COBEM 2025
28TH INTERNATIONAL CONGRESS OF MECHANICAL ENGINEERING

ABCM
Brazilian Society of Mechanical Sciences and Engineering
28th ABCM International Congress of Mechanical Engineering
November 09-14, 2025, Curitiba-PR, Brazil

COB-2025-1258
HIDDEN MARKOV MODELS USAGE IN ELECTRICAL SUBMERSIBLE PUMPS DATA TO FAILURE DETECTION

Paulo Yoshio Kuga
Alberto Luiz Serpa
School of Mechanical Engineering - Universidade Estadual de Campinas (UNICAMP), R. Mendeleyev, 300 - Cidade Universitária - Campinas - SP
p204451@dac.unicamp.br
alserpa@unicamp.br

Abstract. *Currently, unsupervised learning methods are still unexplored in the literature on Electric Submersible Pumps (ESP). The most used approach is to fit supervised learning models on labeled data. Therefore, the objective of this research is to establish a method that can highlight anomalies in data from 57 runs provided by real field data with an unsupervised learning method. This data is noisy and incomplete, leading to inconsistencies when labeling failures. To this end, in this article, the main technique used on the data is a modified Z-score, which uses data filtered by an exponential moving average, the expanding median. This procedure seeks for robustness to outliers, and the expanding standard deviation. This methodology allows comparing pumps with different behaviors in the same reference. Later, it is propposed to use Gaussian Hidden Markov Models (GHMM), a probabilistic model of hidden state transition based on observation of the pump's measured data. It is expected that these states may be an indicator of normality or abnormality of the data provided. In this sense, the model uses a L^2 normalization of the data, seeking to represent the distance of the inputs in relation to the mean. For this distribution, a Gaussian Mixture Model (GMM) is used, with each Gaussian representing a state, providing an initial guess for GHMM. Then, the GHMM is initialized with GMM parameters and it is possible to notice a very high incidence of anomalous states near to the manually registered failure, even before this registration is executed. Therefore, it is concluded that GHMM is a useful tool for detecting anomalies in time series of ESP's.*

Keywords: *ESP failures, Failure Detection, Unsupervised Learning, Hidden Markov Models*

1. INTRODUCTION

In recent years, the development of machine learning techniques has allowed new ways to detect anomalies and failures. This led to new approaches to the improvement of predictive maintenance in several machines. Along the years, plenty of studies have already been done in the context of Electrical Submersible Pump (ESP) operations, intending to use these new methods to detect failures and anomalous behaviours. For this task, most of the literature uses supervised learning models, implying there is a priori knowledge of failures. This means that the data used to train the models is already classified between normal and abnormal.

The models that identify the anomalies are mostly derived from the Support Vector Machine model (SVM), for example, the works of Awsan (2023) and Yang *et al.* (2022b). Both of them have data obtained from a laboratory, which is a controlled environment where it is possible to have greater control over and has greater distinction between characterizing a fault or healthiness situation. The work of Lastra and Xiao (2022) uses combined models, both SVM based models, random forests and K-Nearest Neighbors. They use real operating data in their work, but data labelling was made by experts, which, for a reasonable amount of data shall require a significant amount of work. However, the characterization of failures and anomalies through already known or expertise labelling may not be useful in cases where the operator might not have knowledge about the failure or when it has already happened. Still, Yang *et al.* (2022a) use an outlier technique to classify outliers in their work, but for the classification, they still uses supervised learning, in a way that time-series relations are not considered. Therefore, a method that does not require a priori knowledge and is able to infer time-series relations is desirable.

In the literature, two common types of datasets are noticed: ESP sensor data and field production data. The first one is related to measured features in pumps, such as pressures, temperatures and electrical ones. These measurements are provided by several ESP sensors, allowing real-time monitoring. As for production data, relations such as gas-liquid ratio, water-oil ratio, gas-water ratio, basic sediments and water and pressure, viscosity, and temperature properties of the extracted oil are measured. This would allow to identify failures analysing time-series relations that can predict or identify the health of the pump. Most of the literature uses ESP sensor data, instead of production data, since it might be sensitive to companies to provide this type of data.

The data used in this article is based in real data operation. They are composed by two different datasets. The first one is a database of ESP measurements in a frequency of one hour per entry, which have several pumps properties listed, such as pressures, temperatures, vibrations and electrical properties. The second one is a failure spreadsheet detailing when the failure was reported. Since pumps were not in a controlled environment, where failures could be artificially induced and precisely labelled, the failures in this dataset are due to real operation, often making them less apparent. Nonetheless, they might not exhibit clear patterns or distinct markers, increasing the difficulty of their detection by specialists.

In this article, an unsupervised approach will be applied, establishing an approach to classify anomalies without previous knowledge. The provided data was explored, considering the work of Montgomery *et al.* (2024), about time series, and Kay (2006), on stochastic processes, Gerón (2019) and Gori *et al.* (2023) to provide insights about possible approaches, trying to understand the nature of the time-series and if their distribution is well-behaved. Takacs (2018) is a source about general topics of ESP's, providing a general understanding of measured properties.

It is difficult to find in Machine Learning literature an unsupervised classification method of time-series, because usually, the methods are used in pattern detection without temporal relations. Time-series have a time-relation, where usual classification or prediction problems does not have, as shown in Gori *et al.* (2023). However, through a more general research, it was possible to notice that Hidden Markov Models, established by Baum *et al.* (1970) were used in contexts of sequence de-codification, such as speech recognition (Rabiner (1989)), gene de-codification (Stanke and Waack (2003)), identification of financial patterns (Mamon and Eliott (2010)) and applications in control theory (Eliott *et al.* (2008)).

In this article, an unsupervised approach will be carried with the Gaussian Hidden Markov Model (HMM) (Lebedev *et al.* (2018)), trying to establish an approach to classify anomalies without a previous knowledge. In failure detection contexts, this method was already used to detect anomalies in machines, such as in the article by Smyth (1994) and in the article by Zhou *et al.* (2010). In this way, the time series trains the HMM and attempt to characterize sections of the time series. The objective is to characterize not only the failure point, but its window, trying to indicate to the operators that a failure may be occurring, or it is close to occurring, as indicated by the state of the series.

2. PROBLEM DESCRIPTION

The main objective in this work is to identify anomalies in the submerse centrifugal pumping process using real operation data from the systems involved. This data was provided by a petrol company, which consists of 57 sets of run information, containing properties of the pumps.

A box plot, in Fig. 1 was made to better understand their behaviour. Data was normalized, since their scales were different. It is possible to notice the features do not have identical distributions, neither the outliers follows a same pattern. Current Mean, as a example, has a single point of extreme value (100% related value), influencing its entire box plot. Missing data due to process limitations were not plot into Fig. 1. Some approaches might interpolate a number for them, but this is going to be explored into the next sections.

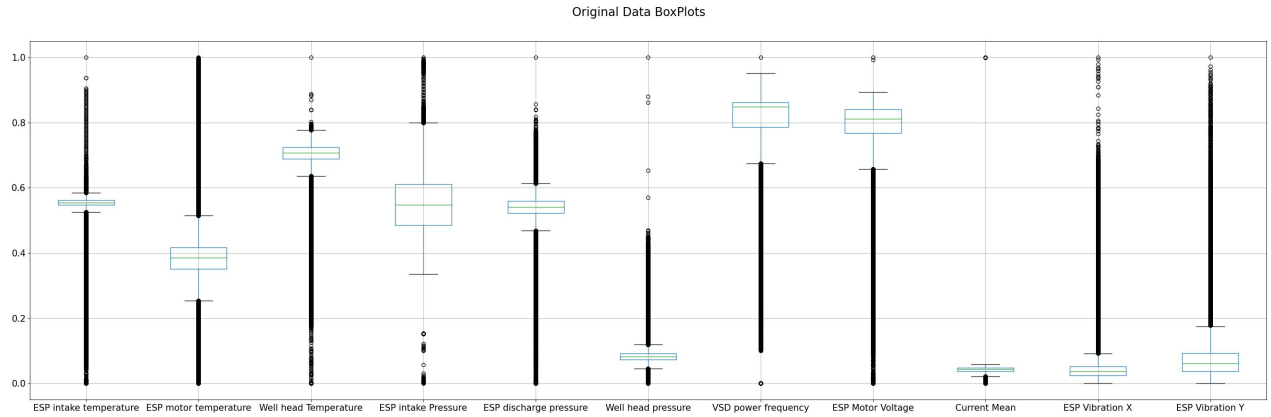

Figure 1: Box-plot of measured features

It is possible to notice that the features do not have a similar concentration of quartiles, and their dispersions are high, leading to difficulties into characterizing the behaviour between normal or abnormal.

Additionally, a failure information sheet was available. One specific difficulty is that the sensor data were recorded by hour frequency and the failure information is not precisely recorded into the same moment that the failure happened. This endorses the fact that, for this dataset, a supervised approach would not be the most suitable.

3. PROPOSED METHODS AND PROCEDURES FOR ANOMALIES DETECTION

3.1 Data filtering

To minimize noise to interfere in the measurement, it is proposed to filter the data before performing the standardization, as a way to smooth outliers. In this work, is is used the Exponential Mean Average (EWM), presented in Kotu and Bala (2019), for smoothing. This filter is described as:

$$S_k = (1-\alpha)S_{k-1} + \alpha x_k \tag{1}$$

where S is the filtered signal, k is the current entry index and α is the weight given to each x_k entry. For a determined span, α is given as $\alpha = \frac{2}{\text{span}+1}$. In this work, the span was set in 24 hours, that corresponds to 24 measured points.

3.2 Modified Z-score

As illustrated in Fig. 1, the dataset exhibits heterogeneous distributions and varying scales among its features. Given that the variables are expressed in different units, it is convenient to normalize them to a common scale to enable the analysis. Although min-max normalization achieves this goal, it is sensitive to new data points and may require rescaling upon the inclusion of additional observations. To address this limitation, this study adopts a modified Z-score transformation, which provides a dimensionless representation based on the standard Gaussian distribution and is more robust to incremental updates. This transformation is described as:

$$Z = \frac{x-\mu}{\sigma} \tag{2}$$

where Z measures the ratio between the difference of the sample x and the mean μ, and the standard deviation σ. This standardization provides a measure of discrepancy between the mean of the data, as a reference, normalized by its standard deviation (Kay (2006)).

Along the time, for continuously registering data, the mean and the standard deviation are changed by new entries. Then, it is possible to update the mean as:

$$\mu_k = \frac{1}{k}x_k + \left(1-\frac{1}{k}\right)\mu_{k-1} \tag{3}$$

A big value of x_k can interfere on the measure of mean μ_k, impacting the measure of Z_k, which is the correspondent Z-value of the k entry. Then, a more robust (less sensitive to outliers) measure of the constant reference may be assumed as the median ($M_k = median(x_k)$), which is the middle value of the set. This measure is independent, because even if the outliers unbalances the mean, the median is dependent from the number of entries in the set, therefore, being more robust.

Also considering the continuous flow of entries, the standard deviation shall be updated as:

$$\sigma_k = \left(\frac{1}{k}x_k^2 + \left(1-\frac{1}{k}\right)\left(\sigma_{k-1}^2 + \mu_{k-1}^2\right) - \mu_k^2\right)^{\frac{1}{2}} \tag{4}$$

where this formula is related to the expectancy of the square of the variable. A more detailed approach can be found in Knuth (1997).

Combining these equations, considering the original definition of Eq. (2), it is possible to rewrite the Z-score as an entry dependent equation, where, when applied to dataset, with the continuous flow of new values, Z_k is calculated as:

$$Z_k = \frac{S_k - M_k}{\sigma_k} \tag{5}$$

Considering the fact that pumps usually have zero values when shutdown, different values of M_k and σ_k are considered for the online and offline pumps. For missing data, the assumed value is zero, since $Z_k = 0 \rightarrow x_k = M_k$, which does not infer any deviation about a data that is unknown.

3.3 L^2 Norm and Gaussian Mixture Model

A L^2 norm is proposed as:

$$\|\boldsymbol{Z}_i\|_2 = \left(\sum_{j=1}^{n} z_j^2\right)^{\frac{1}{2}} \tag{6}$$

where z_j is a vector component of the i entry in the dataset and n is the number of components.. The L^2 norm can be understood as the square root of the total signal energy, being used as a univariate representation of the signal.

Given that the Z-score is derived from a Standard Gaussian distribution, each feature is ideally assumed to follow such a distribution. Consequently, higher Z-score L^2 values would correspond to lower probabilities, and they can be considered indicative of potential anomalies.

Following this, it is possible to fit the resultant distribution of L^2 norm ($\mathcal{D}_{\|\boldsymbol{Z}_k\|_2}(x)$) into a Gaussian Mixture Model (GMM), defined as a sum of Gaussians according to:

$$\mathcal{D}_{\|\boldsymbol{Z}\|_2}(x) = \sum_{i=1}^{p} w_i \mathcal{N}(\mu_i, \sigma_i)(x) \tag{7}$$

where p is the number of states, $\mathcal{N}(\mu_i, \sigma_i)$ are gaussians with mean μ_i and standard deviation σ_i and w are the weights of the gaussians. Each i component is associated with a state of the Hidden Markov Model (HMM), being the GMM a way to provide a first guess to the model.

To fit the L^2 Norm distribution to the proposed model, an instance of the Expectation-Maximization (EM) algorithm (Dempster *et al.* (1977)) is used. Firstly, the parameters are initialized randomly. Then, γ_{is} is defined as the posterior probability as an entry is part of the mixture:

$$\gamma_{is} = \frac{w_s \mathcal{N}(x_i|\mu_s, \sigma_s)}{\mathcal{D}_{\|\boldsymbol{Z}\|_2}(x_i)} \tag{8}$$

where s is one of the gaussians from the GMM model. Equation (8) characterizes the Expectation-step of GMM. For the Maximization-step, the weight and the properties of the GMM are updated as:

$$(w_s, \mu_s, \sigma_s)_{t+1} = \left(\frac{1}{N} \sum_{i=1}^{N} \gamma_{ik} \quad , \quad \frac{\sum_{i=1}^{N} \gamma_{ik} x_i}{\sum_{i=1}^{N} \gamma_{ik}} \quad , \quad \frac{\sum_{i=1}^{N} \gamma_{ik} (x_i - \mu_k)^2}{\sum_{i=1}^{N} \gamma_{ik}} \right) \tag{9}$$

where N is the total number of entries related to x.

In other words, the weight, the mean and the standard deviation of each state is updated as the mean of the probabilities, the weighted average of probabilities through the entry and the averaged mean of probabilities through the deviation, respectively.

3.4 Gaussian Hidden Markov Model

The Gaussian Hidden Markov Model is a model that can infer states over a time-series, providing classifications over each entry. The concept behind this method is the optimization of a probabilistic model that adapts the probability of visualizing a certain value (the observation of the phenomenon) in relation to a hidden state, where the transitions between states can be described by Markov chains. In the present research, we assume the state to be a classification of anomaly, which leads its prediction to the identification of the abnormal state.

This model is defined by the parameters $(\boldsymbol{A}_{ij}, \boldsymbol{B}_i, \boldsymbol{\Pi}_i)_t$, where $\boldsymbol{A}$ is the state transition matrix, of dimension $p \times p$, $\boldsymbol{B}$ is the vector of state observation likelihood, and $\boldsymbol{\Pi}$ the initial probability of a state being set, both with $p \times 1$ size. In this notation, i and j represents the state numbers. In the Gaussian Hidden Markov Model (GHMM) the probability of the observation associated to a state is Gaussian, and for each state there is a Gaussian probability density function associated ($P(O_t|S_t = i) \sim \mathcal{N}(\mu_i, \sigma_i)$). The parameters are defined as:

$$(\boldsymbol{A}_{ij}, \boldsymbol{B}_i, \boldsymbol{\Pi}_i)_t = \begin{pmatrix} P(S_t = j|S_{t-1} = i) & P(O_t|S_t = i) & P(S_0 = i) \end{pmatrix} \tag{10}$$

where S_t is a state, and O_t is an observation, both in time t.

For solving this problem, Baum-Welch algorithm is used, which is a special case of Expectation-Maximization (EM) algorithm to find unknown parameters of Hidden Markov Model (Baum *et al.* (1970)). Firstly, all matrices and vectors are initialized randomly. Then, the Forward-Backward algorithm computes the probability of being in each state, in all time entries. The vector $\boldsymbol{\alpha}$ defines the probability of the sequence until the time t to happens, considering the state related to that line. $\boldsymbol{\beta}$ represents the probability of the most probable sequence for each state. Therefore, considering operator $\odot$ as the Hadamard product, it is possible to write:

$$(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = \begin{pmatrix} (\boldsymbol{A} \cdot \boldsymbol{B}(O_t)) \odot \boldsymbol{\alpha}_{t-1} & (\boldsymbol{A} \cdot \boldsymbol{B}(O_t)) \odot \boldsymbol{\beta}_{t+1} \end{pmatrix} \tag{11}$$

Equation (11) defines the Forward-Backward parameters. This is the Expectation (E) step, where, $\boldsymbol{\alpha}_0 = (\boldsymbol{B}(O_t) \odot \boldsymbol{\Pi})$ and $\boldsymbol{\beta}_n = \mathbf{1}_s$, where $\mathbf{1}$ is a vector of ones. Next, the posterior probability vector is defined as:

$$\boldsymbol{\gamma}_t = (\boldsymbol{\alpha}_t \odot \boldsymbol{\beta}_t) \cdot (\boldsymbol{\alpha}_t^\top \boldsymbol{\beta}_t)^{-1} \tag{12}$$

Leading to state the probability matrix of next time state transitions as:

$$\boldsymbol{\Gamma}_t = (\boldsymbol{A}\boldsymbol{\alpha}_t) \odot (\boldsymbol{B}(O_{t+1})\boldsymbol{\beta}_{t+1}^T) \longrightarrow \boldsymbol{\Xi}_t = \frac{\boldsymbol{\Gamma}_t}{||\boldsymbol{\Gamma}_t||_1} \tag{13}$$

Then, Maximization (M) step begins, with the update equation for the parameters $\boldsymbol{A}$, $\boldsymbol{B}$ $\boldsymbol{\Pi}$. For $\boldsymbol{A}_{t+1}$:

$$\boldsymbol{A}_{t+1} = \sum_{i=0}^{t} \left(\boldsymbol{\Xi}_i \odot \boldsymbol{\gamma}_i^{-1}\right) \tag{14}$$

$\boldsymbol{B}$ gaussians are updated similarly to provided into Eq. (9), however, using the related $\boldsymbol{\gamma}_t$ components from HMM model to find the μ_i and σ_i parameters. Later, $\boldsymbol{\Pi}_{t+1}$ is updated as $\boldsymbol{\gamma}_0$. With the parameters set, it is possible to infer the states for $\|\boldsymbol{Z}_k\|_2$ data using the Viterbi Algorithm (Viterbi (1967)), already implemented in $hmmlearn$. This algorithm infers the most probable sequence of states considering the observations. An explanation of the algorithm is provided in Rabiner (1989). The reader might notice that the weight parameter is not used in GHMM. This is because the model considers subpopulations and not compositions of Gaussians. However, the weight can be used as first guess of $\boldsymbol{\Pi}_0$.

4. RESULTS

The results of this works considers the formulation described and coding using Python scripts. Firstly, the signals were filtered with a span of 24 hours, standardized and missing data was interpolated. The processed features are motor temperature, current mean and voltage, well head pressure and temperature, intake pressure and temperature, pump discharge pressure, Variable Speed Drive (VSD) power frequency, and vibration in X and Y axis of the pump. The L^2 norm is only applied to online data and fitted with the GMM.

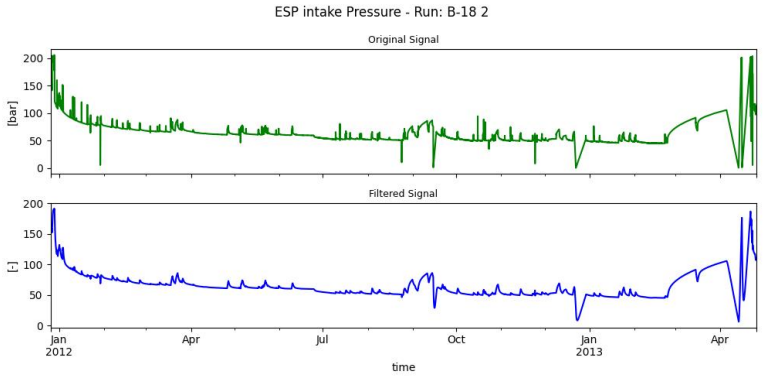

Figure 2: Original (green) and Filtered Signal (blue).

The modified Z-score was calculated for each data entry. Fig. 3 illustrates the process of obtaining the Z-score for ESP intake pressure. First, the filtered signal is adjusted by subtracting the expanding median (light blue), which represents the progressively accumulated median of past values. Then, this difference is divided by the expanding standard deviation (yellow), which accounts for variations in the data over time. The resulting Z-score (red) highlights fluctuations in the signal relative to its historical behaviour, making it easier to identify deviations and potential anomalies.

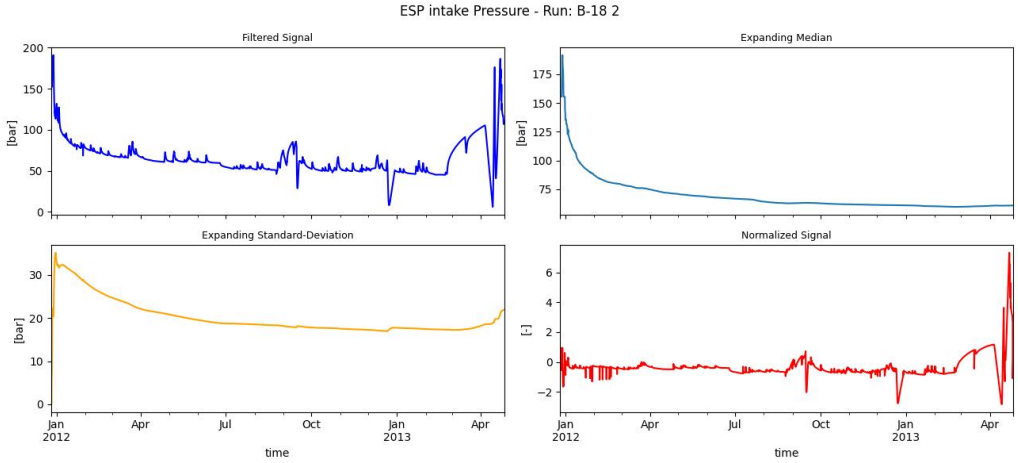

Figure 3: Filtered signal (blue), expanding median (light blue), expanding standard deviation (yellow) and Normalized Signal (red).

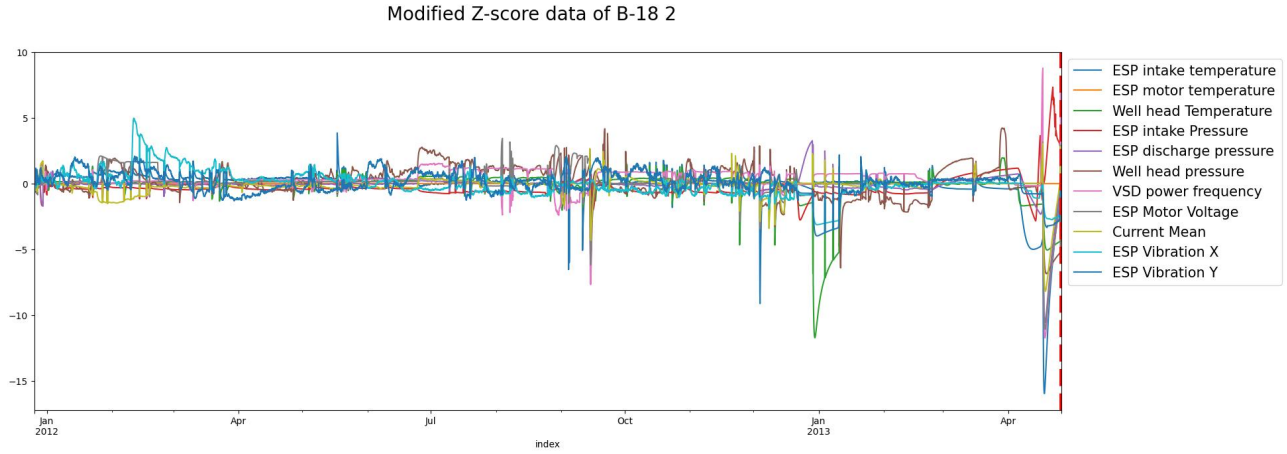

Figure 4: The modified Z-score data for pump run B-18 2.

Calculating the Z-Score for all features, it is possible to obtain the Fig. 4. Notably, an anomalous behaviour appears near the end of the figure. The red dashed line marks the moment when the failure was recorded in the failure spreadsheet.

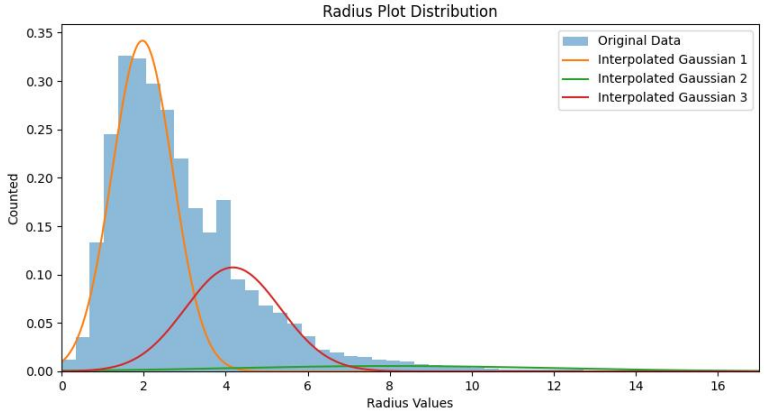

Figure 5: Histograms for the Radius and they correspondent Gaussian Mixture Models

The GMM is interpolated, and the results are shown in Fig. 5. The number of states s should be chosen considering interpretability and the minimization of the Bayesian Information Criteria (BIC). This criterion is a measure of how well the model fits to the distribution (Bishop (2006)). However, with an increase of the number of states, this measure is expected to reduce, what would indicate overfitting. Therefore, the two numbers to be tested were 2 and 3. For 2, BIC is

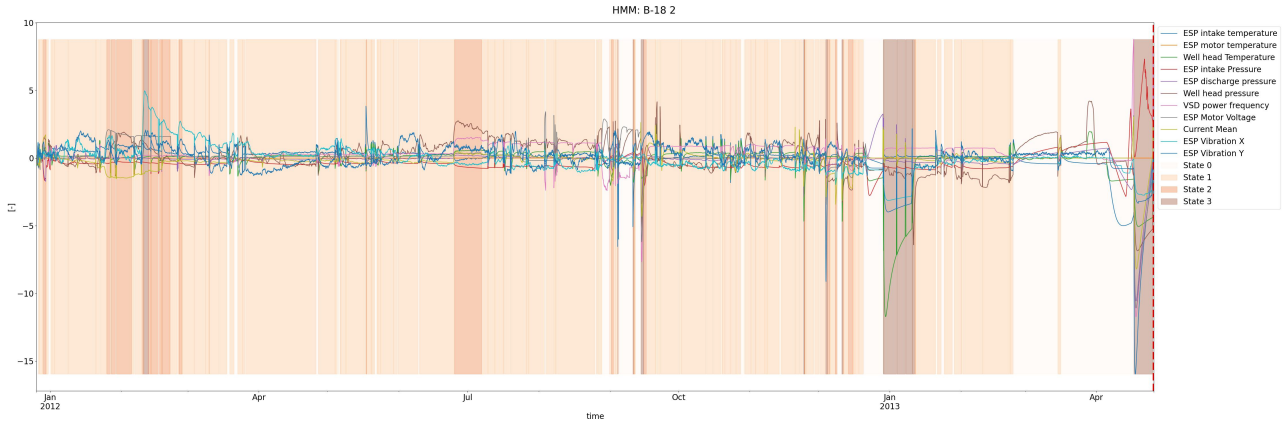

Figure 6: B-18 2 run inferred states from January 2012 to April 2013. Plot based on original features.

3217476, and for 3, BIC is 3114622. Since the criterion is the minimization, and the BIC of 3 is lower than 2, 3 fits well than 2 as the number of states.

Next, the GHMM model was trained with GMM obtained parameters. For the model, an arbitrary State 0 is indicates whether the pump was off or not, and is not a result of the GHMM. The State 1 characterizes normal behaviour along the pump operation and slight anomalies of the pump are represented in State 2. The State 3 is related to severe anomalies. An example is run B-18 2 in Fig. 6, where is possible to notice that State 3 persists before the failure, in the dashed red line. There is a persistence of State 3, that indicates a severe anomaly before the failure.

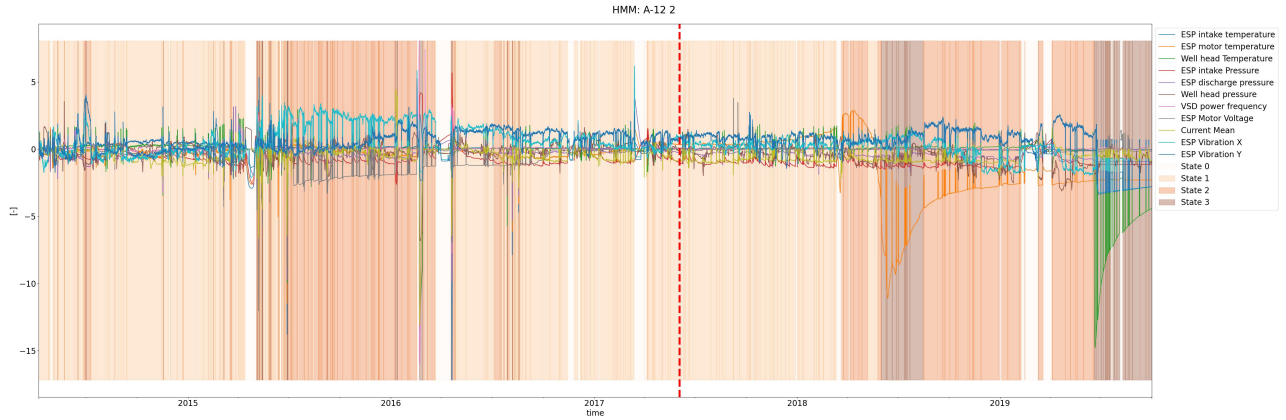

Figure 7: A-12 2 run inferred states from 2014 to 2019. Plot based on original features.

Run A-12 2, shown in Fig. 7, at the moment indicated by the red dashed line, there is no clear persistence of State 3, but instead, a significant persistence of State 2 during the previous year, 2016, and later—towards the end of the time series, both States 2 and 3 remain active for extended periods.

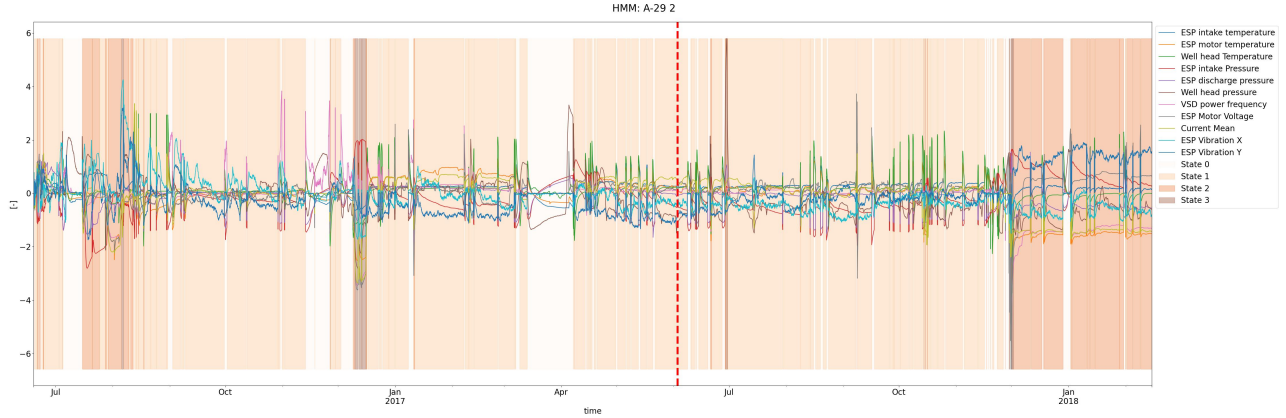

Figure 8: A-29 2 run inferred states from July 2016 to January 2018. Plot based on original features.

Run A-29 2, illustrated in Fig. 8, shows a case where the State 1 is the persistent state in the failure. However, prior to the end of the time series, only State 2 shows notable persistence, with no significant appearance of State 3.

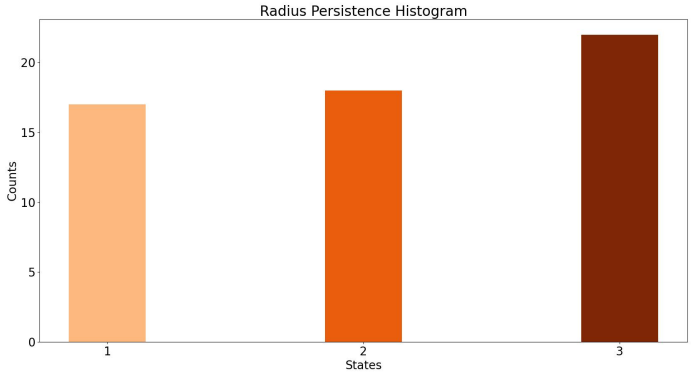

Figure 9: Radius Histogram for seed generated by number 971215

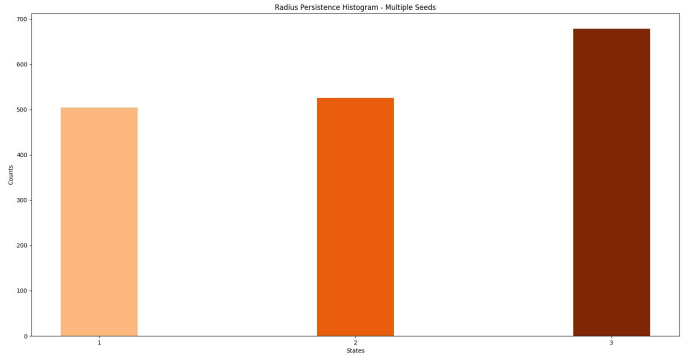

Figure 10: Radius Histogram for multiple random seeds

In these examples, runs were selected to illustrate the GHMM on pump data. After training and running the model for all 57 runs, it is possible to identify the most frequent states occurring within 24 hours of online pump data before a reported failure, as illustrated in Fig. 9. The models were trained using the seed 971215 to initialize the random number generation from $numpy$ library. It is possible to notice that, within the 24 hours prior to the failure, the persistence of States 2 and 3 is higher than State 1.

Then, to account for variations associated with the randomness of parameters initializations, 30 training runs were performed with different seeds from 19971215 to 20210505. The results remained consistent, showing that the proportion of States 1 and 2 is higher, as shown in Fig. 10.

5. DISCUSSION

ESP literature has few articles about anomaly detection without a labelling process. Indeed, previously cited authors, as Lastra and Xiao (2022), Awsan (2023) and Yang *et al.* (2022b) already have knowledge about the failure and the proposed methods are about supervised learning. In literature, Yang *et al.* (2022a) have already used unsupervised techniques considering outliers, and then, using the technique to label data, aiming to use in supervised learning. Nonetheless, their approach is very different from the one presented here, because the unsupervised model they used relies on outlier labelling, and the proposed method relies on time-series sequence classification.

In this work, the persistence plot helps to identify when a specific state remains stable prior to a reported failure, which, in practical scenarios, can be valuable. However, as shown, there are some cases in which, at the moment of the reported failure, the persistent state is State 1 — associated with normal conditions. Even though anomalous states may become dominant by the end of the run, the model was not able to clearly indicate a failure. In this sense, the model is capable of detecting anomalies primarily based on numerical deviations.

It is noticeable that the likelihood of a state occurring is still strongly influenced by the associated Z-score L^2 norm and the Gaussian distribution it most closely matches. However, the difference between only classifying a failure considering the Z-score is the fact that GHMM can model the transitions between their associated gaussians. And, since L^2 norm

can synthetize the intensity of multiple features, if the dataset has enough data to calculate the distribution, the proposed procedure in this article should work.

Another important aspect is that the failure logs were compiled by field operators, who may not have recorded the failure at the exact time they noticed it. It is possible to notice that, for run B-18 2 (Fig. 6) can infer the failure before the report. But, it was cogitated that, in some cases, the failure might have been logged not due to an anomalous pump behaviour, but rather due to anomalies in production output. In this case, the model may not be capable of identifying this failures. This is not a limitation of the model, but from the data.

As stated in the GMM and GHMM model discussions, their parameters are initialized randomly. Therefore, a fixed seed is chosen for assure the consistency of the results. For both models, there were concerns regarding the stability of their EM algorithms, when starting from a random seed. In the case of GMM, the concern was about the distribution of abnormal data may be overlapping the normal data, implying a poor model fit. Ideally, after scaling, the anomaly should be represented by a distinct gaussian from the one modelling the normal behaviour. The concern about GHMM was the model to not converge into a consistent distribution of states. However, in both cases were observed stable results over the states persistences, even with varying initialization seeds.

6. CONCLUSION

This work presented an anomaly detection method using Gaussian Hidden Markov Models (GHMM), an unsupervised learning method that tries to provide an approach that considers relations into the time-series. A modified Z-score was used to measure the difference from the constantly calculated expanding median, being scaled by an expanding standard deviation, considering the continuous acquisition of data. Comparing to the literature, the novelty in this approach is that this method can deal with a multiplicity of data without requiring labelling.

Despite limitations, the formulation is still capable of indicating or providing a characterization of pump condition without requiring prior classification. Furthermore, it offers flexibility to handle various types of data and can be adapted to other contexts. In this work, the use of a simplified norm-2 transformation was preferred over adopting the full multidimensional structure of the Gaussian Hidden Markov Model (GHMM). However, as a suggestion for future research, the use of the Gaussian Mixture Model Hidden Markov Model (GMM-HMM), in which Gaussian mixtures represent the hidden states, could be explored.

The proposed method was effective in identifying patterns and characterizing operational states in ESP data, even in the absence of prior labelling. Its unsupervised nature, combined with the flexibility of the HMM framework, allowed for meaningful insights into pump behaviour. This confirms the viability of the approach as a valuable tool for early anomaly detection and condition monitoring in real-world scenarios.

7. ACKNOWLEDGEMENTS

We gratefully acknowledge the support of EPIC – Energy Production Innovation Center, hosted by the University of Campinas (UNICAMP) and sponsored by FAPESP – São Paulo Research Foundation (2017/15736-3 and 2024/00056-0). We acknowledge the support and funding from Equinor Brazil and the support of ANP (Brazil's National Oil, Natural Gas and Biofuels Agency) through the R&D levy regulation - Project reference number 24177-8. Acknowledgements are extended to the Center for Energy and Petroleum Studies (CEPETRO), School of Mechanical Engineering (FEM), Artificial Lift and Flow Assurance research group (ALFA), Institute of Chemistry and Brazilian Center for Research in Energy and Materials (CNPEM).

Also, we would like to thanks Dr. Heitor Nigro Lopes for the suggestions in the text of this article.

8. REFERENCES

Awsan, M., 2023. "Data driven-based model for predicting pump failures in the oil and gas industry". *Engineering Failure Analysis*, Vol. 145, p. 107019. ISSN 1350-6307. doi:https://doi.org/10.1016/j.engfailanal.2022.107019. URL https://www.sciencedirect.com/science/article/pii/S1350630722009864.

Baum, L.E., Petrie, T., Soules, G. and Weiss, N., 1970. "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains". *The Annals of Mathematical Statistics*, Vol. 41, No. 1, pp. 164 – 171. doi:10.1214/aoms/1177697196. URL https://doi.org/10.1214/aoms/1177697196.

Bishop, C.M., 2006. *Pattern Recognition and Machine Learning*. Springer New York, NY, 1st edition. ISBN 978-0-387-31073-2.

Dempster, A.P., Laird, N.M. and Rubin, D.B., 1977. "Maximum likelihood from incomplete data via the EM algorithm". *J. Roy. Statist. Soc. Ser. B*, Vol. 39, No. 1, pp. 1–38. ISSN 0035-9246. URL http://links.jstor.org/sici?sici=0035-9246(1977)39:1<1:MLFIDV>2.0.CO;2-Zorigin=MSN. With discussion.

Eliott, R.J., Moore, J.B. and Aggoun, L., 2008. *Hidden Markov Models - Estimation and Control*. Springer New York,

1st edition. ISBN 978-0-387-84854-9.

Gerón, A., 2019. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc., 2nd edition. ISBN 978-1-492-03264-9.

Gori, M., Betti, A. and Melacci, S., 2023. *Machine Learning - A Constraint-Based Approach*. Elsevier, 2nd edition. ISBN 978-0-323-89859-1.

Kay, S.M., 2006. *Intuitive Probability and Random Processes using MATLAB®*. Springer New York, NY, 1st edition. ISBN 978-0-387-24158-6.

Knuth, D.E., 1997. *The art of computer programming: Volume 1: Fundamental algorithms*. Addison Wesley, Boston, MA, 3rd edition.

Kotu, V. and Bala, D., 2019. *Data Science - Concepts and Pratice*. Morgan Kaufmann, 1st edition. ISBN 978-0-128-14761-0.

Lastra, R.A. and Xiao, J., 2022. "Machine learning engine for real-time esp failure detection and diagnostics". Vol. Day 2 Wed, October 26, 2022, p. D021S007R001. doi:10.2118/206935-MS. URL https://doi.org/10.2118/206935-MS.

Lebedev, S., Lee, A., Danielson, M. and Varoquax, G., 2018. "Hmmlearn/hmmlearn: Hidden markov models in python, with scikit-learn like api". URL https://github.com/hmmlearn/hmmlearn.

Mamon, R.S. and Eliott, R.J., eds., 2010. *Hidden Markov Models in Finance*. Springer New York, 1st edition. ISBN 978-0-387-71163-8.

Montgomery, D.C., Jennings, C.L. and Kulahci, M., 2024. *Introduction to Time Series Analysis and Forecasting*. John Wiley Sons, 3rd edition. ISBN 978-1-394-18670-9.

Rabiner, L., 1989. "A tutorial on hidden markov models and selected applications in speech recognition". *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257–286. doi:10.1109/5.18626.

Smyth, P., 1994. "Hidden markov models for fault detection in dynamic systems". *Pattern Recognition*, Vol. 27, No. 1, pp. 149–164. ISSN 0031-3203. doi:https://doi.org/10.1016/0031-3203(94)90024-8. URL https://www.sciencedirect.com/science/article/pii/0031320394900248.

Stanke, M. and Waack, S., 2003. "Gene prediction with a hidden markov model and a new intron submodel". *Bioinformatics*, Vol. 19, pp. 215–225. ISSN 1367-4803. doi:10.1093/bioinformatics/btg1080.

Takacs, G., 2018. *Electrical Submersible Pumps Manual*. Gulf Professional Publishing, 2nd edition. ISBN 978-0-128-14570-8.

Viterbi, A., 1967. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". *IEEE Transactions on Information Theory*, Vol. 13, No. 2, pp. 260–269. doi:10.1109/TIT.1967.1054010.

Yang, P., Chen, J., Wu, L. and Li, S., 2022a. "Fault identification of electric submersible pumps based on unsupervised and multi-source transfer learning integration". *Sustainability*, Vol. 14, No. 16. ISSN 2071-1050. doi:10.3390/su14169870. URL https://www.mdpi.com/2071-1050/14/16/9870.

Yang, P., Chen, J., Zhang, H. and Li, S., 2022b. "A fault identification method for electric submersible pumps based on dae-svm". *Shock and Vibration*, Vol. 2022, No. 1, p. 5868630. doi:https://doi.org/10.1155/2022/5868630. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/5868630.

Zhou, Z.J., Hu, C.H., Xu, D.L., Chen, M.Y. and Zhou, D.H., 2010. "A model for real-time failure prognosis based on hidden markov model and belief rule base". *European Journal of Operational Research*, Vol. 207, No. 1, pp. 269–283. ISSN 0377-2217. doi:https://doi.org/10.1016/j.ejor.2010.03.032. URL https://www.sciencedirect.com/science/article/pii/S0377221710002596.

9. RESPONSIBILITY NOTICE

The authors are solely responsible for the printed material included in this paper.