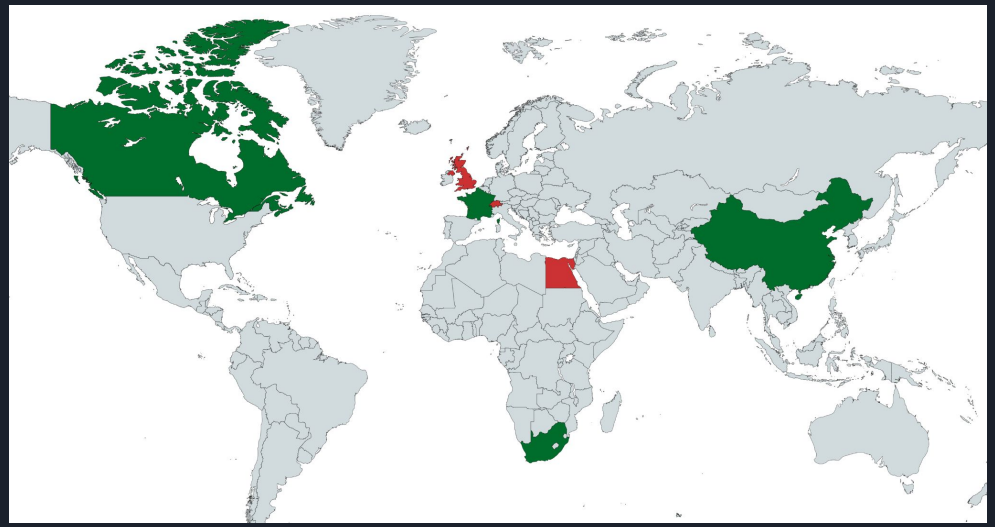
ORLIAn presents: Eln-AI-ny

What is ORLIAn?



Project set up to tackle Challenge 1, made of international volunteers, spanning across different time zones!

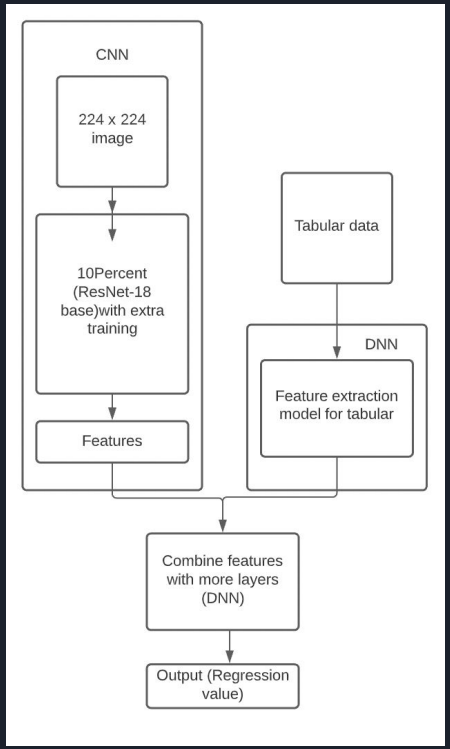
Wide range of skills from data science to medical science, to design skills and consulting with tasks!

I played in all the countries where the core contributors are from, so they named it after me

What did we do?

1. Trained the EI-nAI-ny model using a novel approach: using a pre-trained Res-Net [1] model trained on data from other histological sources as a base and then visualized using GRAD-CAM
2. Conducted analysis of the TIFF files by our team's medical expertise

[1] TenPercent - (Self supervised learning for digital histopathology)



How did we approach model construction

Literature review helped glean some ideas

Vanilla ResNet wasn't good enough, so baseline was an pre-trained ResNet model using histological images from other areas such as breast cancer, lung cancer etc.

Evaluation

Initial experiments suggested metadata and image data alone wasn't sufficient.

Wrote a new model which combined the two and fine tuned for performance, including some feature engineering to help format data more easily

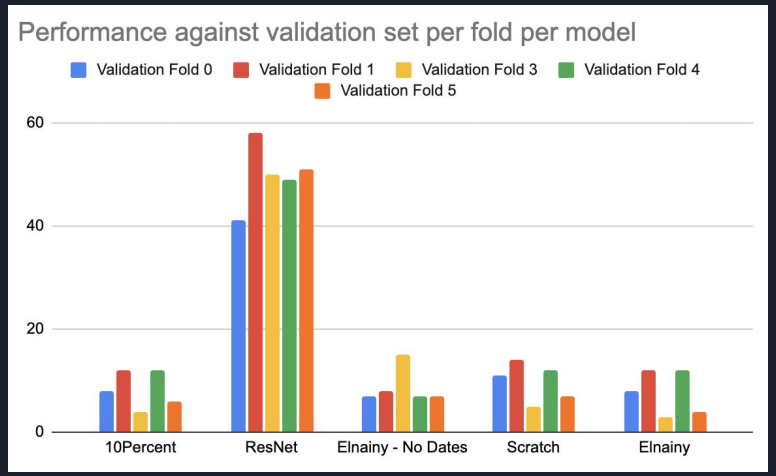
Used K-Fold due to small number of patient IDs, generating folds using the IDs

How did we do?

Best ever model against the Dataiku macro scored 0.829 (MAPE)

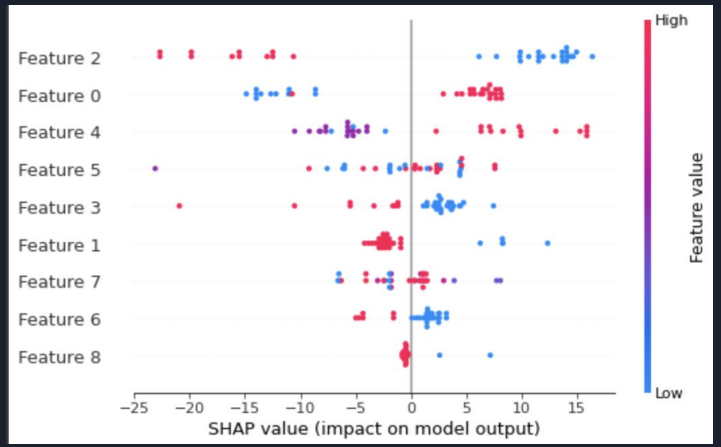
Current model against the Dataiku macro scored 6.00 (MAPE)

Eln-AI-ny outperforms against it's original ResNet-18 model (just the vanilla regressor version, fixed weights) significantly. We used MAE during validation evaluation in each fold



Best ever is better than your current?

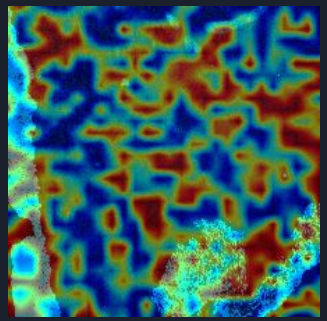
- Experiments suggested decisions coming primarily from tabular
- Turns out our engineered features really help it out! (Feature 2)
- Image data wasn't able to learn too well on its own
- XGBoost with just tabular scored 8.40696 on its validation set (MAE)
- We fixed this exploit by removing the offending features and training again



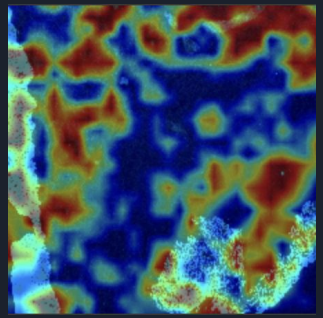
Experiment: Just tabular suggesting feature importance

Far better at grasping cellular information

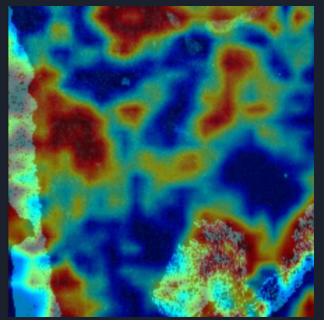
It gets a bit better at task and at identifying cellular structures



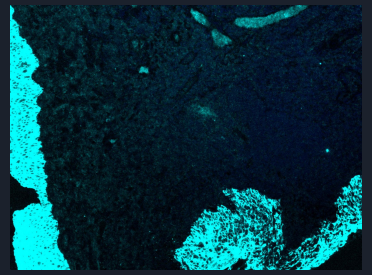
ResNet-18



TenPercent



Eln-AI-ny



Original image

Medical insights

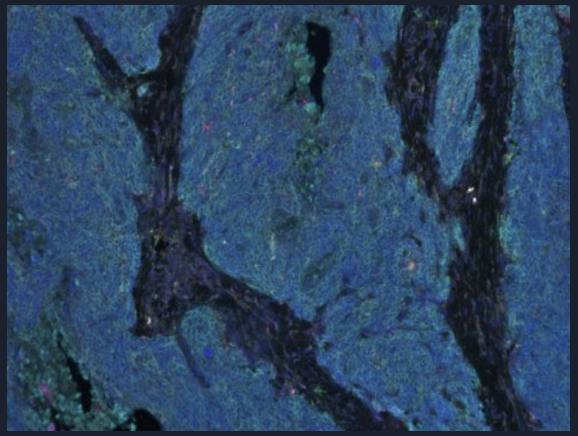
ORLIAn's medical expertise analysed the TIFF images in ImageJ, revealing complex arrangements of inflammatory cells in both tumor/stromal compartments

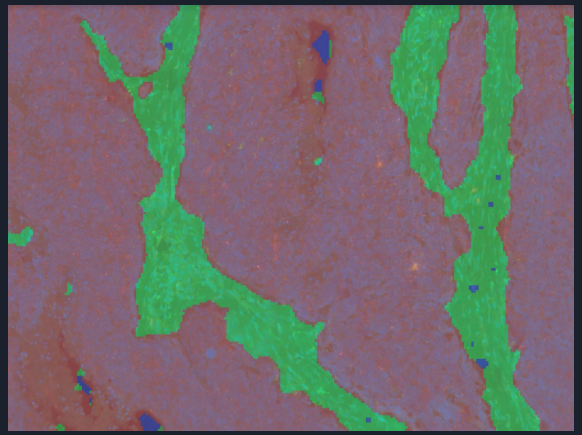
This analysis suggested that these inflammatory cell populations include cells expressing the surface protein PD-1 (programmed cell death 1), acting after activation of T-lymphocytes and persisting in the presence of stimuli.

It is also likely that CD4 and CD8 T-lymphocytes, B-lymphocytes, macrophages, or natural killer (NK) , are highlighted by the histochemical stains

Previous work may thus be validated by the models developed in this challenge.

More detail on these insights can be found here





Retrospective

Insights

1. There is a need for better histological pre-trained models to aid development of models for tasks to provide a foundation.
2. Imaging from the model can be helpful and useful!

Constraints

1. Small dataset really hampered training
2. ID switch up in Dataiku caused time lost trying to get systems to work again on the old IDs etc. which could've gone into figuring out images.
3. Resources and time was hard to find

What are the next steps?

As a proof of concept, we believe we have shown that not only can pathology sections be learned by machines, but also that histologically pretrained models can potentially improve analyses.

- Apply more rigorous statistical testing with more data to test insights
- We would like to investigate the scans the model have produced further and to a higher degree of accuracy
- More clear data is necessary as this was the largest constraint
- Refining the dataset and further explainability