Simple Averaging vs. Learned Stacking in K-Fold Ensembles: Evidence from Pore Pressure Prediction

Pranav Patel
Khoury College of Computer Sciences
Northeastern University
San Jose, CA, USA
pranavjpatel.08@gmail.com

Muhammad Raiees Amjad
Department of Computer Science
Bahria University
Islamabad, Pakistan
mraiees.buic@bahria.edu.pk

Rohan Benjamin Varghese
Khoury College of Computer Sciences
Northeastern University
San Jose, CA, USA
varghese.roh@northeastern.edu

Tehmina Amjad
Khoury College of Computer Sciences
Northeastern University
San Jose, CA, USA
tehmina.amj@gmail.com

***Abstract*—Pore pressure prediction is critical for drilling safety, yet ensemble approaches lack principled methods for model selection and combination under cross-well distribution shift. This study presents a K-Fold stacking framework using well-based GroupKFold cross-validation, where models are ranked by out-of-fold (OOF) R^2, enabling leakage-free ensemble selection without ground truth from new wells. Five architectures (CNN, DFNN, RNN, Random Forest, XGBoost) are evaluated on 186,070 samples from 17 development wells in Pakistan's Potwar Basin. On 4 blind test wells (84,904 samples), the 5-model ensemble achieves R^2 = 0.8905, outperforming the best individual model by 7.03 percentage points. Simple averaging consistently outperforms Ridge meta-learner stacking across all ensemble sizes, with the gap widening from 0.41 to 2.36 percentage points as more models are included. Ridge concentrates weight on fewer models and assigns negative coefficients to others, discarding architectural diversity valuable under distribution shift. Top-K ablation confirms monotonic gains from K=1 through K=5, with the bulk of improvement from combining the first two architecturally distinct models.**

***Index Terms*—pore pressure prediction, ensemble learning, K-Fold stacking, out-of-fold validation, cross-well generalization**

I. Introduction

Precise pore pressure prediction (PPP) is essential for drilling safety in oil and gas operations. Inaccurate pore pressure estimation causes severe consequences, including blowouts, environmental damage, safety risks, and economic losses due to excessive mud weight and reduced drilling efficiency. The economic impact is substantial: drilling hazards related to abnormal pressure zones account for approximately 20% of non-productive time in exploration wells [1]. With abnormally pressured formations constituting about 47.7% of the world's oil fields [2], precise pressure prediction remains a critical challenge.

Traditional physics-based methods such as Eaton's [3] and Bowers' [4] approaches have provided foundational understanding of pressure-compaction relationships. However, these methods assume uniform compaction trends that frequently fail in complex geological settings with heterogeneous rock properties and variable pressure regimes [5]. This heterogeneity makes conventional physics-based methods incapable of fully capturing the complex relationships between well logs and formation pressure [6].

In practice, pore pressure is estimated from well log data (e.g., sonic travel time, resistivity) recorded at successive depth intervals. In ML, each depth-specific sample is treated as a feature vector describing rock properties. The increasing complexity and volume of these geoscientific data have driven a paradigm shift toward machine learning (ML) and deep learning (DL) methods for PPP [7]. ML methods have demonstrated strong performance, with reported R^2 values exceeding 0.95 on validation sets [5], [8]. Recent ensemble approaches combining Random Forest and XGBoost have achieved correlations above 0.97 [9], and hybrid meta-ensemble frameworks integrating CNN, RNN, DFNN, RF, and XGBoost have shown R^2 values of 0.93 on blind wells [6].

Despite these advances, several critical problems persist in ML-based PPP including 1) data leakage in model selection, 2) cross-well generalization failure, 3) neural network instability, and 4) ensemble composition uncertainty.

A recurring issue in the literature is data leakage during model selection, where models or ensemble configurations are chosen based on test-set performance. In such cases, the test set implicitly influences model selection, leading to overoptimistic performance estimates and undermining real-world applicability. This practice is particularly problematic in deployment scenarios, where ground-truth pore pressure is unavailable for new wells and test-based selection is infeasible. Recent analyses indicate that this issue is widespread. Singh et al. [10] revealed that 78% of ML-based PPP studies suffer from improper data splitting, leading to overoptimistic performance estimates.

Another critical limitation concerns cross-well generalization. Models trained on a specific set of wells often fail when applied to geologically distinct targets. Radwan et al. [11] found that model performance degraded significantly with

geological distance from training wells, with RMSE varying from 0.25 to 14.71 psi depending on well proximity. Amjad et al. [6] reported that all five base models (CNN, RNN, DFNN, RF, XGBoost) produced *negative* R^2 values when directly tested on blind wells without domain adaptation, indicating predictions worse than simply using the mean value.

Beyond generalization, neural network instability poses a serious risk in PPP applications. Studies have documented catastrophic failures where neural networks produce predictions many orders of magnitude from ground truth on specific data subsets [6], [9]. Ensuring stable training across diverse geological folds requires careful preprocessing, as discussed in Section III-A.

While ensemble learning has consistently been shown to outperform individual models, two key questions remain underexplored: *how many* models to include and *how* to combine them. Most studies rely on heuristic selection or learned meta-learners without examining whether learned weighting actually helps under cross-well distribution shift. Furthermore, highly correlated models may contribute little additional information despite strong standalone performance, making ensemble composition a non-trivial design choice.

These challenges motivate a principled framework that enables model selection without test data and systematically determines both ensemble composition and combination strategy.

This study addresses these challenges through a K-Fold stacking framework with well-based GroupKFold cross-validation and out-of-fold (OOF) model selection. The key contributions are:

1) **K-Fold stacking with well-based GroupKFold**: Models are ranked by OOF R^2 computed via 5-fold well-grouped cross-validation, ensuring no well appears in both training and validation folds and enabling leakage-free model selection without ground truth.
2) **Simple averaging vs. learned stacking**: A systematic comparison across ensemble sizes shows that simple averaging consistently outperforms Ridge meta-learner stacking, with the gap widening as more models are included. Analysis of Ridge coefficients reveals that learned weighting discards architectural diversity valuable under distribution shift.
3) **Top-K ablation**: Incremental model addition demonstrates monotonic ensemble gains across all five architecturally diverse models, with the 5-model ensemble outperforming the best single model by 7.03 percentage points.
4) **Reproducible evaluation**: Well-based data splitting ensures test wells remain completely unseen during training and model selection, providing realistic cross-well generalization estimates.

The implementation code is publicly available.[1] The well log data from Pakistan's Potwar Basin are available for research purposes upon request to the corresponding author.

[1]https://github.com/pranavpatel08/kfold-GeoScience

II. Related Work

A. Machine Learning for Pore Pressure Prediction

The application of ML to PPP has evolved substantially over the past decade. Ahmed et al. [8] employed artificial neural networks to identify correlations among seven input parameters, achieving a correlation coefficient of 0.999. However, their random data splitting permitted within-well leakage: samples from the same well appeared in both training and test sets, letting models learn well-specific patterns rather than generalizable relationships.

Farsi et al. [5] compared Support Vector Machines and neural networks, achieving R^2 exceeding 0.95. Their subsequent work proposed a Multilayer Extreme Learning Machine optimized with Particle Swarm Optimization (MELM-PSO), achieving RMSE of 11.551 psi in carbonate sequences [12]. Despite strong performance, these studies similarly employed sample-level splitting.

Amjad et al. [2] applied Deep Feedforward Neural Networks to predict geopressure in Pakistan's Potwar Basin, achieving well-specific correlations above 97%. Subsequent work [13] demonstrated that CNN and DFNN models initially failed to generalize to blind wells, requiring fine-tuning and controlled data injection to achieve R^2 values of 0.74. Most recently, Amjad et al. [6] proposed a Hybrid Meta-Ensemble (HME) framework with a meta-learner, achieving R^2 of 0.9382 on blind wells after injecting 15% of blind well data for domain adaptation.

Yu et al. [14] developed a multivariate machine learning methodology for PPP, demonstrating that ensemble methods provide robust solutions for capturing nonlinear dependencies. Krishna et al. [9] analyzed 22,539 data points from the Volve oil field, finding that Random Forest (R^2 = 0.97, RMSE = 2.70 MPa) outperformed complex deep learning solutions, questioning the assumption that deeper architectures necessarily yield better predictions for structured geological data.

B. Cross-Well Generalization and Data Leakage

The challenge of cross-well generalization has received increasing attention. Radwan et al. [11] examined 25,935 data records from four wells in New Zealand's Mangahewa gas field, finding that prediction performance worsened with geological distance from training wells. Their semi-supervised approach only achieved acceptable RMSE for small geological distances, highlighting the domain shift problem.

Singh et al. [10] conducted a critical review revealing that 78% of ML-based PPP studies suffer from improper data splitting. Their analysis demonstrated that sample-level splitting leads to overoptimistic performance estimates that do not reflect true generalization capability. This finding motivates the strict well-based splitting employed in the present study.

Recent work on cross-well sonic log prediction demonstrated that rigorous leakage control is essential for reliable predictions [15]. A strictly leakage-free, features-only strategy using causal lag windows achieved R^2 = 0.895 on blind cross-well tests, though reverse direction predictions showed lower performance, indicating persistent inter-well distribution shift.

C. Ensemble Methods and Stacking

Stacking ensembles combine diverse base learners via a meta-learner trained on their predictions [16]. The standard approach trains the meta-learner on validation predictions, introducing leakage if the same set guides model selection. K-Fold stacking addresses this by generating out-of-fold predictions, where each sample's prediction comes from a model that never saw that sample during training.

Wood [17] evaluated 30 ML algorithms for reservoir property prediction, concluding that ensemble methods outperform individual models but may suffer under distribution shift. Abdelaal et al. [18] showed that ensemble methods provide the most effective balance between predictive accuracy and computational speed for real-time drilling applications, achieving $R^2 > 0.9$ while processing streaming drilling parameters.

The challenge of selecting which models to include in an ensemble, and *how* to combine them, remains underexplored. This study addresses this gap through systematic top-K ablation and a direct comparison of simple averaging against learned Ridge stacking across ensemble sizes.

III. Methodology

A. Dataset and Well-Based Splitting

Well log data from 21 wells across Pakistan's Potwar Basin are used in this study. The Potwar Basin is a tectonically active foreland basin with underpressured, normal, and overpressured zones. Preprocessing applies domain-specific physical constraints to remove anomalous values before training. After preprocessing, all 21 wells are retained with sufficient data coverage.

Wells are partitioned to ensure complete separation between development and evaluation. Table I summarizes the data distribution. The four test wells (RAJIAN-03A, PINDORI-2, TURKWAL DEEP X 2, and Balkassar POL 01) remain completely unseen during training and model selection, providing an unbiased estimate of cross-well generalization.

TABLE I
Data Splits and Pressure Regime Distribution

Split	Wells	Samples	%	Under	Normal	Over
Development	17	186,070	68.7	12.8%	14.6%	72.6%
Test (blind)	4	84,904	31.3	16.2%	13.4%	70.4%
Total	**21**	**270,974**	**100**			

The pressure regime distributions are similar across splits, indicating well-balanced stratification. The test set contains a slightly higher proportion of underpressured samples (16.2% vs. 12.8%), providing a conservative evaluation of model generalization.

B. Feature Engineering

Thirteen predictor features are used, comprising 9 raw well log measurements and 4 physics-engineered quantities:

Raw features: True vertical depth (TVD), compressional wave travel time (DT), normal compaction trend (DT_NCT), gamma ray (GR), porosity (SPHI), hydrostatic pressure (HP), overburden stress (OB), bulk density (RHOB), deep resistivity (RES_DEEP).

Engineered features:

- Eaton ratio: $(DT/DT\_NCT)^3$, capturing departure from normal compaction
- Pressure gradients: $HP/(TVD \times 3.28084)$ and $OB/(TVD \times 3.28084)$ in psi/ft
- Normalized depth: $TVD/\max(TVD)$ for depth-independent learning

The target variable is pore pressure (PPP) in psi.

C. Model Architectures

Five architecturally diverse models are employed to capture different data characteristics. All architectures and hyperparameters follow Amjad et al. [6], where they were optimized for pore pressure prediction on the same geological basin. Although well log data are not images or time series, depth-ordered measurements exhibit local correlations captured by 1D convolutions (CNN) and sequential trends captured by recurrent units (RNN), justifying their inclusion alongside traditional models:

CNN: Two 1D convolutional blocks (64 and 128 filters, kernel size 2) with batch normalization, max pooling, and dropout (0.25), followed by dense layers (128→64→1) with L2 regularization (λ=0.002). Trained for 100 epochs with batch size 32.

DFNN: Fully-connected layers (13→128→64→32→1) with batch normalization and dropout. To address instability, an Adam optimizer with gradient clipping (clipnorm=1.0), L2 regularization (10^{-4}), a RobustScaler for input features, and target clipping at the 1st–99th percentiles are applied. Trained for 100 epochs with batch size 64.

RNN: Two LSTM layers (64 and 32 units) with batch normalization and dropout (0.3), followed by dense layers (64→1). Trained for 130 epochs with batch size 16.

Random Forest: 500 trees with max_depth=15, min_samples_split=5, min_samples_leaf=2.

XGBoost: 1000 estimators with max_depth=6, learning_rate=0.1, subsample=0.8, colsample_bytree=0.8, early stopping after 10 rounds without improvement.

All neural networks use Adam optimizer with EarlyStopping (patience=5) and ReduceLROnPlateau callbacks. Tree models use the original (unscaled) target, while neural networks use standardized targets with inverse transformation for predictions. All experiments were conducted on a single NVIDIA H200 GPU with 64-core CPU.

D. K-Fold Stacking with OOF Model Selection

The core methodology proceeds in three stages:

Stage 1: OOF Prediction Generation. 5-fold GroupKFold cross-validation (scikit-learn) is applied on the development data, using well identifiers as the grouping variable so that no

well appears in both the training and validation folds of any split. In each iteration:

1) StandardScaler is fit on the training folds; the held-out fold is transformed using these statistics.
2) Each model is trained on the training folds.
3) Predictions are generated for the held-out fold only.

Concatenating predictions across all iterations yields OOF predictions for each model, where no sample was seen during training. The 17 development wells are distributed unevenly across folds (1, 3, 4, 4, 5 wells), as GroupKFold assigns whole wells to folds; the single-well fold contains the largest well, contributing a comparable sample count.

Stage 2: Model Ranking and Combination. Models are ranked by OOF R^2. Critically, ranking uses OOF metrics, not test performance. Both simple averaging and a Ridge regression meta-learner (α=1.0) trained on OOF predictions are evaluated as combination strategies, enabling a direct comparison of equal vs. learned weighting (Section IV-C).

Stage 3: Final Model Training. All base models are retrained on the full train+validation data. Test predictions are generated only once for final evaluation.

E. Ensemble Analysis

Top-K ablation starts from the highest OOF R^2 model and incrementally adds models in rank order. At each step K, both simple averaging and Ridge stacking are evaluated on the test set. This dual evaluation reveals: (1) how ensemble performance scales with size, (2) which combination strategy performs better at each K, and (3) how Ridge coefficient behavior changes as ensemble size grows.

Fig. 1 illustrates the complete pipeline, from raw well data through preprocessing, 5-fold cross-validation, OOF ranking, and final ensemble selection via top-K ablation.

IV. Results

A. Cross-Validation Performance

Table II presents OOF R^2, RMSE, and MAE for each model across the 5-fold well-based GroupKFold cross-validation.

TABLE II
Out-of-Fold Performance (5-Fold Well-Based GroupKFold)

Model	OOF R^2	RMSE (psi)	MAE (psi)
XGBoost	**0.8537**	683.09	515.49
DFNN	0.8475	697.38	536.97
CNN	0.8400	714.45	538.08
RNN	0.7905	817.41	592.09
RF	0.7859	826.49	619.09

All five models achieve OOF R^2 between 0.79 and 0.85, with XGBoost ranking first (0.8537), followed by DFNN (0.8475) and CNN (0.8400). The compressed performance range reflects genuine cross-well generalization difficulty: with well-based splits, no model memorizes well-specific patterns, producing realistic but lower R^2 compared to within-well evaluation. Notably, preliminary experiments without physical-constraint-based preprocessing showed neural network sensitivity to outliers; domain-specific filtering resolved this, enabling all five architectures to contribute stably to ensembles.

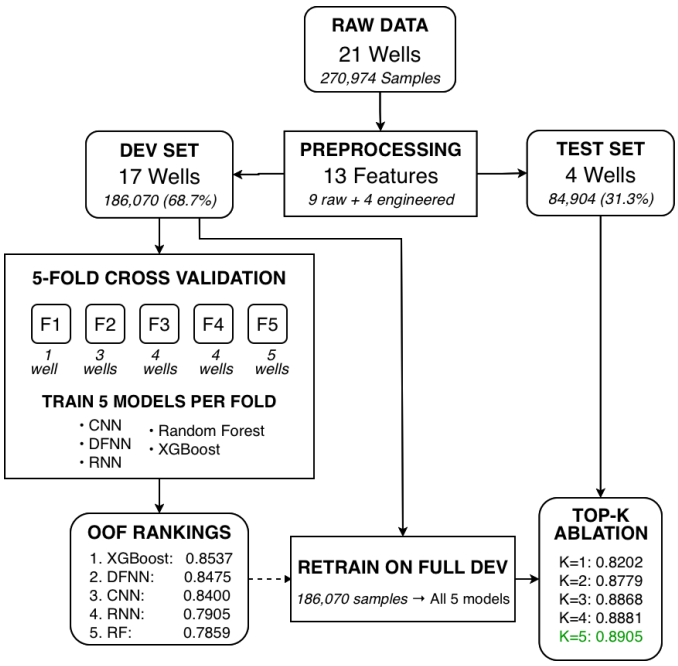

Fig. 1. K-Fold stacking workflow. Development data (17 wells) undergoes 5-fold well-based GroupKFold CV to generate OOF predictions and rank models. Final models are retrained on full development data and evaluated on 4 blind test wells. The 5-model ensemble with simple averaging achieves R^2 = 0.8905.

B. Top-K Ensemble Ablation

Fig. 2 presents test-set performance as models are incrementally added in OOF R^2 order using simple averaging.

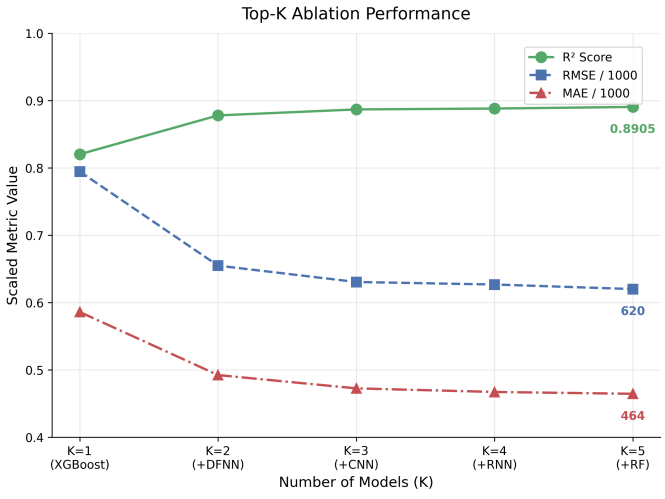

Fig. 2. Top-K Ablation Performance. Models are added in OOF rank order (XGBoost → DFNN → CNN → RNN → RF). R^2 improves monotonically through K=5, with diminishing returns after K=2.

As illustrated in Fig. 2, performance improves monotonically as models are incrementally added, reaching a peak R^2 of 0.8905 with the full 5-model ensemble. The trajectory demonstrates clear diminishing returns. The most substantial marginal improvement (+5.77 percentage points) occurs when combining the first two models, XGBoost and DFNN. This massive gain highlights how merging fundamentally distinct architectures (a tree-based model and a neural network) drives the core benefit of the ensemble. Subsequent additions contribute smaller, though consistently positive, marginal gains: +0.89 pp for the third model, +0.13 pp for the fourth, and +0.24 pp for the fifth.

The full ensemble (K=5) outperforms the best individual model (XGBoost, R^2 = 0.8202) by 7.03 percentage points, demonstrating that combining tree-based, feedforward, convolutional, recurrent, and random forest models provides substantial benefits for cross-well generalization.

C. Simple Averaging vs. Learned Stacking

Table III compares simple averaging against Ridge meta-learner stacking across ensemble sizes K=2 through K=5, alongside Ridge coefficients at each K.

TABLE III
SIMPLE AVERAGING VS. RIDGE STACKING ACROSS ENSEMBLE SIZES

K	Method	R^2	RMSE	MAE
1	XGBoost (best single)	0.8202	794.53	585.87
2	Simple avg	0.8779	654.74	492.15
	Ridge	0.8738	665.59	507.37
3	Simple avg	0.8868	630.41	472.37
	Ridge	0.8819	643.94	492.83
4	Simple avg	0.8881	626.67	467.04
	Ridge	0.8800	649.18	498.46
5	Simple avg	**0.8905**	**619.91**	**464.31**
	Ridge	0.8669	683.53	518.43

Simple averaging outperforms Ridge stacking at *every* ensemble size, with the gap widening as K increases: 0.41 pp at K=2, 0.49 pp at K=3, 0.81 pp at K=4, and 2.36 pp at K=5. This widening gap is notable: as more models are added, Ridge has more diversity to discard, and does so.

The Ridge coefficients reveal why. At K=5, Ridge assigns: XGBoost = 0.759, CNN = 0.346, DFNN = 0.309, RNN = −0.086, RF = −0.200. The negative weights on RNN and RF effectively exclude them from the ensemble despite their positive contributions under simple averaging. As K grows from 2 to 5, Ridge increasingly concentrates weight on XGBoost while suppressing models it considers redundant based on OOF correlations.

The interpretation is straightforward: Ridge optimizes for the OOF distribution, where model predictions are correlated. On blind test wells with distribution shift, the diversity that Ridge discards, namely different architectures making errors in different regions, proves valuable. Equal weighting preserves this diversity by construction.

Fig. 3 provides qualitative validation on blind test well PINDORI-2. The 5-model simple-average ensemble closely tracks the actual pressure profile across the full depth range. Tree models and neural networks capture complementary pressure patterns, explaining the diversity benefit preserved by equal weighting.

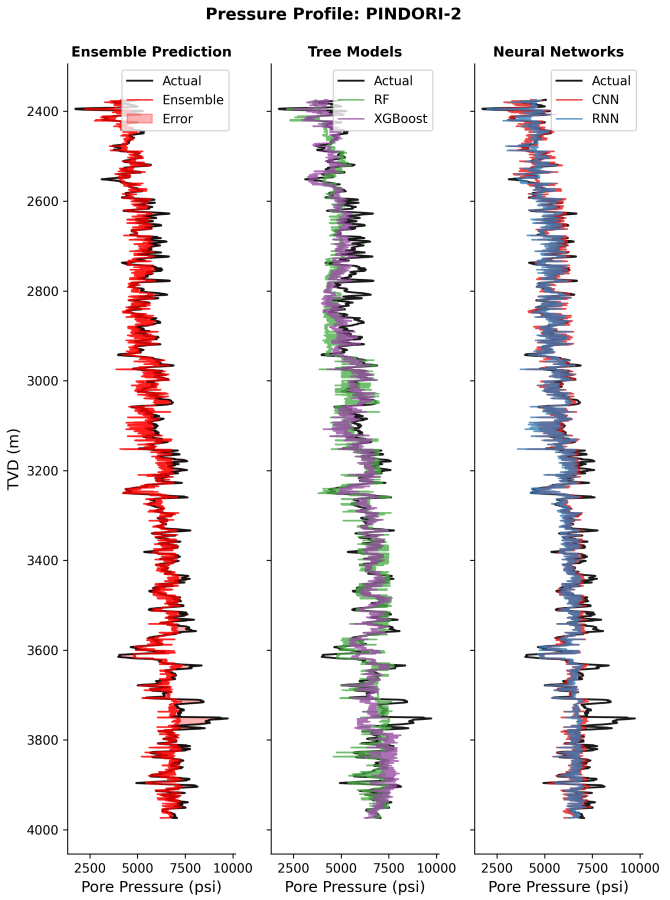

Fig. 3. Pressure prediction profile for blind test well PINDORI-2. Left: 5-model ensemble. Middle: tree model (RF, XGBoost) predictions. Right: neural networks (CNN, RNN, DFNN) capture different pressure variations, illustrating the architectural diversity that simple averaging preserves.

V. DISCUSSION

A. Why Simple Averaging Wins

The consistent advantage of simple averaging over Ridge stacking, with the gap widening from 0.41 pp at K=2 to 2.36 pp at K=5, reveals a fundamental tradeoff between optimizing for the training distribution and preserving diversity for unseen distributions.

Ridge learns combination weights on OOF predictions, where model errors are correlated because all models train on overlapping data from the same geological basin. The learned weights concentrate on models that best fit this shared distribution; XGBoost receives 0.76 weight while RNN and RF receive negative weights. On blind test wells with distribution shift, these suppressed models contribute complementary error

patterns that Ridge has discarded. Simple averaging preserves all contributions equally, providing implicit regularization against distribution shift.

The widening gap with K is mechanistically clear: with K=2 models, Ridge has limited opportunity to discard diversity. As K grows, more models become candidates for suppression, and Ridge increasingly overfits to OOF correlations.

Exhaustive subset search across all 31 combinations further supports this analysis. Prediction correlations are uniformly high (0.92 to 0.97), with RF exhibiting the lowest correlation to neural networks, explaining its disproportionate diversity contribution despite the lowest individual OOF R^2. The best test-set subset under simple averaging (CNN, DFNN, RF; R^2 = 0.8938) outperforms the full 5-model ensemble (R^2 = 0.8905). However, it has a lower OOF R^2 (0.852) than the greedy-optimal K=3 (XGBoost, DFNN, CNN; OOF R^2 = 0.863, test R^2 = 0.887), meaning it would not be identified by OOF selection alone. This confirms that OOF ranking provides a principled but conservative selection criterion. Including all models with equal weighting remains the most robust default when oracle-optimal subsets cannot be identified without test data. Practitioners should favor simple averaging for cross-well prediction tasks where distribution shift is expected.

B. Limitations

Several limitations warrant discussion. First, results are from Pakistan's Potwar Basin; generalization to other geological settings requires validation. Second, only Ridge stacking was compared against simple averaging; more sophisticated meta-learners such as neural network stacking or attention-based combination may yield different results. Third, preprocessing quality is critical: physical-constraint-based filtering was necessary to ensure stable neural network training across folds.

VI. Conclusion

This study presented a K-Fold stacking framework for pore pressure prediction using well-based GroupKFold cross-validation, enabling leakage-free model selection for new wells without ground truth. The 5-model ensemble (XGBoost, DFNN, CNN, RNN, RF) with simple averaging achieves R^2 = 0.8905 on 4 blind test wells, outperforming the best individual model by 7.03 percentage points. Top-K ablation confirms monotonic gains from architectural diversity, with the largest improvement coming from combining the first two architecturally distinct models.

A central finding is that simple averaging consistently outperforms Ridge meta-learner stacking, with the gap widening from 0.41 to 2.36 percentage points as ensemble size increases. Ridge concentrates weight on fewer models and assigns negative coefficients to others, discarding diversity that proves valuable under cross-well distribution shift. For geoscientific prediction tasks with inherent domain shift, equal weighting is preferable to learned combination.

Future work includes Bayesian or conformal prediction for uncertainty quantification, SHAP-based feature explanations for domain expert trust, and multi-basin validation to establish generalization bounds.

Acknowledgment

AI-assisted tools (Claude, ChatGPT) supported code development and manuscript drafting. All experimental design, implementation, and interpretation were conducted by the authors, who bear full responsibility for the scientific content.

References

[1] J. Zhang, "Pore pressure prediction from well logs: Methods, modifications, and new approaches," Earth-Science Reviews, vol. 108, no. 1-2, pp. 50–63, 2011.

[2] M. R. Amjad, M. Zafar, M. B. Malik, and Z. Naseer, "Precise geopressure predictions in active foreland basins: An application of deep feedforward neural networks," Journal of Asian Earth Sciences, vol. 245, p. 105560, 2023.

[3] B. A. Eaton, "The equation for geopressure prediction from well logs," in Fall Meeting of the Society of Petroleum Engineers, SPE, 1975.

[4] G. L. Bowers, "Pore pressure estimation from velocity data: Accounting for overpressure mechanisms besides undercompaction," SPE Drilling & Completion, vol. 10, no. 2, pp. 89–95, 1995.

[5] M. Farsi, N. Mohamadian, H. Ghorbani, D. A. Wood, and S. Davoodi, "Predicting formation pore pressure from well-log data using machine learning algorithms," Journal of Petroleum Science and Engineering, vol. 205, p. 108906, 2021.

[6] M. R. Amjad, R. B. Varghese, and T. Amjad, "Machine learning models for subsurface pressure prediction: A data mining approach," Computers, vol. 14, no. 11, p. 499, 2025.

[7] A. Lawal, Y. Yang, H. He, and N. L. Baisa, "Machine learning in oil and gas exploration: A review," IEEE Access, vol. 12, pp. 19035–19058, 2024.

[8] A. Ahmed, S. Elkatatny, A. Ali, and A. Abdulraheem, "New model for pore pressure prediction while drilling using artificial neural networks," Arabian Journal for Science and Engineering, vol. 44, no. 6, pp. 6079–6088, 2019.

[9] S. Krishna, S. A. Irfan, S. Keshavarz, G. Thonhauser, and S. U. Ilyas, "Smart predictions of petrophysical formation pore pressure via robust data-driven intelligent models," Multiscale and Multidisciplinary Modeling, Experiments and Design, vol. 7, pp. 5611–5630, 2024.

[10] H. Singh, Y. Seol, and J.-H. Choi, "Evaluation biases in machine learning models for pore pressure prediction: A critical review," Reviews of Geophysics, vol. 62, no. 1, p. e2023RG000798, 2024.

[11] A. E. Radwan, D. A. Wood, and A. A. Radwan, "Machine learning and data-driven prediction of pore pressure from geophysical logs: A case study for the Mangahewa gas field, New Zealand," Journal of Rock Mechanics and Geotechnical Engineering, vol. 14, no. 6, pp. 1799–1809, 2022.

[12] M. Farsi et al., "Predicting formation pore-pressure from well-log data with hybrid machine-learning optimization algorithms," Natural Resources Research, vol. 30, pp. 3455–3481, 2021.

[13] M. R. Amjad et al., "Beyond the surface: Deep neural networks for subsurface pore pressure estimation," GSA Connect 2025, vol. 57, no. 6, 2025.

[14] H. Yu, G. Chen, and H. Gu, "A machine learning methodology for multivariate pore-pressure prediction," Computers & Geosciences, vol. 143, p. 104548, 2020.

[15] B. Zare, M. M. Huque, L. A. James, and H. Usefi, "Cross-well machine learning prediction of sonic logs in Newfoundland and Labrador," *Scientific Reports*, vol. 16, no. 1, p. 576, Jan. 2026.

[16] D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.

[17] D. A. Wood, "Comprehensive evaluation of machine learning algorithms for predicting missing reservoir properties," Petroleum Science and Technology, vol. 40, no. 15, pp. 1823–1845, 2022.

[18] A. Abdelaal, S. Elkatatny, and A. Abdulraheem, "Data-driven modeling approach for pore pressure gradient prediction while drilling from drilling parameters," ACS Omega, vol. 6, no. 21, pp. 13807–13816, 2021.