GeoPredictors Report

1. Classification of geological facies and Data Preprocessing

For lithofacies classification, we utilized the Corebreakout library to process core images, extracting core columns and segmenting them into subsections based on facies labels and depth information. Initial challenges included incomplete core detection and depth annotation inconsistencies. To address these, we segmented cores into smaller sections (e.g., core1 into 1a, 1b, etc.), particularly focusing on problematic areas with abrupt transitions or unreliable depth markers. This method resulted in approximately 11,000 labeled images for our training dataset. We also omitted the "bs" (borehole side) class as it is a subset of the "nc" (no core) class.

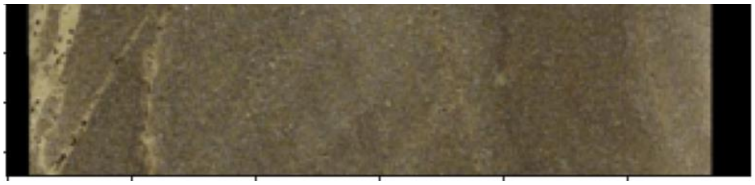

Figure 1: A sample of a sliced core image

Since the task was image classification, our first choice was CNN. Multiple models we tried overfitted and yielded accuracies of no more than 20%. We then chose a pre-trained ResNet34 with custom layers. ResNet34's recognition of complex image patterns and textures, alongside its residual connections that enhance feature extraction without compromising performance, made it an ideal choice. Our model achieved 53% accuracy after training for 30 epochs, although it slightly overfits, and it achieved high accuracy after few training epochs (See Figure below).

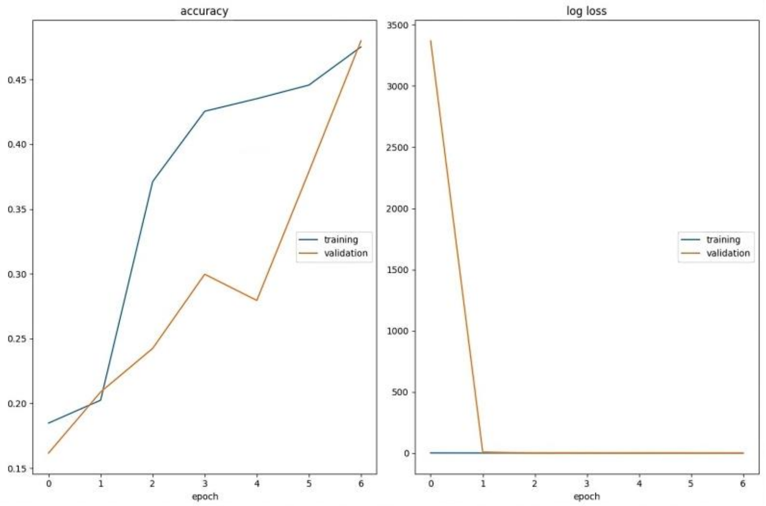

Figure 2: The training curve of the modified ResNet34 after 6 epochs.

During model training, we encountered class imbalance, notably with "nc" images being the most prevalent at 3700, and "ih" the least at around 495. Training on the entire dataset as-is led to a biased 70% accuracy towards "nc" facies. To address this, we balanced the dataset by selecting 495 images from each class, totaling 2970 images for training. This approach, especially with a pretrained model, was deemed sufficient for more accurate and balanced learning.

2. Prediction of Permeability

In predicting permeability, we merged CSV files containing porosity and permeability data with log data, aligning core samples with their corresponding log data at similar depths. Post-EDA and feature selection, we tested ML models such as XGBoost Regressor, Random Forest Regressor and SVR for modelling due to their ability to capture complex and non-linear relationships. We used RMSE and R2 scores for model selection and hyper-parameter analysis to determine the optimal model parameters. The best performing model was the Random Forest Regressor.

To address overfitting, we trained and validated the models on eight of nine wells, reserving one for prediction and iterative parameter adjustments based on R2 and RMSE outcomes. The Random Forest Regressor, yielding the highest R2 Score, was ultimately selected for permeability prediction.

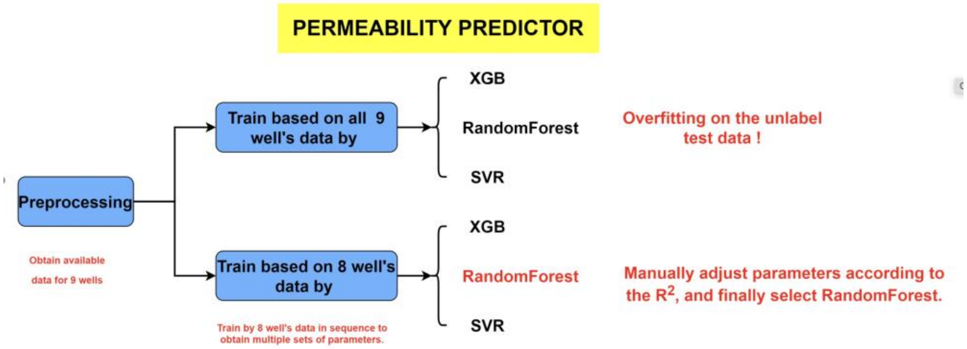
PERMEABILITY PREDICTOR
Preprocessing
Obtain available data for 9 wells
Train based on all 9 well's data by
XGB
RandomForest
SVR
Overfitting on the unlabel test data !
Train based on 8 well's data by
Train by 8 well's data in sequence to obtain multiple sets of parameters.
XGB
RandomForest
SVR
Manually adjust parameters according to the R², and finally select RandomForest.

Figure 3: The training process for the prediction of permeabilities

3. Visualization

Since the data is core data, it made most sense to display the data as wireline data, with the predicted facies being colored and shown next to the core image. (See Figure below)

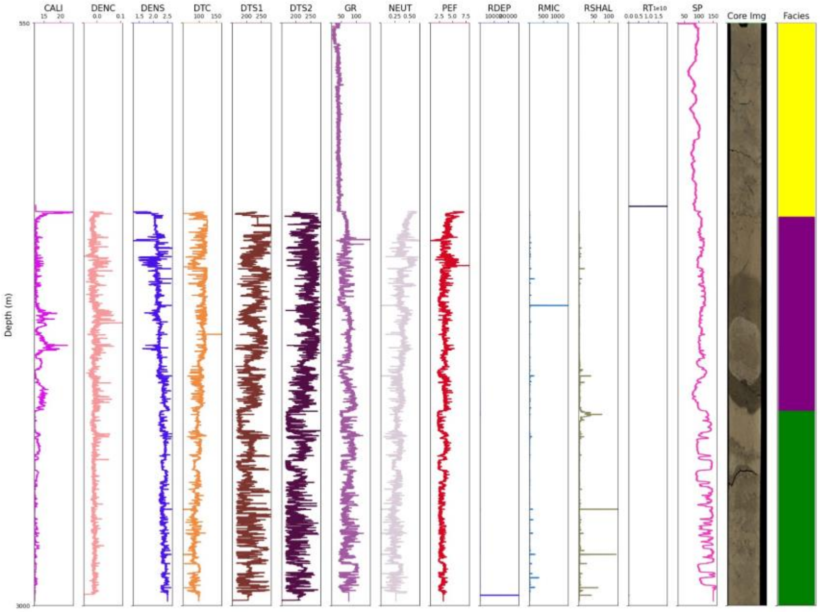

Figure 4: An example output of the visualization tool