Scientific
Research
Publishing

# Effective Electrical Submersible Pump Management Using Machine Learning

**Son Tung Pham\*, Phi Son Vo, Dac Nhat Nguyen**

Faculty of Geology & Petroleum Engineering, Hochiminh City University of Technology-Vietnam National University, Hochiminh, Vietnam
Email: *phamsontung@hcmut.edu.vn

## Abstract

Artificial lift plays an important role in petroleum industry to sustain production flowrate and to extend the lifespan of oil wells. One of the most popular artificial lift methods is Electric Submersible Pumps (ESP) because it can produce high flowrate even for wells with great depth. Although ESPs are designed to work under extreme conditions such as corrosion, high temperatures and high pressure, their lifespan is much shorter than expected. ESP failures lead to production loss and increase the cost of replacement, because the cost of intervention work for ESP is much higher than for other artificial lift methods, especially for offshore wells. Therefore, the prediction of ESP failures is highly valuable in oil production and contributes a lot to the design, construction and operation of oil wells. The contribution of this study is to use 3 machine learning algorithms, which are Decision Tree, Random Forest and Gradient Boosting Machine, to build predictive models for ESP lifespan while using both dynamic and static ESP parameters. The results of these models were compared to find out the most suitable model for the prediction of ESP life cycle. In addition, this study also evaluated the influence factor of various operating parameters to forecast the most impact parameters on the duration of ESP. The results of this study can provide a better understanding of ESP behavior so that early actions can be realized to prevent potential ESP failures.

## Keywords

Machine Learning, Electrical Submersible Pump, Decision Tree, Random Forest, Gradient Boosting Machine

## 1. Introduction

Artificial lifts are widely used in production wells to optimize production flo-

wrate [1]. The Electric Submersible Pump is a popular method applied to about 15 to 20 percent of one million wells worldwide thanks to its outstanding characteristics as it produces a high rate in depth well [2]. However, the failures of ESP usually occur abruptly and they are difficult to predict because of the complex nature of possible causes. The failures lead evidently to the production disruption and a large amount of money for replacement. Hence, the estimation of ESP lifespan is critical to have an early replacement plan in order to avoid the production loss. In parallel, being able to identify the key parameters which have the most impact on ESP failures can contribute to improving operating performance.

In recent years, oil and gas experts have tried to identify the main causes of ESP failures and to predict the life cycle of ESP by different methods such as using the harmonic patterns in the electric supply [3], or real-time ESP monitoring system [4], or considering ESPs failures sample analysis [5]. Guo, *et al.* (2015) built a Support Vector Machine model that used electrical and frequency data to detect anomalies in ESP during operation [6]. Gupta, *et al.* (2016) presented an analytical framework for early health monitoring of ESP based on data-driven modeling [7]. The framework can automatically identify real-time status and assess the health of ESP continuously, thus any detection of abnormal problem can be signaled to operators before it happens. Sherif, *et al.* (2019) [8] used Decision Tree method combined with Principal Component Analysis (PCA) to determine the stable region for ESP operation with parameters taken as intake pressure and temperature, vibrations, system current and frequency.

The use of PCA was also mentioned in the work of Abdelaziz, *et al.* (2017) to predict failure of ESP [9]. The input parameters were also discussed in the research of Popaleny, *et al.* (2018) where the authors presented how ESP mechanical and electrical malfunctions were reflected in the dynamic current spectrum using Motor Current Signature Analysis [10]. In brief, machine learning has been widely used in petroleum industry to predict ESP lifespan in recent years [11]. Various methods were proposed and used separately such as Decision Tree, Linear Regression, Random Forest [12].

This study presents a different way to approach this problem by building a predictive model using different machine learnings algorithms, which are Decision Tree, Random Forest and Gradient Boosting Machine, to predict ESP lifespan with both dynamic and static parameters. A total of 13 operating parameters were collected from 97 ESP. Furthermore, the model also classifies the impact of these parameters on the ESP lifespan. The results can be used to improve the ESP performance by appropriately adjusting the most influential parameters on ESP lifespan.

## 2. Methodology

Machine Learning (ML) is a subset of Artificial Intelligence (AI). The principle

of machine learning is data acquisition and self-learning machines. ML is a data analysis method that automates the building of an analytical model. Using iterative algorithms to learn from data, ML allows computers to find deeply hidden values that cannot be obtained by explicitly programming. The iterative aspect of ML is important because when these models are exposed to new data they can adapt independently. The model can learn from previous calculations to make repeatable and reliable decisions and results.

According to the learning method, Machine Learning algorithms are usually divided into 4 groups: Supervised learning, Unsupervised learning, Semi-supervised lerning and Reinforcement learning. Supervised learning is an algorithm that predicts the outcome of a new data (new input) based on known data (input, outcome).

All three algorithms used in this study, including Decision Tree, Random Forest and Gradient Boosting Machine, are Supervised Learning algorithms.

## 2.1. Decision Tree

Decision Tree is a structured hierarchy that can be used to classify objects based on a series of rules. When giving data about objects with attributes along with its classes, the decision tree will generate rules to predict the class of unknown objects (unseen data). Decision Trees consists of three main parts: a root node, leaf nodes and its branches. The root node is starting point of a decision tree and both the root node and the node contain questions or criteria to be answered. The branch represents the results of the test on the node. For example, the question on the first node asking the answer is "yes" or "no", then there will be one sub-node responsible for the response is "yes", the other node is "no". An example of a decision tree is illustrated in Figure 1. In this research the decision tree is built using Iterative Dichotomiser 3 (ID3) algorithm [13]. The workflow consists of the following steps:

Step 1: Select the best attribute a of the data set S using Information Gain (IG) and Entropy

with

$$\text{Information Gain} = \text{entropy}(\text{parent}) - [\text{weighted average}] * \text{entropy}(\text{children})$$

and

$$\text{Entropy} = \sum p(x) \log p(x)$$

where $p(x)$ is the proportion of the number of elements in class $x$ to the number of elements in set S.

Step 2: Partition the set S into subsets using the attribute for which the resulting entropy after splitting is minimized; or, equivalently, information gain is maximum.

Step 3: Make a decision tree node containing that attribute.

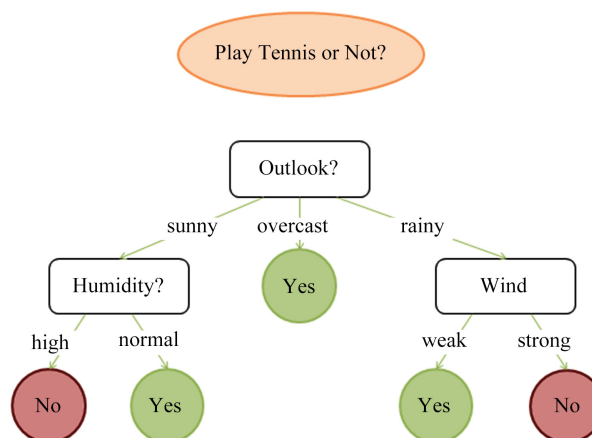Step 4: Recurse on subsets using the remaining attributes.

Figure 1. Example of a decision tree.

## 2.2. Random Forest

Random Forest can build a collection of decision trees and then use voting method to make decisions about the target variable. An example of Random Forest is as follows: suppose you want to go on a British tour and have intention to visit a city like Manchester, Liverpool or Birmingham. To make decision you will need to consult a lot of opinions from friends, travel blogs, tours … Each one corresponds to a decision tree that provides questions like: is the city beautiful, is it possible to visit the stadiums, how much does the visit cost, how long is the duration of the visit … Then you will have a forest of answers to decide which city to visit. The Random Forest evaluates and classifies decision trees using voting to deliver the final results.

Mathematically, the algorithm can be explained as follows: Random Forest is a collection of hundreds of decision trees, where each decision tree is randomly generated from sample re-selection (random selection) part of the data and randomize variables from all variables in the data (Figure 2). With such a mechanism, Random Forest provides an accurate result but the action mechanism of this algorithm cannot be seen due to the complex structure of this model, so this algorithm is one of the Black Box methods which allow us. This constitutes a tradeoff between explanatory power and predictive power.

## 2.3. Gradient Boosting Machine

The Gradient Boosting Machine is a synchronous technique that tries to create a strong model from a number of weak models. Instead of building a prediction model (such as a decision tree) with medium accuracy, we build various predictive models with a weaker accuracy (weak learner) if they work individually but with higher accuracy if they work together. This can be done by building a model from the training data, then creating a second model that tries to correct the error from the first model (Figure 3). Models are added until training data can be perfectly predicted.
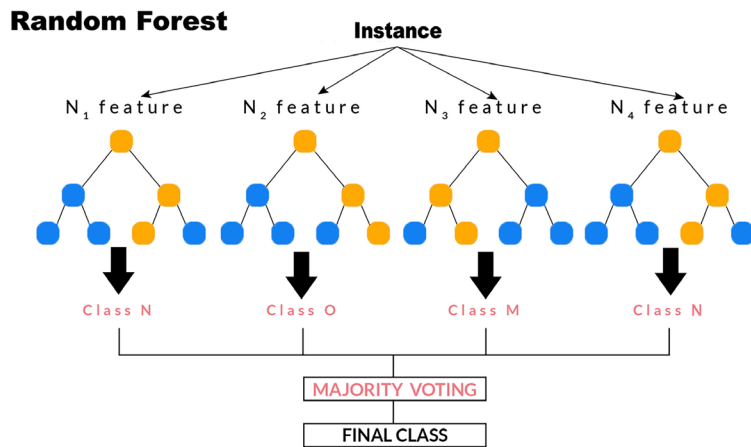
**Random Forest**



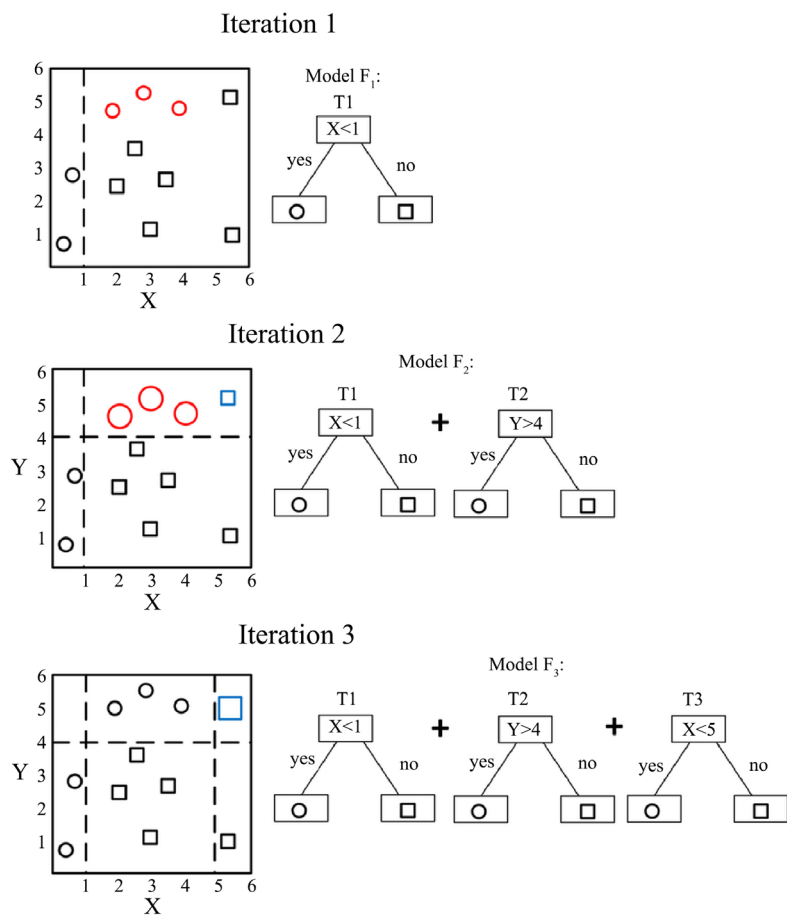Figure 2. Sketch of random forest algorithm.



Figure 3. Sketch of a gradient boosting machine.

We can imagine that each weak learner consists of weak, medium and excellent students, and a teacher. The weight of knowledge of the teacher will be the highest and the one of weak students will be the lowest. When you ask certain questions and need these people to draw conclusions, if many people have the same conclusions or if the weight of knowledge of those who make conclusions

is higher than the total group, then this conclusion may be right.

## 3. Modelling the Electrical Submersible Pump Lifespan

The goal of building a predictive model is to clarify the relationship between a group of input variables, the parameters that affect ESP lifespan, and a target variable, the ESP lifespan itself. The models were built using supervised learning methods: Decision Tree, Random Forest and Gradient Boosting Machine. The accuracy of each model will be evaluated based on the root-Mean Squared Error (RMSE). RMSE measures the difference between forecast value and actual value. In theory, a perfect model would have a RMSE value of 0 meaning absolute prediction, but in practice, the above assumption is not possible with the variable nature of the data. The best model would be the model with the lowest RMSE value.

The dataset used to build the forecast models was collected from 97 ESP. Dataset consists of input variables including static parameters (parameters of wells, fluid properties) and dynamic parameters (operating parameters) of ESP during operation, while the output is number of operating days (lifespan) of ESP (Table 1). 70% of the data is used for training, while the rest is used for testing. Splitting the dataset like that allows to avoid over-fitting phenomenon which can occur when the training result is too good but cannot be applied for a new dataset. During testing procedure, if the error is too large, it is necessary to repeat the workflow again until the error is acceptable. The step-by-step workflow to build predictive models is shown in Figure 4.

**Table 1.** Input parameters for building models.

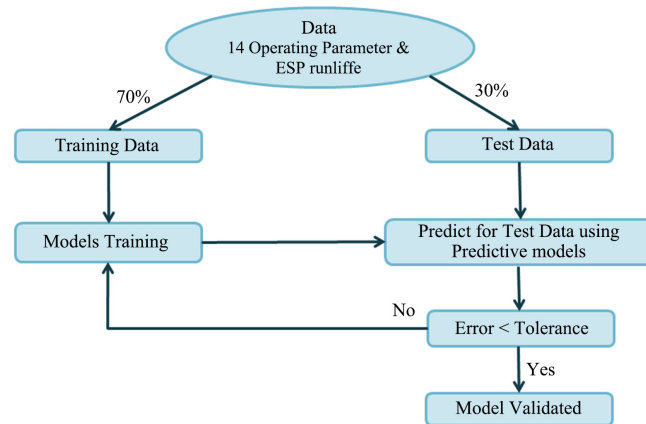| Number | Parameter | Interval of Values | Unit |
| --- | --- | --- | --- |
| 1 | Pump discharge pressure | 1500 - 3000 | psi |
| 2 | Pump intake pressure | 500 - 2500 | psi |
| 3 | Pump intake temperature | 45 - 200 | ˚C |
| 4 | Motor temperature | 70 - 200 | ˚C |
| 5 | Vibration x | 0.1 - 0.3 | inch |
| 6 | Vibration y | 0.05 - 0.2 | inch |
| 7 | Motor current | 30 - 120 | A |
| 8 | Motor speed | 45 - 70 | Hz |
| 9 | Wellhead temperature | 40 - 80 | ˚C |
| 10 | Wellhead pressure | 400 - 500 | psi |
| 11 | Production flow rate | 1000 - 5000 | bpd |
| 12 | Water cut | 20 - 45 | % |
| 13 | Gas oil raito | 100 - 1500 | scf/bbl |
| 14 | Fluid viscosity | 0.2 - 0.5 | cP |
| 15 | ESP lifespan | 30 - 700 | day |

**Figure 4.** Workflow for models building.

## 4. Results and Discussion

The results obtained from the models are presented in Table 2. It is clearly observed that all three models gave forecast results which deviated less than one month from the actual operation time of ESP. Table 2 also showed that the Gradient Boosting Machine model gave the best results when the forecast value is approximately 11.7 days different from the actual operation time of ESP and the RMSE is only about 21.2 days of the actual value.

The graphs comparing the actual values and forecast values of all three models Decision Tree, Random Forest and Gradient Boosting Machine respectively presented in Figures 5-7 confirmed again the above observation, which means the models can be classed from lower to higher accuracy: Decision Tree, Random Forest and Gradient Boosting Machine. Obviously, Figure 5 (Decision Tree) shows that the values are still quite much different from the line $y = x$. The deviation is significantly reduced in Figure 6 (Random Forest) and then the value are quite situated along the line $y = x$ in Figure 7 (Gradient Boosting Machine).

The explanation of this observation can be rooted to the fact that Decision Tree is a single learner, so it might be not suitable for data sets with large numbers of variables that can lead to bigger errors than the other two models. Meanwhile, the Random Forest and Gradient Boosting Machine are both ensemble learning methods, with the accuracy of Random Forest can be improved by performing voting results from hundreds of decision trees, and Gradient Boosting Machine can fix the errors of previous decision trees by the following decision trees.

The ranking of influence factors were extracted from Random Forest and Gradient Boosting Machine and then presented in Figure 8 and Figure 9, respectively. Both Random Forest and Gradient Boosting Machine models gave similar results: the three most influential parameters on ESP lifespan are motor temperature, gas-oil raito and intake pump temperature. Two of these parameters are related to temperature, which indicates that although most pumps are

designed to work under extreme conditions (high temperatures, high pressures, strong corrosive environments), the temperature is the most dangerous factor. This study demonstrated that the high-temperature operating condition leads greatly to ESP failures, hence ESP lifespan will be reduced significantly. Therefore, it is critical to lower the temperature of the ESP system in order to extend the life of the pump.

This study showed that Gradient Boosting Machine can be chosen because it gave the smallest Mean Square Error (MSE), and it can also provide an accurate ranking of the influence parameters on the lifespan of the Electrical Submersible Pumps. The GBM model has AR and RMSE values of 11.7 days and 21.2 days, respectively.

**Table 2.** Accuracy and precise comparison between predictive models.

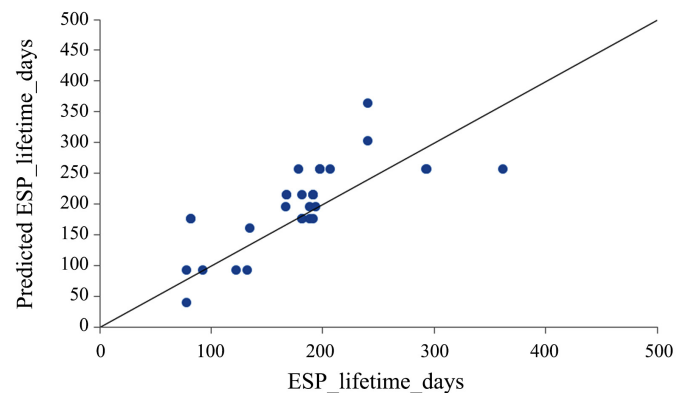| Model | Average Residual (days) | Root Mean Squared Error (RMSE) | Regression Coefficient ($R^2$) |
|---|---|---|---|
| Decision Tree | 28.25 | 51.26 | 0.3697 |
| Random Forest | 18.8 | 34.12 | 0.7186 |
| Gradient Boosting Machine | 11.7 | 21.2 | 0.8908 |



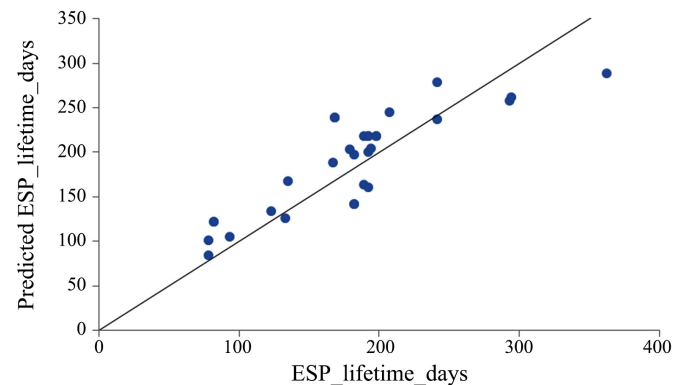**Figure 5.** Predicted vs actual comparison plot, Decision Tree model.



**Figure 6.** Predicted vs actual comparison plot, Random Forest model.
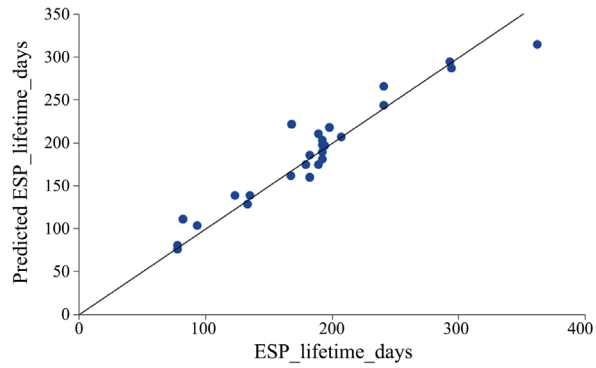
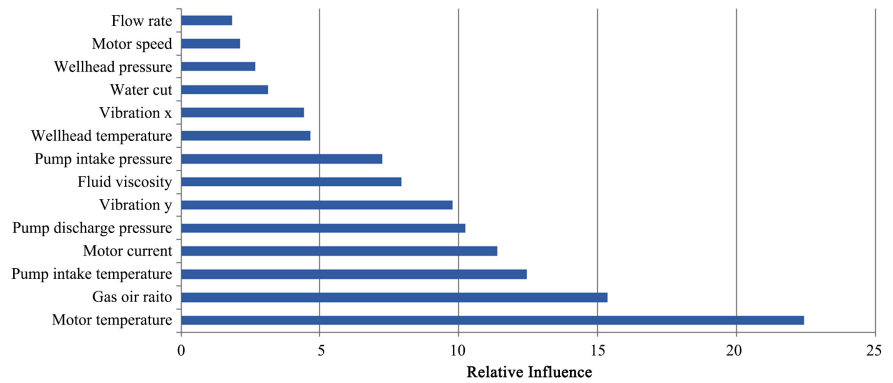**Figure 7.** Predicted vs actual comparison plot, Gradient Boosting Machine model.



**Figure 8.** Relative influence (%) of different parameters on the ESP lifespan given by Random Forest.
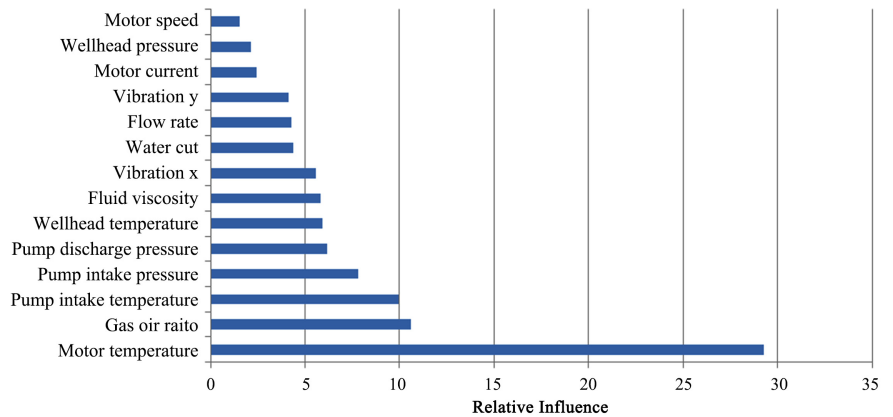


**Figure 9.** Ranking of influence parameters on the ESP lifespan given by Gradient Boosting Machine.

## 5. Conclusion

The paper proposed a proactive approach by building predictive models for Electrical Submersible Pump lifespan based on machine learning algorithms. Unlike previous studies, this study used different methods with the same data set to find out the best method to use in real life. It is concluded that the Gradient Boosting Machine is the best suitable method for predicting ESP life cycle, not

only because it has the most accurate predictive results but also it can give a ranking of influence factors. Although the temperature is known for a long time to have a damaged effect on ESP duration, it has not been demonstrated based on big data mining. Hence, this study is the first to show that the temperature is the most influential factor on ESP run life. This knowledge will help for further improvement of ESP operation worldwide.

## Acknowledgements

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] Gupta, S., Saputelli, L. and Nikolaou, M. (2016) Applying Big Data Analytics to Detect, Diagnose, and Prevent Impending Failures in Electric Submersible Pumps. *Society of Petroleum Engineers Annual Technical Conference and Exhibition*, Dubai, 26-28 September 2016, Article No. SPE-181510-MS. https://doi.org/10.2118/181510-MS

[2] Pragale, R. and Shipp, D.D. (2012) Investigation of Premature ESP Failures and Oil Field Harmonic Analysis. 2012 *Petroleum and Chemical Industry Conference*, New Orleans, 24-26 September 2012, 1-8. 10.1109/PCICON.2012.6549650

[3] Van Rensburg, N. J. (2019) Autonomous Well Surveillance for ESP Pumps Using Artificial Intelligence. *SPE Oil and Gas India Conference and Exhibition*, Mumbai, 9-11 April, Article No. SPE-194587-MS. https://doi.org/10.2118/194587-MS

[4] Takacs, G, (2009) Electric Submersible Pumps Manual: Design Operation, and Maintenance, Gulf Professional Publishing.

[5] Al Maghlouth, A., Cumings, M., Al Awajy, M. and Amer, A. (2013) ESP Surveillance and Optimization Solutions: Ensuring Best Performance and Optimum Value. *the SPE Middle East Oil and Gas Show and Conference*, Manama, 10-13 March 2013. https://doi.org/10.2118/164382-MS

[6] Guo, D., Raghavendra,C.S., Yao, K.-T., Harding, M., Anvar, A. and Patel, A. (2015) Data Driven Approach to Failure Prediction for Electric Submersible Pump Systems. *SPE Western Regional Meeting*, Garden Grove, 27-30 April 2015. https://doi.org/10.2118/174062-MS

[7] Gupta, S., Nikolaou, M., Saputelli, L. and Bravo, C. (2016) ESP Health Monitoring KPI: A Real Time Predictive Analytics Application. *SPE Intelligent Energy International Conference and Exhibition*, Aberdeen, 6-8 September 2016. https://doi.org/10.2118/181009-MS

[8] Sherif, S., Adenike, O., Obehi, E., Funso, A. and Eyituoyo, B. (2019) Predictive Data Analytics for Effective Electric Submersible Pump Management. *SPE Nigeria Annual International Conference and Exhibition*, Lagos, 5-7 August 2019. https://doi.org/10.2118/198759-MS

[9] Abdelaziz, M., Lastra, R. and Xiao, J.J. (2017) ESP Data Analytics: Predicting Fail-

ures for Improved Production Performance. *Abu Dhabi International Petroleum Exhibition* & *Conference*, Abu Dhabi, 13-16 November 2017. https://doi.org/10.2118/188513-MS

[10] Popaleny, P., Duyar, A., Ozel, C. and Erdogan, Y. (2018) Electrical Submersible Pumps Condition Monitoring Using Motor Current Signature Analysis. A*bu Dhabi International Petroleum Exhibition* & *Conference*, Abu Dhabi, 12-15 November 2018. https://doi.org/10.2118/193081-MS

[11] Sneed, J. (2017) Predicting ESP Lifespan with Machine Learning, Austin, Texas, USA: Unconventional Resources Technology Conference. *SPE/AAPG/SEG Unconventional Resources Technology Conference*, Austin, 24-26 July. https://doi.org/10.15530/URTEC-2017-2669988

[12] Van Rensburg, N.J. (2018) Usage of Artificial Intelligence to Reduce Operational Disruptions of ESPs by Implementing Predictive Maintenance. *Abu Dhabi International Petroleum Exhibition* & *Conference*, Abu Dhabi, 12-15 November 2018. https://doi.org/10.2118/192610-MS

[13] Quinlan, J.R. (1986) Induction of Decision Trees. *Machine Learning*, **1**, 81-106. https://doi.org/10.1007/BF00116251