King Fahd University of Petroleum & Minerals

COE 292: Introduction to Artificial Intelligence

Final Report

Project Name:	Lithology Classification Based on Geophysical Logs
Group Number:	1
Date:	12/04/25
Number of students in the group:	4

Student 1:

Name	Abdullah Deya Al-Qisoom
KFUPM ID	202156590
Department	CPG (Geoscience)

Student 2:

Name	Salman Ali Alameer
KFUPM ID	202184310
Department	CPG (Geoscience)

Student 3:

Name	Mohammed Abdelkarim Alqadhib
KFUPM ID	202223600
Department	CPG Petroleum

Student 4:

Name	Mohammed Ahmed Aljafar
KFUPM ID	202260400
Department	Bioengineering

Student 5:

Name	
KFUPM ID	
Department	

Summary of Classification problem

The classification problem involves identifying different rock types (lithology/facies) using well log measurements, which is critical in geoscience and petroleum engineering for understanding subsurface rock properties, reservoir characterization, and well planning. Features like gamma ray, resistivity, and sonic logs are used, but their interrelationships are complex. The challenge lies in the heterogeneous nature of rock formations, their anisotropy (properties changing with measurement direction), and overlapping properties between different facies. A single log's response isn't unique to one lithology, so relying on one log can lead to misinterpretation. Accurate classification is essential for decision-making, such as predicting drilling time in loose sand versus hard consolidated sand or targeting shale in unconventional drilling.

Dataset manipulation

No.	Question	Student Response
1.	How many labeled examples are in your data set?	3232
2.	How many distinct features are in your data set?	7
3.	How many distinct labels are in your data set?	9
4.	For each label, what is the percentage of data?	Class distribution: 1- 8.013614 % 2- 22.834158 % 3- 19.028465 % 4- 5.693069 % 5- 6.714109 % 6- 14.294554 % 7- 3.032178 % 8- 15.408416 % 9- 4.981436 %
5.	Is the dataset balanced based on the above?	[ ] Yes [○] No If No then explain why: class distributions vary significantly, with some classes having much higher representation (Class 2 at 22.83%) while others have very low representation (Class 7 at 3.03%). A balanced dataset should have roughly equal distribution among classes.
6.	Is the dataset related to the major of any group member?	[○] Yes [ ] No If No then explain why:
7.	Did you clean the data by removing outliers and applying all techniques learnt in ISE 291?	[○] Yes [ ] No If No then explain why:

Feature (variable) manipulation

No.	Feature used in file	Short Feature explanation/description
1.	GR (Gamma Ray)	Measures natural radioactivity in formation. Natural occurring radioactive materials (NORM) include the elements uranium, thorium, potassium, radium, and radon.
2.	RELPOS	Relative position
3.	Depth	The depth of lithology
4.	PE	Photoelectric effect log

5.	DeltaPHI	Phi is a porosity index in petrophysics
6.	PNHIND	Average of neutron and density log
7.	NM_M	Nonmarine-marine indicator
8.	ILD_log10	Resistivity measurement
9.		
10.		

Label explanation

No.	Feature used in file	Short Feature explanation/description
1.	**SS**	Nonmarine sandstone
2.	**CSiS**	Nonmarine coarse siltstone
3.	**FSiS**	Nonmarine fine siltstone
4.	**SiSH**	Marine siltstone and shale
5.	**MS**	Mudstone (limestone)
6.	**WS**	Wackestone (limestone)
7.	**D**	Dolomite
8.	**PS**	Packstone-grainstone (limestone)
9.	**BS**	Phylloid-algal bafflestone (limestone)

Data visualization

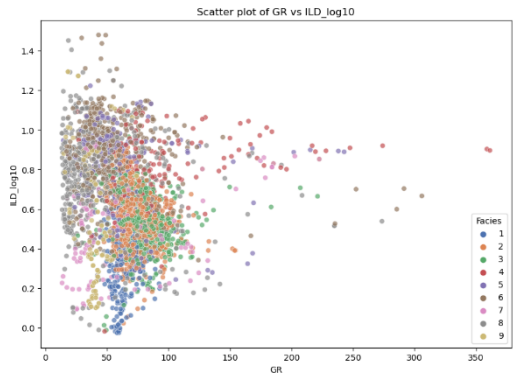


Figure 1

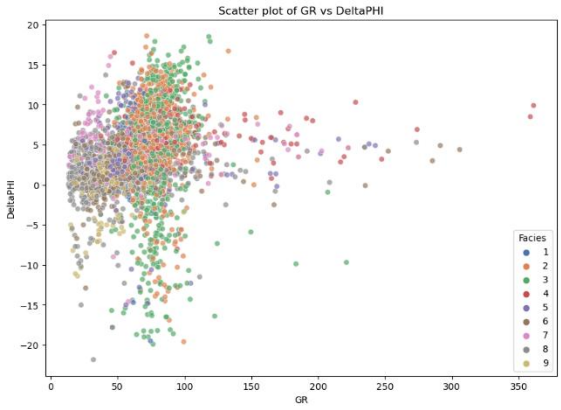


Figure 2

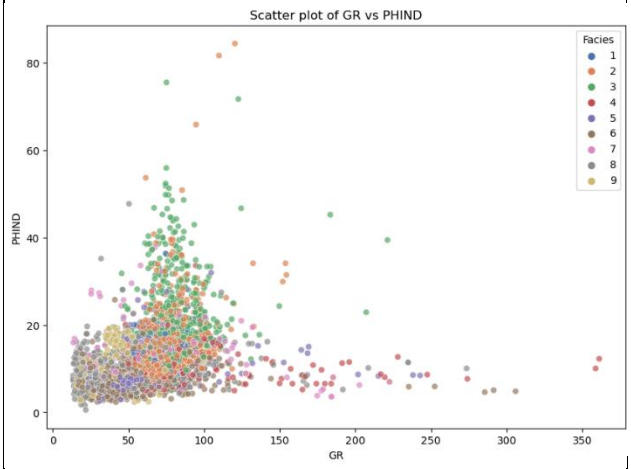


Figure 3

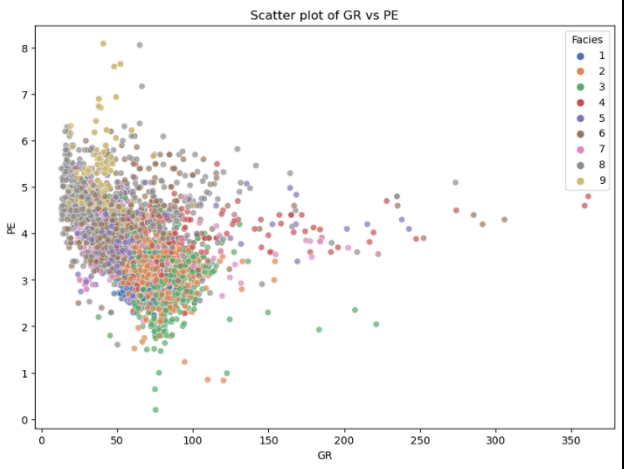


Figure 4

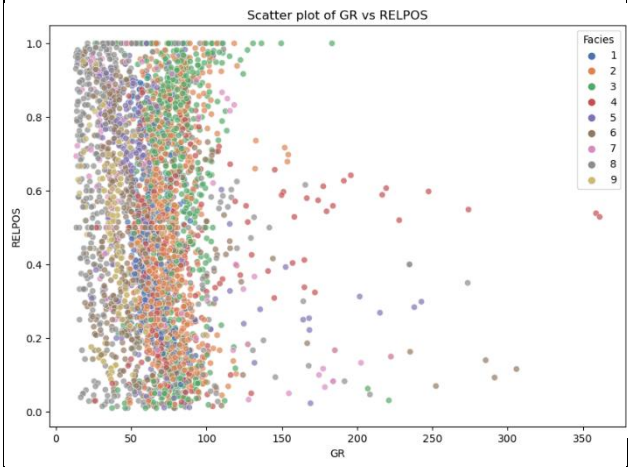


Figure 5

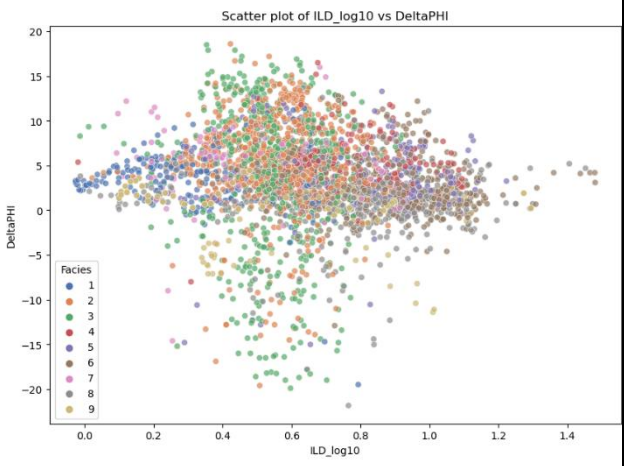
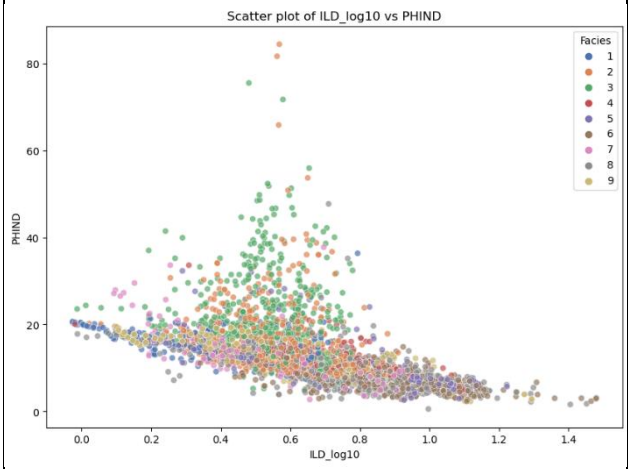


Figure 6



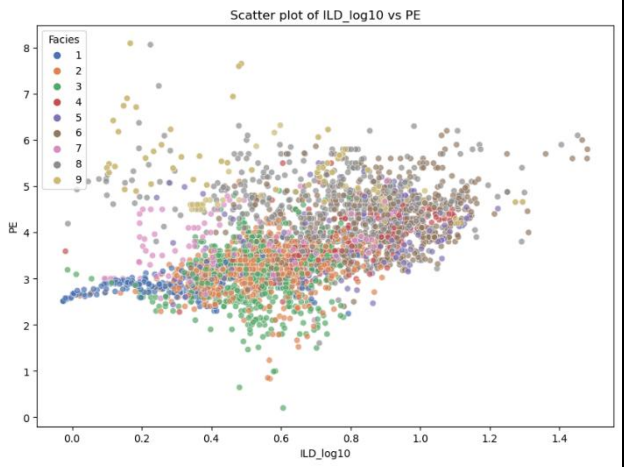


Figure 7

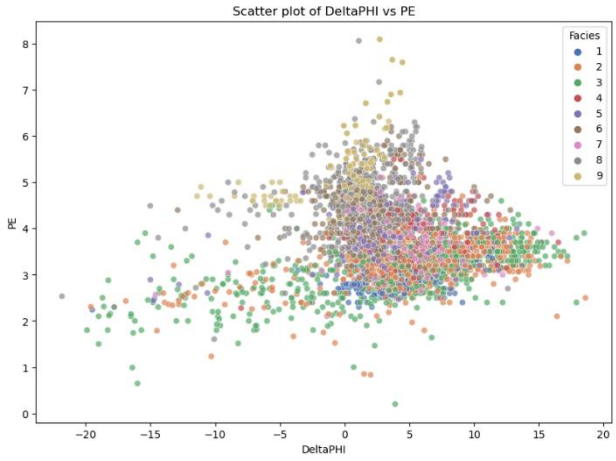


Figure 9

Figure 8

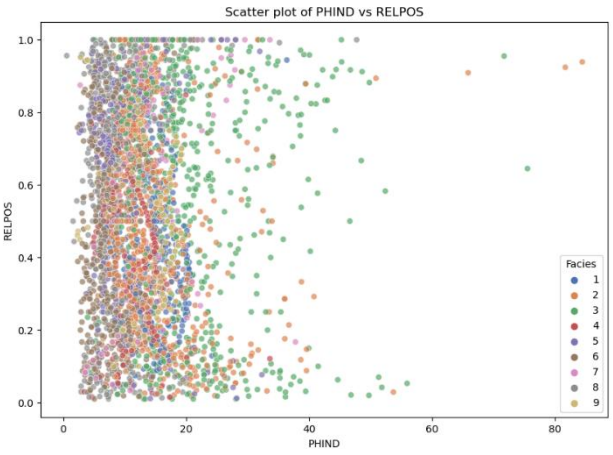


Figure 10

K-NN Algorithm

Dataset Preparation and Feature Scaling

We have used the Euclidean distance method as it is more representative of actual distance than Manhattan distance. All data was scaled using standardscaler which means they are between 0 and 1 thus insuring no single feature is affecting results disproportionately.

Choosing the Right Value of K

We have tested k values 1 to 20 to check for the best accuracy possible, we have found that 1 provides the maximum accuracy; however, it is not appropriate since the graph shows a huge dip in accuracy at k=2 while k=3 is the second highest accuracy with every point after it following a trend downwards

Model Performance

The KNN model achieved an accuracy of 71.8%, with precision and recall around 72%. These are the weighted averages, but performance varied across classes due to dataset imbalance, indicating a need for improvement in minority class prediction.

Cross-Validation

Cross-validation accuracy was 71.2% with a standard deviation of 2.6%, which is acceptable. We used 5-fold cross-validation as it provided stable and consistent results without overfitting.

SVM Algorithm

Dataset Preparation and Feature Scaling

Dataset Preparation and Feature Scaling , StandardScaler was used to normalize inputs to values between 0 and 1. Data was also separated into 70% training, 30% testing and validation (15/15).

Support vectors

Support vectors are the points closest to the boundary and they are the ones that decide where the boundary is. In the soft margin, we had 1770 support vectors, while in the hard margin it was 1507. In the

hard margin case, the support vectors were tighter and gave more accurate predictions than soft margin one.

Kernel Functions

We used the radial basis function kernel (rbf) aka gaussian kernel since our data is nonlinear and not linearly separable and it is the usual choice for data of this size, also it has two parameters that can be easily manipulated to allow for easier fine tuning of the model.

Model Performance

The SVM model got an accuracy of 69%, a precision of 69%, an f1-score of 69% and recall of 69%. While these are the weighted averages, some of the features has deviated from the average substantially to warrant a more specialized model.

Cross-Validation

Accuracy of cross validation was 65% with standard deviation of 3% is acceptable, used k-fold cross validation with 5 folds as it provided the best results.

Deep Learning/CNN Algorithm

Dataset Preparation and Feature Scaling

All data was scaled using Standard Scaler, ensuring a mean of 0 and a standard deviation of 1. This helps the deep learning model converge faster and prevents any single feature from dominating the learning process due to large variations in scale.

Network Architecture Design

The neural network has 7 input features, two hidden layers with 64 neurons each using ReLU activation, and an output layer with 9 neurons using softmax activation. This structure was chosen to balance learning complexity and avoid overfitting, as more layers or neurons could increase model capacity but risk slower training and overfitting.

Activation Functions

We used ReLU activation in the hidden layers because it is simple, fast, and helps avoid the vanishing gradient problem, improving learning speed. Softmax activation was used in the output layer to convert the outputs into probabilities for multiclass classification. A batch size of 32 was chosen to balance training stability and speed. Smaller batches can make training noisier but help generalization, while very large batches can make the model overfit and slow down convergence.

Hyperparameter Tuning

Discuss how different learning rates can affect the convergence of the model.

CNN

ReLU activation was used in the hidden layers to improve learning speed and avoid the vanishing gradient problem, while softmax activation was used in the output layer for multiclass probability output. A batch size of 32 was selected to balance training stability, speed, and generalization.

Model Performance

The deep learning model achieved an accuracy of 64.8%, with balanced precision and recall. Performance varied across classes, with more false negatives in minority facies, indicating the model could improve with more data or further tuning.

Cross-Validation

Cross-validation accuracy was 71.2% with a standard deviation of 2.6%, which is acceptable. We used 5-fold cross-validation as it provided stable and consistent results.

Comparison between KNN, SVM and deep NN/CNN.

KNN achieved the highest test accuracy 74.2%, followed by SVM 71.2% and DNN (64.8%). Overall, KNN performed best because it works well with small, structured datasets. SVM was close and can perform better in cases with clear class boundaries. DNN performed the worst because deep models usually need large datasets to work well. In special cases with more complex patterns and more data, DNN could perform better.

Conclusion

Lithology classification using geophysical logs is essential for geoscience and petroleum engineering as it aids in subsurface characterization, reservoir evaluation, and well planning. An accurate classification model helps geoscientists and engineers make informed decisions regarding drilling operations, hydrocarbon potential assessment, and formation evaluation. By leveraging machine learning techniques, we can improve prediction accuracy, reducing uncertainties associated with traditional lithology identification methods. This advancement enhances efficiency in exploration and production, ultimately optimizing resource extraction and minimizing operational risks. The ability to classify lithology reliably also supports better reservoir management, leading to improved hydrocarbon recovery and cost-effectiveness.