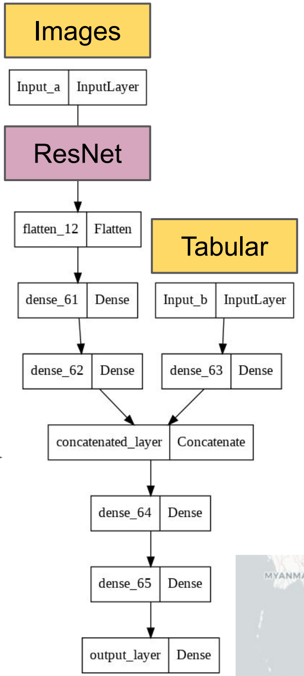
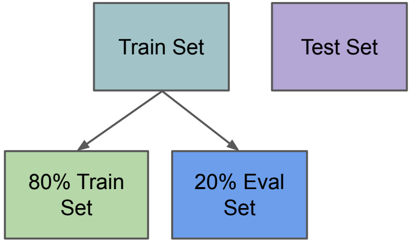
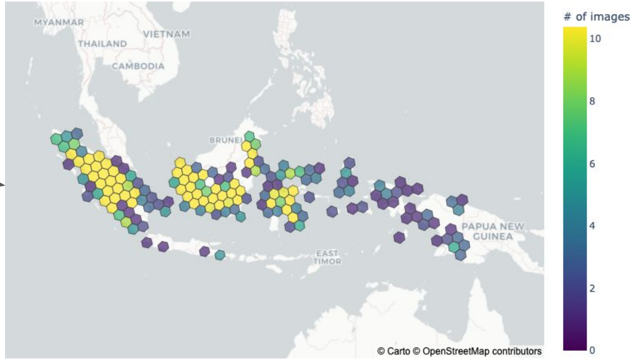
Group: Random For(r)est Gump (DT5). Summary

Idea	Result	Conclusion
Simple data exploration	✅	Data looks good, images from Malaysia, detection of class imbalances
Check F1 score of simple model with tabular data (year, coordinates)	✅	F1 score of around 0.65 with default lightgbm - there is predictive power in the tabular data.
Build CNN network for images and train from scratch (Keras)	😞	Training from scratch (random weights) did not seem to work well, not achieving more than F1 score 0.4.
Build custom Keras model fine-tuning ResNet50 + tabular data concat	🤔	Quite difficult and slow to train - had to be done with a very low learning rate to not have unstable training. F1 score 0.742.
Build FastAI CNN with pretrained resnet50	✅	FastAI gave very nice results, F1 score 0.745.
Test the h3 hexagons of the coords as features	😞	Did not provide any extra predictive power.
Average ensemble of tabular ML models + Pretrained ResNet50	🚀	We got our best results by ensembling the probability predictions of a random forest and the CNN. F1 Score 0.76.



For the experiments:





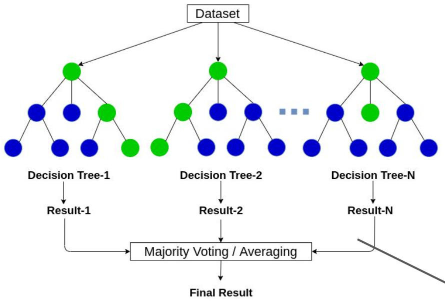
Final model: Ensemble approach

TRAIN DATA

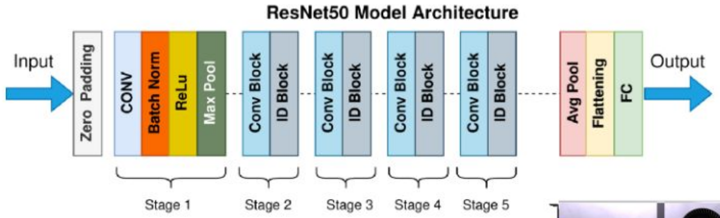
Year, Latitude, Longitude

Images

RandomForest



PreTrained ResNet (on imagenet, fine-tuned)

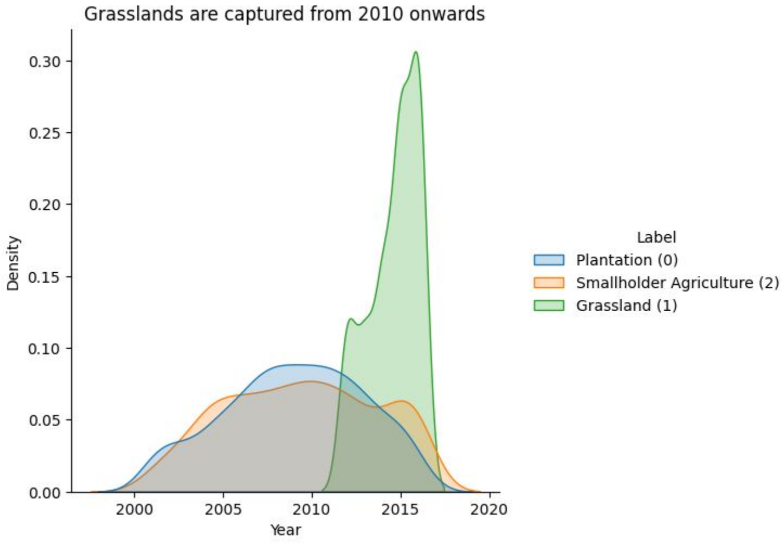


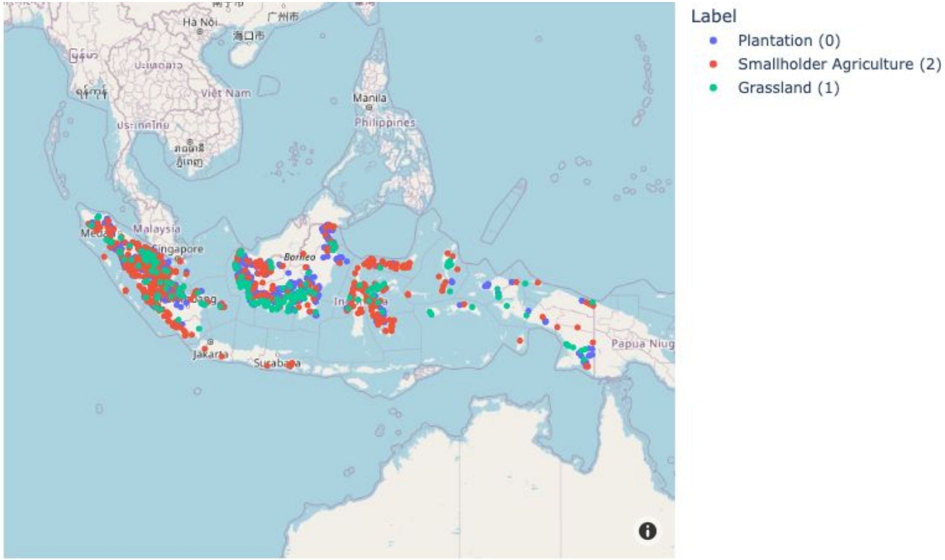
Final predictions = sum(probs)/2

F1 score on 20% evaluation set: 0.759

Final key insights

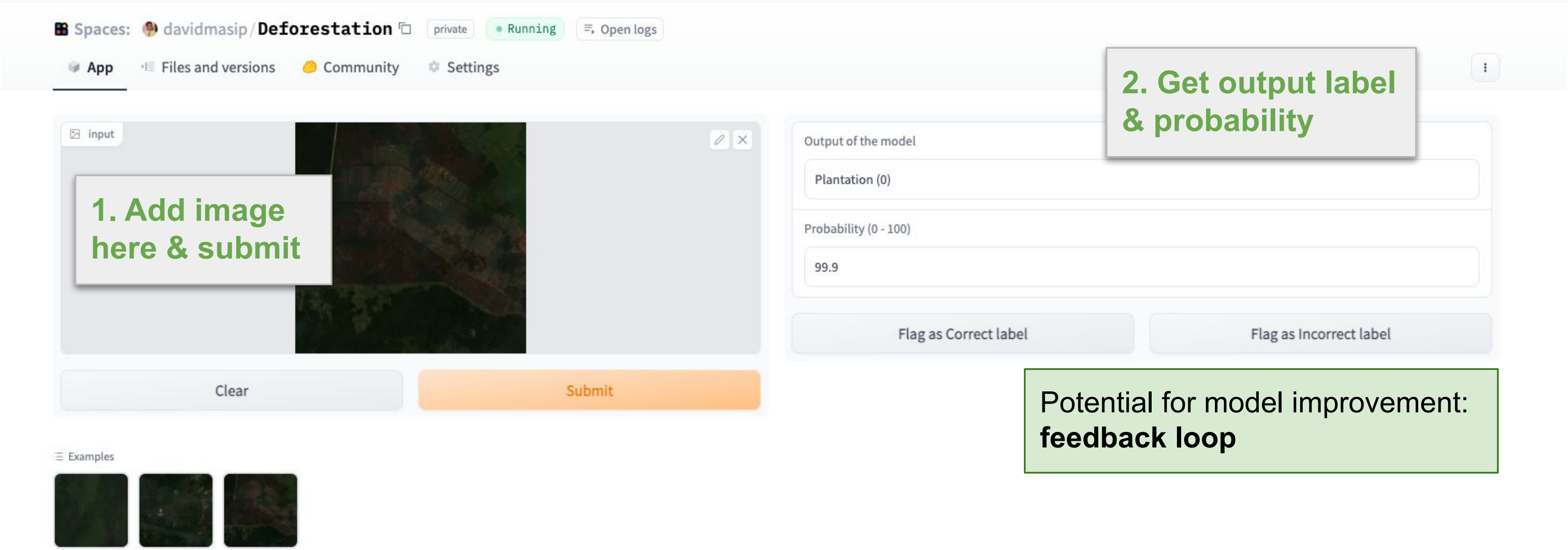
- Since there are very few training samples (1714), **overfitting** is an issue. Data augmentation (image transformations, like flipping, random cropping, etc) and ensembling are critical.
- For the tabular data, **bagging** works better than boosting due to the small size of the dataset.
- Leveraging **pre-trained CNNs** is great to boost performance, be able to iterate fast and get decent results quickly.
- Tabular features matter: year to identify grasslands, each island has their own biodiversity and geography.





NUWE

Live Demo: [Try it yourself here](#) 😃



Potential for model improvement: **feedback loop**

Thank you for organising this challenge, we had fun!

Anne, Miquel & David aka the Random For(r)est Gumpers :)