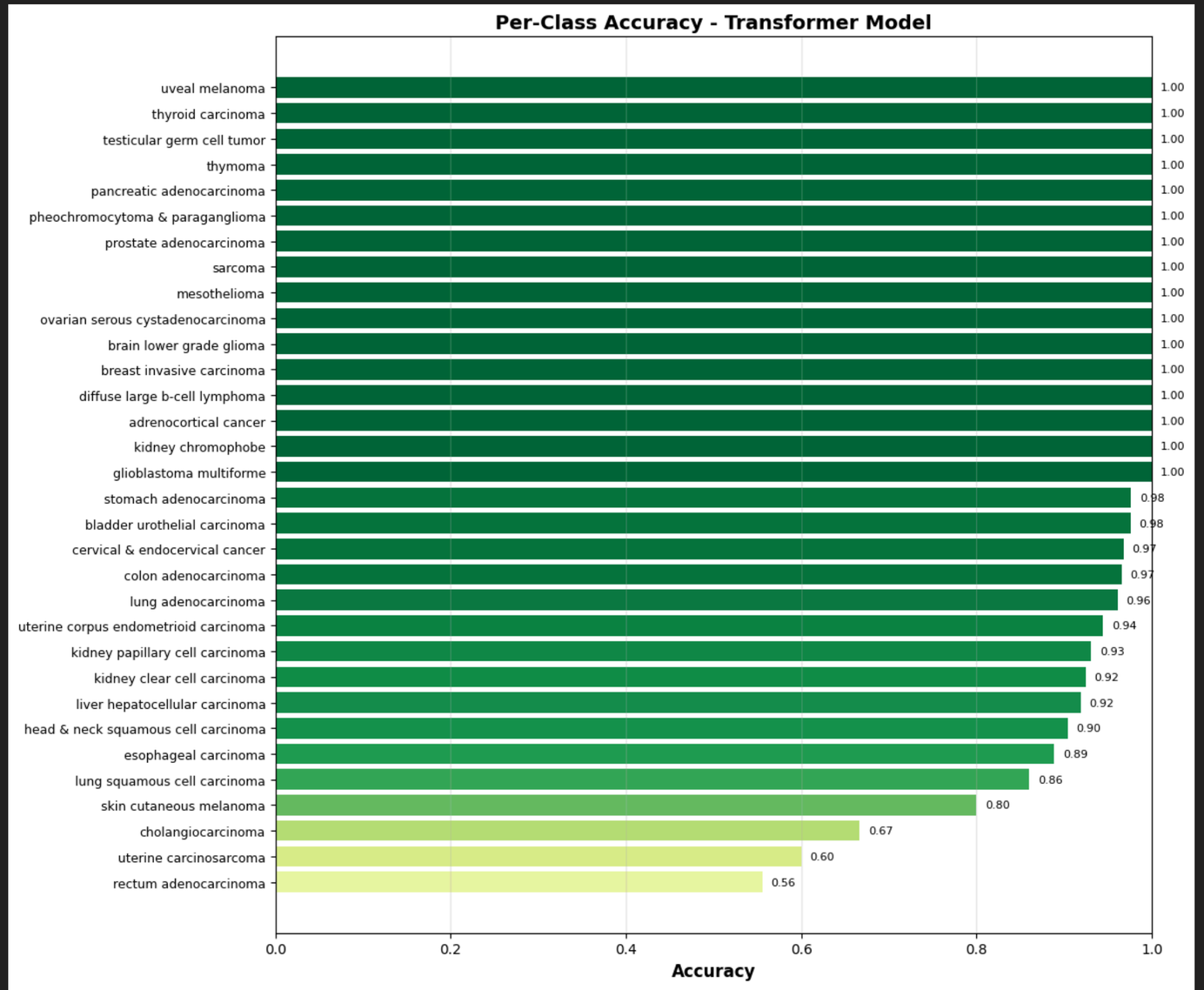

Training and Validation Loss
Train Loss
Val Loss
Loss
Epoch


Training and Validation Accuracy
Train Acc
Val Acc
Accuracy
Epoch

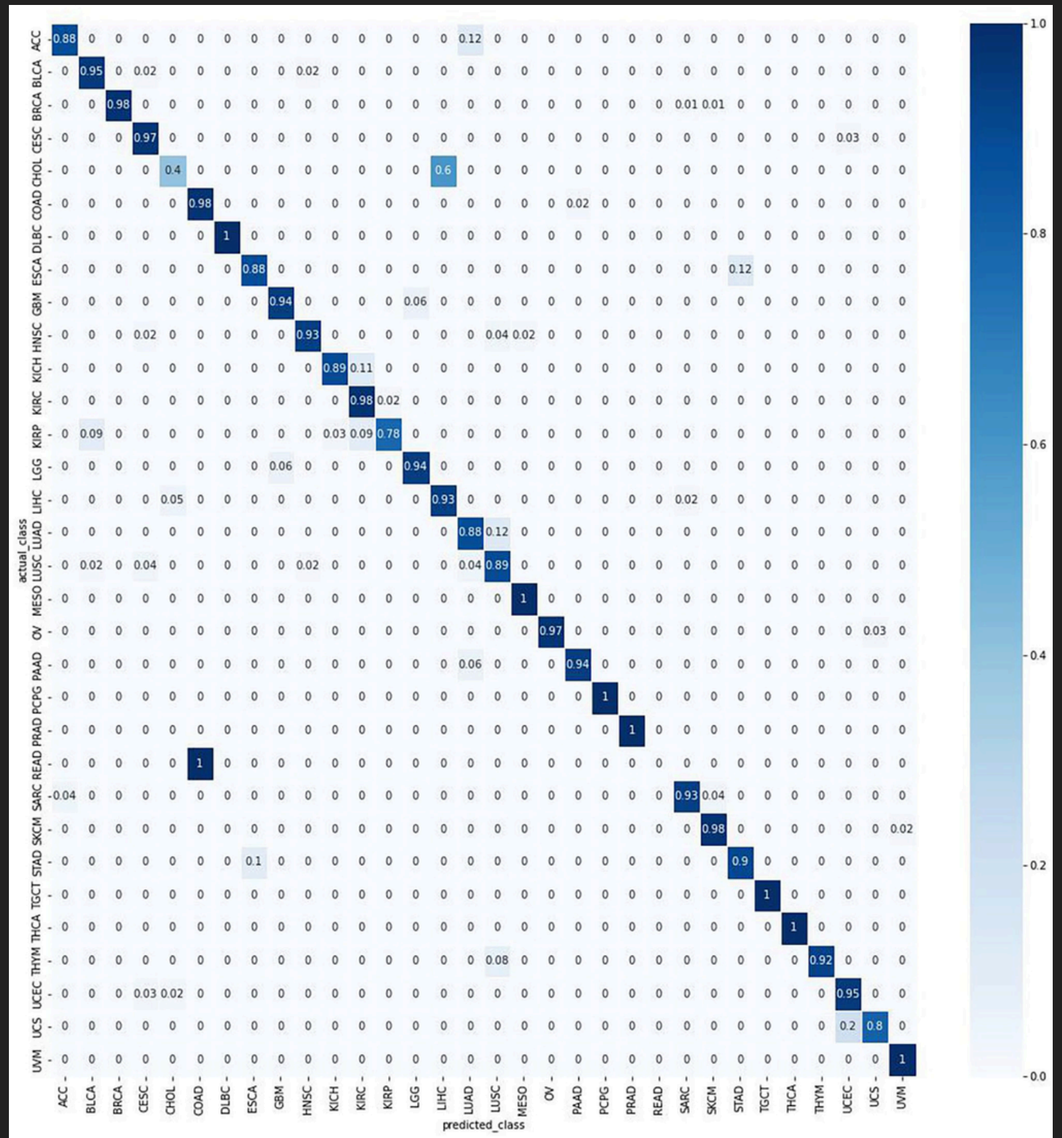
Comparison

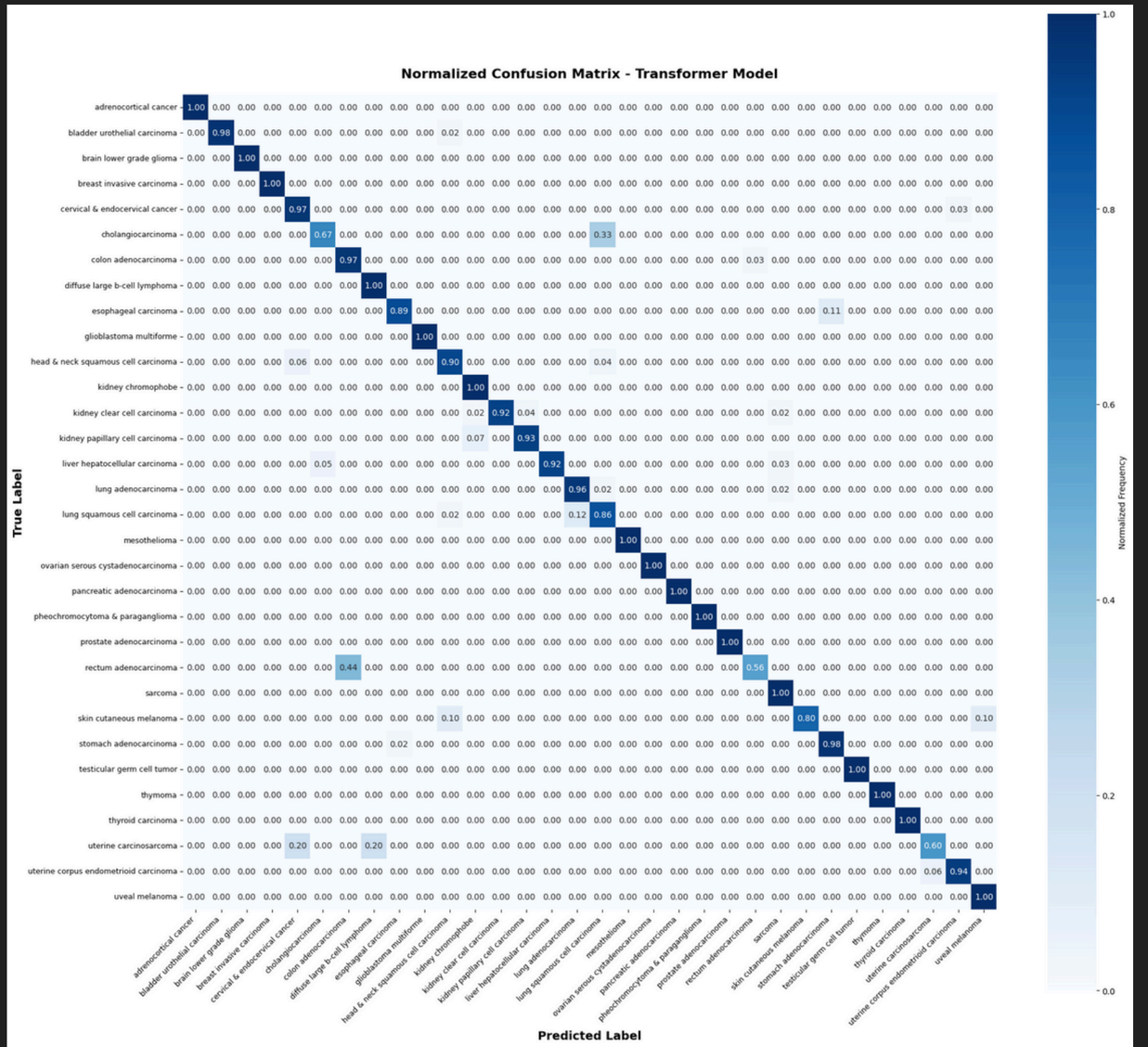

Results
0.96
0.9575
0.955
0.9525
0.95
0.9475
0.945
0.9425
0.94
0.9375
0.935
0.9325
0.93
0.9275
0.925
0.9225
0.9576
0.952
0.9599
0.9576
0.931
0.9574
0.924
Accuracy
Precision
Recall
F1
Transformer
CNN

Types

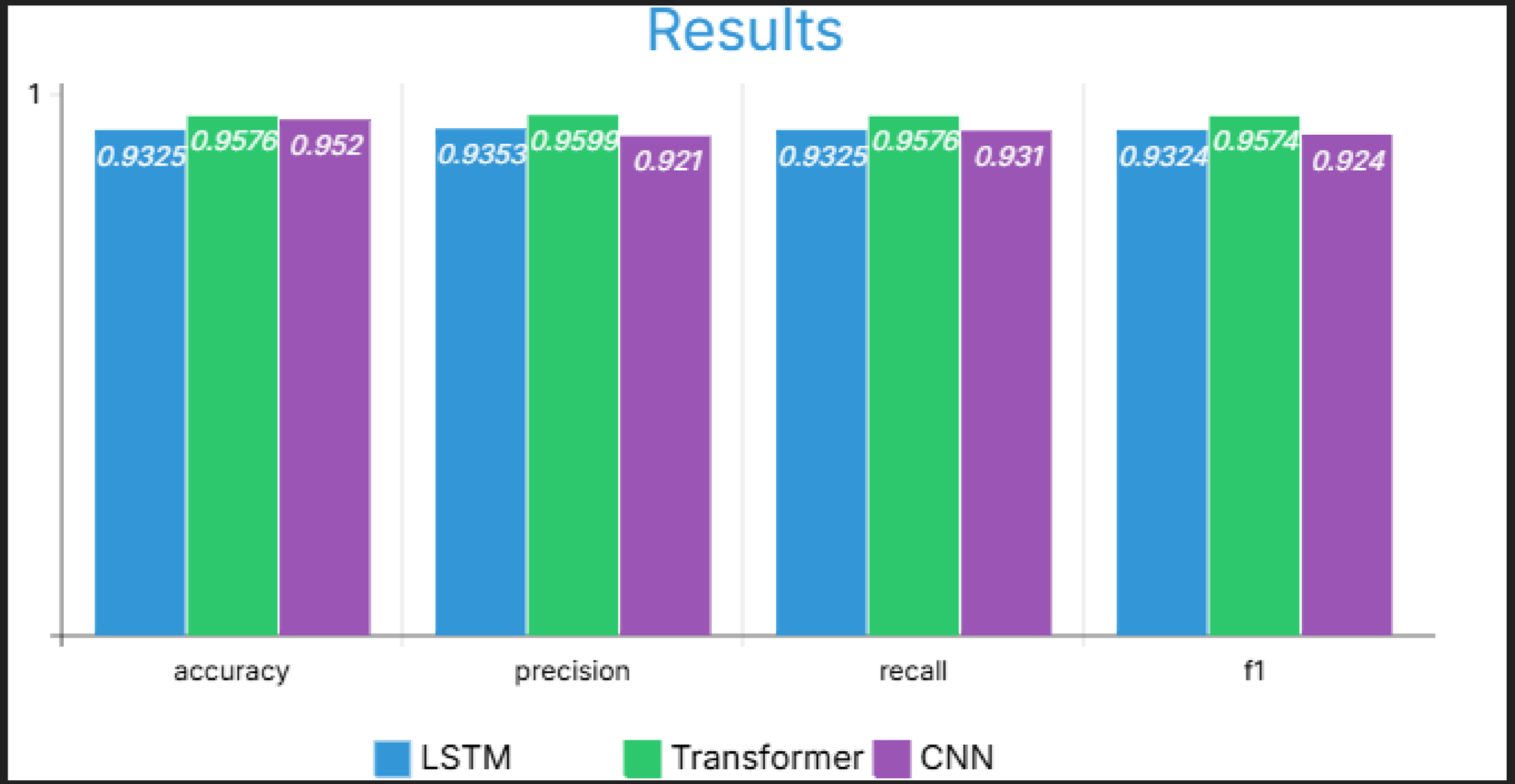

Per-Class Accuracy - Transformer Model
uveal melanoma 1.00
thyroid carcinoma 1.00
testicular germ cell tumor 1.00
thymoma 1.00
pancreatic adenocarcinoma 1.00
pheochromocytoma & paraganglioma 1.00
prostate adenocarcinoma 1.00
sarcoma 1.00
mesothelioma 1.00
ovarian serous cystadenocarcinoma 1.00
brain lower grade glioma 1.00
breast invasive carcinoma 1.00
diffuse large b-cell lymphoma 1.00
adrenocortical cancer 1.00
kidney chromophobe 1.00
glioblastoma multiforme 1.00
stomach adenocarcinoma 0.98
bladder urothelial carcinoma 0.98
cervical & endocervical cancer 0.97
colon adenocarcinoma 0.97
lung adenocarcinoma 0.96
uterine corpus endometrioid carcinoma 0.94
kidney papillary cell carcinoma 0.93
kidney clear cell carcinoma 0.92
liver hepatocellular carcinoma 0.92
head & neck squamous cell carcinoma 0.90
esophageal carcinoma 0.89
lung squamous cell carcinoma 0.86
skin cutaneous melanoma 0.80
cholangiocarcinoma 0.67
uterine carcinosarcoma 0.60
rectum adenocarcinoma 0.56
0.0
0.2
0.4
0.6
0.8
1.0
Accuracy

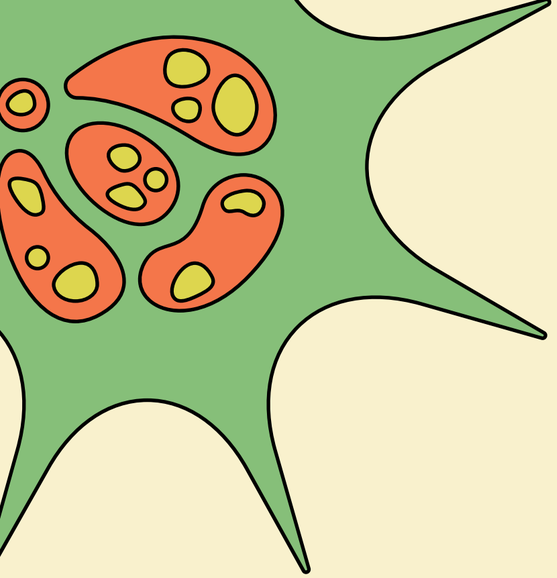
Transformer VS CNN

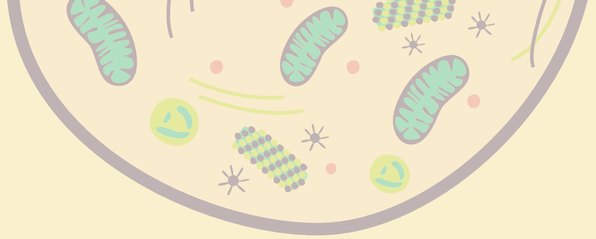


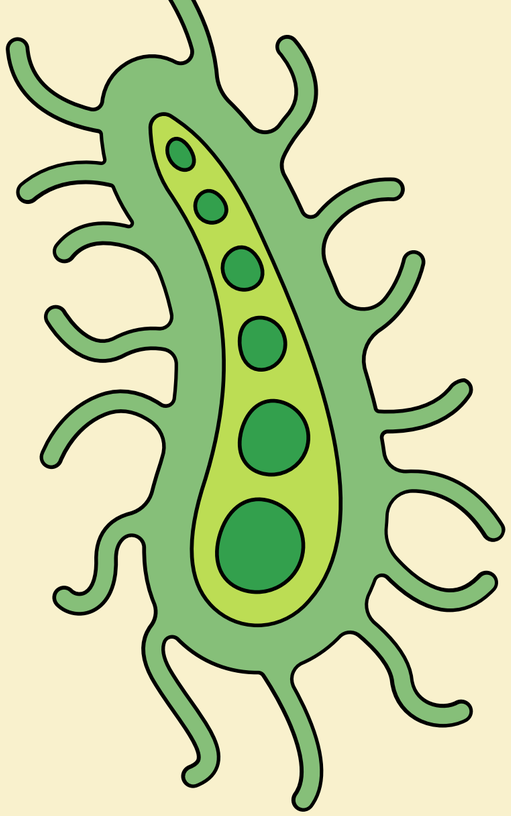


Summary


Results
1
0.9325
0.9576
0.952
0.9353
0.9599
0.921
0.9325
0.9576
0.931
0.9324
0.9574
0.924
accuracy
precision
recall
f1
LSTM
Transformer
CNN







Thank You