ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-TUPDP138

EXPLORATORY DATA ANALYSIS ON THE RHIC CRYOGENICS SYSTEM COMPRESSOR DATASET*

Y. Gao†, K. A. Brown, R. Michnoff, L. Nguyen, B. van Kuik, A. Zarcone
Brookhaven National Laboratory, Upton, NY, USA
A. Tran, Facility for Rare Isotope Beams, East Lansing, MI, USA

Abstract

The Relativistic Heavy Ion Collider (RHIC) Cryogenic Refrigerator System is the cryogenic heart that allows RHIC superconducting magnets to operate. Parts of the refrigerator are two stages of compression composed of ten first and five second-stage compressors. Compressors are critical for operations. When a compressor faults, it can impact RHIC beam operations if a spare compressor is not brought online as soon as possible. The potential of applying machine learning to detect compressor problems before a fault occurs would greatly enhance Cryo operations, allowing an operator to switch to a spare compressor before a running compressor fails, minimizing impacts on RHIC operations. In this work, various data analysis results on historical compressor data are presented. It demonstrates an autoencoder-based method, which can catch early signs of compressor trips so that advance notices can be sent for the operators to take action.

INTRODUCTION

The Relativistic Heavy Ion Collider (RHIC) Cryogenic Compressor System at Brookhaven National Laboratory (BNL) is comprised of ten first-stage compressors, four second-stage compressors, and a redundant compressor that can function as a first, second, or full-stage compressor, as shown in Fig. 1. Initially, the compressors were controlled through 120VAC relay logic with minimal data available for Operations and only a local enunciator to indicate faults during unscheduled shutdowns of a compressor. Since 2014, the compressor controls have been upgraded to a more modern 24VDC PLC-controlled system. To date, six first-stage and all second-stage compressors have been upgraded. A part of the modernization is the increased availability of data for operators to monitor and track the health of each running compressor. The total data acquired is 163 variables for a first-stage compressor and 100 variables for a second-stage compressor, the result of one less motor-compressor set. All data are logged at a one-point-per-second rate. The data focus of this study is on a first-stage compressor, which comprises 27 analog variables, i.e., 19 temperature sensors (names starting with "TT"), 5 pressure transducers (names starting with "PT"), 2 horsepower monitors (M77, M79), and an oil level probe. The oil level probe parameter is omitted in this study since it is not as informative as the other parameters.

* Work supported by Brookhaven Science Associates, LLC under Contract No. DE-SC0012704 with the U.S. Department of Energy.
† ygao@bnl.gov

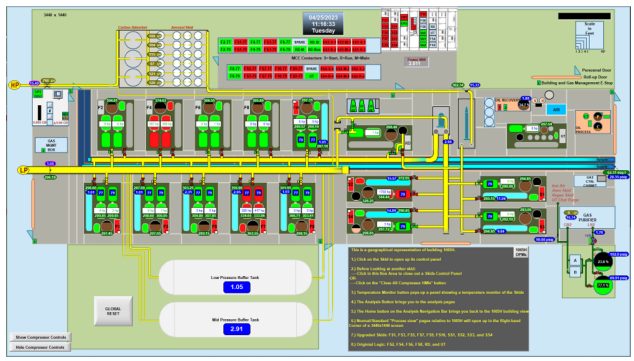

Figure 1: RHIC cryogenic compressor system overview. It comprises ten first-stage compressors, four second-stage compressors, and a redundant compressor.

The 16 to 27 variables per compressor are just a small fraction of the 10,000+ data points an operator must monitor to understand the health of the Cryogenic system. Manually monitoring the system takes valuable time from operators, and it takes much more resources to recover from a system failure than to detect and prevent it beforehand. In this work, we present the initial results of analyzing historical compressor data to determine if developing faults with a compressor can be detected early enough to minimize the impact on operations and to narrow the cause of faults to facilitate quicker repairs, increasing the run-time availability of each compressor.

DATASET AND METHODS

The datasets are collected from the upgraded first and second-stage compressors. Compressor First Stage 1 (FS1) is chosen for analysis because it is the only upgraded compressor with a documented fault during the 493 days of recorded data. The documented trip happened on Apr. 7th, 2022. So the training data were selected from Jan. 15th to Mar. 5th, 2022, and testing data were from Mar. 6th to Apr. 5th, 2022, to test if the algorithm can detect any early precursors. The data were acquired at 1 Hz.

In this work, we focus on analyzing 26 float-type variables, as discussed above. Those time series data are shown in Fig. 2. The "TT"sensors monitor different system parts' temperatures, "PT"sensors monitor different parts' pressures, "M77"and "M79"are the horsepower monitors for the two motors. A pictogram of the FS1 compressor with corresponding parameters is shown in Fig. 3.

In the first step, we applied time series K-means to cluster the datasets to better understand the data patterns. Next, we

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-TUPDP138

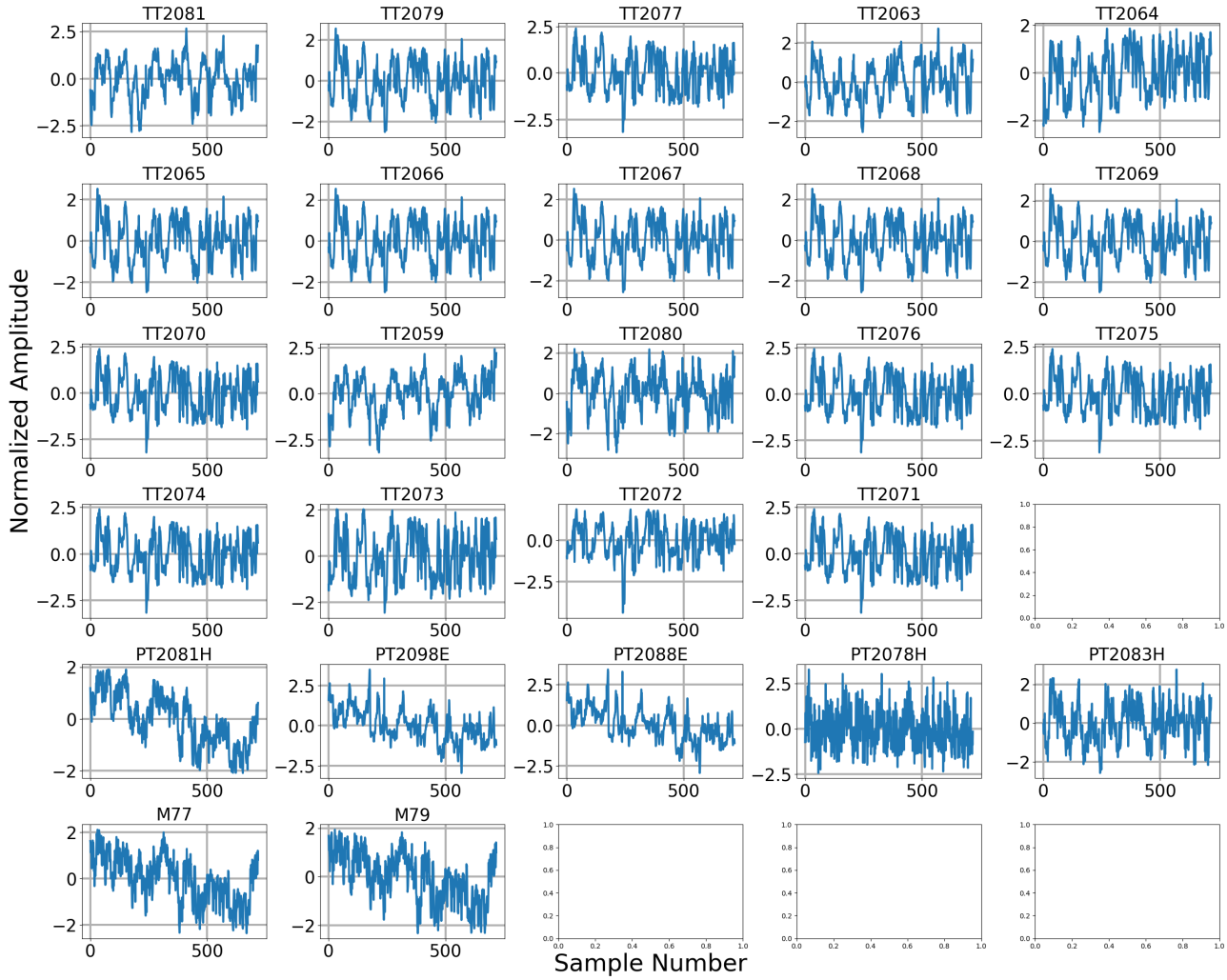

Figure 2: An overview of the 26 time series.

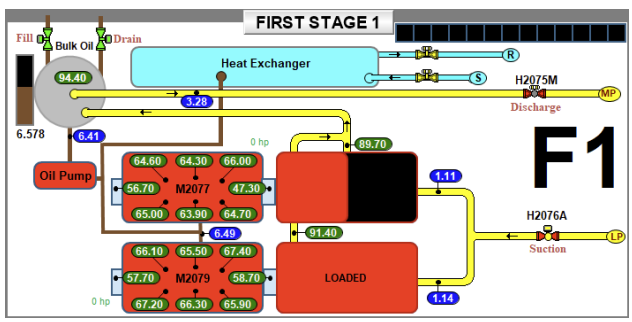

Figure 3: Focus of this study, a pictogram of the compressor First Stage 1 (FS1) with the associate variables.

explore the dataset by applying an LSTM-based autoencoder to detect any anomaly precursors. Moreover, the latent space from the autoencoder is analyzed to gain a deeper understanding of how latent space captures the high-level features of the data.

Time Series K-means Clustering

The k-means method is a type of unsupervised learning algorithm that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid). The problem is computationally difficult (NP-hard). The worst case complexity [1] is given by $\mathcal{O}(n^{(k+2/d)})$, where n is the number of samples and d is the number of features. However, efficient heuristic algorithms can converge quickly to a local optimum, e.g., using either Lloyd's or Elkan's algorithm [2, 3] can solve K-means in $\mathcal{O}(nkdi)$ time, where i is the number of iterations needed until convergence.

As a clustering task aims to group similar objects together, selecting an appropriate distance function to measure the similarity is critical to the algorithm's performance. Traditionally, Euclidean distance is used as the base metric. However, one issue with this metric is that it is not invariant to time shifts which is common in time series data. Thus, a distance metric that is dedicated to time series, Dynamic Time Warping (DTW) [4], is used in our time series clustering task.

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-TUPDP138

The original data is down-sampled to reduce the computation efforts and then standardized to have a 0 mean and unit variance. There are many methods that can be used to find a reasonable number of clusters, e.g., a traditional Elbow test [5]. A good rule of thumb, however, is picking the number of clusters as the square root of the number of features in the training data. In our case, we have $\lceil\sqrt{26}\rceil = 6$ clusters.

The clustering results are shown in Fig. 4. For each cluster, every series is plotted (in gray), and in order to see the main shape of the cluster, the average series is also plotted (in red). The corresponding cluster mapping is shown in Table 1. It gives us an initial idea about the data patterns. Cluster 0 groups all the variables for motor M2079 (bearing/winding temperatures and suction pressure). Correspondingly, Cluster 3 groups all the bearing/winding temperature variables for motor M2077, but its suction pressure data falls into a different Cluster 1, which is distinct from any other variables[1]. Cluster 2 is related to the motor horsepower. Cluster 4 describes the oil tank pressures and Cluster 5 is for the compressor discharge temperatures.

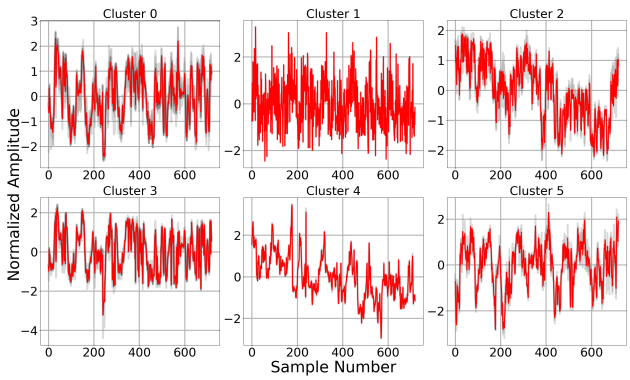

Figure 4: Time series K-means results. It groups the 26 time series into 6 clusters.

Table 1: Cluster Mappings

Cluster	Members
Cluster 0	PT2083H, TT2063, TT2064, TT2065, TT2066, TT2067, TT2068, TT2069, TT2079
Cluster 1	PT2078H
Cluster 2	M77, M79, PT2081H
Cluster 3	TT2070, TT2071, TT2072, TT2073, TT2074, TT2075, TT2076, TT2077
Cluster 4	PT2088E, PT2098E
Cluster 5	TT2059, TT2080, TT2081

LSTM-based Autoencoder Analysis

An autoencoder is a type of neural network with a symmetric architecture. It is composed of an encoder and a decoder, as shown in Fig. 5. The encoder takes in the input data and generates a latent space, usually with a smaller dimension than the input dimension. At the same time, the decoder attempts to reconstruct the inputs from the latent space outputs. Anomalies can be detected if the reconstruction error exceeds a predefined threshold (usually set by the maximum training reconstruction error).

Moreover, if an autoencoder can reconstruct the input using a small latent dimension, this would imply that the dimensionality of the input could be reduced. Studies in the latent space is a popular research area. It helps us gain a deeper understanding of the data. For cases with multiple variables, latent space analysis also gives us insights about which variables are the most important and if we need them all.

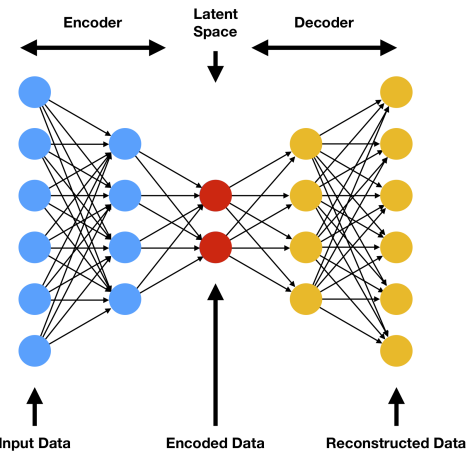

Figure 5: A typical structure of an autoencoder.

Long Short-Term Memory (LSTM) is an architecture used for sequential data. It has been used to predict future steps, such as speech recognition and language modeling [6, 7]. This preceded the recurrent neural network (RNN) since this addresses the vanishing gradient problem [8] of the RNN. LSTM also retains long-term memory, while RNN only has short-term memory, meaning the LSTM should be more robust. In this work, we apply an LSTM-based autoencoder for a more accurate result.

The network has a symmetric structure that steps down from 16 nodes to 2 nodes and a variable encoded length depending on the input dimension. Dropout layers are used to create a denoising effect so the network is more robust to noise. The time sequence for each variable is created using 30 time steps.

RESULTS

The LSTM autoencoder was trained on data from Jan. 15th to Mar. 5th, 2022, and tested on data from Mar. 6th to Apr. 5th, 2022. The documented trip happened on Apr. 7th, 2022, which is due to the discharge temperature sensor TT2059 interlocking the FS1 compressor after it breached a high limit of 125 degrees C for 3 seconds. Technicians

[1] This could be a sign for the operators to check if this sensor works correctly as expected.

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-TUPDP138

found a loose crimp on the sensor, and the compressor was returned to service after repairs.

Anomaly Detection

The reconstruction Mean Absolute Error (MAE) distributions from both training and testing data are shown in Fig. 6. From the plot we can see that the testing data contain parts that have higher reconstruction errors, which indicates the presence of anomalous data patterns.

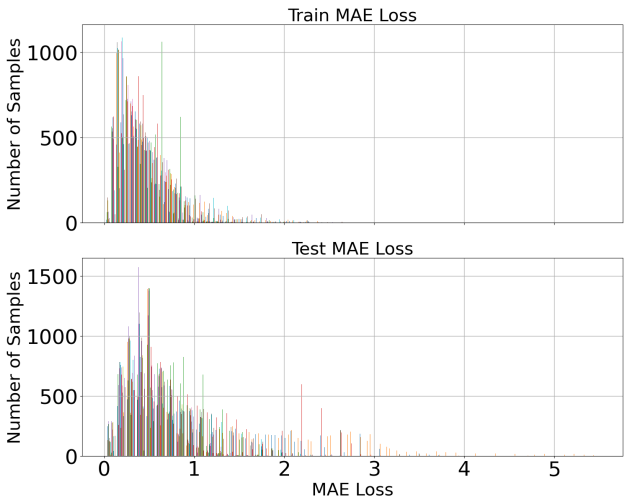

Figure 6: Reconstruction error distributions for both training and testing data.

To further investigate the reconstruction error distribution, the maximum and difference MAE between the train and test data are plotted for each variable in Fig. 7. We can see that the actual cause of the machine failure, the sensor TT2059, gets the highest errors, which may indicate the malfunctioning of that system part already started before the machine trips.

Figure 8 shows the anomaly detection results on the test data of TT2059. The red points are the predicted anomalies. Since the test data are taken before the actual machine trip, we can see that the LSTM autoencoder is able to detect precursors of machine anomalies.

Latent Space Analysis

Autoencoder latent space can capture high-level information about the data. It helps us to gain a deeper understanding of the data patterns [9]. Figure 9 shows the latent space visualization for each variable. The LSTM autoencoder reduces the input dimension to two[2].

The visualization was created by first plotting the training data latent space which serves as the base distribution for all variables, then imposing on it the testing data latent space distribution for each variable (noted by colors). The plot will tell us how the pattern of each variable differs in testing data from a highly abstracted space. We can see that:

[2] The fact that the two latent dimensions almost form a line indicates that one latent variable may be enough to abstract all input information.

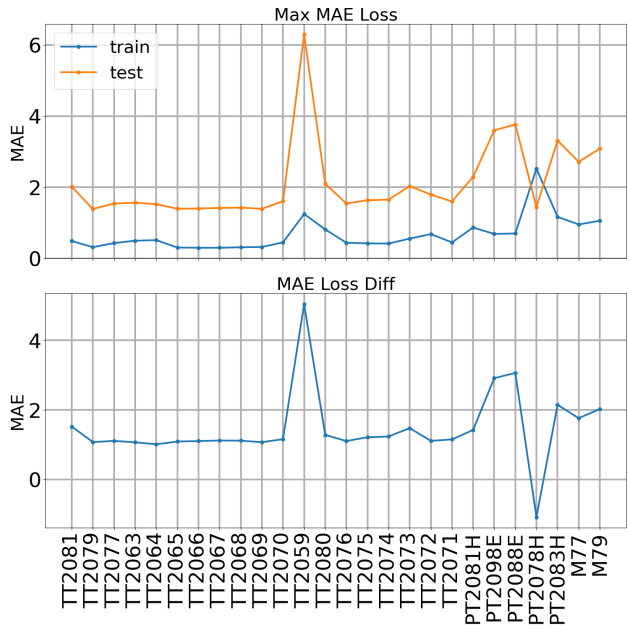

Figure 7: Maximum and difference MAE between train and test data are plotted for each variable. The actual cause of the machine failure, sensor TT2059, gets the highest error value, which could be a warning signal for the operators to check before the machine trips.

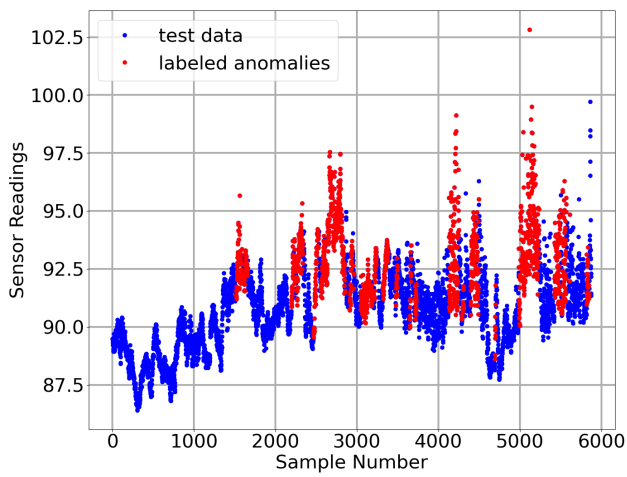

Figure 8: Anomaly detection on the sensor TT2059. We can see that the LSTM autoencoder is able to find precursors and predict anomalies ahead of the machine trip.

- TT2059 has a different data pattern than other "TT" temperature sensors.
- PT2078H and PT2083H do not present obvious data patterns and can be omitted for analysis.

CONCLUSION

In this work, historical compressor data from the RHIC cryogenic refrigerator system are analyzed. Initial data analysis results are presented. It is also demonstrated that machine learning techniques such as LSTM autoencoder can help op-

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-TUPDP138

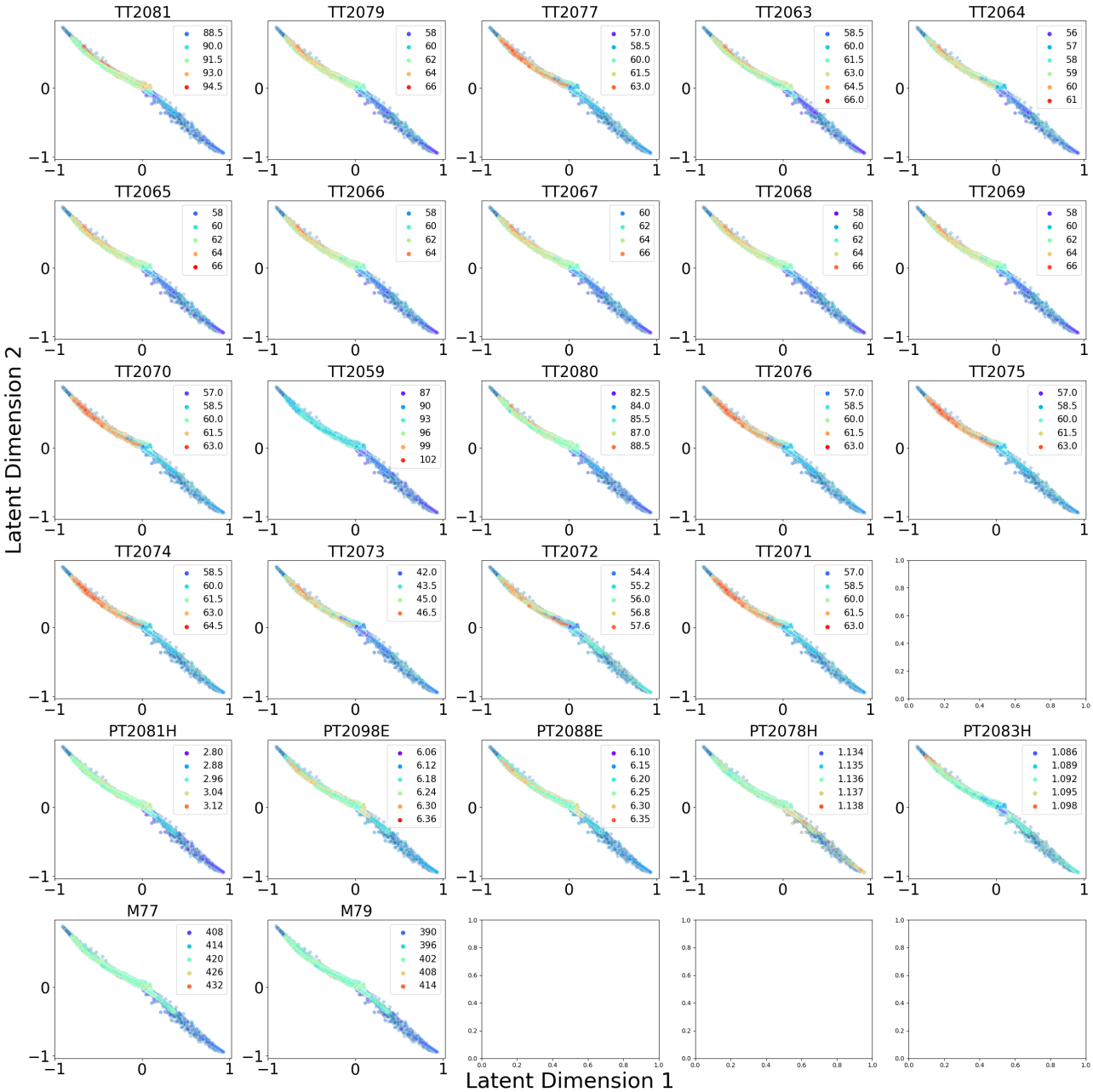

Figure 9: Latent space visualization for each variable. We can see that the sensor TT2059 has a different pattern compared with other temperature sensors. Sensors PT2078H and PT2083H do not have obvious data patterns and can be ignored for analysis.

erators spot machine failure precursors so that proactive actions can be taken to prevent failure and save valuable machine time.

ACKNOWLEDGEMENTS

The first author would like to thank Wan, Jinyu for helpful advice on tuning the LSTM autoencoder.

REFERENCES

[1] D. Arthur and S. Vassilvitskii, "How Slow is the K-Means Method?", in *Proc. SGC'06*, Sedona, AR, USA, Jun. 2006, pp. 144–153. doi:10.1145/1137856.1137880

[2] S. Lloyd, "Least squares quantization in PCM", *IEEE Trans. Inf. Theory*, vol. 82, no. 2, pp. 129–137, 1982. doi:10.1109/TIT.1982.1056489

[3] E. W. Forgy, "Cluster analysis of multivariate data : efficiency versus interpretability of classifications", *Biometrics*, vol. 21, pp. 768–769, 1965.

[4] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans. Acoust. Speech Signal Process.*, vol. 26, no. 1, pp. 43–49, 1978. doi:10.1109/TASSP.1978.1163055

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-TUPDP138

[5] R. L. Thorndike, "Who belongs in the family?", *Psychometrika*, vol. 18, pp. 267–276, Dec. 1953. doi:10.1007/BF02289263

[6] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling", in *Proc. INTERSPEECH'14*, Singapore, Sep. 2014, pp. 338–342. doi:10.21437/Interspeech.2014-80

[7] X. Li and X. Wu, "Constructing Long Short-Term Memory based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition", *arXiv*, Aug. 2018. doi:10.48550/arXiv.1410.4281

[8] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. doi:10.1162/neco.1997.9.8.1735

[9] Y. Liu, E. Jun, Q. Li, and J. Heer, "Latent Space Cartography: Visual Analysis of Vector Space Embeddings", *Comput. Graphics Forum*, vol. 38, Jul. 2019. doi:10.1111/cgf.13672