Computers and Chemical Engineering 152 (2021) 107381

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Adding interpretability to predictive maintenance by machine learning on sensor data

Bram Steurtewagen, Dirk Van den Poel*

Universiteit Gent, Department of MIO/Data Analytics, Tweekerkenstraat 2, 9000 Gent, Belgium

ARTICLE INFO

Article history:
Received 12 April 2021
Revised 8 May 2021
Accepted 23 May 2021
Available online 27 May 2021

Keywords:
Condition-based maintenance
Machine failure prediction
Machine diagnosis
Machine learning
Sensor data

ABSTRACT

Condition-based maintenance (CBM) is becoming more commonplace within the petrochemical industry. While we find that previous research leveraging machine learning has provided high accuracy in the predictive aspect of machine breakdowns, the diagnostic aspect of these approaches is often lacking. This paper implements a supervised machine learning approach, with the goal of both prediction and diagnosis of machinery breakdowns, emphasizing the latter. To achieve this, it uses an XGBoost model trained on a combination of sensor and report data, and enriches the model with Shapley values for diagnostic insights. We show that this combination of statistical methods, combined with a proper data treatment, can be used to great effect and can vastly improve the diagnostic value of machine learning approaches. The insights that follow from the analysis can subsequently be leveraged by plant operators in CBM strategies or root-cause analyses.

1. Introduction

In contemporary refineries, interlinked processes have become more advanced and complex. As a consequence, the machinery used in these processes needs more sophisticated and targeted maintenance strategies (Ahmad and Kamaruddin, 2012). One of the emerging trends is the use of advanced statistical data analysis and modeling techniques on available sensor data (Moraru et al., 2010) to investigate machine behavior and breakdowns (Lee et al., 2014). Leveraging these new methods and utilizing new data inflows come with their own respective challenges (McAfee et al., 2012). The challenges consist of, but are not limited to; gathering, storing, serving, analyzing data, and generating buy-in with on-site personnel. Modern industrial plants exacerbate these challenges by generating vast amounts of data with a high frequency. Moreover, data is gathered from various sources (i.e. sensors, event reporting systems, inspection reports, laboratory results,...) and can either be collected automatically or manually. All of these types of data, however, are seeing increased usage in maintenance applications.

Adding to the above data-related challenges, one finds the continued use of traditional strategies within the industry. The traditional approaches (Bloch and Geitner, 2019) focus on monitoring the 'bad-actor' equipment, repairing the damaged equipment after a breakdown (run-to-failure or corrective maintenance) or on adhering to a strict inspection schedule (preventive maintenance), and both of these outdated extremes can be considered as wasteful (Jardine et al., 2006). However, an evolution in these maintenance strategies is occurring through the application of analytical Condition-based Maintenance (CBM) (Engel et al., 2000). CBM seeks to provide a middle ground between the two traditional views (Tsang, 1995). It avoids the wasteful extreme cases and is being used in real-life scenarios to great effect (Liu and Karimi, 2020). The improvement over the traditional views is mostly attributed to solving the complexity challenges where the ability of humans to identify patterns falls short. In this regard, supervised machine learning is suggested for systematic prediction of faults where the amount of well-defined knowledge is vast and the sequence of steps required to identify the fault is very long (Gelgele and Wang, 1998). This stems from the dependency of machine learning on sizeable amounts of real or simulated (Sobie et al., 2018) data to learn the breakdown patterns. An oil refinery, such as discussed in this paper, fits the paradigm of a data-rich, complex environment where machine learning techniques provide the most benefit (Shah et al., 2020; Kadlec et al., 2009).

In data-centric CBM (Jardine et al., 2006), a predictive model is used to anticipate machine failures or shutdowns (Hashemian, 2010; Mobley, 2002). Fault diagnosis of machinery (Isermann, 2011) is an extensive field covering both theoretical and empirical approaches (Heng et al., 2009). The predictive aspect through empirical data mining is well covered in literature; deep learning neural nets have been used to model bearing failure (Sohaib and Kim, 2018), gas turbine performance (Liu and Karimi, 2020), and

* Corresponding author.
E-mail address: dirk.vandenpoel@ugent.be (D. Van den Poel).

https://doi.org/10.1016/j.compchemeng.2021.107381
0098-1354/

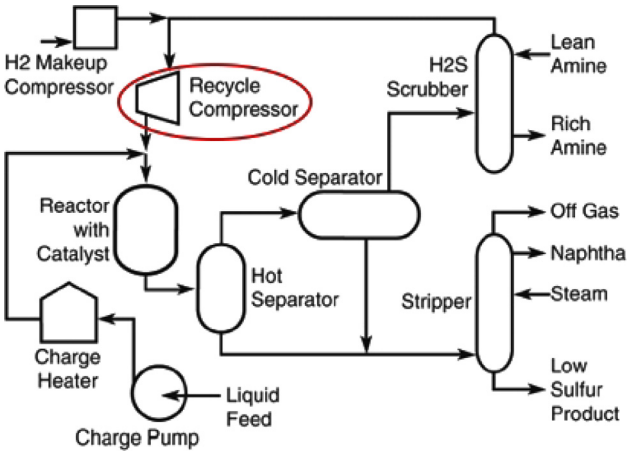

Fig. 1. General ARDS process.

have a proven track record throughout failure predictions (Wen et al., 2017). Other techniques, such as Random Forests, have also proven successful in predicting compressor failures (Aravinth and Sugumaran, 2018; Steurtewagen and Van den Poel, 2019). While these modern statistical techniques are likely to provide high predictive value, they often provide a so called 'black box solution' (Handelman et al., 2019) which is harder to interpret and lacks diagnostic value. Adding to this, identifying key driving factors in complex models is challenging (Carvalho et al., 2019) and diagnosis is confirmed as one of the common issues in data mining (Cao, 2010).

This shortcoming in the diagnostic aspect causes plant operators to attribute little value to an incomprehensible sensor reading stemming from statistical methods (Pyle and San José, 2015). Shapley values (Shapley, 1953) have recently been brought to light as a great interpretative tool for extending machine learning approaches (Messalas et al., 2019; Sundararajan and Najmi, 2019). Our research proposes to add value to the predictive aspect of machine learning by expanding the diagnosis phase and adding a statistical analysis based on the Shapley values to an XGBoost (Chen et al., 2015) classifier which models breakdown behavior.

The paper is structured as follows. It first focuses on exploring a case where a new predictive and diagnostic framework can be applied. Second, it proposes a methodology for collecting data, modeling the machine behavior (predictive), and explaining the predictions by this model (diagnosis). Lastly, it presents the results and conclusions while providing guidance for applicability to any rotary compressor unit with a surrounding network of sensor data.

2. Case description

For the practical testing of the methodology, we focus on two gas compressor units (CU1 and CU2) that are situated within an oil refinery. The compressors are part of an Atmospheric Residue DeSulpherizer (ARDS) unit that is crucial to plant operations. We provide a generalized schematic of the surrounding process in Fig. 1, where the position of the recycle compressors is circled in red. These units operate in a parallel and redundant fashion, in order to guarantee the operating capabilities of the refinery.

Following their crucial position within the process, a precise prediction of performance, breakdowns, and maintenance requirements is paramount. Generally, compressor performance and maintenance maps are needed in order to achieve the necessary accuracy in assessing these requirements. These mappings and specifications are usually proprietary information held by the supplier and can be unknown to the user of the equipment (Liu and Karimi, 2020). Historically, the plant operators supervising this unit have had no success in their assessments, lacking either data from the supplier or interpretability in statistical analysis. It is apparent why there is no buy-in for yet another proprietary or 'black box'-solution.

Table 1
Data sources (per CU).

	Sensors	Unusable	Frequency	Volume change
Compressor Units	10	0	Second	1 / 900
Plant Monitoring	72	5	Minute	1 / 15
Inspection reports	2	0	Monthly	2880/1

Table 2
Data types.

Source:	Features:	Units:
Compressor Units	Temperatures	C
	Pressures	BAR
	Rotational Speed	RPM
	Shaft Movement (x,y,z)	ôm
Plant Monitoring	Temperatures	C
	Pressures	BAR
	Hydrogen Concentrations	%
	Amine Concentrations	%
	Gas Flow Rates	m^3/h
	Gas Bleed Rates	m^3/h
	Buffer Reservoir Levels	%
	Discharge Rates	m^3/h and Kg/h
	Valve States	%
Inspection reports	Time since last inspection	days
	Time since last repair	days

To combat the above pitfall, real data can be leveraged to model the compressor behavior. More specifically, through health monitoring and vibration analysis, it is possible to identify problem periods, performance issues, and potential root causes (Henry and Lalanne, 1974). For different kinds of problems, different trends in vibration are expected. These trends, often based upon the eccentricity of the rotor shaft (Gruwell et al., 1998), can serve as a proxy for issues within the compressors (Kirk and Guo, 2003).

In terms of operating performance of the units, CU1 has been exemplary in its performance and has been reliably operating at near full capacity during its entire lifespan. CU2 however, had multiple breakdowns and high-risk shutdowns over the same time period. The set-up, as described, provides a unique opportunity for comparing the two units.

CU1 and CU2, being redundant parallel units, each have an identical collection of sensors tracking their operating characteristics built into the machines themselves. This primary data source consists of 10 sensor values with readings that are logged every second, amounting to over 600 million sensor readings for the entire period (2005–2015). Secondly, data is collected by the plant monitoring system, which contains information of all input, output, and operating conditions for the two compressor units. The Plant Monitoring System (PI) comprises 72 sensors (i.e. pressures, temperatures, gas concentrations,...) with readings logged on a per-minute basis and contains over 5 million sensor readings for a period ranging from 2005 to 2015. These combined time series provide us with insights on the state of the refinery and the processes surrounding the compressor units. Third, we consider data from manual inspections, in the form of inspection reports.

The data sources and their characteristics are summarized in Table 1, which also provide insight into the results of the aggregation strategy we propose below. For a summary of the features contained within the sources, we refer the reader to Table 2.

The rest of this paper implements a supervised machine learning algorithm, augmented by statistical techniques to improve in-

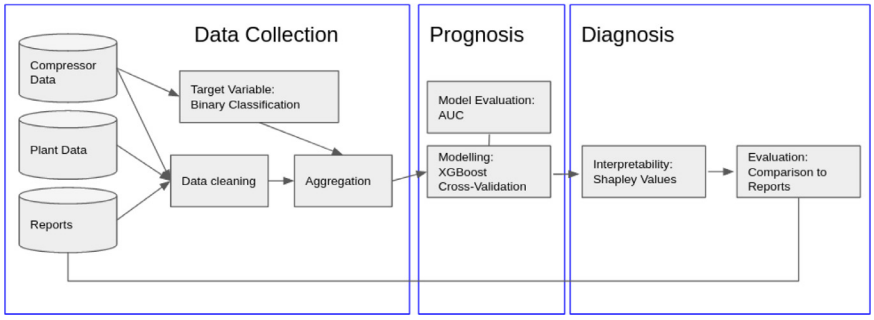

Fig. 2. Proposed methodology.

terpretability and is followed by an example interpretation of the model from both units. The improvement in interpretability leads to an increase in "the degree to which a human can understand the cause of a prediction." Miller (2017) and subsequently to extra buy-in from the plant operators.

3. Methodology

The proposed methodology consists of three main steps (Jardine et al., 2006); Data Collection, Prognosis, and Diagnosis. These three steps are further split into smaller units. Data collection consists of gathering the data from the sources, cleaning it, and transforming it into a usable format through aggregation. The predictions comprise the modeling, with cross-validation and evaluation of the model performance. Lastly, diagnosis consists of calculating the interpretability metrics and comparing these to the inspection reports. An overview of this flow can be found in Fig. 2.

3.1. Data collection and treatment

Since data is coming from different sources on different time granularity, we have to define a common frequency. In discussions with plant operators, it is deemed necessary to shut down failing systems within 15 min and to predict maintenance requirements at least 24 h in advance. Thus, we aggregate all data to the 15-minute level, keeping minimum, maximum, mean, and standard deviation as aggregate metrics. In practice, this leads to having less observations for model training, while keeping 4 features per sensor reading in order to minimize the information loss from the aggregation step (Roy et al., 1982). The expected impact on data volume after aggregation can be found in Table 1.

The primary data source is leveraged to calculate the target metric, namely shaft eccentricity. In literature, when analyzing rotating equipment, eccentricity of the shaft or vibration amplitude is often used as a predictive measure for machine failure (Lei et al., 2018; Deraemaeker and Preumont, 2006). It relates closely to the remaining useful life (RUL) of the machine. When the eccentricity, usually expressed in micrometers, is 0, the shaft operates as expected in the center of the bearing. If the eccentricity rises, the shaft starts rubbing against the inner lining of the bearings which causes local heating and wear. This deterioration leads to further vibrations and an eventual machine breakdown. The expectation for modeling is that the above factors allow for an accurate estimation of breakdowns (Nembhard et al., 2014).

To reduce the complexity and to focus solely on the interpretability of breakdowns or maintenance reasons, we opt to split the target metric into two categories. These categories, 'High Risk' and 'Normal Operation' are defined by plant operators and the supplier of the compressor unit. As soon as the eccentricity reaches a threshold value of 90 micrometers, we define it as 'High Risk'. Any value under this threshold is classified as 'Normal Operation'. This threshold was chosen according to the specifications provided by the compressor manufacturer, machinery of this caliber almost always comes with operating thresholds regarding temperature, shaft eccentricity, rotational speeds, and pressures. We choose to limit this research to the threshold for shaft eccentricity, this limitation is set after an initial inspection of the data. Shaft eccentricity is the only metric that violates the preset thresholds.

In order to generate a predictive model that can incite a safe shutdown or trigger a maintenance of the machine, our target variable needs to be well defined. In conjuction with the plant operators, we opt for the following definition: "Do we encounter a sustained high risk period in the next 24 h of operation?". To clarify, we choose to transform the initial target variable of 'Shaft eccentricity (ôm)' to a boolean variable that answers: "Is the compressor at risk of failure within the next 24 h based on operating thresholds set by the supplier?"

This choice for a binary target variable has benefits and drawbacks. It simplifies our modeling problem from a complex regression, to a binary classification with a predictive element. It does so while maintaining the necessary information for capturing important characteristics in terms of breakdowns and maintenance requirements. Moreover, model evaluation is also simplified; for binary classification problems, we can rely on the proven AUC as a performance metric (Ling et al., 2003). The downside is the possible effect on real-life performance, as not all information contained in the original eccentricity is retained within the transformed target. the features in the primary source include; rotational speed, electrical currents, temperatures and pressures within the compressor unit.

In the secondary data source, we find plant operating data from systems surrounding the compressor unit, as seen in Fig. 1. We note that during initial inspection, 5 of the sensors on both units are unusable due to inconsistent measurements or missing data. This problem was identified autonomously by the plant monitoring system. To avoid issues regarding these sensor readings, it was decided to exclude the affected sensors from the study. All other sensor data was subsequently analyzed for corrupt readings, but the list of corrupted sensors from the plant monitoring system proved to be exhaustive and no extra data corruption was identified. The remaining 68 sensors were aggregated according to aforementioned strategy and added to the model.

In the last data source, we find inspection dates and monthly reports. In these 120 reports, the outcome and the time of inspection can be found. The reports are integrated in our model as the metrics: 'time since last inspection' and 'time since last fault detection'. This allows us to adapt the monthly information flow to a 15-minute basis, increasing the measurement fequency. During the time period of our gathered data, we note that CU2 has incident reports of complete shutdowns due to failures. These inspection reports will be withheld from the training data. We also withhold the data from the three days before these incident reports, as it

Table 3
Feature summary per CU.

	Features	Observations
Compressor Units	36 (+1)	350,400
Plant Monitoring	68	350,288
Inspection reports	2	350,400
CU1 total	306 (+1)	350,288
CU2 total	306 (+1)	350,288
Combined total	306 (+1)	700,576
Withheld validation	306 (+1)	2,016

will be used as validation data to test the final performance of the methodology. Indeed, we also compare the results of our methodology with the results of these reports in order to verify if the diagnostic step can provide value to plant operators. The buy-in from the operators depends greatly on the overlap between our diagnostics and their incident reports.

As noted earlier, aggregation of the measurements is required to align the time frames to a common ground. This aggregation reduces our total number of observations from billions to 350,400 before cleaning (approximately 10 years of data on a 15-minute interval). In terms of features, for the primary and secondary data source, we keep 4 features per variable, this amounts to a total of 306 independent variables (compressor data, plant-data, inspection reports) and 1 dependent variable. The final form of our data consists of 350,288 observations on 15 min intervals. For a clearer detail of this form, we refer to Table 3.

3.2. Model selection

Since our main goal is not to identify a best algorithm, but to provide a comparison across units and to provide interpretability, we opt to use XGBoost (Chen et al., 2015). XGBoost is an implementation of gradient boosted decision trees designed for speed and performance, which has proven effective across multiple industries and use-cases (Dhaliwal et al., 2018; Zhang et al., 2018). It is known for robust predictions and works with non-linear relationships in data that has high degrees of dimensionality and correlation (Nielsen, 2016). To avoid overfitting the model and to increase accuracy, we employ a 10-fold cross validation (Schaffer, 1993) while respecting the continuity principle of time series (Bergmeir and Benítez, 2012). This comes down to respecting the chronological order of the data instead of doing purely random splits, to avoid data leakage from future events. For the practical implementation, we refer to the XGBoost Python package (Chen et al., 2015). Furthermore, as the main goal of our research is to compare models, we opt for the default hyperparameters.

3.3. Model training and comparison

In order to compare the compressor units (CU1 and CU2), the following methodology is put forward:

1. Train models on CU1 and CU2 seperately
2. Cross-check the model performance, use model for CU1 to predict CU2 and vice-versa
3. Compare performances of these models
4. Train model on data from both CU1 and CU2
5. Compare feature importances and feature impacts across models
6. Investigate the reports for the diagnostic analysis

The model performances from step 1 through 4 confirm whether or not the predictive aspect of Condition-based maintenance is proven. It stands to reason that accurate models in the predictive phase are required in order to proceed with diagnostics and maintenance strategies.

For the added interpretability (diagnosis) however, we need stricter requirements. The single model for both units, and the cross-check between the singular units, require accurate results as well. This verifies that we can depend on the comparison of the models and the claimed similarity of both compressor units. Finding a general model for both units, and using the model trained on CU1 for predicting CU2 (and vice-versa), provides us with the needed reference for proceeding with the creation of interpretability metrics across the units.

After an accurate general model has been established, we investigate it in terms of explanatory variables and variable impacts through the use of Shapley values. When successful, the insights from these driving factors provide us with a proper diagnosis for the issue. Lastly, the identified driving factors are linked back to the incident reports for a final verification.

3.4. Model performance evaluation

We use the area under the receiver operating characteristic curve (AUC or AUROC) in order to evaluate the peformance of our chosen models. AUC is a superior metric of model evaluation than accuracy. The AUC is defined as follows:

$$AUC = \int_0^1 \frac{TP}{TP+FN} d\frac{FP}{FP+TN} = \int_0^1 \frac{TP}{P} d\frac{FP}{N} \tag{1}$$

Where;

TP stands for True Positive predictions,
TN for True Negatives,
FN for False Negatives,
TN for True Negatives,
N for Negatives in the data, and P for Positives

Intuitively, the AUC can be seen as the probability that a randomly chosen positive case is ranked higher than a randomly chosen negative case. The AUC has useful range from 0.5 to 1. The former indicates a naive or random prediction, while the latter indicates a perfect model.

3.5. Feature importance and explanation

Accurate predictions are typically obtained by applying learning machines to complex feature spaces. Unfortunately, such feature spaces are hardly accessible to human intuition. From this, it follows that they cannot easily be used to gain insights about the application domain. Practical applications often resort to linear models in combination with variable selection, thereby sacrificing predictive power for presumptive interpretability. To combat this issue, we propose to utilize Shapley values instead (Shapley, 1953). Shapley values have recently gained a lot of traction to combat the limitations of the 'black box solutions'.

The Shapley value is a way to distribute the total prediction to the variables as a sort of payout. According to the Shapley value theory, the share that a given variable i gets given a certain prediction is:

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\ (n-|S|-1)!}{n!} (v(S \cup \{i\}) - v(S)) \tag{2}$$

where n is the total number of variables and the sum extends over all subsets S of N not containing variable i.

Intuitively, Shapley values tell us how to distribute the fair share of predictive power between the variables. Selecting the most important variables based on Shapley values is simply selecting those with the highest absolute contributions. The cost of these intuitive interpretations is a very high computation time during the model training phases, it is very costly to calculate even an approximate solution as we need to compare the contribution of each feature to the entire feature space. It is important to note that the

Table 4
Model performance (AUC%).

Data source:	CU1	CU2	CU1+2	Validation
Model CU1	90.7	82.1	86.4	80.3
Model CU2	81.4	88.2	84.4	87.2
Model (CU1+2)	92.8	89.4	91.1	90.1

Table 5
Confusion matrix.

		Predicted		
		High	Normal	Total
Actual	High	1332	12	1344
	Normal	189	483	672
	Total	1521	495	2016

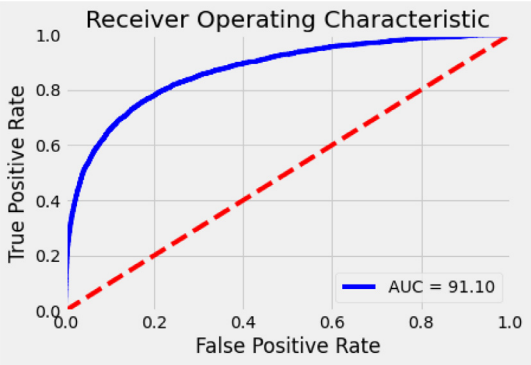

Fig. 3. AUC of the combined model on the combined dataset.

calculation of this value during normal refinery operations after the model training phase, would be significantly reduced and still leave adequate time for predictions of shutdowns. The diagnostic calculations can realistically be done within a minute of making a 'High Risk' prediction, while 'Normal Operation'-predictions do not require diagnostic insights. For the calculation of this metric, we rely on the SHAP package in Python (Lundberg et al., 2020).

4. Results

4.1. Model performance

Initially, we train our models on the CU's seperately. For this initial verification step, we find that our model choice performs well, providing an AUC of 88.2% in the worst case, in line with other research. The specific models that were trained and tested on the same data score 90.7(CU1), 88.2(CU2) and 91.1(CU1+2). The cross-tests between the CU's, are the worst performing models, but they still provide an average AUC of 81.75.

The results of the modeling step are summarized in Table 4. In this table, the column labels represent on which data the model has been tested, the row labels represent on which data the model has been trained. The result of the 'combined model', trained on CU1+2 data and applied to CU1+2 data, is shown in Fig. 3. In the last column of Table 4, we report the real-life performance of the models. This is the performance of each model on the withheld data confirmed to contain a shutdown. Again we see that the combined model (CU1+2) performs best. To further show the performance on this validation set, we also provide the confusion matrix in Table 5, where we see the expected real-life performance in terms of true positives, true negatives, false positives, and false negatives. We deduce from the confusion matrix, that we have a very low chance of missing 'High risk'-situations, with only 12 false negatives. The number of false positives is higher, but excusable as every period leads to an eventual breakdown. Furthermore, the subsequent diagnostic analysis that is triggered by a false positive, even when it is found to be superfluous, can still provide valuable insights into the workings of the compressor unit and the predictive model.

On the whole, we have to remember that the 2016 observations in this validation set, represent seven 3-day periods in which we have a guaranteed breakdown. In light of this, we expect 1344 'High risk' values and 672 'Normal operation' values, which are the row sums in Table 5. From this table, we can calculate the following metrics:

- Sensitivity = TP / (TP + FN): 99.1%
- Specificity = TN / (FP + TN): 71.9%
- Precision = TP / (TP + FP): 87.6%
- Accuracy = (TP + TN) / (P + N) : 90.0%
- F1-score = (2TP) / (2TP + FP + FN) : 93.0%

The confusion matrix leads us to the same conclusion as the AUC. The predictive performance of the model is high, and we can continue our investigation through the use of Shapley values.

It is important to note that specifically training a model on a single compressor unit achieves a slightly lower AUC than training a single model on both units, while cross-testing model performance of one CU on the other has a lower performance. These findings are as expected, the CU's are identical and we have noted that the input data does not deviate much between the two compressor units. Moreover, the specific behavior traits that we are attempting to identify in our diagnosis are highly likely to be the cause of the dips in the cross-tested performance. Remember, CU1 has not had failures and breakdowns, unlike CU2. We also note that the addition of more data and noise, by combining the data from the two compressors, adds performance. Noise in this context needs to be interpreted as detectable characteristics in sensors (i.e. rounding and calibration deviations), adding data from multiple compressors allows the model to focus on the underlying signal. This is in line with other modeling research, as the addition of data and noise is proven to be beneficial when generalizing complex models (Audhkhasi et al., 2016). As a final remark, it is logical that the third column in Table 4 is approximately equal to the mean of columns 1 and 2, barring a small random selection bias, as the combined dataset represents nothing more than an addition of both with equal weights.

The performance of these fitted models, which is confirmed on the withheld validation data from the three days before the incident reports, allows us to continue with the diagnostic step. This is shown in column 4 of Table 4. High accuracy on this validation set was a condition for the continuation of our metholodogy. The AUC of the combined model for this specific case was 90.1%.

4.2. Feature importance and explanation

We calculate the top 10 influencing parameters for the combined model through Shapley values, and show them in decreasing order of importance, in Fig. 4. We adjusted the variable names to human interpretable input. Generally, plant-proprietary sensor names need to be translated through a data dictionary in order for interpretation to take place. The data shows the following factors that contribute to a higher risk of failure, in order of descending impact. We also add the expected value (low, high, or dependent on utilization) under normal conditions:

- An increase in outlet temperature of the unit (low).
- An increase in temperature variance on the outlet side over 15 min (low).
- An increase in inlet temperature in the unit (low).

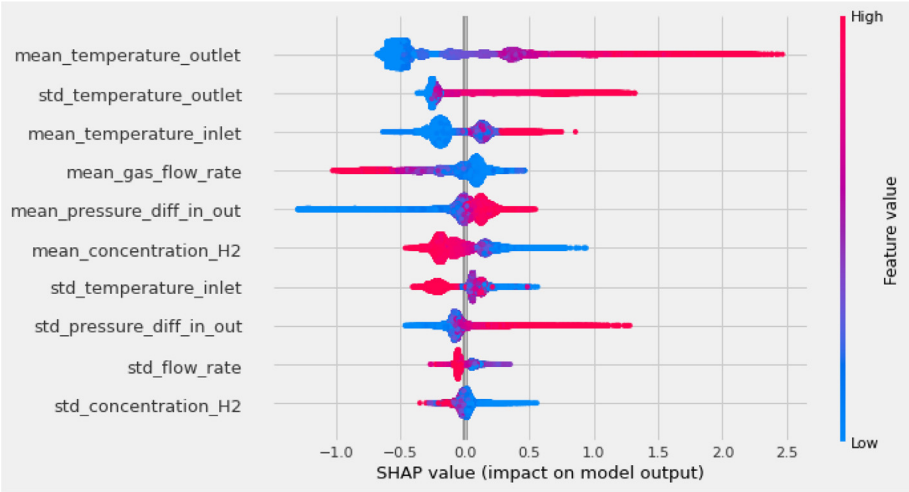

Fig. 4. Shapley values for the combined model.

- A reduction in incoming gas flow rate (depends on utilization).
- A higher pressure differential between input and output(depends on utilization).
- A lower concentration of the H_2 gas (high).
- A decrease in temperature variance on the inlet side over 15 min (low).
- An increase in pressure differential variance over 15 min (low).
- A small decrease in flow rate variance over 15 min (low).
- A small decrease in H_2 concentration variance over 15 min (low).

In terms of the variable spread (or higher concentrations of points on the figure), we note that the denser areas signify a higher amount of observations showing those readings. This higher amount of observations corresponds to “steady states” or normal operation of the compressor, these are obviously more abundant than the breakdown or at-risk states. We also see in Fig. 4 that the “normal operating characteristics” overlap with higher density areas and either have a slightly net negative or neutral effect on the prediction. In general, we expect all variances over 15 min to be low and to subsequently have a low impact, as confirmed by the plot. These machines are part of a slow moving process, entailing that their operating conditions should not show big changes over 15-minute windows. A high variance over these windows signifies a rapid and significant change in the process and this carries a high predictive value.

The plant operators confirm that these findings match their intuition before we move on to comparing the breakdowns to the inspection reports. They note, however, that the last 2 variables in the top 10 do not match their intuition. We note that it is of value in these cases to plot the interaction effect of the top indicator together with the counterintuitive indicator, in order to see hidden relationships. To examine this interaction, Fig. 5 visualizes the combined effect of the outlet temperature with the H_2 variance. This interaction between the features shows that the impact of the H_2 variance is to be taken into account together with the outlet temperature; when the outlet temperature is nominal (low), the impact of the H_2 variance is severely lessened. Vice-versa, as the temperature increases, we see that the impact of the H_2 variance increases. During breakdowns, or when maintenance is required, less variance in the H_2 concentration is combined with a high temperature. In this combined situation, pollution has been building up for some time (low variance, but also low concentration) and has already impacted outlet temperature. The investigation of the interaction effect between flow rate variance and outlet temperature shows a similar pattern.

4.3. Comparisons to inspection reports

For the comparisons, we proceed as follows; we take the diagnostic data from the inspection reports and compare them to the calculated Shapley values and predictions for that period. The combination of previous research on compressor breakdowns, combined with plant operator insights is compared to the statistical output. As previously mentioned, these two factors should be in line in order to generate the necessary buy-in.

For the first group of shutdowns(3) in our validation set, we find reports that point towards damage of the shaft and impeller. Our models indicate that the top predictors for this event were related to the operating efficiency of the compressor: pressure ratio and flow rate of the gas. In order to keep stability in the compressing system, these parameters can be adjusted with control valves. When incoming gas flow rate is too low, at constant pressure ratio, the compressor will be driven towards a surge working region. Surging causes higher pressure in the pipe receivers than gas pressure in the compressor and therefore reversal flows are induced (Bloch and Geitner, 1997). These reversal flows heavily damaged the equipment. Given the conclusion of the inspection reports and our identified variables, the plant operators deem this a successful statistical diagnosis of the breakdown cause.

For the second group of shutdowns(3) found in the reports, we find that the model identifies temperature, and hydrogen percentage as the main contributor to the predictions, together with radial movement of the shaft. These parameters indicate a bearing issue as mentioned earlier, the hydrogen percentage playing a role indicates that the compressor unit underwent a period where fouled gas entered the unit. This fouled gas built up pollutants around the shaft, causing an increased wear on the bearings and an increase in temperature (Jombo et al., 2018). For this breakdown type, plant operators also find our statistical diagnosis to be adequate.

For the last group of shutdowns(1), the predictions were clear, but the diagnostic results were inconclusive. The indicated important variables were not linked to reports by the plant operators. The predictions were a mixture of the above two causes, indicating that there might have been an occurrence of both factors combined.

In summary, we find that our model was able to accurately predict all withheld breakdowns. We also conclude that diagnostics were feasible on the majority of these breakdowns and that the

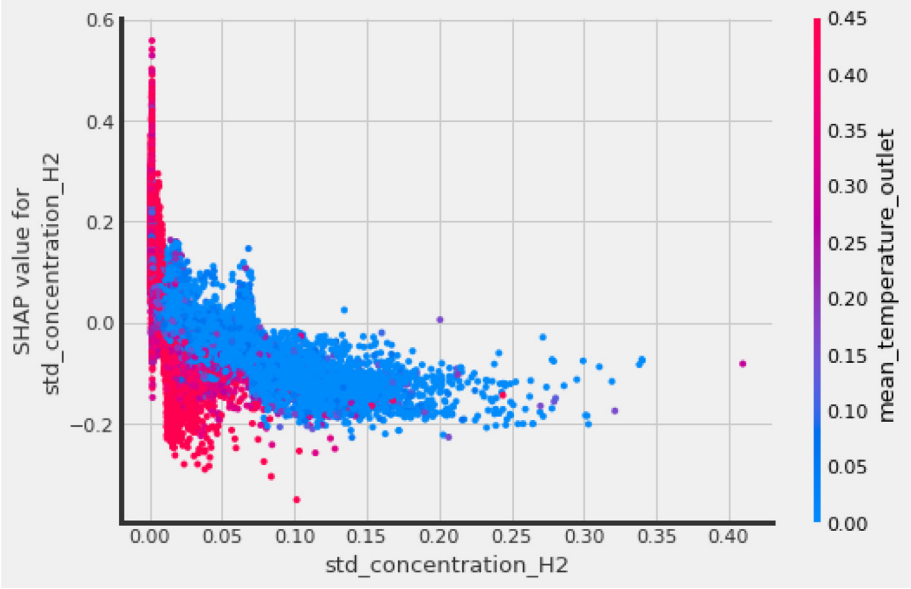

Fig. 5. Interaction effect of H_2 variance with outlet temperature.

insights from these statistical methods were in line with the inspection reports and plant operator intuition. In any case, the diagnostic step adds statistically supported insights to a 'High Risk' prediction of the model and shows the added value of the technical diagnosis on top of or in support of inspection reports.

5. Conclusion

The main topic of this paper was utilizing predictive modeling supported by statistical methods in order to achieve a predictive model with an added diagnostic analysis of machinery breakdowns. While we found that a lot of work has already been done, industrial applications are lagging. The 'black box'-nature of many statistical solutions was identified to be one of the leading causes of this lag and for the lack of buy-in of plant operators. Therefore, we modeled the behavior of multiple compressor units with a supervised machine learning model, achieving a high baseline predictive value. In addition to this, we augmented the predictive model with explanatory Shapley values in order to increase the diagnostic value and to reduce the 'black box'-nature generally associated with these techniques.

First, regarding the modeling aspect of this research, we conclude that we are able to accurately model the behavior of both compressor units and are also able to construe one generalized model for both units. Due to the creation of these models with a high predictive value, plant operators can now foresee periods with a high risk. As expected from the machine learning techniques, they provide high predictive performance. The chosen algorithm proved to be particularly effective for this case.

Second, regarding the diagnostic phase, the enrichment of model insights with Shapley values in order to provide a more accurate view towards the plant operators has succeeded. The methodology provided accurate insights into the discussed breakdowns through comparing inspection reports with the results from the statistical analysis. The Shapley values identified the leading causes of high-risk situations and can be utilized in the future to provide extra support for the plant operators in their inspection and maintenance.

This improved diagnostic value has many benefits. It leads to better maintenance planning, through insights into the timing of possible breakdowns of the compressors. It avoids the risk of downtime for the refinery, through a more accurate planning of maintenances and the possibility to anticipate breakdowns. It enables more to-the-point repair interventions, leading to faster execution of maintenance works. All of these factors lead to cost-savings and efficiency gains for the refinery, as they move away from the traditional run-to-failure models and towards a CBM approach (Jardine et al., 2006).

Future research

While this framework serves as a ground for future diagnostic efforts, we would like to point out some initial limitations in our work. First, we did not fine-tune models specifically for the compressor unit, so we expect predictive value to be higher still. In terms of modeling, we also see the possibility for a non-binary model in practical applications and for other model implementations such as neural networks.

Second, we limited the diagnostic aspect to the generalized model, as it was the best performing one for both compressors. If we had found CU2 to be modeled better by a dedicated model, it would stand to reason to utilize that model for the diagnostic aspect as well.

Third, our definition of target variable was driven by input from the supplier of the compressor and plant operators. For specific use-cases, we highly recommend changing the definition of the target variable. We did not investigate the impact of changing the time frame with which we look ahead for possible breakdowns, nor did we investigate the impact of choosing different aggregation windows other than 15-minutes. These last points were not investigated as analyzing these amounts of data is very computationally intensive and resources were limited. The expected increase in accuracy did not weigh up to the expected increase in computational difficulty, as our baseline already performed adequately.

Declaration of Competing Interest

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

CRediT authorship contribution statement

Bram Steurtewagen: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - review & editing, Visualization. **Dirk Van den Poel:** Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

All persons who have made substantial contributions to the work reported in the manuscript (e.g., technical help, writing and editing assistance, general support), but who do not meet the criteria for author-ship, are named in the Acknowledgments and have given us their written permission to be named. If we have not included an Acknowledgments in our manuscript, then that indicates that we have not received substantial contributions from non-authors.

References

Ahmad, R., Kamaruddin, S., 2012. An overview of time-based and condition-based maintenance in industrial application. Comput. Ind. Eng. 63 (1), 135–149. doi:10.1016/j.cie.2012.02.002. http://www.sciencedirect.com/science/article/pii/S0360835212000484.

Aravinth, S., Sugumaran, V., 2018. Air compressor fault diagnosis through statistical feature extraction and random forest classifier. Prog. Ind. Ecol. Int.J. 12 (1-2), 192–205.

Audhkhasi, K., Osoba, O., Kosko, B., 2016. Noise-enhanced convolutional neural networks. Neural Netw. 78, 15–23.

Bergmeir, C., Benítez, J.M., 2012. On the use of cross-validation for time series predictor evaluation. Inf. Sci. 191, 192–213.

Bloch, H.P., Geitner, F.K., 1997. Practical Machinery Management for Process Plants: Volume 2: Machinery Failure Analysis and Troubleshooting. Elsevier.

Bloch, H.P., Geitner, F.K., 2019. Chapter 1 - machinery maintenance: an overview. In: Bloch, H.P., Geitner, F.K. (Eds.), Machinery Component Maintenance and Repair (Fourth Edition). In: Practical Machinery Management for Process Plants. Gulf Professional Publishing, pp. 3–12. doi:10.1016/B978-0-12-818729-6.00001-0. https://www.sciencedirect.com/science/article/pii/B9780128187296000010.

Cao, L., 2010. Domain-driven data mining: challenges and prospects. IEEE Trans. Knowl. Data Eng. 22 (6), 755–769. doi:10.1109/TKDE.2010.32.

Carvalho, T.P., Soares, F.A.A.M.N., Vita, R., da P. Francisco, R., Basto, J.P., Alcalá, S.G.S., 2019. A systematic literature review of machine learning methods applied to predictive maintenance. Comput. Ind. Eng. 137, 106024. doi:10.1016/j.cie.2019.106024. https://www.sciencedirect.com/science/article/pii/S0360835219304838.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al., 2015. XGBoost: extreme gradient boosting. R package version 0.4-2 1 (4).

Deraemaeker, A., Preumont, A., 2006. Vibration based damage detection using large array sensors and spatial filters. Mech. Syst. Signal Process. 20 (7), 1615–1630. doi:10.1016/j.ymssp.2005.02.010. http://www.sciencedirect.com/science/article/pii/S0888327005000312.

Dhaliwal, S.S., Nahid, A.-A., Abbas, R., 2018. Effective intrusion detection system using XGBoost. Information 9 (7), 149.

Engel, S., Gilmartin, B., Bongort, K., Hess, A., 2000. Prognostics, the real issues involved with predicting life remaining. In: 2000 IEEE Aerospace Conference. Proceedings (Cat. No.00TH8484), Vol. 6, pp. 457–469vol.6. doi:10.1109/AERO.2000.877920.

Gelgele, H.L., Wang, K., 1998. An expert system for engine fault diagnosis: development and application. J. Intell. Manuf. 9 (6), 539–545. doi:10.1023/A:1008888219539.

Gruwell, D.R., Zeidan, F.Y., et al., 1998. Vibration and eccentricity measurements combined with rotordynamic analyses on a six bearing turbine generator. In: Proceedings of the 27th Turbomachinery Symposium. Texas A&M University. Turbomachinery Laboratories.

Handelman, G.S., Kok, H.K., Chandra, R.V., Razavi, A.H., Huang, S., Brooks, M., Lee, M.J., Asadi, H., 2019. Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. Am. J. Roentgenol. 212 (1), 38–43.

Hashemian, H.M., 2010. State-of-the-art predictive maintenance techniques. IEEE Trans. Instrum.Meas. 60 (1), 226–236.

Heng, A., Zhang, S., Tan, A.C.C., Mathew, J., 2009. Rotating machinery prognostics: state of the art, challenges and opportunities. Mech. Syst. Signal Process. 23 (3), 724–739. doi:10.1016/j.ymssp.2008.06.009. http://www.sciencedirect.com/science/article/pii/S0888327008001489.

Henry, R., Lalanne, M., 1974. Vibration analysis of rotating compressor blades. J. Eng. Ind. 96 (3), 1028–1035. doi:10.1115/1.3438403.

Isermann, R., 2011. Fault-Diagnosis Applications: Model-Based Condition Monitoring: Actuators, Drives, Machinery, Plants, Sensors, and Fault-Tolerant Systems. Springer Science & Business Media.

Jardine, A.K.S., Lin, D., Banjevic, D., 2006. A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mech. Syst. Signal Process. 20 (7), 1483–1510. doi:10.1016/j.ymssp.2005.09.012. http://www.sciencedirect.com/science/article/pii/S0888327005001512.

Jombo, G., Pecinka, J., Sampath, S., Mba, D., 2018. Influence of fouling on compressor dynamics: experimental and modeling approach. J. Eng. Gas Turbines Power 140 (3).

Kadlec, P., Gabrys, B., Strandt, S., 2009. Data-driven soft sensors in the process industry. Comput. Chem. Eng. 33 (4), 795–814.

Kirk, R.G., Guo, Z., 2003. Expert system source identification of excessive vibration. Int. J. Rotating Mach. 9 (2), 63–79.

Lee, J., Kao, H.-A., Yang, S., 2014. Service innovation and smart analytics for industry 4.0 and big data environment. Procedia Cirp 16, 3–8.

Lei, Y., Li, N., Guo, L., Li, N., Yan, T., Lin, J., 2018. Machinery health prognostics: a systematic review from data acquisition to RUL prediction. Mech. Syst. Signal Process. 104, 799–834. doi:10.1016/j.ymssp.2017.11.016.

Ling, C.X., Huang, J., Zhang, H., et al., 2003. AUC: a statistically consistent and more discriminating measure than accuracy. In: Ijcai, Vol. 3, pp. 519–524.

Liu, Z., Karimi, I.A., 2020. Gas turbine performance prediction via machine learning. Energy 192, 116627. doi:10.1016/j.energy.2019.116627. https://www.sciencedirect.com/science/article/pii/S0360544219323229.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable ai for trees. Nat. Mach. Intell. 2 (1), 2522–5839.

McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D.J., Barton, D., 2012. Big data: the management revolution. Harvard Bus. Rev. 90 (10), 60–68.

Messalas, A., Kanellopoulos, Y., Makris, C., 2019. Model-agnostic interpretability with Shapley values. In: 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA). IEEE, pp. 1–7.

Miller, T., 2017. Explanation in artificial intelligence: insights from the social sciences. CoRR. arXiv preprint arXiv:1706.07269.

Mobley, R.K., 2002. An Introduction to Predictive Maintenance. Elsevier.

Moraru, A., Pesko, M., Porcius, M., Fortuna, C., Mladenic, D., 2010. Using machine learning on sensor data. J. Comput. Inf. Technol. 18 (4), 341–347.

Nembhard, A.D., Sinha, J.K., Pinkerton, A.J., Elbhbah, K., 2014. Combined vibration and thermal analysis for the condition monitoring of rotating machinery. Struct. Health Monit. 13 (3), 281–295.

Nielsen, D., 2016. Tree boosting with XGBoost-why does XGBoost win “every” machine learning competition?

Pyle, D., San José, C., 2015. An executive's guide to machine learning. McKinsey Q. 3, 44–53.

Roy, J.R., Batten, D.F., Lesse, P., 1982. Minimizing information loss in simple aggregation. Environ. Plann. A 14 (7), 973–980.

Schaffer, C., 1993. Selecting a classification method by cross-validation. Mach. Learn. 13 (1), 135–143.

Shah, D., Wang, J., He, Q.P., 2020. Feature engineering in big data analytics for IoT-enabled smart manufacturing - comparison between deep learning and statistical learning. Comput. Chem. Eng. 141, 106970. doi:10.1016/j.compchemeng.2020.106970. https://www.sciencedirect.com/science/article/pii/S0098135420300363.

Shapley, 1953. A value for n-Person Games. Princeton University Press.

Sobie, C., Freitas, C., Nicolai, M., 2018. Simulation-driven machine learning: bearing fault classification. Mech. Syst. Signal Process. 99, 403–419. doi:10.1016/j.ymssp.2017.06.025.

Sohaib, M., Kim, J.-M., 2018. Reliable fault diagnosis of rotary machine bearings using a stacked sparse autoencoder-based deep neural network. Shock Vibr. 2018.

Steurtewagen, B., Van den Poel, D., 2019. Root cause analysis of compressor failure by machine learning. In: 2019 Petroleum and Chemical Industry Conference Europe (PCIC EUROPE), pp. 1–5. doi:10.23919/PCICEurope46863.2019.9011628.

Sundararajan, M., Najmi, A., 2019. The many Shapley values for model explanation. Comment: 9 pages. http://arxiv.org/abs/1908.08474.

Tsang, A.H., 1995. Condition-based maintenance: tools and decision making. J. Q. Maint. Eng. 1 (3), 3–17.

Wen, L., Li, X., Gao, L., Zhang, Y., 2017. A new convolutional neural network-based data-driven fault diagnosis method. IEEE Trans. Ind. Electron. 65 (7), 5990–5998.

Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., Si, Y., 2018. A data-driven design for fault detection of wind turbines using random forests and XGBoost. IEEE Access 6, 21020–21031.