

Vol. 2 No. 2 (2024)

INTERNATIONAL JOURNAL OF INDUSTRIAL AND PRODUCTION ENGINEERING (IJIPE)

JOURNAL HOMEPAGE: https://journals.unizik.edu.ng/index.php/ijipe/

Machine Condition Monitoring with Gaussian Mixture Model-Probabilistic Clustering for Pumps

Newyear Oghenekome Ezeghare, Eyere Emagbetere, Peter Ufuoma Anaidhuno

Department of Mechanical Engineering, Federal University of Petroleum Resources, Effurun, Nigeria

newforreal@yahool.com, emagbetere.eyere@fupre.edu.ng, anaidhuno.ufuoma@fupre.edu.ng

Abstract

Pump systems play a critical role in various industries, and ensuring their reliability and timely maintenance is paramount. This research investigated the application of Gaussian Mixture Models (GMM) to condition monitoring and fault detection for pumps. The research begins by collecting and pre-processing extensive pump data from the Warri Refinery Petroleum Company, encompassing 374 samples of pre-processed vibration signals on various operating conditions and fault scenarios. The data were statistically analyzed, and then GMM, renowned for their ability to model complex data distributions, and K-means, a traditional clustering technique, was employed to cluster the dataset. The GMMs and K-means clustering were implemented by using suitable libraries on Python 3.0 software. The optimum hyper-parameters were determined using a grid search method. Then the clusters were created using both models, and their performance was investigated by calculating the silhouette and BIC scores. The obtained clusters were then assessed for their uniqueness to identify fault types and other pump conditions. Based on a hyper-parameter grid search, the optimum number of clusters was found to be 6 and a random state of 54. Comparative analyses revealed that GMMs outperform K-means having silhouette scores of 0.68 and 0.51, respectively. The application of GMMs showcases their potential for proactive maintenance, by identifying different anomalies such as those resulting from faulty sensors, and outboard and inboard faults. This study demonstrates the effectiveness of GMM-based clustering in accurately identifying different operational states and detecting anomalies within pump data. The application of GMMs provides a practical and effective means of enhancing pump system reliability and maintenance strategies.

1.0 INTRODUCTION

Pumps are among the essential equipment needed in many different industries, including oil and gas, chemical processing, water treatment, and others (Carravetta, et al., 2018). To sustain the general effectiveness and productivity of these industries, pumps must operate properly. Unanticipated pump failures can lead to serious disruptions, monetary losses, and safety risks (Yang, et al., 2022). To proactively identify anomalies and potential pump failures, it is crucial to adopt effective machine condition monitoring systems.

The use of Gaussian Mixture Models (GMMs) for probabilistic clustering and anomaly detection is a promising method. It is a potent statistical tool for simulating complex data distributions (Yu & Deng, 2014). This makes them well-suited for simulating the behavior of pumps, which can exhibit several operational modes and variable circumstances throughout time, as they can capture detailed linkages and fluctuations in the data. Utilizing GMMs makes it feasible to locate distinct clusters of data, each of which corresponds to a different aspect of the pumps' operation or condition.

Several traditional techniques are applied to machine condition monitoring, and the predefined threshold-based methods are the most frequently used of them all. However, traditional methods are generally not sensitive enough to identify small changes in pump behavior or may cause false alarms (Li, et al., 2018). The state of machinery has traditionally been monitored using pre-established criteria and regulations, which may not be sensitive enough to identify minute irregularities or adjust to shifting operating conditions (Black, Richmond, & Kolios, Condition monitoring systems: a systematic literature review on machine-learning methods improving offshore-wind turbine operational management, 2021)**.** Consequently, there is an increasing interest in using more efficient cutting-edge methodologies that can easily identify intricate patterns and variations in machine performance.
As a result of Industry 4.0 and the development of the Industrial Internet of Things (IoT), there is an increasing tendency toward data-driven, predictive maintenance practices. These methods use statistical modeling and machine learning to interpret the massive volumes of data produced by sensors and devices linked to machinery. Several such methodologies have been investigated and reported for a variety of pump applications, including fault detection, classification, and prediction (Qi, et al., 2022). These established data-driven methodologies present different forms of challenges, such as computational complexities and limitations of applications, for applications to pump condition monitoring (Eltouny, et al., 2023).

Potentially, GMMs can be used to group sensor data from pumps into several modes or states for machine condition monitoring. The various circumstances such as regular operation, wear and tear, or impending breakdown can then be connected to these clusters. The probabilistic characteristics of GMMs also offer a measure of uncertainty, which is useful in situations where it may be difficult to distinguish between various pump conditions.

The reliable operation of machinery, especially pumps, is crucial for sustaining efficient industrial processes and preventing costly downtime. Pumps in particular must run consistently for industrial processes to continue operating effectively and to avoid expensive downtime. There are several issues with current methodologies for machine condition monitoring since conventional traditional monitoring techniques often struggle to capture subtle anomalies and adapt to dynamic operating conditions, leading to compromised maintenance decisions. As the industrial landscape embraces data-driven approaches and predictive maintenance strategies, there is a need for advanced methodologies that can effectively analyse the complex sensor data generated by pumps and provide accurate insights into their condition. Thus, there is a high demand for cutting-edge procedures that can efficiently assess the complex sensor data produced by pumps and give precise insights into their condition as the industrial environment embraces data-driven approaches and predictive maintenance plans. This research investigated a robust and accurate machine condition monitoring system for a group of pumps using Gaussian Mixture Models (GMMs)-based probabilistic clustering 3.

2. Review of Literature

The shift towards data-driven approaches in machine condition monitoring is a transformational change influenced by the emergence of Industry 4.0 and the Industrial Internet of Things (IoT). It involves leveraging advanced data analytics, machine learning, and real-time connectivity to enhance the accuracy, efficiency, and effectiveness of monitoring the health and performance of industrial machinery. This shift represents a departure from traditional, schedule-based maintenance towards more proactive, predictive, and informed maintenance strategies (Tao, et al., 2018).

In essence, the shift towards data-driven approaches in machine condition monitoring driven by Industry 4.0 and IIoT is revolutionizing maintenance practices. It empowers organizations to make informed decisions based on real-time data, optimize maintenance activities, and ensure the reliability and longevity of their machinery, ultimately leading to enhanced operational efficiency and competitiveness (Buhr & Schicktanz, 2022)**.**

Gaussian Mixture Models (GMMs) are powerful statistical tools used for modeling complex data distributions by representing them as a combination of multiple Gaussian (normal) distributions. GMMs are widely employed in various fields, including machine learning, pattern recognition, data analysis, and probabilistic clustering (Huang, et al., 2023).

Gaussian Mixture Models (GMMs) provide a powerful and flexible framework for probabilistic clustering. By modeling data as a combination of Gaussian distributions, GMMs can capture complex data patterns, provide soft clustering assignments, and handle uncertainty in data. The Expectation-Maximization algorithm is commonly used to estimate GMM parameters, enabling data-driven insights and enhanced understanding of underlying structures in the data (Huang, et al., 2023).. They provide a flexible framework to capture intricate data patterns and uncover underlying structures in data.

The complexity in monitoring the condition of pumps is due to their operational variability, failure modes, and critical roles in industrial processes. Addressing these challenges requires robust sensor placement, advanced data analytics, real-time monitoring capabilities, and effective integration with existing systems. Overcoming these hurdles is essential to ensure the reliability, performance, and longevity of pumps and the systems they support (Nardi, et al., 2021).

The importance of vibration analysis in pump condition monitoring cannot be overemphasized using vibration analysis techniques, and signal processing methods, and provides insights into integrated approaches for pump condition monitoring, combining vibration analysis, acoustic monitoring, and other sensor-based methods. This study presented an integrated approach using multiple sensors for fault detection and diagnosis in centrifugal pumps (Romanssini, et al., 2023)**.**

Fault detection and diagnosis framework for centrifugal pumps using vibration signals and machine learning algorithms demonstrated more effectiveness in identifying specific faults (Vishwakarma, et al., 2017)**.**

Applied machine learning to acoustic emission signals for cavitation detection in centrifugal pumps demonstrated the potential of detecting and quantifying cavitation severity. The study also investigated pump fault detection by fusing data from multiple sensors, including

temperature, pressure, and vibration sensors. It demonstrated the potential of sensor fusion for accurate fault detection (Ghazali & Rahiman, 2021).

3. Methodology

3.1 Data collection process

The data comprises vibration signals and other pump parameters collected at the Warri Refining and Petrochemicals Company (WRPC) Limited, Ekpan-Warri, Delta State. The vibration signals were measured from the different pumps located within different departments at the refinery. It covered a span of 4 years (2015 -2018). During the data collection process, experts were utilized throughout.

3.1.1 Measuring device

The IRD digital vibrometer was used for the data collection. It is a specialized equipment developed for recording accurate vibration signals from machinery, such as pumps, to assess their operational conditions. It comprises a sensor, a data acquisition unit, a display interface, and an internal processing and storage unit. The sensors are attached to the pumps to take vibration measurements, the data acquisition unit collects the digitalized signal. It comes with a user-friendly interface that displays real-time vibration data. This interface allows users to visualize the recorded vibration signals and other details. Typically, its internal processing and storage system supports external storage devices to save the collected vibration data and allow the data to be transferred to a computer, while the signal processing capabilities allow users to apply filters, spectral analysis, and other processing techniques to the vibration data. These tools help extract meaningful information from the raw signals. There is an attached battery for powering it. The IRD vibrometer used for taking the vibration readings and further analysis is shown in Figure 1.

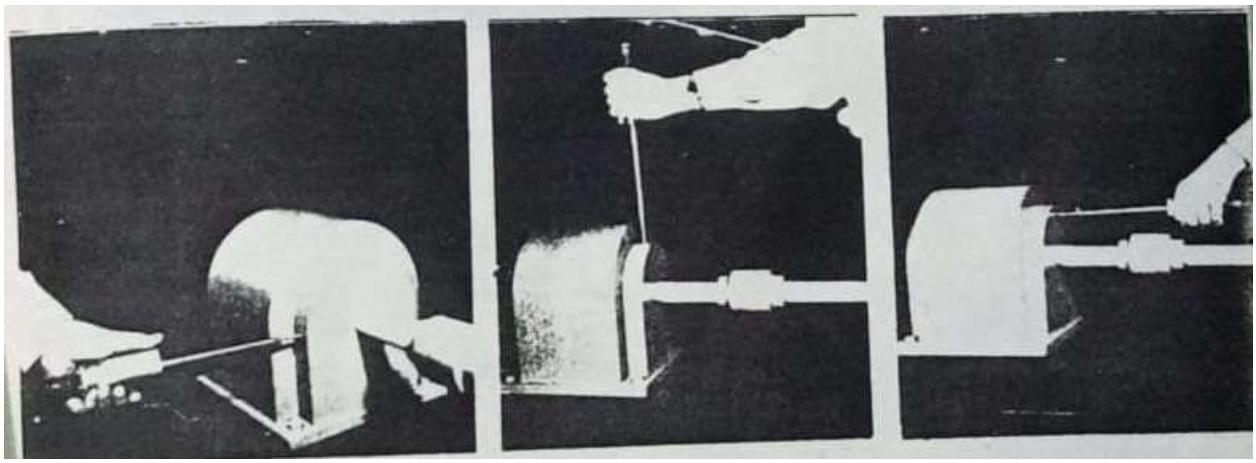


Figure 1: Vibration reading.

3.1.2 Setting up the vibrometer

The sensor of the vibrometer was attached to the pump using a pickup cable attached to the end of the vibrometer's receptacle point. Then the other end of the pickup cable is positioned at a desired point for collecting the vibration signals. The vibration signals were recorded for both the inboard and outboard of the device bearings in different directions (horizontal, vertical, and

axial directions). The positioning of the pickup cable to get horizontal, vertical, and axial readings, is shown in Figure 2.



Horizontal (H) Vertical (V) Axial (A)

Figure 2: Positioning the device for taking vibration Readings

3.1.3 Data Preprocessing

Initial pre-processing steps were performed on the data to address noise and unwanted signals. The IRD digital vibrometers included signal processing capabilities that allowed the application of filters, spectral analysis, and other processing techniques to the vibration data. The tool was then used to extract meaningful information from the raw signals, which in this case is the amplitude of vibrating velocity.

3.2 Data Labelling

The data included all necessary details. These included the pump specifications described by their tags, date, power of machine and pickup point. Since the analysis was going to be unsupervised learning, acknowledged problems or failures were excluded. These information for suspected problems were excluded during measurements because they were just speculations. Obtained values of vibration signals were in two categories, which are in-board and out-board signals. The readings were taken twice for each pickup point of for the three (3) different directions (horizontal, vertical and axial). The data was labelled as H, V and A for horizontal, vertical, and axial readings, respectively, and it is accompanied by a number 1 or 2 which indicates the reading as 1st or second, at the in-board or out-board, respectively. For instance, H1 means horizontal in-board reading.

The first part, which is usually a number such as 20 in 20-PM-05A stands for Topping Unit (the location where the equipment is installed).
The second part which is usually a letter two letters stands for the equipment type. For example, P in 20-P-05A stands for pump, which indicates it is the pump.
The last part which contains a number and a letter combined, such as it is 05A in the case of 20-PM-05A connotes the successive number of such equipment installed in the unit.

3.3 Data Quality Assurance

Quality assurance checks were implemented to ensure that the collected vibration data was accurate and reliable. This involved periodic sensor calibration and validation against known standards. These were done from time to time at the WRPC where the data was collected.

3.4 Data cleaning

The data cleaning process involves locating and removing the missing values from the recorded dataset. Missing vibration signals that were due to typing error were replaced while the others that were missing from the field were removed from the data set. No imputation or interpolation was done during the data-cleaning process. In the end, the data was reduced to a total of 424 samples.

Next error readings, such as values that cannot possibly be the velocity of vibrating rotating machines were identified and removed. Any evident data errors or outliers that could skew the analysis's findings were removed or corrected. A simple formula shown in Equations (1) and (2) that computes outliers was used to calculate and remove all outliers from the readings, and in the end, only 374 samples were left. Vibration signals greater than the upper limit or lesser than the lower limit were then removed.

$$Lower\ limit = Q1 - 1.5 * IQR \quad (1)$$

$$Upper\ Limit = Q3 + 1.5 * IQR \quad (2)$$

Q1 is first quartile

IQR is the interquartile range

Q3 is the third quartile

3.4 Data transformation:

To comply with the suppositions of the factor analysis and K-means clustering algorithms, the data set was codified into dummy variables where necessary. Then all the measured variables were transformed by calculating their z-scores as appropriate. Calculating the Z-score helps to address the issues that may affect the analysis as a result of differing scales. This helps in preventing bias toward variables with higher magnitudes. Calculated Z-scores are known effective way of scaling data. The Z-scores were calculated using Equation (3). The z-score was calculated for all the variables by coding on Microsoft Excel

$$z = \frac{x-\mu}{\sigma} \quad (3)$$

Z is the z-score

X is the measured signal being transformed

μ is the mean

σ is the standard deviation

3.5 Gaussian Mixture Model (GMM):

In this study, Gaussian Mixture Models (GMMs) were implemented in the Python 3.0 software Skip-learn library. GMMS is a powerful statistical model used in machine learning for representing and analyzing complex data distributions. They are particularly useful when dealing with data that does not follow a single simple distribution but is composed of multiple underlying patterns or clusters. GMMs are advantageous for clustering because they can capture complex cluster shapes and densities. Unlike some other clustering algorithms, GMMs also provide a probability distribution for each data point's membership in each cluster, which can be useful for uncertain or overlapping clusters.

In a GMM, the data is assumed to be generated from a mixture of several Gaussian (normal) distributions, each representing a distinct cluster or component of the data. These Gaussian distributions are combined with different weights to form the overall mixture. GMMs capture both the means and variances of these underlying Gaussian components, allowing them to model data with various shapes and complexities.

GMMs have various applications in machine learning. However, in this work, it was used for unsupervised clustering, where they can automatically identify and group similar data points into clusters. Each Gaussian component corresponds to a cluster, and the model assigns data points to the component that best explains their distribution.

3.4.1 GMM algorithm

The steps in carrying out GMM involve: Initialization, Expectation-Maximization (EM) Algorithm, Iterative Refinement, Cluster Assignment, and Number of Clusters

3.4.2 Mathematical formulation of GMM

The Gaussian Mixture Model (GMM) is mathematically formulated as a mixture of multiple Gaussian distributions. Let's break down the components of this formulation:

3.4.3 Gaussian Distribution (Normal Distribution)

The Gaussian distribution is a fundamental probability distribution (P(x) commonly used to model continuous data (x). It is defined by two parameters: the mean (μ) and the covariance (Σ). Where σ is the standard deviation for a one-dimensional Gaussian distribution, the probability density function (PDF) is given by:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}}.e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (4)$$

In the case of multivariate data, the multivariate Gaussian distribution is used, and the PDF is given by:

$$P(x) = \sum_{k=1}^{K} \pi_K.N(x;\ \mu_K, \Sigma_K) \quad (5)$$

where:

μ_K is the distribution mean for component K, π is the weight (or mixture coefficient)

3.4.4 Expectation-Maximization (EM) Algorithm:

The EM algorithm is used to estimate the parameters of the GMM, including the means, covariances, and weights. It involves two main steps: the E-step (Expectation) and the M-step (Maximization).

E-step (Expectation):

In this step, the algorithm calculates the posterior probabilities (responsibilities) that each data point belongs to each Gaussian component. The posterior probability of data point ω_{ik} belonging to component k is given by:

$$\omega_{ik} = \frac{\pi_k.N(x_i;\mu_K,\Sigma_K)}{\sum_{j=1}^{K}\pi_j.N(x_i;\mu_j,\Sigma_j)} \quad (6)$$

M-step (Maximization):

In this step, the algorithm updates the parameters of each Gaussian component based on the calculated responsibilities. The updated parameters are given by:

$$\mu_K = \frac{\sum_{i=1}^{N} \omega_{ik}.X_i}{\sum_{i=1}^{N} \omega_{ik}} \quad (7)$$

$$\Sigma_K = \frac{\sum_{i=1}^{N} \omega_{ik}.(X_i-\mu_K)(X_i-\mu_K)^T}{\sum_{i=1}^{N} \omega_{ik}} \quad (8)$$

$$\pi_k = \frac{1}{N}\sum_{i=1}^{N} \omega_{ik} \quad (9)$$

The E-step and M-step are alternated iteratively until convergence, where the parameters stabilize. Convergence is typically determined by changes in the log-likelihood or after a predetermined number of iterations.

The EM algorithm helps GMMs find the optimal parameters that maximize the likelihood of the observed data under the model. It's important to note that the EM algorithm can be sensitive to initialization, and multiple runs with different initializations might be needed to find a good solution.

3.5 Grid Search parameter optimization

Grid search is a systematic approach for hyperparameter optimization that involves evaluating a combination of hyperparameters across a predefined grid of possible values. In this study, grid search was applied to optimize hyperparameters for Gaussian Mixture Model (GMM) clustering of pump vibration signals following the steps below:

a. **Definition of hyperparameter Grid:**
b. **Iterate through Hyper parameter Combinations:**

Iterate through all possible combinations of hyperparameters from your defined grid. For each combination, the following steps were performed: Pre-process Data, Train GMM, Evaluate Clustering, Record Results, and Repeat for all combinations:

4.0 Results/Discussion

4.1 Data description

The data used for the analysis after cleaning is summarized using different descriptive statistical tools and presented in Table 1. The total records used are 344 readings taken at different periods. The power rating ranged from 4 kw to 750 kw. The maximum reading obtained is 20.37 mm/s taken at the axial outboard position (A3). A closer look at the mean and maximum values showed that the values of outboard signals (A3, H3, A3, V4, H4, and A4) were higher than those of the inboard readings (A1, H1, A1, V2, H2, and A2).

Table .1: Summary of the different variables used for the analysis

Statistics	Power Rating	V1	H1	A1	V2	H2	A2	V3	H3	A3	V4	H4	A4
Count	344	344	344	344	344	344	344	344	344	344	344	344	344
Mean	104.93	1.73	1.72	1.09	2.49	2.33	2.20	3.01	3.79	2.22	2.44	3.07	2.24
Std	105.41	1.71	1.21	1.17	2.74	1.89	2.60	2.43	3.32	2.36	2.11	2.90	2.31
Min	4.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	40.00	0.73	0.90	0.47	0.90	1.13	0.71	1.28	1.45	1.07	1.20	1.24	0.85
50%	75.00	1.25	1.47	0.76	1.59	1.89	1.25	2.39	3.01	1.74	1.95	2.30	1.56
75%	132.00	2.04	2.20	1.29	2.89	2.99	2.67	4.41	5.13	2.97	3.19	4.39	3.15
Max	750.00	14.91	6.41	7.77	14.72	11.83	14.33	14.99	19.41	20.37	10.09	17.75	18.45

After cleaning, most of the datasets utilized were readings taken in the year 2015. Other years in which vibration signal data were included are 2014, 2016, 2017, and 2018, as shown in Figure 1. These were the years the facility was fully in operation and the machines were utilized and accessed. Overall, the data read was least for year 2014, followed by 2017.

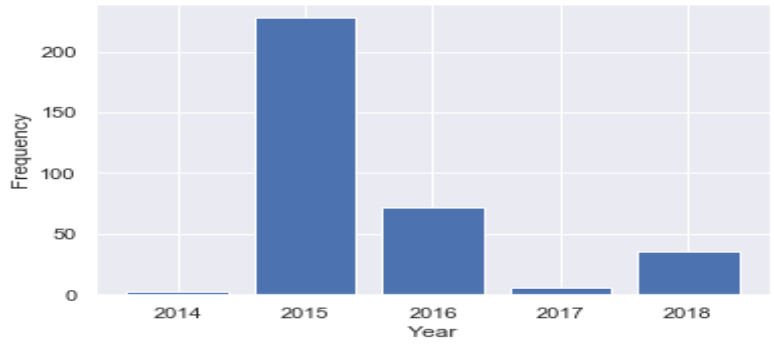


Figure 3: Year of reading for the different data points

Data on the location of the equipment being assessed was also recorded, the frequency of recordings taken at the different locations is shown in Figure 2. The essence is to be sure which equipment maintenance personnel took the recording at the time. This can also be used to assess how often equipment would encounter failure in each location. S can be observed, that most of the readings were taken at location "10," indicating several of the equipment reside there. Also, location "15" and "16" houses a good number of equipment as well.

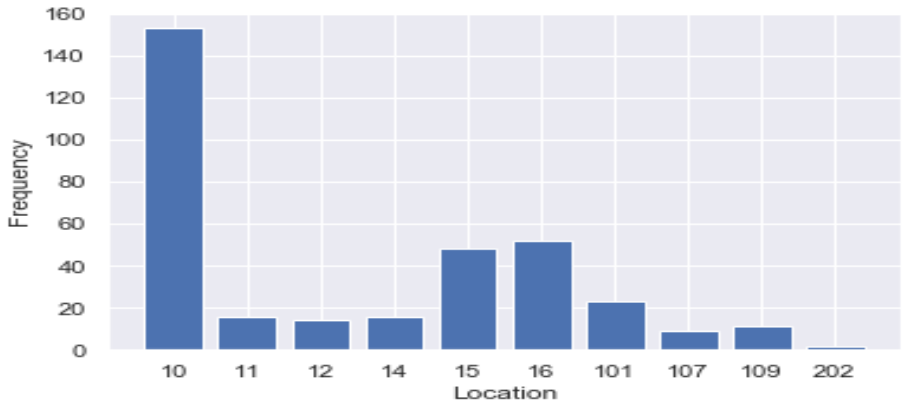


Figure 4: Location of equipment

The distribution of power ratings of the different pumps from which the data was taken is shown using a histogram in Figure 3. As can be observed, most of the motors are of power rating less than 300 kW, with small power pumps (less than 50 kW) having the highest frequency. A few equipment, however, are of a very high power rating, ranging between 250 kw to 750 kw.

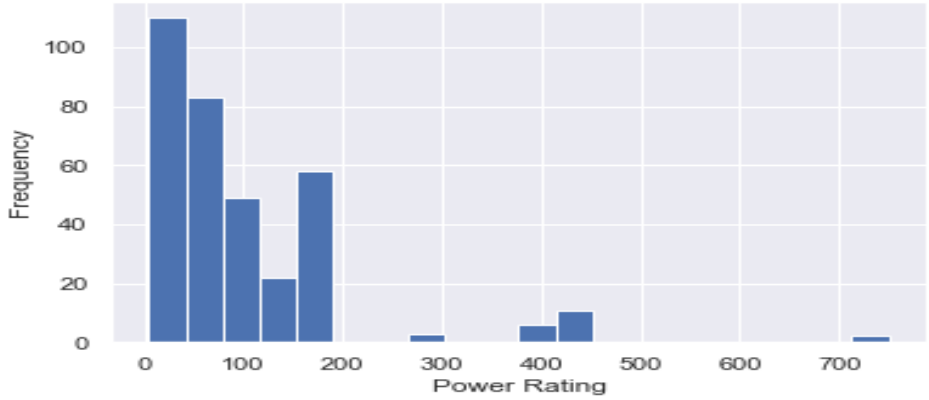


Figure 5: Histogram of power rating

The frequency distribution of the different signals in mm/s is shown given subplots (Figures 4, 5, and 6). The histogram bins are higher for lower values, showing an exponential distribution for all the readings. The frequency of these low values is up to 200 for axial directions in all the charts. However, the frequency is slightly lesser (about 170) for the inboard readings of the horizontal reading. It is far less (about 120 for the inboard reading of the vertical readings. These higher values are indications that many of the readings taken having lower vibration signals are healthy, and fewer pumps are in a bad state needing attention. Overall, the outboard readings are higher, having more readings that are greater than 10 mm/s.

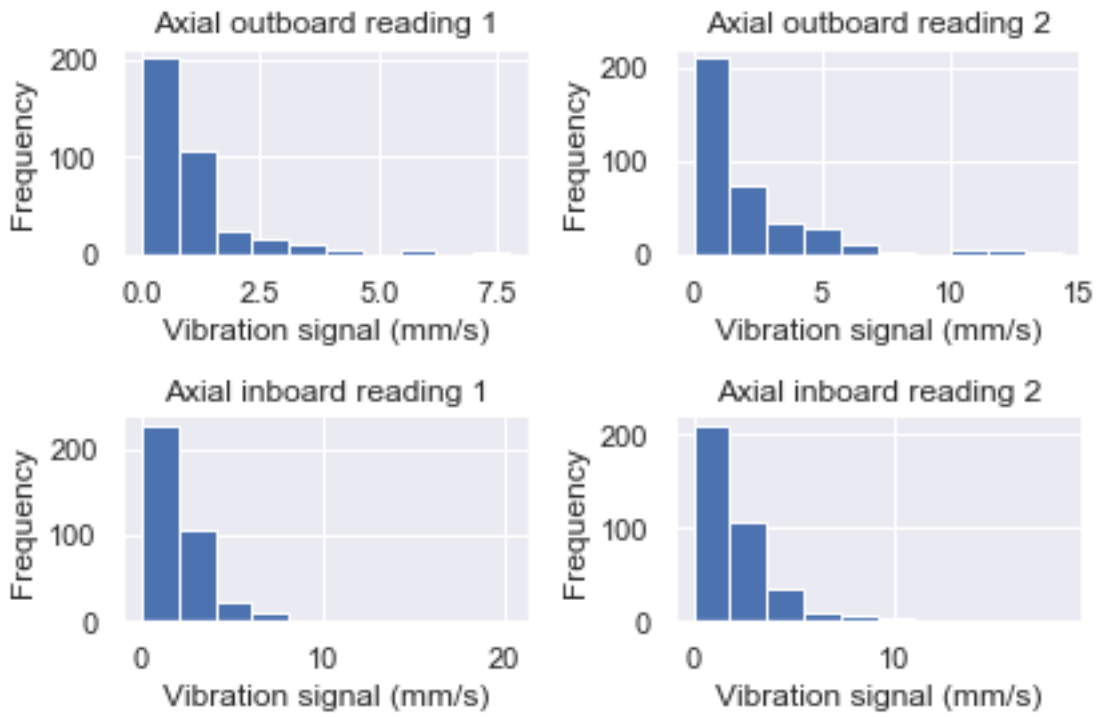


Figure 6: Axial readings taken at the inboard and outboard positions

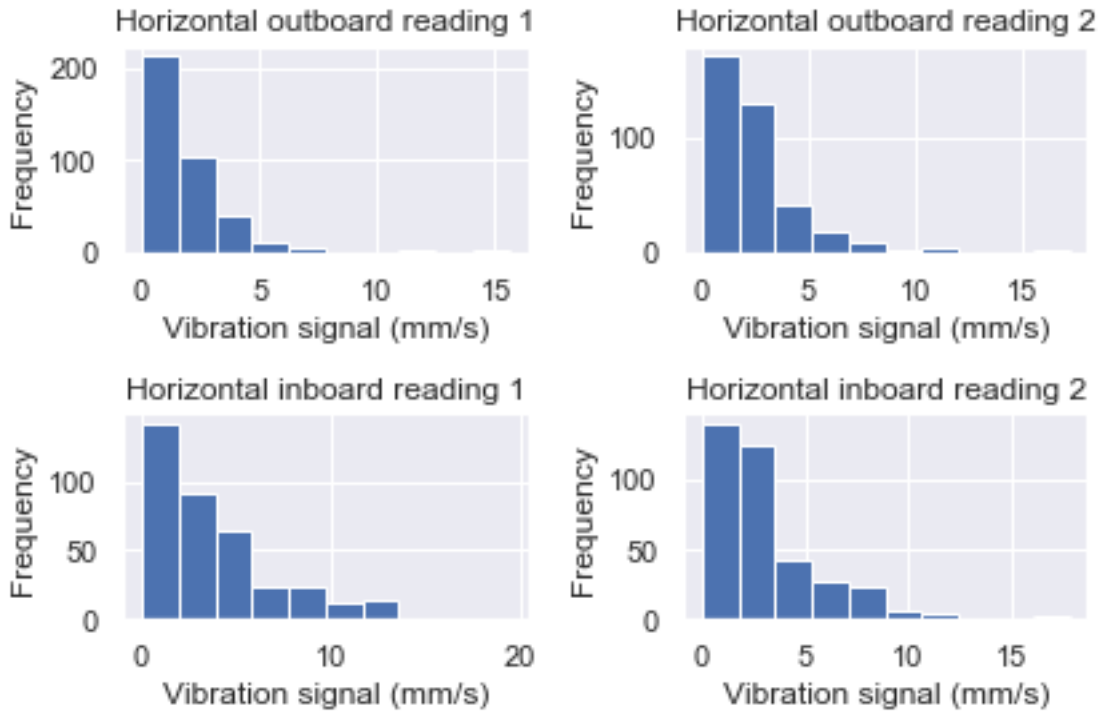


Figure 7: Histogram of the vertical outboard and inboard readings

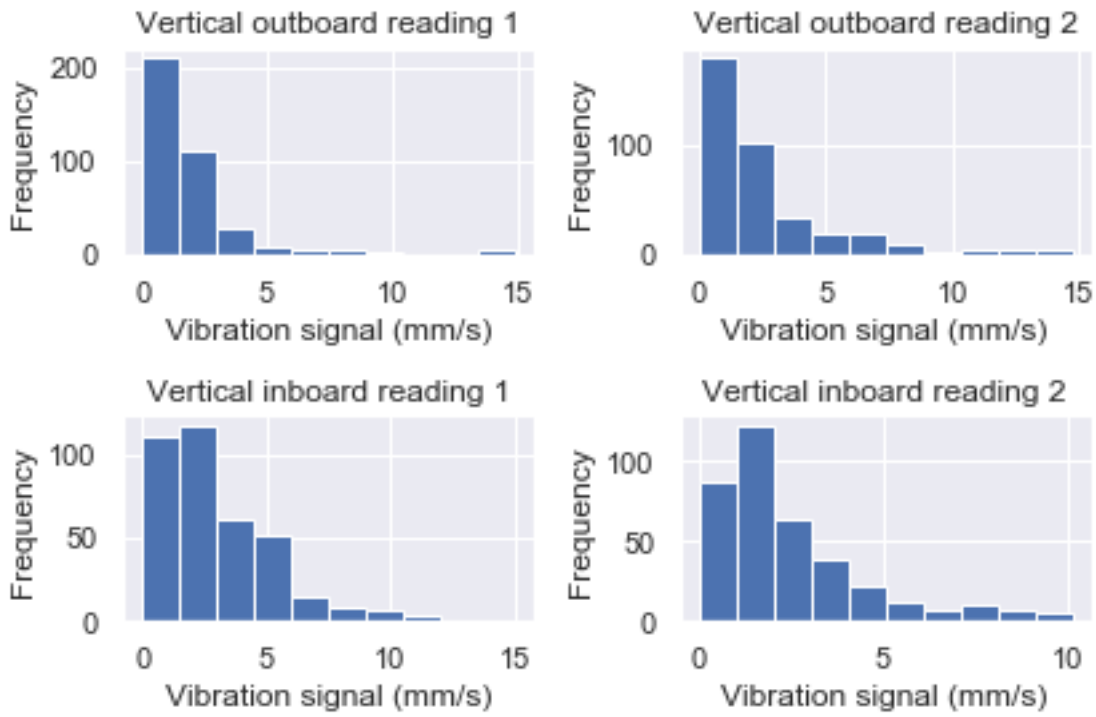


Figure 8: Histogram of the horizontal inboard and outboard readings

4.2 Optimum Hyperparameter

Three different indicators were used to assess which number of clusters were most suitable for the analysis. They are the silhouette scores, BIC scores, and the log-likelihood index. It was estimated for different numbers of clusters and different values of random states. The findings for the calculated indicator scores are plotted against the different random states and number of clusters. The values obtained using log-likelihood were not feasible, since they didn't show any form of convergence. The scores increased continually as the number of clusters increased. However, a closer inspection showed that 7 clusters had a good score for both silhouette and BIC plots, and the best random state for that number of clusters was 54. So, 7 and 54 were selected for the number of clusters and random states, respectively.

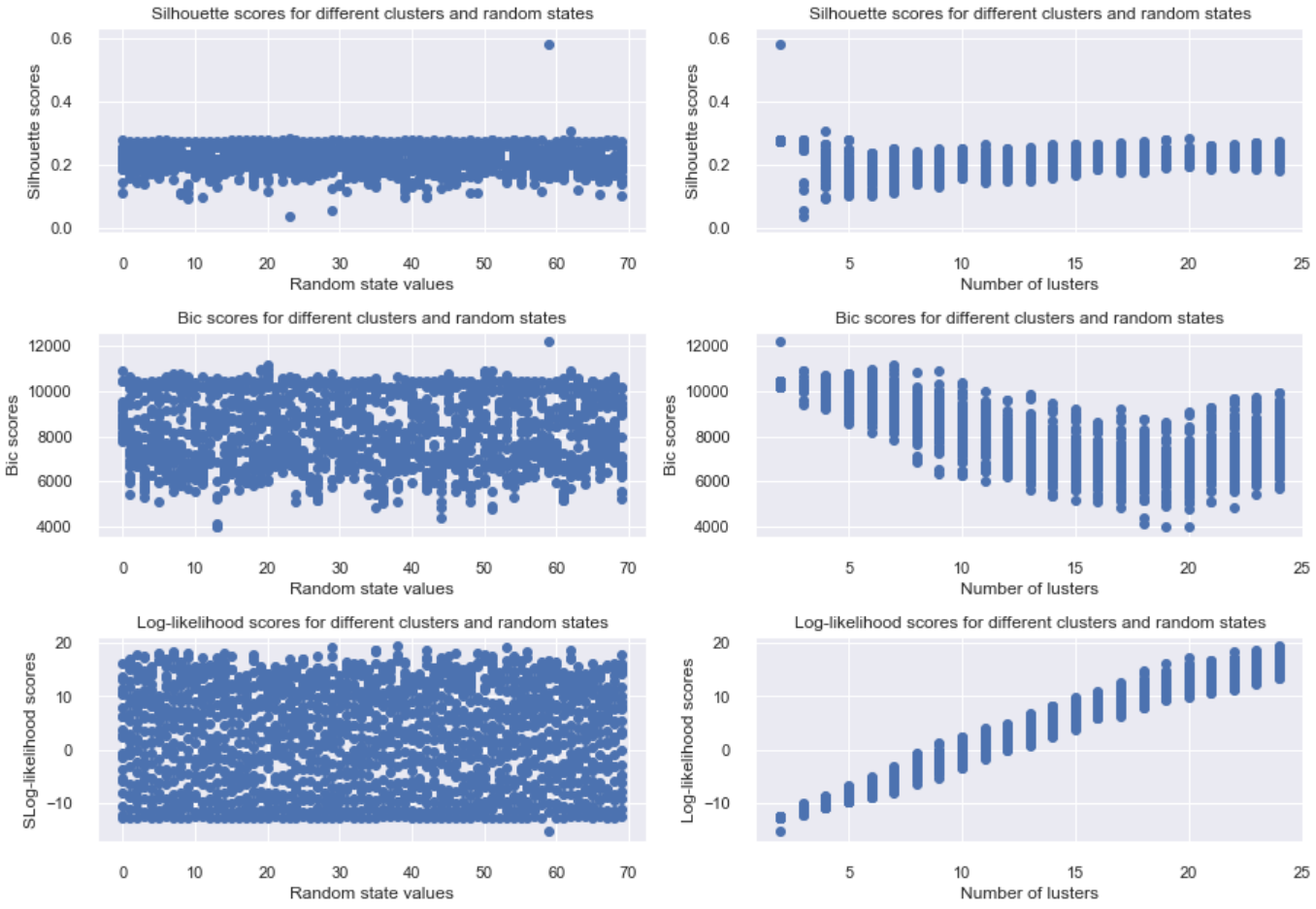


Figure 9: Different performance scores for the investigated number of clusters and random states

4.3 Performance characteristics of GMM

Based on the pre-determined number of clusters (7) and the best value for the random state (54), the data was divided into the different clusters firstly using the Gaussian mixture model, and then the K-means algorithm. The performance of each algorithm was assessed using three different indices, namely silhouette scores, Log-likelihood score, and the Bayesian information criteria (BIC). The result is shown in Table 2. For BIC scores the model with the lower scores has a better performance whereas reverse is the case for silhouette scores. Therefore, it is obvious that for seven number of clusters, GMM had a better result with a higher silhouette score and lower BIC.

Table 2: Performance analysis for GMM and K-Means clustering

Indices	GMM	K-Means
Silhouette	0.68	0.51
BIC	9367	12178

The number of signals grouped into each cluster using each of the model. The first GMM cluster had the highest number of data while the fifth K-means cluster had the highest number of data set. For both models, there are some groups with very few numbers of recordings (GMM

cluster 2 &4 and K-means cluster 7& 5). The other clusters for both models had similar number of data set grouped into them.

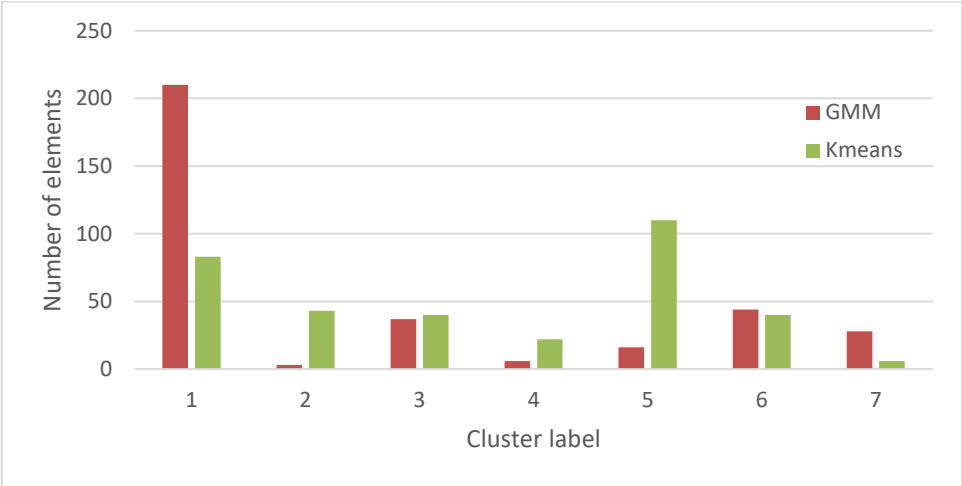


Figure 10: The number of signals grouped into each cluster

4.4 Fault detection with GGM and K-means

The different clusters were assessed to identify their unique characteristics, and it was found that different unique characteristics can be found and discussed as follows. The different clusters were named from 0-6 and combined with the model that created it. Cluster GMM0 means the cluster tagged 0 created by the GMM model while cluster Kmeans0 means the cluster tagged zero created by the K-means model.

4.4.1 Good machines

Four clusters contained machines with signals that indicated that they were in good order. They are GMM0, Kmeans0, Kmeans1 and Kmeans2. The GMM model classified all the good machines into a single cluster containing 210 sets of machine readings, whereas K-means created three different clusters for good machines. The number of machines contained in each of the clusters and other summaries are shown in Table 3. The differences in the different groups created by K-means were the location and year. The k-means clustering technique took into consideration the year the data was collected and the machine's location to cluster the data, whereas it was ignored by the GMM technique.

The GMM technique included about 21 classes of pumps based on their power rating, whereas the K-means technique considered 15, 9, and 11 categories in clusters 0, 1, and 2, respectively. Minimum values were zero for three of the clusters indicating that some machines had no values which represented faulty sensors. The average values of all inboard and outboard readings were relatively low (< 2.1 m/s).

Table 3: Summary of data for machines in clusters of good machines

	GMM0	Kmeans0	Kmeans1	Kmeans2
No of Equip in the cluster	210	83	43	40
Min Power rating	4	30	30	30
Max Power rating	450	750	170	180
No of different categories	21	15	9	11
Minimum signal read	0	0	0.67	0

Average inboard vertical	1.37	1.82	1.37	1.16
Average inboard horizontal	1.91	1.84	1.81	1.89
Average inboard axial	1.34	1.42	2.84	1.48
Average outboard vertical	1.51	2.09	1.42	1.56
Average outboard horizontal	1.38	1.86	2.09	0.79
Average outboard axial	0.92	1.57	1.85	1.16
Location	10, 11, 12,14, 15, 16, 101	10, 15, 16, 101	10,15,101	10, 11, 15, 16
year read	2014 - 2018	2015-2016	2015-2018	2017-2018

4.4.2 Bad Sensors

Three different clusters had machines with several zero values, indicating that they contained several faulty sensors, hence no reading was obtained at that point. The information from these groups is summarized in Table 4. The three clusters were GMM5, Kmeans4, and Kmeans5. The k-means model created two clusters belonging to this category, whereas GMM created just one cluster in this category. The total number of machines in theGMM5 cluster was 44, whereas all the machines classed into this category by the K-means model summed up to 62. Thus, more machines with bad sensors were left out by the GMM model. All 14 categories of machines based on their power rating had bad sensors.

The Kmeans4 differs from the Kmeans5 cluster Kmeans5 cluster had machines with bad sensors and also some readings that were moderately high compared to the Kmeans4 values. This is obvious from their average values.

Table 4: Summary statistics of clusters of machines with bad sensors

	GMM5	**Kmeans4**	**Kmeans5**
No of Equip in cluster	44	22	40
Min Power rating	14	4	14
Max Power rating	286	160	286
No of different categories	12	15	9
Minimum signal read	0	0	0
Average inboard vertical	1.24	0.9	2.36
Average inboard horizontal	0.57	1.07	1.68
Average inboard axial	0.7	1.23	2.13
Average outboard vertical	1.9	1.06	3.45
Average outboard horizontal	1.73	1.09	4.01
Average outboard axial	1.97	0.81	4.6
Location	10, 14, 15, 16, 101, 107	10, 11, 12, 14, 15, 16, 101, 109	10, 14, 15, 101, 107
year read	2015 - 2018	2014-2016	2015-2018

4.4.3 Defect from inboard readings

A group was created by the GMM model that contained pumps with signals indicating that there was a fault that is from the inboard signals. The summary which is for cluster tagged

GMM3 is shown in Table 4. As can be observed, there are just 6 pump data in this category with power ranging from 30-75 kw. The average readings from the inboard areas are higher compared to those of the outboard values. Also, the minimum signal read in this group is 0, indicating that a machine with a faulty sensor was also included in this group.

Since the power rating is low, one can deduce that these pumps are faulty. All the pump readings in this group were measured in 2015

Table 5: Summary of the data for pumps with faults indicated by the inboard reading

	GMM3
No of Equip in cluster	6
Min Power rating	30
Max Power rating	75
No of different categories	3
Minimum signal read	0
Average inboard vertical	7.04
Average inboard horizontal	3.21
Average inboard axial	6.915
Average outboard vertical	2.19
Average outboard horizontal	2.36
Average outboard axial	2.72
Location	10, 101
year read	2015

4.4.5 Defects from Outboard Readings

There were two clusters created that contained pumps with outboard readings which indicated that the pump had faults, while the K-means technique created just a single cluster with potential fault indicated by the outboard readings. The difference between the two clusters created by the GMM model is that one (GMM2) contains pumps with faults indicated by the all out-board readings taking in different directions, while the second GMM6 contains pumps whose faults were indicated by just readings in the horizontal and axial directions. Overall, as seen in Table 6, the total number of pumps in this category is 10, 5, and 9 for the GMM2, GMM6, and Kmeans3 clusters, respectively. The Kmeans3 and GMM2 clusters had signals with value 0 indicating that some pumps with faulty signal readings were included. The pumps in these categories ranged from 30 to 400 kw in capacity. The location of the pumps spanned around the site and were readings taken in different years from 2015 to 2018.

Table 6: Summary of pump characteristics for clusters containing faulty pumps based on outboard readings

	GMM2	**GMM6**	**Kmeans3**
No of Equip in cluster	37	28	22
Min Power rating	30	30	30
Max Power rating	400	90	400
No of different categories	10	5	9
Minimum signal read	0	0.4	0
Average inboard vertical	2.21	2.22	1.92
Average inboard horizontal	2.23	2.08	2.32

Average inboard axial	2.04	2.71	2.44
Average outboard vertical	5.03	2.9	5.1
Average outboard horizontal	6.51	6.04	6.74
Average outboard axial	8.21	5.32	7.78
Location	10, 14, 15, 16, 101	101, 107, 109, 202	10, 14, 15, 16, 101, 109
year read	2015 - 2018	2015-2017	2015-2018

4.4.6 Defective from all kinds of reading

A last category of the cluster, GMM3, by the GMM model is that which had just a set of readings from a particular pump of 400 kw with readings. The pumps had high inboard and outboard readings taken from all directions. They were located at position "15" and were assessed in 2015. There were readings of zero values as well, indicating that pumps with bad sensors were also included in this cluster.

4.4.7 Irregular combinations

The last category of clusters was those with irregular combinations, and the summary of the different clusters is shown in Table 4.7. The GMM and K-means model could create such kind of clusters, demonstrating their weakness. The GMM had a higher number of pumps classed into the GMM4 while K-means produced a cluster, kmeans6 with just 6 pumps in this category. Power ranged between 30 and 400 kw. Several sets of issues were identified among pumps in this group, namely bad sensors, and high values of signals from various directions. The readings included in both clusters were taken within the same year and locations.

Table 7: irregular readings from different points

	GMM5	**Kmeans6**
No of Equip in the cluster	16	6
Min Power rating	30	30
Max Power rating	170	400
No different categories	6	3
Minimum signal read	0.61	0
Average inboard vertical	5.2	8.22
Average inboard horizontal	4.32	5.49
Average inboard axial	2.16	6.35
Average outboard vertical	4.97	4.21
Average outboard horizontal	5.84	4.76
Average outboard axial	3.89	3.67
Location	10,16,101	15, 16, 101
year read	2015 - 2016	2015-2016

In conclusion, the findings from this research indicate that Gaussian Mixture Models hold considerable promise for addressing the challenges of machine condition monitoring and fault detection in pump systems. The ability to accurately cluster data, detect anomalies, and provide insights into fault types underscores the potential impact of this research on industrial maintenance practices.

References

Black, I. M., Richmond, M. & Kolios, A., 2021. Condition monitoring systems: a systematic literature review on machine-learning methods improving offshore-wind turbine operational management. *International Journal Of Sustainable Energy,* 40(10), pp. 923-946.

Buhr, L. & Schicktanz, S., 2022. Individual benefits and collective challenges: Experts' views on data-driven approaches in medical research and healthcare in the German context. *Big Data & Society,* 9(1), p. 15.

Carravetta, A., Houreh, S. D. & Ramos, H. M., 2018. *Pumps as Turbines.* s.l.:Springer Cham.

Christensen, H. I. & Hager, G. D., 2008. Sensing and Estimation. In: *Springer Handbook of Robotics.* Berlin: Springer.

Fan, C. et al., 2021. A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Front. Energy Res.,* Volume 9.

Ghazali, M. H. M. & Rahiman, W., 2021. Vibration Analysis for Machine Monitoring and Diagnosis: A Systematic Review. *Shock and Vibration,* p. 25.

Li, L., Hua, C. & Xu, X., 2018. Condition monitoring and fault diagnosis of electric submersible pump based on wellhead electrical parameters and production parameters. *System Science and Control Engineering,* 6(3), pp. 253-261.

Nardi, L., Lavinia, C., Iatauro, D. & Calabrese, N., 2021. Field study on heat pump monitoring: challenges and opportunities. *E3S Web of Conferences,* p. 12.

Pourbahrami, S., Balafar, M. A., Khanli, L. M. & Kakarash, Z. A., 2020. A survey of neighborhood construction algorithms for clustering and classifying data points. *Computer science review,* Volume 38, p. 100315.

Qi, R., Zhang, J. & Spencer, K., 2022. A REVIEW ON DATA-DRIVEN CONDITION MONITORING OF INDUSTRIAL EQUIPMENT. *ALGORITHMS,* 16(1), p. 9.

Romanssini, M., Aguirre, P. C. C. d., Compassi-Severo, L. & Girardi, A. G., 2023. A Review on Vibration Monitoring Techniques for Predictive Maintenance of Rotating Machinery. *Eng,* 4(3), pp. 1797-1817.

Tao, F., Qi, Q., Liu, A. & Kusiak, A., 2018. Data-driven smart manufacturing. *Journal of Manufacturing Systems,* Volume 48, pp. 157-169.

Viswanathan, P. C. et al., 2023. Deep Learning for Enhanced Fault Diagnosis of Monoblock Centrifugal Pumps: Spectrogram-Based Analysis. *Machines,* Volume 11, p. 874.

Yang, Y. et al., 2022. Current Status and Applications for Hydraulic Pump Fault Diagnosis: A Review. *Sensors,* 22(24), p. 9714.

Yu, B. et al., 2023. A Network Traffic Anomaly Detection Method Based on Gaussian Mixture Model. *Electronics,* 12(6), p. 1397.

Yu, D. & Deng, L., 2014. *Gaussian Mixture Models.* s.l.:Springer, London.