Sector Learning with Data Sharing – Case Studies for the Geothermal Industry

Version 23.05.2025

Sector Learning with Data Sharing – Case Studies for the Geothermal Industry

Authors:		**Reviewers:**	
Company:	TNO	**Company:**	TNO
From:	Pejman Shoeibi Omrani	**From:**	Ryvo Octaviano, Laura Precupanu
E-mail:	pejman.shoeibiomrani@tno.nl	**E-mail:**	ryvo.octaviano@tno.nl, laura.precupanu@tno.nl
Project coordinator	Laura Precupanu, laura.precupanu@tno.nl		

Subject: Sector Learning with Data Sharing – Case Studies for the Geothermal Industry

This project is executed with funding of the National Growth Fund of the Ministry of Economic Affairs, implemented by the Netherlands Enterprise Agency. The specific funding for project *Data gedreven optimalisatie van aardwarmte systemen* (NWNI10005) concerns the funding rules 2022.

Management Summary

This report explores the role of data sharing technologies in advancing digitalization within the geothermal energy sector, with a focus on enabling secure and collaborative analytics across multiple stakeholders. In the geothermal sector, data is often fragmented due to privacy concerns, confidentiality, and competitive boundaries. The report provides an overview of several emerging approaches to address these challenges, including secure multi-party computation (SMPC) and federated learning (FL), each offering unique mechanisms to facilitate collaboration without compromising data ownership or confidentiality.

A key value of these technologies lies in their ability to enable stakeholders, such as geothermal plant operators, equipment manufacturers, and researchers, to jointly analyze operational data while keeping sensitive information protected. This opens up opportunities to develop more accurate, data-driven insights into critical processes to improve efficiency and economics of the geothermal plants. By securely pooling knowledge from distributed sources, these approaches help overcome data sparsity and variability that often limit the performance of traditional models developed in isolation. A set of potential case studies is identified, illustrating how different stakeholders in the geothermal sector can securely share data and derive mutual benefits through collaborative analytics and predictive models. Beyond individual plant benefits, the integration of diverse datasets significantly strengthens the derived analytics and model generalizability and reliability, supporting better-informed decisions at both the operational and strategic levels. Collaborative learning across multiple environments allows for more robust forecasting, better understanding of degradation patterns, and improved equipment lifecycle management.

This report highlights how secure data sharing technologies can serve as a foundation for collective innovation in geothermal energy. By enabling trust-based collaboration without sacrificing data control, they offer a path forward for digital transformation in the sector. The findings underline the potential for these technologies to not only optimize individual plant performance but also enhance system-wide efficiency and resilience, contributing meaningfully to the energy transition.

Table of Contents

Management Summary 2
Introduction 6
Data Sharing Technologies 8
Secure Multi-Part Computation (SMPC) 8
Federated Learning (FL) 9
Case Studies in the Geothermal Sector 12
Predictive Maintenance for Geothermal Plant Equipment 13
Corrosion and Scaling Mitigation 13
Plant Design and Optimization 14
Drilling Optimization 14
Reservoir Characterization and Management 15
Exploration and Resource Assessment 15
Challenges and Limitations 17
Conclusions and Future Outlook 19
References 20

Introduction

Geothermal energy is an emerging and increasingly important contributor to the global transition toward sustainable energy (IEA, 2021). As the demand for renewable heat and power continues to rise, geothermal systems are expected to play a more prominent role in achieving decarbonization goals. However, the sector is still navigating a steep learning curve. Its inherent complexitiessuch as - subsurface conditions, high upfront capital costs, and long development timelines – pose significant barriers to widespread adoption (Goetzl et al., 2021). Accelerating the learning process is therefore essential to improving the cost-competitiveness and reliability of geothermal energy.

Efficient design and operation are key to reducing risks and enhancing the economic viability of geothermal projects (Wasch et al., 2019). These projects require informed decisions at every stage, from exploration and well placement to reservoir management and plant operation. Challenges such as resource uncertainty, scaling and corrosion, and pump degradation can severely impact performance. Leveraging high-quality data and learnings from past and ongoing projects can significantly support more effective decision-making and risk mitigation.

The geothermal industry is becoming increasingly data-rich, benefitting from both the growing number of geothermal installations and the knowledge base of adjacent sectors like oil and gas (Weers and Anderson, 2015; Vrijlandt et al., 2019). The availability of operational, geological, and performance-related data presents a significant opportunity to extract actionable insights – if these data sources can be effectively shared and utilized. Advances in digital technologies now make it possible to collect, process, and analyze data at a greater scale and frequency than ever before.

Artificial intelligence (AI) and machine learning (ML) are emerging as powerful tools for enhancing geothermal operations. Applications range from subsurface modeling and drilling optimization to failure detection and predictive maintenance (Shoeibi Omrani et al., 2025). However, many geothermal sites individually generate limited volumes of data, resulting in machine learning models with limited accuracy and poor generalizability. This limitation underscores the importance of data-sharing approaches that can harness collective knowledge from multiple installations.

Despite the recognized benefits of sharing data, practical implementation remains a key challenge. Concerns over confidentiality, intellectual property, and competitive advantage often discourage open data exchange. As a result, geothermal operators tend to work in silos, reducing opportunities for cross-learning and sector-wide improvement. This lack of collaboration not only slows innovation but also contributes to inefficiencies in project development and operation.

To address these issues, this report investigates the value of data-sharing technologies – particularly secure and decentralized machine learning approaches – for the geothermal sector. These technologies offer ways to collaborate on model development without transferring raw data, preserving privacy while still enabling shared learning. Approaches such as secure multi-party computation (SMPC) and federated learning (FL) allow organizations to build predictive models using distributed data across multiple stakeholders. These techniques can support advanced use cases such as equipment failure prediction, reservoir optimization, and emissions reduction – while maintaining data sovereignty.

The report provides a detailed overview of available data-sharing frameworks, their technical foundations, and their applicability to the geothermal context. It also explores the broader opportunities and limitations of data-driven collaboration in the sector. As part of this work, a curated list of potential case studies is presented, outlining how different types of stakeholders – geothermal operators, service providers, and researchers – can jointly benefit from secure data sharing. While no specific technology is demonstrated in this report, the case study scenarios highlight real-world applications where collaborative data utilization could provide tangible value, such as predictive maintenance of pumps, cross-site performance benchmarking, or corrosion monitoring under different fluid chemistries.

In conclusion, this report emphasizes that secure data-sharing technologies represent a critical enabler

for innovation, cost reduction, and performance optimization in the geothermal sector. By reducing barriers to collaboration, these approaches have the potential to significantly accelerate learning across projects and organizations. Moving forward, the implementation of digital solutions that respect data ownership while enabling collective intelligence will be key to scaling up geothermal energy as a reliable and competitive component of the global energy transition.

Data Sharing Technologies

In this section, we go over relevant data-sharing technologies and elaborate on our selection, highlighting their advantages, limitations, and applicability to geothermal energy systems. Recent advancements in AI and ML offer solutions to challenges associated with data privacy and interoperability, enabling efficient models for performance prediction and predictive maintenance. Secure Multi-Party Computation (SMPC) and Federated Learning (FL) are two prominent technologies designed to address these issues by enabling collaborative learning without compromising data privacy.

Secure Multi-Part Computation (SMPC)

Secure Multi-Party Computation (SMPC) is a foundational cryptographic approach designed to enable multiple parties to collaboratively compute a function over their respective private datasets without revealing those datasets to one another. Initially introduced through Yao's "Millionaires' Problem" (Yao, 1982), and further formalized by Goldreich (2004), SMPC has since evolved into a powerful framework for privacy-preserving data analysis and secure collaboration. At its core, SMPC guarantees that all parties learn only the final outcome of the computation, and nothing else about the other participants' inputs. This makes it particularly attractive in sectors where data sensitivity is crucial, e.g. in healthcare, finance, defense, and now increasingly, in energy systems and industrial applications. A schematic of the SMPC workflow is shown in Figure 1.

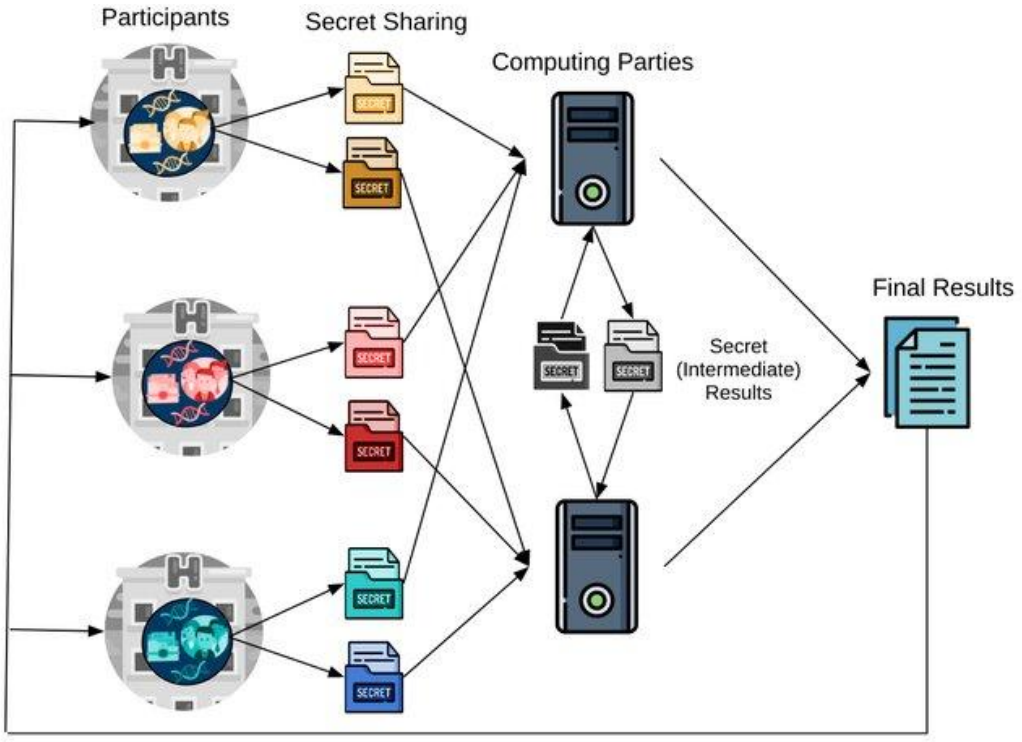

Figure 1. Secure multi-party computation: each participant sharing their data with the computing party and computing party calculate intermediate results to be securely shared with each other. The figure is adapted from Torkzadehmahani et al., 2020

One of SMPC's key advantages lies in its ability to support distributed trust models: no central party or data protector is required, thereby reducing the risk of single-point breaches. In practice, SMPC has been applied to a range of privacy-critical use cases, such as federated biometric matching, encrypted genomics processing, and secure statistical reporting (Rahaman et al., 2024; Lindell et al., 2008). Recent research has also explored its potential in energy and infrastructure sectors, where competing stakeholders may wish to collaborate on model development without disclosing proprietary operational or design data (Gamiz et al., 2025; Bonawitz et al., 2017).

However, despite its promise, SMPC is often constrained by computational and communication overheads, particularly when applied to large-scale or real-time machine learning tasks. Protocols must frequently exchange encrypted messages or secret shares across all parties, which can become prohibitively slow and bandwidth-intensive as the model complexity or dataset size increases. For instance, secure training of deep learning models or even moderately complex ensemble models under SMPC typically incurs non-linear increases in latency and computation time, limiting scalability. Furthermore, the need for synchronous execution and trusted setup environments in many SMPC implementations adds additional engineering burdens. To mitigate these challenges, hybrid privacy-preserving techniques are being proposed, such as combining SMPC with homomorphic encryption or trusted execution environments. These hybrid models attempt to leverage the fine-grained privacy control of SMPC while reducing the computational load by outsourcing certain operations or integrating lighter-weight privacy guarantees.

In summary, SMPC remains a promising technology for privacy-preserving computation, offering unmatched security in collaborative scenarios where data confidentiality is non-negotiable. While its computational overhead currently limits its use in high-frequency and real-time settings, ongoing advancements in protocol efficiency, hardware acceleration, and hybrid architectures promise to extend its applicability across sectors where collaborative learning without data exposure is becoming increasingly critical.

Federated Learning (FL)

Federated Learning is a decentralized ML approach where models are trained collaboratively across multiple devices or institutions without centralizing the raw data. Instead, local models are trained on individual datasets and only model updates are shared with a central aggregator, preserving data privacy (Konečnỳ et al., 2016). A schematic of the federated learning workflow is shown in Figure 2. Federated learning enables the development of a highly accurate and robust model while ensuring data privacy. Each participant trains a local model on their own data, which remains secure and undisclosed. A global model, accessible to all participants, is continuously improved by integrating generic insights from multiple local models. This collaborative learning approach enhances the performance of individual models while complying with data security and privacy policies, ensuring that sensitive information is never shared.

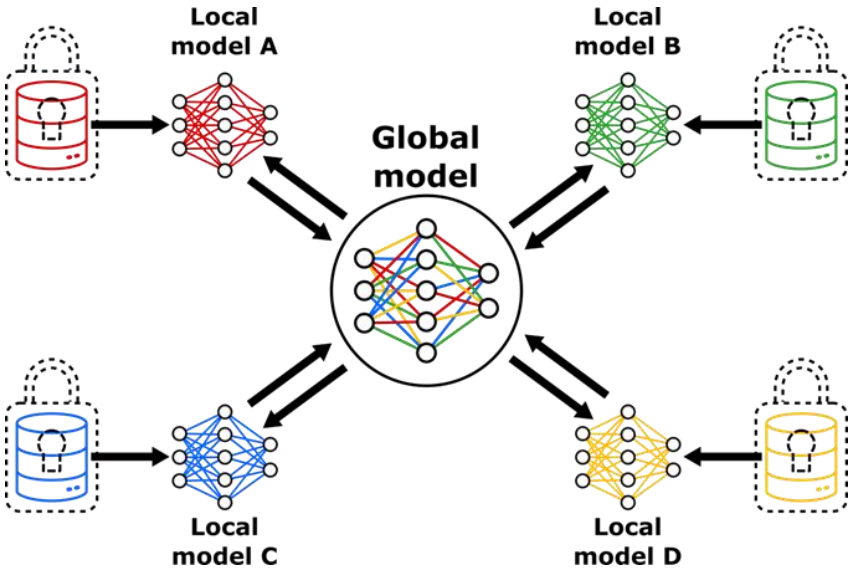

Figure 2. A schematic of the federated learning approach, developing a global model based on local models trained on the data of the stakeholder/clients.

Federated learning techniques are categorized into horizontal (sample-based) FL and vertical (feature-based) FL. Both methods are schematically shown in Figure 3. In horizontal federated learning (HFL), participants possess datasets with the same features but different samples. Vertical federated learning applies when parties hold different features about the same entities. Each approach enables privacy-preserving collaboration tailored to the structure of data distribution among stakeholders.

Horizontal FL is used when multiple parties have datasets with the same features (attributes) but different samples. In geothermal plants, this often means several operators each have similar types of data (sensor time-series, well performance metrics, etc.) from different plants or wells. A horizontal FL setup allows these operators to train a shared model (for example, a performance monitoring model for pump failure prediction) on the aggregated experience of all their facilities without exposing any facility's data. Vertical FL applies when different parties hold different features about the same sample entities. In geothermal contexts, this could occur if, for instance, an equipment OEM holds design specifications for a set of geothermal plants while the plant operators hold the operational performance data for those same plants – each party has distinct information on the same projects. Using vertical FL, they can jointly train models (e.g. relating design parameters to performance outcomes) by aligning on common identifiers (the plants) and keeping each party's feature set private. Both horizontal and vertical FL facilitate cross-institutional collaboration. Horizontal FL tends to be common when many operators or sites wish to pool learning on a common task (each site acts as a client in the federation with an identical model structure), whereas vertical FL is ideal for combining complementary data from different stakeholders (e.g. geological survey data with drilling results). In all cases, a central server orchestrates the training rounds, aggregating model updates from clients.

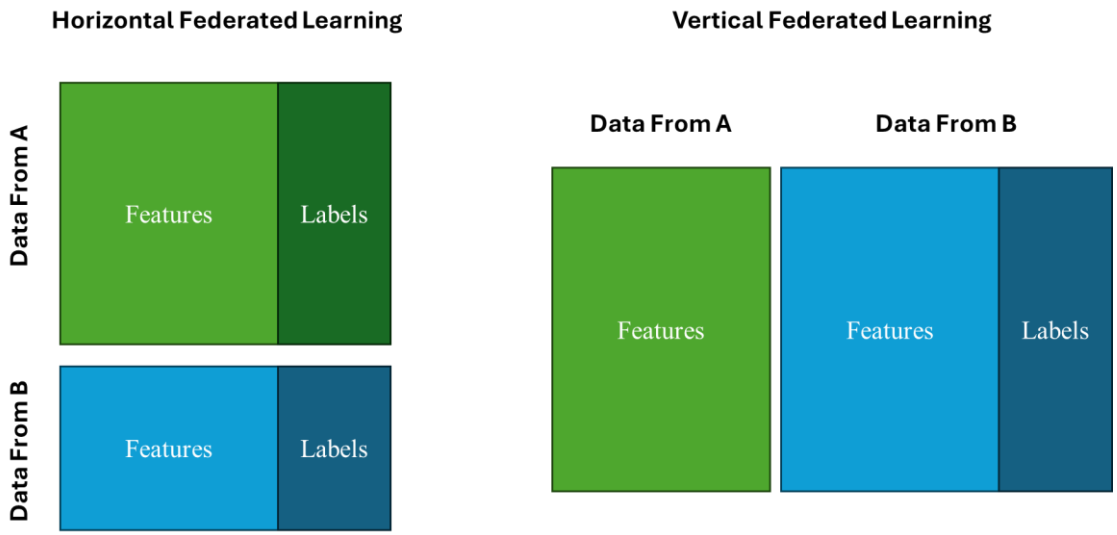

Figure 3. Schematic of two different type of federated learning (left) horizontal and (right) vertical federated learning

Comparison between SMPC and FL

Table 1 highlights key differences between Secure Multi-Party Computation (SMPC) and Federated Learning (FL) in terms of privacy, scalability, and communication. While SMPC offers strong cryptographic guarantees, it comes with high computational and communication overhead. FL, in contrast, enables efficient, privacy-conscious model training across decentralized data sources with lower resource requirements. For the geothermal industry, where data is distributed across multiple operators and sites, and real-time decision-making is important, FL seems to be the more suitable choice for enabling data sharing across the different stakeholders.

In the following section, we present several case studies showcasing how Federated Learning can be applied in the geothermal industry to enable collaborative data sharing without compromising data

privacy. These examples highlight the practical value of FL in environments where sensitive operational data is distributed across multiple stakeholders.

Table 1. Comparison of Secure Multi-Party Computation (SMPC) and Federated Learning (FL) with a focus on data privacy, scalability, and computational needs

Aspect	SMPC	FL
Data handling	Encrypted or partitioned data shared across parties.	Raw data stays local; only model updates shared.
Computational overhead	High computation and communication costs.	Lower overhead; more efficient for ML tasks.
Privacy	Strong cryptographic guarantees.	Depends on techniques like differential privacy.
Scalability	Limited to few participants.	Scales to thousands of devices.
Use cases	Secure joint analytics (e.g., finance, health).	Distributed ML (e.g., mobile apps, energy systems).

Case Studies in the Geothermal Sector

In the geothermal sector, data relevant to performance and reliability (e.g. well logs, chemical analyses, sensor readings, failure records) are often distributed across different companies and institutions. Data privacy and security are critical: operators may consider operational data confidential, and regulatory agencies enforce strict data protection. Data sharing technologies, and specifically federated learning, addresses these concerns by training models in a distributed manner: each participant (client) keeps their dataset local and only shares model parameters or gradients, not sensitive raw data. This approach has been shown to enable collaborative training across multiple clients "without sharing the involved training datasets," effectively overcoming the lack of data sharing that hinders innovation in renewable energy. From a security standpoint, modern FL implementations can employ encrypted communications and secure aggregation protocols so that even the model updates exchanged are shielded from interception or misuse. This privacy-by-design approach builds trust among geothermal stakeholders, encouraging cross-institutional projects. For example, a regulatory agency or research consortium could act as a neutral FL coordinator, allowing several geothermal operators and OEMs (original equipment manufacturer) to train a joint model (for equipment health, reservoir behavior, etc.) that benefits all participants. Each party gains improved predictive capability from the broader data pool, while their proprietary data remain local and confidential. This collaborative paradigm shift can accelerate learning across the geothermal industry, where data scarcity at any single site has traditionally limited machine learning effectiveness.

In this section, we provide a list of potential case studies that can be defined and implemented for data sharing in the geothermal sector. The case studies are categorized based on different disciplines, and within each section potential case studies for both horizontal and vertical FL is discussed. An overview of the case studies discussed in this report is demonstrated in Figure 4.

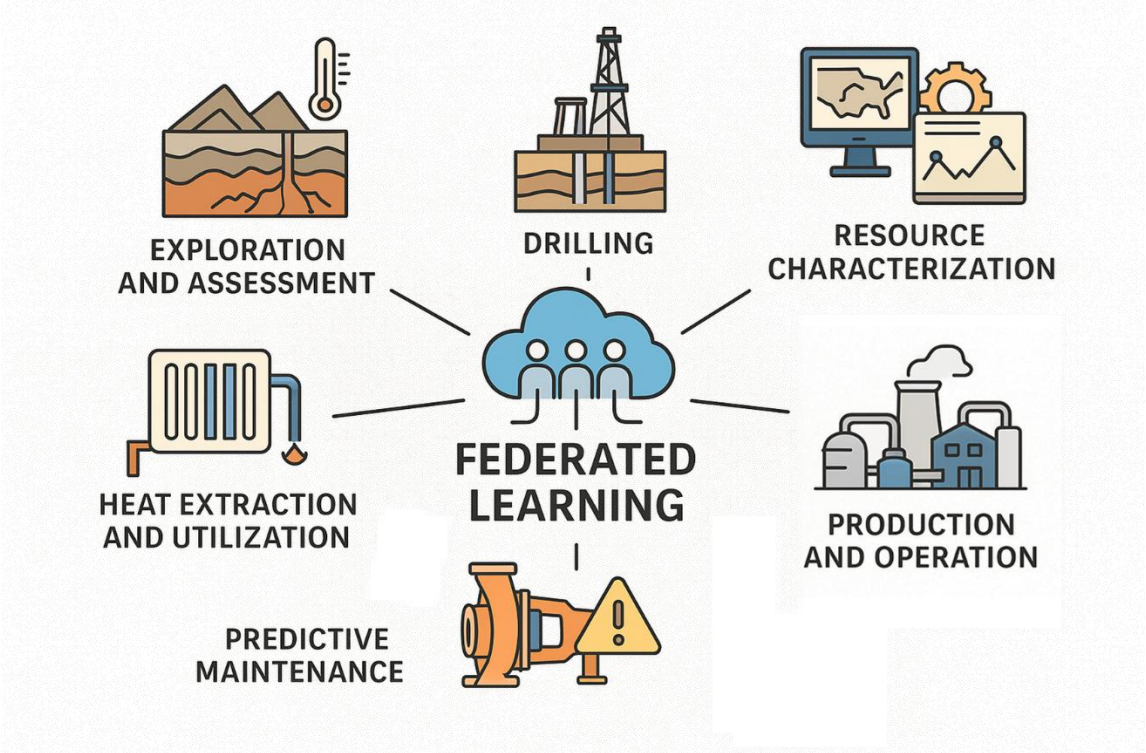

Figure 4. Overview of potential use cases and areas in geothermal energy that can benefit from data sharing and federated learning (the figure is generated by AI)

Predictive Maintenance for Geothermal Plant Equipment

Predictive maintenance is crucial for improving the uptime and safety of geothermal operations, especially in heating plants where equipment failures can cause costly heat supply interruptions. Key assets include electrical submersible pumps, surface/injection pumps, heat exchangers, separators, and filters. These components experience stress from high saline fluids, temperature cycling, and continuous operation, making proactive fault detection essential. Horizontal federated learning can dramatically enhance predictive maintenance in geothermal plants. Multiple operators each train a local model on their equipment sensor data (vibrations, pressures, temperatures, flow rates, etc.), and a central aggregator combines the learned patterns. Because all sites use a similar feature space (time-series sensor readings and maintenance logs), a horizontally federated model can learn from a much broader range of operating conditions than any single operator's data alone. For instance, one plant's pump might have experienced early-stage failure signatures that others have not, while another plant logged a rare motor failure mode, their data teach a more comprehensive failure-prediction model. FL ensures no raw sensor or failure data leaves the plant premises, easing concerns about exposing operational vulnerabilities or proprietary condition monitoring strategies. This collaborative approach directly supports predictive maintenance by yielding a model that can detect anomalies or predict failures with high confidence. In fact, federated anomaly detection models have achieved notable success in analogous industrial settings. For example, an FL-based pump monitoring model in a manufacturing context reached over 97% accuracy in detecting anomalies leading to failures (Ahn et al., 2023).

Such performance underscores the potential for geothermal operators to achieve early fault prediction (e.g. foreseeing a pump seizure or clogging weeks in advance) by pooling their datasets without centralizing sensitive information. Horizontal FL is appropriate here because each operator's data share the same structure – time-series sensor inputs and known failure events – and the value comes from aggregating many such examples. The result is improved maintenance scheduling, optimized spare parts management, and avoided unplanned shutdowns across the participating geothermal facilities. In short, FL enables a "virtual plant" of geothermal equipment to be monitored and learned from collectively, improving reliability industry-wide. Notably, if an OEM that manufactures the pumps also joins the federation with its test or design data, a vertical FL extension could enhance the model further; however, the primary paradigm for multi-operator predictive maintenance is horizontal FL.

Corrosion and Scaling Mitigation

Mineral scaling (precipitation of solids like calcite, silica, etc.) and corrosion in pipes and heat exchangers are pervasive problems in low-enthalpy geothermal operations involving sedimentary aquifers. The occurrence and severity of scaling and corrosion depend on a complex interplay of water chemistry (salinity, gas content, pH), temperature-pressure conditions, and mitigation measures (like chemical inhibitors). Operators typically collect water sample analyses and track when and where scaling or corrosion incidents occur (e.g. deposition in a reinjection well or thinning of a pipe wall). These data are often considered sensitive, as they can indicate a plant's efficiency or the need for costly workovers. Horizontal FL offers a way for multiple geothermal operators (or operation and maintenance (O&M) contractors) to collaboratively train a model to predict scaling and corrosion risks, without revealing their raw chemical data or frequency of problems to other operators. In a FL setup, each operator could use their historical data – for instance, inputs like fluid composition, temperature drop in the heat exchanger, flow rate, and operational changes, with labels indicating if significant scaling or corrosion was observed under those conditions. A global model aggregated from these local models could learn generalizable patterns that any participant can then apply on their own site. By using horizontal FL, all participants share the same modeling objective (e.g. classification of "scaling likely" vs "not likely" for given conditions, or a regression of expected scale deposition rate) and feature space, but each with their site-specific examples.

The benefit is a model that covers diverse geochemical regimes and operational strategies: one aquifer's water chemistry might teach the model about sulfate scaling, while another provides examples of carbonate scaling, etc. The FL process maintains privacy and builds collective knowledge on mitigating these issues. This can inform operators when to adjust injection strategies or dosing of inhibitors. It also aids OEMs and chemical suppliers by highlighting conditions where their corrosion inhibitors succeed or not, without each company directly sharing proprietary performance data. In summary, federated

learning helps create a comprehensive corrosion and scaling risk model drawing on industry-wide experience.

Plant Design and Optimization

Designing efficient geothermal plants involves selecting and sizing equipment (well design, screen selection, pumps, heat exchangers, filters) and operational setpoints to match the reservoir characteristics. Historically, design optimization has been limited by the few number of projects and the fact that detailed performance data from existing plants are rarely shared openly. Vertical federated learning can bridge this gap by allowing collaboration between the entities that design geothermal plants and those that operate them. Consider an engineering firm or an OEM that has designed multiple geothermal installations (thus possessing features such as equipment specifications, well depths, intended flow rates, design temperatures), and the various plant operators who have data on the actual performance metrics of those installations (achieved flow, output temperatures, efficiency, downtime statistics, etc.). Individually, neither party can easily build a predictive model relating design decisions to outcomes across many projects, since they each hold only half of the picture. Using vertical FL, however, the design firm and the operators can train a joint model on the combined feature set without ever exchanging their proprietary data. They align on each geothermal plant (the common entity) as a sample: the design firm provides its design parameters for that plant as input features, while the operator provides the realized performance indicators as output labels. A training iteration involves computing partial model updates that are securely exchanged and merged by a coordinating server (perhaps run by a neutral research institute). Through this feature-sharing federated approach, the partners can develop models that, for example, predict the long-term performance or cost-efficiency of a plant given certain design parameters and reservoir properties. Such a model could help optimize future plant designs (choosing equipment sizes or configurations that on average perform best) and also help operators fine-tune existing plants. Vertical FL is critical here because the data is partitioned by feature between two groups – each plant's full data is only revealed when the model is aggregated. This form of collaboration can also include researchers providing additional features (e.g. simulation-based performance metrics or geological context) and regulators interested in general design best-practices. Ultimately, federated learning in plant design leads to better design optimization guidelines and digital twins that are informed by real-world outcomes across the industry.

Drilling Optimization

Drilling geothermal wells in sedimentary basins can be challenging, often corresponds with issues like variable rock hardness, lost circulation zones, or wellbore instability (Shoeibi Omrani et al., 2025). Many factors affect drilling performance (rate of penetration, bit wear, occurrence of problems) including drilling parameters (weight on bit, rotary speed, mud properties) and geological conditions (lithology, formation pressures, depth). Gathering enough data to build predictive models for drilling optimization (e.g. to predict penetration rate or risk of complications) is challenging for any single operator, as individual companies drill a limited number of wells. Horizontal federated learning can enable cross-company drilling data analysis in a privacy-preserving manner. In this scenario, each participating entity (could be geothermal operators, or drilling service companies) uses its historical drilling datasets to train a local model that predicts outcomes like penetration rate given certain parameters and geologic log data, or classifies problematic zones. The local models' parameters are periodically sent to a central server and averaged (or otherwise aggregated) to form a global model that benefits from a much larger effective dataset encompassing wells from different fields and regions. Importantly, none of the raw drilling data (which may include sensitive information on drilling costs, techniques, or proprietary mud formulations) is shared, addressing competitive and confidentiality concerns. Using horizontal FL here is natural because all parties are training the same kind of model (e.g. a prediction model for drilling speed or torque) with the same set of input features (sensor readings, drilling settings, bit type, etc.), just drawn from different well sites. The aggregated model can capture trends that one company alone might miss. The collaborative model might also incorporate rare events, like unexpected high-pressure kicks or stuck pipe incidents, thereby improving drilling risk forecasting. This federated approach can be facilitated by industry consortia or research projects. By leveraging horizontal FL in drilling, the geothermal industry can reduce costs and improve safety, as improved models lead to shorter drilling times and more efficient processes.

Reservoir Characterization and Management

Understanding and predicting subsurface reservoir properties is key for efficient geothermal production. Reservoir characterization tasks include predicting parameters like permeability, porosity, temperature distribution, and pressure decline, as well as understanding the aquifer's response to long-term production and injection. Machine learning could assist these tasks (for instance, learning to estimate permeability from well logs), but geothermal operators often have limited well data per field (Okoroafor et al., 2022). Here, horizontal federated learning can play a transformative role by allowing multiple operators or field projects to collaboratively train reservoir models. Each operator may contribute a few data points – e.g. well log curves and core sample analyses paired with measured permeability or well test results. A shared model can be trained across all these data points via FL, greatly increasing the training sample size without centralizing the data. The horizontal FL approach fits because each data contributor has the same types of features (well log measurements, geological markers) and labels (reservoir properties or performance metrics) for different wells scattered across the region. The collaborative model effectively captures the collective geoscience knowledge of multiple sedimentary geothermal fields.

Exploration and Resource Assessment

Exploring for new geothermal resources or characterization of resources can be further enhanced with data sharing and federated learning. Successful exploration often requires correlating these large-scale datasets with ground-truth outcomes from existing wells (e.g. whether a test well actually encountered sufficient temperature and permeability). However, exploration data is typically fragmented. Federated learning provides a secure framework for combining these complementary data for improved exploration models. As an example for vertical FL, one party could serve as the holder of input features (for instance, seismic attributes or basin modeling results), while the exploration companies contribute the labels (e.g. "successful discovery" vs "dry well" or actual measured temperature and flow rate at those coordinates) for the locations where they have drilled. By aligning on common geographic locations or prospects, a vertical federated model can be trained to predict the probability of geothermal success or the expected resource quality in a target location. This FL-assisted exploration model has significant benefits. Data privacy and security are maintained, which is critical since exploration outcomes are often confidential and geophysical datasets can be proprietary or sensitive. At the same time, cross-institution collaboration is achieved: the public sector's data and private sector's data work in concert to improve understanding of geothermal potential. The resulting model might help identify high-potential sites (for leasing or further study) much more effectively than conventional methods, because it leverages far more data than any single entity possesses.

An example of employing FL for resource assessment in petroleum fields is demonstrated by Peng et al. (2024). As an example for the vertical federated learning, an oil company and an exploration institute collaborate to develop a machine learning model for predicting reservoir productivity. The oil company, with access to labels and engineering data, acts as the active party, while the exploration institute, possessing only geological features, serves as the passive client aiming to build its own predictive model. The schematic of the workflow is presented in the paper, as shown in Figure 5.

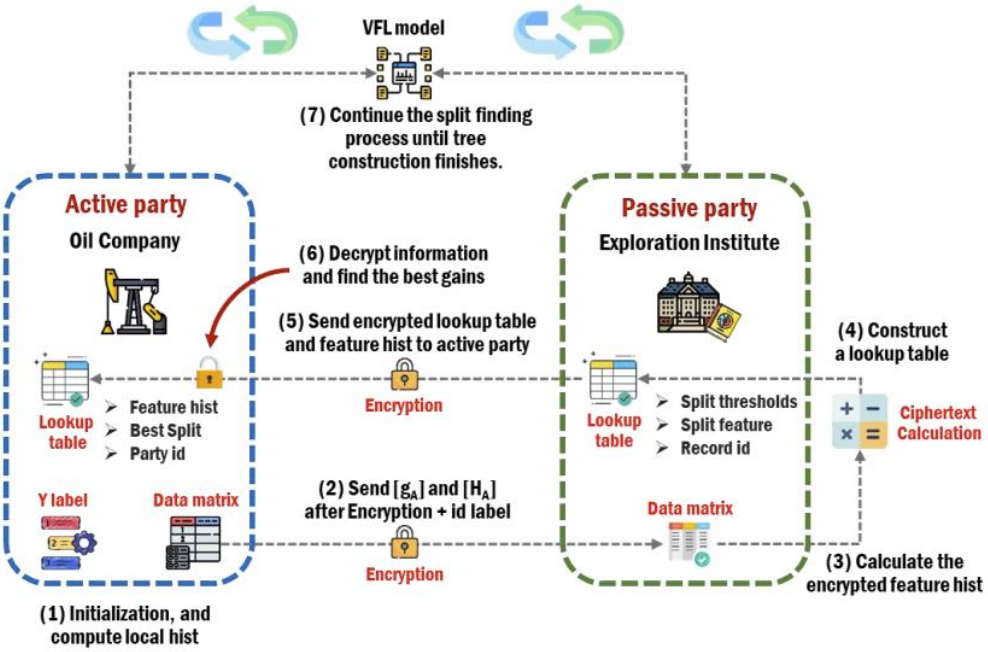

Figure 5. Workflow for an example of vertical federated learning for resource assessment in geoenergy systems (figure adapted from Peng et al., 2024, arXiv:2404.18527)

Beyond this specific arrangement, horizontal FL could also be applied among companies themselves (if multiple companies each have their own geophysical interpretations and drilling results, they could horizontally aggregate a model without involving a third party). In fact, a regulator might encourage several companies operating in the same basin to use FL to collectively de-risk exploration, improving the overall success rate of geothermal projects in the region. Such a move could lower costs and accelerate development of geothermal heating projects, aligning with policy goals, all while respecting competitive boundaries through privacy-preserving ML.

Challenges and Limitations

Despite the potential and added value of using data sharing through SMPC or federated learning technologies in the geothermal sector, there are several limitations and challenges in deploying these digital technologies. One significant challenge is **data heterogeneity and inconsistency**. Geothermal data originates from a multitude of diverse sources, including drilling logs, seismic surveys, well tests, and production data. This data often varies considerably in format, quality, and measurement units, which makes standardization and integration across different datasets a challenging task. Complementing this, **data scarcity and proprietary nature** pose a substantial hurdle. High-quality geothermal data, particularly from deep wells, is not only scarce but also expensive to acquire. Companies frequently consider such data proprietary, leading to a reluctance to share it, which in turn limits the scope for collaborative efforts.

Furthermore, **confidentiality and security** concerns are vital. Companies are understandably hesitant to share sensitive operational data due to intellectual property concerns, the risk of competitive disadvantages, and the ever-present threat of data breaches. This is compounded by a widespread lack of standardized data governance in the sector, meaning there are often no common protocols, legal frameworks, or ethical guidelines for data ownership, access, usage, and sharing among different organizations. This absence of clear rules contributes to interoperability issues, making it difficult to integrate data from various software platforms, sensors, and legacy systems that were not designed to communicate with each other.

The **technical complexity** of FL or any data sharing implementation itself presents a challenge. Successfully implementing FL requires specialized technical expertise, robust IT infrastructure, and handling of model aggregation and privacy-preserving techniques to ensure data remains secure while insights are gained. Building trust and collaboration gaps among competitors or different stakeholders (industry, academia, government) is also crucial for successful data sharing and collaborative learning, but this often proves to be a significant challenge.

For successful implementation of data sharing frameworks, **ethical** and operational considerations must be carefully addressed to ensure effective and responsible use. One concern is the lack of clarity in how decisions are made within data or model sharing frameworks. These models can be complex and difficult for participants to fully understand, making it challenging to detect biases or assess model decisions. Even with advanced predictive capabilities, the outcomes may not provide actionable insights for each individual participant. In addition, shared datasets might not be representative of the entire geothermal landscape, leading to biased models if certain geological settings or operational conditions are underrepresented. To mitigate this, it is crucial to adopt practices that improve the explainability of the models, allowing stakeholders to trust the process and results.

A critical concern in federated learning is the risk of **untrustworthy participant behavior**. Since FL involves collaboration across multiple sites, there is the potential for individuals or organizations to manipulate or falsify local data, or attempt to reverse-engineer shared parameters (adversarial attacks). Such actions could severely compromise the integrity, fairness, and security of the entire system, especially in critical infrastructure sectors like geothermal. To mitigate this, clear participation agreements, technical safeguards (such as differential privacy, secure aggregation, and anomaly detection), and a robust governance framework are essential. These should be accompanied by transparent auditing processes and accountability mechanisms to ensure ethical compliance and build trust among stakeholders.

Moreover, **model drift and generalizability** need to be assessed. Geothermal reservoirs are highly site-specific, meaning models trained on federated data might struggle with generalizability to new, unseen sites if the underlying geological characteristics vary significantly. This is closely related to data bias and skewness. As data originates from various geothermal sites, imbalances may arise, influencing model outcomes. If certain participants are selectively included or data from some sites is overrepresented, it could result in skewed or inaccurate models. Ensuring fair inclusion of all relevant parties in the collaboration is crucial for minimizing bias and producing reliable models. The complex issue of **attribution and incentivization** also needs to be addressed, as determining fair attribution for contributions to a shared model and creating incentives for participation can be complex, especially when data sharing is

not mandatory.

Ownership of shared data and/or trained models needs to be clarified at very early stage of the development. When data is contributed to a shared pool or used to train a federated model, clarifying who maintains ownership of the original shared data, the derived insights, and the resulting trained models becomes crucial. Establishing clear intellectual property rights and usage agreements among all participants is essential to prevent disputes and encourage participation in collaborative efforts.

Conclusions and Outlook

Data sharing in the geothermal sector has the potential to revolutionize how insights are generated across exploration, characterization, drilling, and operational stages. By enabling the aggregation of knowledge from geographically diverse fields and systems, collaborative modeling (especially when enabled by privacy-preserving frameworks like federated learning) can produce highly accurate and generalizable models. These global models outperform those trained on isolated, site-specific datasets by capturing a broader range of geological and operational scenarios. This leads to improved predictive capabilities, allowing for better-informed decisions in resource assessment, maintenance planning, and operational optimization, even for operators with limited data or underrepresented geological conditions. As demonstrated, such global models are essential in bridging data silos and reducing the reliance on large local datasets, particularly in mid- and low-enthalpy geothermal systems where project-specific data may be sparse.

From the technological perspective, both Secure Multi-Party Computation and Federated Learning offer valuable approaches to privacy-preserving data collaboration. However, given the distributed nature of geothermal operations, the need for scalability, and the focus on model training rather than secure computation, FL emerges as the more practical and suitable technique for the geothermal industry.

Moreover, federated learning enables this high level of model performance without compromising the confidentiality of sensitive data, addressing one of the key barriers to collaboration in the industry. Through decentralized training, stakeholders can jointly develop robust predictive tools while retaining control over their proprietary datasets. This advances trust and facilitates wider participation, encouraging innovation across the sector. The enhanced generalization capabilities of shared models reduce uncertainties in geothermal development and operations, ultimately leading to increased efficiency, reduced costs, and improved sustainability. As most of the sectors, including energy sector, move towards more data-driven decision-making, the added value of data sharing will be a basis of future competitiveness and resilience in geothermal energy systems.

The first step toward realizing the benefits of data sharing through federated learning is the establishment of a trusted collaborative framework among stakeholders. This includes geothermal operators, equipment manufacturers, research institutions, and regulators, all agreeing on a shared vision for secure and ethical data collaboration. Key components of this framework should include standardized data formats, clear governance protocols, participation agreements outlining data use and protection, and the deployment of technical safeguards such as secure aggregation and anomaly detection. Pilot projects should be launched to validate the framework in a controlled environment, demonstrating both the feasibility and the benefits of federated learning in real-world geothermal contexts. These initial pilots will help build confidence in the approach, refine governance mechanisms, and lay the groundwork for broader adoption across the sector.

References

- Ahn, J., Lee, Y., Kim, N., Park, C., & Jeong, J. (2023). Federated Learning for Predictive Maintenance and Anomaly Detection Using Time Series Data Distribution Shifts in Manufacturing Processes. Sensors, 23(17), 7331. https://doi.org/10.3390/s23177331
- Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Parcollet, T., de Gusmão, P. P. B., & Lane, N. D. (2020). Flower: A friendly federated learning research framework. arXiv preprint arXiv:2007.14390.
- Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1175–1191.
- Čaušević, S., Sharma, S., Ben Aziza, S., van der Veen, A., & Lazovik, E. (2023). LV grid state estimation using local flexible assets: a federated learning approach. IET Conference Proceedings, 2023(6), 1045-1049. https://doi.org/10.1049/icp.2023.0624.
- Gamiz, I., Regueiro, C., Lage, O., & others. (2025). Challenges and future research directions in secure multi-party computation for resource-constrained devices and large-scale computations. International Journal of Information Security, 24(27)
- Goldreich, O. (2004). Foundations of Cryptography: Volume 2, Basic Applications. Cambridge University Press.
- IEA (2021), Geothermal Power, IEA, Paris https://www.iea.org/reports/geothermal-power
- Lindell, Y., Pinkas, B. (2008). Secure Multiparty Computation for Privacy-Preserving Data Mining, Journal of privacy and confidentiality, issue 1
- McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. Y. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. In Artificial Intelligence and Statistics (pp. 1273–1282). PMLR.).
- Okoroafor, E. R., Smith, C. M., Ochie, K. I., Nwosu, C. J., Gudmundsdottir, H., & Aljubran, M. (2022). Machine learning in subsurface geothermal energy: Two decades in review. Geothermics, 102, 102401.
- Peng, W., Gao, J., Chen, Y., & Wang, S. (2024). Bridging data barriers among participants: Assessing the potential of geoenergy through federated learning. arXiv. https://arxiv.org/abs/2404.18527
- Rahaman, M., Arya, V., Orozco, S. M., & Pappachan, P. (2024). Secure multi-party computation (SMPC) protocols and privacy. In Innovations in Modern Cryptography (pp. 190–214). IGI Global.
- Shoeibi Omrani, P., Poort, J., & Shahmohammadi, S. (2025). Artificial intelligence in the geothermal energy systems. In Geothermal Energy Engineering (pp. 349-377). Elsevier.
- Torkzadehmahani, R., Nasirigerdeh, R., Blumenthal, D., Kacprowski, T., List, M., Matschinske, J., Späth, J., Wenke, N., Bihari, B., Frisch, T., Hartebrodt, A., Hauschild, A.-C., Heider, D., Holzinger, A., Hötzendorfer, W., Kastelitz, M., Mayer, R., Nogales, C., Pustozerova, A., & Baumbach, J. (2020). Privacy-preserving artificial intelligence techniques in biomedicine. arXiv. https://doi.org/10.48550/arXiv.2007.11621
- Vrijlandt, M. A. W., Struijk, E. L. M., Brunner, L. G., Veldkamp, J. G., Witmans, N., Maljers, D., & van Wees, J. D. (2019). ThermoGIS update: A renewed view on geothermal potential in the Netherlands. European Geothermal Congress 2019.
- Wasch, L., Creusen, R., Eichinger, F., Goldberg, T., Kjoller, C., Regenspurg, S., ... & van Pul-Verboom, V. (2019). Improving geothermal system performance through collective knowledge building and technology development. In European Geothermal Congress.
- Weers, Jon, & Anderson, Arlene (2015). DOE Geothermal Data Repository: Getting More Mileage Out of Your Data: Preprint.
- Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated Machine Learning: Concept and Applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-19.
- Yao, A. C. (1982). Protocols for Secure Computations. In 23rd Annual Symposium on Foundations of Computer Science (SFCS 1982) (pp. 160-164). IEEE