Lithology-Aware Conditional Variational Autoencoder for Synthetic Well Log Generation in Petroleum Reservoirs (Student Abstract)

Aline Cambri Fredere[1,2], Gabriel De Oliveira Ramos[1], Luciano Garim Garcia[1,2], Mateus da Rocha Simionato[2], José Manuel Marques Teixeira de Oliveira[2], Ariane Santos da Silveira[2]

[1] Graduate Program in Applied Computing
[2] Center of Excellence in Geological and Numerical Modeling (MGEN)
Universidade do Vale do Rio dos Sinos, São Leopoldo, Brazil
{afredere, gdoramos, lucianogarim, mrsimionato, josem, ariane}@unisinos.br

Abstract

Machine learning applications in reservoir modeling are hindered by the limited availability of well log data, a common challenge in the oil and gas industry. We propose VAEc-tMC, a domain-informed Conditional Variational Autoencoder that generates synthetic well-log data conditioned on rock type. Addressing a critical gap by existing generative models that rely solely on statistical reconstruction, our model embeds geological domain knowledge into the latent space, and optimizes a modified objective with an adaptive Student-t reconstruction loss and a β-weighted KL regularizer, improving stability under heavy-tailed data. When used for data augmentation, the synthetic samples preserve inter-log dependencies and substantially enhance downstream classification, accuracy 39→63%, F1-score 36→68%, AUC 0.46→0.80 on a held-out well. Beyond the geological context, the proposed approach illustrates a generalizable strategy where domain-aware generative models with adaptive loss functions provide a robust solution for data-efficient learning in scientific domains facing data scarcity, noise, and heavy-tailed distributions.

Introduction

The characterization of petroleum reservoirs depends on the integration of multiple data sources, including seismic surveys, laboratory tests, and geophysical well logs. Well logs are continuous records acquired by logging tools along the borehole that measuring physical properties of rocks and fluids including density, porosity, and gamma-ray response, plotted as a function of depth. However, well logs are scarce and expensive to acquire, which limits the accuracy of geological models and constrains the use of machine learning reservoir studies.

The generation of synthetic well logs has evolved significantly from regression-based predictive models, ranging from traditional neural networks to sequential and hybrid architectures (Du et al. 2008; Rolon et al. 2009; Kim et al. 2020; Zhang, Chen, and Meng 2018; Shan et al. 2021), to more complex generative frameworks such as VAEs (Jeong et al. 2021) and GANs (Garcia et al. 2024; Al-Fakih et al. 2025), which aim to model the complete data distribution rather than point-wise predictions. State-of-the-art approaches have refined these generative methods with custom loss functions, such as those based on PCA, to better preserve the multivariate correlations between logs (Garcia et al. 2024). However, existing work shows that, despite increasing sophistication, existing generative models largely disregard geological knowledge. They reproduce only average statistical patterns, failing to capture petrophysical dependencies. Historically, lithology has been treated as a secondary variable or applied externally to the synthesis process (Parapuram, Mokhtari, and Hmida 2018; Oliveira and de Carvalho Carneiro 2020; Jeong et al. 2021), revealing a critical research gap. This work addresses this limitation by proposing VAEc-tMC, a Conditional Variational Autoencoder (CVAE) that explicitly integrates rock type (lithology) as a condition to capture intrafacies variability and enforce geologically consistent inter-log relationships.

The main contributions of this work are as follows:

- We introduce VAEc-tMC, the first CVAE for synthetic well log generation that explicitly incorporates rock type and integrates geological knowledge.
- We devise a loss function adapted from Evidence Lower Bound (ELBO) employing Student's t-distribution for the reconstruction term, enabling modeling of heavy-tailed, non-Gaussian data.
- We demonstrate that the proposed architecture enables statistically faithful and geologically plausible data augmentation, leading to a substantial boost in downstream lithology classification performance

Architecture VAEc-tMC

The *VAEc-tMC* extends the CVAE (Sohn, Lee, and Yan 2015) specifically designed for lithology-aware well log generation. Input features are normalized using a scaler based on the median and the Median Absolute Deviation (MAD). Lithology is encoded through a one-hot vector, whose dimensionality equals the number of lithology classes. The dataset is divided into training (80%) and testing (20%) subsets.

The encoder concatenates both inputs (scaled log features and the one-hot lithology vector) and applies two dense layers (256 and 128 units, *Swish* activation) to map data into an 8-dimensional latent space parameterized by mean (μ) and

Lithology	Synthetic	Synthetic + Real	Real
1	2000	2445	445
2	2000	2055	55
3	2000	2043	43
4	2000	4055	2055
5	2000	2888	888

Table 1: Sample distribution of datasets by lithology type before and after data augmentation.

log-variance ($\log \sigma^2$). Latent variables are sampled via the reparameterization trick with Gaussian noise.

The decoder, receives the concatenated latent vector and lithology condition and feeds it into two dense layers (128 and 256 neurons, Swish activation) followed by a linear output layer matching the number of input features.

Model optimization follows modified Evidence Lower Bound (ELBO), comprising two principal adaptations: (i) the reconstruction likelihood adopts independent univariate Student-t distributions ($\nu = 3$), accommodating heavy-tailed marginals and outliers; and (ii) the Kullback–Leibler (KL) divergence is weighted by β = 0.48, prioritizing reconstruction fidelity over posterior regularization. The reconstruction term is estimated via multi-sample Monte Carlo using L = 10 samples, providing lower-variance gradient estimates compared to single-sample approximations. Hyperparameters β and ν were empirically tuned.

Training uses the Adam optimizer with an exponential learning-rate decay (initial rate 10^{-3}, decay rate 0.9 every 100 steps) and gradient clipping (norm = 1.0). The model is trained for up to 200 epochs with batches of 128 samples.

Experiments and Results

We conducted two experiments to assess the statistical fidelity and practical utility of the synthetic well logs generated. Statistical fidelity was evaluated via distribution-based analyses (histograms, correlations, t-SNE). Practical utility was assessed by training XGBoost (XGB) and Multi-Layer Perceptron (MLP) classifiers on real, synthetic, and combined data, with performance evaluated on an independent real test set using accuracy, F1-score, and the area under the ROC curve (AUC).

Synthetic Well Log Generation

Our dataset consists of 3,483 samples from five wells in the Santos Basin, encompassing the Pre-salt interval. The selected features were `DEPTH`, `GR`, `RHOB`, `NPHI`, `DT`, `PE`, `HTHO`, `HURA`, `RT10`, `RT90` logs. Although small, the dataset reflects typical data scarcity in this domain. The VAEc-tMC was trained using the features and conditioned on five rock types, where the lithologies were obtained through geological interpretation. A total of 2,000 synthetic samples were generated for each lithology class (Table 1).

Downstream Task

We evaluated VAEc-tMC architecture in a downstream lithology classification task by training XGB and MLP on real, synthetic, and combined datasets (Table 1), and validating them on an independent and real well.

Model	Synthetic			Synthetic + Real			Real		
	Acc.	F1	AUC	Acc.	F1	AUC	Acc.	F1	AUC
XGB	0.62	**0.64**	0.79	**0.63**	**0.64**	**0.80**	0.39	0.36	0.55
MLP	0.64	0.67	**0.83**	**0.66**	**0.68**	0.80	0.54	0.50	0.46

Table 2: Overall classification performance of XGB and MLP on Real, Synthetic, and augmented (Synthetic + Real) datasets. All metrics are weighted averages across lithology classes.

The real dataset (Table 1) presents strong class imbalance and limited size. The proposed generative model effectively mitigates both limitations. By producing 2,000 realistic samples per lithology, the augmented dataset (Synthetic + Real) expands the feature distribution, enabling the classifiers to better capture inter-class boundaries.

The inclusion of synthetic data consistently improved all metrics, substantially increasing AUC, accuracy, and F1-scores for both classifiers (Table 2). All scenarios using synthetic data (synthetic-only or augmented) demonstrated significantly superior performance compared to the baseline trained only on real data. The results demonstrate that the generated samples successfully capture lithology-specific patterns and enrich the feature space representation.

Notably, synthetic only training achieves competitive performance (MLP AUC=0.83). This suggests VAEc-tMC successfully learned the lithology-specific distributions. The balanced synthetic dataset (2,000 samples per class) enables better boundary estimation than the severely imbalanced real data. However, augmented data shows more balanced performance across classifiers, validating VAEc-tMC's ability to generate geologically consistent data for practical use.

Consequently, VAEc-tMC operates as a knowledge-aware generative model that encodes lithology-specific information in the latent space, generating synthetic well-log data that preserve multivariate dependencies to support machine learning workflows in reservoir characterization.

Conclusion

This study introduced VAEc-tMC, the first CVAE for lithology-conditioned synthetic well log generation. It employs a custom loss optimized for stable gradient estimation with heavy-tailed data. The synthetic data improved model generalization in a downstream task, suggesting that domain-informed generative models hold potential for addressing data scarcity in reservoir characterization. Beyond the geological context, the proposed architecture suggests a generalizable strategy for knowledge-informed data augmentation in domains characterized by limited, noisy, and heavy-tailed datasets. Future work will benchmark VAEc-tMC against conditional GANs, TVAEs, and diffusion models to better contextualize its performance.

Acknowledgments

We thank the reviewers for their valuable feedback and Brazilian National Agency of Petroleum, Natural Gas and Biofuels (ANP) for providing the database used in this study. This research was partially supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (grant 313845/2023-9) and Petrobras (grant 4600675263).

References

Al-Fakih, A.; Koeshidayatullah, A.; Mukerji, T.; Al-Azani, S.; and Kaka, S. I. 2025. Well log data generation and imputation using sequence based generative adversarial networks. *Scientific Reports*, 15(1): 11000.

Du, Y.; Tan, W.-a.; Jiang, C.; Lu, D.; and Li, D. 2008. An Effective Hash-based Method for Generating Synthetic Well Log. In *2008 Third International Conference on Pervasive Computing and Applications*, volume 2, 1017–1020.

Garcia, L. G.; Ramos, G. D. O.; de Oliveira, J. M. M. T.; and Silveira, A. S. D. 2024. Enhancing Synthetic Well Logs with PCA-Based GAN Models. In *2024 International Conference on Machine Learning and Applications (ICMLA)*, 1350–1355.

Jeong, J.; Park, E.; Emelyanova, I.; Pervukhina, M.; Esteban, L.; and Yun, S.-T. 2021. Application of conditional generative model for sonic log estimation considering measurement uncertainty. *Journal of Petroleum Science and Engineering*, 196: 108028.

Kim, S.; Kim, K. H.; Min, B.; Lim, J.; and Lee, K. 2020. Generation of synthetic density log data using deep learning algorithm at the Golden field in Alberta, Canada. *Geofluids*, 2020: 5387183.

Oliveira, L. A. B. D.; and de Carvalho Carneiro, C. 2020. Synthetic geochemical well logs generation using ensemble machine learning techniques for the Brazilian pre-salt reservoirs. *Journal of Petroleum Science and Engineering*, 196: 108080.

Parapuram, G.; Mokhtari, M.; and Hmida, J. B. 2018. An artificially intelligent technique to generate synthetic geomechanical well logs for the bakken formation. *Energies*, 11(3): 680.

Rolon, L.; Mohaghegh, S. D.; Ameri, S.; Gaskari, R.; and McDaniel, B. 2009. Using artificial neural networks to generate synthetic well logs. *Journal of Natural Gas Science and Engineering*, 1(4-5): 118–133.

Shan, L.; Liu, Y.; Tang, M.; Yang, Y.; and Bai, X. 2021. CNN-BiLSTM hybrid neural networks with attention mechanism for well log prediction. *Journal of Petroleum Science and Engineering*, 205: 108838.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning Structured Output Representation using Deep Conditional Generative Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28.

Zhang, D.; Chen, Y.; and Meng, J. 2018. Synthetic well logs generation via recurrent neural networks. *Petroleum Exploration and Development*, 45(4): 629–639.