# POLITECNICO DI TORINO

## Master of Science in Petroleum and Mining Engineering

### Master's Degree Thesis

**Machine Learning (ML) for subsurface geothermal resource analysis and development**

**Candidate:** *Salman Mirzayev (S301465)*



**University Supervisor:**
*Glenda Taddia*

***Co-supervisor:***
*Martina Gizzi*

**Company Supervisor:**
*Ivo Colombo*

***Co-supervisors**:*
*Ilgar Hasanov*
*Elisabetta Billotta*

*Academic Year 2021/2022*

# Abstract

This dissertation presents a comprehensive examination of the integration of Machine Learning (ML) methodologies to enhance the analysis and development of subsurface geothermal resources, a pivotal element within the renewable energy spectrum. Despite geothermal energy's substantial potential for sustainable energy production, its exploitation is impeded by numerous geological and technical obstacles. This research endeavors to overcome these impediments by harnessing sophisticated ML algorithms, aiming to augment the predictability and operational efficiency in the exploration and characterization of geothermal resources.

Utilizing an extensive dataset from the Northeastern United States, the study conducts a thorough analysis and refinement of prevailing approaches in data preprocessing, feature engineering, and hyperparameter optimization. It assesses the efficacy of various ML models in forecasting subsurface temperatures and geothermal gradients, underscoring the significance of meticulous data examination and model refinement strategies, including outlier detection, data normalization, and the employment of grid search techniques for hyperparameter fine-tuning.

The outcomes reveal that ML applications can markedly improve the precision and dependability of predictions concerning geothermal resources, thereby diminishing the financial and technical uncertainties inherent in geothermal project development. The enhanced predictive models formulated through this research facilitate more strategic decision-making and resource allocation within the geothermal energy domain.

By melding conventional geothermal resource assessment methodologies with the latest in ML innovations, this work establishes a foundational framework for significant advancements in the sustainability and efficacy of geothermal energy, reinforcing its role as an indispensable component of the global renewable energy portfolio.

## Acknowledgements

I hereby extend my profound appreciation for the unwavering support and guidance that I have been fortunate to receive throughout the course of my thesis work. This endeavor has not only culminated in the completion of this thesis but has also significantly contributed towards a cause deeply aligned with my conviction for fostering a more sustainable future.

My sincere thanks are directed towards my esteemed supervisors, Taddia Glenda and Ivo Colombo, whose dedicated involvement and invaluable guidance have been instrumental from inception to the culmination of this thesis. Additionally, my co-supervisors, Martina Gizzi, Ilgar Hasanov, and Elisabetta Bilotta, deserve special recognition for their astute insights and steadfast support throughout this scholarly journey.

I am particularly grateful to the staff at Geolog International for their technical support, which has been pivotal in enhancing the quality of my research. Their expertise and willingness to assist have been of paramount importance.

Concluding my academic pursuits with this thesis fills me with a sense of satisfaction and eagerness for the future endeavors that lie ahead. This milestone, however, would not have been attainable without the enduring encouragement and support of my parents and mentors, to whom I owe an immense debt of gratitude. Their belief in my abilities and continuous encouragement have been the cornerstone of my achievements.

# Index

# List of Tables

# List of Figures

# Chapter I

## Introduction

### 1.1 General Introduction

Since its initial development in Larderello, Italy, in 1904, geothermal energy has emerged as a pioneering, entirely renewable energy source for the generation of electricity [1]. Characterized by its continuous availability, geothermal power provides a dependable foundation for electricity production, offering direct heating capabilities and featuring a comparatively low levelized cost of electricity. Moreover, the process of harnessing geothermal energy from terrestrial sources is associated with negligible emissions, underscoring its environmental benefits [2].

However, the geothermal sector encounters complex geological and engineering obstacles throughout the exploration and exploitation phases of geothermal field development. These challenges are primarily due to significant subsurface technical uncertainties, which consequently elevate the financial risks during the initial phases of geothermal resource exploration. Figure 1.1 delineates the sequential stages of geothermal resource development, highlighting the incremental percentage of costs attributed to each phase alongside the concomitant project and financial risks.

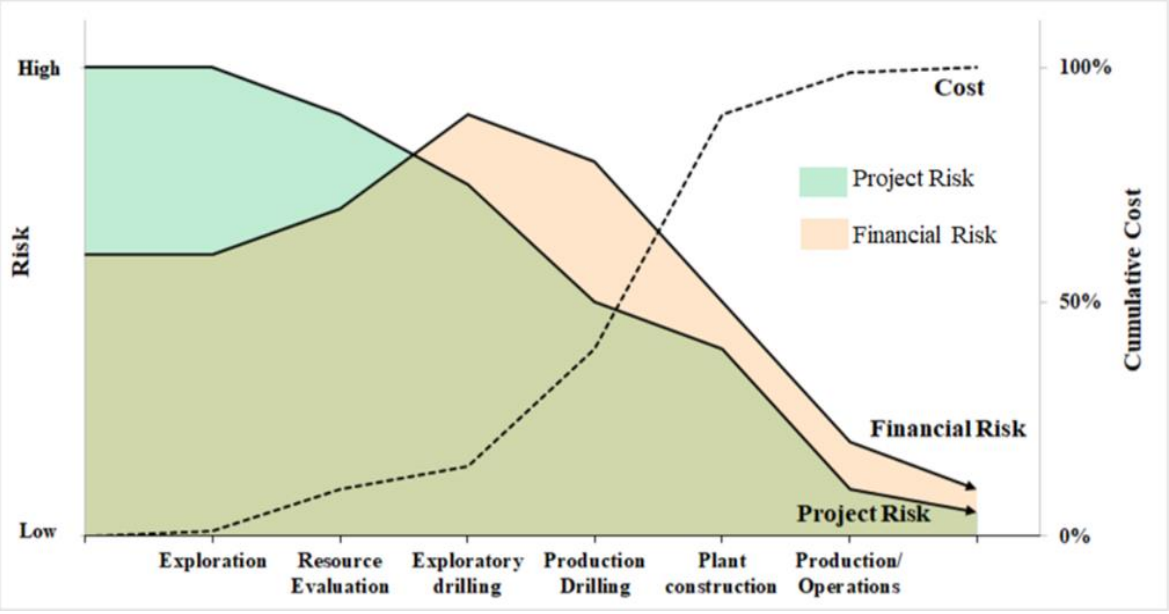


**Figure 1.1.** The stages of geothermal resource development, the estimated cumulative cost involved (as a percentage of the overall project cost), and the risks associated with each stage. Figure adapted from ESMAP (2012) [14], Witherbee (2012) [15], and the World Bank (2019) [16].

Here, project risks refer to the likelihood of failing to identify a geothermal resource of viable commercial magnitude, whereas financial risks denote the potential for financial losses or yields falling short of projections. The figure outlines the medium to high project and financial risks, particularly in the initial stages of geothermal resource development. This high-risk environment poses significant obstacles to securing private investment funds for geothermal projects until risks are mitigated to an acceptable level. This mitigation typically occurs after the drilling of an exploratory well confirms the presence of a commercially viable resource. Consequently, many geothermal projects remain undeveloped due to the inadequacy or capital intensity of existing technologies for resource discovery and development risk reduction.

Most geothermal brownfields in existence were located by performing drills close to geothermal manifestations at the surface, such as hot springs, fumaroles, and deposits from past geothermal activity. Occasionally, these areas were also discovered by entities in search of water, minerals, or fossil fuels. The challenge with brownfields primarily lies in the reduced efficiency of geothermal power stations, which is due to entropy-related inefficiencies stemming from temperature discrepancies [3].

Innovations in technology, such as the development of double flash, triple flash, hybrid geopressure/geothermal systems, and binary cycle plant designs, have played a pivotal role in enhancing the global installed capacity of geothermal energy. By 2020, these advancements helped raise the total geothermal capacity to an estimated 16 gigawatts of electrical energy worldwide [4]. Particularly, binary cycle power generation systems have markedly decreased the threshold temperatures required for resource exploitation. This technological advancement has enabled the economically viable development of geothermal resources previously classified as "low-grade," thereby expanding the scope of geothermal energy production to include previously untapped sources [5]. However, despite these advancements, current technologies have not fully tapped into the extensive potential offered by geothermal energy sources.

In the context of the United States, current forecasts indicate that adhering to conventional operational strategies will result in a geothermal generation capacity of

approximately 6 gigawatts-electric (GWe) by the year 2050. Nevertheless, the acceleration of development timelines for geothermal projects could lead to a substantial increase in capacity, potentially exceeding 13 GWe. Furthermore, the synergistic effect of expedited project timelines combined with advancements in technology could amplify the geothermal generation capacity to an estimated 60 GWe [6]. This highlights the considerable yet unexploited potential within the geothermal industry, pointing to the critical necessity for inventive strategies to accelerate its growth. Significant investigative efforts into technological progressions for the exploitation of geothermal resources have led to the emergence of pioneering Enhanced Geothermal Systems (EGS), innovative operational models for power plants such as hybrid systems that capitalize on the co-production capabilities from pre-existing oil and gas infrastructures, and the recovery of essential materials from geothermal brines produced during extraction [7]. Concurrently, there has been a significant increase in the deployment of sensors, along with the enhancement of data acquisition, storage, and processing capabilities, all aimed at improving the characterization of the subsurface [8].

The escalating requirement for the integration of vast quantities of geologic and geophysical information has prompted a paradigm shift among geothermal industry leaders. They are moving away from traditional reliance on expert judgment, standard modeling techniques, and statistical analysis, towards the adoption of advanced Artificial Intelligence (AI) technologies. It stands to reason that AI could play an indispensable role in accelerating the timelines for geothermal project development and in the improvement of geothermal energy extraction technologies.

The geothermal energy domain encounters technical obstacles similar to those faced by the oil and gas sector, where Artificial Intelligence (AI) has been utilized to mitigate risks and curtail expenses. Sircar et al. illustrated the capacity of AI to diminish exploration risks and augment the success rates of exploration wells [9]. Furthermore, Crow et al. highlighted the role of AI in advancing automated drilling technologies, which has led to enhancements in penetration rates, tripping speeds, and a reduction in drilling expenses [10]. Heghedus et al. presented the significant contribution of AI in the monitoring of reservoirs, while Zhang et al. elucidated how AI can streamline the processes of rock physics inversion [11] [12].

Given the demonstrated success of AI in mitigating risks, enhancing operational efficiencies, and reducing costs in the oil and gas sector, there exists a compelling rationale to explore how similar advantages could be applied within the geothermal energy field. This exploration aims specifically to pinpoint where AI might exert the most profound effects. The investigative efforts have included a detailed review and enhancement of a conceptual framework for assessing the applicability of machine learning techniques in forecasting subsurface temperatures and geothermal gradients, particularly across the Northeastern United States. The core objective of this research focused on leveraging bottom-hole temperature information from extensive oil and gas well datasets to create complex models of heat flow and maps of temperature at various depths. Such models and maps are instrumental in identifying and outlining areas with high geothermal potential within the targeted geographic locale. Driven by a steadfast commitment to deepening scientific knowledge, our study embarked on an extensive journey into the complex world of machine learning models, meticulously designed for the precise prediction of temperature-at-depth and geothermal gradient parameters. This venture was initiated in response to the acknowledgment of existing uncertainties and the simplifying assumptions prevalent in modern physics-based models, an understanding that formed the foundation of our research efforts.

Integral to the overarching thesis, our investigation ventured into the nuanced intricacies of model performance, conducting rigorous analyses and assessments aimed at unraveling the predictive capabilities embedded within the chosen models. This multifaceted approach was underpinned by a strategic objective: to discern the most efficacious machine learning methodologies capable of providing precise insights into the complex interplay of temperature distribution and geothermal characteristics within the Northeastern United States.

This pursuit assumed heightened significance in light of the aforementioned uncertainties and simplifications associated with prevailing physics-based models. Acknowledging the challenges posed by data heterogeneity, model complexity, and inherent subsurface uncertainties, our research undertook a methodical investigation. Innovative solutions were strategically implemented, encompassing the harmonization of diverse datasets, streamlining the intricacies of machine

learning model architecture, and devising robust methodologies specifically tailored to address the unique uncertainties embedded in subsurface conditions.

The central achievement of this study is the pinpointing and correction of gaps in the current body of research regarding the exploratory analysis of machine learning techniques for the prediction of subsurface temperature and geothermal gradient across the Northeastern United States [13]. By systematically addressing coding errors and implementing an alternative approach, this research has significantly enhanced the accuracy and efficacy of predictive models in this domain. This not only advances the understanding of exploratory analyses within geothermal studies but also provides a more robust and precise framework for future investigations.
Through a meticulous approach that involves scrutinizing and refining methodologies employed in previous studies, various aspects, such as outlier removal, data preprocessing, and hyper-parameter tuning, were systematically reevaluated. These revisions were underpinned by the recognition that methodological choices profoundly impact model outcomes. By applying a novel methodology, this research has yielded more precise models and effective solutions, surpassing the limitations of prior approaches.

Data collection involved a comprehensive analysis of relevant literature and datasets specific to the Northeastern United States. This information served as the foundation for identifying shortcomings in existing methodologies and implementing strategic changes. The conclusions drawn from these analyses not only shed light on the nuances of subsurface temperature and geothermal gradient prediction but also offer valuable insights for researchers and practitioners in the geothermal energy sector. The refined methodology introduced in this work contributes to the robustness of predictive models, laying the groundwork for improved decision-making processes and sustainable utilization of geothermal resources.

## 1.2 Foundational Insights into Artificial Intelligence

Artificial Intelligence (AI) encompasses a specialized area within science and engineering aimed at crafting machines that demonstrate intelligent behavior, especially in the form of smart computer programs. The field has seen a notable expansion in its application across numerous industries recently. As detailed by the AI Index, which meticulously tracks the impact of AI on different fronts, including the economic growth of nations, job creation, diversity, and academic research, there's been an extraordinary increase in AI-related activities. A key indicator of this trend is the marked upsurge in global investment in AI capabilities [17].

The AI Index delineates a pronounced escalation in worldwide investment towards AI, signifying an enhanced emphasis and recognition of AI advancements. Illustratively, contractual engagements of the U.S. government with various federal entities experienced a notable augmentation, surging from approximately 500 million in 2017 to 1.8 billion in 2020. This trend evidences the substantial fiscal allocations and commitments undertaken to exploit the potential of artificial intelligence in diverse sectors [18].

## 1.3 Understanding of Machine Learning

Machine Learning (ML), a distinct discipline within artificial intelligence, adopts statistical methodologies to empower computational models to derive insights from data, as depicted in Figure 1.2. At their core, ML algorithms embody advanced mathematical constructs, endowed with a multitude of parameters, that establish associations between input variables (features) and resultant outputs (targets). These targets encompass a wide array of outputs, from quantitative assessments and probabilistic determinations to the discernment of patterns or the establishment of innovative classifications. The essence of training an ML model lies in the iterative refinement of these parameters, aiming to refine the model's capability to accurately correlate inputs with their anticipated outcomes [19].

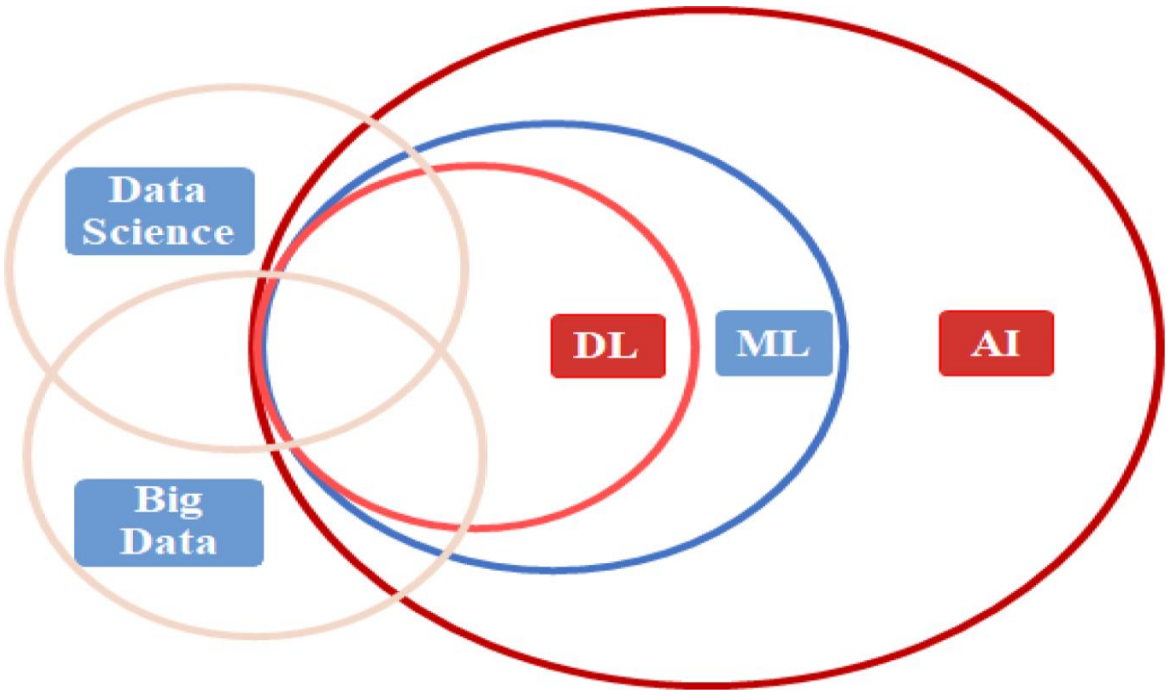


**Figure 1.2.** Hierarchical Representation of AI, ML, DL, Data Science, and Big Data

Deep Learning (DL), a facet within machine learning, employs artificial neural networks inspired by the structure of the human brain. These networks, often comprising numerous layers (referred to as deep layers), are trained using extensive datasets. In technical and earth science fields, two key categories of machine learning algorithms are integral: supervised and unsupervised learning. In supervised learning, models are trained to make predictions by correlating measurable features with labels through numerous examples and employing suitable algorithms. Conversely, unsupervised learning operates without the need for labeled data; instead, algorithms discern patterns and connections within the dataset.

### 1.3.1 Random Forest Algorithm

Random forests, alternatively referred to as random decision forests, represent an ensemble learning approach utilized across classification, regression, and a spectrum of other tasks. This technique revolves around the creation of multiple decision trees during the training stage. In classification endeavors, the collective decision of these trees determines the class assignment. Conversely, for regression tasks, it consolidates predictions from individual trees by means of averaging, effectively counteracting the propensity of decision trees to overfit their training data.

The inception of random decision forests traces back to 1995, credited to Tin Kam Ho, who introduced the method employing the random subspace technique. Ho's utilization of "stochastic discrimination" for classification draws inspiration from the conceptual framework originally proposed by Eugene Kleinberg.

The algorithm enhancements introduced by Leo Breiman and Adele Cutler culminated in the establishment of "Random Forests" as a trademark in 2006. Their adaptation integrates Breiman's "bagging" concept with random feature selection, an idea initially introduced by Ho and later independently explored by Amit and Geman. This synthesis endeavors to generate an ensemble of decision trees with regulated variance, thereby advancing upon the principles underlying decision tree learning. Notably, decision trees offer inherent advantages in various machine learning applications owing to their scalability, resistance to irrelevant features, and the capacity to produce interpretable models. However, standalone trees, particularly when extensively grown, frequently encounter overfitting issues, characterized by the assimilation of intricate patterns that inadequately generalize to new data. Random forests address this concern through the aggregation of results from multiple trees, each trained on subsets of the same dataset. This strategy effectively mitigates variance without imposing substantial bias or compromising model interpretability. Consequently, this trade-off typically engenders a noteworthy enhancement in model accuracy. [20].

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set X = $x_1$, ..., $x_n$ with responses Y = $y_1$, ..., $y_1$ bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For b = 1, ..., B:

1. Sample, with replacement, n training examples from X, Y; call these $x_b$, $y_b$

2. Train a classification or regression tree fb on $x_b, y_b$

After training, predictions for unseen samples $x'$ can be made by averaging the predictions from all the individual regression trees on $x_1$:

$$f = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

or by taking the plurality vote in the case of classification trees.

Improved model performance is achieved through the bootstrapping procedure, which effectively decreases model variance without introducing bias. This means that while a single tree's predictions are highly sensitive to noise within its training set, the average predictions of multiple trees are not, as long as these trees remain uncorrelated. Training multiple trees on the same dataset could lead to strong correlations among them (or even identical trees if the training algorithm is deterministic); however, bootstrap sampling serves to alleviate this issue by providing each tree with distinct training sets.

Additionally, an estimate of the uncertainty of the prediction can be made as the standard deviation of the predictions from all the individual regression trees on x':

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - f)^2}{B-1}}$$

The number of samples/trees, B, is a free parameter. Typically, a few hundred to several thousand trees are used, depending on the size and nature of the training set. An optimal number of trees B can be found using cross-validation, or by observing the out-of-bag error: the mean prediction error on each training sample $x_i$, using only the trees that did not have $x_i$ in their bootstrap sample. The training and test error tend to level off after some number of trees have been fit [21]. The preceding method outlines the original bagging algorithm designed for trees. However, random forests incorporate an additional bagging technique: they utilize a modified tree learning algorithm that, at each potential split during the learning phase, opts for a random subset of features. This method, also known as "feature bagging", aims to tackle the issue of tree correlation within a typical bootstrap sample. Specifically, if certain features strongly predict the response variable (target output), they are consistently chosen across many trees in the ensemble, resulting

in correlated trees. Ho provides an analysis of how bagging and random subspace projection contribute to accuracy improvements under various circumstance [22].

Typically, for a classification problem with p features, $\sqrt{p}$ (rounded down) features are used in each split. For regression problems the inventors recommend $p/3$ (rounded down) with a minimum node size of 5 as the default. In practice, the best values for these parameters should be tuned on a case-to-case basis for every problem [23].

The inclusion of an additional layer of randomization gives rise to extremely randomized trees, denoted as ExtraTrees. While sharing resemblances with traditional random forests as an ensemble of individual trees, two primary differentiating factors emerge: firstly, each tree is trained utilizing the entire learning sample as opposed to a bootstrap sample; secondly, the top-down splitting process within the tree learner undergoes randomization. Instead of computing the locally optimal cut-point for each considered feature, a random cut-point is selected. This selection is drawn from a uniform distribution within the empirical range of the feature within the tree's training set. Subsequently, among all the randomly generated splits, the one yielding the highest score is elected to split the node. Analogous to conventional random forests, the number of randomly selected features to be assessed at each node can be stipulated. Default values for this parameter are $\sqrt{p}$ for classification and p for regression, where p is the number of features in the model [24]. As the field of machine learning advances, random forests maintain their pivotal role within the ensemble learning framework. Current scholarly endeavors are focused on optimizing their efficiency, scalability, and predictive accuracy.

### 1.3.2 XGBoost Algorithm

XGBoost (eXtreme Gradient Boosting) stands as an open-source software library that offers a sophisticated gradient boosting framework across various programming languages, including C++, Java, Python, R, Julia, Perl, and Scala. Its project description underscores the objective of providing a "Scalable, Portable and Distributed Gradient Boosting (GBM, GBRT, GBDT) Library." It demonstrates versatility by operating seamlessly on both single machines and distributed processing frameworks like Apache Hadoop, Apache Spark, Apache Flink, and Dask. Throughout the mid-2010s, XGBoost garnered widespread acclaim and attention, emerging as the algorithm of choice for numerous victorious teams in machine learning competitions [25].

The distinctive characteristics of XGBoost that set it apart from other gradient boosting algorithms encompass several key aspects [26];[27]; [28]:

- **Clever Penalization of Trees:** XGBoost applies a regularization term to its objective function, penalizing complex models to control over-fitting. This regularization term is a key differentiator from other gradient boosting frameworks that might not directly address model complexity in their optimization process. The regularization includes both L1 (lasso regression) and L2 (ridge regression) terms, contributing to the production of simpler, more generalizable models.
- **Proportional Shrinking of Leaf Nodes:** XGBoost employs a technique known as "shrinkage" or "learning rate" that scales down the weights of new trees added to the model. This approach helps in making the boosting process more conservative, reducing overfitting and allowing for more robust models. The idea is to give subsequent trees a chance to learn from the residuals left by the predecessors, improving the model iteratively.
- **Newton Boosting:** Unlike traditional gradient boosting, which uses only first-order gradient information, XGBoost utilizes second-order information (Hessians) in its optimization algorithm. This method, often referred to as Newton Boosting, allows for a more accurate search for the minimum loss function, especially in cases where the loss function's curvature is significant.
- **Extra Randomization Parameter:** XGBoost introduces an additional layer of randomness by allowing subsampling of the training data and features at each split. This feature, akin to the randomization in Random Forests, helps in preventing overfitting and adds another layer of variance reduction to the model.

- **Implementation on Single, Distributed Systems, and Out-of-Core Computation:** XGBoost is designed for efficiency and scalability. It can run on a single machine, take advantage of multicore CPUs, operate on distributed systems like Hadoop and Spark, and even handle datasets that do not fit into memory, thanks to its out-of-core computation capabilities.
- **Automatic Feature Selection:** During the training process, XGBoost automatically handles feature selection by assigning importance scores to each feature. Features contributing significantly to the model's predictive power are utilized, while less important ones are pruned away. This built-in feature selection mechanism simplifies the modeling process and can lead to more parsimonious models.
- **Theoretically Justified Weighted Quantile Sketching for Efficient Computation:** XGBoost implements an advanced algorithm for quantile sketching, which is crucial for efficiently finding the best split points in the presence of weighted data. This feature is particularly important for handling large datasets and ensures that XGBoost remains computationally efficient even as data size grows.
- **Parallel Tree Structure Boosting with Sparsity:** XGBoost's algorithm can build trees in parallel, significantly speeding up the training process. Furthermore, it is designed to handle sparse data natively, optimizing both the storage and computation for sparse features. This capability makes XGBoost highly suitable for handling high-dimensional data with many missing values.
- **Efficient Cacheable Block Structure for Decision Tree Training:** XGBoost organizes data into a block structure, optimized for cache usage on modern CPUs. This design choice leads to high computational efficiency, as it minimizes memory access times during the construction of trees, making the algorithm faster and more scalable.

XGBoost diverges from conventional gradient boosting methodologies by adopting a Newton-Raphson methodology within function space. Unlike the prevalent gradient descent approach, XGBoost leverages a second-order Taylor approximation within the loss function to establish its association with the Newton-Raphson technique [25].

### 1.3.3 LightGBM Algorithm

Originally developed by Microsoft, LightGBM, an acronym for light gradient-boosting machine, is a freely available and open-source distributed gradient-boosting framework designed for machine learning. Rooted in decision tree algorithms, it caters to tasks such as ranking, classification, and various other machine learning endeavors. The primary focus of its development lies in enhancing performance and scalability, rendering LightGBM an invaluable tool for tackling large-scale and computationally intensive learning tasks [29].

LightGBM, like XGBoost, boasts numerous advantages in terms of optimization, flexibility, and performance enhancement techniques. However, it stands out from XGBoost in its tree construction strategy, which significantly influences its efficiency and memory consumption.

- **Tree Construction Approach:** LightGBM adopts a leaf-wise tree growth strategy rather than the traditional level-wise approach used by many other boosting implementations, including XGBoost. In the leaf-wise strategy, instead of expanding the tree level by level, LightGBM selects the leaf node that leads to the maximum reduction in loss. This approach tends to produce deeper trees with fewer nodes, allowing for more intricate representations of the data distribution [30].
- **Decision Tree Learning Algorithm:** Unlike XGBoost and other implementations that use sorted-based decision tree learning, LightGBM employs a histogram-based decision tree learning algorithm. This algorithm offers significant advantages in terms of both efficiency and memory usage. Instead of sorting feature values to find the best split points, LightGBM discretizes the continuous features into bins and constructs histograms for efficient computation of split points [31].
- **Gradient-Based One-Side Sampling (GOSS):** LightGBM introduces the Gradient-Based One-Side Sampling technique to improve training speed while maintaining accuracy. GOSS identifies and retains only the instances with large gradients for constructing each decision tree, effectively reducing the computational overhead while focusing on the most informative data points [32].
- **Exclusive Feature Bundling (EFB):** Exclusive Feature Bundling is another innovative technique employed by LightGBM to enhance efficiency. EFB combines

correlated features into exclusive bundles, reducing the number of features and thus speeding up the training process. By bundling features intelligently, LightGBM reduces redundancy and computational overhead without sacrificing predictive performance [32].

### 1.3.4 Neural Networks

Characterized as an artificial neural network with multiple layers positioned between the input and output layers, a deep neural network (DNN) comprises essential components such as neurons, synapses, weights, biases, and functions. These components collectively emulate the functionality of the human brain and are amenable to training, akin to other machine learning algorithms.

For instance, in the context of image classification, a DNN trained to recognize dog breeds undergoes a process where it evaluates the provided image, computes the probability distribution across various breed categories, and subsequently provides probabilistic predictions. Users have the flexibility to review and filter these probabilities based on specified thresholds, facilitating the determination of the proposed label. Each mathematical operation within a DNN, representing a layer, contributes to its depth, giving rise to the designation "deep" networks. Notably, DNN architectures excel in capturing complex non-linear relationships, generating compositional models that express objects as layered compositions of primitives [33] [34] [35].

In the realm of image classification, a Deep Neural Network (DNN) trained to recognize dog breeds provides a prime example. Within this context, the network meticulously evaluates the input image, computing the probability associated with each potential breed. Users are then empowered to review and refine the displayed probabilities based on specified criteria, such as a predetermined threshold, facilitating the determination of the proposed breed label for the image. Each computational operation within the DNN, symbolizing a distinct layer, contributes to the network's depth. Complex DNNs often feature numerous layers, earning them the designation of "deep" networks. This depth enables the amalgamation of features extracted from lower layers, potentially enabling the modeling of complex data with fewer units compared to shallower networks while maintaining comparable performance. Fundamentally, DNNs possess the capability to model intricate non-linear relationships, a feature attributed to their compositional architecture. This architectural paradigm facilitates the representation of objects as layered compositions of fundamental elements, thereby endowing the network with a profound depth of interpretative prowess. Significantly, research has underscored the exponential advantage of DNNs over shallow networks in approximating sparse multivariate polynomials, underscoring the efficacy of deep architectures in capturing and interpreting complex data structures [36] [37].

An array of deep architectures exists, each representing variants of fundamental approaches tailored to excel in specific domains. Comparing the performance of these architectures often proves challenging unless they undergo evaluation on identical datasets. Deep Neural Networks (DNNs), for instance, typically adhere to a feedforward structure, characterized by the unidirectional flow of data from input to output layers without recurrent connections. Initially, a DNN establishes a virtual map of neurons, assigning random numerical values, termed "weights," to the connections between them. These weights are subject to multiplication with inputs, yielding outputs within the range of 0 to 1. In cases where the network fails to accurately recognize a specific pattern, an algorithm intervenes to adjust the weights accordingly. In this manner, the algorithm iteratively adjusts certain parameters to enhance their influence until it discerns the precise mathematical manipulations necessary for comprehensive data processing [38].

While Deep Neural Networks (DNNs) excel in capturing complex patterns within data, their extensive depth and complexity can render them vulnerable to overfitting. This phenomenon occurs when the model mistakenly learns noise or irrelevant features present in the training data. To counteract this issue, diverse regularization techniques and optimization strategies are employed throughout the training phase [39]:

1. **Regularization Methods:**

- Ivakhnenko's unit pruning and weight decay are regularization techniques used to prevent overfitting by penalizing large weights or reducing the complexity of the model [39]
- Dropout regularization randomly drops out units (neurons) from the hidden layers during training, forcing the network to learn redundant representations and reducing the likelihood of overfitting [40].

2. D**ata Augmentation:**

- Data augmentation techniques such as cropping and rotating artificially expand the training dataset by generating variations of existing samples. This helps expose the model to a broader range of scenarios and reduces overfitting by providing more diverse training examples [41].

3. **Optimization Tricks:**

- DNNs require tuning various hyperparameters, including the network size (number of layers and units per layer), learning rate, and initial weights [42]. Exhaustively searching through this parameter space for optimal configurations

is often impractical due to computational constraints. Instead, techniques like batching, where gradients are computed on multiple training examples simultaneously, help accelerate training by exploiting parallelism in modern hardware architectures like GPUs or the Intel Xeon Phi [43]; [44].

### 1.3.5 Hyperparameter tuning

In the realm of machine learning, hyperparameter optimization, also known as tuning, involves selecting the most suitable set of hyperparameters for a learning algorithm. Hyperparameters are parameters utilized to govern the learning procedure [45].

The aim of hyperparameter optimization is to identify a combination of hyperparameters that leads to an optimal model, minimizing a predefined loss function on a given dataset of independent data. This process typically involves defining an objective function that accepts a set of hyperparameters as input and computes the associated loss [46]. To estimate the model's generalization performance, cross-validation is frequently employed, aiding in the selection of hyperparameter values that maximize this performance metric [47].

The conventional method for hyperparameter optimization has traditionally involved using grid search, also known as parameter sweep. This approach entails exhaustively exploring a predefined subset of the hyperparameter space of a learning algorithm [48]. To conduct grid search effectively, it's essential to have a performance metric guiding the process, usually evaluated through techniques like cross-validation on the training set or assessment on a separate hold-out validation set [49].

Given that the parameter space of a machine learning model can encompass real-valued or unbounded values for certain parameters, it may be necessary to establish manual bounds and discretization before employing grid search. For example, a typical soft-margin SVM classifier equipped with an RBF kernel has at least two hyperparameters that need to be tuned for good performance on unseen data: a regularization constant C and a kernel hyperparameter γ.

$$C \in \{10,100,1000\}$$
$$\gamma \in \{0.1,0.2,0.5,1.0\}$$

Grid search proceeds by training a Support Vector Machine (SVM) using every combination of (C, γ) from the Cartesian product of these two sets. The performance of each SVM is then assessed using a held-out validation set or through internal cross-validation on the training set, which involves training multiple SVMs per pair. Eventually,

the grid search algorithm identifies the configurations that yield the highest score in the validation process.

Despite its effectiveness, grid search is challenged by the curse of dimensionality. However, it often benefits from being embarrassingly parallel, as the hyperparameter settings it assesses are typically independent of each other.

# Chapter II

## Literature Overview

### 2.1 Advancements in Geothermal Exploration: From BHT Measurements to Machine Learning Applications

The measurement of Bottom-hole temperature (BHT) has emerged as a crucial factor in the analysis of subsurface temperatures, particularly in the evaluation of geothermal resources in the United States [50]. Extracted primarily from oil and gas wells, BHT data play a pivotal role, with the highest recorded temperature typically reported at the terminal drilled depth. Blackwell and Richards (2010) integrated BHT data with stratigraphic information in the northeastern United States, utilizing a simplistic thermal conductivity model to generate surface heat flux and temperature-at-depth maps [51]. Subsequent analyses by Jordan et al. extended this work to assess the risks and potentials of geothermal resources in New York, Pennsylvania, and West Virginia.

Despite the traditional concentration of geothermally active regions in the western United States, Jordan et al. (2016) highlighted the untapped potential in low-temperature geothermal regions in the northeast for various direct-use applications [52]. Snyder et al. further emphasized the potential reduction in electricity consumption through industrial and residential applications of geothermal energy, but financial challenges hinder the establishment of geothermal sites in the northeastern states [53]. Key parameters for geothermal exploration, such as heat flux and temperature-at-depth, are traditionally computed using a generalized thermal conductivity model which involves the correction of BHT data and the approximation of geological formation thickness and thermal conductivity values at each well's location through the Correlation of Stratigraphic Units of North America (COSUNA).

However, the long-standing applicability of the physics-based model is not without limitations, as noted by Stutz et al. (2012) and Blackwell and Richards (2010). The model's assumptions introduce uncertainties, particularly due to the lack of an easily applicable method for independently measuring the heat flux parameter, necessitating its

approximation through the thermal conductivity model using BHT data, as expressed in Equation (1).

$$Q_s = k(\frac{dT}{dZ}) \quad (1)$$

Recognizing the challenges in geothermal exploration, there is a growing reliance on machine learning and geostatistics to mitigate risk and uncertainty [54][55][56]. Despite the scarcity of comprehensive surveys focusing on risk analysis and geothermal site development machine learning has significantly contributed to the geothermal energy field [57][58][59][60][61][62].

Various machine learning and deep learning methodologies have been explored, particularly in the exploration and drilling phases. Progress has been observed in feedforward neural networks with one or two hidden layers, contributing to characterizing geomechanical properties, fault detection and interpretation, inversion of geophysical data, and lithofacies classification [63][64][65][66][67]. Perozzi et al. and Rezvanbehbahani et al. demonstrated the application of machine learning in geological interpretations and estimation of geothermal heat flux, respectively, with notable success [68].

In a unique approach, Assouline et al. utilized machine learning to map very shallow geothermal potential, focusing on the Random Forest method for predicting critical thermal variables across Switzerland. The dynamic landscape of technological advancements within geoscience practices has paved the way for machine learning applications to play a prospective role in shaping the future of geothermal energy development [69].

In light of this, our comprehensive review spanning the period from 2001 to 2022 aims to ascertain the nuanced utilization of algorithms in the geothermal energy field.

# Chapter III

## Methodology

In this study, we introduce a novel methodology employing advanced machine learning techniques for the prediction of subsurface temperatures, leveraging Bottom Hole Temperature (BHT) data derived from an extensive dataset comprising over 20,750 oil and gas wells in the northeastern United States [70]. To rigorously evaluate the performance of our machine learning models, a comparative analysis is conducted with an additional dataset containing vertical temperature profiles from 58 wells located in West Virginia [71].

The following workflow figure 3.1 encapsulates the systematic approach adopted in our study:

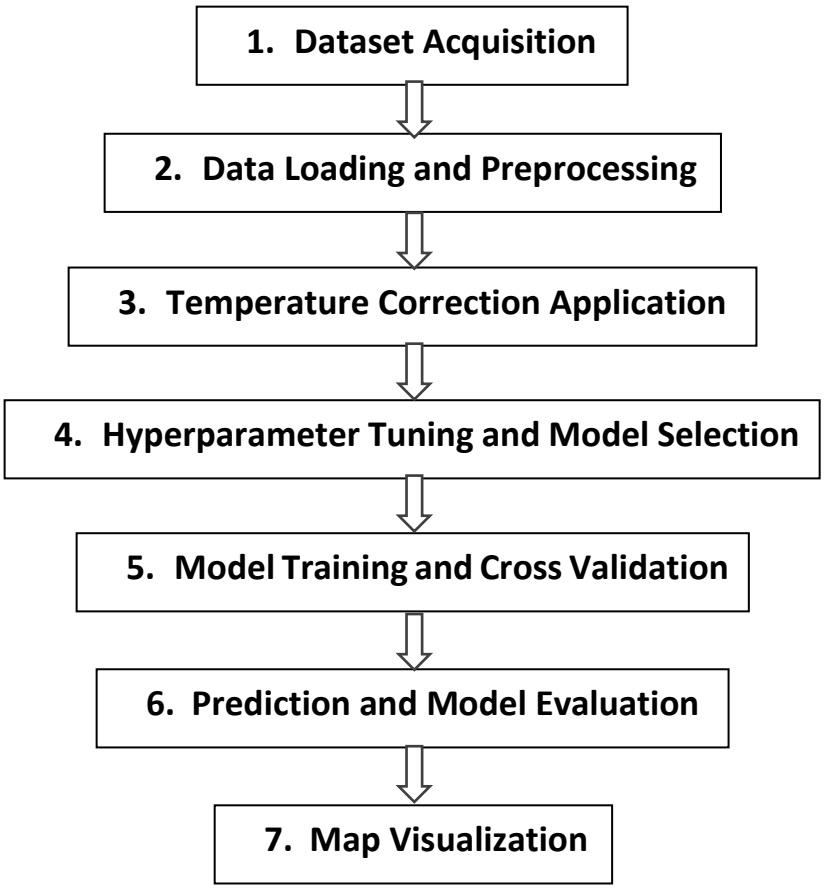


**Figure 3.1.** Machine Learning Pipeline.

For the primary analysis, we meticulously curated a comprehensive dataset featuring raw and corrected BHT, surface temperature, well identification numbers (API), latitude, longitude, geological setting information (including layer thickness and conductivity), and

various other pertinent parameters sourced from the expansive pool of 20,750 oil and gas wells in the northeastern region. Critical to the integrity of our machine learning models was the rigorous process of data loading and preprocessing. This step ensured the refinement of our datasets, where we methodically cleaned, normalized, and transformed the raw data to create a structured and analysis-ready dataset. This stage is instrumental in minimizing potential biases and errors that could compromise the model's performance. A pivotal aspect unique to geothermal data analysis is the application of temperature correction to the bottom-hole temperature (BHT) data. Given the perturbation caused by drilling, we employed state-of-the-art correction algorithms to adjust the BHT data, thereby ensuring that the temperatures used in our models reflected the true geothermal gradient. With the preprocessed data, we delved into hyperparameter tuning and model selection. Leveraging techniques such as cross-validation and grid search, we meticulously explored a variety of model configurations to ascertain the most efficacious parameters that would yield the most accurate predictions. The core of our methodological approach was the training of our machine learning models. By implementing a robust cross-validation framework, we systematically trained and validated our models, ensuring their ability to generalize and perform consistently across different data subsets. Following training, we progressed to predictions and model evaluations. Utilizing a selection of metrics tailored to our research objectives, we critically assessed the performance of our models. This phase was crucial in confirming the predictive prowess of our models and their applicability in real-world geothermal energy exploration. The culmination of our methodological journey was the visualization of predictions. By creating detailed geospatial maps depicting the predicted geothermal gradients, we provided a vivid and intuitive representation of our findings.

## 3.1. Data Description

### 3.1.1 Dataset - 1

Table 3.1 provides a comprehensive overview of crucial parameters after the removal of outliers, comprising a total of 55 features as specified. Particularly noteworthy is the integration of geological attributes by calculating the product of conductivity and thickness for each formation (features 6–55), which aligns with the tenets of thermal conductivity theory (Equation (1)). The dataset, sourced from the Geothermal Data Repository, underscores the value of open-access repositories in facilitating geothermal research endeavors [70].

| Variable number | Name | Unit | Source | Description | Type |
|---|---|---|---|---|---|
| 1 | BHTCorr | °C | Well log report | Corrected bottom-hole temperature | Label |
| 2 | LatDegree | - | Well log report | Lat degree of the well's location | Feature |
| 3 | LongDegree | - | Well log report | Long degree of the well's location | Feature |
| 4 | MeasureDepth | M | Well log report | The depth where BHT is recorded | Feature |
| 5 | SurfTemp | °C | Annual average temperature | Surface temperature at the well's location | Feature |
| 6 to 55 | KH | W/(°K) | Approximated from the data reported in Correlation of Stratigraphic Units of North America | Multiplication product of each layer's thickness with its corresponding thermal conductivity | Feature |

**Table 3.1.** Statistical summary of important parameters [13].

### 3.1.2 Dataset - 2

A distinct temperature-profile dataset was meticulously assembled, featuring data from an additional 58 wells strategically located across the West Virginia region, denoted by blue points in Figure 3.2 This dataset furnishes temperature profiles for each well within a defined depth interval, presenting mean and standard deviation values of 1167 and 511 meters, respectively. Data were meticulously sourced from the West Virginia Geological and Economic Survey provided in the LAS file format [71].

The temperature measurements, along with a suite of geological parameters, were systematically documented at various depths. This supplementary dataset assumes a pivotal role in benchmarking our machine learning models against results derived from a physics-based model. Noteworthy is the inclusion of 11 wells from the 58, whose BHT points are already present in the primary dataset (20,750 wells). The remainder of the wells were judiciously selected as new additions to facilitate a comparative analysis between physics-based and machine learning methodologies.

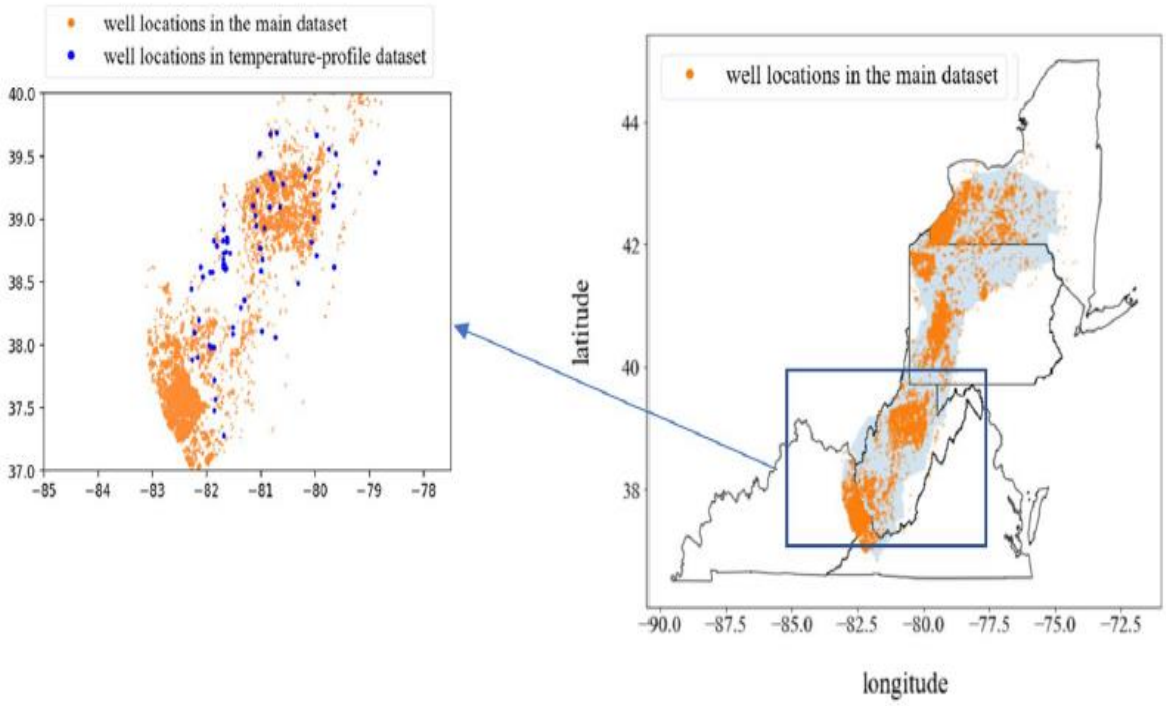


**Figure 3.2** Right plot represents the spread of oil and gas wells in the first dataset (containing 20750 BHT data points). In the left plot, the locations of newly obtained wells (with full temperature profile) are annotated using blue color [13].

## 3.2 Data Preprocessing

### 3.2.1 Bottom-hole temperature correction

In the context of Bottom-hole temperature (BHT) correction methodologies, Jordan et al. (2016) adopted a geographically stratified approach within the Appalachian Basin. This involved the partitioning of the basin into three distinct regions: West Virginia, Pennsylvania Rome Trough, and Allegheny Plateau. The authors crafted exclusive correction correlations tailored to the specific information available in each of these delineated regions. Notably, the Allegheny Plateau region provided a more comprehensive dataset, incorporating details on drilling fluids that were absent in the West Virginia section.

The correction process involved a meticulous statistical evaluation of a limited set of equilibrium well-log temperature measurements for each region. Subsequently, the authors formulated new sets of contextually appropriate BHT corrections. In the West Virginia region, a Generalized Least Square (GLS) regression model was meticulously fitted using Eq. (2) to encapsulate the nuanced relationships influencing BHT. Conversely, for the Pennsylvania Rome Trough, an absence of statistically significant relations with depth precluded the application of any adjustments.

$$T\_WVA = -1.99 + 0.00652z,\ 305\ m<z<2606m \quad (2)$$

Fortunately, in the Allegheny Plateau, the availability of drilling fluid data facilitated the formulation of correlation equation 2 tailored to different fluids. This bespoke approach, considering regional variations and the specificities of available, exemplifies the authors' commitment to precision and contextual relevance in refining BHT measurements within each distinct segment of the Appalachian Basin.

### 3.2.2 Outlier removal approach

Our approach to outlier removal involves the application of the Interquartile Range (IQR) technique, a method meticulously designed to identify outliers within our geothermal dataset, with a specific focus on the heat flow parameter. This methodology has been chosen over alternative techniques, such as the 3-sigma rule, due to its distinct advantages in outlier detection and its alignment with the unique characteristics of geothermal datasets.

The process commences with the implementation of the IQR method on our geothermal dataset, where outliers are identified based on their deviation from the median value of the heat flow parameter. Unlike the 3-sigma rule, which relies on standard deviation, the IQR method offers a more robust and versatile approach, allowing for a comprehensive capture of deviations within the dataset.

One of the key benefits of utilizing the IQR method is its ability to enhance the precision and accuracy of outlier identification. By focusing on the interquartile range, which represents the middle 50% of the data, the IQR method effectively identifies outliers that fall outside this range, providing a more refined understanding of the dataset's variability.

Furthermore, the adoption of the IQR method underscores our commitment to scientific rigor within the geothermal research domain. By systematically identifying outliers and removing them from the dataset, we ensure that subsequent analyses and interpretations are based on reliable and accurate data, thereby enhancing the credibility of our findings.

Overall, the IQR method serves as a robust tool for outlier removal in geothermal datasets, enabling us to uncover hidden patterns and insights that may have been obscured by noisy or erroneous data points. By employing this methodological approach, we strive to enhance the integrity and reliability of our geothermal exploration endeavors, ultimately contributing to the advancement of knowledge in this field.

- **IQR Outlier Removal Approach**

In our IQR (Interquartile Range) Outlier Removal Approach, we meticulously selected key features crucial for our analysis, including latitude, longitude, measurement depth, surface temperature, and heat flow. This deliberate choice aimed to capture diverse aspects of geothermal data, facilitating a comprehensive examination of the 'HeatFlow' values within our dataset.

To delve deeper into our dataset, we employed a line plot to visually inspect the raw 'HeatFlow' values (refer to Figure 3.3). This graphical representation served as our initial exploration tool, offering a visual overview of the distribution of 'HeatFlow' values across the dataset. By scrutinizing the line plot, we aimed to uncover insights into the underlying patterns and trends within the data, allowing us to identify any potential outliers or irregularities that warranted further investigation.

Through this visual exploration, we sought to gain a nuanced understanding of the variability present in the 'HeatFlow' values and identify data points that exhibited notable deviations from the expected patterns. By leveraging visual analytics, we aimed to uncover hidden insights and anomalies within the dataset, providing valuable context for our subsequent outlier removal process.

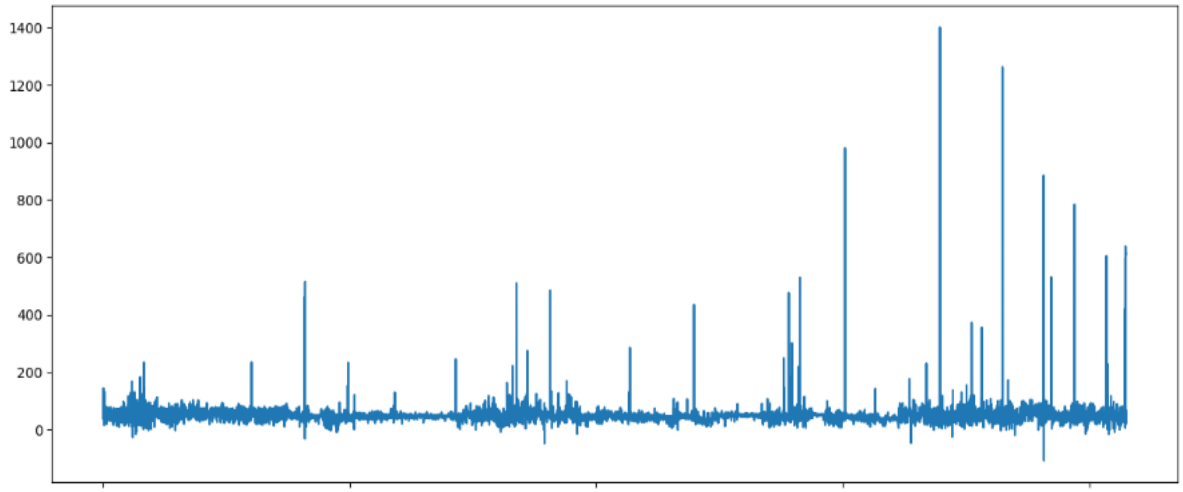


**Figure 3.3.** Raw Heat Flow Values.

In the analytical process, each data point represented on the line plot corresponds to an individual observation within the dataset, facilitating a detailed examination of the behavior of the 'HeatFlow' variable. This visualization method enables a rapid assessment of the overall trend displayed by the variable, aiding in the identification of potential outliers by highlighting any deviations from the expected pattern. By visually

analyzing the line plot, it becomes possible to identify data points that significantly deviate from the general trend, indicating their potential status as outliers that require further investigation.

Additionally, to gain deeper insights into the distribution of 'HeatFlow' values, a kernel density plot was generated using the seaborn library Figure 3.4 Unlike traditional histograms, which discretize data into intervals, the kernel density plot offers a continuous and smoothed representation of the data distribution. This visualization technique provides valuable insights into the central tendencies and potential skewness present within the 'HeatFlow' data, allowing for the detection of underlying patterns and characteristics that may not be immediately evident from other graphical representations. By utilizing the kernel density plot, a more comprehensive understanding of the distributional properties of the 'HeatFlow' variable is attained, thereby enriching the analytical approach and informing subsequent outlier identification and removal procedures.

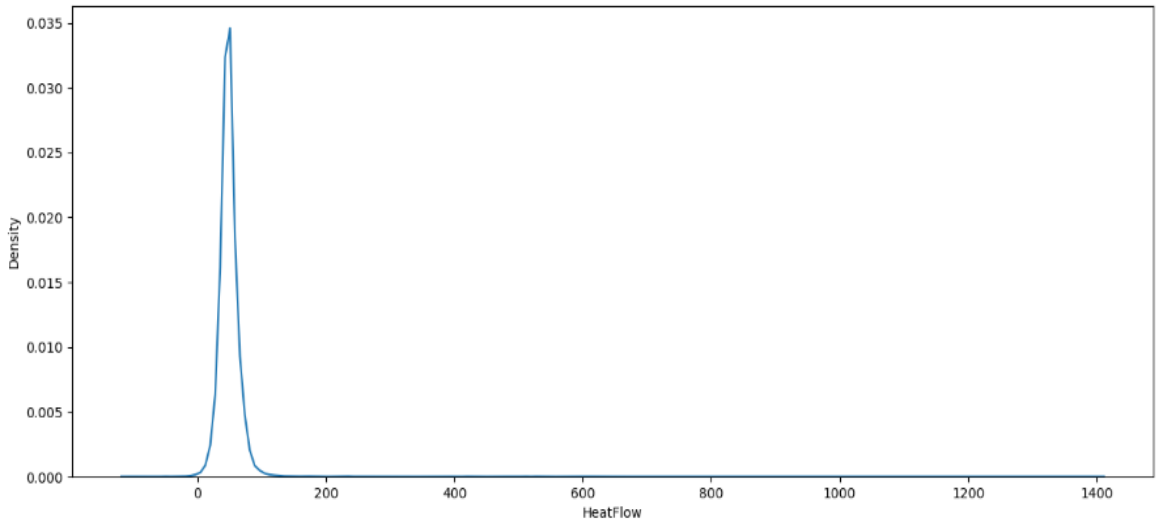


**Figure 3.4.** Kernel Density Plot.

After thoroughly examining the initial graphical plot showcasing the unprocessed heat flow values, it became evident that a methodical approach to outlier removal was imperative. Hence, the decision was made to employ the Interquartile Range (IQR) method. This method offers a systematic means of identifying and removing outliers, ensuring the integrity and accuracy of subsequent analyses.

The subsequent elaboration provides a detailed overview of the procedural intricacies involved in implementing the Interquartile Range (IQR) method for outlier removal within the dataset. This comprehensive explanation aims to elucidate the steps undertaken, offering deeper insights into the underlying principles and algorithms utilized for this purpose.

Quartiles and IQR:

To initiate the computational process, the method first computes the first quartile (Q1) and the third quartile (Q3) using the quantile method. These quartiles represent key statistical measures that delineate the dataset into four equal parts, with Q1 denoting the value below which 25% of the data lies, and Q3 representing the value below which 75% of the data lies. Subsequently, the Interquartile Range (IQR) is calculated as the difference between Q3 and Q1, providing a robust measure of the spread or dispersion of the middle 50% of the dataset.

Outlier Bounds:

To identify outliers within the dataset, the method establishes lower and upper bounds by subtracting and adding 1.5 times the IQR, respectively, from Q1 and Q3. These bounds serve as thresholds beyond which data points are considered to be potential outliers. By utilizing a multiplier of 1.5 times the IQR, the method adopts a conservative approach to outlier detection, allowing for the identification of extreme values while minimizing the inclusion of false positives.

Creation of Masks:

Within the method's implementation, two distinctive masks are generated to categorize data points based on their outlier status. The outliers_mask is designed to identify and isolate data points that are considered outliers, based on their deviation from the established bounds. Conversely, the non_outliers_mask distinguishes data points that are deemed non-outliers, indicating their adherence to the expected distributional characteristics of the dataset. These masks facilitate the segregation and selective treatment of outliers, enabling further analysis or removal as deemed appropriate.

- **3-Sigma Rule for Outlier Removal**

The 3-sigma rule, a widely recognized statistical method, was implemented using the remove_outliers_3sigma function to identify and remove outliers within our dataset. This conventional approach relies on fundamental principles of statistics to detect data points that significantly deviate from the mean of the 'HeatFlow' parameter.

The methodology begins with the calculation of the mean and standard deviation of the 'HeatFlow' parameter. The mean serves as a measure of central tendency, representing the average value of the dataset, while the standard deviation quantifies the dispersion or spread of the data around the mean.

Subsequently, outliers are identified based on their deviation from the mean by three times the standard deviation. According to the 3-sigma rule, approximately 99.7% of the data in a normally distributed dataset should fall within three standard deviations from the mean. Therefore, data points that exceed this threshold are considered outliers and are flagged for further investigation or removal. This approach leverages well-established statistical principles to enhance the integrity and reliability of our dataset, ensuring that subsequent analyses are based on accurate and representative data. However, it's important to note that the effectiveness of the 3-sigma rule may vary depending on the distributional characteristics of the data and the specific context of the analysis.

## 3.3. Hyperparameter tuning and Model Development

### 3.3.2 Hyperparameter tuning with LightGBM

Hyperparameter tuning with LightGBM represents a pivotal aspect of our analytical framework, as it harnesses the algorithm's intrinsic capabilities to optimize predictive performance, especially within the intricate domain of geothermal data analysis. LightGBM stands out due to its efficient gradient boosting implementation, leveraging a histogram-based approach that is particularly well-suited for handling large datasets with complex features—a hallmark of geoscientific data.

Our methodology integrates the robust outlier removal technique of IQR with LightGBM hyperparameter tuning, showcasing a commitment to scientific rigor. This synergistic approach aims not only to encapsulate the nuanced complexities of geophysical processes but also to demonstrate the predictive prowess inherent in LightGBM's algorithmic design.

To begin, we initiate the process by normalizing specific columns through Min-Max scaling, a critical step ensuring a uniform scale for features. This normalization mitigates the risk of any particular feature dominating the subsequent model training process, thus enhancing the model's stability and performance.

Following data normalization, we define the features (X) and the target variable (y) based on the normalized dataset. These features encompass a range of geospatial and geological information, including geographical coordinates, layer thickness, conductivity, measured depth, and surface temperature. The target variable is defined as the corrected bottom-hole temperature, which serves as the focal point for prediction in our analysis.

Subsequently, the dataset undergoes a meticulous three-way split, with 80% allocated for training, 10% for testing, and an additional 10% for validation. This balanced distribution is essential for robustly assessing the model's generalization performance and ensuring its ability to accurately predict unseen data.

With the train-test-validation split established, we define a hyperparameter grid (param_grid) comprising various combinations, including learning rate, number of leaves, number of estimators, maximum depth, and minimum child samples. An

exhaustive grid search is then performed, systematically exploring each hyperparameter combination to identify the optimal configuration.

For each hyperparameter combination (Table 3.2), a LightGBM model is trained on the training set and evaluated on the validation set. The mean squared error (MSE) serves as the performance metric, guiding the selection of the best hyperparameters that yield the lowest error rate and optimal model performance.

The best hyperparameters are subsequently employed to train the final LightGBM model, which is then evaluated on the designated test set. The evaluation process includes calculating the root mean squared error (RMSE), a key metric used to assess the model's predictive accuracy on unseen data.

### 3.3.3 Hyperparameter tuning with XGBoost

In the cited study, the authors utilized XGBoost, a widely adopted gradient boosting algorithm renowned for its effectiveness in predictive modeling tasks, particularly in estimating subsurface temperature and geothermal gradient. To ensure robust model evaluation and generalization, they adopted a 90:10 data splitting strategy, reserving 90% of the dataset for training and allocating the remaining 10% for testing.

As part of the preprocessing pipeline, the authors applied MinMax scaling to normalize specific columns independently. This preprocessing technique is commonly employed to rescale features within a consistent range, mitigating the risk of certain features disproportionately influencing the model training process. By normalizing the data, the authors aimed to enhance the stability and convergence of the XGBoost model during training, thereby improving its overall performance.

Following data preprocessing, the XGBoost model was trained on the training data, leveraging the powerful gradient boosting framework to iteratively optimize model predictions. Once the training phase was completed, the trained model was used to make predictions on the test data subset.

To assess the predictive performance of the XGBoost model, the authors computed the model error using the root mean squared error (RMSE) metric. This metric quantifies the average squared difference between the predicted values generated by the model and the actual values observed in the test dataset.

### 3.3.4 Hyperparameter tuning with Random Forest

In this segment of our methodology, we aimed to enhance the accuracy and reliability of our geothermal modeling approach by incorporating hyperparameter tuning with Random Forest alongside outlier removal using the Interquartile Range (IQR) method. This methodological adaptation signifies a departure from previous research practices, where hyperparameter tuning with Random Forest was typically combined with the 3 sigma rule for outlier removal.

Hyperparameter tuning is a crucial step in optimizing the performance of machine learning algorithms. In our case, we focused on tuning the hyperparameters of the Random Forest Regressor, which is a powerful ensemble learning algorithm. By adjusting parameters such as the number of estimators, maximum depth, minimum samples per leaf, and minimum samples per split, we aimed to find the optimal configuration that would result in the most accurate predictions for our geothermal modeling task.

For instance, the number of estimators refers to the number of decision trees in the Random Forest ensemble. By setting this parameter to 500, we aimed to create a diverse ensemble that would capture a wide range of patterns in the data. Similarly, the maximum depth parameter controls the maximum depth of each decision tree in the ensemble, preventing overfitting by limiting the complexity of individual trees.

Additionally, we set the minimum samples per leaf and split parameters to 2, which govern the minimum number of samples required to create a leaf node or perform a split in a decision tree, respectively. These parameters help prevent overfitting by ensuring that each decision tree in the ensemble generalizes well to unseen data.

After configuring the Random Forest Regressor with these hyperparameters (Table 3.2), we divided our dataset into training and testing sets to evaluate the model's performance. The training set, comprising 80% of the data, was used to train the model, while the testing set, comprising the remaining 10% of the data, and validation set 10% of the data was used to assess the model's accuracy.

Predictions were made on the test set using the trained model, and the root mean squared error (RMSE) was calculated to quantify the difference between the predicted and actual values.

### 3.3.5 Hyperparameter tuning with DNN

In the quest to refine our geothermal predictive modeling framework, a comprehensive hyperparameter tuning process for a Deep Neural Network (DNN) was undertaken utilizing the Keras framework. This methodological endeavor, conducted within the academic context of a university research project, was strategically aimed at optimizing the DNN's capacity to forecast the corrected bottom-hole temperature ('CorrBHT') based on a diverse array of geothermal features. The rationale behind this initiative stemmed from the recognition of the paramount importance of accurate geothermal resource assessment in the context of sustainable energy development.

The process commenced with meticulous data preprocessing, a critical step in ensuring the quality and integrity of subsequent analyses. Independent Min-Max scaling was applied to specific columns within the dataset, including 'LatDegree,' 'LongDegree,' 'MeasureDepth_m,' and 'SurfTemp,' in adherence to established data normalization practices. Simultaneously, normalization was performed on the target variable 'CorrBHT,' aimed at enhancing the model's convergence and stability during training.

The dataset was then judiciously partitioned into distinct training, validation, and testing subsets, adhering to the industry-standard 80:10:10 ratio. This principled approach to data partitioning served to facilitate robust model evaluation, allowing for rigorous assessments of performance across diverse datasets. Such meticulousness was deemed essential in validating the model's efficacy and generalizability beyond the confines of the training data.

The architecture of the DNN model was meticulously crafted to accommodate the inherent complexities of geothermal data, reflecting a sophisticated understanding of both machine learning principles and domain-specific knowledge. Customization of critical hyperparameters (Table 3.2), including the choice of optimizer, dropout rates, hidden units, and the number of layers, was informed by a comprehensive review of literature and best practices in the field of deep learning.

Following the delineation of the model's architecture, a systematic hyperparameter tuning loop was initiated, embodying a rigorous empirical approach to model optimization. This iterative process spanned a broad range of hyperparameters, including various optimizers, dropout rates, hidden units, layer counts, epochs, and

batch sizes, reflecting a commitment to exhaustively explore the hyperparameter space and identify configurations conducive to optimal model performance.

Once the best hyperparameters were identified, the DNN model underwent rigorous training on the entire designated training set. Subsequent evaluation on the validation set involved the meticulous assessment of key performance metrics, root mean squared error (RMSE), serving as robust indicators of the model's predictive accuracy and effectiveness in capturing the underlying patterns within the geothermal dataset.

| Model | Hyper-parameter | Best Parameter |
|---|---|---|
| LightGBM | learning_rate | 0.1 |
| LightGBM | num_leaves | 30 |
| LightGBM | n_estimators | 200 |
| LightGBM | max_depth | 15 |
| LightGBM | min_child_samples | 5 |
| Random Forest | n_estimators | 500 |
| Random Forest | max_depth | 12 |
| Random Forest | min_samples_split | 2 |
| Random Forest | min_samples_leaf | 2 |
| XGBoost | max_depth | 10 |
| XGBoost | n_estimators | 2000 |
| XGBoost | learning_rate | 0.01 |
| XGBoost | gamma | 0.1 |
| XGBoost | regression_lambda | 10 |
| Deep Neural Network | optimizers | 'sgd' |
| Deep Neural Network | dropout_rates_1 | 0.01 |
| Deep Neural Network | dropout_rates_2 | 0.001 |
| Deep Neural Network | hidden_units_list | 50 |
| Deep Neural Network | num_layers_list | 2 |
| Deep Neural Network | epochs_list | 20 |
| Deep Neural Network | batch_sizes | 64 |

**Table 3.2.** Hyper-parameters related to LightGBM, Random Forest, XGBoost and DNN models.

# Chapter IV

## Results

In this section, we provide an in-depth exploration of the results derived from our research. This examination is focused on elucidating the outcomes obtained through the application of diverse analytical techniques and methodologies within the context of subsurface geothermal resource assessment. Through a rigorous and meticulous analysis of the findings, our objective is to dissect the intricacies associated with the evaluation of subsurface geothermal resources. Additionally, we endeavor to furnish substantive insights that significantly augment the collective comprehension of this specialized domain. This scholarly endeavor not only aims to contribute to the academic discourse surrounding geothermal resource analysis but also seeks to inform future research directions and methodologies in the field. By leveraging advanced machine learning algorithms and data analytics, this study underscores the potential for innovative approaches to enhance the accuracy and efficiency of geothermal resource assessment, thereby facilitating more informed and strategic development initiatives in this vital sector of renewable energy.

Continuing from the aforementioned exploration, our investigation further employs the Interquartile Range (IQR) method as a strategic approach for outlier detection within the extensive geothermal dataset under scrutiny. Through this analytical procedure, a considerable total of 1177 data points were categorically identified as outliers, specifically within the context of the heat flow parameter. This rigorous application of the IQR method highlights its remarkable efficacy in the critical discernment and subsequent exclusion of anomalous data points. Such a meticulous process significantly bolsters the overall integrity and precision of our dataset, ensuring a robust foundation for subsequent analyses.

To elucidate the substantive impact of the outlier removal process, we have painstakingly developed a visual representation, denoted as Figure 3.5, which meticulously showcases the 'HeatFlow' values post the exclusion of identified outliers.

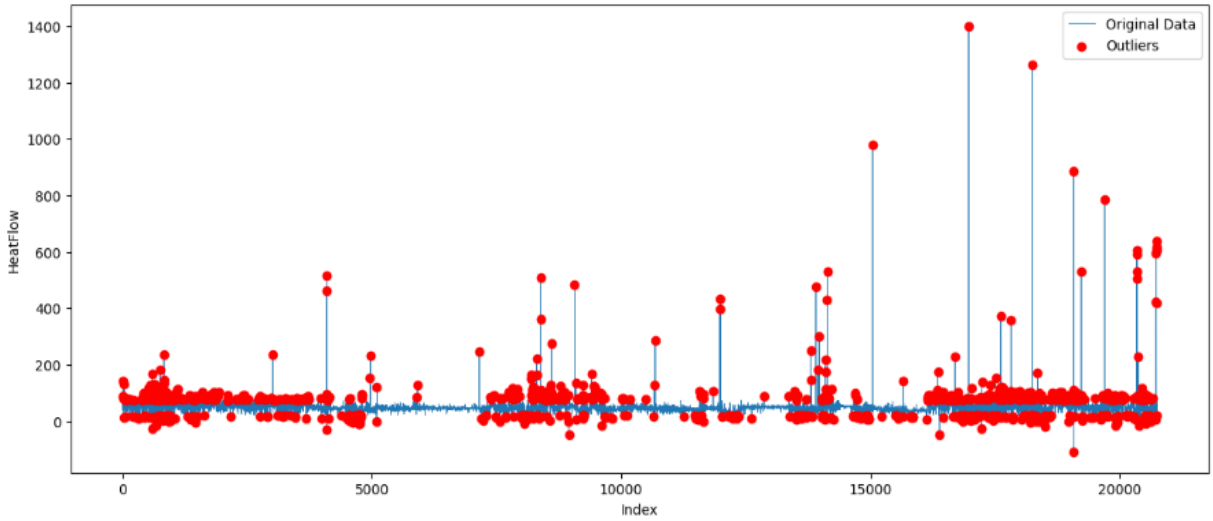


**Figure 3.5.** Heat flow original data with outliers (IQR method).

This graphical illustration serves as a compelling demonstration of the consequential refinement achieved through the IQR method, emphasizing its pivotal role in ensuring the reliability and accuracy of subsequent analyses within the geothermal domain. By systematically purging outliers, we ensure that our dataset accurately reflects the underlying geothermal characteristics, thus fortifying the foundation upon which further scientific exploration and interpretation can be conducted.

Conversely, previous research endeavors have utilized the 3-sigma rule, identifying a notably lower count of outliers, totaling 101 within the dataset. While the 3-sigma rule aligns with conventional statistical norms, its inherent sensitivity to extreme values may potentially compromise its efficacy, particularly in datasets exhibiting non-normal distributions. This observation is elucidated through visual aids (Figure 3.6), wherein outliers identified by the 3-sigma rule are delineated in red within line plots, offering valuable insights into the distributional patterns and the relative performance of outlier detection methodologies.

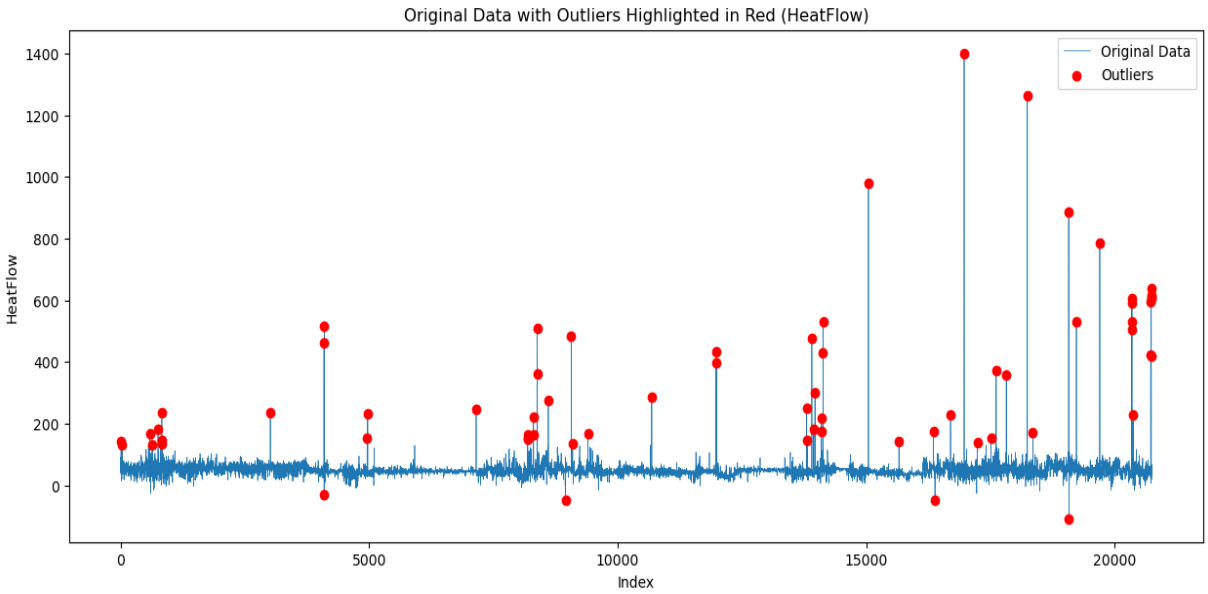


**Figure 3.6.** Heat flow original data with outliers (3-sigma rule).

By conducting a comparative analysis between the outcomes derived from our IQR-based outlier detection approach and those obtained through the application of the 3-sigma rule, this study not only accentuates the robustness inherent in the IQR method but also illuminates the nuanced considerations pivotal to the selection of outlier detection methodologies. Through an exhaustive examination of various outlier removal techniques and their consequential impacts on the refinement of the dataset, our research makes a substantial contribution to the ongoing discourse regarding data integrity and reliability within the sphere of geothermal research.

This analytical endeavor extends beyond mere method comparison to delve into the intricacies of dataset preprocessing, shedding light on the complex interplay between statistical methodologies and the quality of geothermal data analysis. The comprehensive discussion articulated herein not only fosters a deeper understanding of the challenges and considerations involved in dataset preprocessing but also underscores the critical importance of methodological rigor in enhancing the reliability of research findings.

By systematically elucidating the strengths and limitations of distinct outlier detection strategies, our study paves the way for more informed decision-making processes in the context of scientific advancement within the geothermal domain. It is anticipated that the insights garnered from this comparative analysis will serve as a valuable resource for researchers and practitioners alike, guiding the selection of appropriate

statistical techniques for dataset refinement and ultimately contributing to the advancement of knowledge and understanding in the field of geothermal resource analysis.

## 4.1. Performance of Machine Learning Models

After the completion of the initial step involving the removal of outliers utilizing the Interquartile Range (IQR) method, we embarked on an exhaustive hyperparameter tuning process for various machine learning algorithms, including LightGBM, Random Forest, and a Deep Neural Network (DNN). This endeavor encompassed the integration of the IQR technique for outlier removal, a method deeply entrenched in our pursuit of methodological refinement and precision augmentation within the realm of geothermal predictive modeling.

### 4.1.1. Ensemble Models

To provide a qualitative assessment of the model predictions, we generated scatter plots, juxtaposing the predicted values against the actual values on the test set for LightGBM and XGBoost, as depicted in Figure 4.1. Notably, previous research reported a model error of 5.099, obtained through XGBoost, while our meticulous approach yielded a root mean squared error (RMSE) of 3.354, indicative of a significantly higher level of predictive accuracy (Table 4.1).

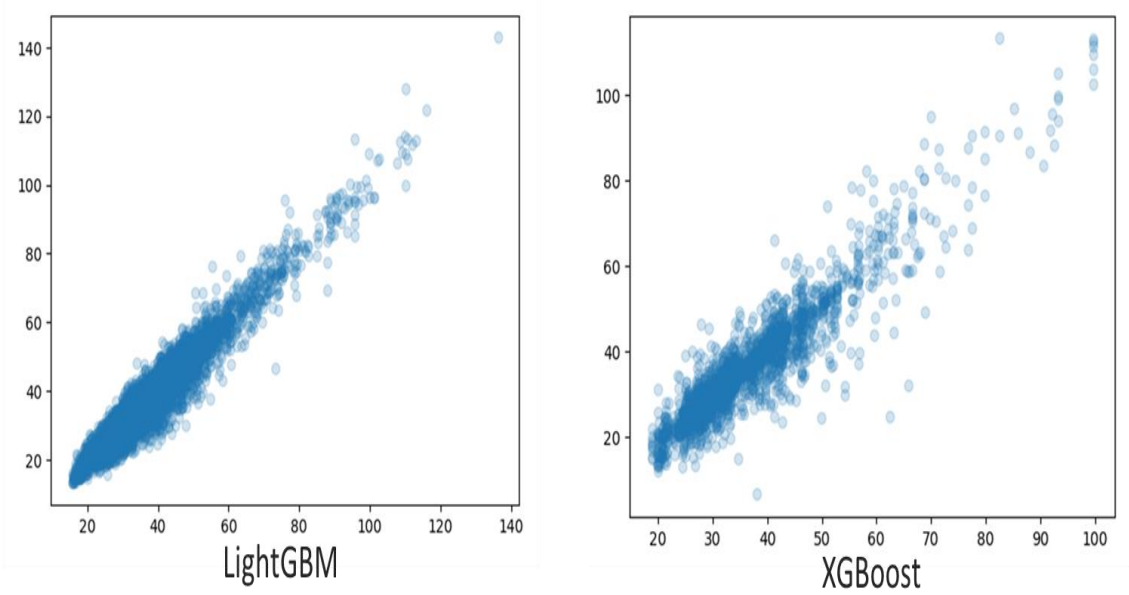


**Figure 4.1.** Scatter Plot: Predicted vs Actual Values on Test Set.

This methodical and systematic approach ensures the optimization of the LightGBM model for the geothermal dataset, thereby enabling dependable predictions in subsequent geothermal exploration analyses. Furthermore, the adoption of an 80:10:10 train-test-validation split methodology bolsters the robustness of the model evaluation

process, culminating in an overall enhancement of the reliability and accuracy of our geothermal exploration endeavors.

It is pertinent to acknowledge that while XGBoost is a commonly utilized and effective algorithm, it may pose limitations in terms of interpretability and computational efficiency, particularly when dealing with large datasets. Moreover, the utilization of the 3 sigma rule for outlier removal, as observed in previous studies, may lack the robustness offered by more advanced techniques such as the IQR method.

In our investigation, we meticulously compared the efficacy of the Random Forest algorithm in conjunction with outlier removal utilizing the Interquartile Range (IQR) method against its application with the 3 sigma rule, as documented in previous studies.

The RMSE value obtained through our approach (Table 4.1), specifically 4.08, underscores the enhanced predictive accuracy achieved through the utilization of the IQR-based outlier removal technique. This stands in stark contrast to the RMSE of 5.01 reported in prior research, where the Random Forest algorithm was combined with the 3 sigma rule.

The scatter plot depicted in Figure 4.2 provides a visual representation of the improved predictive accuracy attained through our methodology, highlighting the superiority of our approach over conventional methods.

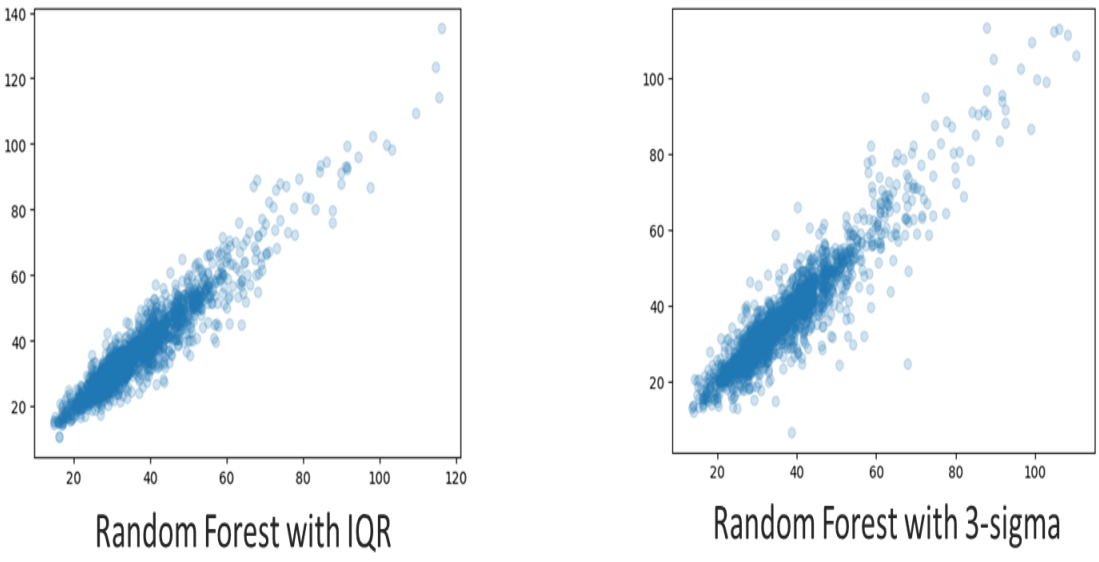


**Figure 4.2.** Scatter Plot: Predicted vs Actual Values on Test Set.

In contrast to previous research where hyperparameter tuning (HP) for Deep Neural Networks (DNNs) was deemed impractical due to computational constraints, our study took a different approach. Recognizing the significance of HP tuning in enhancing model performance, particularly in geothermal predictive modeling, we meticulously engaged in this process.

### 4.1.2 Neural Networks

The decision to conduct HP tuning for the DNN model stemmed from our acknowledgment of its potential to significantly improve predictive accuracy. This strategic maneuver aimed to optimize the DNN's ability to predict geothermal parameters by iteratively adjusting critical hyperparameters such as learning rates, batch sizes, and activation functions.

Deploying the trained DNN model on the test set yielded promising results, with predictions meticulously compared against actual values. A visual representation of the model's performance was provided through a scatter plot (Figure 4.3), offering insights into its predictive efficacy.

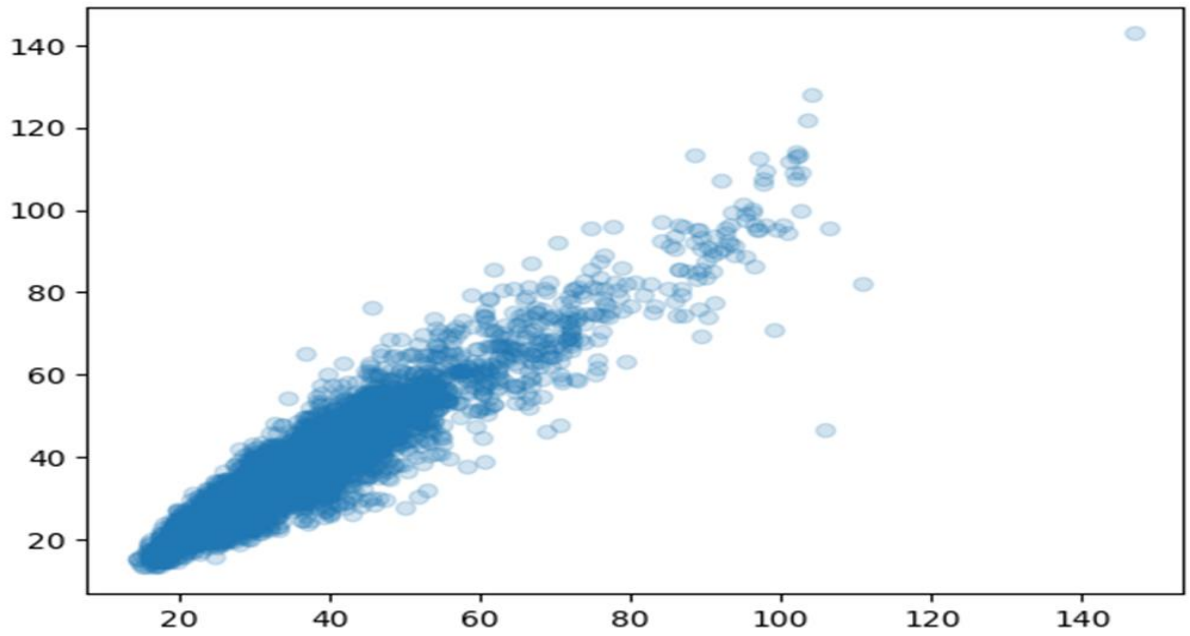


**Figure 4.3** Scatter Plot: Predicted vs Actual Values on Test Set (DNN).

Furthermore, to gain deeper insights into the DNN's convergence and generalization capabilities, we employed a loss curve—a graphical representation of the model's performance during training (Figure 4.4). This curve played a pivotal role in monitoring convergence, detecting overfitting, and guiding adjustments to hyperparameters throughout the training process.

In contrast to prior research, where default hyperparameters were utilized, resulting in a Root Mean Squared Error (RMSE) value of 5.08, our deliberate investment of approximately 8 hours in HP tuning yielded notable improvements. Through meticulous parameter optimization, we achieved a refined DNN model with a reduced RMSE of 4.54. This reduction underscores the tangible benefits of HP tuning in enhancing the accuracy and reliability of geothermal predictive models.

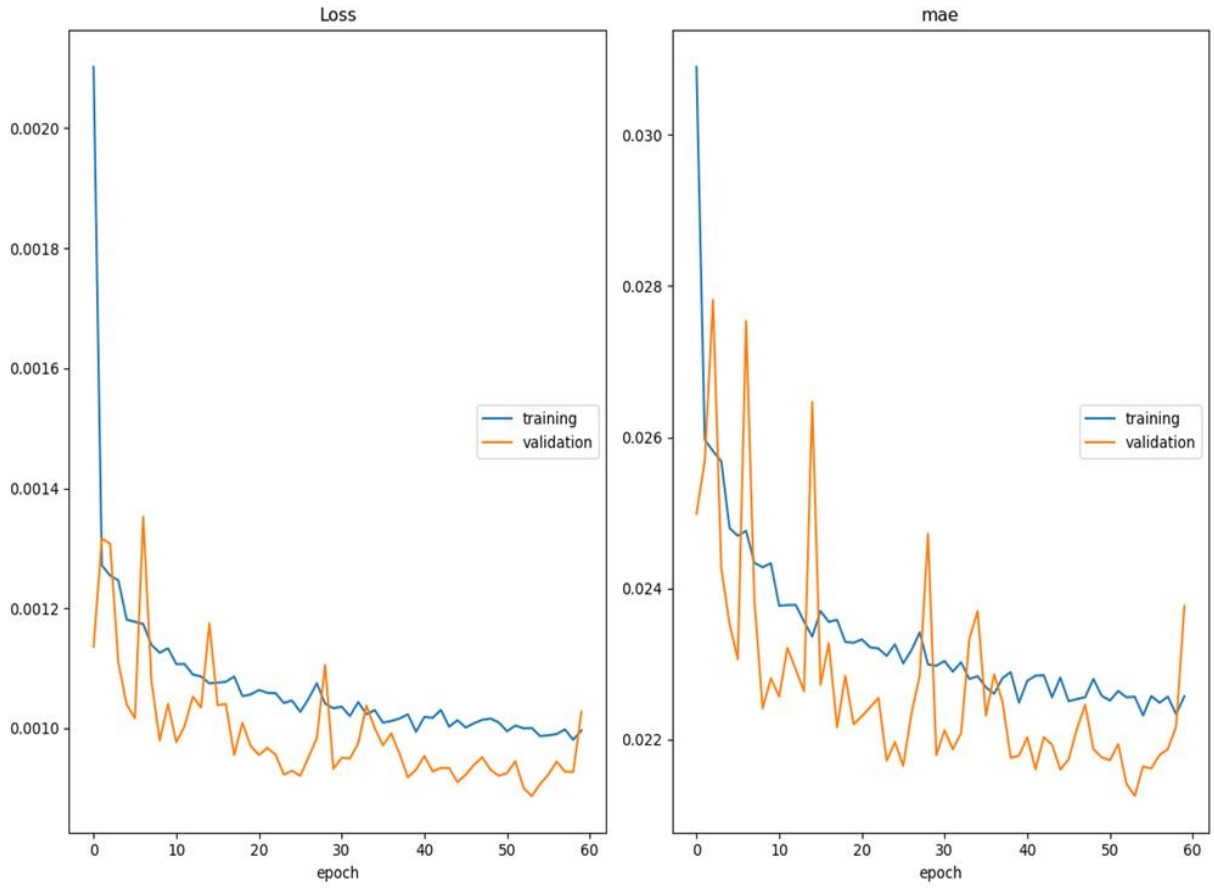


**Figure 4.4.** Training and Validation Loss Curves over Epochs.

| | **LightGBM** | **Random Forest** | **XGBoost** | **DNN** |
|---|---|---|---|---|
| Root Mean Squared Error (RMSE) | 3.354 | 4.086 | 5.099 | 4.546 |

**Table 4.1.** Model Evaluation Metrics.

## 4.2. Predictions of Machine Learning Models

### 4.2.1 Prediction of Temperature profile

In this part of work, we processed a comprehensive dataset, beginning with the random sampling of 10,000 data points from new well data, with depths rounded to facilitate matching with existing temperature profiles. Employing the Nearest Neighbors algorithm, we spatially matched each sampled well with the closest counterpart from a main dataset, providing a relevant basis for our predictive models.

Temperature predictions were initially established using a physics-based model at corresponding depths, serving as a baseline for comparison. We further refined our predictive capabilities by extracting optimal hyperparameters for the Gaussian kernel from a pre-generated output, enabling accurate spatial interpolation of subsurface properties across 49 distinct layers using k-Nearest Neighbors regression. This process was vital for capturing the intricacies of subsurface thermal properties.

Surface temperatures at each well location were predicted using a similar KNN regression approach, completing the set of temperature conditions necessary for deeper modeling. A suite of models—including LightGBM regressor, Deep Neural Network (DNN), and Random Forest regressor—was trained using the historical dataset, now enhanced with the newly interpolated properties. These models were subsequently employed to predict temperatures for the sampled wells, leveraging their respective strengths in handling non-linear patterns and interactions within the data.
The proximity of each sampled well to its historical counterpart was quantified, providing contextual depth to the analysis and highlighting the geospatial dimension of our modeling approach.

In an extensive evaluation of machine learning (ML) models for geothermal resource analysis, we scrutinized the predictive accuracy of these models against actual temperature measurements and physical model predictions within 58 randomly selected wells from the northeastern United States. This analysis sought to uncover the efficacy of ML models in capturing the complex thermal profiles characteristic of subsurface environments. From this dataset, three wells—API#4700100668, API#4705900805, and API#4709300104—were selected for a more detailed examination to extract deeper insights into the performance and applicability of ML in geothermal temperature prediction.

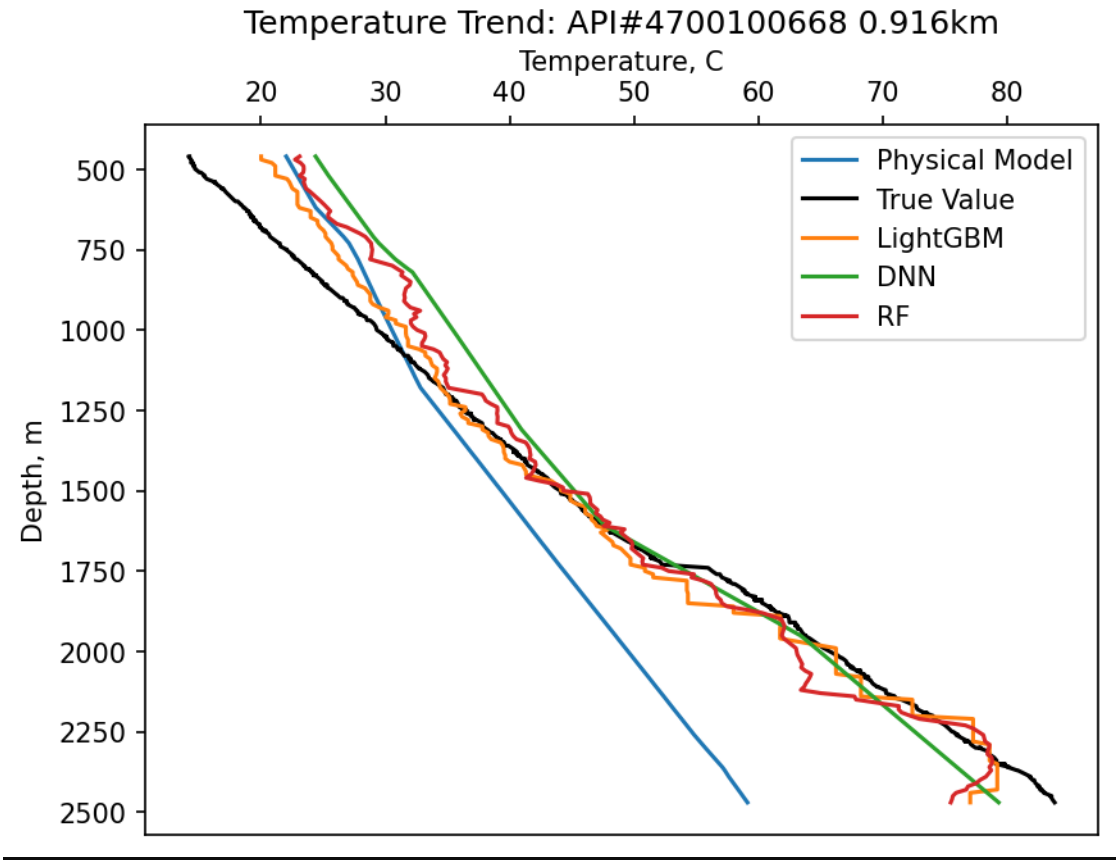


**Figure 4.5** Temperature-profile predictions.

- **Well API#4700100668: A Study in Spatial Proximity**

Well API#4700100668, in close proximity to its reference well at a distance of 0.916km, showcased the precision that ML models can achieve (Figure 4.5). The predictions from the LightGBM, DNN, and RF models not only adhered closely to the true temperature values but also maintained this consistency across varying depths. This accuracy is indicative of the models' ability to utilize local geological information, thereby suggesting that spatial correlation is a significant determinant of predictive success. The convergence of model predictions with the true values, especially in the deeper geological strata, indicates a high level of sophistication in the models' capacity to internalize and replicate complex thermal gradients.

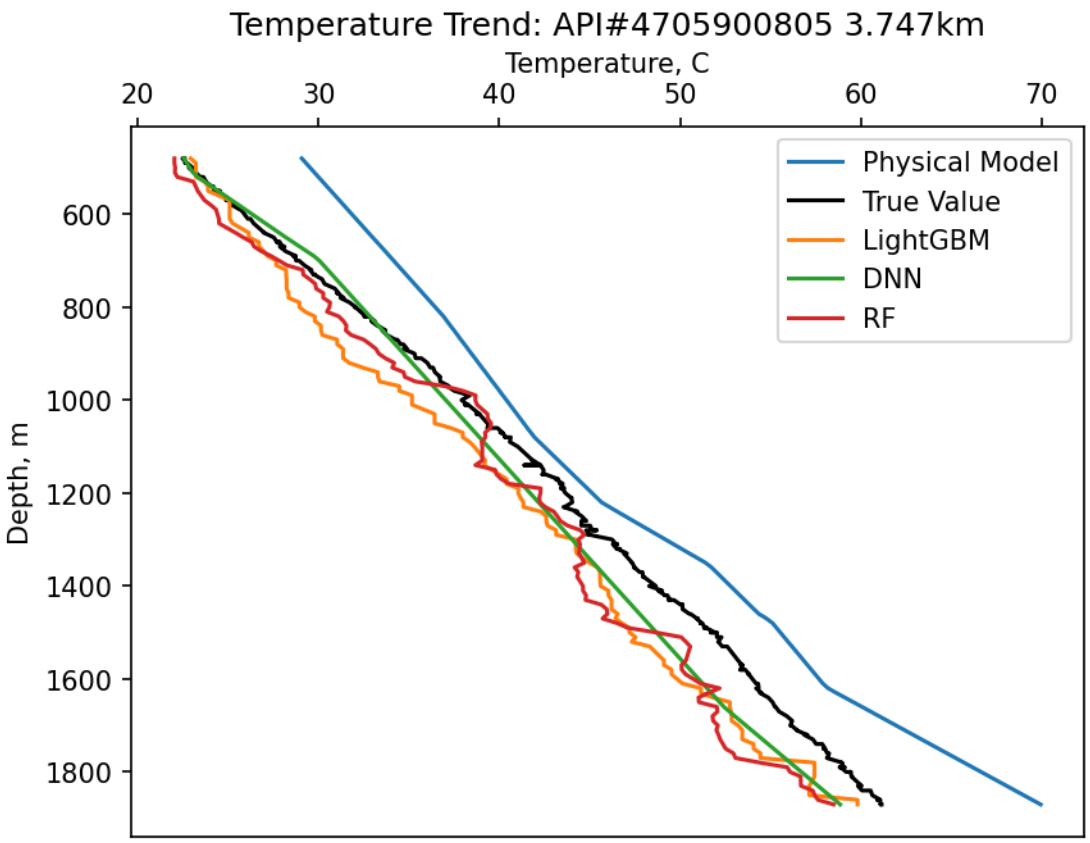


**Figure 4.6** Temperature-profile predictions.

- **Well API#4705900805: Distance and Predictive Discrepancies**

Contrasting with the previous well, API#4705900805, at a distance of 3.747km from the nearest well from the main dataset, presented a case where ML model predictions began to show discrepancies, particularly in the deeper subsurface regions (Figure 4.6). While the models generally conformed to the actual temperature measurements, their slight divergence at greater depths highlights the challenges ML models face when extrapolating beyond the more immediate geographical similarities found in training data. This well serves as a prime example of the limitations inherent in ML models when operating under conditions of reduced geological affinity, emphasizing the need for robust, geographically diverse training datasets to enhance model resilience and accuracy.

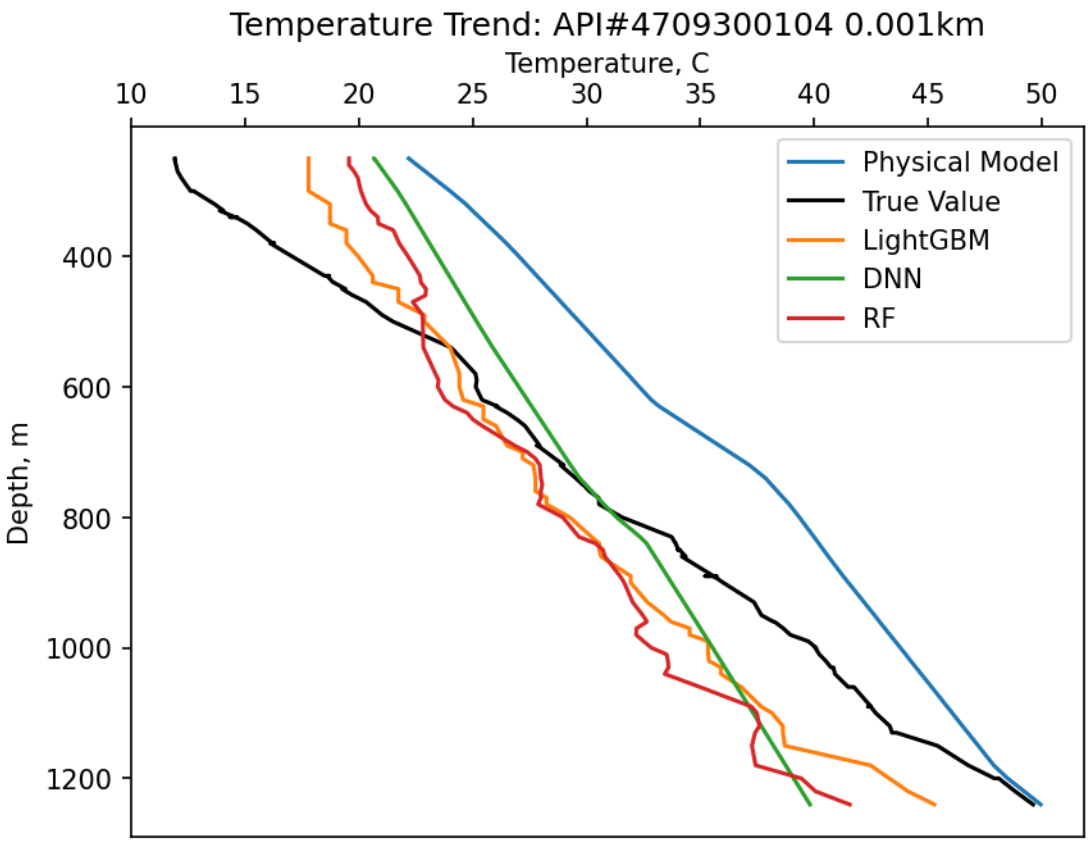


**Figure 4.7** Temperature-profile predictions.

- **Well API#4709300104: Robustness in the Face of Geological Variability**

Well API#4709300104, at an extremely short distance of 0.001km from its nearest reference well, provided a distinct perspective (Figure 4.7). Here, the DNN and LightGBM models offered predictions that closely aligned with the actual temperatures, demonstrating their robustness even when faced with the complex geological variability that is often characteristic of geothermal systems. The RF model, while accurate, showed slight deviations at certain points, which may be attributed to the model's inherent random processes or to limitations in capturing some of the more nuanced geothermal properties at those specific depths.

Through these case studies, several critical insights emerge:

- **Influence of Spatial Correlation:** The proximity of wells to their historical counterparts significantly impacts ML predictions, with closer wells typically yielding more accurate results. This is likely due to the geospatial correlation in subsurface properties, which ML models can effectively leverage.

- **Model Performance at Varying Depths:** All models performed admirably at shallower depths; however, the true test of their predictive power was observed at greater depths, where geological conditions become increasingly complex. Models like LightGBM and DNN showed remarkable adaptability, maintaining high accuracy where simpler models might fail.

- **Complexity of Geological Interpretation:** The physical model's underperformance, as seen in well API#4705900805, suggests that traditional geothermal models may not always capture the intricacies of subsurface conditions. ML models, equipped with a multitude of data-driven features, can offer a more intricate interpretation of the subsurface conditions.

- **Integration with Geothermal Development Strategies:** The insights provided by ML models can be integrated into broader resource management and development strategies. Their predictive capabilities allow for better planning of drilling operations, optimization of resource extraction, and minimization of financial and environmental risks.

### 4.2.2 Depth-Stratified Subsurface Temperature Profiling

Temperature-at-depth maps are fundamental tools in geothermal energy exploration, offering insights into temperature distributions at various depths. In our study, we have extended this practice to create temperature-at-depth maps for different depths across the northeastern United States. These maps serve as vital resources for stakeholders and investors, providing additional predictive information for potential geothermal projects.

Our newly developed machine learning-based temperature maps offer an opportunity for comparative analysis with existing thermal conductivity models. By comparing these datasets, researchers can identify similarities and differences, enhancing our understanding of subsurface temperature dynamics.

To construct these temperature prediction maps, it is essential to have access to relevant features across different geographical locations with varying latitudes and longitudes. To address this requirement, we employed an interpolation process to extrapolate essential features across the northeastern region. Utilizing the LightGBM algorithm, we conducted this interpolation, carefully considering geographical variations and feature interactions utilizing a robust dataset consisting of 20,750 data points. This rigorous approach ensured that the algorithm's parameters were fine-tuned to maximize predictive accuracy and reliability, thereby enhancing the quality of the generated temperature-at-depth maps.

Delving deeper into the dataset, it provided anticipated underground temperatures grounded in physics at the specific geographical coordinates of each well, spanning various depths. Leveraging this dataset in conjunction with the LightGBM algorithm, we endeavored to approximate physics-based values across a spectrum of latitudes, longitudes, and depths, thus enriching our understanding of subsurface temperature distribution. Illustrated in Figure 4.8 are temperature prediction for 1000 depth generated through the implementation of LightGBM models.

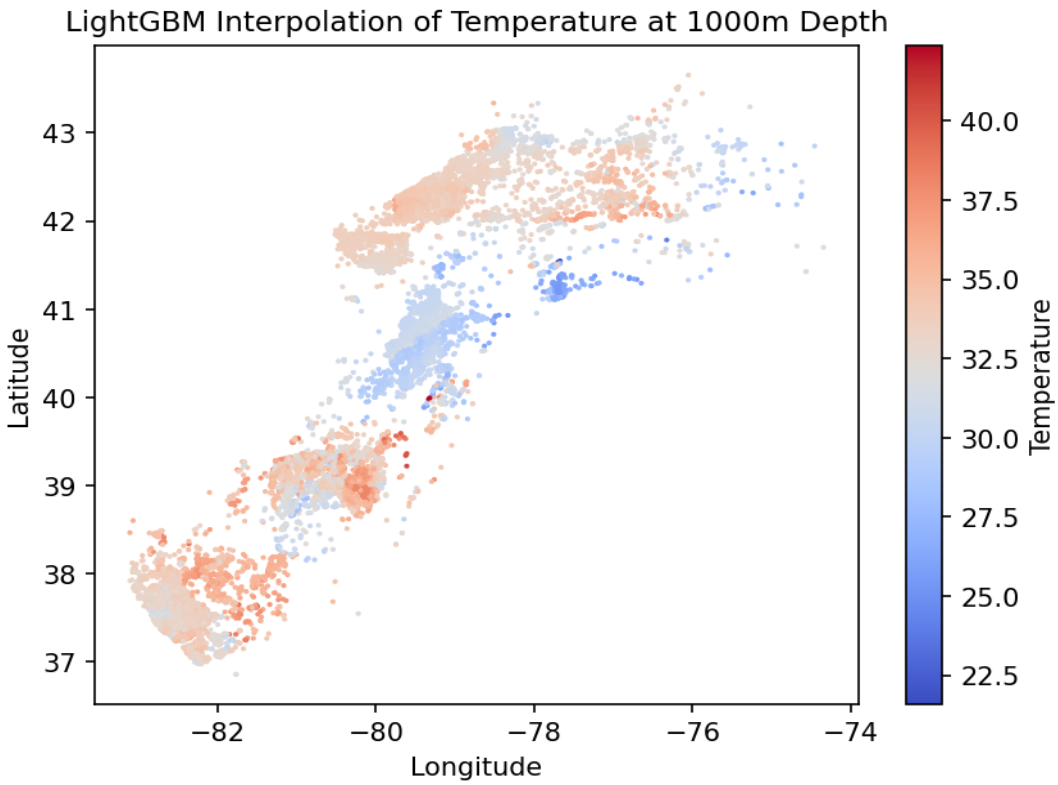


**Figure 4.8.** Temperature Depth Map.

The 1000m depth temperature map (Figure 4.8) is particularly intriguing as it represents the uppermost section of the subsurface environment where geothermal gradients are more susceptible to surface influences and near-surface geological formations. In this map, temperatures span from a relatively cool 22.5 degrees to a modestly warm 40 degrees. This range suggests a complex interplay of geological and hydrological factors that could have significant implications for geothermal resource development.

At this shallow depth, the thermal regime is often influenced by recent geological processes. For instance, the presence of karst formations, common in parts of the northeastern USA, can lead to rapid vertical movement of water, which in turn can affect the local temperature distribution. Similarly, the thermal conductivity of near-surface rocks can vary considerably, with sedimentary layers often acting as insulators, while metamorphic and igneous rocks can lead to higher thermal transmissivity.

The temperature anomalies detected at the 1000m level may also be indicative of subsurface hydrological processes. Groundwater flow can redistribute heat through advection, which can create local zones of higher or lower temperatures than the

regional average. Such variations are crucial for identifying potential geothermal reservoirs, especially for low-enthalpy systems suitable for direct-use applications.
The detailed temperature distribution at 1000m depth offers a promising outlook for the development of shallow geothermal systems. Such systems could be harnessed for district heating, greenhouse agriculture, or other direct-use applications. The spatial resolution of the temperature map allows for the identification of sites with the highest potential, thus minimizing exploratory drilling costs and reducing the risk associated with geothermal projects.

In transitioning from the 1000m depth temperature map to the 2000m depth map (Figure 4.9) in the context of northeastern USA, we observe notable changes that are indicative of the subsurface thermal regime's complexity. The spatial distribution of temperature anomalies at 2000m displays a more pronounced pattern, with the warmer areas becoming more apparent. This could be attributed to deeper geologic structures such as basin boundaries, fault zones, or areas with higher radioactive decay generating heat within the crust. Several factors could contribute to the observed temperature distribution at 2000m depth. Thermal conductivity of the rocks, which dictates how heat is transferred through the Earth’s crust, is one such factor. Areas with higher thermal conductivity may show a more uniform temperature distribution, while low conductivity rocks can result in localized temperature anomalies.
Additionally, the presence of water-filled fractures can also play a role in redistributing heat through convective processes.

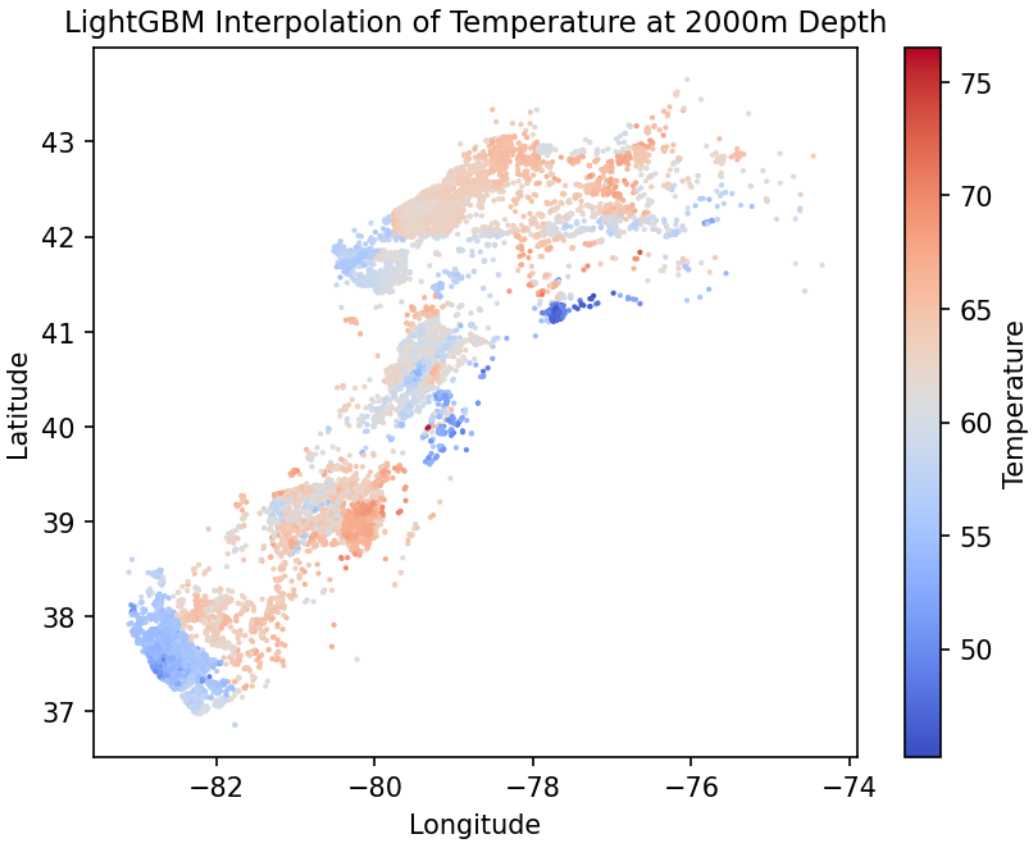


**Figure 4.9.** Temperature Depth Map.

The LightGBM interpolation at a depth of 3000 meters (Figure 4.10) uncovers distinct geothermal characteristics within the northeastern United States. Notably, the highest temperature readings are concentrated in the northern part of Pennsylvania and the southeast region of New York. These localized thermal anomalies are of significant interest for several reasons. The elevated temperatures in these areas may be attributed to the unique geological history of the region. Pennsylvania and New York are known for their varied geological formations, including the presence of the Appalachian Basin, which could impact geothermal gradients. These areas may have enhanced geothermal properties due to a combination of factors such as residual heat from historical tectonic activity, radiogenic heat production from granite bodies, or deep-seated fractures that facilitate heat flow.

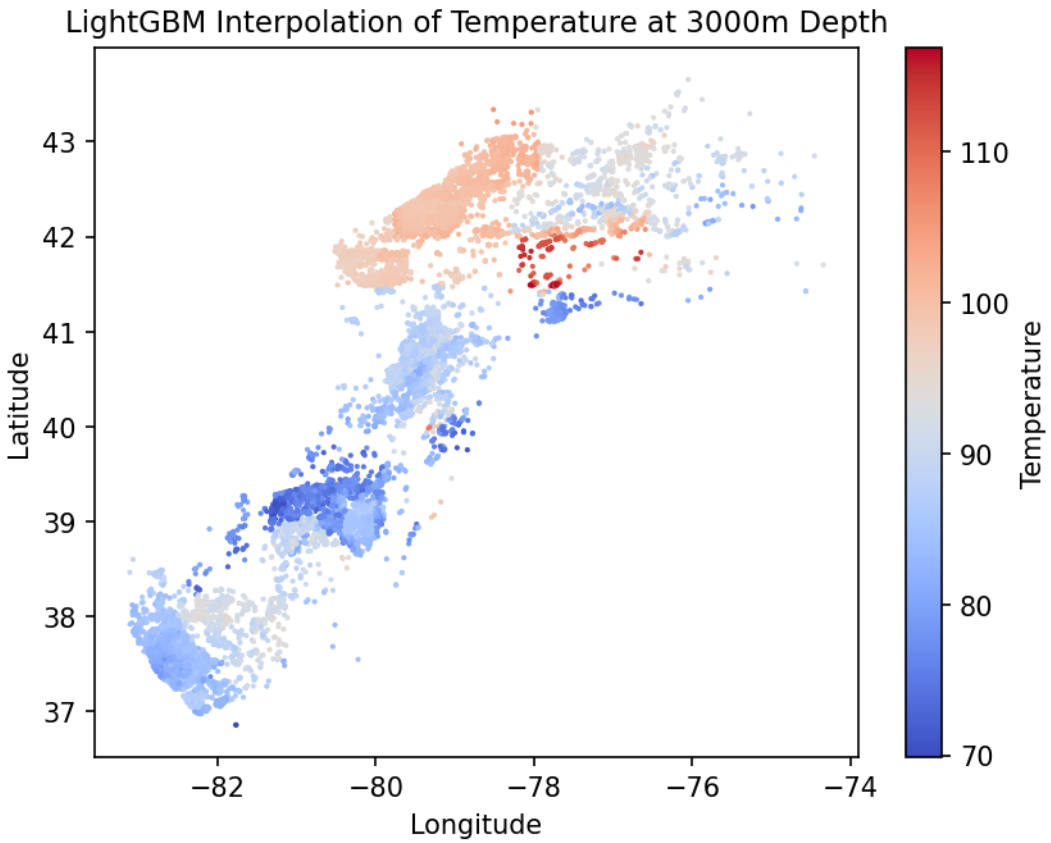


**Figure 4.10.** Temperature Depth Map.

The three maps, representing depths of 1000m, 2000m, and 3000m, collectively offer a multi-layered perspective of the thermal state beneath the surface, illustrating a clear increase in temperature with depth and highlighting areas of particular geothermal interest. Predictive temperature-profile charts for additional wells can be found in our GitHub repository [72].

# Chapter V

## Conclusion

This thesis has embarked on a pioneering exploration into the integration of Machine Learning (ML) methodologies with geothermal energy exploration and development, a crucial frontier within the renewable energy spectrum. Through a rigorous examination of an extensive dataset from the Northeastern United States, this research has not only demonstrated the potential of ML to revolutionize the predictability and operational efficiency of geothermal resource characterization but has also unveiled the complexities and challenges inherent in this endeavor.

The study meticulously analyzed and refined prevailing approaches in data preprocessing, feature engineering, and hyperparameter optimization, revealing that the adaptability and precision of ML models significantly enhance the forecasting of subsurface temperatures and geothermal gradients. Notably, the employment of sophisticated outlier detection, data normalization, and grid search techniques for hyperparameter fine-tuning emerged as pivotal elements in augmenting the accuracy of predictive models. Such methodological innovations underscore the nuanced understanding required to effectively harness ML in geothermal resource exploration.

One of the cardinal challenges encountered in this research was the heterogeneity and complexity of geothermal data, which necessitated the development of robust ML algorithms capable of accommodating diverse data types and structures. By addressing these challenges, the research contributes significantly to the body of knowledge, offering a comprehensive framework for future investigations in the domain.

The implications of this work extend beyond the immediate realm of geothermal energy, suggesting a paradigm shift in how renewable resources are explored and developed. The enhanced predictive models facilitate strategic decision-making and resource allocation, potentially leading to a reduction in exploratory costs and a more sustainable approach to energy generation.

Looking ahead, the integration of real-time data acquisition and the implementation of ML monitoring systems stand out as promising avenues for future research. Such

advancements could further refine the predictive capabilities of ML models, ensuring their applicability in dynamic geological environments. Additionally, exploring the interoperability of ML methodologies with other renewable energy sources could yield comprehensive insights into a holistic energy sustainability strategy.

In conclusion, the work presented in this thesis not only reaffirms the indispensable role of geothermal energy within the renewable energy portfolio but also sets a precedent for the innovative application of ML methodologies in environmental science and engineering. It serves as a testament to the power of interdisciplinary research in addressing some of the most pressing challenges of our time, paving the way for future advancements in renewable energy exploration and development.

# References


1. Parri, R. and F. Lazzeri. *Larderello: 100 years of geothermal power plant evolution in Italy*. Geothermal Power Generation, 2016. p. 537-590.
2. Lazard., *Levelized cost of energy, levelized cost of storage, and levelized cost of hydrogen*. Lazard website, 2020.
3. Zarrouk, S. and H. Moon, *Efficiency of geothermal power plants: A worldwide review.* Geothermics, 2014. p. 142–153.
4. ThinkGeoEnergy, *Global Geothermal Power Plant Map. ThinkGeoEnergy website,* 2020.
5. Hettiarachchi, M., et al., *Optimum design criteria for an Organic Rankine Cycle using low-temperature geothermal heat sources.* Energy, 2007. p. 1698-1706.
6. Office, G.T., *GeoVision: Harnessing the Heat Beneath Our Feet. U.S. Department of Energy Office of Energy Efficiency & Renewable Energy.* Geothermal Technologies Office, 2019.
7. Robins, J.C., et al., *2021 U.S. Geothermal Power Production and District Heating Market Report*, 2021.
8. Misra, S., et al., *Machine Learning Tools for Fossil and Geothermal Energy Production and Carbon Geo-sequestration—a Step Towards Energy Digitization and Geoscientific Digitalization.* Circular Economy and Sustainability. Springer, 2021.
9. Sircar, A., et al., *Application of machine learning and artificial intelligence in oil and gas industry.* Petroleum Research, 2021.
10. Crow, D., et al., *Impact of Drilling Costs on the US Gas Industry: Prospects for Automation.* Energies, 2018. p. 2241.
11. Heghedus, C., A. Shchipanov, and C. Rong, *Advancing Deep Learning to Improve Upstream Petroleum Monitoring.* IEEE Access, 2019. p. 1-1.
12. Zhang, J., et al., *Prediction method of physical parameters based on linearized rock physics inversion.* Petroleum Exploration and Development, 2020. p. 59-67.
13. Shahdi, A., et al., *Exploratory analysis of machine learning methods in predicting subsurface temperature and geothermal gradient of Northeastern United States.* Springer, 2021.
14. ESMAP, *Planning and Financial Power Generation. Energy Sector Management Assistance Program*, Geothermal Handbook, 2012.
15. Witherbee, K., *Overview of Geothermal Energy Development*, in *Webcast: Overview of Geothermal Energy Development slides 54-55*. Office of Indian Energy, 2012.
16. Bank, T.W., *International bank for reconstruction and development*. World Bank, 2019.
17. McCarthy, J., *What is Artificial Intelligence?* ResearchGate, 2004.
18. Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Perrault, and R., *The AI index 2021 annual report*, C.A.I.S.C. Stanford, Human-Centered AI Institute, Stanford University, 2021.
19. Bortnik, J., Camporeale, E, *Ten ways to apply machine learning in earth and space sciences*. Eos website, 2021.
20. Wikipedia. *Random Forest Algorithm*. Retrieved from: https://en.wikipedia.org/wiki/Random_forest.
21. Tibshirani, G.J.D.W.T.H.R., *An Introduction to Statistical Learning.* Springer, 2013: p. 316-321.
22. Ho, T., *A Data Complexity Analysis of Comparative Advantages of Decision Forest Constructors.* Pattern Anal. Appl., 2002. p. 102-112.
23. Hastie, T.T., Robert; Friedman, Jerome *The Elements of Statistical Learning.* Springer, 2008.

24. Geurts, P., D. Ernst, and L. Wehenkel, *Extremely Randomized Trees.* Machine Learning, 2006. p. 3-42.
25. Wikipedia. *XGBoost Algorithm*. Retrieved from: https://en.wikipedia.org/wiki/XGBoost.
26. Gandhi, R., *Gradient Boosting and XGBoost*. Medium, 2019.
27. Science, T.D., *Boosting algorithm: XGBoost*. ResearchGate, 2017.
28. Chen, T.G., Carlos *XGBoost: A Scalable Tree Boosting System*. ReseacrhGate, 2016.
29. Wikipedia. *LightGBM Algorithm*. Retrieved from: https://en.wikipedia.org/wiki/LightGBM.
30. Manu, J., *The Gradient Boosters IV: LightGBM*. Deep & Shallow website, 2020.
31. Ye, A., *XGBoost, LightGBM, and Other Kaggle Competition Favorites*. Medium, 2020.
32. Guolin Ke, Q.M., Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu, *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. Neural Information Processing Systems conference. 2017.
33. Bengio, Y., *Learning Deep Architectures for AI.* Foundations, 2009. p. 1-55.
34. Schmidhuber, J., *Deep learning in neural networks: An overview.* Neural Networks, 2015. p. 85-117.
35. Gavrilova, Y., *A Guide to Deep Learning and Neural Networks*. Serokell website, 2020.
36. Szegedy, C., A. Toshev, and D. Erhan, *Deep Neural Networks for Object Detection*. ResearchGate, 2013. p. 1-9.
37. Rolnick, D. and M. Tegmark, *The power of deeper networks for expressing natural functions.* ResearchGate, 2017.
38. Hof, R.D., *Deep Learning*. *MIT Technology Review*, 2013.
39. Ivakhnenko, A., *Polynomial theory of complex systems*. Springer, 1971. p. 364-378.
40. Dahl, G., *Broadcast Language Identification & Subtitling System (BLISS)*. ResearchGate, 2013.
41. Andrew, Ng., *Data Augmentation*. Coursera. Retrieved from: https://www.coursera.org/lecture/convolutional-neural-networks/data-augmentation-AYzbX.
42. Aleksander, I., et al., *A brief introduction to Weightless Neural Systems*. ResearchGate, 2009.
43. You, Y., *Scaling deep learning on GPU and knights landing clusters.* ResearchGate, 2017.
44. Viebke, A., et al., *CHAOS: a parallelization scheme for training convolutional neural networks on Intel Xeon Phi.* The Journal of Supercomputing, 2019. p. 197-227.
45. Feurer, M. and F. Hutter, *Hyperparameter Optimization*. ResearchGate, 2019. p. 3-33.
46. Claesen, M. and B. De Moor, *Hyperparameter Search in Machine Learning.* ResearchGate, 2015.
47. Bergstra, J. and Y. Bengio, *Random Search for Hyper-Parameter Optimization.* The Journal of Machine Learning Research, 2012. p. 281-305.
48. Hsu, C.-w., C.-c. Chang, and C.-J. Lin, *A Practical Guide to Support Vector Classification Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin.* ResearchGate, 2003.
49. Chicco, D., *Ten quick tips for machine learning in computational biology.* BioData Mining, 2017. p. 35.
50. Blackwell D, R.M., *New geothermal resource map of the northeastern US and technique for mapping temperature at depth*. ResearchGate, 2010.
51. University, C., *Appalachian Basin play fairway analysis: thermal quality analysis in low-temperature geothermal play fairway analysis.* Geothermal Data Repository, 2015.
52. Jordan T, R.M., Horowitz F, Camp E, *Low Temperature geothermal play fairway analysis for the appalachian basin*. Geothermal Data Repository, 2016.

53. Snyder DM, B.K., Young KR, *Update on geothermal direct-use installations in the United States*, P. of and v. forty-second workshop on geothermal reservoir engineering, Stanford University, 2017.
54. Witter, J., W. Trainor-Guitton, and D. Siler, *Uncertainty and risk evaluation during the exploration stage of geothermal development: A review.* Geothermics, 2019. p. 233-242.
55. Lukawski, M., R. Silverman, and J. Tester, *Uncertainty analysis of geothermal well drilling and completion costs.* Geothermics, 2016. p. 382–391.
56. Bloomquist G, N.P., El-Halabi R, Löschau M, *The AUC/KFW Geothermal Risk Mitigation Facility (GRMF)–A Catalyst for East African Geothermal Development*, G. Transactions, 2012.
57. Assouline, D., et al., *A machine learning approach for mapping the very shallow theoretical geothermal potential.* Geothermal Energy, 2019.
58. G, B., *Data fusion and machine learning for geothermal target exploration and characterisation*, N.I.A.L. (NICTA), 2014.
59. Faulds JE, B.S., Coolbaugh M, Deangelo J, Queen JH, Treitel S, Fehler M, Mlawsky E, Glen JM, Lindsey C, Burns E., *Preliminary report on applications of machine learning techniques to the nevada geothermal play fairway analysis*. 45th workshop on geothermal reservoir engineering, 2020.
60. Rezvanbehbahani, S., et al., *Predicting the Geothermal Heat Flux in Greenland: A Machine Learning Approach.* Geophysical Research Letters, 2017.
61. Tut Haklidir, F. and M. Haklidir, *Prediction of Reservoir Temperatures Using Hydrogeochemical Data, Western Anatolia Geothermal Systems (Turkey): A Machine Learning Approach.* Natural Resources Research, 2019.
62. Shi, Y., X. Song, and G. Song, *Productivity prediction of a multilateral-well geothermal system based on a long short-term memory and multi-layer perceptron combinational neural network.* Applied Energy, 2021. p. 116046.
63. Keynejad, S., M.L. Sbar, and R.A. Johnson, *Assessment of machine-learning techniques in predicting lithofluid facies logs in hydrocarbon wells.* Interpretation, 2019. p. SF1-SF13.
64. Ma, Y., et al., *A deep-learning method for automatic fault detection*. ReserachGate, 2018.
65. Zhang, C., et al., *Machine-learning Based Automated Fault Detection in Seismic Traces.* ResearchGate, 2014.
66. Araya, M., et al., *Deep-learning tomography.* The Leading Edge, 2018. p. 58-66.
67. Hall, B., *Facies classification using machine learning.* The Leading Edge, 2016. p. 906-909.
68. Perozzi L, G.L., Moscariello, A, *Minimizing Geothermal exploration costs using machine learning as a tool to drive deep geothermal exploration*. AAPG European Region, 3rd Hydrocarbon Geothermal Cross Over Technology Workshop, 2019.
69. Gunderson, K.L., Holmes, R. C., & Loisel, J, *Recent digital technology trends in geoscience teaching and practice*. The Geological Society of America , 2020.
70. *Dataset*- 1: *Appalachian Basin Play Fairway Analysis: Thermal Quality Analysis in Low-Temperature Geothermal Play Fairway Analysis (GPFA-AB)*. Geothermal Data Repository, 2015. Retrieved from: https://gdr.openei.org/submissions/638.
71. *Dataset-2: West Virginia Geological & Economic Survey. Retrieved from: https://www.wvgs.wvnet.edu*.
72. Github, *MS-thesis, Available from: https://github.com/salmanmirzayev/MS-thesis/blob/main/well-prediction.ipynb*.