Unsupervised Machine Learning Methods to Estimate a Health Indicator for Condition Monitoring using Acoustic and Vibration Signals: A Comparison Based On a Toy Data Set From a Coffee Vending Machine

Yonas Tefera[1], Maarten Meire[1], Stijn Luca[2], and Peter Karsmakers[1]

[1] KU Leuven, Department of Computer Science, DTAI-ADVISE, Kleinhoefstraat 4 2440 Geel, Belgium
{yonas.tefera,maarten.meire,peter.karsmakers}@kuleuven.be
[2] Ghent University, Department of Data Analysis and Mathematical Modelling, Coupure Links 653, 9000 Gent, Belgium
stijn.luca@ugent.be

Abstract. Automating the task of assessing an asset's status based on sensor data would not only relieve trained engineers from this time intensive task, it would also allow a continuous follow-up of assets, potentially resulting in a fine-grained view on the asset's status. In this work three unsupervised machine learning approaches that define a Health Indicator (HI) based on acoustic and vibration signals were empirically assessed. Such a HI indicates the similarity of the current measured state to the baseline/normal operational state. The lower the HI the worse the asset's condition is. In this way the condition of an asset can be automatically monitored. Gaussian mixture models, Variational Autoencoders (VAE) and One Class Support Vector Machine (OC-SVM) were considered for this task. To enable the empirical assessment, a toy data set was created in which vibration and acoustic data was recorded simultaneously from a coffee vending machine with rotating elements in the bean grinder and water pump, relatively fast changing levels in the water and bean containers, and several stages in the coffee making cycle was used. Experiments were performed to analyse whether subtle changes in the sensor data due to changing container levels could be automatically detected and discriminated. Moreover, it was studied if a change could be rooted back to a cause (being a low level in the water or bean container). A set of time and spectral domain features were extracted and considered, while experiments were also performed by fusing the acoustic and vibration signals. The applied models achieved a comparable performance in terms of detecting low and empty container levels, with VAE using convolutional layers and OC-SVM achieving a further better discrimination of the different container levels when using the fused signals. It was also determined that the root cause of a level change can be determined by looking at the HI in the various stages.

Keywords: Gaussian Mixture Models (GMMs) · One Class Support Vector Machine (OC-SVM) · Variational Autoencoder (VAE) · Condition Monitoring · Health Indicator · Data Driven Modeling.

1 Introduction

In condition monitoring assets are continuously tracked by sensors to follow-up their operational status and identify possible changes that might indicate future faults. In this way interruption due to failure of the asset is prevented and maintenance can be applied only when it is required which reduces the down time.
Nowadays, a multitude of sensors are installed to monitor an assets condition. This work focuses on the use of an acoustic sensor, which is contactless and retrofittable, and a vibration sensor, which requires contact with the asset. Such sensors are typically applied when the asset contains rotating elements. Manually inspecting the large amount of data these sensors generate is not feasible. Therefore, robust algorithms that automatically identify anomalous behavior within the data are required.
When data-driven modeling (machine learning) techniques are used to detect faulty conditions in most cases example data from both normal as anomalous cases are assumed. For example, in [6] the authors propose a feature learning model for condition monitoring based on convolutional neural networks and vibration signals for rotary machinery. However, in practice the type of anomalies that can occur is not clearly defined and when data of anomalies is available it is scarce. Other approaches define a Health Indicator (HI) which uses a model that was estimated only (or mainly) based on data acquired when the asset operated in a normal way. Such HI should indicate the similarity of the current measured state to the baseline/normal operational state. A low HI relates to a poor asset's condition while a high HI points to a healthy condition. In [7] the authors defined a HI based on the Mahalanobis distance of vibration signals to indicate the health condition of a cooling fan and induction motor. Deep statistical feature learning based on Gaussian-Bernoulli deep Boltzmann machine from vibration measurements of rotating machinery was used in [9] as a fault diagnosis technique. In [11] acoustic signals were used to define a HI based on the residual errors of an autoencoder to detect abnormalities in a Surface-Mounted Device. Other examples that use acoustic signals are found in [3,12,4].
The main contributions of this work focus on an empirical assessment of three unsupervised machine learning approaches to generate a HI based on acoustic, vibration and fused signals. Two generative methods, Gaussian Mixture Models (GMM) and Variational Autoencoders (VAE), and a discriminative method, One Class Support Vector Machines (OC-SVM), were considered. Moreover, different types of features were compared. In case of VAE convolutional layers were added to let the model automatically discover higher level features based on the handcrafted lower level features. Using the models built, a root cause analysis that pin points a change in HI to a specific operational state of the asset (where specific parts are being used) was performed. As an input to the models, a toy data set is collected that enables the empirical assessment. To the best of our knowledge no data set is publicly available that recorded both vibration and acoustic data simultaneously from an asset that has rotating elements as well as relatively fast behavioral changes and contains several operational states (or contexts). In this work, a coffee vending machine with rotating elements in the

bean grinder and water pump, relatively fast changing levels in the water and bean containers, and several stages in the coffee making cycle were used to generate the data set.
The remaining part of the paper is organized as follows. Section 2 discusses the collected toy data set, the feature extraction, modeling algorithms and model formulation. Sections 3 discusses the experimental setup. In Section 4 the obtained results are discussed in detail. Finally, Section 5 concludes the paper by summarizing the results obtained from the set of experiments.

2 Toy data set for condition monitoring of assets

In order to study the machine learning methods that generate an asset's HI based on vibration, acoustic or both signals, a toy data set from a coffee vending machine, was recorded in an office environment that has:

a) synchronized recordings of both vibration and acoustic signals to capture the dynamics of rotating elements in the bean grinder and water pump,
b) relatively fast changes in the underlying physics (the change in levels of the bean and water container),
c) several operational states or contexts (different stages in the coffee making cycle).

First, the operational characteristics of the coffee vending machine will be introduced. Then, the sensing mechanisms used to collect a toy data set and the models built to calculate the HI will be briefly reviewed.

2.1 The Coffee Vending Machine

The coffee vending machine used in this research work is a **Bosch-TCA53**. It is a fully automatic espresso machine with desirable characteristics suitable for the planned experiments. The important concepts and components in the condition monitoring of the coffee vending machine are:

1) **Process States:** The coffee preparation process of Bosch-TCA53 vending machine goes through the set of states shown in Fig. 1 from start to finish, when the machine is operating in normal conditions.

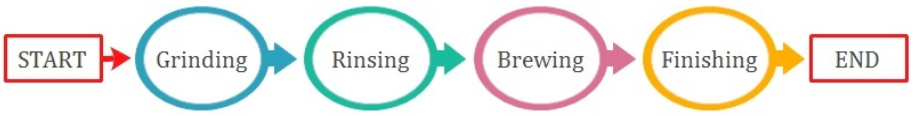

Fig. 1: Coffee preparation states of the vending machine.

In the grinding state, the machine will grind enough coffee beans to prepare a cup of coffee. Then it progresses to the rinsing state which will wash the

brewing tank and make it ready to brew the coffee. After brewing, the final state is finishing where optional operations are performed (like adding sugar and/or cream). Each input data stream is annotated in the pre-processing phase to extract the different states.

2) **Sensors used for data collection:** The two sensors used are:
   (i) *Accelerometer*: A three axis-accelerometer is used to collect the vibration data generated by the vending machine. This sensor is attached to the side of the machine in direct contact with the bean container. The sampling rate of the sensor is 1037Hz.
   (ii) *Microphone*: is used to collect the acoustic data generated by the vending machine. This sensor was positioned next to the machine roughly in 2cm distance without being in direct contact with it. The sampling rate of the sensor is 48kHz.

   Fig. 2 shows the acoustic and vibration signals acquired during a normal coffee preparation process.

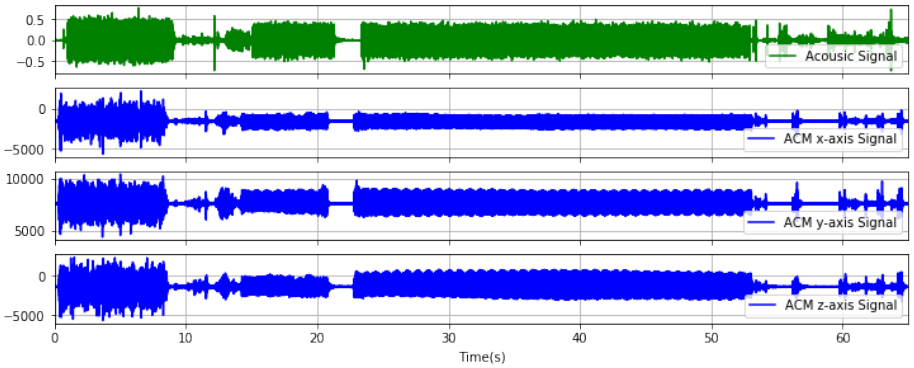

Fig. 2: Acoustic and vibration signals during a normal coffee preparation process.

3) **Monitored Tasks:** The main interest of using the coffee vending machine in an experimental setup originates from detecting subtle changes in signals received from the sensors which correspond to a change in operating behavior. Two tasks were defined based on the sensor data:
   a) to discriminate the situation of above half-full bean and water containers from all other situations where at least one container is below half level;
   b) to output a HI that correlates with the different container levels.

   Lower container levels should correspond to lower HI meaning the condition of the machine is further apart from the normal situation.

4) **Data collection:** To ease the analysis three discrete container levels were defined:
   – **Normal Set**: This data set represents a condition when the bean and water in the containers are above half of the respective tank levels.

– **Low Level Set**: This data set represents a low, below half, bean and/or water container level condition.
– **Empty Set**: This data set represents an empty bean or water container level condition.

Table 1: Collected training and test data set

Cycle Combinations	**# of Cycles**
Normal Bean/Normal Water	9
Low Water /Normal Bean	6
Low Bean/Normal Water	7
Low Bean/Low Water	4
Empty Bean/Normal Water	1
Empty Water/Normal Bean	4

In total 31 full cycle espresso coffee samples were collected for experimental purpose. The toy data collected from the coffee vending machine tried to capture all possible combinations of the specified status conditions as shown in Table 1. For the sake of having a more balanced number of abnormal cycles to be used in the test phase, the bean and water containers were filled, unfilled and refilled with no-specific order in the collection procedure.

2.2 Feature Extraction

When extracting features from sensor data, one should try to capture the relevant information from the input data as much as possible. The features are extracted to capture both the time and spectral domain signal properties of the input data, collected from the sensors. Prior to calculating the features, the sensor signals first pass through a framing operation, which transforms the raw signals into short overlapping segments or frames. In this work the following set of features are assessed:

1. **Time Domain Energy**
   For a sequence of samples in a frame, its energy is calculated as a sum of the squares of the samples in the frame [2].

2. **Spectral Centroid**
   The spectral centroid measures the spectral position of a signal. This measure is obtained by evaluating the "center of gravity" using the Fourier transform's frequency and magnitude information. The individual centroid of a spectral frame is defined as the average frequency weighted by their corresponding amplitudes, divided by the sum of the amplitudes [2].

3. **Linear and mel spectra**
   After transforming the signal from the time to the frequency domain, the resulting spectra generally have too high dimensionality. These spectra are then further compacted by passing them through a number of linear or mel filter banks. Mel spectra are commonly used in acoustic signal processing, as can be seen in [1]. Linear spectra are chosen for the accelerometer signal since less dimensionality reduction is needed, due to a lower signal bandwidth.

4. **Mel-frequency cepstral coefficients**
   Mel-frequency cepstral coefficients (MFCCs) are used in the representation of acoustic and low frequency dominant signals. The MFCC feature extraction technique basically includes windowing the signal, applying the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the inverse DCT as can be seen in detail from [15].

2.3 Modeling Algorithms

Normal profiles are created using the data obtained from baseline operating conditions of the machine, reflecting the fact that no faulty condition occurs in the coffee making cycles. A model is built from the normal profiles using unsupervised algorithms to automatically label significant deviations. The algorithms used are:

1. **Gaussian Mixture Model**
   A Gaussian mixture model (GMM) is a weighted sum of M component Gaussian densities as given by [13],

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{1}$$

   where $\mathbf{x}$ is a D-dimensional data vector (i.e. measurement or features), w_i, $i = 1, \ldots, M$, are the mixture weights, and $g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, \ldots, M$, are the Gaussian densities of each component that have the following D-variate form,

$$g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} exp\Big\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\Big\} \tag{2}$$

   with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma_i}$.
   The HI (h_{gmm}) of an observation is directly tied to it's weighted log probability ($\log(p(\mathbf{x}|\lambda))$).

2. **One-Class Support Vector Machine**
   One-Class Support Vector Machine (OC-SVM)[14] is a variant of SVM method trained on data from only a single class by computing a bounding hypersphere that encompasses as much of the training data as possible. Given

training vectors $\mathbf{x}_i \in R^n$, $i = 1, \ldots, l$ without any class information, OC-SVM is defined as:

Primal problem:

$$\begin{aligned} \underset{w,\xi,\rho}{min} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} - \rho + \frac{1}{\nu l}\sum_{i=1}^{l}\xi_i \\ subject\ to \quad & \mathbf{w}^T\phi(\mathbf{x}_i) \geq \rho - \xi_i, \\ & \xi_i \geq 0,\ i = 1, \ldots, l. \end{aligned} \tag{3}$$

Dual problem:

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{min} \quad & \frac{1}{2}\boldsymbol{\alpha}^T Q \boldsymbol{\alpha} \\ subject\ to \quad & 0 \leq \alpha_i \leq \frac{1}{\nu l}, i = 1, \ldots, l, \\ & \boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1. \end{aligned} \tag{4}$$

where $Q_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ is the kernel function. Then an RBF kernel ($\exp(-\gamma \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2), \gamma > 0$) is used, which is a popular and mostly used kernel in practice. The decision function is defined as:

$$f(\mathbf{x}) = sgn((w^T\phi(\mathbf{x}_i)) - \rho) = sgn(\sum_{i=1}^{l}\alpha_i K(x, x_i) - \rho) \tag{5}$$

The resulting HI (h_{OCSVM}) of an observation is measured by the score of each sample.

3. **Variational Autoencoder**

As described in [10], the goal of a standard autoencoder (AE) is to use an encoding network ($\mathcal{E}$) to create a compact representation $\mathbf{z}$ from an input $\mathbf{x}$ and then use a decoding network ($\mathcal{D}$) to make a reconstruction $\hat{\mathbf{x}}$.

$$\mathbf{z} = \mathcal{E}(\mathbf{x}|\theta_E), \tag{6}$$

$$\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z}|\theta_D). \tag{7}$$

Variational autoencoders are a modification of this AE to a generative model. This is done by replacing the representation $\mathbf{z}$ of an input $\mathbf{x}$ by a posterior distribution $q(\mathbf{z}|\mathbf{x})$, $\mathbf{z}$ will then be sampled from this distribution.

$$q(\mathbf{z}|\mathbf{x}) = \mathcal{E}(\mathbf{x}|\theta_E), \tag{8}$$

This posterior is usually chosen as a Gaussian, where the mean and variance are the output of the encoder. To ensure valid outputs when sampling from the posterior, the Kullback-Leibler (KL) divergence from a prior distribution p(z) is used as regularization. During inference the sampling of $\mathbf{z}$ is fixed to the mean of the posterior distribution q($\mathbf{z}|\mathbf{x}$), to remove randomness in the reconstruction. Using this reconstruction, the error with the original input can be calculated using Equation 9, we will call this error the HI h_{VAE}.

$$h_{VAE} = -\Sigma_{i=1}^{n}(\mathbf{x}_i - \hat{\mathbf{x}}_i)^2 \tag{9}$$

2.4 Sensor Data Fusion

Observational data collected by sensors can be combined, or fused, at a variety of levels for an improved performance of the detection system. This fusing can take place at the raw data (or observation) level, feature level, or at the decision level [5]. Raw sensor data can be directly combined if the sensors are commensurate (i.e. if the sensors are measuring the same physical phenomena). Conversely, if the sensors data are noncommensurate, the data can be fused at a feature or decision level. In this work, since we are using two sensor data streams that are not necessarily commensurate, fusing is performed at the feature level for performance comparison purposes with the individual sensor HI results.

3 Experimental Setup

This section presents and discusses the empirical results obtained by using GMM, OC-SVM and VAE models for HI estimation and related fault detection on the acoustic and vibration data sets. A comprehensive set of experiments are performed to compare and demonstrate the detection accuracy and asset change trends in HI by the models using vibration, acoustic and fused features.

3.1 Experimental Settings

From the acoustic and vibration data collected in different operating conditions, two categories of features are extracted to be used as input to the developed models.

Category 1: uses higher level features of time domain energy, spectral centroid and MFCCs. For this category, frames of size 500ms equating to a window size of 24000 for the acoustic signal and 518 for the accelerometer signal, with an overlap of 50% were used. For the MFCC features, 128 mel bands are used with the first 13 cepstral coefficients being retained.
Category 2: uses linear and mel spectra as an input to the models. The window size is reduced to 100ms and the overlap remained 50%. This reduction was done to provide a more detailed input to the convolutional network, since it will discover its own features. A total of 64 mel and linear bands were used for acoustic and accelerometer signals respectively. In this category the spectrograms are divided into frames of 32 timesteps, which roughly equates to 1.5s of data, and are then used as input for the machine learning model, allowing it to automatically discover features.

The extracted features are standardized to have a zero mean and unit variance so that they will have an equally weighted effect in the modeling phase. For the second category features this is done before the division into frames. Since we have 9 coffee making cycles in the normal data set, a 9-fold split is created with each fold containing 6 cycles in the training, 2 in the validation and 1 in the test set. This approach prevents data from the same cycle to appear in the different sets.

When applying GMM models, the optimal number of components are obtained by calculating the Bayesian Information Criterion (BIC) value which achieved a minimum for 4 number of components. For OC-SVM, the hyper-parameter settings which consistently gave a better performance for the different experiments are a 0.1, 1/30 for the fused signals and 1/15 for the other signals and RBF kernel values for nu, gamma and the kernel type respectively. The evaluation will be done both on a frame level, for each separate input, and on a run level, for a complete coffee making cycle. The run level results are obtained by averaging the frame level results.

3.2 Autoencoder architectures

In this research two kinds of VAE will be used: one with convolutional layers (VAE-Conv) and one without (VAE-FC). Due to these layers, VAE-Conv will be able to automatically discover features from low level features (e.g. mel spectra). The VAE-FC uses only fully connected layers on the first category of features. The model consists of 6 fully connected layers, 3 in the encoder and 3 in the decoder, with 16, 8, 4 neurons and 4, 8, 14 neurons respectively. The neurons are doubled when using the fused signal.
In the architecture that uses the second category of features, convolutional and deconvolutional layers are needed. This VAE model consists of 5 2D convolutional layers and a fully connected layer in the encoder and a fully connected layer and 5 2D deconvolutional layers in the decoder, with 8, 16, 32, 64, 128, 32 and 256, 64, 32, 16, 8, 1 neurons in the encoder and decoder respectively.
In both models, the activation functions are relu in all but the final layers of the encoder and decoder, where linear functions are used instead. L2 regularization and the adam [8] optimizer are used with a factor of $1e^{-4}$ and $1e^{-3}$ respectively.

4 Results and Discussion

In this section the results obtained with the various models will be discussed. Initially, the performance of individual and fused sensor input is evaluated with respect to discriminating normal and abnormal conditions. The performance in terms of correct classifications is given on the frame and complete run level. Then additional analysis is done to further discriminate between the various container levels on a chosen model. Finally, a frame level analysis of the HI outcomes is performed to determine the root cause of the observed trends.

4.1 Model comparison

As explained in Section 3.1 two categories of features are used. The results of the experiments are summarized in Table 2. All results shown are calculated by taking the mean and standard deviation of the AUC score across the 9 folds. Looking at the results from the models, on the frame level a quite good distinction can be made between the normal and abnormal cycles using acoustic

signals, while the accelerometer signals provide a weaker distinction. However, by fusing both signals the majority of the models attained a slight performance increase. On the run level, an even better distinction is achieved compared to the frame level. This difference is likely attributed to the lowered HI in the grinding phase compared to the other phases. This will be discussed in more detail in Section 4.3. In general we see that using only the acoustic signal provides a close performance compared to the fused signal, while the accelerometer signal has a lower performance.

Table 2: Status prediction results where the models in an orange cell use the category 1 features and the model in the green cell uses category 2 features.

		Sensors		
		ACM	**Audio**	**Fused**
GMM	**Frame Level**	0.703 ± 0.082	0.826 ± 0.081	0.854 ± 0.071
	Run Level	0.871 ± 0.043	0.945 ± 0.060	0.945 ± 0.060
OC-SVM	**Frame Level**	0.774 ± 0.079	0.885 ± 0.056	0.874 ± 0.032
	Run Level	0.894 ± 0.141	0.990 ± 0.019	0.994 ± 0.027
VAE-FC	**Frame Level**	0.660 ± 0.061	0.827 ± 0.081	0.855 ± 0.046
	Run Level	0.919 ± 0.093	0.990 ± 0.019	1.000 ± 0.000
VAE-Conv	**Frame Level**	0.704 ± 0.117	0.811 ± 0.123	0.802 ± 0.058
	Run Level	0.904 ± 0.087	0.939 ± 0.171	0.995 ± 0.014

4.2 Level/Trend Analysis

This section investigates whether there is a correlation between the HIs delivered by the considered models and the levels of the bean and water containers. Focusing on the results from OC-SVM and VAE-Conv, with the remaining algorithms showing similar trends, the accelerometer performed poorly in distinguishing abnormal data when compared to the acoustic signals as shown in Table 2. Based on these results, one may be inclined to conclude that using only the acoustic signals might be enough or better to purely determine if the current situation deviates from the normal operation.
However, when considering the HI in relation with the container levels no clear correlation is observed for the acoustic signals, however the accelerometer seems to perform slightly better in this task, as can be seen on Fig. 3 and 4. Another interesting aspect of the result is observed from the fused signal of Fig. 4, revealing a better correlation with the container levels, meaning that the cycles are becoming more anomalous as the container levels decrease.

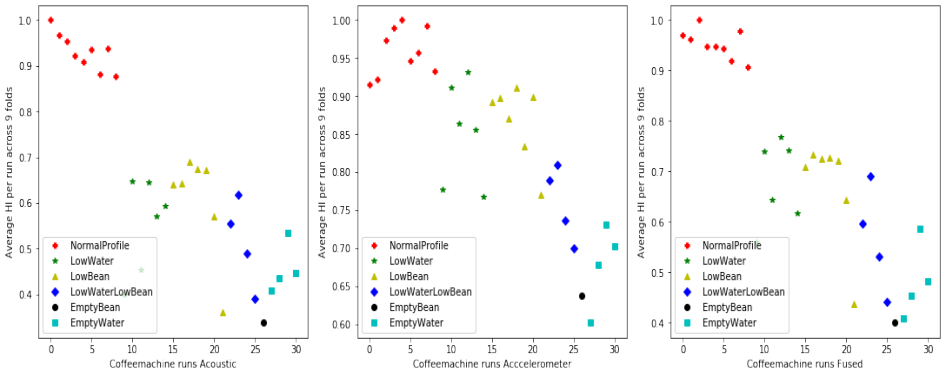

Fig. 3: OC-SVM: average of the run level results for category 1 features across all folds

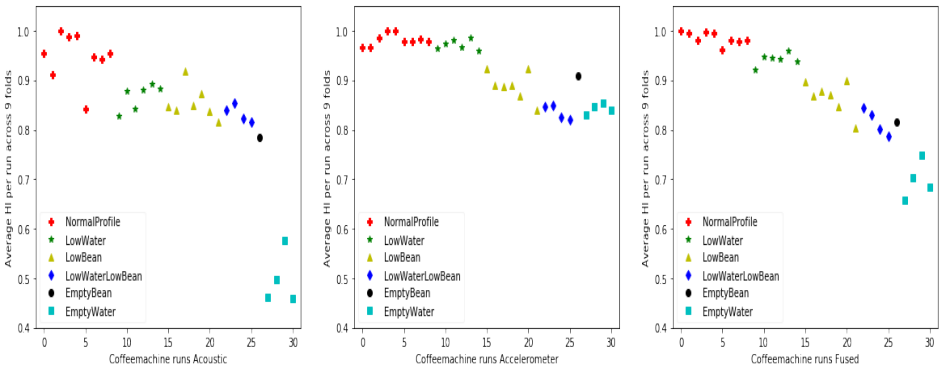

Fig. 4: VAE-Conv: average of the run level results for category 2 features across all folds

A more detailed view of the HI for the different cycles is shown in Fig. 5 and 6. The most noticeable observations are:

- For OC-SVM model with category 1 features, there is an overlap between the three low container level runs of the acoustic signal. When we look at the accelerometer signals, we notice a better discrimination of the various levels compared to the acoustic signals. Looking at the fused signals, they achieve a much better discrimination between almost all the various cycles, with an overlap only in between the LowBean and LowWater scenarios.
- For VAE-Conv model with category 2 features, there is an overlap between four low and empty container level runs of the acoustic signal. This indicates that the acoustic signal is less correlated with the container level. When we look at the accelerometer signals, we notice a clear difference in HIs for the various levels. However, the accelerometer is not able to properly separate

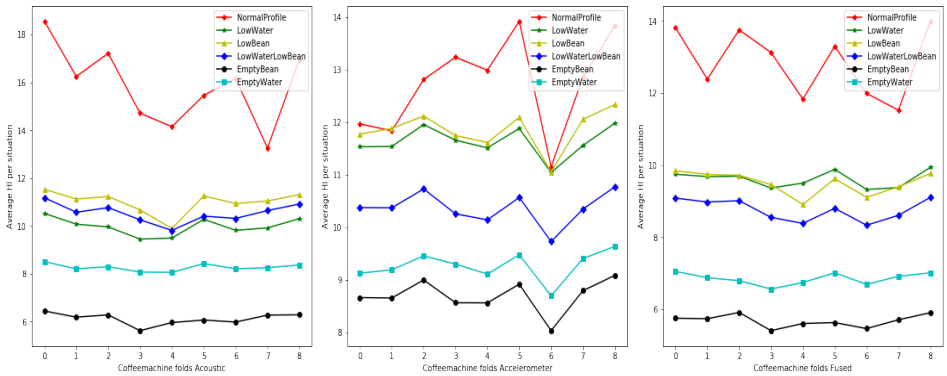

Fig. 5: OC-SVM: average HI of all cycles per scenario for each fold

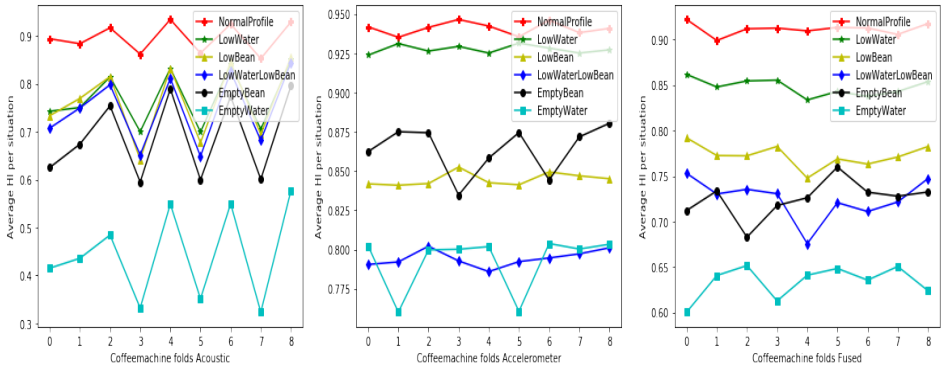

Fig. 6: VAE-Conv: average HI of all cycles per scenario for each fold

whether the resulting conditions are due to the water or bean container levels. Finally, when looking at the fused signals, they achieve a much better discrimination between almost all the various cycles, with only an overlap between the EmptyBean and LowWaterLowBean scenarios. A possible explanation could be found in the in-depth analysis of the HIs in Section 4.3.

In conclusion, the fusion of both signals provides a complementary solution in achieving a high performance in normal/abnormal cycles separation and a clearer discrimination between the abnormal cycles in both models and feature categories.

4.3 Causality

In the previous sections we discussed both a quantitative comparison of the performance of our models based on the AUC score and an analysis of the ability to differentiate the various abnormal cycles on the run level. However, another way

to approach this problem is to examine if the root cause of the abnormal cycle (e.g. a low bean container) can be found based on the shift in HI in the states on the frame level. The VAE-Conv model will be used for this examination.
In Figure 7, a comparison between the different abnormal and normal container situations is shown. The first image shows the average HI of the cycles in the normal situation, and the other images show how the average HI of these situations compares to the normal situation. These HIs are all scaled by dividing them by the maximum of the HI of the normal situation. Firstly, we noticed that the HI

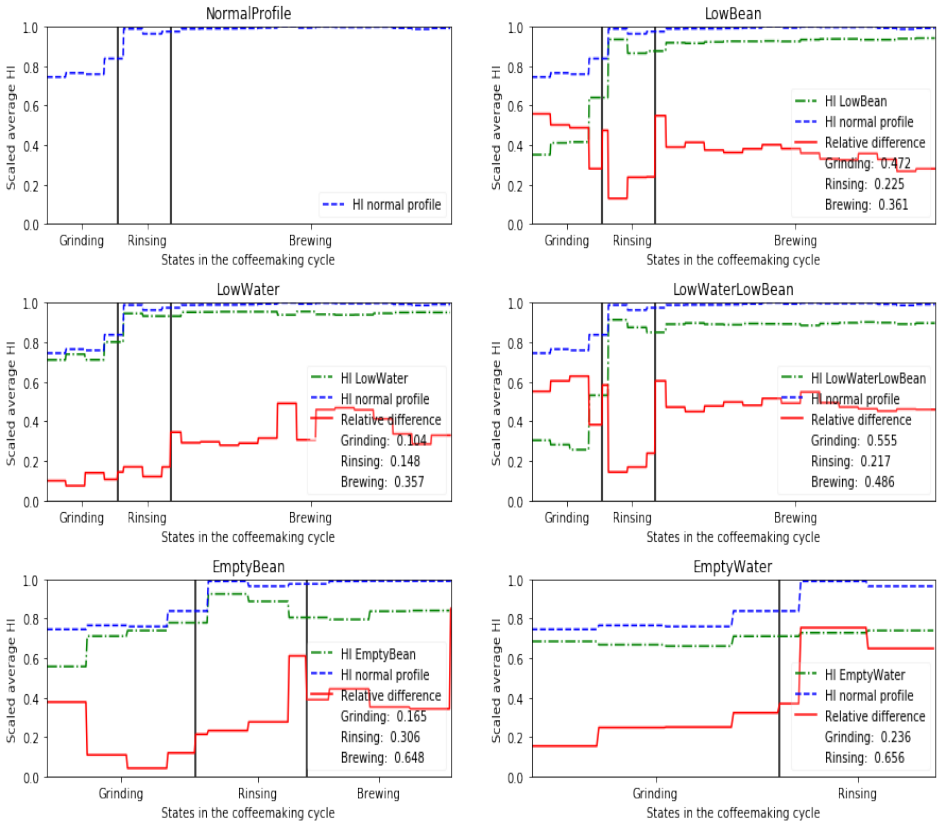

Fig. 7: Average HI across folds per container situation, compared to the the normal situation. The black lines indicate the state changes. The average relative distance in HI between the normal and selected situation is shown in the legend.

for the first state "grinding" is noticeably lower than for the other states. This can likely be attributed to two factors. The first factor could be the imbalance in the data, with the brewing phase being roughly five times as long as the grinding phase. Due to this imbalance, the VAE will train more on the brewing phase and

be able to reconstruct it better compared to the grinding phase, resulting in a difference in HI. A second factor is the amount of energy in the different states, with the grinding state features having more energy compared to the brewing state, which could contribute to the lower HI.
Secondly we examined the different states in these situations to determine if the correct container situation could be selected based on the average HIs in the states. These HIs are once again scaled, so we can calculate a relative instead of absolute difference, to achieve a fairer comparison. For this scaling we attached more importance to deviations on a high HI, so we flipped $(1 + h_{VAE})$ the scores before scaling them. One noticeable thing is that for both situations with empty containers a clear distinction can be made. When looking at the situations with low containers, this distinction is less clear, but still visible. These observations show that the root cause of the abnormal cycle can possibly be found by looking at the frame level results.

5 Conclusion

In this paper, three machine learning methods GMM, OC-SVM and VAE were used to define a HI. The HI generated by each method was analyzed on a newly collected toy data set that includes acoustic and vibration signals from a coffee vending machine. Different features were extracted from the input data enabling in identifying a behavior that deviates from the normal situation. It was observed that in most cases an improved detection performance can be achieved when vibration and acoustic data was fused compared to using a single modality. A sensor can have a robust performance in detecting deviating behavior, while another sensor can give a good discrimination of the different types of abnormal conditions. The experimental results obtained on this toy data set indicate this property. While acoustic signals enable a better discrimination of the normal and abnormal operational conditions, the vibration signals are better in identifying subtle differences in the abnormal signals leading to an enhanced root cause analysis. Furthermore, leveraging on the use of both sensors, a sensor fusion approach consistently outperformed the case when a single sensor was used.

Acknowledgment

We are grateful to Magics Instruments for providing the platform to collect the data used in this experiment and for their thoughtful and detailed feedback which has helped greatly in improving this work. This research received funding from the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” programme.

References

1. Cao, Y., Kong, Q., Iqbal, T., An, F., Wang, W., Plumbley, M.: Polyphonic sound event detection and localization using a two-stage strategy. In: Proceedings of

the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019). pp. 30–34. New York University, NY, USA (October 2019)
2. Giannakopoulos, T., Pikrakis, A.: Introduction to Audio Analysis: A MATLAB Approach (2014). https://doi.org/10.1016/C2012-0-03524-7
3. Gong, C.s.A., Lee, H.c., Chuang, Y.c., Li, T.h., Su, C.h.S., Huang, L.h., Hsu, C.w., Hwang, Y.s., Lee, J.d., Chang, C.h.: Design and Implementation of Acoustic Sensing System for Online Early Fault Detection in Industrial Fans **2018** (2018)
4. Gu, D.S., Choi, B.K.: Machinery Faults Detection Using Acoustic Emission Signal (2011)
5. Hall, D.L.D.L., Member, S., Llinas, J.: An introduction to multisensor data fusion. Proceedings of the IEEE **85**(1), 6–23 (1997). https://doi.org/10.1109/5.554205
6. Janssens, O., Slavkovikj, V., Vervisch, B., Stockman, K., Loccu, M., Verstockt, S., Walle, R.V.D., Hoecke, V.: Convolutional Neural Network Based Fault Detection for Rotating Machinery (2016). https://doi.org/10.1016/j.jsv.2016.05.027
7. Jin, X., Chow, T.W.: Anomaly detection of cooling fan and fault classification of induction motor using Mahalanobis-Taguchi system. Expert Systems with Applications **40**(15), 5787–5795 (2013). https://doi.org/10.1016/j.eswa.2013.04.024
8. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (12 2014)
9. Li, C., Shanchez, R.V., Zurita, G., Cerrada, M., Cabrera, D.: Fault diagnosis for rotating machinery using vibration measurement deep statistical feature learning. Sensors (Switzerland) **16**(6) (2016). https://doi.org/10.3390/s16060895
10. Meire, M., Karsmakers, P.: Comparison of deep autoencoder architectures for real-time acoustic based anomaly detection in assets. In: 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). vol. 2, pp. 786–790 (Sep 2019). https://doi.org/10.1109/IDAACS.2019.8924301
11. Oh, D.Y., Yun, I.D.: Residual Error Based Anomaly Detection Using Auto-Encoder in SMD Machine Sound pp. 1–14 (2018). https://doi.org/10.3390/s18051308
12. Oh, H., Azarian, M., Pecht, M.: Estimation of fan bearing degradation using acoustic emission analysis and Mahalonabis distance. In: Proceedings of the Applied Systems Health Management Conference. pp. 1–12 (2011)
13. Reynolds, D.a.: Gaussian Mixture Models. Encyclopedia of Biometric Recognition **31**(2), 1047–64 (2008). https://doi.org/10.1088/0967-3334/31/7/013
14. Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the Support of a High-Dimensonal Distribution. Microsoft Research, Redmond, WA **TR 87**(November) (1999)
15. Zheng, F., Zhang, G., Song, Z.: Comparison of different implementations of MFCC. Journal of Computer Science and Technology **16**(6), 582–589 (2001). https://doi.org/10.1007/BF02943243