Explainable multimodal learning for predictive maintenance of steam generators

Duc An Nguyen[1], Sagar Jose[1], Khanh Nguyen[1], and Kamal Medjaher[1]

[1] *Laboratoire Génie de Production, Ecole Nationale d'Ingénieurs de Tarbes (ENIT), Tarbes, 65000, France*
duc_an.nguyen@enit.fr
sagar.jose@enit.fr
thi-phuong-khanh.nguyen@enit.fr
kamal.medjaher@enit.fr

ABSTRACT

Prognostics and Health Management (PHM) is identified as an important lever for enhancing the development of predictive maintenance to ensure the reliability, availability, and safety of industrial systems. However, the efficiency of data-driven PHM approaches is dependent on the quality and quantity of data. Therefore, exploiting multiple data sources can provide additional, useful information than single-modal data. For instance, by incorporating multiple data sources, including condition monitoring data, images from cameras, and texts from maintenance technicians' reports, multi-modal learning can provide a more comprehensive and accurate understanding of the system's health. However, multi-modal deep learning is complex to understand. To address this complexity, it is crucial to incorporate explainable artificial intelligent techniques to provide clear and interpretable insights into how the model makes decisions. In this light, this paper proposes the application of the model-agnostic-explanation approach, i.e., SHAP, to explain the working mechanism of multimodal learning for the prediction of industrial steam generator degradation. Particularly, we determine the important features of each data modality and investigate how multimodal learning can overcome the issues of low-quality data from a single modality due to the additional information from other data modalities.

Keywords: Explainable AI, SHAP, multimodal learning, predictive maintenance, degradation prediction, steam generators.

1. INTRODUCTION

Rapid advancements in data-driven PHM have been propelled by the emergence of machine learning and deep learning techniques (Nguyen, Medjaher, & Tran, 2023). Nevertheless, these techniques require large amounts of data, and the quality of their outcomes depends on both the quality and quantity of training data. Although the majority of industrial applications focus on unimodal sensor-based monitoring data, alternative sources of information, such as images from cameras and textual reports from inspections, also exist. Harnessing these data sources to develop a data-driven prognostic solution presents considerable challenges due to the greater complexity involved in processing multimodal data. Despite this, the potential benefits of learning from multimodal data make it a research direction worth exploring (Jabeen et al., 2023).
The efficacy of multimodal learning can be attributed to the intricate interplay between diverse data modalities in the course of training a prognostic function. However, interpreting and explaining these interactions is difficult due to the diverse nature of the data and the distinct processing methods required (Joshi, Walambe, & Kotecha, 2021). In industries, it is vital that stakeholders can understand how a prognostic model generates its predictions for it to be considered reliable and trustworthy as a maintenance decision support tool. To achieve this, it is essential to integrate explainable artificial intelligence (XAI) techniques into the development and deployment of prognostic models. XAI enables the interpretation and communication of the model's underlying reasoning, thereby enhancing stakeholders' trust in the model and its predictions.
The use of XAI in prognostics has emerged as a noteworthy subject in recent years. For instance, in (Amin, Brown, Stephen, & McArthur, 2022), the authors used SHapley Additive exPlanations (SHAP) to explain prognostic models built from nuclear power station data. Besides, (Nor, Pedapati, Muhammad, & Leiva, 2022), authors built Bayesian deep learning models and apply SHAP to determine the contribution of data from sensors. For bearing fault prognosis, (Sanakkayala et al., 2022) employed a Convolutional Neural Network (CNN) model, specifically VGG16, in conjunc-

First Author et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

tion with Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016) to detect faults through spectrogram images that were transformed from vibrational signals, and identified the parts of the image used by the CNN algorithm. McKinley (Mckinley, Somwanshi, Bhave, & Verma, 2020) used XAI to explain an XGBoost model that predicts the failure of transit bus Nitrogen Oxides sensors. In the context of Chiller Fault-Detection Systems, XAI was applied by using the local explainability of Lime to a machine learning model (Srinivasan et al., 2021).

Despite the considerable interest in the application of XAI to the field of prognostics, the extant literature has neglected the importance of explainable multimodal learning techniques. To address this research gap, the present study seeks to provide a comprehensive account of the working mechanism of multimodal learning for predicting the degradation of industrial steam generators. Specifically, this study aims to answer three following research questions: (RQ1) Which features are important for multimodal learning? (RQ2) How does each feature in each modality contribute to the prediction results? and (RQ3) How does multimodal learning improve wrong predictions made by unimodal learning?

To do this, the paper is organized as follows: Section 2 will introduce the concepts and overview of multimodal learning and XAI, and the investigation methodology. Our approach will also be presented in this section. Section 3 is dedicated to presenting the case study and discussing the obtained results. Finally, in Section 4, the paper will conclude with a summary of the key findings and a discussion on the study's perspective.

2. Explainable multimodal learning

2.1. Multimodal learning

Multimodal learning is a dynamic approach that integrates multiple kinds of data to enhance the learning experience and improve information retention. In this study, a deep learning model is examined that processes three data modalities: image, text, and numerical data (refer to Figure 1). This model employs distinct learning branches for each data modality. Image data is processed using a convolutional neural network (CNN) (O'Shea & Nash, 2015), while text data is first transformed into an embedding (Jiao & Zhang, 2021) and subsequently passed through a CNN. Numerical data, on the other hand, is processed via a fully connected network. Intermediate attention layers are incorporated between these branches to facilitate cross-modal communication at an intermediate level of abstraction. These attention layers adhere to the query, key, and value implementation of the attention mechanism as described in (Vaswani et al., 2017). In the proposed model, text attends to image, image attends to numerical data, and numerical data attends to text. Given that this study primarily focuses on explaining the multimodal model and elucidating the crossmodal interactions during learning, an in-depth discussion of the architecture's implementation details is beyond the scope of this research.

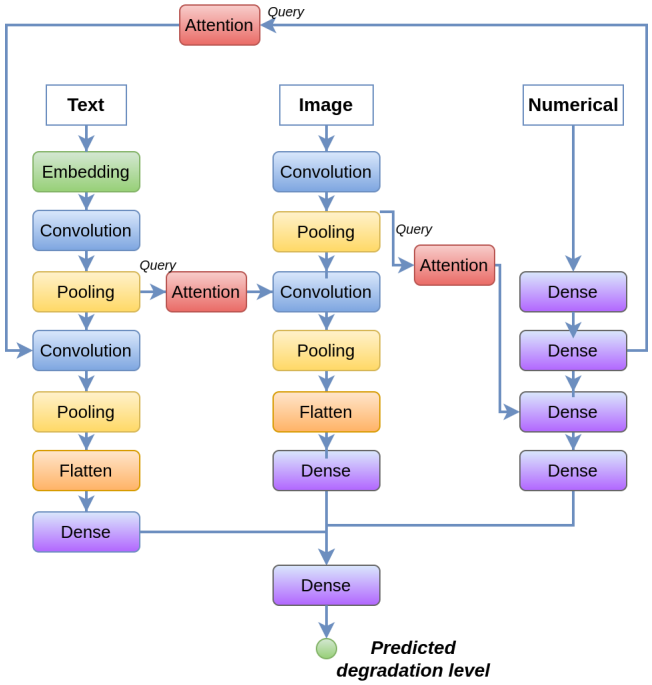


Figure 1. Structure of the multimodal deep learning model investigated in this work.

2.2. Explainable artificial intelligent techniques

In the literature, according to the model's dependency, XAI can be classified into two groups: model-agnostic and model-specific methods (Joshi et al., 2021). The first group, which is independent and irrespective of the explained model, has general modularity in design and can be applied to diverse kinds of models. Meanwhile, model-specific explanations only apply to a specific model.

In this study, we consider the multimodal model (Figure 1) which comprises diverse deep learning (DL) modules with intricate internal interactions. As a result, an explainable method with model-agnostic attributes is necessary to enable the explanation of any DL model. Moreover, to address the three research questions posited in the introduction, the proposed explainable method must be capable of expounding feature effects within a model while also performing both local and global explanations. Given these prerequisites, the Shapley Additive Explanations (SHAP) approach has been selected in this paper to explicate the investigated multimodal learning.

SHAP is a game-theoretic approach for interpreting machine learning models that assigns a value to each feature based on its contribution to the final output. By calculating Shapley values for each feature and combining them, SHAP provides

a detailed explanation of how a model arrives at its decisions. It works by comparing predictions with and without each feature and weighting the differences by the proportion of feature combinations that include that particular feature. For more details of SHAP, one can consult the paper (Lundberg & Lee, 2017)

2.3. Investigation methodology

Figure 2 presents an overview of the proposed method to address the research questions presented in the introduction. As SHAP is agnostic to the internal structure of the model but needs a prediction model to calculate feature effects. In this paper, a pre-trained multimodal model, three data modalities, and background factors are used as the input of SHAP for explainability.

Definition of multimodal background: Applying SHAP to a multimodal model is challenging due to inconsistent data dimensions. Therefore, the background for every modality needs to be converted into the same dimension. Particularly, the text, numeric, and image backgrounds taken from the training dataset are converted into a 1D vector with n_t, n_n, and n_i elements, respectively.

Image background: Instead of using full-size RGB images for the background, the proposed method segments the image into n_i areas using Simple Linear Iterative Clustering algorithm. This algorithm allows clustering pixels based on their color similarity and proximity in the image plane (Achanta et al., 2010). From the n_i areas segmented in the image, a vector with the size of $1 \times n_i$ values 0, is used as a background for SHAP, in which each element is representative of an area. This enhances SHAP's computational efficiency and interpretability since an area has more significant effects on the prediction results than a pixel.

Text background: To ensure consistency in the dimensions of the SHAP background, we employed a bootstrapping procedure (Efron, 1992) to resample the training text data into a single sample with a size of $1 \times n_t$, where n_t denotes the length of each sentence in the training set.

Numeric background: The average value of each feature in the training dataset is calculated to create a background vector with a size of $1 \times n_n$. This method enhances the meaning of the dataset by effectively capturing its overall characteristics through the average values.

Three backgrounds of text, numeric, and image data are then concatenated to create a vector of length of $1 \times (n_i + n_t + n_n)$ that is used as the multimodal background for this study.

Formation of multimodal data that need to be investigated: Similar to the backgrounds, the image, text, and numeric data are first converted to the same dimensions prior to investigation. The image is segmented into distinct parts using a segmentation algorithm, and each part is represented as a feature of SHAP input in the form of a vector with a total size of $1 \times n_i$. It is noteworthy that if an element in this vector is set to 1, its corresponding region in the image is not masked. The text and numeric data remain unchanged with dimensions of n_t and n_n, respectively. Subsequently, all the input data, including the image, text, and numeric data, are concatenated to create a vector with a size of $1 \times (n_i + n_t + n_n)$, which is analogous in structure to the SHAP's multimodal background.

Multimodal model that needs to be explaines: This block serves two primary purposes. Firstly, it ensures that the SHAP input is appropriately formatted for use in the prediction model. For image data, a mask function denoted as $h(x, origImg)$ is applied to the SHAP features to generate a meaningful image $maskImg = h(x, origImg)$. Here, x represents the Shap input, $origImg$ is the original image that needs to be explained, and $maskImg$ is the resulting image after replacing values. If the area-representative-element in the input image vector is 0, then all the values in that area are replaced with 255; otherwise, no replacement occurs. In contrast, text and numeric data do not require a masking function. Secondly, after processing the input data, the multimodal block is used to generate a prediction value, which is then explored by SHAP to calculate the SHAP values.

Resolving research questions: The outputs of SHAP serve to address the three research questions posed in the introduction. To answer for RQ1, the mean absolute SHAP (MAS) value is calculated for each feature across the entire training dataset. Notably, in the case of text data, where features change in each sentence, the MAS is calculated for each "word" appearing in the training data. From that, a bar plot is used to visualize the importance of each feature from three data modalities. For RQ2, the training dataset is divided into different groups according to the system's degradation level. The MAS values of each feature for each group are then calculated separately for both the multimodal and unimodal models. Afterward, a comparison between the unimodal and multimodal models is conducted within each group to determine how each feature contributes to the prediction results in each type of model. With the final question RQ3, the initial step involves the identification of specific instances exhibiting significant Mean Absolute Error (MAE) within the context of unimodal learning. Then, the SHAP's waterfall plot is used for investigating the origins of the aforementioned error, while also determining which features serve as potential drivers of improved predictive outcomes in the context of multimodal learning.

3. CASE STUDY

3.1. Description of case study

The case study utilized in this paper pertains to the degradation level prediction of steam generators (SG), which is achieved through the implementation of a multimodal model. The model categorizes the degradation levels into three groups: good (0,40], medium (40,80], and bad (80,+∞). It is designed to process input data in the form of images, tex-

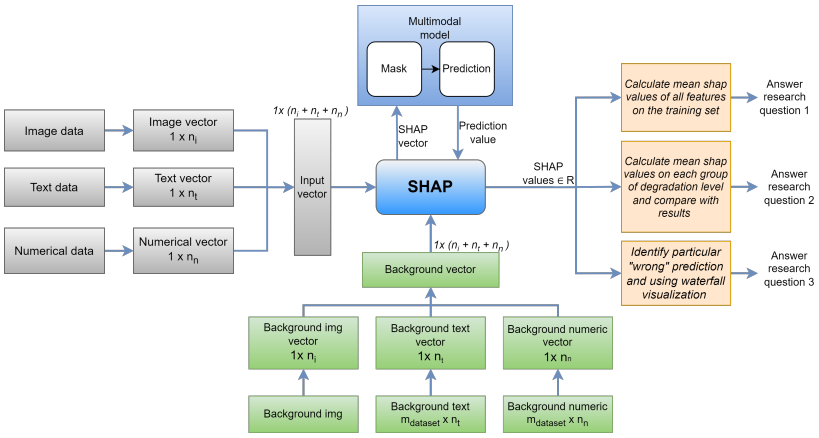


Figure 2. Overview of the investigation methodology.

tual information, and numerical data. Specifically, the images are constructed based on the Wide Range Level (WRL) signal (Girard, 2014), which is a measure of the pressure difference between the top and bottom of the SG downcomer. Text data is written by technicians after each inspection, including qualitative descriptions of the machine's deterioration before maintenance, the maintenance performed, and the machine's condition after maintenance. Numerical data consists of maintenance information including the "number of chemical cleaning", "number of mechanical cleaning", "time since the last repair", "next inspection time", and working "time" of SG. The details about the case study can be found in (Yang, Baraldi, & Zio, 2021), in which, an example of each type of data is visualized in Figure 3.

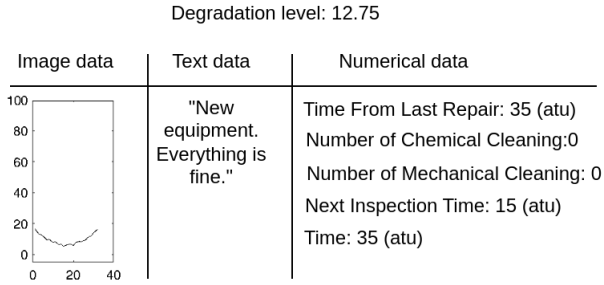


Figure 3. An example of the types of data used in this paper. ("atu": arbitrary time units)

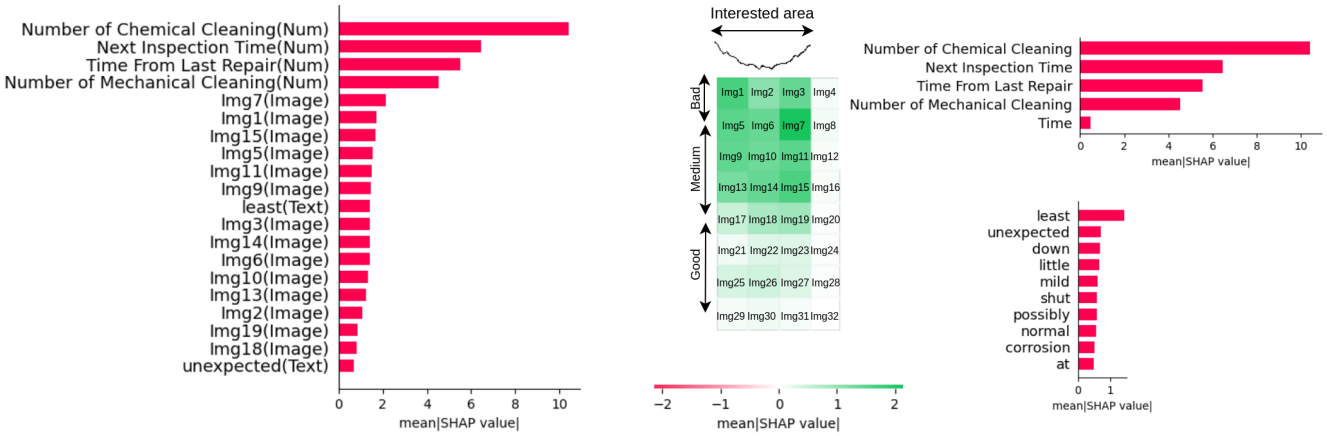
3.2. Results and discussions

This section aims to represent the achieved results by applying SHAP for both multimodal and unimodal models. The representation will include 3 parts corresponding to research questions (RQ1, RQ2, and RQ3) that were mentioned in the introduction of the research. To address RQ1 and RQ2, we analyzed the training dataset to determine the contribution of features to model development. RQ3 was addressed using the test dataset to investigate erroneous predictions in the unimodal model and to identify how the multimodal model improves prediction accuracy.

Answer the RQ1: Figure 4 presents the results obtained after evaluating MAS values for each feature on the entire dataset. Specifically, Figure 4a depicts the average contribution of the 20 most important features from all three modalities to the predictions. Furthermore, Figure 4b presents specific visualizations for the important order of features in each modality. The image data visualization represents 32 distinct features corresponding to 32 areas of the image, as determined by the segmentation algorithm. Meanwhile, the text data visualization displays the top 10 features with the highest impact on the results. And the numeric modality is shown with all its features mentioned in the previous section.

Considering Figure 4a, one can see that the numeric modality demonstrates the highest importance, with significantly higher SHAP values compared to other modalities. After the numeric modality, the image data also contributes to the predictions with values ranging from 2 to 3. However, the text data appears to have a relatively minor impact on the predictions, with all values below 1.5.

Regarding the important order of features in each modality, see Figure 4b, one can see that in the case of image data, SHAP values concentrate on the top regions of the image with darker green colors, but are less pronounced towards the bottom. This phenomenon may be attributed to a dataset imbalance, where the medium and bad groups have nearly twice as many instances as the good group, as indicated in Table 1. Furthermore, in the images, as the degradation curve is not present in these areas on the right side, multimodal learning does not allocate its attention toward them. For numeric data, only the "time" feature, indicating the current working time, holds limited meaning in degradation. The remaining 4 features are important as they reflect the degradation of the steam generator. The text modality, when compared with other types of data, has a small impact on the results. De-



(a) Order of the 20 most important features from 3 modalities

(b) Importance order of features in each modality (Num, Image, Text)

Figure 4. Importance order of features for multimodal learning

spite the "least" feature being deemed the most important but lacks meaningful information to reflect degradation levels accurately.

Table 1. Number of samples from each group

Group	Number of samples
Bad	398
Medium	542
Good	284

Answer the RQ2: This study conducts a comparative analysis of the feature importance in the construction of unimodal and multimodal models for each SG's degradation group. Notably, each type of data is used for the corresponding unimodal learning while these 3 data modalities are the 3 inputs of the same multimodal model. The results of this analysis are presented in Figure 5.

One can see that for a given degradation group and data modality, the SHAP values of features in the unimodal model are higher than those in the multimodal model. This is due to the distribution of contribution across features from other modalities in the multimodal model, resulting in a comparatively lower amplitude than in the unimodal model.

In imaging modality, both uni- and multimodal learning focus upwards, but unimodal learning emphasizes the left side of the image while multimodal learning distributes contribution evenly across the object's position. The histogram on the right indicates degradation density, with the multimodal model achieving higher accuracy by focusing on areas with high degradation density, while the unimodal model does not emphasize these areas as much.

For text modality, the top 10 important features with the highest SHAP values are displayed. In contrast to the unimodal model, the text data has a limited impact on the prediction accuracy in the multimodal model. The unimodal model leverages all available information to produce more accurate results, making the explainable information from the unimodal model more meaningful. In the good group, the unimodal model focuses on features such as "fine", "OK", and "new", while other features appear to work well except for "failure" and "heavy". These two words do not indicate a problem in the good group but are often accompanied by "not" in a sentence, indicating that the machine is still fine. The model does not focus on this feature "not", leading to misunderstandings in the explanation. In the medium group, the unimodal model focuses on words that denote some sign of degradation, such as "degrades", "failure", and "unexpected". In the bad group, words with more serious meanings, such as "needed" and "carefully" are prominent. However, some words that lack meaning, such as "and", "thoroughly", and "some," contribute to a high mean absolute error in the model. Besides, the text modality has limited meaning in the multimodal model at each group level.

For numeric data, the "number of chemicals" has the greatest impact on prediction in the multimodal learning for all 3 groups. Meanwhile, this impact varies across the different groups in unimodal learning. Additionally, in multimodal learning, the "time" feature has a negligible impact on model prediction, as indicated by its SHAP values consistently being under 1 across all three groups. This similarity also holds true for unimodal groups, where the "time" feature has the lowest contribution out of the five features.

Answer the RQ3: To address this research question, we analyzed specific cases with high Mean Absolute Error (MAE) in the prediction of unimodal learning and visualized them in Figure 6. We compared the results of the unimodal and multimodal models in these cases and provided detailed explanations.

Figure 6a compares a unimodal (based on the image data) and a multimodal model for the same sample. The unimodal

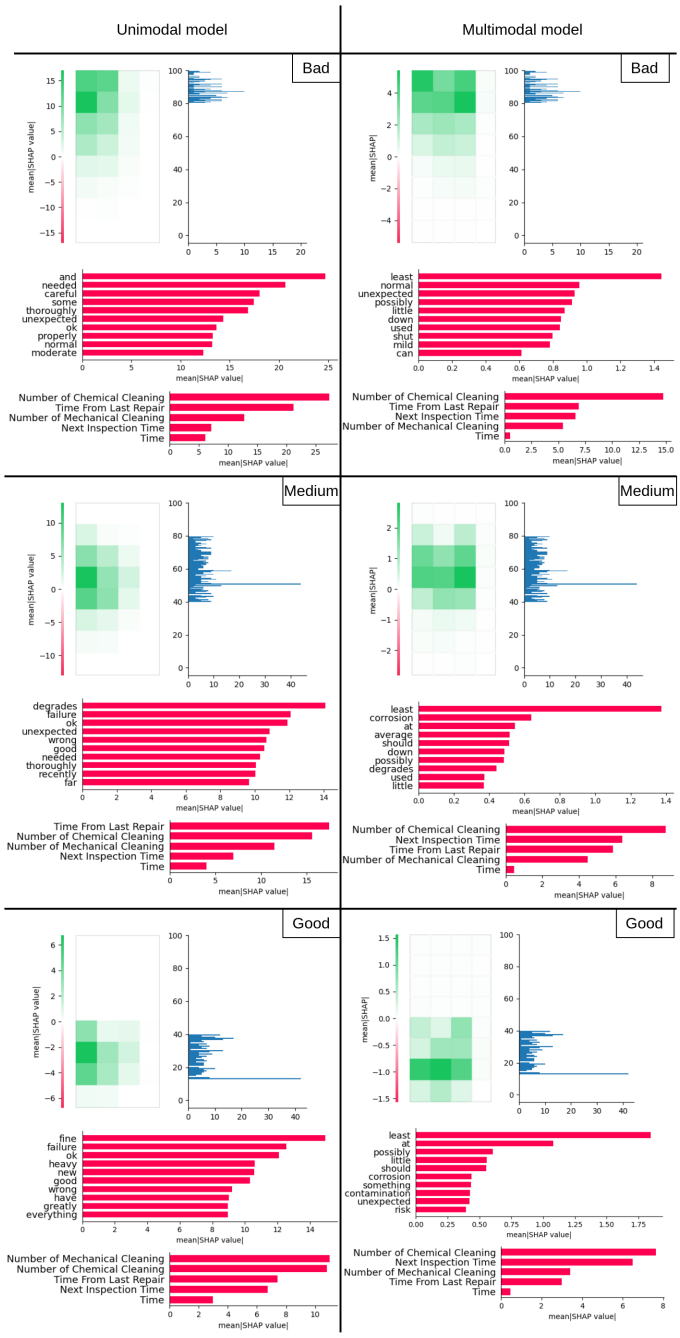


Figure 5. Contribution of features in each degradation group

model has a high MAE of 25.52 because there exists wrong information from image data (i.e. the degradation curve in the image is at the position between the medium and good groups while the true degradation level belongs to the bad group). Besides, multimodal learning leverages useful information from additional numeric features, resulting in a significantly lower MAE of 1.65 in the prediction result.

Figure 6b displays the explanations and results for a sample of unimodal learning based on numerical data and the corresponding result when using multimodal learning. The unimodal model has a high MAE of 39.9 due to a lack of information from the image data, which is included in the multimodal model. The "number of mechanical cleanings", "number of chemical cleanings", and "time from last repair" are important features in both models, but the multimodal model also considers the "time" and "next inspection time" features, which play a significant role in determining the machine's state. The order of these features in each explanation contributes to the higher accuracy of the multimodal model.

4. Conclusion

This paper explored explainable multimodal learning for steam generator degradation prediction. SHAP was used to determine important features of each modality and investigate how multimodal learning can overcome low-quality data from a single modality. Results showed that the numerical modality was the most important, followed by image and text data. The multimodal model achieved higher accuracy in areas with high degradation density for the image modality, while the text modality had limited impact in the multimodal model. Among numeric features, the "number of chemicals" had the greatest impact on prediction in the multimodal model for all degradation groups. From these findings, this study has identified potential avenues for improving the accuracy of the predictions of the SG's degradation by using multimodal learning.

This paper serves as a foundation for future research in applying multimodal learning to industrial applications. Further investigation into the inner mechanisms of the multimodal model, including the interaction between individual nodes and modalities, could yield valuable insights for enhancing model accuracy.

Acknowledgement

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant ANR-22-CE10-0011-01 (project X-IMS)

References

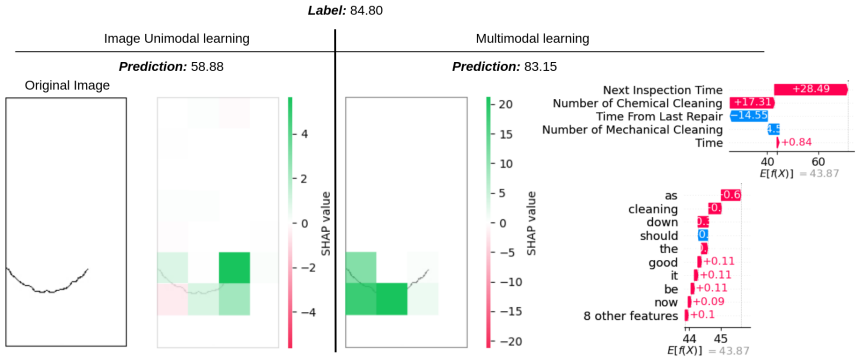
Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2010, 06). Slic superpixels. *Technical report, EPFL*.

Amin, O., Brown, B., Stephen, B., & McArthur, S. (2022). A case-study led investigation of explainable ai (xai) to support deployment of prognostics in the industry. In *Phm society european conference* (Vol. 7, pp. 9–20).

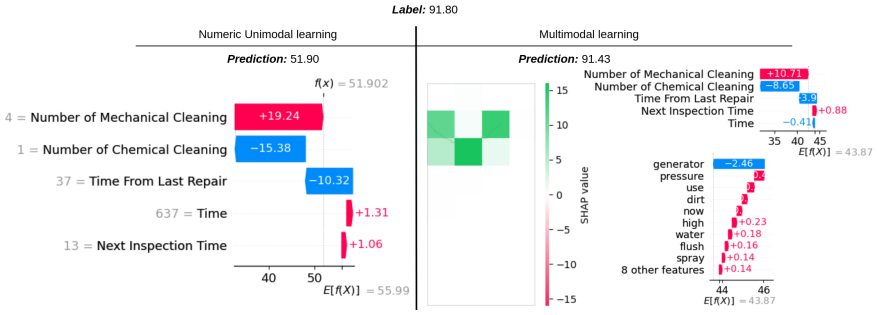
Efron, B. (1992). *Bootstrap methods: another look at the jackknife*. Springer.

Girard, S. (2014). *Physical and statistical models for steam generator clogging diagnosis*. Springer.

Jabeen, S., Li, X., Amin, M. S., Bourahla, O., Li, S., & Jab-



(a) Illustration of a specific case with high MAE in unimodal learning based on image data



(b) Illustration of a specific case with high MAE in unimodal learning based on numeric data

Figure 6. Explaination of some particular "wrong prediction" cases in unimodal lerning

bar, A. (2023). A review on methods and applications in multimodal deep learning. *ACM Transactions on Multimedia Computing, Communications and Applications*, *19*(2s), 1–41.

Jiao, Q., & Zhang, S. (2021). A brief survey of word embedding and its recent development. In *2021 ieee 5th advanced information technology, electronic and automation control conference (iaeac)* (Vol. 5, pp. 1697–1701).

Joshi, G., Walambe, R., & Kotecha, K. (2021). A review on explainability in multimodal deep neural nets. *IEEE Access*, *9*, 59800–59821.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Mckinley, T., Somwanshi, M., Bhave, D., & Verma, S. (2020). Identifying nox sensor failure for predictive maintenance of diesel engines using explainable ai. In *Phm society european conference* (Vol. 5, pp. 11–11).

Nguyen, K. T. P., Medjaher, K., & Tran, D. T. (2023, April). A review of artificial intelligence methods for engineering prognostics and health management with implementation guidelines. *Artificial Intelligence Review*, *56*(4), 3659–3709.

Nor, A. K. M., Pedapati, S. R., Muhammad, M., & Leiva, V. (2022). Abnormality detection and failure prediction using explainable bayesian deep learning: Methodology and case study with industrial data. *Mathematics*, *10*(4), 554.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).

Sanakkayala, D. C., Varadarajan, V., Kumar, N., Soni, G., Kamat, P., Kumar, S., … Kotecha, K. (2022). Explainable ai for bearing fault prognosis using deep learning techniques. *Micromachines*, *13*(9), 1471.

Srinivasan, S., Arjunan, P., Jin, B., Sangiovanni-Vincentelli, A. L., Sultan, Z., & Poolla, K. (2021). Explainable ai for chiller fault-detection systems: Gaining human trust. *Computer*, *54*(10), 60–68.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Yang, Z., Baraldi, P., & Zio, E. (2021). A multi-branch deep neural network model for failure prognostics based on multimodal data. *Journal of Manufacturing Systems*, *59*, 42–50.