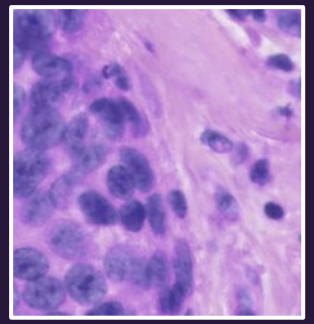
DATA3888 (IMAGE)

DISCIPLINARY ASSIGNMENT 2

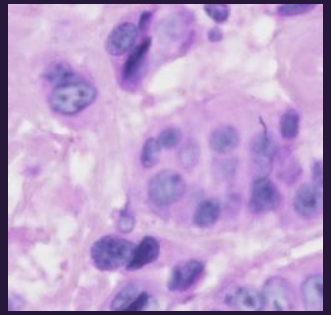
Building a model to classify tumour cells from immune cells in an H&E image

OVERVIEW

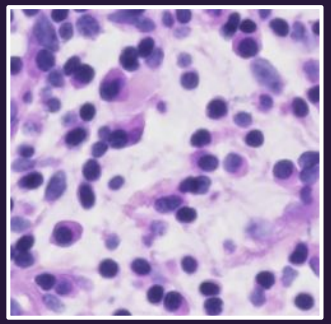


TUMOUR

- Size: 1,000



?



IMMUNE

- Size: 976
  (4 blank images deleted)

DATA PROCESSING

TRAIN-TEST SPLIT

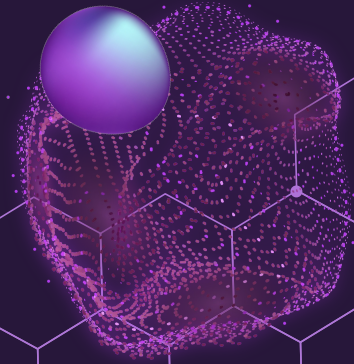
- Test set: random 20% of each
- Train set: the rest 80%

=> negligibly imbalance

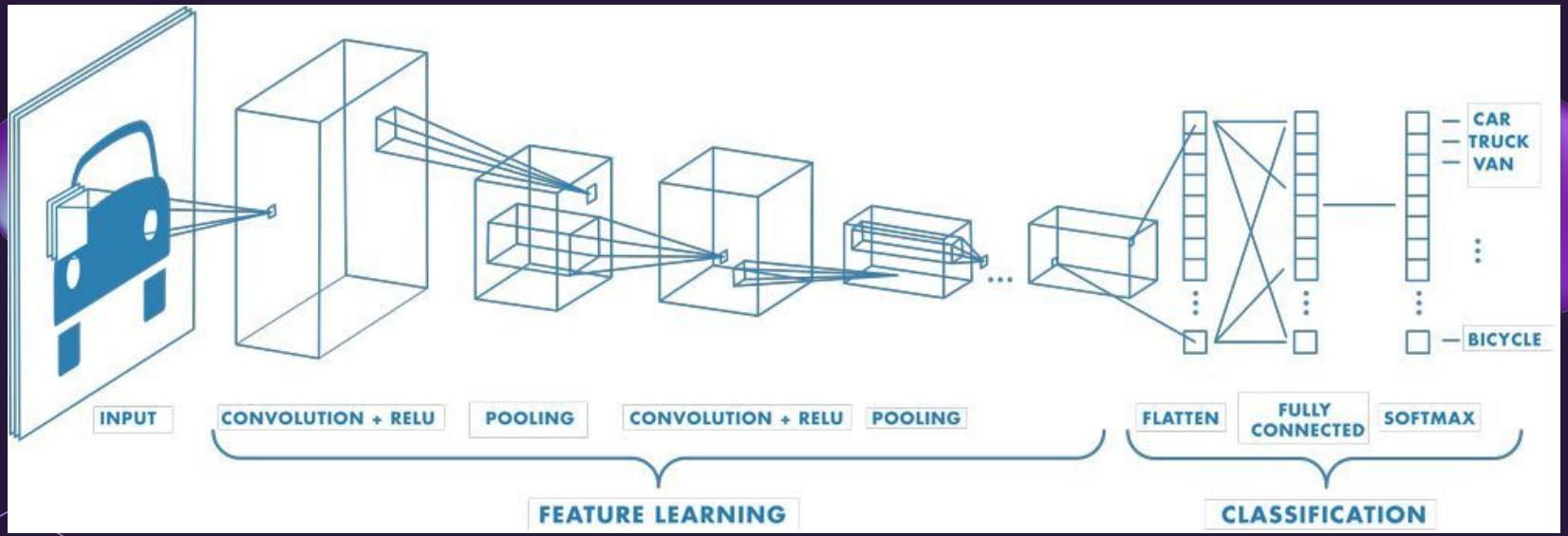
CROSS VALIDATION

- Repetitively train on the train set to select the best fine-tune model

BASE MODELS

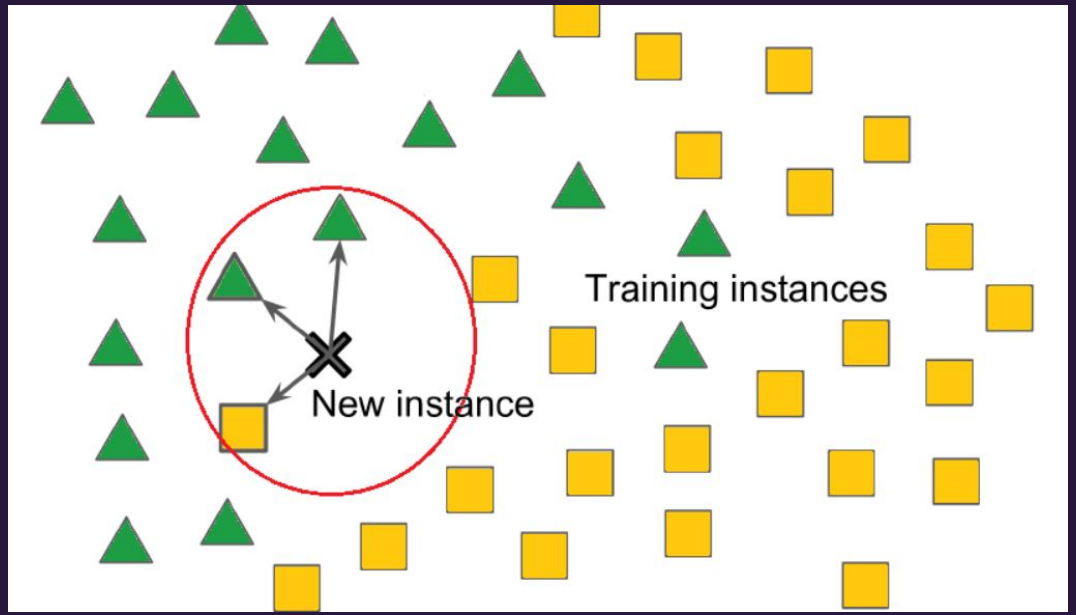


CONVOLUTIONAL NEURAL NETWORK



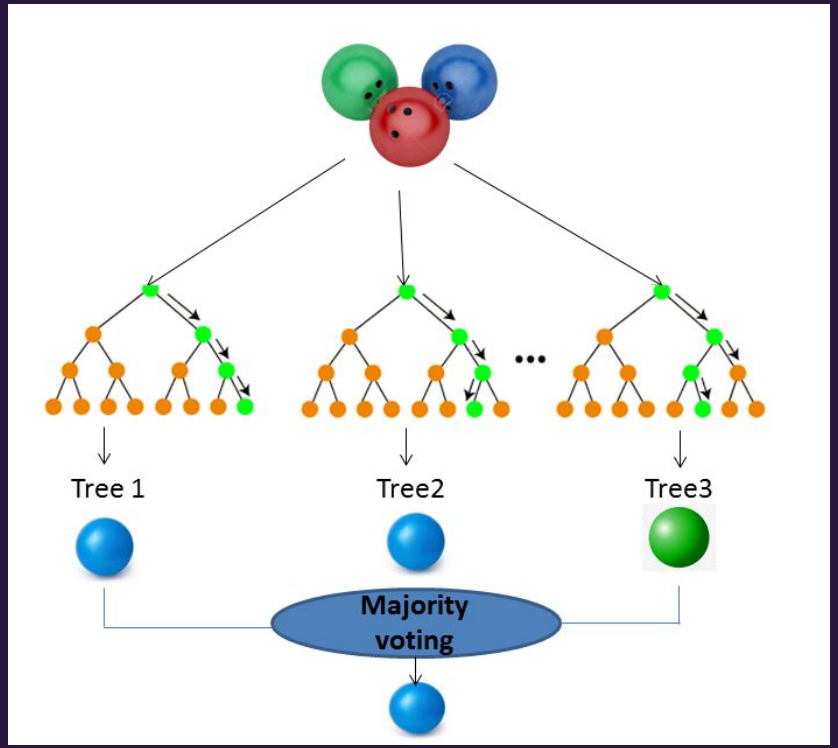
Source: The MathWorks, Inc. (n.d.)

K-NEAREST NEIGHBOURS



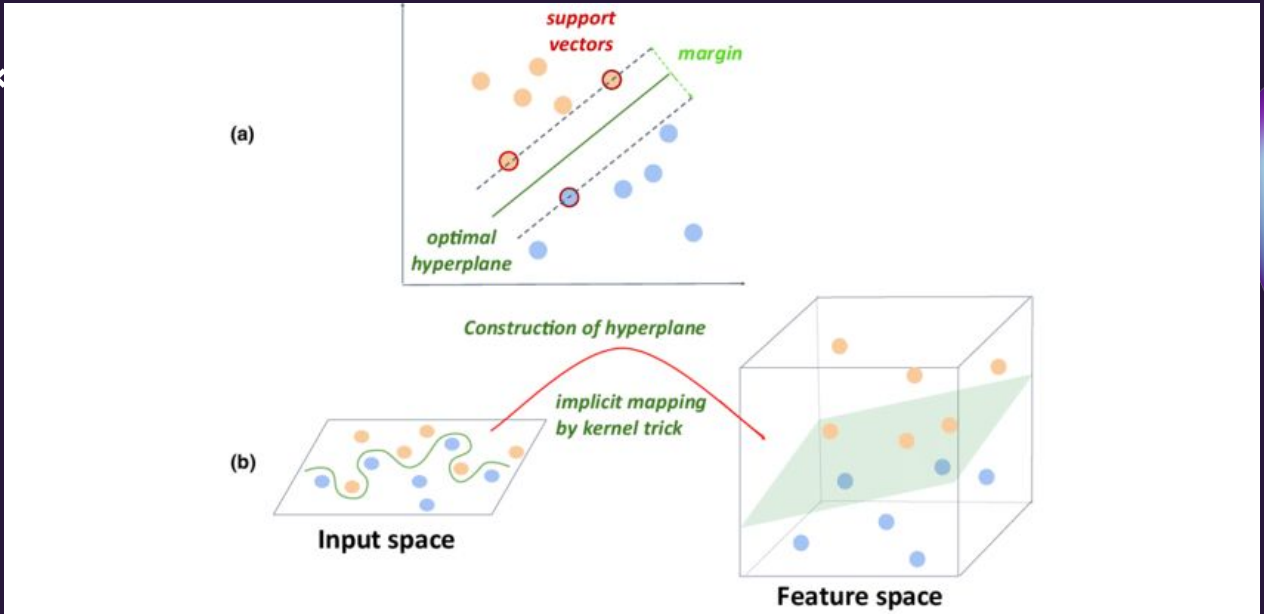
Source: Raafat (2022)

RANDOM FOREST



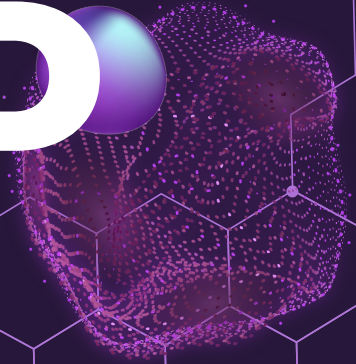
Source: Shiksha Online (2022)

SUPPORT VECTOR MACHINE



Source: Badillo et al. (2020)

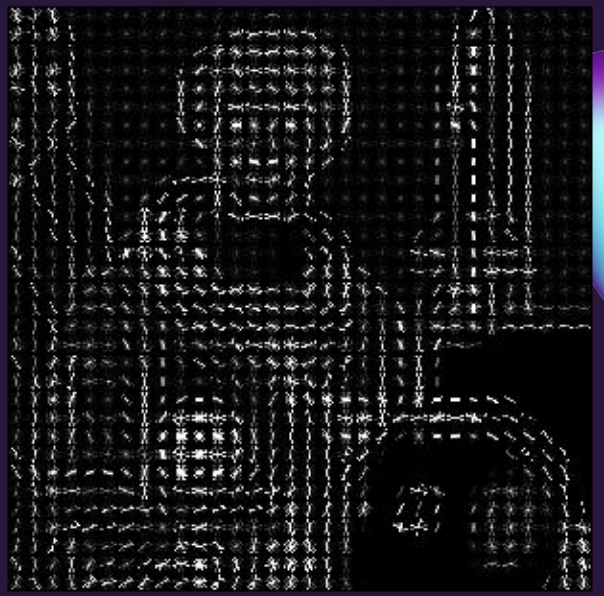
FEATURES EXTRACTED



HISTOGRAM OF GRADIENTS

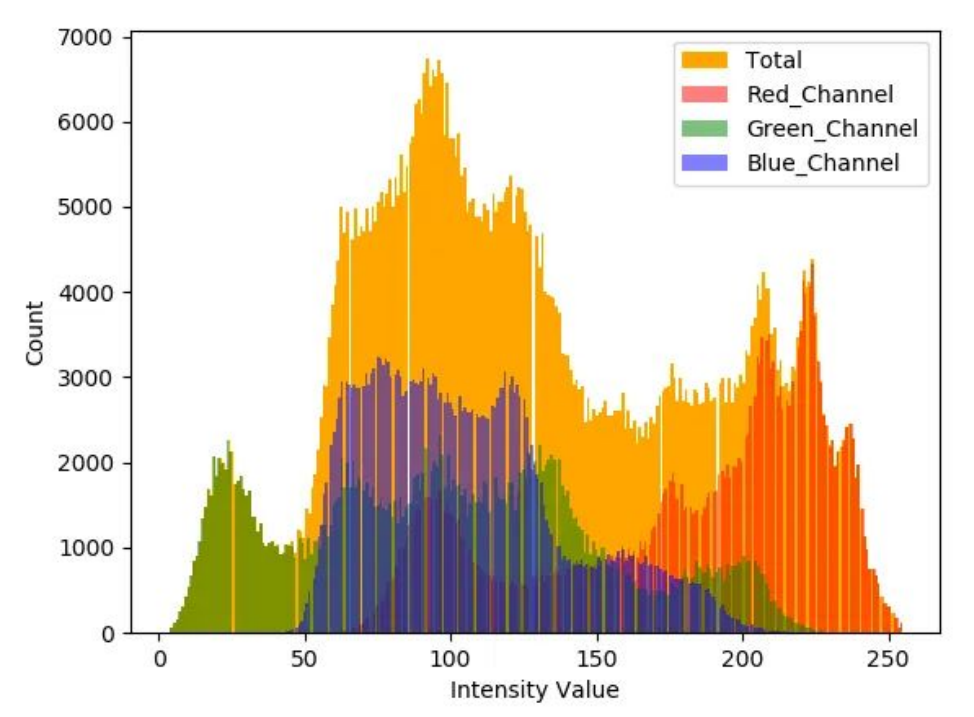
Input Image

Histogram of Oriented Gradients



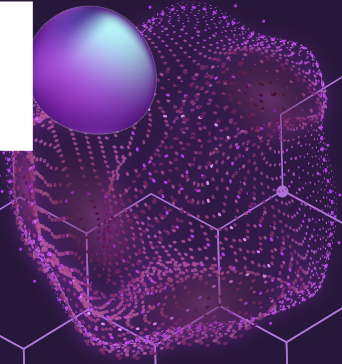
Source: Mittal (2020)

HISTOGRAM OF COLOURS

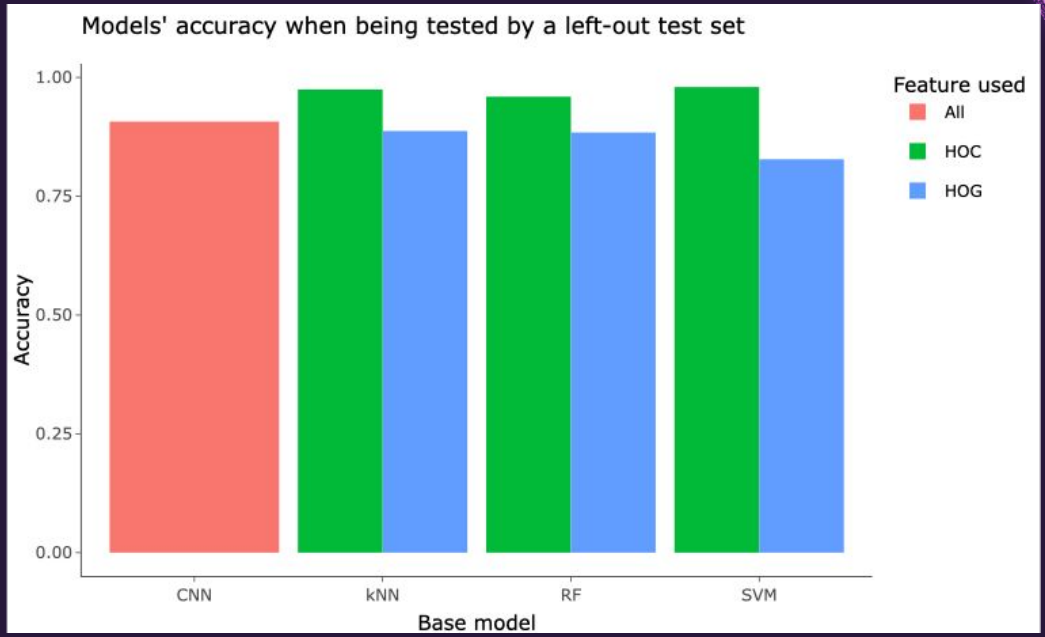


Source: Sajid (2024)

MODEL SELECTION



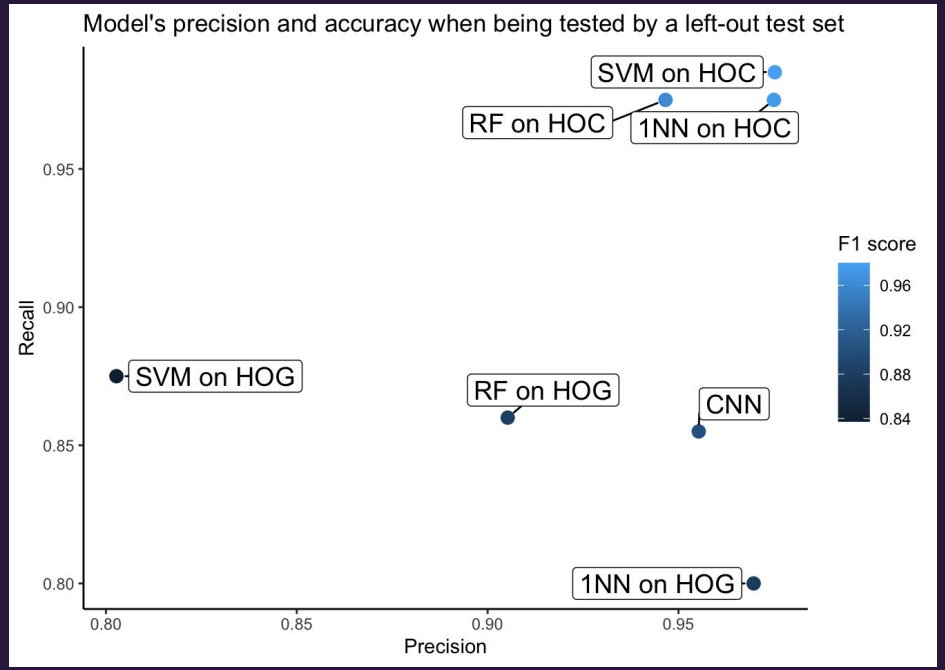
ACCURACY



i.e. the proportion of both tumour and immune cells that the model correctly classifies

RECALL - PRECISION - F1 SCORE

i.e. the proportion of actual tumour cells that the model correctly identified



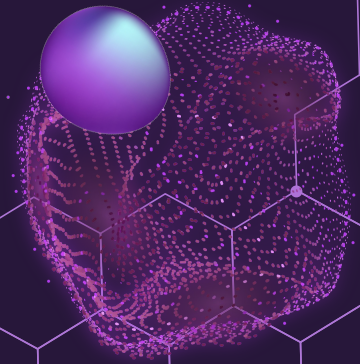
i.e. the balance of precision and recall - it's high if and only if the other 2 are high!

i.e. the proportion of cells the model predicted as 'tumour' that were actually tumour

SUMMARY TABLE

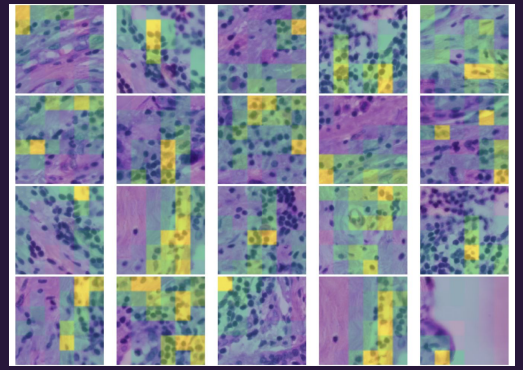
	model	accuracy	prec	recall	f1
7	SVM on HOC	0.98	0.975	0.985	0.98
3	1NN on HOC	0.975	0.975	0.975	0.975
5	RF on HOC	0.959	0.947	0.975	0.961
1	CNN	0.907	0.955	0.855	0.902
2	1NN on HOG	0.887	0.97	0.8	0.877
4	RF on HOG	0.884	0.905	0.86	0.882
6	SVM on HOG	0.828	0.803	0.875	0.837

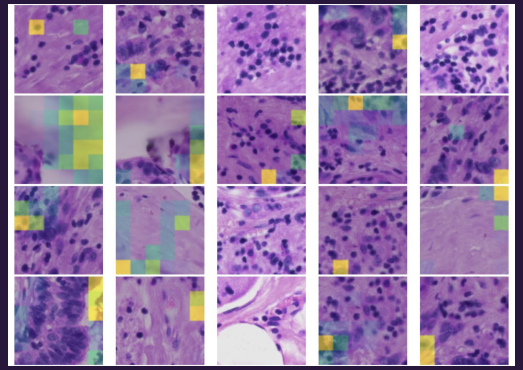
LIMITATIONS & NEXT STEPS



WHAT'S WRONG WITH CNN?

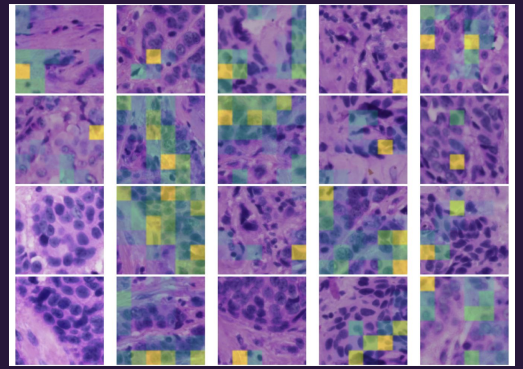
Predict
= actual
= tumour

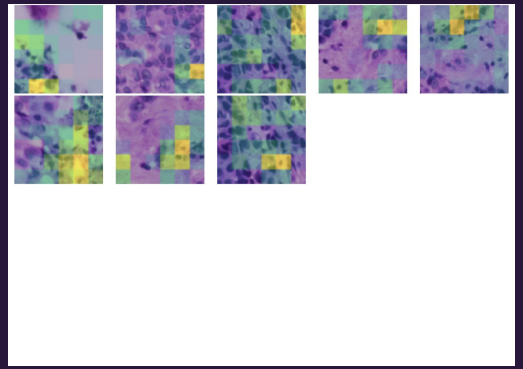




Predict
= actual
= immune

Predict
= tumour
≠ immune





Predict
= immune
≠ tumour

OTHER LIMITATIONS

COMP COST

Computational cost
=> the hyperparameters are only best at its capable range

TIME LIMIT

There might be some other base models and features that can perform the task better

THE MORE THE MERRIER

With more data being used, we can be hopeful about the future of this project

DOMAIN KNOWLEDGE

With more domain knowledge, we could potentially go on a better track to improve the model


thank
for
watching