Predicting Subsurface Lithology: A Deep Learning Approach on the FORCE 2020 Dataset

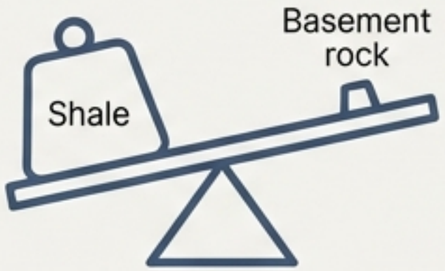
A concise technical summary of the data exploration, modeling pipeline, and performance results.

The FORCE 2020 Dataset Presents Three Core Technical Challenges

>1.1M

Data points from 98 distinct wells, requiring a scalable and efficient pipeline.

~7,000:1



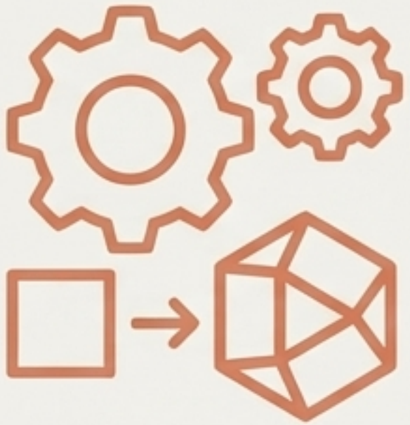
Extreme class imbalance, with Shale (61.6%) being nearly 7,000 times more frequent than Basement rock (0.01%).

~30%

Pervasive missing data, with critical logs like SGR and DTS having over 85% null values.

Any successful model must strategically address data scale, extreme imbalance, and widespread missingness.

A Three-Pillar Strategy to Overcome the Data Hurdles

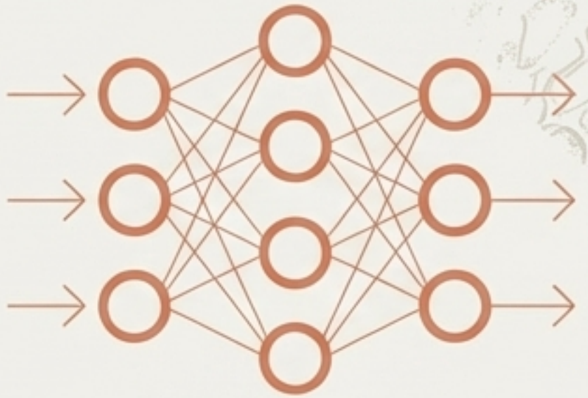


Stage 1

Create high-value petrophysical and spatial features to inject domain knowledge and contextual signal.

Stage 2

Build a robust pipeline to impute missing values (mean), encode categoricals, and standardize all features for the neural network.



Stage 3

Design a deep Multi-Layer Perceptron (MLP) with Dropout and Batch Normalization to learn complex patterns while preventing overfitting.

Engineering Signal from Petrophysics and Spatial Context

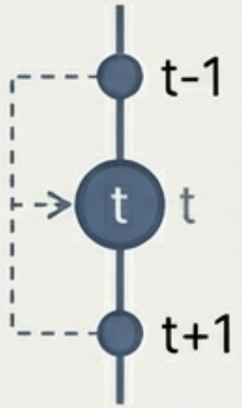
Domain-Specific Features

Log Transforms (e.g., RDEP_LOG)

Domain Ratios (e.g., GR / RHOB)

Polynomial Features (e.g., PEF_SQ)

Vertical Context Features



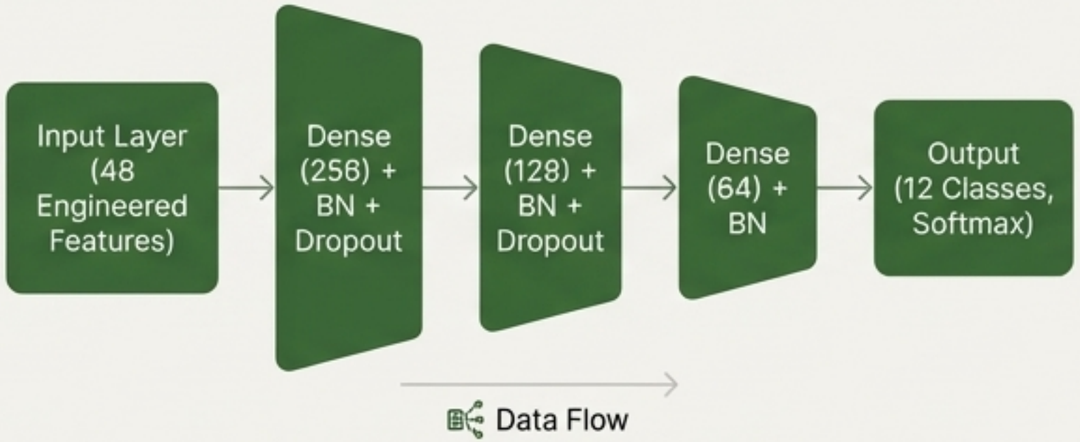
Lag/Lead Features (e.g., GR_prev_1, GR_next_1)

Local Gradient (e.g., GR_grad)

This step enriches the dataset by capturing physical relationships and vertical geological trends critical for accurate interpretation.

A 3-Layer MLP Trained with 10-Fold Stratified Cross-Validation

Model Architecture



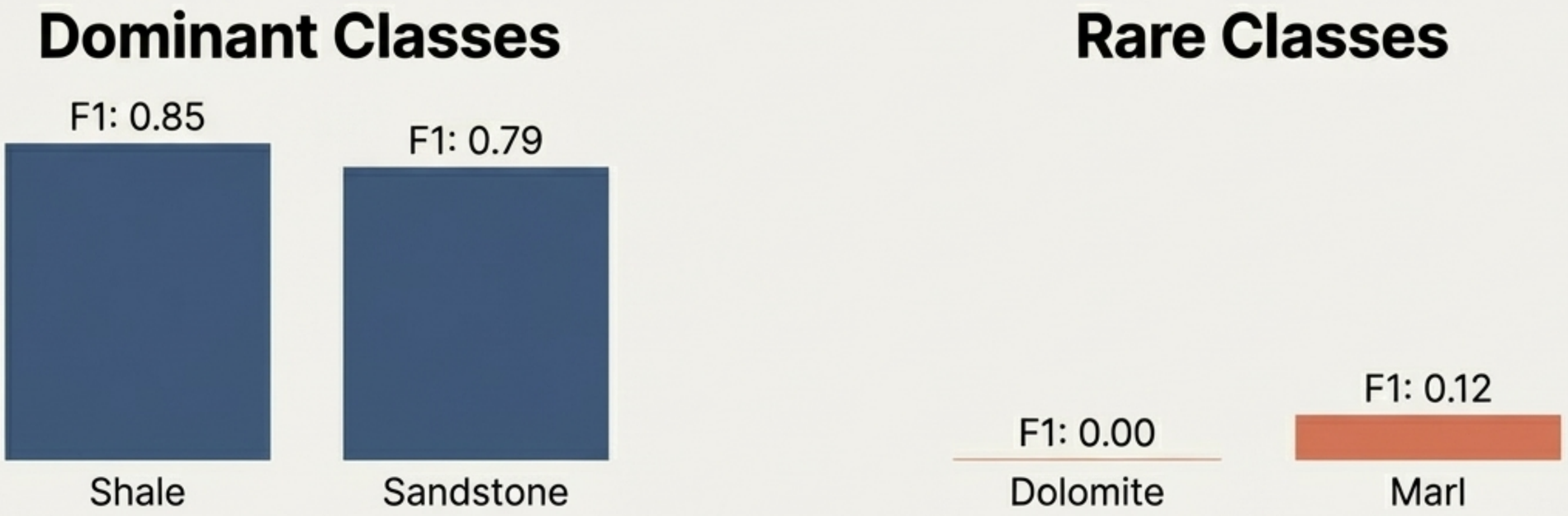
Training Protocol

- **Method:** Stratified 10-Fold Cross-Validation
- **Optimizer:** Adam
- **Regularization:** Dropout (0.3), Batch Normalization
- **Callbacks:** Early Stopping (patience=5), ReduceLROnPlateau (patience=2)

Average Cross-Validation Accuracy: **89.4%**

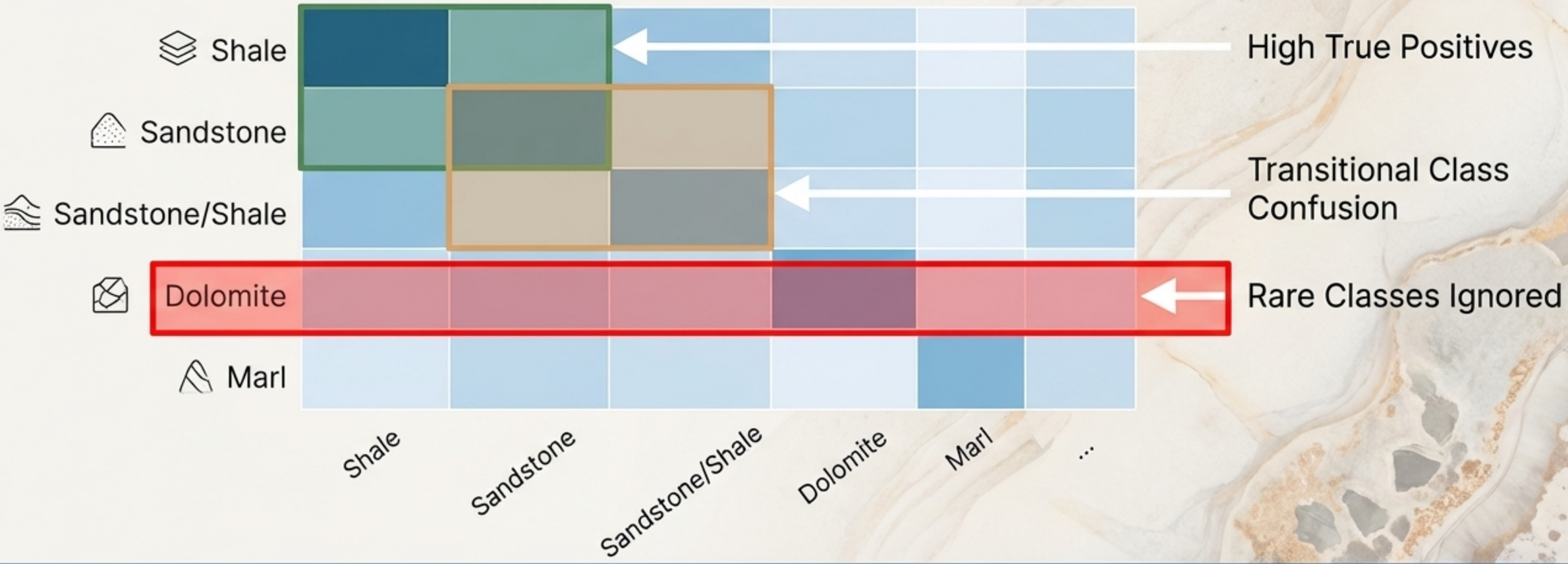
Test Set Accuracy Reaches 74%, With Strong Performance on Majority Classes

Overall Test Set Accuracy: **74%**



The model excels at identifying common lithologies but fails on rare classes—a direct consequence of the extreme data imbalance shown earlier.

The Confusion Matrix Visually Confirms Class Performance



Key findings: clear success on primary classes, significant confusion on transitional classes, and systematic failure to identify rare classes

A Strong Baseline with Clear Directions for Improvement

✔ Summary & Achievements

- Developed an end-to-end deep learning pipeline achieving **74%** accuracy on a complex dataset.
- Demonstrated that standard ANNs, while powerful, are highly susceptible to severe class imbalance.
- Established a valuable baseline with excellent performance on majority lithofacies.

→ Future Directions

- **Imputation Strategy**: Explore advanced methods (e.g., KNN, MICE) to better handle the ~30% missing data.
- **Class Imbalance**: Implement specific techniques like SMOTE or class-weighted loss functions to boost minority class performance.
- **Model Architecture**: Experiment with sequence-aware models (LSTMs, Transformers) to better leverage the data's inherent vertical structure.