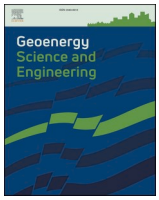
249 (2025) 213775

Contents lists available at ScienceDirect

Geoenergy Science and Engineering

journal homepage: www.sciencedirect.com/journal/geoenergy-science-and-engineering



Real-time model-based condition monitoring of geothermal systems under uncertainties – Case study on electrical submersible pumps

Pejman Shoeibi Omrani [a,b,*], Yifan Yang [a], Huub H.M. Rijnaarts [a,c,d], Shahab Shariat Torbaghan [a]

[a] Environmental Technology, Wageningen University & Research, Bornse Weilanden, Wageningen, the Netherlands
[b] Heat transfer and Fluid Dynamics Department, TNO, Kesslerpark 1, Rijswijk, the Netherlands
[c] Institute for Circular Society of the EWUU Alliance, The Netherlands
[d] Universities of Eindhoven, Wageningen, Utrecht and University Medical Centre Utrecht, The Netherlands

ARTICLE INFO

Keywords:
Condition monitoring
Proactive maintenance
Geothermal facilities and equipment
Data-driven decision support systems
Uncertainty quantification

ABSTRACT

Monitoring the condition of geothermal facilities and equipment (GFE) is crucial for ensuring reliable and cost-effective operations. This work emphasizes the importance of real-time data-driven condition monitoring for proactive operation and maintenance (O&M) planning in geothermal assets. Recognizing that operational planning can be significantly impacted by uncertainties, a novel framework is proposed to monitor the performance of geothermal assets under these conditions. The approach combines machine learning (ML), statistical methods, and expert knowledge to account for uncertainty in evaluating the degradation or onset of failure in GFE. This method was applied to field data from a geothermal plant to monitor Electrical Submersible Pumps (ESPs) and tested for the accuracy and robustness of the framework. Additionally, the framework provides explainability, aiding in understanding the factors influencing equipment condition and degradation. The framework was capable of systematically detecting the onset of the ESP degradation up to six months prior to its failure, with an accuracy of more than 95% in estimating the performance of ESP during normal operation. The explainability layer provided insights on the cause of the failure which was not attributed to ESP malfunction but to a restriction in production inflow into the well. The framework's ability to accurately assess equipment condition under uncertainty supports more informed maintenance decisions, ultimately improving GFE operational reliability and efficiency.

1. Introduction

Low and mid-enthalpy geothermal energy is an increasingly important part of the heat transition, making the heating of moderate climate horticulture sector and built environment free from natural gas (IEA, 2021). Geothermal plants consist of several equipment, parts, and components both in the subsurface and at the surface including wells, downhole pumps, separator, filters, heat exchanger, and valves. Despite an extensive operational experience in geothermal systems, characterized by a steep learning curve with ongoing operational knowledge acquisition, the sector is still in an emerging phase. The operation of geothermal assets is often associated with problems in the GFE caused by the chemistry of the geothermal brine (Ocampo-Díaz et al., 2005), production conditions such as pressure and temperature (Wasch et al., 2019), variability in the heat demand (Lund and Lienau, 2009) and operational errors (van't Spijker et al., 2016). In the current systems, occasional equipment degradation and malfunctions are still inevitable. It is necessary to monitor the assets and equipment condition constantly to prevent unplanned shut-ins and costs associated with the inspection, repair, and downtime. Tools providing insight and delivering predictability in GFE condition deterioration and malfunctioning are still in development (Siratovich et al., 2020).

Several technologies have been developed to monitor operational performance, utilizing either sensor data alone or in combination with models, a method known as model-based monitoring (Jaber, 2016). The latter relies on the continuous comparison of real-time data with predictive models to detect anomalies, optimize system performance, and enhance operational reliability (Surucu et al., 2023). These technologies can provide insights into equipment performance, up to real-time, and detect component and system anomalies (Chandola et al., 2009; Loh et al., 2018; Poort et al., 2020; Octaviano et al., 2020). The operation monitoring and maintenance planning is shifting gradually from

* Corresponding author. Environmental Technology, Wageningen University & Research, Bornse Weilanden, Wageningen, the Netherlands.
E-mail address: pejman.shoeibiomrani@tno.nl (P. Shoeibi Omrani).

https://doi.org/10.1016/j.geoen.2025.213775
Received 10 October 2024; Received in revised form 1 February 2025; Accepted 15 February 2025
Available online 18 February 2025

reactive and corrective to preventive and eventually predictive. Condition-based monitoring is a critical paradigm in the realm of predictive maintenance. Corrective maintenance refers to the process of repairing or restoring a system, equipment, or facility to its proper functioning state after a failure or malfunction has occurred which is a reactive maintenance strategy (Molęda et al., 2023). The preventive maintenance involves scheduled inspection and maintenance to prevent potential failures in the equipment which has lower inspection cost but lead to a higher failure cost for unplanned instances (Soh et al., 2021). The predictive maintenance allows for proactive maintenance prior to the failure by utilizing data analysis and monitoring techniques. Currently, most of the maintenance and operation planning is performed by operators' knowledge based on available data, generally without model-based decision support assistance.

As an example, an Electrical Submersible Pump (ESP) is a critical equipment in geothermal assets due to its vital role ensuring the desired production rates and high costs associated with their inspection, maintenance, and replacement. ESP can often suffer from performance degradation or failure due to changes in their efficiency, suboptimal production from the reservoirs, wrong operational settings, or an operator mistake. It has been observed that ESPs in geothermal systems on average have a shorter lifetime and they are dealing with several operational issues which are unique for geothermal systems (Shoeibi Omrani et al., 2021) due to utilization of larger pumps with a higher horsepower combined with high temperature and high salinity of the geothermal fluid. Apart from the production downtime due to ESP maintenance or replacement, there is a large cost associated with the replacement of ESPs up to 3% of the CAPEX (Capital Expenditure) (Octaviano et al., 2022). A robust ESP operational decision support system is therefore needed to operate these pumps in an optimal way, and to enable operators to perform early failure detection for improving maintenance planning.

There are several sensors providing monitoring data for the entire GFE, including different components such as ESP. This data can be used and processed to provide real-time insights on the performance of the geothermal production. Machine learning (ML) techniques are increasingly being applied to model and monitor equipment in process installation and energy assets (Shoeibi Omrani et al., 2018; Alatrach et al., 2020). Based on the provided example of ESPs, they come equipped with various sensors, including temperature, pressure, frequency, vibration, voltage, and current, enabling the utilization of complex ML methods. However, the multitude of potential failure modes associated with ESP systems poses challenges for condition monitoring and predictive maintenance (Shoeibi Omrani et al., 2021). Examples of prior work in this field, which are mainly from the petroleum sector, are summarized in Table 1. The publications are compared based on four different criteria, data-driven, expert knowledge, explainability, and uncertainty. Previous studies have employed a wide range of algorithms and methodologies for ESP monitoring, from detecting ESP failure patterns using principal component analysis (Gupta et al., 2016; Adesanwo et al., 2017; Bhardwaj et al., 2019) to leveraging deep learning techniques, such as convolutional neural networks (CNNs) (Lastra et al., 2021). Additional applications of ML in ESP systems include predictive maintenance (Abdalla et al., 2022), failure detection and diagnostics (Lastra et al., 2022), and the establishment of a digital ESP monitoring framework (Lastra, 2019). Few studies also focused on development of digital twin for ESPs to predict the remaining useful lifespan of ESP or ESP components such as stator windings (Don et al., 2024). Most publications focus on the application of individual machine learning and data-driven methods for monitoring or detecting ESP failures. However, as evident from the literature, the aspects of uncertainty and the use of ensemble models have received limited attention.

The challenges of data-driven condition-based monitoring are mainly two folds, lack of explainability of the data-driven models and dealing with uncertainties in the processes. The monitoring of GFE can be greatly impacted by the inherent uncertainties in monitoring data, uncertainties in the performance and lifetime of equipment components, and the intricate complexities involved in modeling the dynamic behavior of such equipment (Kullick et al., 2017). These factors complicate the operational decisions and maintenance planning and necessitate sophisticated techniques and robust predictive models to accurately anticipate and mitigate potential operational problems in geothermal systems. Applying ensemble ML techniques can provide more confidence in the prediction made by ML models for condition monitoring (Surucu et al., 2023). By employing data-driven models, the challenge of modelling the complex production behavior in geothermal assets can be partially tackled, however these models are lacking explainability. Explainable AI enhances data-driven decision-making by providing transparent and interpretable insights into the models' predictions, enabling users to understand, trust, and effectively act on the insights from data-driven models (van Gerven et al., 2019).

To date, limited efforts have been made to explore the impact of uncertainties in monitoring geothermal systems or predicting component performance and failures. In a thorough and systematic review of application of data-analytics in the operation of geothermal systems (Abrasaldo et al., 2024), the word 'uncertainty' was hardly mentioned in the literatures reviewed in this publication. Looking into other sectors, in one of the studies performed by Dussi et al. (2022), Bayesian neural networks were employed to predict ESP degradation in oil wells across different forecasting horizons, achieving remarkably low error scores of 2.5% for a ten-day prediction horizon by explicitly incorporating causal relationships between input features and the associated probability of the specific failure. A recent study by Costa et al. (2024) demonstrated the integration of uncertainties with deep learning models for simulating ESP systems in oil wells. The study detailed the development and testing of NARX DNN models for ESP simulation, where synthetic data was generated to meet the data requirements of this approach, and Bayesian inference was applied for uncertainty assessment. The results confirmed the accuracy of NARX DNN models in predicting ESP system parameters. However, the incorporation of uncertainties into real-time condition monitoring of geothermal assets and facilities, where fast and robust methods are essential for handling field data, has not yet been explored.

ML models are black-box, and their performance can be impacted by several parameters such as data availability, data variability, and data quality. Hence, operational decisions based on a single ML model without considering uncertainties and explainability can be misleading. To monitor geothermal plant equipment's performance and assist with operational decisions, we propose a novel data-driven framework. In this paper, we developed a framework for model-based condition monitoring of geothermal assets by integrating data-driven models, with explainability and expert knowledge by leveraging ensemble ML models and uncertainty metrics. The framework was tested on the data of a geothermal ESP in a low-enthalpy geothermal asset to demonstrate its accuracy and performance.

Unlike traditional approaches that primarily focus on model selection or rely on multi-variate data analysis, the proposed method adopts a heuristic process to design and implement an ensemble-based solution. By integrating multiple statistical metrics and explainability tools, the framework aims to offer robust real-time anomaly detection and degradation monitoring, providing insights into potential failure causes. This work not only highlights the reasoning behind model design but also expands the scope of condition monitoring by demonstrating the practical application of uncertainty metrics beyond conventional prediction error metrics. It is important to emphasize that the primary focus of this paper is not on optimizing machine learning model architecture or hyperparameters for system performance evaluation.

The organization of this paper is as follows: first, the Methodology section provides a schematic overview of the developed framework, detailing each component, including model training, explainability, ensemble ML models for uncertainty analysis, metrics calculations, and operator knowledge in anomaly detection. Next, the case study for

Table 1
Overview of studied literature and publications related to condition monitoring of geothermal systems and ESPs (both in geothermal and petroleum applications) compared based on four criteria: data-driven modelling, integration of expert knowledge, explainability of the approach and dealing with uncertainties.

Reference	Data-driven	Expert knowledge	Explainability	Uncertainty
Karnik et al. (2021); Adesanwo et al., 2016**; Gupta et al., 2016**; Adesanwo et al., 2017**; Bhardwaj et al., 2019**; Sherif et al., 2019**; Peng et al., 2021**; Nanavaty (2024); Hamedi Shokrlu et al., 2024	✓	x	✓	x
Guo et al. (2015); Andrade Marin et al. (2019); Jansen van Rensburg et al. (2019); Zulkarnain et al. (2019); Alamu et al. (2020); Abdurakipov (2021); Lastra et al., 2021, 2022; Abdalla et al. (2022); Alhashem et al. (2024)	✓	x	x	x
Zhao et al. (2006); Xi (2008); Li et al. (2008); Tao et al. (2011); Zhao (2011); Zhang et al. (2017)	x	✓	x/ ✓*	x
Rauber et al. (2017); Sharma et al. (2022)	✓	✓	x	x
Tandazo et al. (2022); Octaviano et al. (2022); Irl et al. (2023)	✓	✓	✓	x
Costa et al. (2021)	x	✓	x	✓
Dussi et al. (2022); Mello et al. (2022); Costa et al. (2024)	✓	x	x	✓

* The methods based on expert knowledge and physics-based modelling are inherently explainable, however they do not directly provide parameters' importance on observed processes, like explainability applied to ML learning models.
** These studies utilized Principal Component Analysis (PCA) for event detection in ESPs, leveraging a data-driven approach that provides a certain degree of explainability in the results.

which the framework was tested and validated is described. Subsequently, the results obtained from the framework testing are presented and analyzed. In the conclusion section the highlights of the results and proposed next steps are presented.

2. Methodology

The schematic and steps of the proposed framework is shown in Fig. 1. The framework depicted in Fig. 1 outlines an approach to use ML models for condition monitoring under uncertainty and explainability. The framework consists of five steps and below is a detailed description of each step.

Step 1: Model training (Data collection and ML models trainings using historical data).

Initially, the historical data needs to be used to establish the base model for condition monitoring. The focus is on historical data to train the ML models representing the optimal state of the component or system. Here, 'optimal' does not refer solely to maximum efficiency

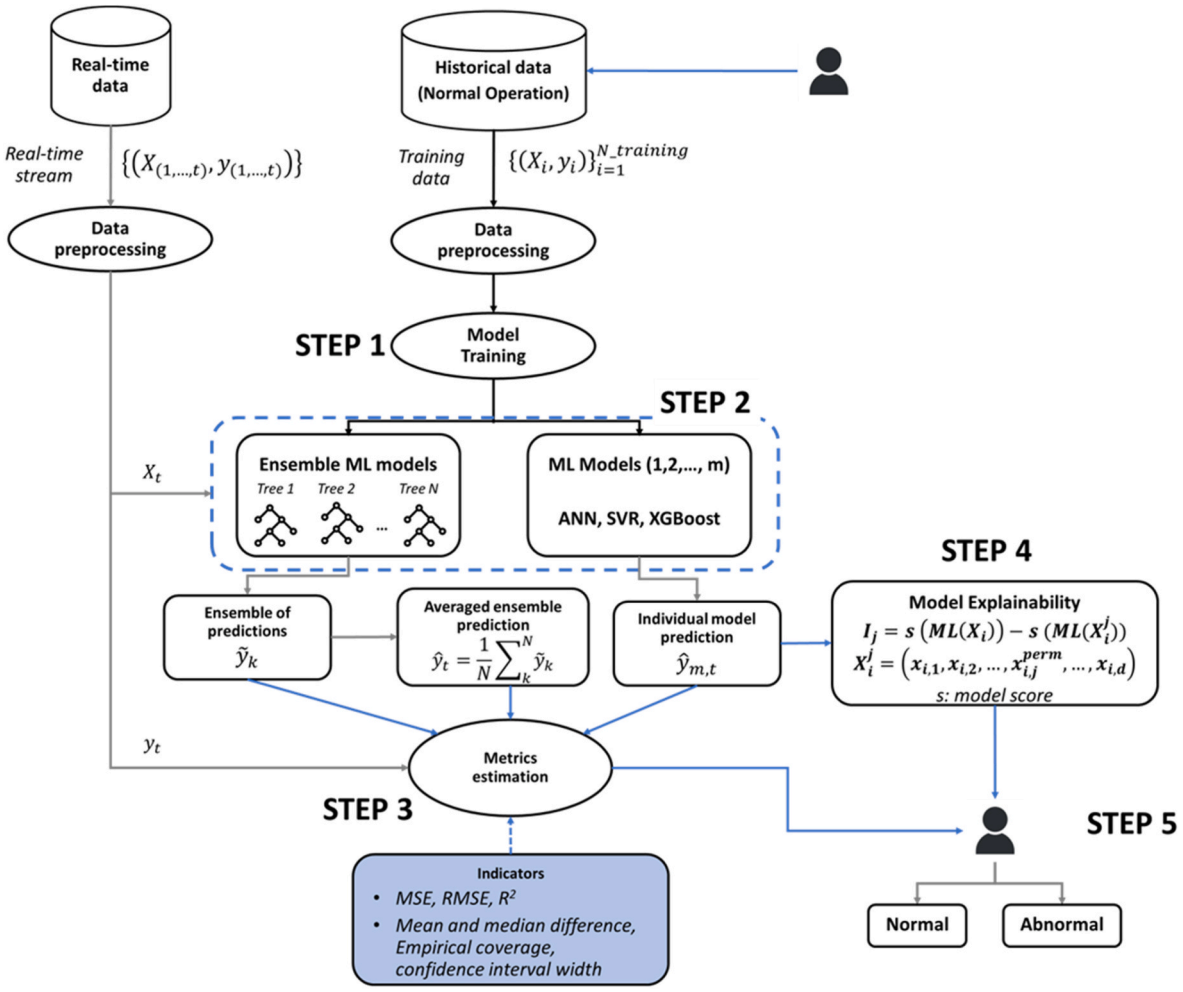


Fig. 1. Schematic of the framework for real-time monitoring of geothermal assets and facilities, X, y stands for the input and outputs of the models, respectively, to be determined based on the quantities to be monitored, I: imputation based on the score of the ML model, Subscript t: time, m: model, k: index of trees in the ensemble ML model (in this case quantile forest regression).

operation but encompasses operation across various points within the operating envelope where no degradation or malfunction of the equipment is observed. Thus, selection of the historical data to be used for ML model is of great importance. This process typically begins with consultation with subject matter experts (SMEs) or the plant operator to determine which signals' period best represent the healthy period of the equipment. If such information is not available, unsupervised machine learning such as clustering or dimensionality reduction can be utilized to provide insights on degrading state of the equipment. Once the selection of the normal period for the model training is done, a ML model is trained following a set of procedures typically involve data preprocessing, feature selection, model training, and hyperparameter tuning to ensure robust and reliable performance on unseen data.

The preprocessing step is employed to smoothen the data noise, address missing values, and identify and rectify any unphysical data anomalies. The choice of the appropriate filter (such as moving average or Savitzky-Golay filter), interpolation method (e.g. linear or polynomial), and data imputation (e.g. listwise deletion or regression based), and denoising techniques (e.g. wavelet transforms) depends on the specific problem, as well as the type and quality of the data. The selection of these techniques must align with the data characteristics to ensure meaningful and accurate preprocessing results. After the selection of the historical dataset, the parameters of interest and features need to be selected to support the monitoring of the plant or equipment. The essential elements for building a baseline model representing the normal state of the equipment is prepared. In this framework, we propose the use of ML models for condition monitoring due to the requirements for calculation speed (to be deployed in real-time) and complexities of the equipment and processes involved.

Two sets of ML models are trained on historical data, a set of individual ML models and ensemble ML models. For the individual ML models, a variety of ML models (ML Models (1, 2, …, m)) are trained on the historical data and the ones used in this study are random forest (RF) regressor, support vector regression (SVR), artificial neural networks (ANN), and extreme gradient boosting (XGBoost). RF, an ensemble method using bootstrap aggregation of decision trees, improves prediction accuracy through averaging and cross-validation, making it robust against bias and variance (Breiman, 2001). SVR aims to find a hyperplane that best fits data points within certain margins, focusing on support vectors to minimize regression errors (Smola et al. 2004). ANN, inspired by neural connections in the human brain, uses a multi-layer perceptron architecture where weights and biases are optimized to minimize prediction errors (Hastie et al., 2008). XGBoost, an algorithm within the gradient boosting framework, sequentially builds decision trees to correct errors from previous iterations, making it effective for sparse data (Chemura et al., 2020) and enhancing predictive performance through iterative improvements (Lu et al. 2020).

These models were chosen due to their general applicability and well-established performance in regression tasks in similar applications (Abdalla et al., 2022). While these models can handle temporal data,

they do not explicitly model temporal dependencies. Models such as Long Short-Term Memory (LSTM) networks, which are better suited for time-series and temporal behavior, could also be considered in future studies. It is important to note that LSTM would require substantial amount of data for an effective training.

For the ensemble ML models to quantify uncertainties and confidence bounds of the prediction, quantile regression forest was employed. Quantile regression forest is an extension of the RF regression model that allows for the estimation of conditional quantiles, rather than just the conditional mean (Hastie et al., 2008). The quantile forest algorithm works by growing an ensemble of regression trees, where each tree is trained to predict a specific quantile of the target variable's distribution. This provides a more complete picture of the conditional distribution. Quantile regression forests have been shown to outperform standard random forests in applications where the conditional distribution exhibits heteroscedasticity or skewness (Meinshausen, 2006). The selection of this method was due to its robust mechanism to estimate prediction intervals and allowing for quantification of uncertainty and statistical analysis in anomaly detection (Li et al. 2023). Each model is optimized to predict the target variable and the optimized and trained models are used in the following steps.

Step 2: Prediction step (ML model testing using real-time data).

Following the establishment of the baseline model, real-time data is connected for condition monitoring. For this purpose, model prediction emerges as a key component of the framework, facilitating the detection of deviations between predicted and measured values. Real-time data which is continuous data streams are collected from sensors or other data sources and will be fed directly into the ML models for processing and prediction. To enhance the effectiveness of this predictive approach, various trained ML models outlined above are deployed for error estimation. Several predictions are made in this step. Each individual ML model (denoted by m) generates its own prediction ($\hat{y}_{m,t}$) using the same (real-time) input data. The individual predictions provide additional insights into the status of GFE. Another set of predictions is performed using ensemble ML models. From such ensemble ML models, ensemble of predictions can be generated ($\widetilde{y}_{k,t}$) based on real-time data inputs. In addition, the predictions from the ensemble are averaged to yield a single prediction value at time t ($\hat{y}_t$). The output of the model's prediction is further forwarded to the next steps for metrics estimation (Step 3) and model explainability (Step 4).

Step 3: Metrics estimation.

Uncertainty can arise from several sources, including the selection of model forms and errors inherent in the model itself, which aligns with the approach taken in this work to assess model reliability through statistical metrics. In this study, uncertainty is defined as the dispersion of model predictions within the ensemble of machine learning models, similar to the concept described by Der Kiureghian and Ditlevsen (2009), where uncertainty is viewed as the variability in model outputs (a.k.a. variance). This approach follows a common practice in machine learning, where uncertainty is expressed by analyzing the spread of predictions around a central value, providing a statistical measure of the confidence in the model's predictions, with wider dispersions indicating greater uncertainty.

The ML models prediction will be compared with the data streams which are recorded from the GFE. To detect whether the GFE is under a degradation or an anomalous state, several metrics are required to compare the estimation of the ML model with the actual measured value from the sensors. The choice to use multiple metrics, rather than just one, has been extensively studied across various applications. Research has shown that relying on a single metric often provides an incomplete representation of processes, particularly in real-world scenarios (Ribeiro et al., 2016). The detection of anomalies based on models' prediction on the real-time data is performed using several metrics, including.

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Coefficient of Determination (R^2)
- Mean and median differences (for ensemble ML models)
- Empirical coverage (for ensemble ML models)
- Confidence interval width (for ensemble ML models)

Firstly, the definitions of the first three metrics which are used for each individual ML model prediction (with subscript m) are given below:

$$MSE_m = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_{m,i}\right)^2$$

$$RMSE_m = \sqrt{MSE_m}$$

$$R_m^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{m,i}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}$$

In which.

n the number of data points in the prediction time window.

y_i the actual value of the target variable for the ith data point.

$\hat{y}_{m,i}$ the predicted value of the target variable for the ith data point for the ML model m.

$\overline{y}$ mean of the actual target variable values (also denotes as μ).

The subscript m is dropped in the results and analysis section as these metrics are reported for each ML model separately. It is important to note that these metrics are often used to evaluate the accuracy of the models. However, in this context we compare the estimated value of the ML model with the actual measured value from the real-time data stream to detect the mismatch between these values and potentially flag it as an alarm to the operator of the plant.

For the ensemble ML model predictions from the quantile regression forest, several additional metrics can be derived. These indicators were evaluated for a selected prediction time window and were namely, absolute difference between the mean and median of the predicted values, width of the confidence interval bound and empirical coverage. Each of the metrics are formulated as follows:

$$Absolute\ Difference = |\mu - M|$$

$$Confidence\ interval\ width = F^{-1}\left(0.5 + \frac{q}{2}\right) - F^{-1}\left(0.5 - \frac{q}{2}\right)$$

$$Empirical\ coverage = \left(\frac{\sum_{i=1}^{n} I(x_i \in A)}{n}\right),$$

$$I(x_i \in A) = \begin{cases} 1, & if\ x_i \le F^{-1}\left(0.5 + \frac{q}{2}\right) and\ x_i \ge F^{-1}\left(0.5 - \frac{q}{2}\right) \\ 0, & otherwise \end{cases}$$

In which.

M median of the predicted variable values.

μ mean of the predicted variable values.

F^{-1} Inverse of the cumulative distribution function of the predicted values.

q desired confidence interval, a value between 0 and 1 (in this case 0.95).

I an indicator function as defined above, to count the number of data points between the upper and lower bound of the confidence interval.

The first metric provides information on the skewness of the data distribution as one of the metrics for deviation from the expected behavior. One of the common figures for anomaly detection is based on median absolute deviation (MAD) which is described as the median of the absolute deviations from the data's median (Dodge et al. 2010). In the context of anomaly detection in this paper, anomalies are identified by analyzing a window of predicted data points rather than individual points, focusing on structured mismatches between the measured and estimated values (Sagoolmuang et al., 2017). Thus, instead of using

MAD as a measure of anomaly, the mean and median of the predicted data are calculated to detect these anomalies. Another metric which is evaluated on the ensemble ML model is the confidence interval width. If the confidence interval width is increasing over time, this is an indication for growing uncertainty in the predictions as time progresses and the model accuracy is decreasing (Hochenbaum et al., 2017). The final metric is empirical coverage of the confidence interval (Schall, 2012). This parameter refers to the proportion of confidence intervals that contain the estimated values. By having a higher coverage of the estimated data points within the confidence interval of the ensemble ML model prediction, higher the probability of the system or equipment to be in the normal condition. The metrics above are all calculated for the trained models using real-time data. At this stage, no decision is made on the state of the system or equipment and only different indicators are estimated and passed into the next step for explainability and anomaly detection.

Step 4: Model explainability.

Explainability is a crucial aspect for condition monitoring as they provide insights into the factors driving equipment performance and potential failures, enabling operators to prioritize resources and interventions effectively. Understanding which features have the most significant impact on equipment condition, not only enhances predictive accuracy but also improves transparency and trust in the monitoring system's decision-making process. In this framework, permutation

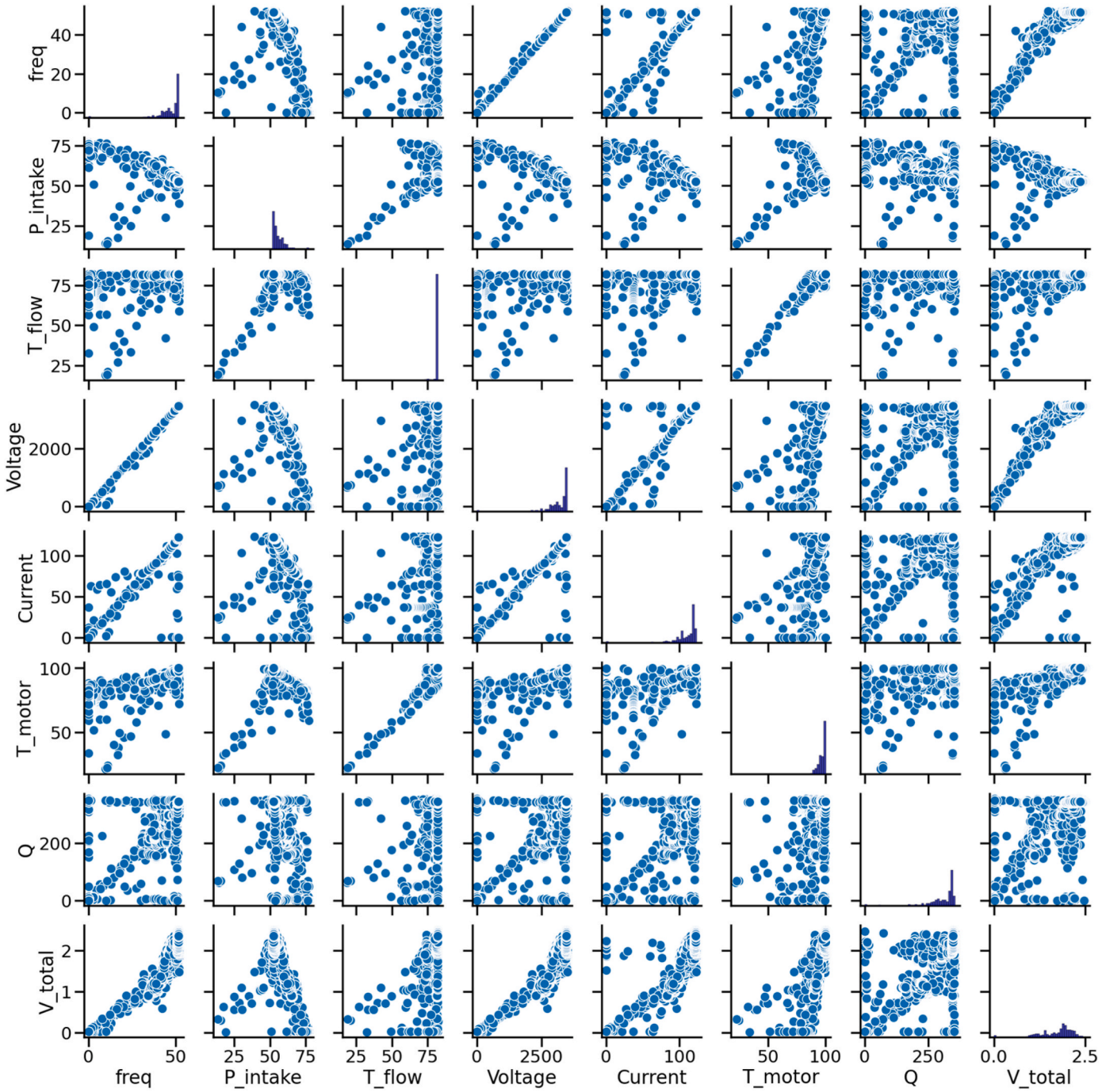


Fig. 2. Pairwise relationship between the variables in the dataset (frequency of the pump, intake pressure of the pump, flow temperature, voltage, current, motor temperature, flow rate and total vibration). All the variables are included in both x and y axis with the same order. Each individual scatterplot (off-diagonal plots) illustrates the relationship between two variables. The diagonal plots represent the distribution of each variable as histograms.

importance was used as a method to assess the importance of features in the developed and trained ML models. The method works by randomly shuffling the values of a feature and measuring the drop in the model's performance (Altmann et al., 2010). This method was selected due to being model-agnostic and easy to interpret which is key in operational decisions. This step will enable targeted intervention strategies for inspection and maintenance by providing insights on the most influential parameters causing a deviation from the normal operation state.

Step 5: Anomaly detection with expert knowledge.

The predictions and their associated metrics are analyzed to determine the state of the equipment or asset. The outputs of Step 3 (metrics estimation) and Step 4 (explainability) are used as inputs for the anomaly detection. This step is the point where expert knowledge is integrated with the data-driven model, uncertainties, and explainability which enable targeted intervention strategies an operator (or in the future an automated system) can do. In the context of a decision support system, the operator reviews the model outputs and all the calculated metrics together with the model explainability results to classify the status as either "Normal" or "Abnormal".

As the first step, the calculated metrics are used for the detection of anomalies and events. Any change in the estimated metrics over time can be indicated and flagged as an anomaly. In order to minimize the number of alarms to the operator, thresholds can be set. Operators can set custom thresholds for each of the estimated metrics for anomaly detection, ensuring that the system triggers alarms based on the sensitivity and risk tolerance appropriate for the specific operational context. Selecting a threshold for each indicator is critical and impacts the performance of the condition monitoring tool (Choi et al., 2021). If the threshold is set at a high value (higher error to be detected as event), it may lead to a high number of undetected events and if a lower value is set, then a minute deviation from the threshold caused by e.g. noise in the data can also be detected as an event and leads to overwhelming the operators. In this study, we did not aim to recommend a threshold for the detection of anomalies in geothermal systems but rather provide an overview of different indicators trends near the degradation or anomaly regimes.

The operator has the flexibility to configure alarms for anomaly detection based on operational protocols, or criticality of the process, allowing them to decide whether alarms should trigger when any single indicator signals an anomaly or only when multiple indicators align. This decision is based on predefined thresholds and rules derived from historical analysis or the trends which are available in the handbooks (Octaviano et al., 2022).

In the next step, the model explainability is used to support the operator for confirming the anomaly. Since data-driven models are used in this framework, their performance and accuracy will be impacted by extrapolation. Thus, a mismatch between the prediction and measured values can be caused by extrapolation errors, e.g. when in the evaluation phase it is the first time that the pump is ramped up to its maximum frequency and the ML models for ESP were not trained in this condition. The feature importance estimated by explainability can provide further insight on the actual reason of the mismatch prediction by the model. Operator and experts' wisdom will be used to analyze the model explainability to flag it as a true event or not.

3. Case study

3.1. Dataset and pre-processing

In this study, the ESP in a low-enthalpy hydrothermal geothermal well was used as a case study to demonstrate the proposed monitoring framework. Two years of data from this well was provided with an hourly data acquisition frequency. Several sensor data of the geothermal production was provided, including flow rate, wellhead pressure, wellhead temperature and separator pressure. For the ESP, an extensive dataset was provided including pump frequency, intake pressure, motor temperature, voltage, current, and vibration in two horizontal and vertical directions along pump shaft (x and z). An overview of the provided data and pairwise relationships between variables in the dataset is shown in Fig. 2 (using Seaborn Python Library (Waskom et al., 2017)).

Missing data and entries (e.g. containing NaN values) were replaced using linear interpolation. Outliers were identified based on a statistical threshold, defined as values exceeding three standard deviations from the mean, and were replaced with interpolated values. The data received from the system had already undergone a moving average filter, and no additional data smoothing was performed. Finally, min-max scaling was applied to normalize the data within the range of 0–1, ensuring consistency across all features. For other applications or systems, data preprocessing may require alternative algorithms or filters depending on the specific data quality and characteristics, such as noise levels or the frequency of missing values.

3.2. Data splitting and feature selection

According to the proposed framework and to construct a data-driven model for the ESP (Step 2), several supervised learning techniques were employed consists in training the model based on known input-output pairs, before using the model on 'unseen' data. Since, the goal of this method is to derive a model-based condition monitoring algorithm to detect abnormalities or degradation in the geothermal systems, the training period should consist of data associated to the normal condition of the production, meaning that no degradation or malfunctioning occurred yet. The historical data containing the normal operation of the equipment is used for training the ML models. This dataset is split into the training and validation subsets using a fixed random seed and no shuffling to preserve the temporal order. From the total historical dataset, 80% was used for training and 20% for validation. The current dataset only contains 0.7% of missing data.

This evaluation should be done carefully to ensure that the trained model can predict the production performance under the new conditions otherwise the mismatch can also be associated to models' inaccuracy. It is crucial to test the model using data in a similar range as the training set, and for this purpose the distribution of the data in the training and test set was compared to ensure the data in both sets have a similar distribution. A large difference of means and distribution range between training and testing set were not observed. Based on the advice from the pump operator, the degradation happened at an uncertain period in the second year of operation, therefore, it was assumed that the first year of operation are data of the ESP being in normal condition. The remaining data was used for testing model's capability in predicting normal ESP performance and detecting off-normal behavior and/or degradation. The histograms in Fig. 3 illustrate the distribution of feature values in the training and test datasets. Understanding the similarities and differences in feature distributions between these datasets is essential for robust model generalization and performance evaluation. Differences in dataset distributions, potentially attributed to degradation of the production, are observable; however, notable overlaps suggest that errors in the test set may not solely arise from extrapolation.

Three ESP parameters were selected to assess the condition of the ESP in real-time, namely motor temperature, total vibration, and intake pressure which are typical in condition monitoring of ESPs (Mohamad et al., 2022). Motor temperature often provides indication on ESP performance, overheating (often resulted from wear, lubrication or electrical faults) and insulation degradation (Hoevenaars et al., 2021). The total vibration is a direct measure for mechanical condition of the pump resulted from imbalances, misalignments, or bearing issues and ESP intake pressure can hint towards flow assurance, blockages, or reservoir performance (Iranzi et al., 2024). For each output parameter, a certain collection of input features from the original dataset was selected, as shown in Table 2.

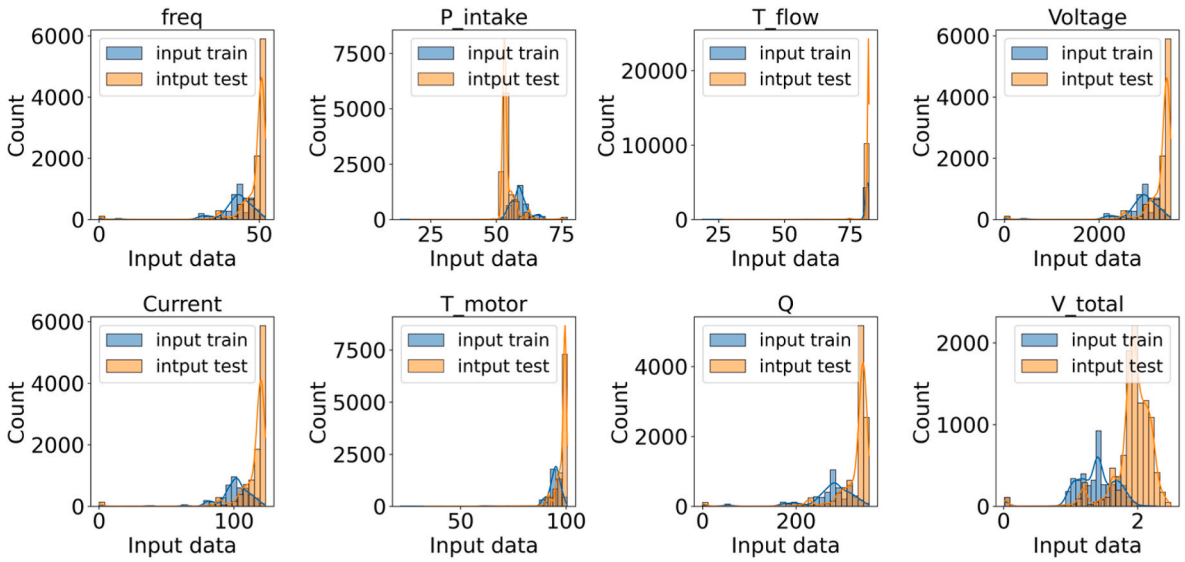


Fig. 3. Comparison of variables' distributions between the training and test datasets for all the 8 parameters in the dataset, including ESP frequency (freq), ESP intake pressure (P_intake), flow temperature (T_flow), voltage, current, ESP motor temperature (T_motor), mass flow rate (Q), and total vibration (V_total).

Table 2
The input features and output parameters of ML models.

Input features	Output parameter
$f, P_{in}, T_{flow}, Voltage, current, V_x, V_z, Q$	Motor temperature: T_{motor}
$f, P_{in}, T_{flow}, Voltage, current, T_{motor}, Q$	Total vibration: V_{total}
$f, Q, P_{wh}, T_{flow}, T_{motor}, Voltage, current$	ESP intake pressure: P_{in}

3.3. ML models' hyperparameter settings

This section provides a brief overview of the ML models used to predict ESP parameters. For ANN regression, the output layer was tailored to generate continuous values as predictions for the parameters of interest. The ANN architecture includes an input layer with dimensions twice the number of features, two hidden layers with the same dimensions and ReLU activation functions, and an output layer with a single node and a linear activation function. The model is compiled with the Adam optimizer. Key hyperparameters include 300 epochs, a batch size of 10, and an early stopping criterion with a patience of 10 epochs, monitoring the loss metric and restoring the best weights. SVR models were constructed utilizing a linear kernel function. The key hyperparameter tuned was the regularization parameter, which controls the trade-off between achieving a low error on the training data and maintaining generalization to unseen data. The SVR model was evaluated with three different regularization parameters, 0.1, 1, and 5 and the optimum result was found for the regularization parameter of 1 (based on MSE). This configuration was subsequently used to evaluate the model on the test set.

Furthermore, XGBoost model the primary hyperparameters tuned for this model were the number of estimators [50,100, 200], and the learning rate [0.01, 0.1, 0.2]. The tuning was carried out using 5-fold cross-validation, with performance evaluated based on the mean squared error (MSE) on the validation set. The final hyperparameter values were comprised ensembles of 100 trees with a learning rate of 0.1. The early stopping rounds were set to 10 without performing any sensitivity. For the RF model, no further hyperparameter study was performed and the model was selected with ensembles comprising of 100 trees that were constructed with the random state parameter set to 3 to ensure reproducibility.

As explained, Mean Squared Error (MSE) is used as the loss function and in the final step R^2, MSE and RMSE were used as a comparison for the performance of the models. Prior to selecting the final hyperparameters, k-fold cross-validation with 5 folds was applied to the

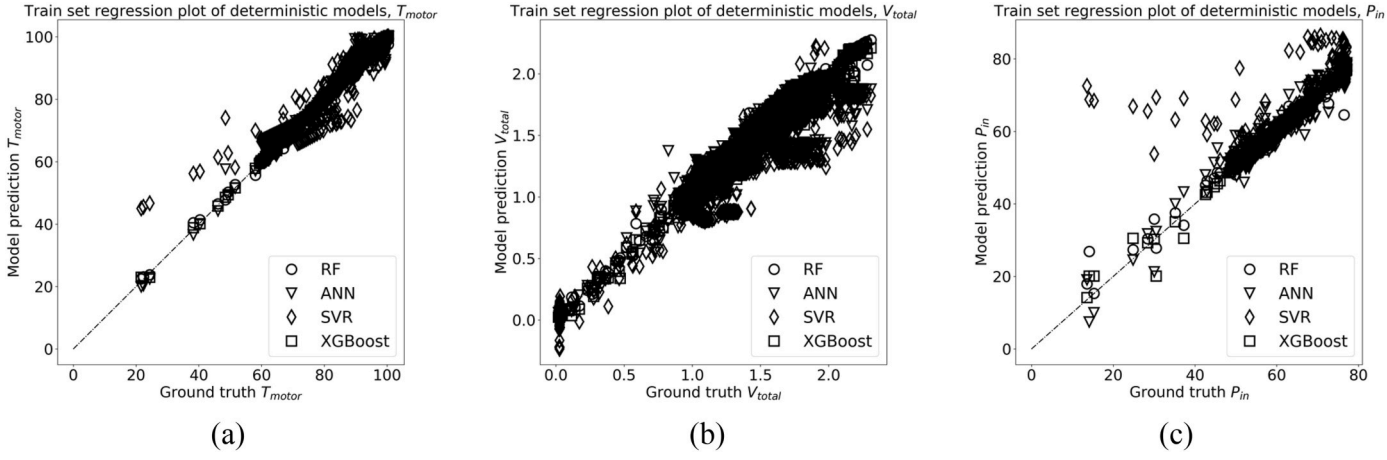


Fig. 4. Regression plots of the actual versus predicted values of the different trained ML models (denoted in the legend) for (a) ESP motor temperature (b)total vibration (c) pump intake pressure.

models. All ML models are implemented within a Scikit-learn pipeline in Python (Pedregosa et al., 2011).

The focus of the manuscript is not on selecting the best-performing ML model, but rather on demonstrating the application of ensemble ML models and statistical metrics for condition monitoring. These specified architectures and hyperparameters were carefully selected and tested to ensure robust and effective model performance within the context of the study's objectives. A broad selection of ML models was necessary to thoroughly understand the problem space. Firstly, it was important to conduct a comprehensive evaluation that covers a range of model complexities and approaches. Secondly, comparing the performance of different models provides valuable insights into the impact of data size on model effectiveness. The results derived from these models are indicative given the selected architecture and hyperparameters and no generic conclusion can be made on the performance of each model from this study.

4. Results and discussion

4.1. ML model training results

This section summarizes the results from Step 1, as proposed in the framework. After several changes in the size of the training set and the sampling size for the training and validation sets, the training set (the first year of data) was split into 60% for the training and 40% for validation. The sampling was done randomly and no significant change in the accuracy of the model was found for a smaller training set. The results of the prediction values versus actual values in the validation set for four different ML models of three target variables are shown in Fig. 4. The regression analysis for T_{motor} indicates that most of the models yielded precise predictions, except for some deviations observed at lower T_{motor} values in the predictions made by the SVR model. Similarly, the prediction of V_{total} was largely accurate, although both the ANN and SVR models demonstrated a tendency to underpredict at higher V_{total} values. The intake pressure (P_{in}) of the ESP was predicted with high accuracy by nearly all models, with the SVR model exhibiting the lowest prediction accuracy for P_{in}.

A summary of the accuracy indicators is provided in Fig. 5. The results showed that RF and XGBoost exhibited superior predictive performance compared to SVR. Specifically, RF and XGBoost demonstrated lower MSE and RMSE values and higher coefficient of determination (R^2) values indicating better accuracy in predicting the target variable. In terms of accuracy, motor temperature and intake pressure were predicted accurately almost by all four models. In addition, it was shown that while ANN did not outperform other models, its inclusion in the analysis highlights the data-dependency of model performance. SVR showed the lowest performance in the current dataset, likely due to its sensitivity to the limited dataset size and potentially difficulty in capturing the underlying patterns with the available data and using a relatively high number of features. Based on the derived results, it can be concluded that using RF and XGBoost will lead to a monitoring accuracy of more than 95% for all the quantities of interest during the normal operation of the system.

4.2. Results of ML models for condition monitoring

To assess the condition monitoring framework, after the ML model training is performed, the real-time data is introduced to the ML model to estimate the quantities of interest (Step 2). In this study, the second year of data is treated as real-time data. The output of the individual ML models and ensemble ML models are used to calculate the metrics in Step 3. In the subsections below, first the outcome of metrics from the individual ML models are described and afterwards the metrics calculated from the ensemble ML models.

4.3. Condition monitoring with individual ML models

The output of each individual ML model for all the quantities of interest in the second year of data (test data) is calculated. Based on the case study data, while the exact point of failure was identified, the initiation of the degradation leading to this failure remained unknown. As the first step, the change in individual ML models' metrics between

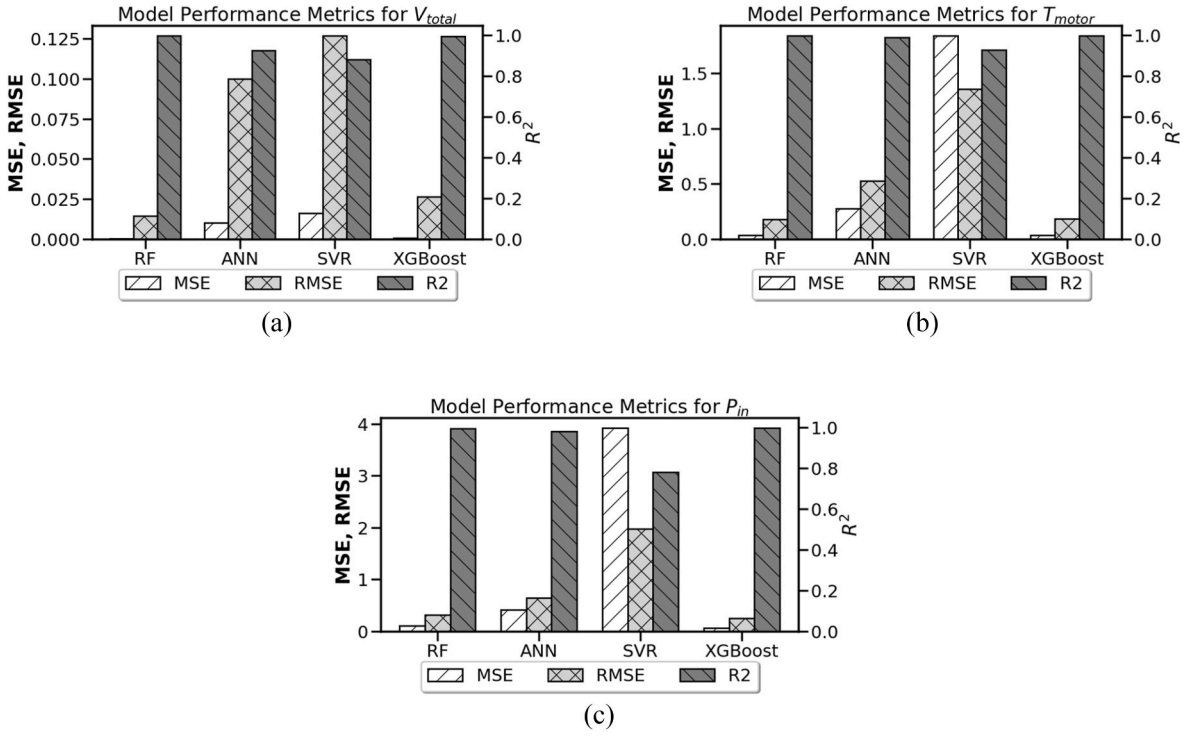


Fig. 5. Metrics, including MSE, RMSE and R^2, for different ML models (random forest, artificial neural networks, support vector regression, XGboost) predicting (a) total vibration, (b) Motor temperature, (c) intake pressure.

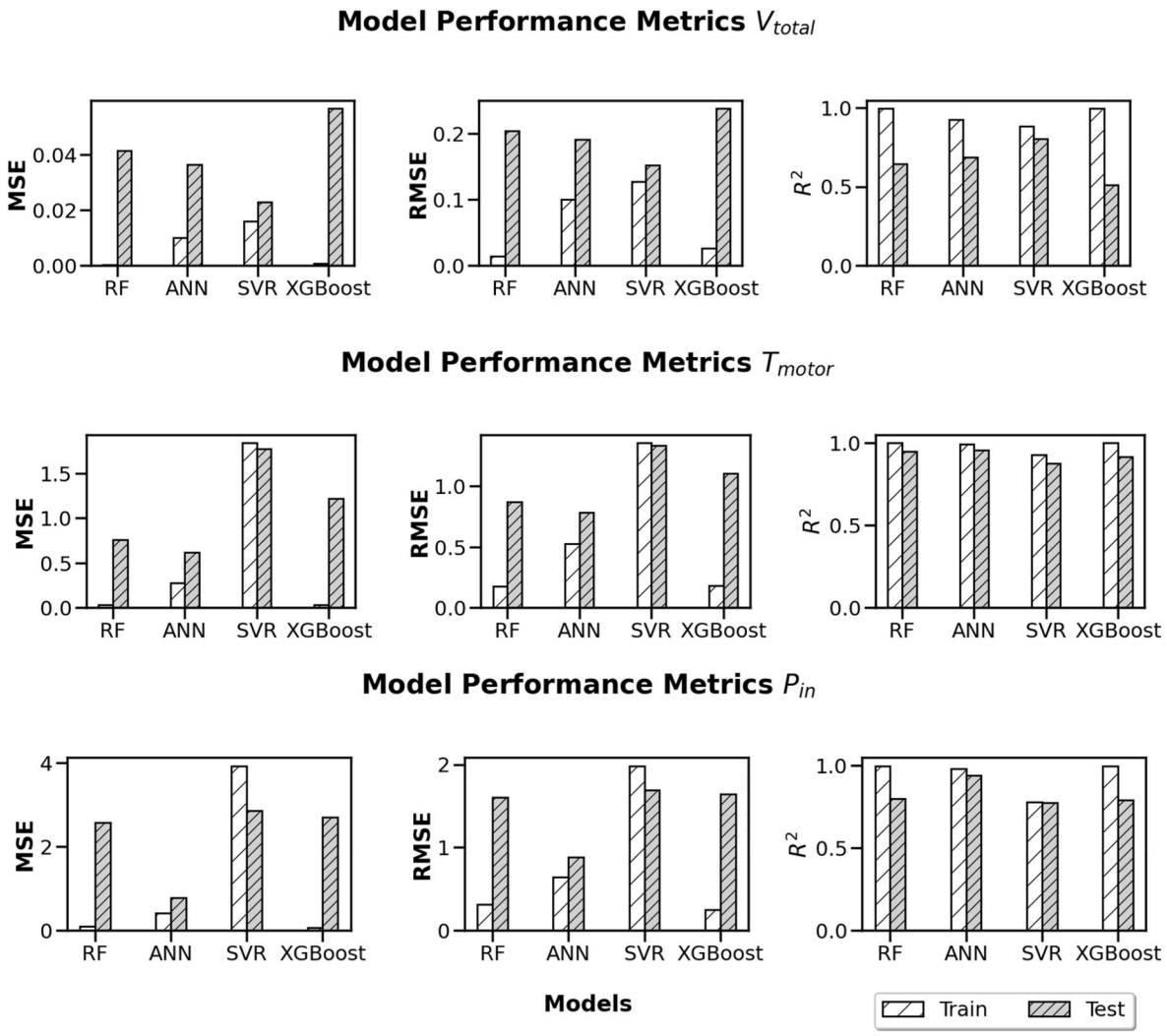


Fig. 6. Model performance metrics for all the ML models compared between the training and test set for total vibration, motor temperature and intake pressure.

the training and test sets was compared. The results of Step 3, illustrated in Fig. 6, show the three metrics (*MSE, RMSE, R^2*) for all three quantities of interest (V_{total}, T_{motor}, P_{in}) and all the four ML models trained. Based on the initial observation on the results, it is indicated that MSE and RMSE for the test set compared to the training set is increasing. Additionally, the test set exhibited a lower R^2 value for the test set. This observation is consistent for all the models, irrespective of the model accuracy on the training set, and in all three quantities of interests. These metrics have also been studied in other applications for the condition monitoring and similar trends were observed (Surucu et al., 2023).

Such a change in the metrics for individual ML models, can be caused either by the change in the data structure due to the degradation of the equipment or lack of generalization of the model. Since analyzing the distribution of the dataset, showed that the train and test sets do have an overlap in distribution and range (Fig. 3), hence data structure was not at the base of this mismatch. As mentioned in the results of ML models training, the accuracy of the ML model for different training set was not significantly changed that hints towards the model's robustness and consistent performance across various data subsets (Freiesleben et al. 2023). This stability suggests that the models have generalization capabilities and the underlying patterns in the data are likely caused by changes and degradation in the production.

4.4. Condition monitoring with ensemble ML models

In this section, we present the results of the ensemble ML models along with the statistical metrics derived from the analysis (Step 3). Analysis was performed on all three parameters described in the previous section but only shown in detail for ESP intake pressure, illustrating how the framework detects performance degradation over time. As previously indicated, quantile regression forest method was used for the ensemble ML models.

The results of the condition monitoring under uncertainty framework are illustrated in Fig. 7. The top graph shows field data and predictions for the train and test sets over a span of nearly two years (including the training and test sets). The field data is represented by blue lines, with the training set indicated by solid lines and the test set by dashed lines. The predictions from the averaged of ensemble model predictions are depicted in red dashed lines, and the confidence intervals at 50% and 95% levels are shown in green and blue shaded areas, respectively. The close alignment between the predicted values and actual field data, particularly within the confidence intervals, demonstrates the model's accuracy and robustness in capturing the performance of the ESP.

The middle and bottom graphs focus on the test set data for the first and last months of the observation period, respectively. These graphs provide a more granular view of the model's performance over shorter timeframes. The middle graph, representing the first month of the test

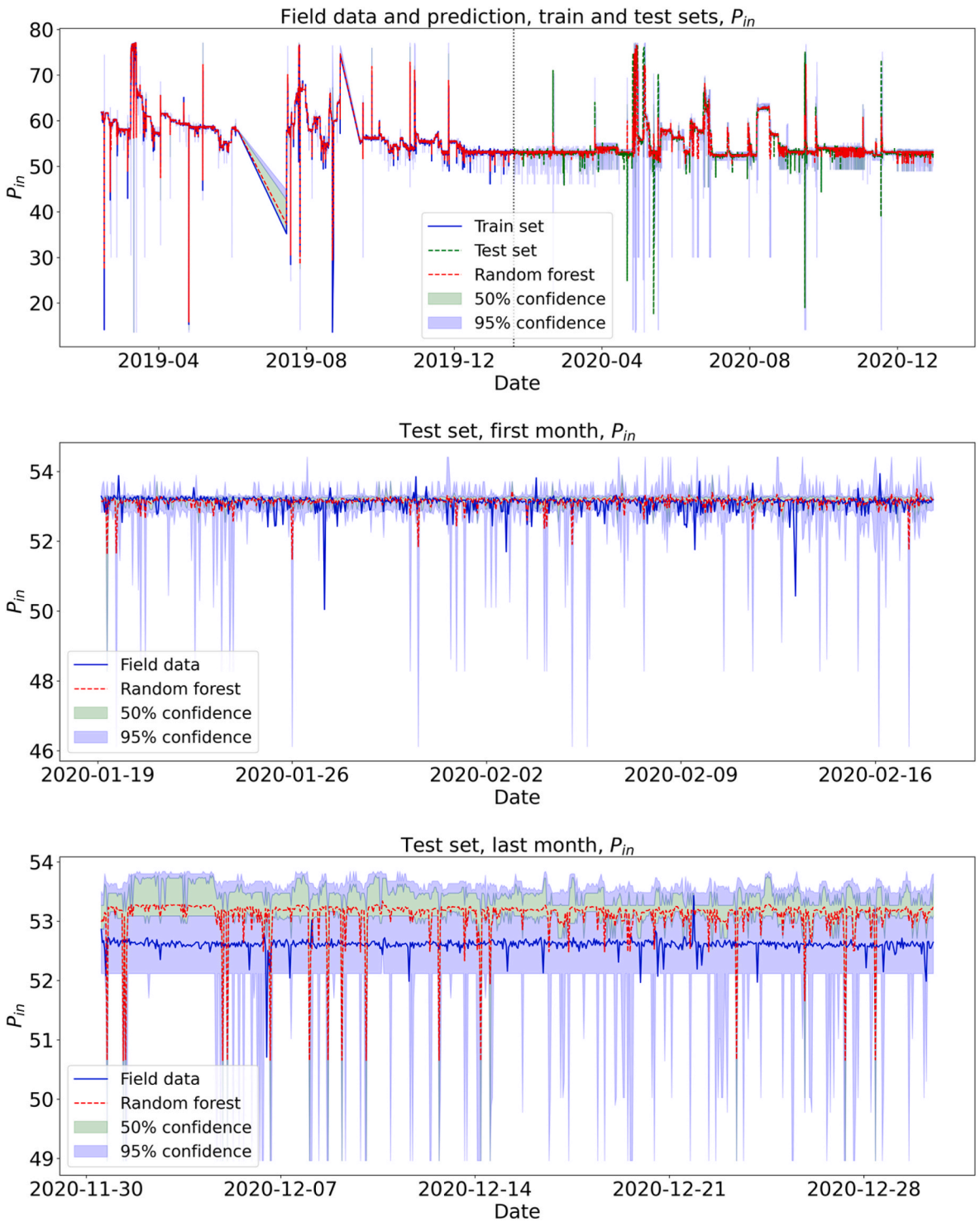


Fig. 7. Condition monitoring of ESP using a quantile forest regression model under uncertainty. The top panel shows field data and model predictions for both the training and test sets over nearly two years, with confidence intervals at 50% and 95%. The middle panel focuses on the first month of the test set, highlighting the model's performance and accuracy. The bottom panel shows the last month of the test set, indicating potential degradation in ESP performance as evidenced by increased deviations between field data and predictions for the ESP intake pressure.

set, shows a tight clustering of field data around the predicted values, with most data points falling within the 95% confidence interval. This indicates a strong initial performance of the ESP. The bottom graph, representing the last month of the test set, reveals a slight increase in deviations between the field data and predictions, with more data points outside the confidence intervals, especially almost all the data falls outside the 50% confidence bound. This divergence suggests potential degradation in the production and ESP's performance over time. In addition, by observing the bounds of the 50% and 95% confidence interval, a shift in the distribution of the data within the bounds is observed which suggests the mismatch between the mean and median of the prediction, which will be quantified in the next part. The size of the 50% and 95% confidence bound also increased in the last month of monitoring prior to the failure.

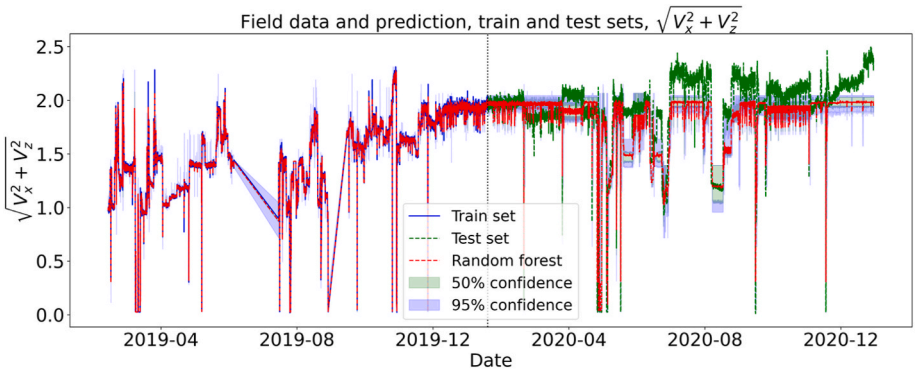


Fig. 8. Ensemble ML models used for condition monitoring of the total vibration demonstrated in both train and test sets.

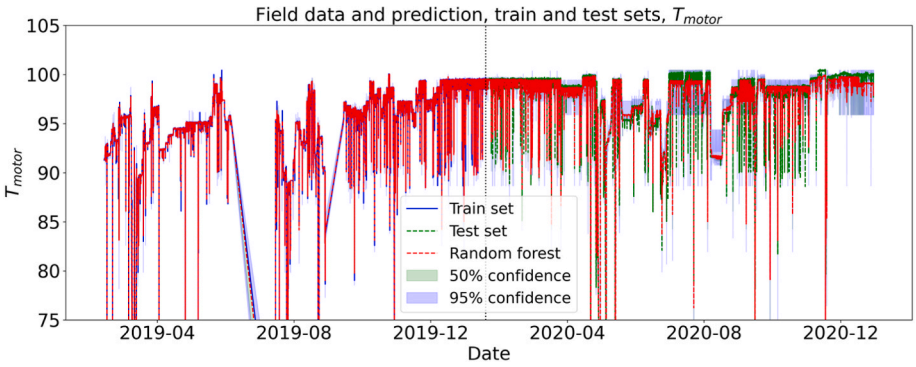


Fig. 9. Ensemble ML models used for condition monitoring of the ESP motor temperature demonstrated in both train and test sets.

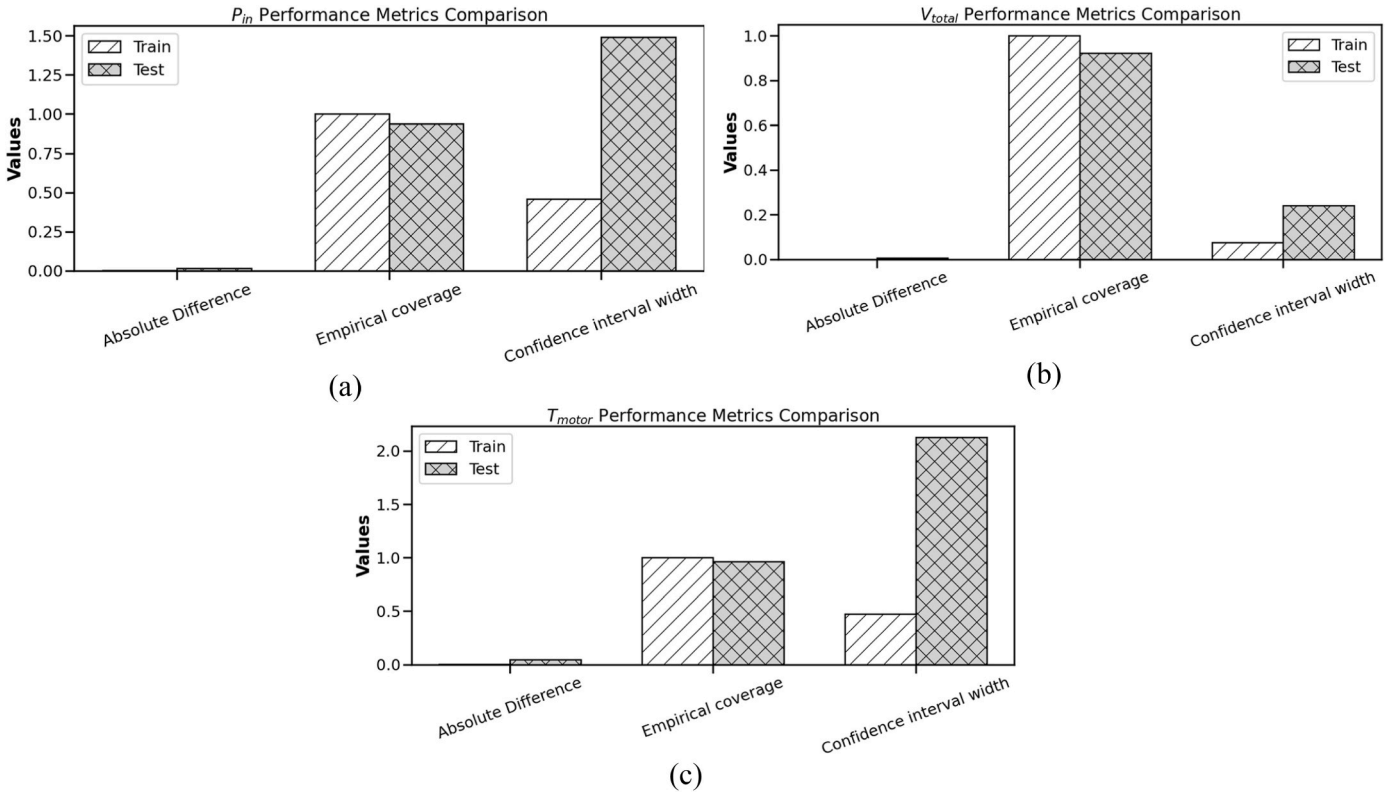


Fig. 10. Comparison of performance metrics (Absolute Difference, Empirical Coverage, and Confidence Interval Width) for training and testing datasets for three different parameters (a)P_{in}, (b) V_{total} and (c)T_{motor}.

The prediction results for the other two parameters, total vibration, and motor temperature, are shown in Figs. 8 and 9, respectively. For the total vibration, a more noticeable discrepancy between the predicted values and the confidence bounds is observed, with data consistently falling outside the bounds approximately six months before the failure. Additionally, there is a significant increase in total vibration one month prior to the failure, which the model, including both 50% and 95% confidence bounds, failed to capture. In contrast, the motor temperature parameter exhibits less pronounced and visible changes, with a slightly larger 95% confidence interval bound width closer to the failure point. There are more fluctuations in the motor temperature field data, but these remain within the confidence bound range, as observed around August and September 2020. However, the motor temperature data fall outside the confidence bounds a few days before the event.

Overall, these detailed visualizations highlight the effectiveness of the proposed framework in real-time monitoring. To make a step from qualitative to quantitative analysis, the statistical metrics which were previously defined are calculated for all three parameters and are shown in Fig. 10, both for the train and test sets. A systematic observation is that a clear difference between all three indicators on the train and test sets can be seen, Absolute Difference between mean and median increases, empirical coverage decreases and confidence interval width increases.

Despite the difference in Absolute Difference between the two

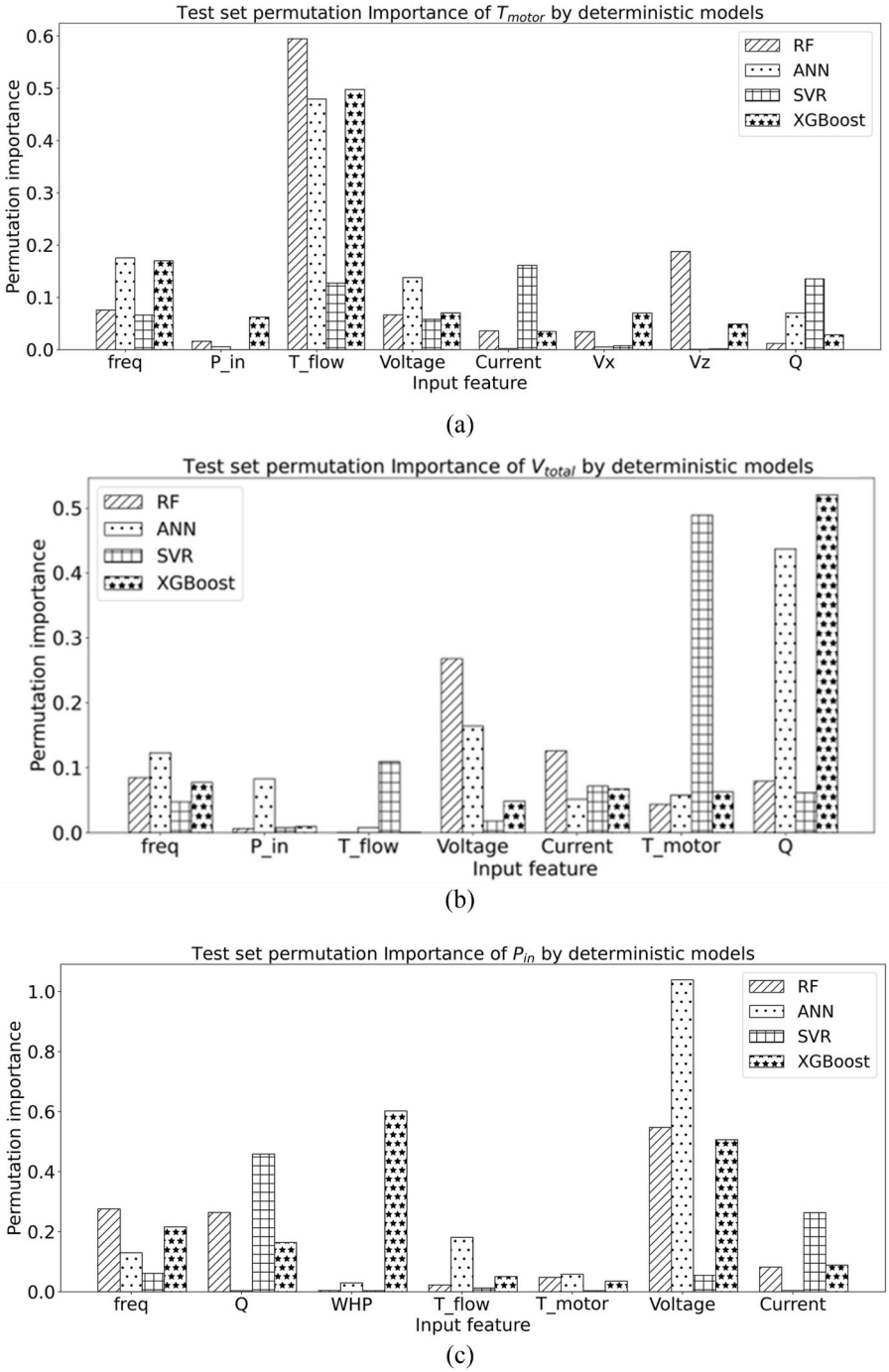


Fig. 11. Permutation importance of input features for predicting (a) T_{motor}, (b)V_{total} and (c)P_{in} across four ML models, RF, ANN, XGBoost and SVR. The graph illustrates the varying significance of each feature in contributing to the models' prediction.

datasets, the variation is marginal, and only using this metric can be unreliable for detecting performance degradation. The difference in empirical coverage and confidence interval width of train and test sets provide more noticeable changes in the datasets. In addition, Empirical Coverage shows moderate values for both training and testing, suggesting that the model captures a proportion of the true parameter values within the predicted confidence intervals. The Confidence Interval Width is significantly larger for the testing set compared to the training set, indicating greater uncertainty in the model's predictions when applied to unseen data. For real-time applications in the field, it is suggested to estimate these metrics in real-time on a moving window, the size of the window to be determined by the operator, to ensure an accurate detection of degradation onset.

4.5. Model explainability

To provide explainability on the derived results (Step 4), a permutation importance of various input features in predicting the three output parameters was performed. The feature importance provides insights into which input features significantly contribute to the prediction performed by each model. This step was performed on the test dataset. The outcome is shown in Fig. 11 for T_{motor}, V_{total} and P_{in}. From the graph, it is evident that the feature importance varies notably across different models. By considering T_{motor} as an example, and looking into the RF and XGBoost explainability, T_{flow} exhibits the highest importance, followed by frequency and voltage. This indicates that the RF model relies heavily on these features to predict T_{motor} in the test set. In contrast, the ANN model assigns the greatest importance to T_{flow} and voltage, highlighting a different feature interaction pattern compared to the RF model. The SVR model, which performed the worst in terms of predictive accuracy, shows a more evenly distributed importance across several features. This distribution suggests that SVR might struggle to leverage specific features effectively due to the limited dataset size, inherent complexity in the data or lack of convergence.

When analyzing explainability results, both the ranking of the features and their influence on the prediction parameters, as indicated by the amplitude, must be taken into account. If the feature importance ranking remains unchanged but their amplitudes changes, it indicates that the equipment is following the same underlying correlation, but the influence of parameters has been impacted. This might reflect a change in operational conditions or component performance. In addition, a change in the feature importance ranking signals a shift in the equipment's behavior, possibly due to altered system dynamics or emerging issues, highlighting the need for further investigation. By analyzing the explainability using the most accurate ML models, the following explainability for the degradation monitoring can be derived.

- Total vibration exhibited the most prominent mismatch, started around 6 months prior to the failure and the features that supports the prediction of this mismatch are flow rate, voltage, and current. Based on the operation of an ESP, it is well known that the changes in the flow rate, voltage and current can impact the total vibration.
- Motor temperature mismatch seems to be mainly correlated to the voltage, flow temperature, and pump frequency. Such a correlation can also be found in during the normal operation of the ESP. Motor temperature is significantly correlated with the flow temperature and an increase in the frequency or voltage can lead to a higher motor temperature due to a higher power consumption.
- ESP intake pressure degradation is captured through the impact of voltage and wellhead pressure followed by flow rate and ESP frequency.

In order to better understand the impact of features on three monitored parameters, a comparison was made between the feature importance calculated based on the training and test set (a month prior to the failure). This comparison for RF, as one of the most accurate model in this case study, is shown in Fig. 12. For ESP intake pressure and motor temperature, the ranking of features remains stable across both datasets,

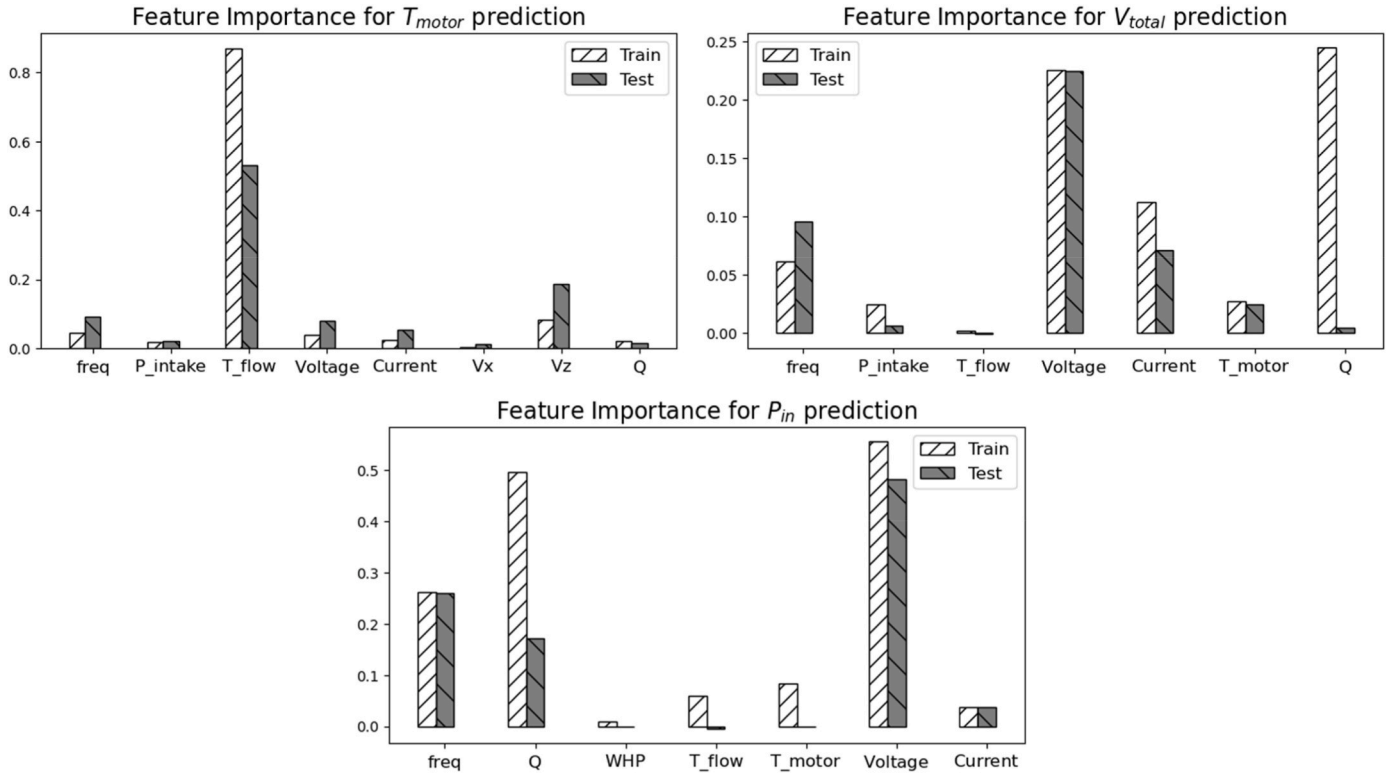


Fig. 12. Comparison of feature importance between the training set and a portion of test data (last month prior to the failure) of the trained random forest model to identify the features contribution to the mismatch detected in motor temperature, total vibration and intake pressure of ESP.

but the amplitude of feature permutation varies. This suggests that while the relative importance of each feature stays the same, the magnitude of their impact on model predictions differs between the training and test sets. On the other hand, for total vibration, although voltage, current, and frequency continue to be important features, the correlation between flow rate and total vibration significantly weakens in the test set. This shift indicates that the model's ability to capture the flow rate's impact on vibration diminishes over time, which may be due to changes in the system dynamics or the model's reduced ability to generalize the relationship between these features as the failure approaches. The explainability observed demonstrates that the relationship between inputs and outputs captured by the accurate ML models is logical and remains valid during both normal and abnormal operations. While some correlation is expected, excessive or unexpected changes in the monitoring parameters with the features could signal potential issues. Such deviations may indicate system degradation or the early onset of a failure and is provided from the metrics calculated in Step 3.

4.6. Anomaly detection with expert knowledge

In the final step of the framework, the expert needs to decide about the status of the system. As explained in the Step 5 of the framework, the detection of anomalies and degradation is done by the operator of GFE based on in-house operational practices or handbooks to setup the threshold and is dependent on the criticality of the equipment. In the case study, all the metrics and model mismatches provide insights on the initiation of the degradation prior to the failure which can vary between 1 and 6 months and with different level of severity. Based on the visualization of ML models, the initiation of the ESP degradation was identified between 1 month, for ESP intake pressure, up to 6 months, for the total vibration, prior to the failure of the pump. In case, an accurate model for some of the monitoring values cannot be made, other parameters can be used to detect anomalies and degradation in the equipment.

Along with the metrics, visualization, and explainability, it was found that in the test data containing ESP degradation, the ESP intake pressure was higher than the field data, the predicted vibration was lower than the actual field value, and the field motor temperature was higher than predicted values. By considering the changes predicted by the models and comparing with known trend analysis for ESP degradation (Awaid et al., 2014), the analysis suggested a potential blockage at the inflow and perforation, due to a higher vibration and motor temperature while the intake pressure was lower than anticipated. In a similar example shown in the reference paper by Awaid et al. (2014), an ESP in a well showed trends of declining intake pressure and higher motor temperature which were associated to blockage at perforations. The observed behavior in the case study was consulted with the operator and the analysis was confirmed that the ESP degradation was mainly caused by inflow restriction.

In conclusion, the proposed framework showed promising result to serve as a real-time tool for surveillance of the ESP and geothermal production systems. In case a degradation is observed, the geothermal operator or engineers could receive an alarm to evaluate the condition and take corrective actions to resolve the situation. In addition, when a failure or trip occurred, the additional analysis provided by the framework could support the engineers to perform a root-cause failure analysis and improve the understanding of the situation.

In real-time conditions, the uncertainty intervals should be interpreted as a range of potential outcomes, informed by the diversity in predictions from the ensemble models. These intervals represent epistemic uncertainty due to variations in model performance, helping to provide a more nuanced understanding of the prediction's reliability. Managing uncertainty in model predictions has significant practical contributions in real-time decision-making. By incorporating uncertainty intervals, decision-makers can assess the confidence in predictions and make more informed choices in dynamic environments. This is especially important in complex systems where decisions must account for potential variability and model limitations. Uncertainty management improves the robustness of decisions and reduces the risk of overconfidence in model outputs.

5. Conclusions

In this paper we introduce a novel methodology in the realm of monitoring and anomaly detection by leveraging an ensemble of ML models, quantifying the model confidence, integrating with explainability and expert knowledge. Unlike existing frameworks that typically rely on a single model and metric for event detection, the ensemble approach combines the strengths of multiple models, such as RF, ANN, SVR, and XGBoost, to generate more robust and accurate predictions. This strategy mitigates the limitations of individual models and enhances predictive performance.

Ensemble ML models were also integrated into the framework to quantify the uncertainties and confidence bound in the predictive models for an enhanced detection of events in the data in an objective manner. This framework is crucial for risk management and improved operational planning, as it enables the identification of anomalies with a quantifiable degree of confidence. Furthermore, the incorporation of model explainability adds a layer of interpretability, allowing for a deeper understanding of the factors contributing to predictions. By focusing on both prediction accuracy and model transparency, this approach facilitates better decision-making in anomaly detection scenarios.

Introducing several metrics for individual and ensemble ML models provides a comprehensive view on the monitoring of the geothermal systems based on the developed models as each metric has a different implication of the detected event, e.g. MSE and RMSE quantifies the average magnitude of prediction errors, confidence interval width indicates the precision of the model's predictions, and empirical coverage helps validating the model's prediction and its confidence bound coverage for the measured values. This approach goes beyond current practice, which often focuses solely on metrics like R^2 and RMSE, by incorporating additional indicators that provide deeper insights into model reliability and event implications.

For the case study, the framework could predict the onset of the ESP degradation from 1 month up to 6 months prior to the occurrence of the failure with more than 95% confidence and the explainability layer together with the predicted trends provided potential cause of the failure which was due to the inflow restriction and production decline into the geothermal well. Compared to existing monitoring approaches, this methodology is more comprehensive due to its combined use of ensemble modeling, detailed explainability, and robust uncertainty quantification, offering an advanced solution for real-time anomaly detection. The proposed framework for real-time monitoring, combining ensemble ML models and uncertainty metrics, is inherently generic and can be adapted to any geothermal plant with adequate training data and the case study presented in the paper serves solely as a demonstration example. The results of this research can support the operators and engineers of geothermal plants to maintain reliable an efficient operation of the geothermal systems, as well as equipment manufacturers and service providers to improve the maintenance and inspection of these systems.

To ensure practical usability, the framework should supports informative visualizations such as model predictions with uncertainty bounds and feature importance trends over time. These tools highlight the key parameters influencing predictions and their temporal changes, as illustrated in the manuscript, and needs to be implemented for the operators to enable real-time decisions.

While the proposed framework shows promise in using ensemble models for monitoring geothermal systems, its conclusions are limited by the choice of employed ML models, selected indicators, and expert-provided inputs. The system's performance is heavily influenced by the selection of machine learning models, which must be carefully tailored to the specific geothermal systems, processes, or equipment based on their unique operational conditions. For new systems with

limited data or vastly different operating conditions, the framework's applicability may be constrained, highlighting the need for further research. Additionally, reliance on sensor data and statistical metrics can limit the framework's ability to fully capture the complexity of real-world scenarios, particularly under extreme conditions or significant operational variability. Issues such as missing or noisy data further impact the framework's effectiveness, emphasizing the importance of robust preprocessing techniques.

To address variability in operating conditions, methods like transfer learning could be explored to adapt existing ensemble models to new environments. These limitations should be carefully considered when interpreting the results and assessing the framework's broader applicability.

For future works, the framework is suggested to be tested on a larger well database to estimate remaining useful lifetime of ESP or other vital equipment in the geothermal plant, providing therefore predictive capability to the monitoring tool. Further integration of the framework with the digital twin of geothermal assets can provide confidence in the monitoring and detection of anomalous production behavior. While this study demonstrates the applicability of the proposed framework through a single case study, future work will focus on validating its use across diverse geothermal systems, including different pump types in various geothermal plants, heat exchangers, and filter clogging scenarios. This will help assess its general applicability and further enhance its robustness. Benchmarking the proposed framework with other existing monitoring approaches from the literature in a geothermal site will be another future research topic. Finally, integrating measurement uncertainties in the workflow is foreseen as a next step.

CRediT authorship contribution statement

Pejman Shoeibi Omrani: Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Yifan Yang:** Writing – original draft, Visualization, Validation, Methodology. **Huub H.M. Rijnaarts:** Writing – review & editing, Validation, Supervision. **Shahab Shariat Torbaghan:** Writing – review & editing, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the research partners for constructive discussions during the project meetings and supplying data for performing the study.

Nomenclature

ANN	Artificial neural networks
CNN	Convolutional Neural Networks
ESP	Electrical Submersible Pumps
GFE	Geothermal facilities and equipment
LSTM	Long Short-Term Memory
MAD	Median Absolute Deviation
ML	Machine Learning
MSE	Mean squared error
O&M	Operation and maintenance
RF	Random Forest
RMSE	Root mean square error
SME	Subject Matter Expert
SVM	Support vector machine
SVR	Support vector regression
UQ	Uncertainty quantification
XGBoost	Extreme Gradient Boosting

Data availability

Data will be made available on request.

References

Abdalla, R., Samara, H., Perozo, N., Carvajal, C.P., Jaeger, P., 2022. Machine learning approach for predictive maintenance of the electrical submersible pumps (ESPs). ACS Omega 7 (21), 17641–17651.

Abdurakipov, S., 2021. Increasing the efficiency of electric submersible pumps by using big data processing and machine learning technologies. J. Phys. Conf. 2119. Article 012109.

Abrasaldo, P.M.B., Zarrouk, S.J., Kempa-Liehr, A.W., 2024. A systematic review of data analytics applications in above-ground geothermal energy operations. Renew. Sustain. Energy Rev. 189 (Part B), 113998. https://doi.org/10.1016/j.rser.2023.113998.

Adesanwo, M., Denney, T., Lazarus, S., Bello, O., 2016. Prescriptive-based decision support system for online real-time electrical submersible pump operations management. SPE Intelligent Energy International Conference and Exhibition. https://doi.org/10.2118/181013-MS.

Adesanwo, M., Bello, O., Olorode, O., Eremiokhale, O., Sanusi, S., Blankson, E., 2017. Advanced analytics for data-driven decision making in electrical submersible pump operations management. SPE Nigeria Annual International Conference and Exhibition. https://doi.org/10.2118/189119-MS.

Alamu, O.A., Pandya, D.A., Warner, O., Debacker, I., 2020. ESP data analytics: use of deep autoencoders for intelligent surveillance of electric submersible pumps. Paper Presented at the Offshore Technology Conference. https://doi.org/10.4043/30468-MS. Houston, Texas, USA.

Alatrach, Y., Mata, C., Shoeibi Omrani, P., Saputelli, L., Narayanan, R., Hamdan, M., 2020. Prediction of well production events using machine learning algorithms. Paper Presented at the Abu Dhabi International Petroleum Exhibition & Conference. Abu Dhabi, UAE.

Alhashem, M., Lastra, R., Ahmed, M., Ghouti, L., 2024. Evaluation of machine learning techniques for ESP diagnosis using a synthetic time series dataset. Paper Presented at the International Petroleum Technology Conference. https://doi.org/10.2523/IPTC-24210-MS. Dhahran, Saudi Arabia, February 2024.

Altmann, A., Toloşi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26 (10), 1340–1347.

Andrade Marin, A., Busaidy, S., Murad, M., Balushi, I., Riyami, A., Jahwari, S., Ghadani, A., Ferdiansyah, E., Shukaili, G., Amri, F., Kumar, N., Marin, E., Gala, R., Rai, R., Venkatesh, B., Bai, B., Kumar, A., Ang, E., Jacob, G., 2019. ESP well and component failure prediction in advance using engineered analytics - a breakthrough in minimizing unscheduled subsurface deferments. Paper Presented at the Abu Dhabi International Petroleum Exhibition & Conference. Abu Dhabi, UAE.

Awaid, A., Al-Muqbali, H., Al-Bimani, A., Yazeedi, Z., Al-Sukaity, H., Al-Harthy, K., Baillie, A., 2014. ESP well surveillance using pattern recognition analysis, oil wells, petroleum development Oman. Paper Presented at the International Petroleum Technology Conference. Doha, Qatar. https://doi.org/10.2523/IPTC-17413-MS.

Bhardwaj, A.S., Saraf, R., Nair, G.G., Vallabhaneni, S., 2019. Real-time monitoring and predictive failure identification for electrical submersible pumps. Abu Dhabi International Petroleum Exhibition & Conference.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly Detection: A Survey, vol. 41. ACM computing surveys (CSUR), pp. 1–58.

Chemura, A., Rwasoka, D., Mutanga, O., Dube, T., Mushore, T., 2020. The impact of land-use/land cover changes on water balance of the heterogeneous Buzi sub-catchment. Zimbabwe Remote Sens. Appl. Soc. Environ. 18. Article 100292.

Choi, K., Yi, J., Park, C., Yoon, S., 2021. Deep learning for anomaly detection in time-series data: review, analysis, and guidelines. IEEE Access 9, 120043–120065, 2021.

Costa, E.A., Rebello, C. de M., Santana, V.V., Reges, G., Silva, T. de O., Abreu, O. S. L. de, Ribeiro, M.P., Foresti, B.P., Fontana, M., Nogueira, I.B., dos, R., Schnitman, L., 2024. An uncertainty approach for Electric Submersible Pump modeling through deep neural network. Heliyon 10 (2), e24047. https://doi.org/10.1016/j.heliyon.2024.e24047.

Costa, E.A., de Abreu, O.S.L., Silva, T. de O., Ribeiro, M.P., Schnitman, L., 2021. A Bayesian approach to the dynamic modeling of ESP-lifted oil well systems: an experimental validation on an ESP prototype. J. Petrol. Sci. Eng. 205, 108880. https://doi.org/10.1016/j.petrol.2021.108880.

Der Kiureghian, A., Ditlevsen, O., 2009. Aleatory or epistemic? Does it matter? Struct. Saf. 31 (2), 105–112.

Dodge, Yadolah, 2010. The Concise Encyclopedia of Statistics. Springer, New York. ISBN 978-0-387-32833-1.

Don, M.G., Liyanarachchi, S., Wanasinghe, T.R., 2024. A digital twin development framework for an electrical submersible pump (ESP). Arch. Adv. Eng. Sci. 3, 1–10.

Dussi, S., Octaviano, R., Shoeibi Omrani, P., 2022. Bayesian networks applied to ESP performance monitoring and forecasting. In: SPE Annual Technical Conference and Exhibition. Houston.

Freiesleben, T., Grote, T., 2023. Beyond generalization: a theory of robustness in machine learning. Synthese 202, 109. https://doi.org/10.1007/s11229-023-04334-9.

Guo, D., Raghavendra, C.S., Yao, K.-T., Harding, M., Anvar, A., Patel, A., 2015. Data driven approach to failure prediction for electrical submersible pump systems. SPE Western Regional Meeting 2015. SPE-174062-MS.
Gupta, S., Nikolaou, M., Saputelli, L., Bravo, C., 2016. ESP health monitoring KPI: a real-time predictive analytics application. SPE Intelligent Energy International Conference and Exhibition.
Hamedi Shokrlu, Y., Bazile, J., 2024. Improving ESP production for unconventional wells through real-time machine learning based changepoint detection. Presented at SPE Western Regional Meeting. Palo Alto, California, US.
Hastie, T., Tibshirani, R., Friedman, J., 2008. The Elements of Statistical Learning, pp. 587–604.
Hochenbaum, J., Vallis, O., Kejariwal, A., 2017. Automatic anomaly detection in the cloud via statistical learning. ArXiv preprint arXiv:1704.07706.
Hoevenaars, A., McGraw, M., Burley, C., Bierhaus, E., 2021. Improving motor performance and runtime in ESP applications with novel sinewave filter. Paper Presented at the SPE Gulf Coast Section Electric Submersible Pumps Symposium. Virtual and The Woodlands, Texas, USA. https://doi.org/10.2118/204493-MS.
IEA, 2021. Geothermal Power. IEA, Paris. https://www.iea.org/reports/geothermal-power.
Iranzi, J., Wang, J., Lee, Y., et al., 2024. Evaluating the intake plugging effects on the electrical submersible pump (ESP) operating conditions using nodal analysis. J. Pet. Explor. Prod. Technol. 14, 1071–1083. https://doi.org/10.1007/s13202-024-01754-2.
Irl, M., Schifflechner, C., Wieland, C., Spliethoff, H., 2023. Advanced monitoring of geothermal organic rankine cycles. Renew. Energy 217, 119124. https://doi.org/10.1016/j.renene.2023.119124.
Jaber, A.A., 2016. Design of an Intelligent Embedded System for Condition Monitoring of an Industrial Robot. Springer.
Jansen Van Rensburg, N., Kamin, L., Davis, S., 2019. Using machine learning-based predictive models to enable preventative maintenance and prevent ESP downtime. Paper Presented at the Abu Dhabi International Petroleum Exhibition & Conference. Abu Dhabi, UAE.
Karnik, S., Yenuganti, N., Jusri, B.F., Gupta, S., Nirgudkar, P., Mohajer, M., Malik, A., 2021. Automated ESP failure root cause identification and analyses using machine learning and natural language processing technologies. Paper Presented at the SPE Gulf Coast Section Electric Submersible Pumps Symposium. Virtual and The Woodlands, Texas, USA. https://doi.org/10.2118/204519-MS. October 2021.
Kullick, J., Hackl, C.M., 2017. Dynamic modeling and simulation of deep geothermal electric submersible pumping systems. Energies 10 (10), 1659. https://doi.org/10.3390/en10101659.
Lastra, R., Jinjiang, X., 2022. Machine learning engine for real-time ESP failure detection and diagnostics. In: SPE Middle East Artificial Lift Conference and Exhibition.
Lastra, R., Tulbah, F., 2021. Computer vision for real-time ESP failure detection and diagnostics. In: World Petroleum Congress (WPC), Houston.
Lastra, R., 2019. Electrical submersible pump digital twin, the missing link for successful condition monitoring and failure prediction. In: Abu Dhabi International Petroleum Exhibition and Conference. Abu Dhabi, 2019.
Li, Z., van Leeuwen, M., 2023. Explainable contextual anomaly detection using quantile regression forests. Data Min. Knowl. Discov. 37, 2517–2563.
Li, J.J., Zhang, G.S., Song, S.Q., Yu, Y.Y., Duan, J., 2008. Development and application of macro-control diagram on oil production with electric submersible pump. Fau. Bl. O. & G. 15 (6), 121–122.
Loh, K., Shoeibi Omrani, P., van der Linden, R., 2018. Deep Learning History Matching for Real-Time Production Forecasting. European Association of Geoscientists & Engineers.
Lu, H., Ma, X., 2020. Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere 249. Article 126169.
Lund, J.W., Lienau, P.J., 2009. Geothermal district heating projects. International geothermal days, conference and summer school, Slovakia, 2009.
Meinshausen, N., 2006. Quantile regression forest. J. Mach. Learn. Res. 7, 983–999.
Mello, L.H.S., Oliveira-Santos, T., Varejão, F.M., Ribeiro, M.P., Rodrigues, A.L., 2022. Ensemble of metric learners for improving electrical submersible pump fault diagnosis. J. Petrol. Sci. Eng. 218, 110875.
Mohamad, T.S.N.F.T., Afrizal, N., Daud, M.Z., Awal, M.R., 2022. Review on advance monitoring of electrical and mechanical failure in submersible pump. In: 2022 4th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE). Kuala Lumpur, Malaysia, pp. 1–8.
Molęda, M., Małysiak-Mrozek, B., Ding, W., Sunderam, V., Mrozek, D., 2023. From corrective to predictive maintenance-A review of maintenance approaches for the power industry. Sensors (Basel) 23 (13), 5970. https://doi.org/10.3390/s23135970. Jun 27.
Nanavaty, R., 2024. Exploring autoencoders and XGBoost for predictive maintenance in geothermal power plants. Proceedings of the 49th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California, February 12-14, 2024 (SGP-TR-227).
Ocampo-Díaz, J.D.D., Valdez-Salaz, B., Shorr, M., Sauceda, M.I., Rosas-González, N., 2005. Review of corrosion and scaling problems in Cerro Prieto Geothermal Field over 31 years of commercial operations. In: Proceedings of the World Geothermal Congress 2005, vol. 1. Antalya, Turkey, 24-29 April 2005.
Octaviano, R., Dussi, S., de Zwart, H., Shoeibi Omrani, P., van Pul-Verboom, V., Elewaut, K., van Schravendijk, B., 2022. Model-based monitoring of geothermal assets. WarmingUP program final report.
Octaviano, R., Hornstra, E., Poort, J., Shoeibi Omrani, P., van Linden, R., Slot, H., Dechelette, B., 2020. Improving well production performance by using realtime wash desalting planning tool WDPT in the North Sea. Paper Presented at the Abu Dhabi International Petroleum Exhibition & Conference. Abu Dhabi, UAE.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Peng, L., Han, G., Pagou, A.L., Zhu, L., Ma, H., Wu, J., Chai, X., 2021. A predictive model to detect the impending electric submersible pump trips and failures. SPE Annual Technical Conference and Exhibition.
Poort, J., Shoeibi Omrani, P., Vecchia, A.L., Visser, G., Janzen, M., Koenes, J., 2020. An automated diagnostic analytics workflow for the detection of production events-application to mature gas fields. Abu Dhabi International Petroleum Exhibition and Conference, p. 2020.
Rauber, T.W., Oliveira-Santos, T., de Assis Boldt, F., Rodrigues, A., Varejão, F.M., Ribeiro, M.P., 2017. Kernel and random extreme learning machine applied to submersible motor pump fault diagnosis. 2017 International Joint Conference on Neural Networks (IJCNN), pp. 3347–3354.
Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Model-agnostic Interpretability of Machine Learning arXiv preprint arXiv:1606.05386.
Sagoolmuang, A., Sinapiromsaran, K., 2017. Median-difference window subseries score for contextual anomaly on time series. 2017 8th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES), pp. 1–6.
Schall, R., 2012. The empirical coverage of confidence intervals: point estimates and confidence intervals for confidence levels. Biom. J. 54 (4), 537–551.
Sharma, A., Songchitruksa, P., Sinha, R.R., 2022. Integrating domain knowledge with machine learning to optimize electrical submersible pump performance. Paper Presented at the SPE Canadian Energy Technology Conference. Calgary, Alberta, Canada.
Shoeibi Omrani, P., Dobrovolschi, I., Belfroid, S., Kronberger, P., Munoz, E., 2018. Improving the accuracy of virtual flow metering and back-allocation through machine learning. Paper Presented at the Abu Dhabi International Petroleum Exhibition & Conference. Abu Dhabi, UAE.
Shoeibi Omrani, P., Van der Valk, K., Bos, W., Nizamutdinov, E., Van der Sluijs, L., Eilers, J., Pereboom, H., Castelein, K., Van Bergen, F., 2021. Overview of opportunities and challenges of electrical submersible pumps ESP in the geothermal energy production systems. Paper Presented at the SPE Gulf Coast Section Electric Submersible Pumps Symposium. Virtual and The Woodlands, Texas, USA, October. https://doi.org/10.2118/204524-MS.
Siratovich, P.A., Blair, A., Weers, J., 2020. GOOML: geothermal operational optimization with machine learning. GRC Transactions 44 (2020).
Sherif, S., Adenike, O., Obehi, E., Funso, A., Eyituoyo, B., 2019. Predictive data analytics for effective electric submersible pump management. SPE Nigeria Annual International Conference and Exhibition.
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14, 199–222.
Soh, J., DaeEun, K., 2021. Condition monitoring with time series data based on probabilistic model. 24th International Conference on Electrical Machines and Systems (ICEMS) Oct 31-Nov 3, 2021. Hybrid, Korea.
Surucu, O., Gadsden, S.A., Yawney, J., 2023. Condition monitoring using machine learning: a review of theory, applications, and recent advances. Expert Syst. Appl. 221, 119738.
Tandazo, S., Hu, H., Corredor, F., 2022. Monitoring Real-Time Geothermal Data Leads to Innovative Machine Learning Improvements in ESP Operations and Diagnostics, European Association of Geoscientists & Engineers EAGE GET 2022. Nov. 2022.
Tao, F., Liu, G., Xi, W., 2011. Research on the fault diagnosis of excess shaft ran of electric submersible pump. Advances in Multimedia, Software Engineering and Computing 1 (128), 509–513, 2011.
Van't Spijker, H., Ungemach, P., 2016. Definition of Electrosubmersible Pump (ESP) Design and Selection Workflow, Public Final Report as a Part of Kennisagenda Aardwarmte Program.
van Gerven, M.A.J., Seeliger, K., Güçlü, U., Güçlütürk, Y., 2019. In: Samek, W., et al. (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer International, pp. 379–394.
Wasch, L., Creusen, R., Eichinger, F., Goldberg, T., Kjoller, C., Regenspurg, S., Mathiesen, T., Shoeibi Omrani, P., van Pul-Verboom, V., 2019. Improving geothermal system performance through collective knowledge building and technology development. In: European Geothermal Congress 2019.
Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., et al., 2017. mwaskom/seaborn: v0.8.1. Zenodo. https://doi.org/10.5281/zenodo.883859.
Xi, W.J., 2008. Research on Fault Diagnosis of Electric Submersible Pumps Based on Vibration Detection. China University of Petroleum (East China), Dongying, China. Master's Thesis.
Zhang, P., Chen, T., Wang, G., Peng, C., 2017. Ocean economy and fault diagnosis of electric submersible pump applied in floating platform. Int. J. e-Nav. Mari. Econ. 6, 37–43.
Zhao, X.J., Li, A., Yang, F., 2006. Fault analysis of electric submersible pump based on using FTA. Mod. Manuf. Technol. and Equip. 4, 29–32.
Zhao, P., 2011. Study on the Vibration Fault Diagnosis Method of Centrifugal Pump and System Implementation, vol. 2011. North China Electric Power University, Beijing, pp. 56–58.
Zulkarnain, I., Surjandari, R.R., Bramasta, R., Laoh, E., 2019. Fault detection system using machine learning on geothermal power plant. In: 2019 16th International Conference on Service Systems and Service Management (ICSSSM), pp. 1–5. https://doi.org/10.1109/ICSSSM.2019.8887710. Shenzhen, China.