ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Machine Learning Approach for Predictive Maintenance of the Electrical Submersible Pumps (ESPs)

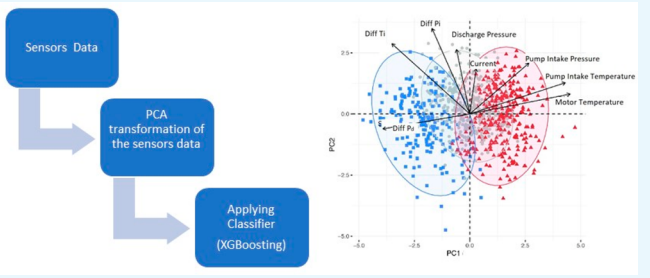
Ramez Abdalla,* Hanin Samara, Nelson Perozo, Carlos Paz Carvajal, and Philip Jaeger

Cite This: *ACS Omega* 2022, 7, 17641−17651

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Electrical submersible pumps (ESPs) are considered the second-most widely used artificial lift method in the petroleum industry. As with any pumping artificial lift method, ESPs exhibit failures. The maintenance of ESPs expends a lot of resources, and manpower and is usually triggered and accompanied by the reactive process monitoring of multivariate sensor data. This paper presents a methodology to deploy the principal component analysis and extreme gradient boosting trees (XGBoosting) in predictive maintenance in order to analyze real-time sensor data to predict failures in ESPs. The system contributes to an efficiency increase by reducing the time required to dismantle the pumping system, inspect it, and perform failure analysis. This objective is achieved by applying the principal component analysis as an unsupervised technique; then, its output is pipelined with an XGBoosting model for further prediction of the system status. In comparison to traditional approaches that have been utilized for the diagnosis of ESPs, the proposed model is able to identify deeper functional relationships and longer-term trends inferred from historical data. The novel workflow with the predictive model can provide signals 7 days before the actual failure event, with an F1-score more than 0.71 on the test set. Increasing production efficiencies through the proactive identification of failure events and the avoidance of deferment losses can be accomplished by means of the real-time alarming system presented in this work.



1. INTRODUCTION

Recently, the trends of automation and digitalization, artificial intelligence (A.I.), and machine learning have gained momentum. Also, oil field digitization is considered a whole new opportunity for further production optimization in the oil and gas industry.[1] The key question arises of how to implement these tools in such a way that all known risks are managed, value is genuinely delivered, and the actual results make a measurable difference to the profitability of the operation and, of course, that they are applicable to specified and predefined production optimization goals.

Previous research in this area promised to further revolutionize key aspects of oil production applications including well monitoring and control, reservoir management,[2−4] production optimization,[5,6] artificial lift,[7−10] flow assurance,[11,12] and predictive maintenance. In general, this research area focuses on the utilization of machine learning in order to understand the status of equipment so as to facilitate predictive maintenance and to avoid operation downtime.

One of the most widely used artificial lift technologies is the electrical submersible pump (ESP).[13] They are installed in many producing wells that are subject to harsh environments and need to pump complex fluid mixtures that on their turn undergo changes in composition, pressure, and temperature over time. For assuring a reliable fluid delivery, on time interventions are required in case of upcoming problems. Hence, strong efforts are undertaken in the area of a "digital oil field" that focus on deploying machine learning and data-driven models in the area of predictive pump maintenance of electrical submersible pumps.[14]

Beyond only creating a relation between continuous data, a data-driven model can be used to understand the internal relations between the parameters generating these data. Therefore, the key to perform fault detection on the ESP can be better defined as a problem to build an accurate data-driven model that describes the ESP system dynamics. Table 1 shows various contributions in the area of predictive maintenance of electrical submersible pumps.

The previous literature includes applications that only deal with statistical analysis in a descriptive way,[15−17,36] while the rest are diagnostic analyses. The diagnostic analysis literature can be divided into two groups. The first group encompasses

Received: January 25, 2022
Accepted: April 29, 2022
Published: May 19, 2022

ACS Publications

© 2022 The Authors. Published by American Chemical Society

Table 1. Summary of the Most Relevant Studies Related with This Paper

author, year	relevant work
(Zhao et al., 2006); (Li et al., 2008); (Zhang, 2017)	ESP fault tree diagnosis through a proposed qualitative and quantitative method.[15−17]
(Xi, 2008)	the use of a traditional mechanical fault diagnosis and wavelet analysis
	realization of excessive shaft thrust and wear fault characteristic extraction to investigate the fault diagnostics of the centrifugal pump[18]
(Wang, 2004)	use of Neuro-Fuzzy Petri nets and extracted features for the identification of eccentric wear of both the impeller and bearing as well as the sand plug of the impeller[19]
(Zhao, 2011)	ESP vibration signal analysis, feature extraction, and establishment of typical fault vibration mechanical models[20]
(Tao, 2011)	data analysis and application of vibration signals based on wavelet analysis and wavelet transform in the ESP.[21]
(Guo et al., 2015)	utilization of the support vector method in the prediction of anomalous operation[22]
(Wang, 2013) (Peng, 2016)	utilization of back propagation (BP) neural networks for ESP diagnosis.[23,24]
(Jansen Van Rensburg, 2019)	exploration of surveillance-by-exception on ESP using a train model with normal yet good quality data[25]
(Andrade Marin et al., 2019)	analysis of random forest to obtain a high value of accuracy and recall of ESP failure prediction in 165 cases[26]
(Adesanwo et al., 2016); (Adesanwo et al., 2017); (Gupta et al., 2016); (Abdelaziz et al., 2017); (Bhardwaj et al., 2019); (Sherif et al., 2019); (Peng et al., 2021); (Zhang et al., 2017); (Yang et al., 2021)	application of principal component analysis (PCA) for anomaly detection and failure prediction for the identification of correlations in the dynamic ESP parameters such as intake pressure and temperature, discharge pressure, vibrations, motor and system current and frequency measured by means of a variable speed drive (VSD) at regular time intervals[27−35]

applications that rely on ammeter charts, which is an old technology in ESP troubleshooting.[18,37,38] The second group includes applications that depend on pump-deployed sensors.[27−34]

Regarding the second group, it is noticeable that the majority of the applications attempts to use the sensor data transformation on principal component analysis (reduction of the dimensionality of large data sets). Then, the data are projected to map the sensor readings. The objective in this case is to group data from pump sensors based on their downhole conditions. The majority of the recent research either only used PCA as an unsupervised learning technique for real-time diagnosis of the ESP or applied surveillance-by-exception on the system (detection of disruptive events), but none of them was a predictive approach. Surveillance-by-exception is done by using a normal range of the sensor data to train the algorithms. Then, the algorithm is used on the test data to detect points located outside the predication confidence interval.[33,39]

In this paper, a methodology along with its implementation is presented for the application of PCA using the so-called extreme gradient boosted trees machine learning technique, in order to provide an intuitive way of predicting downhole failures of the ESP system 7 days ahead, before the workover. The workflow is arranged in this paper as follows: first the proposed methodology and its implementation are explained in detail, followed by introduction of an evaluation technique and finally presentation and discussion of the results of its application.

2. PROPOSED METHODOLOGY AND IMPLEMENTATION

This research intends to develop a model that can predict downhole electrical submersible pump problems, so that proper actions might be taken proactively to avoid the occurrence of such problems. The approach of the supervised learning is used to train the model. This model will be able to predict the probability of some abnormal conditions or class label a few days before events. Finally, its reliability and accuracy will be tested.

Supervised learning algorithms will be used to analyze the training data. These algorithms produce an inferred function capable of mapping the training examples. Also, they will be allowed to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.[40]

The study is executed using the Knowledge Discovery in Databases (KDD) process.[41] This process is used to show (1) data collection, (2) data preprocessing, (3) how to extract the features, (4) the use of the proper classifier and its relevant hyperparameter tuning, and (5) the evaluation of the results. Figure 1 shows the (KDD) phases. In this study, downhole

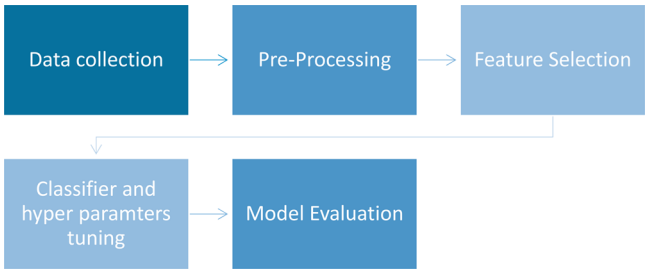


Figure 1. Knowledge discovery in database workflow.

conditions are considered as the dependent variable. On the other hand, the dynamometer cards data are considered as the independent variables.

2.1. Data Collection and Preprocessing. Time-series data are collected from sensors on electrical submersible pumped wells. The reported measurements are pump frequency (FRQ), pump discharge pressure (PDP), pump intake pressure (PIP), well head pressure and temperature (WHP/WHT), motor temperature (MT), casing head pressure (CHP), and variable speed drive output current (Current). These measurement data have different frequencies. Also, well status sheets for the same wells are gathered on a daily basis at the same time periods. These data were collected from a field undergoing polymer flooding. Based on the status sheets, pumps exhibited two main problems. These two problems were motor downhole failures (MDHFs) or

https://doi.org/10.1021/acsomega.1c05881
ACS Omega 2022, 7, 17641−17651

electrical downhole failures (EDHFs); therefore, both failures are categorized as electrical pump failures.

Electric failure of the downhole facilities constitutes failure of any of the electrical components in the ESP assembly including the electric cable, the motor electrical components such as the stator, and the downhole sensor. Failures associated with the cable were mainly caused by electric cable failure, cable insulation failure due to corrosion, material failure, and abrasion, and cable failure due to overload. Meanwhile, electrical failures associated with the motor are usually a resultant of the stator failure. The stator has been reported to fail due to overheating. As the motor is the hottest point in the well, this appears to worsen polymer deposition on the motor body. This in turn reduces heat dissipation, leading to increasing motor winding temperature, which in turn makes the deposition worse and causes an eventual ramp down of the ESP frequency when maximum motor temperatures are reached. In addition, the high temperatures around the motor aids the precipitation of solid polymer in fluids flowing past the motor and are the source of polymer plugging in the pump inlet.

The workflow in predictive modeling starts with the data cleaning process, known as cleansing. On one hand, it is important to eliminate unphysical values (e.g., negative or enormous pressure values), remove further outliers, and align units. On the other hand, it is a critical step for handling noise data while maintaining the realistic anomalies that may identify downhole problems of the pumps.

After visual inspection of the data, a pipeline of a preprocessing strategy is created. First, it starts by resampling the data using a moving median in 1 h steps. Figure 2 shows

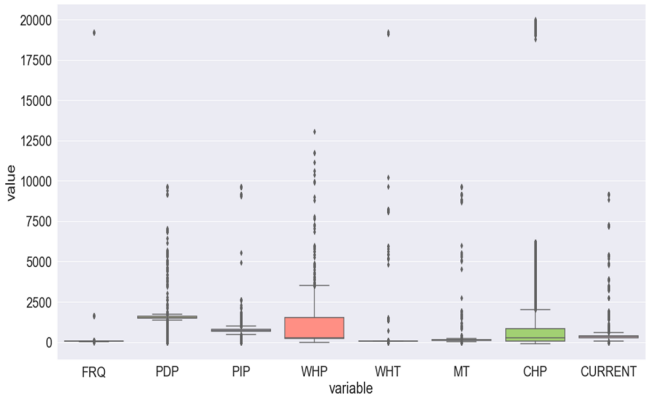


Figure 2. Data box plot before outlier detection.

the box plot of the data after resampling and before outlier removal. It is obvious that some measurements include unreasonable values. For example, the well head temperature reaches 18 500 °F, which is obviously a measurement error. Therefore, the second step is removing outliers. It includes first removing measurements where oil production is zero; then, outlier removal by limits is applied.

Outlier removal by limits depends mainly on quartiles; therefore, we used box plots. They summarize sample data using the 25th, 50th, and 75th quartiles. The midspread or middle 50th, or technically H-spread, is a measure of statistical dispersion, being equal to the difference between the 75th and 25th percentiles. It is called the interquartile range (IQR). IQR is somewhat similar to Z-score in terms of finding the distribution of data and then keeping some threshold to identify the outlier. To define the outlier, a base value is defined above and below the normal range of a data set, namely the upper and lower bounds. The upper and the lower bounds are calculated according to eqs 1 and 2.

$$upper = \mathrm{Q3} + 1.5{*}\mathrm{IQR} \tag{1}$$

$$lower = \mathrm{Q1} - 1.5{*}\mathrm{IQR} \tag{2}$$

Afterward, a standard scaler is employed (subtracting the mean from each point and dividing by the variance), transforming the mean value to zero and scaling the data to unit variance. Finally, the moving difference is applied on all sensor measurements. Figure 3 shows the box plot after outlier removal, and Figure 4 shows the box plot after normalization.

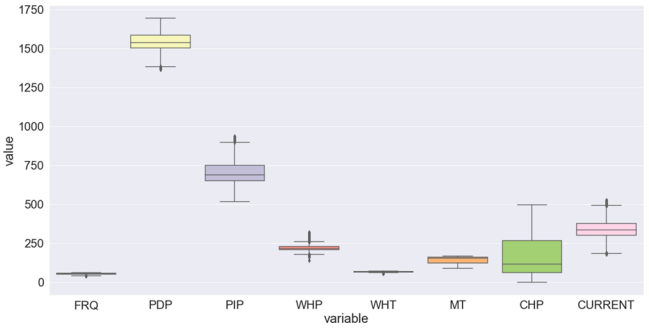


Figure 3. Box plot after outlier removal.

Table 2 describes the main signals after outliers and zero production points are removed without standardization. Table 3 shows the number of available data points after mapping the sensor data. These data points are classified, based on workover sheets, into "normal data", "preworkover", and "workover". Preworkover data are data points that are reported 7 days before the workover day. Workover events are the data points made available on workover day.

2.2. Principal Component Analysis Application. PCA is defined as an unsupervised dimensionality reduction technique. It reduces large dimensionality data sets into lower dimensions called principal components. This happens while preserving as much information as possible. It makes use of the interdependence of original data to build a PCA model. This results in reducing the dimensions of production parameters by making the most of the linear combinations and by generating a new principal component space (PCs).[42]

2.2.1. Principal Component Analysis Calculations. The process of obtaining a PCA model from a raw data set is divided into four steps as follows:

First, the covariance matrix ($\sum$) of the whole data set is computed. It is important to see whether there is a relationship between contributing features.

$$\sum = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \widetilde{x_j})(X_{ik} - \widetilde{x_k}) \tag{3}$$

Eq 3 is used to find the covariance between each pair of data set columns.

The second step is to calculate eigenvectors and corresponding eigenvalues. Let $\boldsymbol{A}$ be the covariance matrix that has been computed in the first step, ν be a vector, and λ be a scalar that satisfies $\boldsymbol{A}\nu = \lambda\nu$; then, λ is the eigenvalue corresponding to the eigenvector ν of $\boldsymbol{A}$. This step is

https://doi.org/10.1021/acsomega.1c05881
ACS Omega 2022, 7, 17641−17651

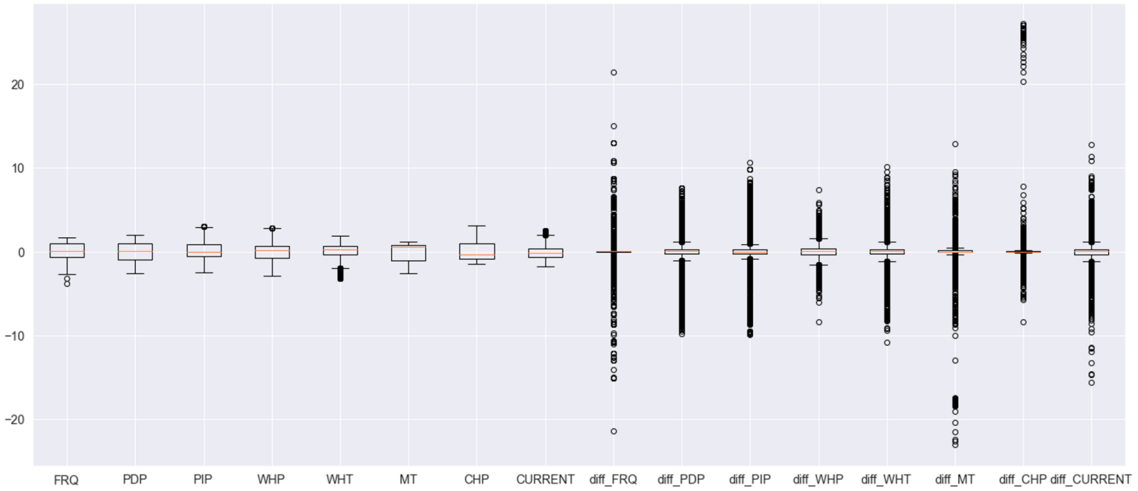


Figure 4. Box plot after outlier removal, normalization, and use of moving difference signals.

Table 2. Data Exploration

	FRQ (Hz)	PDP (Psi)	PIP (Psi)	WHP (Psi)	WHT (F)	MT (F)	CHP (Psi)	CURRENT (A)
mean	52.68	1522.96	855.89	642.20	62.38	137.11	922.13	349.65
std	4.29	187.46	211.08	608.61	11.49	32.39	1883.03	86.90
min	35.00	1086.51	610.45	194.06	60.54	120.62	0	101.59
25%	49.70	1506.13	660.90	213.38	63.23	122.20	91.54	317.00
50%	52.76	1543.80	721.80	243.58	65.99	154.19	261.38	354.00
75%	56.58	1595.04	778.89	547.33	67.65	160.60	405.28	394.73
max	64.96	1893.23	1578.28	845.86	122.84	169.30	986.74	598.21

Table 3. Data Points Classification

condition	reported data points
normal	339 089
preworkover	1728
workover	288

considered the calculation of the principal components of the data.

The third step is determining the number of principal components. The eigenvectors only define the directions of the new axis, while the eigenvalues represent the variance of the data along the new feature axes. Therefore, we sort the eigenvectors based on the eigenvalues. Hence, a threshold is chosen on the eigenvalues, and a cutoff is made on the eigenvectors to select the most informative lower dimensional subspace. In other words, lower variance dimensions are omitted. This is because they possess the least information about the data's distribution.

The fourth step consists of transforming the samples into the new subspace. In this last step, the lower dimensional subspace **W** is selected. In the current step, the data set samples are transformed into this new subspace via the equation $\boldsymbol{Y} = \boldsymbol{W}' \cdot \boldsymbol{X}$ where $\boldsymbol{W}'$ is the *transpose* of the matrix $\boldsymbol{W}$. In the following, two principal components are computed, and the data points are reoriented onto the new subspace. Figure 5 shows the simple geometric meaning of PCA.

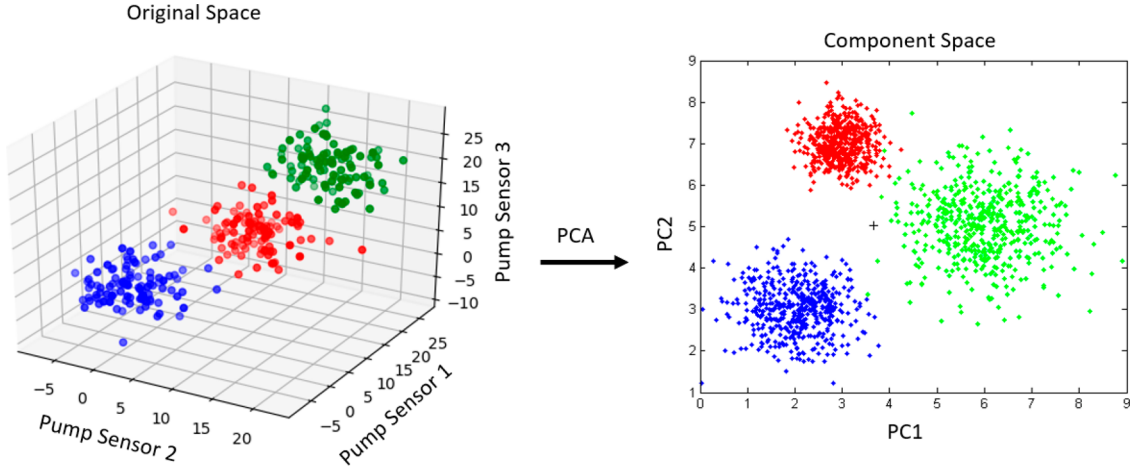


Figure 5. Geometric meaning of PCA.

https://doi.org/10.1021/acsomega.1c05881
ACS Omega 2022, 7, 17641−17651

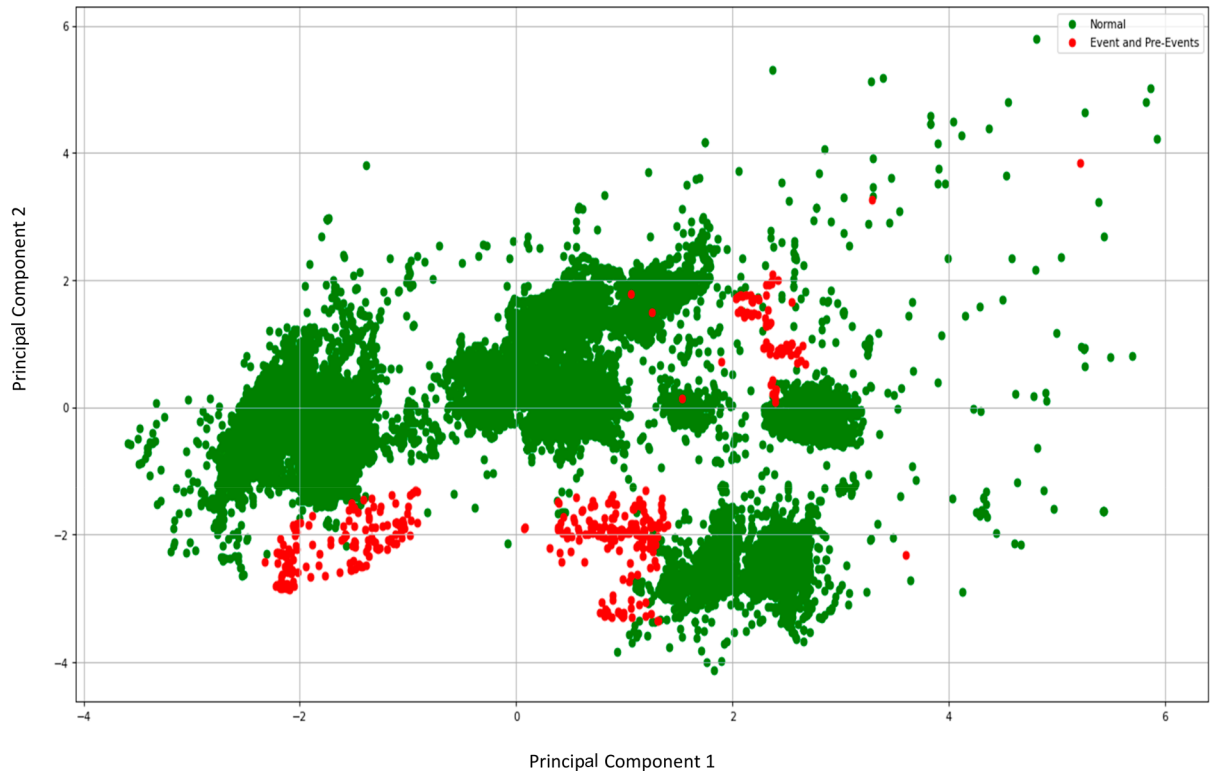


Figure 6. Principal component analysis of ESP wells.

Table 4. Loading for Input Parameters

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
FRQ	−0.90	−0.05	0.06	0.05	−0.05	−0.26	0.05	0.01
PDP	−0.54	−0.14	−0.70	0.32	−0.07	−0.08	0.04	0.05
PIP	0.03	−0.17	−0.47	0.14	−0.05	0.81	−0.18	−0.04
WHP	0.10	−0.02	−0.79	0.37	−0.01	−0.15	0.23	−0.28
WHT	−0.63	0.00	0.35	−0.30	0.18	0.22	0.12	−0.32
MT	−0.79	−0.12	−0.07	0.04	−0.12	0.25	−0.16	0.22
CHP	−0.85	−0.15	−0.07	0.15	−0.03	−0.27	0.07	0.12
CURRENT	−0.84	−0.10	0.21	−0.17	0.07	0.21	−0.06	−0.19
diff_FRQ	−0.14	0.78	0.03	0.27	0.09	0.02	−0.22	−0.17
diff_PDP	−0.05	0.09	−0.45	−0.54	0.21	−0.19	−0.40	0.28
diff_PIP	0.06	−0.35	−0.34	−0.67	−0.21	0.06	0.18	0.11
diff_WHP	−0.03	0.11	−0.42	−0.50	0.36	−0.15	−0.05	−0.42
diff_WHT	−0.10	0.43	−0.08	−0.11	0.37	0.23	0.69	0.26
diff_MT	−0.15	0.84	−0.10	−0.10	−0.34	0.03	−0.03	−0.03
diff_CHP	0.03	−0.29	0.03	0.30	0.85	0.01	−0.14	0.11
diff_CURRENT	−0.13	0.80	−0.14	−0.04	0.16	0.05	−0.09	0.20

2.2.2. Application of Principal Component Analysis in Electrical Submersible Pumps. In ESP systems, sensor data are generally highly correlated, e.g., wellhead pressure is directly proportional to discharge and intake pressures. However, when a downhole problem occurs or is about to occur, anomalous data can be identified, because it breaks certain rules in the input signals and their relative changes, i.e., if there is a tubing leak, the annulus discharge pressure decreases, while intake pressure and annulus pressure increase, etc.

Principal component analysis then serves an engineer's purpose in creating an anomaly detection system. This is mainly because it makes use of the interdependence of original data to build a model. The primary goal of this step is to create clusters out of the data.

As discussed earlier, the selection of the principal components is made based on the maximum variance criterion. The highest variance is captured in the first principal component, while the next highest variance is captured in the second principal component, where information from the first principal component has already been removed. In a similar manner, consecutive principal components (third, fourth, ⋯, kth) can be constructed to evaluate the original system.

The PCA model finds the kth principal component to construct the PCs, where most of the information belonging to

https://doi.org/10.1021/acsomega.1c05881
ACS Omega 2022, 7, 17641−17651

the initial system is contained. The kth principal component is represented in eq 4 below, where PC1 is given as an example.

$$PC_k = a_{1k}{}^*P_{intake\ pressure} + a_{2k}{}^*P_{discarge\ pressure} + a_{3k}P_{motor\ temperature} + \ldots + a_{pk}P_{intake\ pressure\ moving\ difference} + \ldots\text{etc} \quad (4)$$

Figure 6 shows the projection of ESP well sensor data on the principal components. The developed model is used also to evaluate near failure conditions. The problematic days and 7 days before workover, sensor data clearly show specific failure patterns in line with the reported motor (MDHF) and electric (EDHF) downhole failures.

The goal of the PCA is to come up with optimal weights from each sensor measurement. That means capturing as much information as possible from the input signals, based on the correlations among those variables. The loadings are the correlations between the variables and the component. We compute the weights in the weighted average from these loadings. To compute the loading matrix, namely the correlations between the original variable and the principal components, the cross-covariance matrix needs to be computed using eq 5.

$$cov(X, Y) = V\sqrt{E} \quad (5)$$

where X represents the original variables, Y represents the principal components, V represents the principal axes, and E represents its eigenvalues.

Table 4 represents the load factor for each input parameter in the relevant principal component up to the eighth principal component. However, we are mostly interested in parameter loading factors on the first and second principal components, because they explain approximately 0.6 of the data variance (see Figure 7). Large loadings (positive or negative) indicate

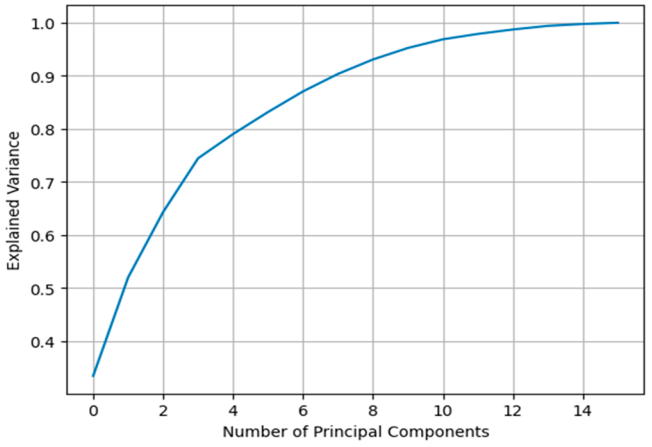


Figure 7. Explained variance of the proposed model.

that a particular variable has a strong relationship to a particular principal component. The sign of a loading indicates whether a variable and a principal component are positively or negatively correlated.[43] Hence, the parameters that exhibit the highest correlation with the first principal component are pump frequency, casing head pressure, current, motor temperature, and well head temperature.

2.3. Extreme Gradient Boosting (XGBoost). XGBoost is a tree-based ensemble model. Ensemble learning is a systematic solution that combines the predictive abilities of multiple models, eventually resulting in a single model. This single model provides the aggregated output of several models that, on their turn, only perform slightly better than random guessing. Therefore, extreme gradient boosting (XGBoost) is an ensemble set of predictors, with a unified objective of predicting the same target variable. A final prediction is performed through the combination of these predictors.

2.3.1. Extreme Gradient Boosting (XGBoost) Calculations. Building an XGBoost model has the following sequence. It starts with a single root (contains all the training samples). Then, an iteration is performed over all features and values per feature, and subsequently, each possible split loss reduction is evaluated. Eqs 6 and 7 represent the objective function (loss function and regularization, respectively) at each iteration that is needed to be minimized.

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, p_i + O_{value}) + \frac{1}{2}\lambda O_{value}^2 \quad (6)$$

$$l(y_i, p_i) = -[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \quad (7)$$

where y_i is the true value required to be predicted of the i-th instance; p_i is the prediction of the i-th instance; $l(y_i p_i)$ is the loss function for a typical classification problem; O_{value} is the output of the new tree, and $\frac{1}{2}\lambda O_{value}^2$ is the regularization term.

Chen stated "XGBoost objective function cannot be optimized using traditional optimization methods in Euclidean space".[44] Therefore, in order to be able to transform this objective function to the Euclidean domain, the second-order Taylor approximation is using enabling traditional optimization techniques to be employed. Eqs 8 and 9 represent the Taylor approximation of the loss function.

$$\mathcal{L}^{(t)} \cong \left[\sum_{i=1}^{n} l(y_i, p_i) + g_i O_{value} + \frac{1}{2}h_i O_{value}^2\right] + \frac{1}{2}\lambda O_{value}^2 \quad (8)$$

where g_i is the gradient and calculated by $g_i = \frac{\partial}{\partial p_i} l(y_i, p_i)$ and h_i is the Hessian and calculated by $h_i = \frac{\partial^2}{\partial p_i^2} l(y_i, p_i)$.

Finally, removing the constant parts, the simplified objective to minimize at step t, results in

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} g_i O_{value} + \frac{1}{2}h_i O_{value}^2 + \frac{1}{2}\lambda O_{value}^2 \quad (9)$$

Eqs 10 and 11 show how to minimize that function

$$\frac{\mathrm{d}}{\mathrm{d}O_{value}} \sum_{i=1}^{n} g_i O_{value} + \frac{1}{2}h_i O_{value}^2 + \frac{1}{2}\lambda O_{value}^2 = 0 \quad (10)$$

$$O_{value} = -\frac{\sum_{i=1}^{n} g_i}{\sum_{i}^{n} h_i + \lambda} \quad (11)$$

By combining eq 10 with the first and the second derivatives of the classification loss functions g_i *and* h_i, the similarity equation is derived. The similarity score is calculated as follows in eq 12

$$similarity = \frac{\sum residual_i}{\sum previous\ probability_i{}^*(1 - previous\ probability_i) + \lambda} \quad (12)$$

https://doi.org/10.1021/acsomega.1c05881
ACS Omega 2022, 7, 17641−17651

Table 5. Hyperparameter Tuning

parameter	reference to	sampling type	range
max_depth	control of overfitting, higher depth facilitates such that the model learns relations that are specific to a particular sample	suggest integer value	2, 10
min_child_weight	a minimum sum of weights is defined for all observations required in a child	log uniform	1e−10, 1e10
colsample_bytree	the subsample ratio of columns when constructing each tree	uniform	0, 1
learning_rate	overfitting prevention through step size shrinkage in updates	uniform	0, 0.1
gamma	specification of the minimum loss reduction required to make a split	suggest integer value	0, 5

The similarity score is calculated for a "leaf" of the "tree". Various thresholds are used to split the tree into more leaves. The similarity score is calculated for each new leaf followed by calculating the so-called gain as presented in eq 13 below

$$gain = left_{similarity} + right_{similarity} - root_{similarity} \quad (13)$$

Then, thresholds continue to be set until higher gain thresholds are reached and the tree keeps growing. There is a minimum number of residuals in each leaf where the tree stops growing. This number is determined by calculating a parameter called cover. It is defined as the denominator of the similarity score minus lambda. During boosting, the operation is performed such that trees are sequentially constructed. Each tree reduces the error of its predecessor and learns from it while simultaneously updating the residual errors. As a result, each tree growing in the sequence will learn from a version of the residuals that is already been updated.

Further, in boosting, the base learners are weak due to their high bias, and their predictive power has only a slight improvement over random guessing. Nevertheless, some vital information for prediction is supplied by each of these weak learners. By means of boosting, a strong learning effect is produced through combining these weak learners into a single strong learner that reduces both the bias and the variance.

2.3.2. Extreme Gradient Boosting (XGBoost) Application. In our proposed model, principal component analysis (PCA) for sensor measurements and moving difference is pipelined with XGBoost and k-folds cross-validation to identify near failure regions. The data set is divided into two groups: a training data set containing 70% of the data and a black box testing set with the remaining 30% of the data.

The importance of principal components is evident, because it shows to which extent this component is able to explain the variance in the data set. Therefore, Figure 7 shows the cumulative explained variance with each principal component. It is shown that eight principal components will include more than 90% of the explained variance in the data set of ESP sensors and their derived features.

In the cross-validation algorithm, the data set is divided into three components as follows: a training set constituting 70% of the data, a validation set constituting 15% of the data, and a testing set constituting the remaining 15% of the data. Each model is then trained on the training subset only, in order to infer some hypothesis. Finally, the hypothesis with the smallest error on the cross-validation set is selected.

A better estimation of each hypothesis is achieved through testing a set of examples (validation set) that the models were not trained on. A true generalization error is also obtained. As a consequence, a single model possessing the smallest estimated generalization error can be then proposed. Upon validation set error minimization, this can be further expanded such that the proposed model is retrained on the entire training set, including the validation set. It is worth noting that some risk exists in selecting validation points, which may contain a disproportionate amount of difficult and obscure examples. Therefore, the k-fold cross validation maybe applied to avoid such occurrences.

A K-fold cross validation algorithm aims at selecting validation sets. Initially, the data set is randomly divided into (k) disjoint subsets. In each subset, the number of readings is equal to the total number of data points (m) over (k). These subsets are indicated by m_1 to m_k. Then, subset is evaluated for each model as follows:

All these subsets are used to train the XGBoost model, with the exception of the subset m_j. The intention behind excluding this subset is to infer a hypothesis that is eventually tested on (m_j). As such, the error of testing the hypothesis on the subset (m_j) is calculated, and the estimated generalization error of the model is calculated by averaging over (m_j). Afterward, the selection of the model with the lowest estimated generalization error is performed, and last, the selected model is retained on the entire training set (m). The hypothesis resulting from such operation would be the final answer. When performing cross validation, It is typically a standard that the chosen number of folds is equal to 10 $(k = 10)$.[45]

Hyperparameter tuning is considered one of the important steps while creating any data-driven model to get the best results from the deployed algorithm. Regarding the XGBoost algorithm, hyperparameters are divided into three categories. These categories are known as general parameters, booster parameters, and learning task parameters.

General hyperparameters define the type of algorithm to be either linear or tree-based, the verbosity to print results, and the number of threads to run on. Booster parameters include the main tuned parameters for the algorithms such as the learning rate, the minimum sum of weights of all observations required in an internal node in the tree, and the learning parameters to specify the minimum loss reduction required to make a split.[46] These parameters are used to define the optimization objective and the metric to be calculated at each step. Table 5 shows ranges that are used for hyperparameters tuning.

3. MODEL EVALUATION

3.1. Evaluation Metrics. Some questions are vital for understanding the classifier performance. One of which is the number of signals that have been classified correctly among the entirety of those that have been classified as "preworkover". The answer lies in inspecting the model's precision. Precision is the ratio of positives that have been correctly classified to the sum of both positives and negatives. This is the percent of the true alarms, which is an important measure to eliminate the false preworkover alarms as much as possible.[41] Eq 14 shows the precision as per below

https://doi.org/10.1021/acsomega.1c05881
ACS Omega 2022, 7, 17641−17651

$$precision = true\ positive/(true\ positive + false\ positive) \quad (14)$$

Another common question is the proportion of correctly classified preworkover signals (TP) to the total preworkover signals (TP + FN) that are identifiable and nonidentifiable by the model. This is the recall, or the true positive rate which indicates how capable the model is of finding the preworkover signals.[41] Eq 15 below shows the recall

$$recall = true\ positive/(true\ positive + false\ negative) \quad (15)$$

The F1-score is the harmonic mean of the precision and recall. F1-score is used for model validation. F-measure has an intuitive meaning. It describes how precise our classifier is (how many events are classified correctly) as well as how robust it is, i.e., not missing a significant number of events.

3.2. Diagnostic Tools. A receiver operating characteristic curve (ROC) is applied as a diagnostic tool where the performance of a classification model is summarized with respect to the positive class. The false positive rate is the *x*-axis, and the true positive rate is the *y*-axis.

The true positive rate is the ratio of the total number of true positive predictions to the sum of the true positives and the false negatives (e.g., all examples in the positive class). The true positive rate is referred to as the sensitivity or the recall as shown in eq 16.

$$true\ positive\ rate = true\ positives/(true\ positives + false\ negatives) \quad (16)$$

The false positive rate is the ratio of the total number of false positive predictions to the sum of the false positives and true negatives (e.g., all examples in the negative class).[47] Eq 17 calculates the false positive rate.

$$false\ positive\ rate = false\ positives/(false\ positives + true\ negatives) \quad (17)$$

4. RESULTS AND DISCUSSION

To reduce the false alarms in our model, data exploration is performed, and then, raw-sensor data are preprocessed. Afterward, the "cleaned standardized" time-series data with their moving difference are entered into feature engineering transformation through the use of PCA. Finally, an ML model is used to classify the operating points. The upcoming results are reported in two different processes. First, validation results are reported for the 10 folds of the data set. Along with model training, model validation intends to locate an ideal model with the best execution. The model performance is optimized using training and the validation data set. Therefore, ROC curves are reported for the 10 folds of the data set and their mean value. Then, the model generalization performance is tested using the testing set. The test data set remains hidden during the model training and model performance evaluation stage. In this regard, the precision recall curve is used.

Figure 8 shows the fraction of correct predictions for the positive class depicted on the *y*-axis versus the fraction of errors for the negative class depicted on the *x*-axis. For interpreting the ROC curve, a single score can be given for a classifier model through the so-called "ROC area under curve" (AUC), which is attained, as the name implies, by integrating the area under the curve. The score has a value ranging between 0.0 and 1.0, which indicates a perfect classifier. Figure 8 also shows the ROC curves for our proposed model with 10-fold validation sets and its mean curve. The mean value for the ROC AUC is 0.95.

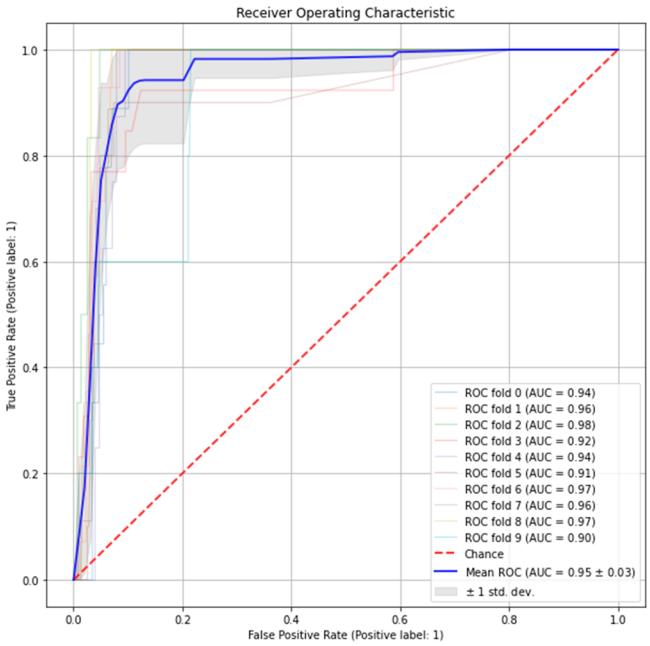


Figure 8. ROC for the proposed model.

As mentioned earlier, the second process was testing the proposed model against a testing set using the precision-recall curve (PRC), which is a valuable diagnostic tool particularly when classes are very imbalanced. The PRC trade-off between a classifier's precision, a measure of result relevancy, and recall, a measure of completeness for every possible cutoff, is depicted. Figure 9 shows a precision recall curve (PRC) for the preworkover and workover class.

It is clear that the data set is unbalanced. For this reason, it is important to check the precision and recall for each class of the pumping conditions for better evaluation of the classifier. From Figure 9 and Table 6, the precision and the recall for preworkover and workover condition is less than those in normal conditions. This is mainly due to a higher number of data points supporting the normal labeled status. This is an effect of using an unbalanced data set. One approach to addressing the imbalanced data sets is to oversample the minority class. The simplest approach involves duplicating examples in the minority class, although these examples do not add any new information to the model. Instead, new examples can be synthesized from the existing examples. This is a type of data augmentation for the minority class and is referred to as the synthetic minority oversampling technique (SMOTE). This can be part of further work. However, such procedures are inherently dangerous, because they may result in overfitting of the model.

5. CONCLUSION

In this application, sensor measurements with a moving difference are applied to the data set in order to predict the pumping condition. Then, a dimensionality reduction technique is used, and the whole data set has been projected to the new lower dimensions. Finally, these new transformed data have been pipelined with a supervised algorithm, which is XGBoosting in our application. The training data set consists of inputs (PCA projected features) paired with the

https://doi.org/10.1021/acsomega.1c05881
ACS Omega 2022, 7, 17641−17651

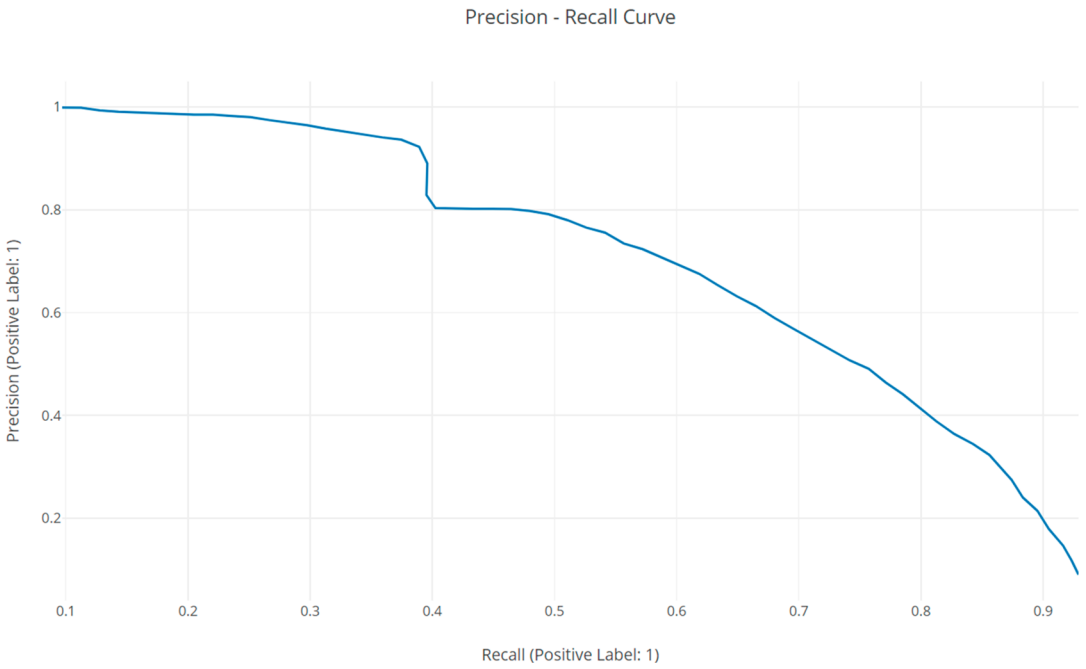


Figure 9. Precision recall curve.

Table 6. Precision, Recall, and F1-Score

	precision	recall	F1-score	support
normal	0.99	1.00	1.00	101 726
7 days or less pre-event	0.80	0.60	0.71	604

representative outputs (in case of ESP failure prediction, the labeled outputs are the 7 days before the reported failures). Each of these input−output pairs should be seen as a "data point" that can be used to train, validate, and test the proposed model.

Regarding the validation set, the proposed model has a mean AUC for the 10-fold validation equal to 0.95, which in turn means that the model has an adequate performance and can be tested on the upcoming processes against test sets.

Regarding testing sets, the proposed model can report the preworkover and workover classes with 0.8 precision and 0.6 recall. The model has high precision on testing sets and hence a small number of false alarms. Of course, the relevant recall is small, which means not all of the 7 days before the event are marked as a yellow alarm (preworkover and workover events). In other words, the model will report alarms with high precision but not for all days before the workover, which is acceptable, because it is not necessary that all days before the event will exhibit a sign of an upcoming workover.

AUTHOR INFORMATION

Corresponding Author

Ramez Abdalla − *Clausthal University of Technology, Institute of Subsurface Energy Systems, 38678 Clausthal-Zellerfeld, Germany;* orcid.org/0000-0001-5118-0011; Email: ramez.abdalla@tu-clausthal.de

Authors

Hanin Samara − *Clausthal University of Technology, Institute of Subsurface Energy Systems, 38678 Clausthal-Zellerfeld, Germany;* orcid.org/0000-0002-8662-6666

Nelson Perozo − *Clausthal University of Technology, Institute of Subsurface Energy Systems, 38678 Clausthal-Zellerfeld, Germany*

Carlos Paz Carvajal − *Clausthal University of Technology, Institute of Subsurface Energy Systems, 38678 Clausthal-Zellerfeld, Germany*

Philip Jaeger − *Clausthal University of Technology, Institute of Subsurface Energy Systems, 38678 Clausthal-Zellerfeld, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.1c05881

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge the financial support by the Open Access Publishing Fund of Clausthal University of Technology.

REFERENCES

(1) Diker, G.; Frühbauer, H.; Bisso Bi Mba, E. M. Development of a Digital ESP Performance Monitoring System Based on Artificial Intelligence. *Abu Dhabi International Petroleum Exhibition & Conference* **2021**, SPE-207929-MS.

(2) Bai, Y.; Li, J.; Zhou, J.; Li, Q. Sensitivity Analysis of the Dimensionless Parameters in Scaling a Polymer Flooding Reservoir. *Transp Porous Med.* **2008**, *73* (1), 21−37.

(3) Hou, J.; Zhou, K.; Zhang, X.-S.; Kang, X.-D.; Xie, H. A review of closed-loop reservoir management. *Pet. Sci.* **2015**, *12* (1), 114−128.

(4) Ma, H.; Yu, G.; She, Y.; Gu, Y. Waterflooding Optimization under Geological Uncertainties by Using Deep Reinforcement Learning Algorithms. *SPE Annual Technical Conference and Exhibition* **2019**, SPE-196190-MS.

(5) Forouzanfar, F.; Della Rossa, E.; Russo, R.; Reynolds, A. C. Life-cycle production optimization of an oil field with an adjoint-based gradient approach. *J. Pet. Sci. Eng.* **2013**, *112*, 351−358.

(6) Hidayat, A.; Prakoso, N. F.; Sujai, A.; Medco, P. T. Production and Cost Optimization in a Complex Onshore Operation Using

https://doi.org/10.1021/acsomega.1c05881
ACS Omega 2022, 7, 17641−17651

Integrated Production Model. *SPE Symposium: Production Enhancement and Cost Optimisation* **2017**, SPE-189223-MS.
(7) Shafiei, A.; Dusseault, M. B.; Zendehboudi, S.; Chatzis, I. A new screening tool for evaluation of steamflooding performance in Naturally Fractured Carbonate Reservoirs. *Fuel* **2013**, *108*, 502−514.
(8) Liu, Y.; Yao, K.-T.; Raghavenda, C. S.; Wu, A.; Guo, D.; Zheng, J.; Olabinjo, L.; Balogun, O.; Ershaghi, I. Global Model for Failure Prediction for Rod Pump Artificial Lift Systems. *SPE Western Regional & AAPG Pacific Section Meeting 2013 Joint Technical Conference* **2013**, SPE-165374-MS.
(9) Liu, Y.; Yao, K.; Liu, S.; Raghavendra, C. S.; Lenz, T. L.; Olabinjo, L.; Seren, B.; Seddighrad, S.; Dinesh Babu, C. G. Failure Prediction for Rod Pump Artificial Lift Systems. *SPE Western Regional Meetin* **2010**, SPE-133545-MS.
(10) Abdalla, R.; El Ela, M. A.; El-Banbi, A. Identification of Downhole Conditions in Sucker Rod Pumped Wells Using Deep Neural Networks and Genetic Algorithms (includes associated discussion). *SPE Pro & Oper* **2020**, *35* (02), 435−447.
(11) Wylde, J. J.; Fell, D. Scale Inhibitor Solutions For High Temperature ESP Lifted Wells In Northern California: A Case History Of Failure Followed By Success. *SPE International Oilfield Scale Conference* **2008**, SPE-113826-MS.
(12) Toma, P.; Vargas, E.; Kuru, E. Prediction of Slug-to-Annular Flow Pattern Transition (STA) for Reducing the Risk of Gas-Lift Instabilities and Effective Gas/Liquid Transport From Low-Pressure Reservoirs. *SPE Prod & Oper* **2007**, *22*, 339−346.
(13) Dunham, C. Summary of Presentations. *2013 27th ESP Workshop*. Artificial Lift R&D Council, April 2013. https://www.spegcs.org/media/files/files/cebfcc3a/2013-ESP-Workshop-Summary-of-Presentations.pdf.
(14) Alamu, O. A.; Pandya, D. A.; Warner, O.; Debacker, I. ESP Data Analytics: Use of Deep Autoencoders for Intelligent Surveillance of Electric Submersible Pumps. *Offshore Technology Conference* **2020**, OTC-30468-MS.
(15) Zhao, X. J.; Li, A.; Yang, F. Fault analysis of electric submersible pump based on using FTA. *Mod. Manuf. Technol. and Equip.* **2006**, *4*, 29−32.
(16) Li, J. J.; Zhang, G. S.; Song, S. Q.; Yu, Y. Y.; Duan, J. Development and application of macro-control diagramon oil production with electric submersible pump. *Fau. Bl. O. & G.* **2008**, *15* (6), 121−122.
(17) Zhang, P.; Chen, T.; Wang, G.; Peng, C. Ocean Economy and Fault Diagnosis of Electric Submersible Pump applied in Floating platform. *Int. J. e-Nav. Mari. Econ.* **2017**, *6*, 37−43.
(18) Xi, W. J. Research on fault diagnosis of electric submersible pumps based on vibration detection. Master's Thesis, China University of Petroleum (East China), Dongying, China, 2008.
(19) Wang, K.; Lei, B. Using B-spline neural network to extract fuzzy rules for a centrifugal pump monitoring. *J. Intell. Manuf.* **2001**, *12* (1), 5−11.
(20) Zhao, P. *Study on the vibration fault diagnosis method of centrifugal pump and system implementation*; North China Electric Power University: Beijing, 2011; pp 56−58.
(21) Tao, F.; Liu, G.; Xi, W. Research on the Fault Diagnosis of Excess Shaft Ran of Electric Submersible Pump. *Advances in Multimedia, Software Engineering and Computing Vol.1* **2011**, *128*, 509−513.
(22) Guo, D.; Raghavendra, C. S.; Yao, K.-T.; Harding, M.; Anvar, A.; Patel, A. Data Driven Approach to Failure Prediction for Electrical Submersible Pump Systems. *SPE Western Regional Meeting* **2015**, SPE-174062-MS.
(23) Peng, K. Fault diagnosis of electric submersible pump based on BP neural network. *J. Petrochem. Sci. Eng.* **2016**, *29* (1), 76−79.
(24) Wang, J. H. *Fault diagnosis of centrifugal oil pump based on BP neural network*; Hebei University of Engineering: Handan, China, 2013.
(25) Jansen Van Rensburg, N.; Kamin, L.; Davis, S. Using Machine Learning-Based Predictive Models to Enable Preventative Maintenance and Prevent ESP Downtime. *Abu Dhabi International Petroleum Exhibition & Conference* **2019**, SPE-197146-MS.
(26) 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. ESP Well and Component Failure Prediction in Advance using Engineered Analytics - A Breakthrough in Minimizing Unscheduled Subsurface Deferments. *Abu Dhabi International Petroleum Exhibition and Conference* **2019**, SPE-197806-MS.
(27) Adesanwo, M.; Denney, T.; Lazarus, S.; Bello, O. Prescriptive-Based Decision Support System for Online Real-Time Electrical Submersible Pump Operations Management. *SPE Intelligent Energy International Conference and Exhibition* **2016**, SPE-181013-MS.
(28) Adesanwo, M.; Bello, O.; Olorode, O.; Eremiokhale, O.; Sanusi, S.; Blankson, E. Advanced Analytics for Data-Driven Decision Making in Electrical Submersible Pump Operations Management. *SPE Nigeria Annual International Conference and Exhibition* **2017**, SPE-189119-MS.
(29) Gupta, S.; Nikolaou, M.; Saputelli, L.; Bravo, C. ESP Health Monitoring KPI: A Real-Time Predictive Analytics Application. *SPE Intelligent Energy International Conference and Exhibition* **2016**, SPE-181009-MS.
(30) Abdelaziz, M.; Lastra, R.; Xiao, J. J. ESP Data Analytics: Predicting Failures for Improved Production Performance. *Abu Dhabi International Petroleum Exhibition & Conference* **2017**, SPE-188513-MS.
(31) Bhardwaj, A. S.; Saraf, R.; Nair, G. G.; Vallabhaneni, S. Real-Time Monitoring and Predictive Failure Identification for Electrical Submersible Pumps. *Abu Dhabi International Petroleum Exhibition & Conference* **2019**, SPE-197911-MS.
(32) Sherif, S.; Adenike, O.; Obehi, E.; Funso, A.; Eyituoyo, B. \Predictive Data Analytics for Effective Electric Submersible Pump Management. *SPE Nigeria Annual International Conference and Exhibition* **2019**, SPE-198759-MS.
(33) Peng, L.; Han, G.; Pagou, A. L.; Zhu, L.; Ma, H.; Wu, J.; Chai, X. A Predictive Model to Detect the Impending Electric Submersible Pump Trips and Failures. *SPE Annual Technical Conference and Exhibition* **2021**, SPE-206150-MS.
(34) Zhang, P.; Chen, T.; Wang, G.; Peng, C. Ocean Economy and Fault Diagnosis of Electric Submersible Pump applied in Floating platform. *Int. J. e-Nav. Mari. Econ.* **2017**, *6*, 37−43.
(35) Yang, J.; Li, W.; Chen, J.; Sheng, L. Fault diagnosis of electric submersible pump tubing string leakage. *E3S Web of Conferences* **2021**, *245*, 01042.
(36) Li, L.; Hua, C.; Xu, X. Condition monitoring and fault diagnosis of electric submersible pump based on wellhead electrical parameters and production parameters. *Systems Science & Control Engineering* **2018**, *6*, 253−261.
(37) Wang, H.; Chen, P. Fault diagnosis of centrifugal pump using symptom parameters in frequency domain. *CIGR E. J.* **2007**, *9*, IT 07 005.
(38) Rajakarunakaran, S.; Devaraj, D.; Rao, K. S. Fault detection in centrifugal pumping systems using neural networks. *Inter. J. Model. Iden. Cont.* **2008**, *3*, 131.
(39) Khabibullin, R. A.; Shabonas, A. R.; Gurbatov, N. S.; Timonov, A. V. Prediction of ESPs Failure Using ML at Western Siberia Oilfields with Large Number of Wells. *SPE Russian Petroleum Technology Conference* **2020**, SPE-201881-MS.
(40) Frasier, K. E. A machine learning pipeline for classification of cetacean echolocation clicks in large underwater acoustic datasets. *PLOS Comp. Bio* **2021**, *17* (12), e1009613.
(41) Kotu, V. *Data Science: Concepts and Practice*, 2nd ed.; Morgan Kaufmann Publishers: Cambridge, MA, 2018.
(42) Peres-Neto, P. R.; Jackson, D. A.; Somers, K. M. How Many Principal Components? Stopping Rules for Determining the Number of Non-Trivial Axes Revisited. *Comp. Stat. Data Anal.* **2005**, *49*, 974−997.
(43) Jolliffe, I. T. Principal Component Analysis and Factor Analysis. *Principal Component Analysis* **1986**, 115−128.

https://doi.org/10.1021/acsomega.1c05881
ACS Omega 2022, 7, 17641−17651

(44) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**, 785−794.
(45) Nti, I. k.; Nyarko-Boateng, O.; Aning, J. Performance of Machine Learning Algorithms with Different K Values in K-fold Cross-Validation. *Inter. J. Info. Technol. Comp. Sci.* **2021**, *13*, 61−71.
(46) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M.. Optuna: A Next-generation Hyperparameter Optimization Framework. *arXiv*, 2019.
(47) Marins, M.; Barros, B.; Santos, I.; Barrionuevo, D.; Vargas, R.; de M. Prego, T.; de Lima, A.; de Campos, M.; Da Silva, E.; Netto, S. Fault detection and classification in oil wells and production/service lines using random forest. *J. Pet. Sci. Eng.* **2021**, *197*, 107879.

https://doi.org/10.1021/acsomega.1c05881
ACS Omega 2022, 7, 17641−17651