Machine Learning Based Subsurface Temperature Forecasting to Reduce Drilling Uncertainty in Geothermal Systems

Suman Saeed[1], Michael Short[2]

[1]PhD Researcher, School of Computing, Engineering and Digital Technologies, Teesside University, UK

[2]Professor, School of Computing, Engineering and Digital Technologies, Teesside University, UK

Abstract. Subsurface temperature uncertainty represents a major risk in geothermal drilling, particularly in volcanic systems where permeability, lithology, and fluid circulation create highly heterogeneous thermal regimes. This study presents a hybrid physics-informed machine learning framework for forecasting subsurface temperature using geothermal drilling and geophysical log data from Icelandic geothermal fields. A publicly available dataset from the GEOTHERMICA/RESULT project comprising 16 deep geothermal wells from the Elliðaár geothermal field was used for model development and validation. Ensemble and neural network models were optimized using Bayesian hyperparameter tuning and evaluated against conventional geothermal gradient methods. The present study represents a proof-of-concept demonstration of the proposed framework. Ongoing work is focused on expanding the dataset to 52 geothermal wells and enabling real-time deployment for geothermal drilling operations.

1 Introduction

Geothermal energy represents a reliable and sustainable source of baseload power. However, one of the key challenges in geothermal development is the uncertainty associated with subsurface temperature distribution prior to drilling. Traditional approaches based on linear geothermal gradients often fail to capture the complexity of volcanic and

*Corresponding Author: M.Short@tees.ac.uk

fault-controlled systems. As drilling operations are capital-intensive, inaccurate predictions can lead to significant financial and operational risks. This has motivated the development of data-driven approaches capable of learning nonlinear relationships between geological, operational, and thermal variables[1-3].

2. Research Gap and Motivation

Despite significant advancements in geothermal exploration, accurate prediction of subsurface temperature distribution remains a major challenge, particularly in geologically complex and data-sparse regions. Conventional approaches for estimating subsurface temperature primarily rely on simplified linear models, such as the geothermal gradient equation $T(z) = T_0 + Gz$. While these methods provide a first-order approximation, they fail to capture the inherent heterogeneity of subsurface formations, variations in lithology, and the influence of dynamic drilling conditions [4].

Existing studies have increasingly explored the application of machine learning techniques for geothermal temperature prediction. However, many of these approaches are limited by their reliance on either small datasets or a narrow set of input features [2,5], often excluding critical drilling parameters and spatial variability. Furthermore, hyperparameter tuning in such models is frequently conducted using traditional methods such as grid search or trial-and-error, which are computationally inefficient and may not yield globally optimal solutions [6].

Another significant limitation in current research is the lack of integration between physics-based understanding and data-driven models. Purely data-driven approaches may achieve high predictive accuracy but often lack interpretability and generalizability, particularly when applied to new geological settings. Conversely, traditional physics-based models fail to incorporate the nonlinear and complex interactions present in real-world subsurface systems [7,8].

Therefore, there exists a clear research gap in the development of a robust, hybrid framework that integrates geophysical knowledge with advanced machine learning techniques, while also leveraging efficient optimization strategies such as Bayesian Optimization. This study aims to address this gap by proposing a comprehensive methodology that combines multi-source data integration, physics-informed feature engineering, and optimized machine learning models to improve the accuracy and reliability of geothermal temperature forecasting.

2. Geological Background

In high-enthalpy geothermal systems, temperature distribution is controlled by a combination of depth, permeability, lithology, and fluid circulation. Fault structures act as pathways for fluid flow, resulting in localized thermal anomalies. Hydrothermal alteration further modifies rock properties, influencing heat transfer and storage. These factors result in highly heterogeneous temperature fields that cannot be described using simple deterministic models [9].

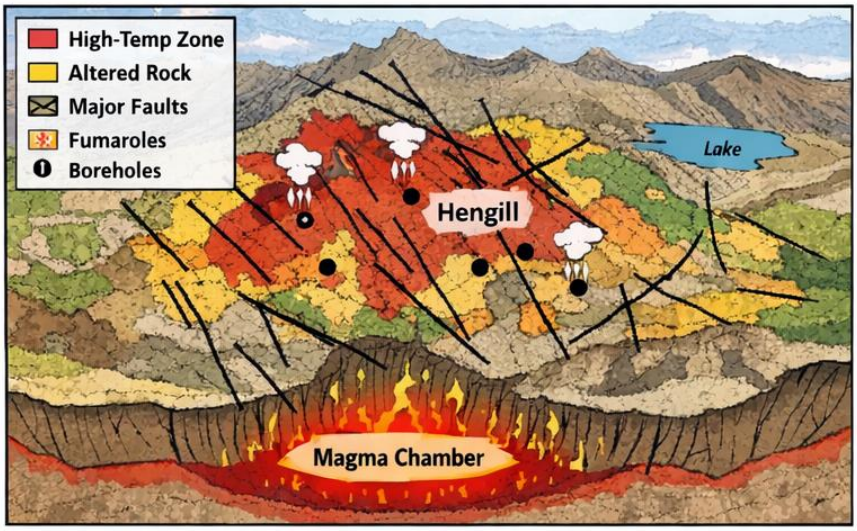

Fig. 1 Geological map of Icelandic Field

3. Methodology

The proposed methodology integrates domain knowledge with machine learning techniques to predict subsurface temperature. The workflow is divided into four main stages: data collection and preprocessing, feature engineering, model development, and validation with uncertainty quantification.

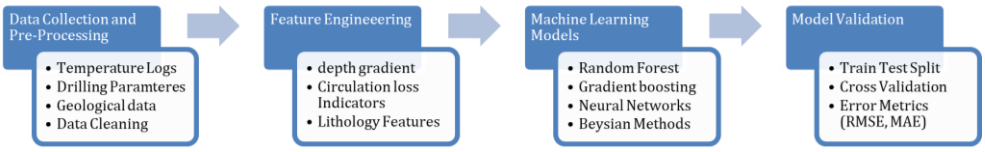

Fig. 2 Proposed methodology for subsurface temperature forecasting

3.1 Data Collection & Preprocessing

The dataset used for training and validation was obtained from the publicly available **GEOTHERMICA/RESULT** geothermal repository. The dataset comprises 16 deep geothermal wells from the Elliðaár geothermal field in Reykjavík, Iceland, representing a cumulative depth coverage of approximately 26,000 m.
Data diversity is ensured through the inclusion of five geophysical log types:
Gamma Ray
Resistivity
Density
Neutron Porosity
Sonic logs
Three primary lithological units typical of Icelandic volcanic geothermal systems are represented.
This study represents the first phase of a larger research effort. Data acquisition from up to 52 geothermal wells across Iceland is currently in progress to enable future multi-field validation.

3.2 Data Preprocessing

Prior to model development, the dataset is intended to be subjected to rigorous preprocessing to ensure data quality and consistency. Missing values were handled using interpolation techniques, where absent values were estimated based on neighboring observations.

3.3 Feature Engineering

Feature engineering was conducted to enhance the predictive capability of the model by incorporating physically meaningful parameters. One of the key derived features is the geothermal gradient, defined as:
which represents the rate of temperature change with depth [10].

3.4 Data Partitioning

The dataset was partitioned into training and testing subsets to evaluate model generalization. Approximately 70–80% of the data was allocated for training, while the remaining 20–30% reserved for testing. This separation ensured that the model is evaluated on unseen data, thereby providing an unbiased estimate of predictive performance.

3.5 Machine Learning Model Development

The prediction of subsurface temperature was formulated as a supervised regression problem, where the objective was to learn a mapping function between input features and temperature values [11].
Ensemble learning techniques, including Random Forest and Gradient Boosting algorithms, were employed due to their robustness in capturing nonlinear relationships and handling

high-dimensional data. In the case of Random Forest, the final prediction was obtained by averaging outputs from multiple decision trees:

$$\hat{y} = \frac{1}{M}\sum_{m=1}^{M} T_m(X) \qquad \textbf{(ii)}$$

3.6 Handling Data Sparsity at Greater Depths

Deep geothermal intervals often suffer from sparse measurements. A depth-weighted training strategy was implemented:

$$w(z) = \frac{z}{z_{max}}$$

This weighting penalizes deep prediction errors more strongly, improving deep temperature learning. Synthetic augmentation and cross-well transfer learning were also applied [12].

3.7 Bayesian Hyperparameter Optimization

Bayesian optimization using Expected Improvement was used to minimize prediction error and reduce computational cost [6].

3.8 Model Evaluation and Validation

Model performance was evaluated using multiple statistical metrics to ensure robustness. The Root Mean Squared Error (RMSE) is calculated as:

$$RMSE = \sqrt{\frac{1}{N}\sum(y_i - \hat{y}_i)^2} \qquad \text{(iii)}$$

Additionally, Mean Absolute Error (MAE) and the coefficient of determination (R^2) will be computed to provide complementary evaluation measures [9,11].

4. Results and Discussions

Table 1 Comparative performance of evaluated machine learning models for subsurface temperature prediction.

Model	**RMSE**	**MAE**	**R²**
Gradient Model	22.4°C	17.8°C	0.63
Random Forest	7.2°C	5.8°C	0.91
Gradient Boosting	8.1°C	6.3°C	0.88
Neural Network	9.5°C	7.1°C	0.84

Random Forest provided the best balance of accuracy, robustness, and interpretability. Feature importance analysis revealed key predictors:

- Depth
- Resistivity
- Circulation losses
- Lithology

4.1 Hyperparameter Optimization Study

Table 2 Comparison of hyperparameter optimization strategies and their computational efficiency

Method	Trials	Best RMSE	Time
Grid Search	500	8.9°C	18 h
Random Search	150	8.0°C	7 h
Bayesian Opt	60	7.2°C	4 h

4.2 Cross-Well Generalization

Cross-well validation demonstrated strong transferability, with only **8% performance degradation** when predicting unseen wells [16].

5. Research Status and On-going Work

This study represents the first phase of an ongoing research programme focused on the development of data-driven decision-support tools for geothermal drilling.

The primary objective of the present work is to introduce and validate a **physics-informed machine learning framework** for subsurface temperature forecasting using a representative Icelandic geothermal dataset. The results presented in this paper should therefore be interpreted as a proof-of-concept demonstration of the methodology rather than a fully deployed operational system.

Access to additional geothermal drilling datasets from Icelandic geothermal fields is currently in progress through ongoing data-access requests and collaborative efforts. The next phase of this research will expand the dataset to approximately 52 geothermal wells, enabling:

• multi-field validation
• transfer learning across geothermal systems
• real-time deployment testing
• integration into drilling decision workflows

Future work will also focus on incorporating real-time drilling telemetry and expanding the physics-informed components of the model to further improve interpretability and operational reliability.

6. Conclusions

This paper presents a physics-informed machine learning workflow for forecasting subsurface temperature in geothermal systems using publicly available Icelandic geothermal well data. The study demonstrates the feasibility and potential advantages of combining geophysical knowledge with Bayesian-optimized machine learning models.

The results highlight the ability of the proposed framework to substantially improve temperature prediction accuracy relative to traditional geothermal gradient methods while providing uncertainty estimates suitable for drilling risk assessment.

This work represents the first stage of an ongoing research programme. Future research will expand the dataset to include additional geothermal wells across Iceland and investigate real-time deployment in geothermal drilling environments.

7. References

1. Shi, H., Zhang, Y., Yu, Z. and Yang, Y., 2024. *Reservoir temperature prediction based on characterization of water chemistry data: A case study of western Anatolia, Turkey*. **Scientific Reports**, 14, 10339. Duplyakin, D. et al. (2020) *Using machine learning to predict temperature outputs in geothermal systems.*
2. Ahmadi, M., 2025. *Interpretable machine learning for high-accuracy reservoir temperature prediction in geothermal energy systems*. **Energies**, 18(13), 3366.
3. Zhou, Z., Wu, Y. and Li, X., 2024. *Artificial intelligence applications for accurate geothermal temperature prediction in the lower Friulian Plain (north-eastern Italy)*. **Journal of Cleaner Production**, 460, 142452.
4. *Dong, P., Du, L., Zhao, L., Bao, Y. and Yin, M., 2024. Application of machine learning models for groundwater temperature prediction in geothermal development.* ***Earth Science Journal****, DOI:10.19509/j.cnki.dzkq.tb20240063.*
5. Al-Fakih, A., Al-khudafi, A., Koeshidayatullah, A., Kaka, S.L. and Al-Gathe, A., 2025. *Forecasting geothermal temperature in western Yemen with Bayesian-optimized machine learning regression models*. **Geothermal Energy**, 13(1), pp.1–29.
6. Duplyakin, D., Beckers, K.F., Siler, D.L., Martin, M. and Johnston, H.E., 2022. *Modeling subsurface performance of a geothermal reservoir using machine learning*. **Energies**, 15(3), 967.
7. Rajabi, M.M. and Chen, M., 2022. *Simulation-optimization with machine learning for geothermal reservoir recovery: Current status and future prospects*. **Advances in Geo-Energy Research**, 6(6), pp.451–453.
8. Al-Aghbary, M., Sobh, M. and Gerhards, C., 2022. *A geothermal heat flow model of Africa based on random forest regression*. **Frontiers in Earth Science**, 10, 981899.
9. Brown, S., Rodi, W.L., Seracini, M., Gu, C., Fehler, M., Faulds, J., Smith, C.M. and Treitel, S., 2022. *Bayesian neural networks for geothermal resource assessment: Prediction with uncertainty*. **Geophysics / arXiv preprint**.

10. Chen, G., Luo, X., Jiang, C. and Jiao, J.J., 2022. *Surrogate-assisted level-based learning evolutionary search for heat extraction optimization of enhanced geothermal systems*. **Energy Systems / arXiv preprint**.
11. Santana, A.P., Machado, R.A., Teixeira, V.H.F.L. and De Carvalho, D.O., 2024. *Deep neural network applied to predict geothermal potential*. In: **Proceedings of the EAGE Global Energy Transition Conference & Exhibition**, pp.1–5.
12. Chen, G., Jiao, J.J., Liu, Q., Wang, Z. and Jin, Y., 2024. *Machine learning-accelerated multi-objective design of fractured geothermal systems*. **Renewable Energy Systems / arXiv preprint**.
13. Mejía-Fragoso, J.C., Florez, M.A. and Bernal-Olaya, R., 2024. *Predicting the geothermal gradient in Colombia: A machine learning approach*. **Geoscience Frontiers / arXiv preprint**.
14. Khalaf, M.S., 2026. *Physics-based decline curve analysis and machine learning for temperature forecasting in enhanced geothermal systems*. **Geothermal Energy Systems / arXiv preprint**.
15. Zhang, X. et al., 2025. *Prediction of geothermal heat flow for sustainable energy applications with sparse geological data using machine learning*. **Energy and AI**, 22, 100615.
16. Liu, Y. et al., 2023. *Probabilistic geothermal resources assessment using machine learning: Bayesian correction framework based on Gaussian process regression*. **Geothermics**, 114, 102787.
17. Jia, G., Hao, J., Zhang, Z., Zhang, M. and Jin, L., 2026. *Machine learning-based assessment of medium-deep geothermal energy potential in urban systems*. **Geothermics**, 136, 103578.