**The Use of Machine Learning For the Interpretation of Wellbore Data**

Presenter: Cassandra Goldberg1

Mentors: Roman Y. Makhnenko2, Nikita Bondarenko2, Priyam Mazumdar3

1 NSF REU FoDOMMaT Fellow, Bowdoin College, Brunswick, ME, USA

2 Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA

3 Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA

**Abstract**

The Illinois Basin has an excellent history of pilot- and demonstration-scale CO₂ injection projects, positioning it as a promising potential hub for carbon capture and storage (CCS). Since the presence of fractures increases the risk of an induced seismic response and CO₂ leakage, it is essential that these injection sites are properly characterized before project deployment. To address this challenge, this research project focuses on utilizing machine learning techniques to efficiently and accurately infer the presence of fractures in geological formations.

The data labels were extracted through manual analysis of FMI logs, which provide high-resolution images by measuring rock resistivity. Various machine learning approaches, including a feed-forward neural network were implemented to map petrophysical and mineralogical data to binary classifications indicating the presence of fractures at different depths. The best performing model achieved a 79% test accuracy and 54% test F1 score when trained on well VW1 and tested on well CCS1. Similarly, training on well CCS1 and testing on well VW1 yielded a 75% test accuracy and 50% test F1 score.

Despite the current model’s limitations, data analysis techniques revealed highly correlated features and clustering patterns based on rock formations. These findings suggest that a more systematic approach to feature selection and the incorporation of geological context may enhance feature prediction capabilities. Further research incorporating additional wells and optimizing feature combinations may lead to more accurate predictions, enabling practical applications to geoenergy projects.

**Introduction**

Fractures play a critical role in geoenergy projects because they enable fluids to flow through rock formations. In the context of carbon capture and storage (CCS), fractures can compromise the integrity of storage sites, acting as potential leakage pathways for CO2 (Rutqvist, 2015). Moreover, the presence of pre-existing fractures is associated with a higher risk of induced seismic activity (Kivi et al., 2023). Detecting fractures with characteristic size smaller than ten meters proves challenging due to the low-resolution of seismic surveys. Since fractures of this size can still impact the processes happening in the subsurface, there is a pressing need for more advanced fracture detection and prediction techniques.

Various petrophysical and mineralogical properties of subsurface formations are interdependent and affected by the presence of fractures, though their precise relationships remain largely unknown. Machine learning algorithms provide a promising framework for analyzing these intricate relationships, as they do not rely on explicit equations to make predictions. Moreover, ML eliminates the need for subjective interpretation and allows for real-time predictions, which is advantageous for fracture analysis during the project implementation.

Due to its successful history of CO₂ injection projects, the Illinois Basin has been positioned as a promising location for larger CCS endeavors. Reliable fracture prediction techniques that operate quickly and at a regional-scale will play an invaluable role in guaranteeing the safety of various injection site locations. This research project aims to explore machine learning techniques that can be leveraged to improve the quality of geophysical wellbore data interpretation for the upcoming CO2 injection projects in the Illinois Basin.

**Methodology**

Machine learning requires high-quality input data and ground truths for effective training and performance evaluation. The input data for this project was collected from one-dimensional wellbore logs that includes various petrophysical and mineralogical features. Geophysical measurements were recorded every half foot from approximately 5,000-7,000 ft below the surface, across four deep wells. This dataset was retrieved from an Energy Data eXchange open repository.

In order to generate the label set, Formation MicroImager (FMI) logs were examined manually. FMI sensors operate by measuring the electrical resistivity of rock formations along the wellbore as a function of azimuthal angle. Since most of the fractures contain water, which lowers resistivity of rock, azimuthal resistivity measurements enable the production of high-resolution images of high water content zones.

Various machine learning models were explored including Logistic Regression, K-Nearest Neighbors, Random Forest, XGBoost, Feed-Forward Neural Networks, and Long Short-Term Neural Networks. Implementation was carried out using PyTorch and Scikit-Learn packages, with each model designed to map geological input data to binary classifications indicating the presence of fractures at different depths. To assess and compare the models’ performance, visualization plots and metrics such as accuracy and F1 scores were employed.

To ensure the quality and interpretability of the results, data exploration techniques were applied. High multicollinearity in the dataset could lead to overfitting and reduced computational efficiency. To address this, correlation strength between each variable pair was assessed, visually represented through correlation heatmaps. Additionally, variance inflation factors (VIFs) were calculated, which is a quantitative measure indicating how correlated a variable is with all of the other variables. A random search algorithm was then implemented to optimize the feature set, ensuring VIF scores below 5 while maintaining a maximal amount of features.

T-Distributed Stochastic Neighbor Embedding (TSNE) was used to visualize clustering patterns in the input log data. TSNE is an optimized dimension reduction technique that aims to maintain the distances and relationships between neighboring points in the high-dimensional space in two-dimensional space.

Furthermore, XGBoost’s feature importance capabilities were used to identify key features related to fracture presence. The XGBoost generates feature importance scores by aggregating information gain from each feature across decision trees and considering the weighted frequency of feature splits.

**Machine Learning Model Results**

The test results for models trained on well VW1 and tested on well CCS1 are presented in Figure 1. The plots show the known depths of fractures (indicated by a gray line) and the predicted fracture depths (indicated by blue line). Additionally, the test accuracy and F1 scores for each model are provided. Among the models evaluated, the two neural network models (feed-forward and long short-term memory) achieved the best F1 scores, with the feed-forward neural network outperforming the long short-term memory network by a small margin.

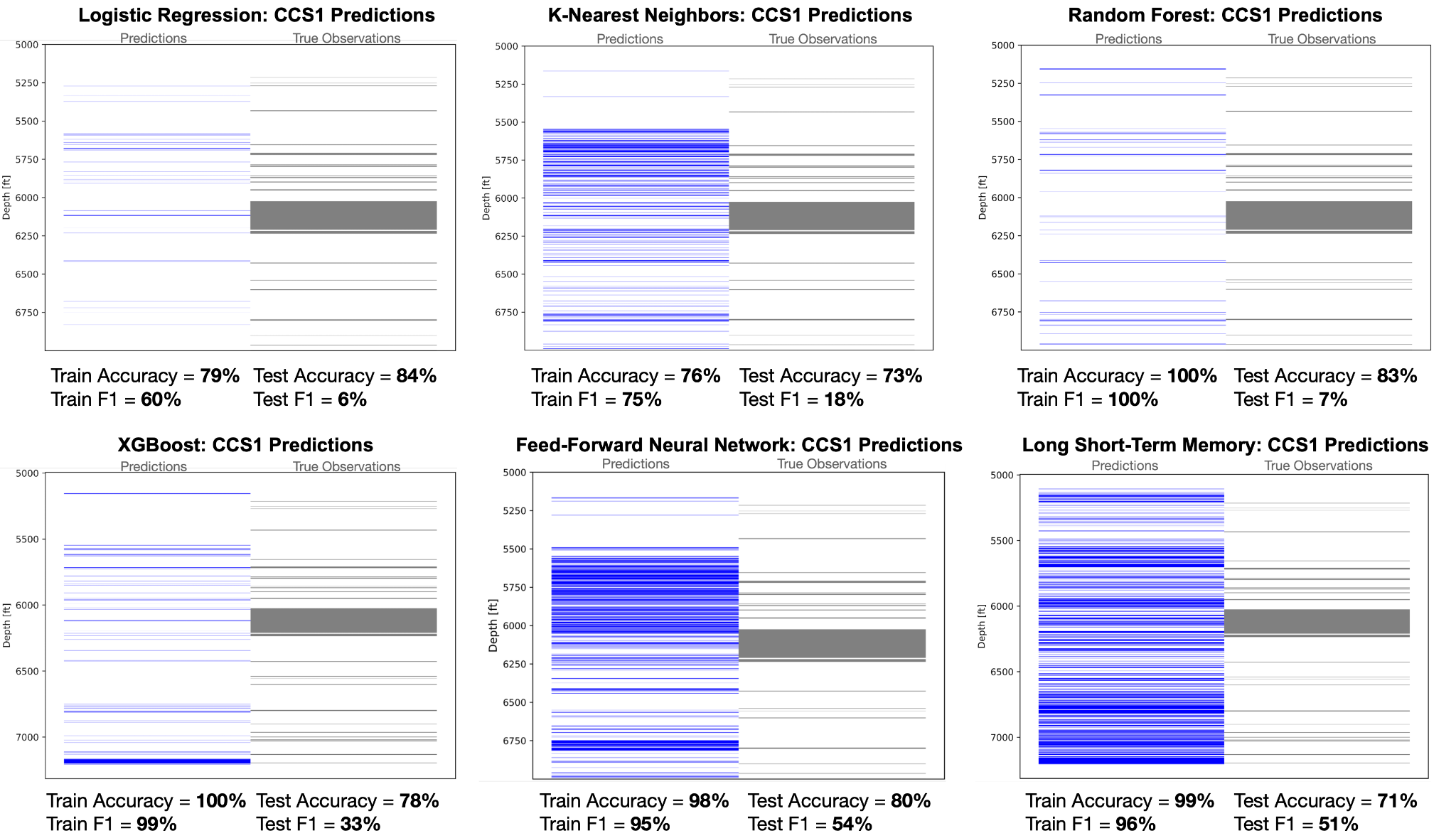


Figure 1: ML Model Results For Fracture Prediction of Well VW1

The best performing model, the feed-forward neural network, was designed two hidden layers containing 32 and 16 neurons, respectively. It utilized a ReLU activation function, a mean squared error loss function, the Adam optimizer, and a learning rate of 0.001. Dropout and batch normalization was also employed during training. The model underwent 500 epochs with a batch size of 50 samples. The model’s performance remains relatively consistent when trained on CCS1 and tested on VW1, as shown in Figure 2.

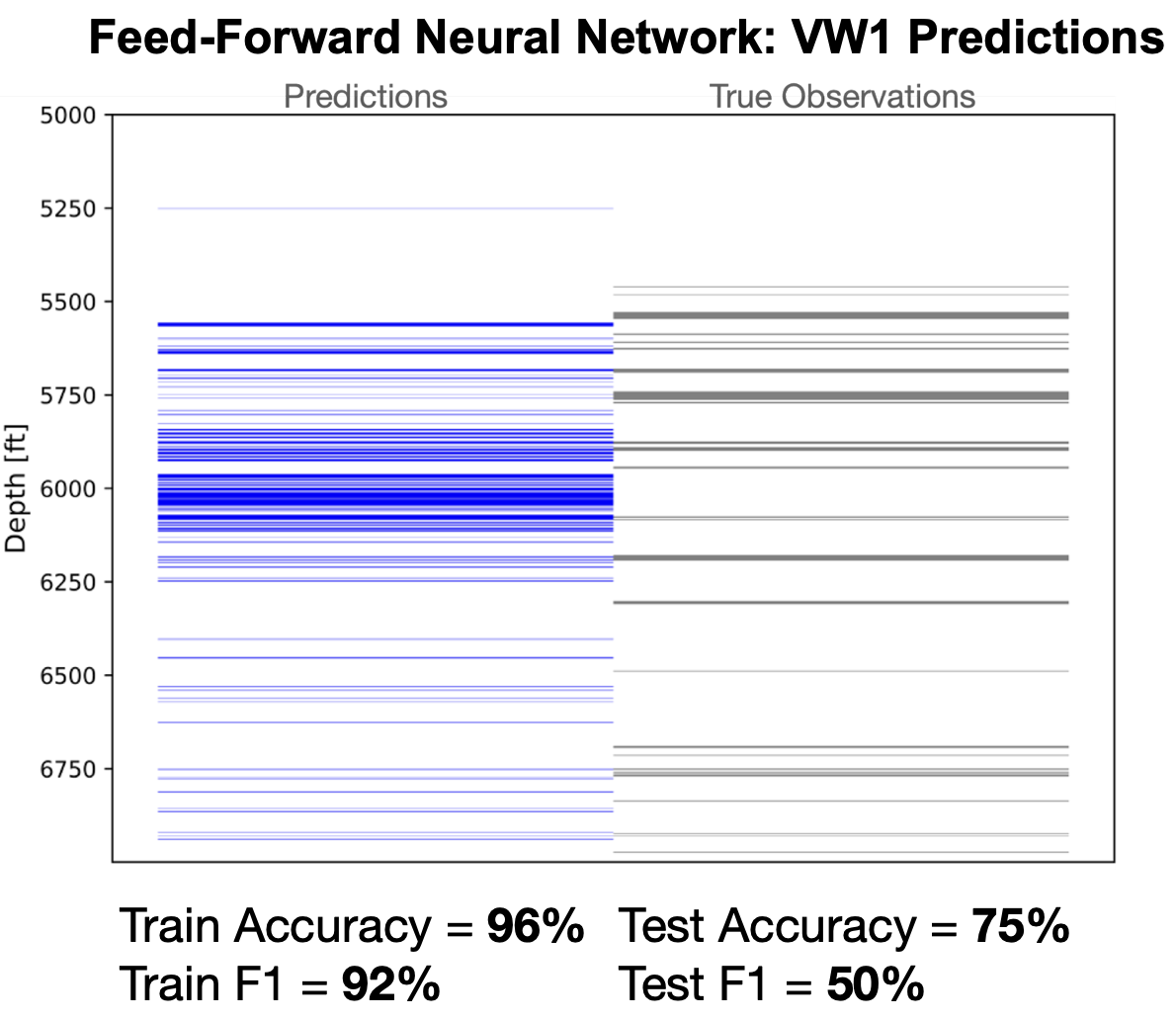


Figure 2: Feed-Forward Neural Network Fracture Prediction For Well VW1

**Data Exploration Results**

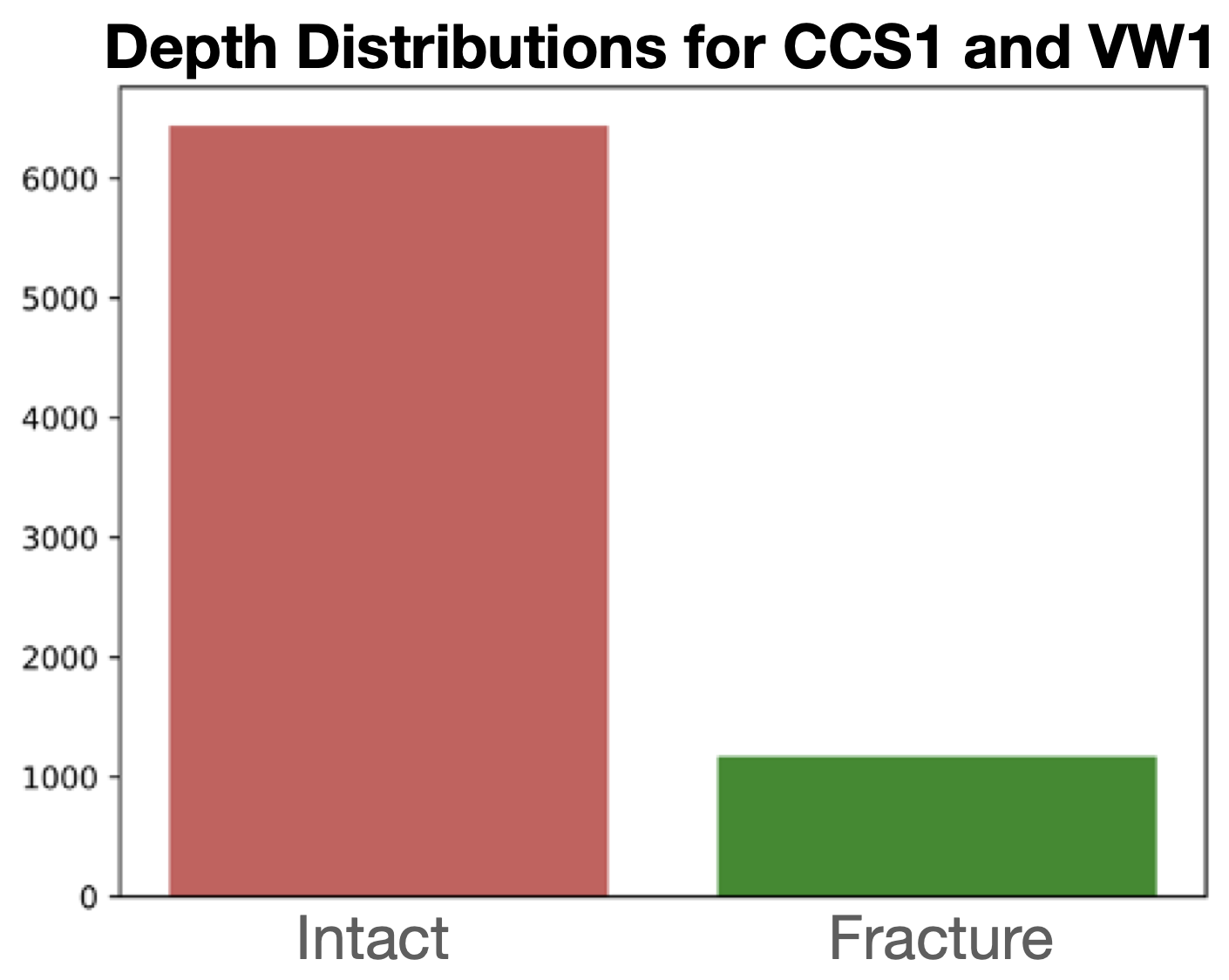
The dataset is imbalanced, with only about 10% of the recorded depths contain fractures (Figure 3).

Figure 3: Depth Distribution of Intact vs Fractured Depths

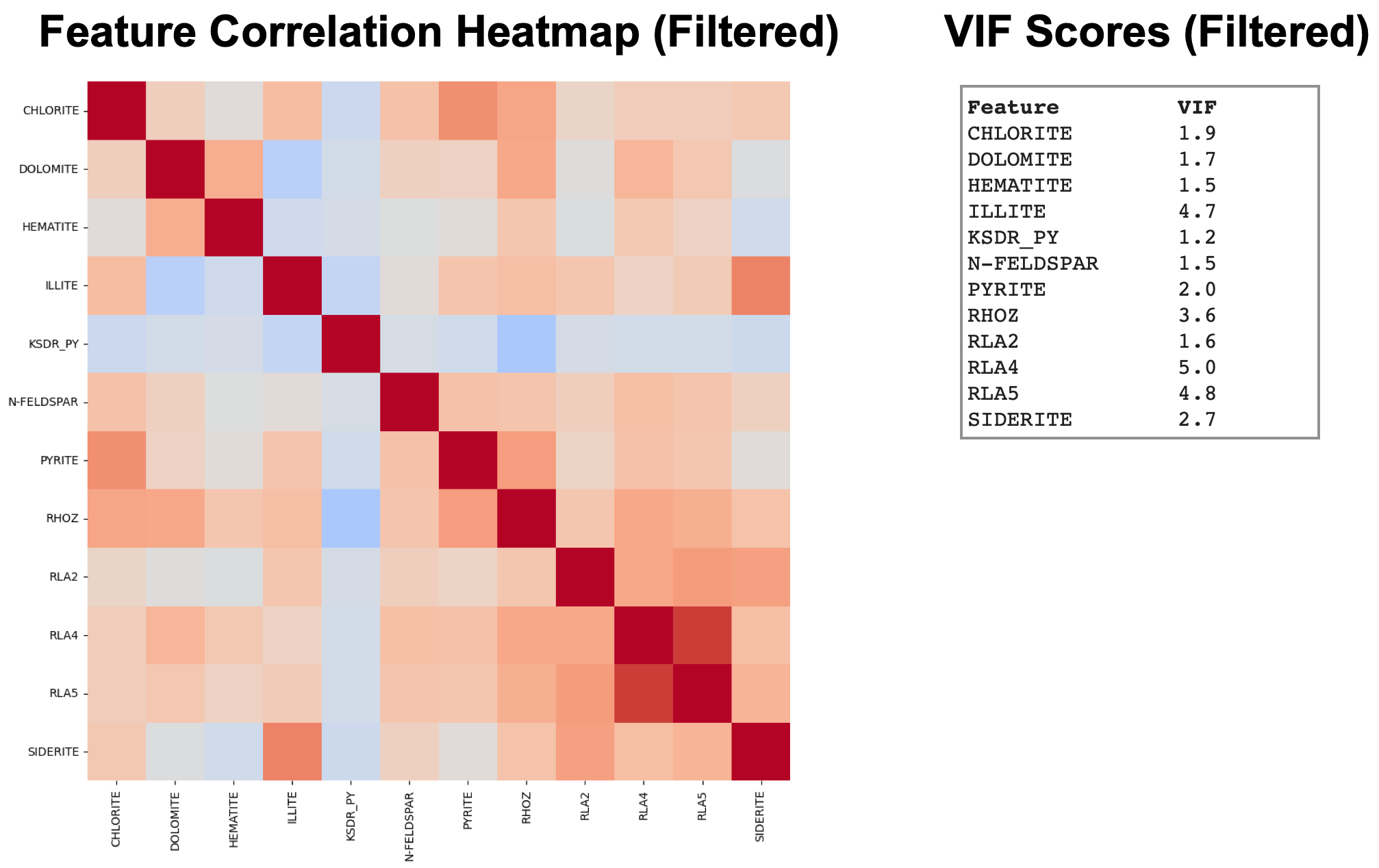
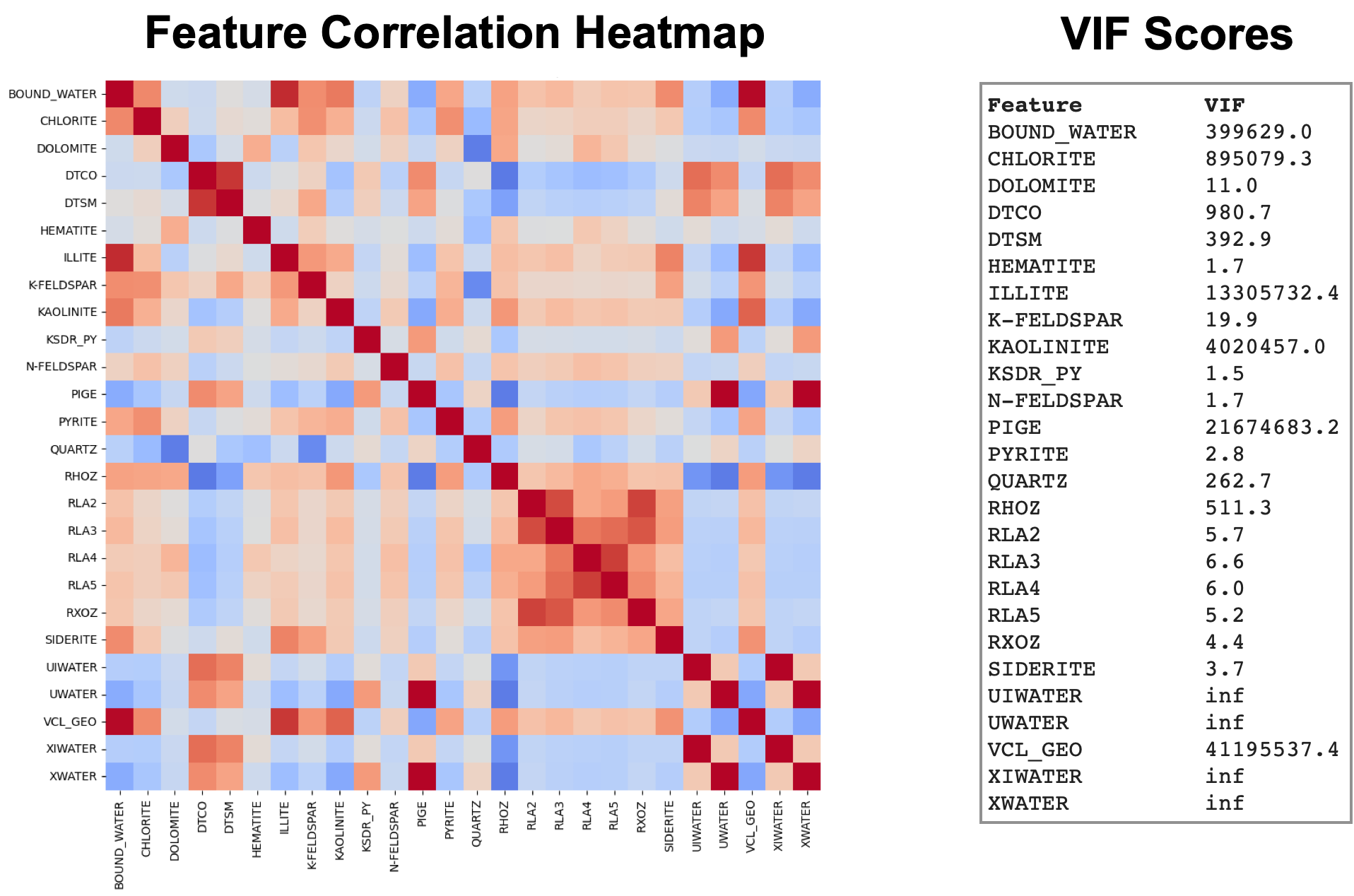
Also, significant correlations were detected among many of the petrophysical and mineralogical features considered in the models. The correlations were examined in the raw data, as well as in the filtered features after applying a random search to reduce multicollinearity (Figure 4). 

Figure 4: Feature Correlation in Raw and Filtered Dataset

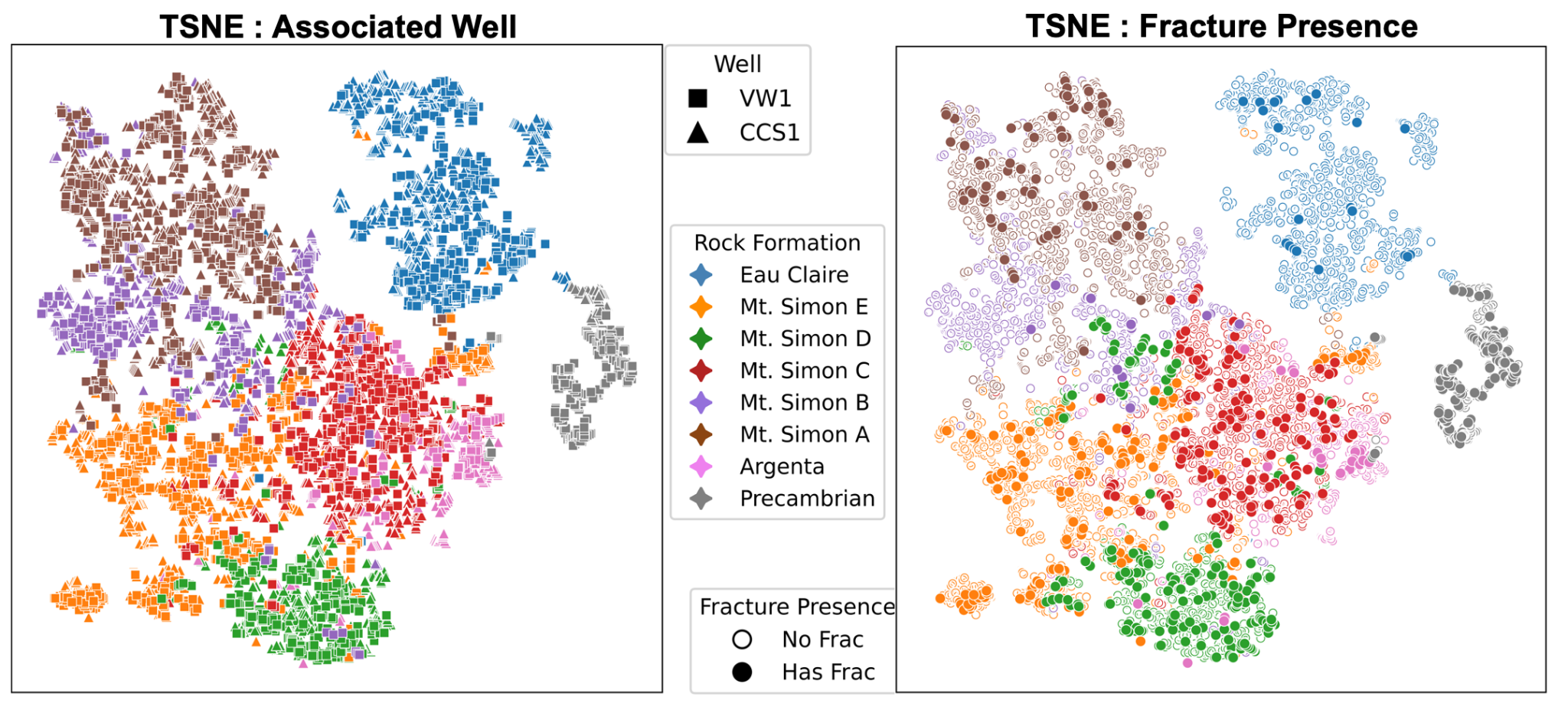
The TSNE analysis reveals that the data exhibits clustering primarily based on rock formation type, rather than being clustered according to the associated well or the presence of fractures (Figure 5).

Figure 5: TSNE Analysis of Clustering in the Dataset

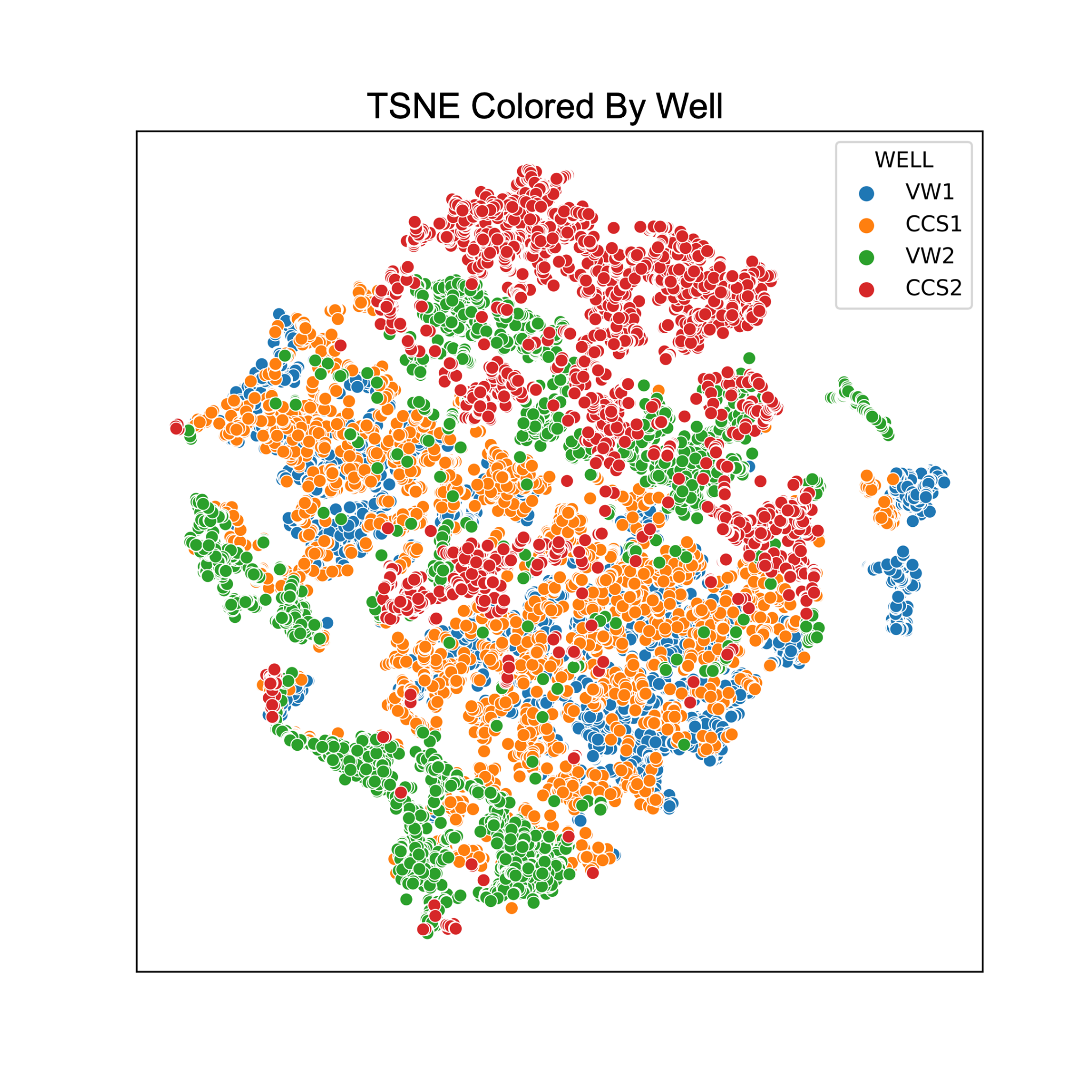
The TSNE analysis displayed in Figure 6 also showed that the VW1 and CCS1 data has a lot of overlap and therefore appears to have similar data composition. However, wells VW2 and CCS2 are distanced from all other considered wells, indicating dissimilarity. 

Figure 6: Dissimilarity Of Wells VW1 and VW2

Even though Figure 5 may suggest similarities between CCS1 and VW1 data, the importance scores generated by the XGBoost model reveal that the wells have different most important features for predicting fractures.

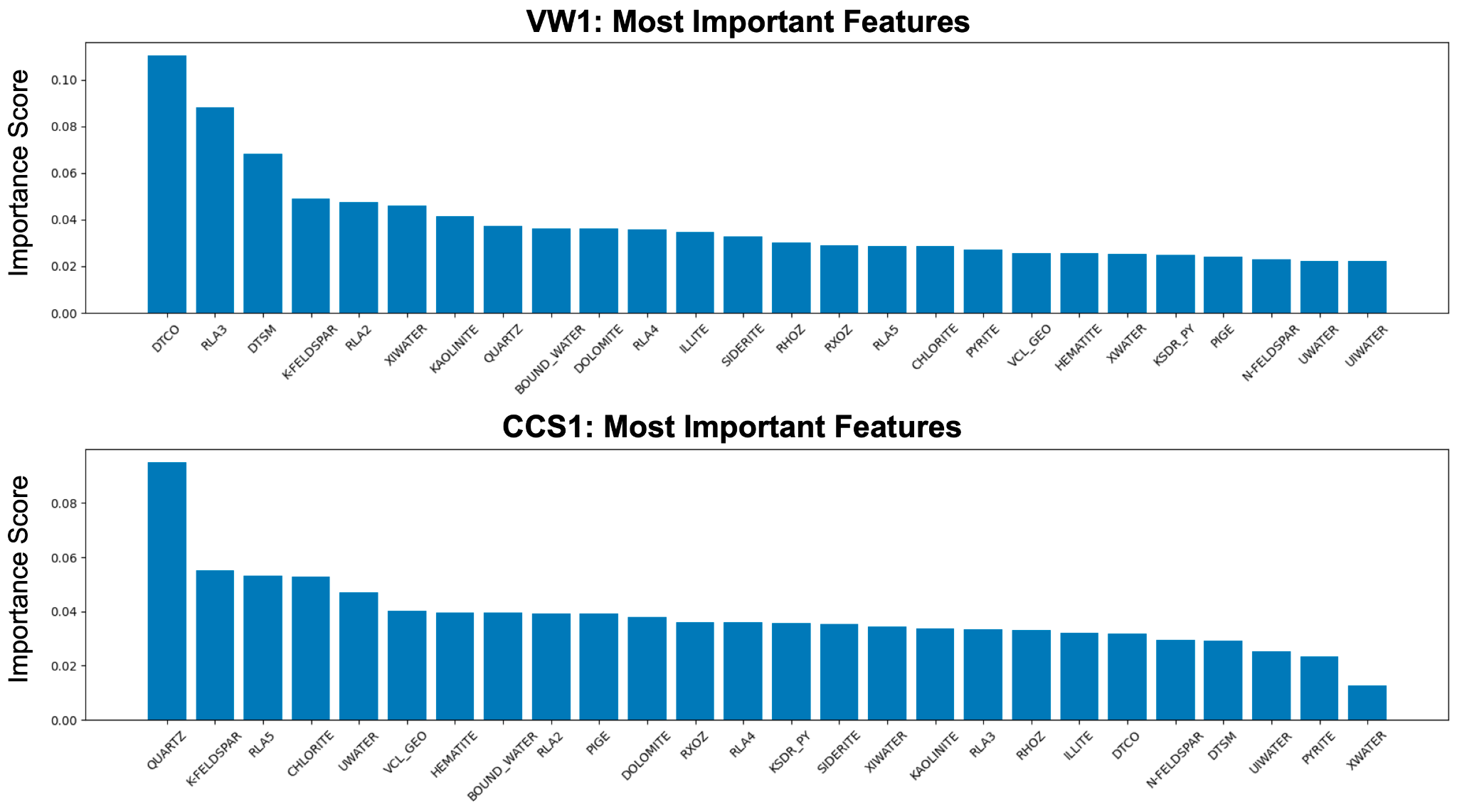


Figure 7: Comparative Feature Importance Generated by XGBoost

**Discussion**

The feed-forward neural network, as depicted in Figure 1, produced the most promising results among the models evaluated. However, despite achieving relatively high test accuracies from around 75% to 80%, this level of accuracy is not a sufficient metric for practical fracture detection in binary classification tasks, particularly given the dataset's inherent imbalance (as evident in Figure 3). When a model performs well on the majority class but poorly on the minority class, it can result in misleadingly high accuracies. To address this, F1 scores, which are less affected by class distribution, were employed for a more meaningful evaluation. The model's F1 scores were around 50%, which is relatively low. The plots in Figure 1 and 2 reaffirm this evaluation, clearly revealing that the model’s predictions do not align well with the labels. Moreover, the notably high training results suggest a potential issue of overfitting, where the model is fitting the noise of the training data and therefore cannot generalize to unseen wells.

Despite the current models’ inaccuracies, delving deeper into the data exposed valuable insights related to the composition of the dataset and how geophysical factors are related to fracture presence. Though the current models yield inaccurate predictions, further exploration of the dataset revealed that it may be possible to extract useful insights related to the presence of fractures from geophysical wellbore data. For instance, the Figure 4 analysis demonstrates that many features within the current dataset exhibit high correlations. Addressing this redundancy by removing highly correlated features, particularly those with higher VIF scores, can lead to a more refined dataset that retains all essential information.

Furthermore, the T-SNE plots in Figure 5 reveal that the current dataset’s variations are more strongly associated with rock formation types rather than specific wells or the presence of fractures. It is encouraging that the data is not clustered by wells, as the overarching objective of the project is to develop wellbore data interpretation tools that can be broadly applied to different wells. This also indicates that incorporating information about the rock formations can enhance a model’s prediction capabilities.

The feature importance analysis also provided promising findings, indicating that the XGBoost model is able to identify which features are most related to fracture presence based on the current dataset. For well VW1, factors such as compressional/shear velocity (DTCO/DTSM) and resistivity (RLA3), which are known to be significantly correlated to the presence of fractures, have the highest importance scores. For well CCS1, the highest importance scores were assigned to mineralogical features, which are believed to be less directly related to the presence of fractures. This discrepancy suggests that certain wells may serve as more reliable data sources, while others may encode information that potentially conflicts with established physics knowledge, impeding generalization to new wells.

**Conclusion and Recommendations**

Overall, while the current findings suggest that machine learning models may not be able to effectively predict fractures based on the current dataset, further exploration of the dataset reveals promising avenues to enhance fracture prediction capabilities.

To improve prediction performance, it would be beneficial to implement a more systematic and experimental approach for selecting input features. While the current feature selection is based on domain knowledge, adopting statistical approaches to identify the most optimal feature combinations could yield more accurate results. Additionally, addressing feature redundancy by eliminating highly correlated variables could lead to improved model performance.

Furthermore, incorporating additional wells would allow the model to train on a larger, more comprehensive dataset and help mitigate its overfitting issues. As evidenced by the clustering patterns based on rock formation types, the inclusion of geological context alongside geophysical data has the potential to bolster fracture prediction abilities. This integration may uncover uncharacterized relationships and enhance the model’s ability to generalize across different wells.

**References**

Greenberg S, Whittaker S, and McDonald S. On the Path to Commercial CCS: Scaling from Field Demonstration to Regional Hub. In Proceedings of 14th Greenhouse Gas Control Technologies Conference Melbourne (2018).

Kivi, I. R., Boyet, A., Wu, H., Walter, L., Hanson-Hedgecock, S., Parisio, F., and Vilarrasa, V. Global physics-based database of injection-induced seismicity, Earth Syst. Sci. Data Discuss. preprint in review, (2023).

Rutqvist, J. Fractured rock stress-permeability relationship from in situ data and effects of temperature and chemical-mechanical couplings. Geofluids, 15(1–2), 48–66 (2015).