Multimodal learning strategies for industrial machine health diagnostics and prognostics under data scarcity

Sagar Jose

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Université de Toulouse

Doctorat de l'Université de Toulouse

préparé à Toulouse INP

Stratégies d'apprentissage multimodal pour le diagnostic et le pronostic de la santé des machines industrielles dans un contexte de manque de données

---

Thèse présentée et soutenue, le 14 octobre 2024 par

Sagar JOSE

École doctorale
SYSTEMES

Spécialité
Génie Industriel

Unité de recherche
LGP - Laboratoire Génie de Production

Thèse dirigée par
Kamal MEDJAHER et Thi Phuong Khanh NGUYEN

Composition du jury
M. Christophe BÉRENGUER, Président, Université Grenoble Alpes
M. Marcos ORCHARD, Rapporteur, Universidad de Chile
M. Piero BARALDI, Rapporteur, Politecnico di Milano
M. Ali ZOLGHADRI, Examinateur, Université de Bordeaux
Mme Louise TRAVÉ-MASSUYÈS, Examinatrice, CNRS Occitanie Ouest
M. Kamal MEDJAHER, Directeur de thèse, UTTOP
Mme Khanh T. P NGUYEN, Co-directrice de thèse, UTTOP

Membres invités
M. Ryad Zemouri, Institut de recherche d'Hydro-Québec
M. Jayant Sen Gupta, Airbus

Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Institut National Polytechnique de Toulouse (INP Toulouse)*

Présentée et soutenue le *14/10/2024* **par :**

Sagar JOSE

Multimodal Learning Strategies for Industrial Machine Health Diagnostics and Prognostics under Data Scarcity

JURY

Christophe Bérenguer	Professeur d'Université (Univ. Grenoble Alpes)	Président du Jury
Marcos Orchard	Professeur d'Université (Univ. de Chile)	Membre du Jury
Piero Baraldi	Professeur d'Université (Politecnico di Milano)	Membre du Jury
Ali Zolghadri	Professeur d'Université (Univ. de Bordeaux)	Membre du Jury
Louise Travé-Massuyès	Directice de Recherche (LAAS,CNRS)	Membre du Jury
Ryad Zemouri	Chargé de Recherche (Institut de recherche d'Hydro-Québec)	Invité
Jayant Sen Gupta	Directeur de Recherche (Airbus AI)	Invité

École doctorale et spécialité :
EDSYS : Informatique 4200018
Unité de Recherche :
Laboratoire Génie de Production (LGP)
Directeur(s) de Thèse :
M. Kamal MEDJAHER et *Mme. Khanh T. P. NGUYEN*
Rapporteurs :
M. Marcos ORCHARD et *M. Piero BARALDI*

Acknowledgments

I am immensely grateful to my doctoral advisors, Prof. Kamal Medjaher and Dr. Khanh T.P Nguyen. Khanh, I must thank for showing me a work ethic that inspired me. She is the one who has listened to my vague ideas and made actual action plans out of them. Kamal has been a guide and mentor throughout. The wisdom you've shared and the example you've set are deeply valuable. On multiple occasions during the thesis, you've reminded me that progress in research is not linear, but a step function. Just keep working even without any visible progress, and one day, you reach the next step up, you said. Those words kept me going during some of the more difficult moments of the thesis. I've always been lucky to work with good bosses. But Khanh and Kamal have set the bar so high that whoever I work with in the future have very big shoes to fill.

Dr. Ryad Zemouri, who I worked closest with in Montreal, has contributed a lot to this thesis. Our discussions would usually start at 7 a.m in the bus and end only when we get off the metro after work. Your ambitious plans and daring gave me courage to attempt long shots. Prof. Antoine Tahan facilitated a lot of the work in Montreal. Though very busy, he is generous with his time and has a keen ability to cut right to the heart of the matter. Dr. Mélanie Lévesque, thank you for sharing your vast expertise on electrical systems with us. Without your knowledge and cooperation, this thesis would not be what it is.

I also thank the members of my jury for their insightful review of my work, their feedback and thought provoking questions during the defense.

Weikun Deng, bro, you're an inspiration in life and science. I'm so glad that I got you moved to my office, and hope that our journey is only beginning here. Yan Yan, thank you holding the rope firmly in our climbing adventures. To all those who helped out when I first came here: Fabio, Martin, Anthony, Ajdin, Eric, Pierre, Sylvain, Caroline, Lola, Charlotte, and other friends and colleagues at the lab, I thank you all.

I thank my mom for being the strongest example of perseverance, and for insisting I go to university (when I had other plans which I will never admit to). Dad, any ability I have in math came from you. Sis and lil bro, thank you both for your belief in me.

Monique, Sophie and Jean-Louis, thank you for your kindness these years. Hai Bo, thank you for the inspiration and for making Montreal awesome.

To all who have helped me directly and indirectly over the years, I thank all of you. Science is a long road. It is a difficult road. But it does not have to be a lonely road. Thank you for walking with me.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
CBM	Condition Based Maintenance
CM	Condition Monitoring
CNN	Convolutional Neural Network
DL	Deep Learning
EDA	Exploratory Data Analysis
EMD	Earth Mover's Distance
FDD	Fault Detection and Diagnostics
GAN	Generative Adversarial Network
GNN	Graph Neural Network
HI	Health Indicator
LLM	Large Language Model
$LSTM$	Long-Short Term memory
MAE	Mean Absolute Error
$MAPE$	Mean Absolute Percentage Error
ML	Machine Learning
MSE	Mean Square Error
NLP	Natural Language Processing
PCA	Principal Component Analysis
PD	Partial Discharge
PDA	Partial Discharge Analyzer

PDC	Polarizing and Depolarizing Currents
PHM	Prognostics and Health Management
$PRPD$	Phase Resolved Partial Discharge
$RMSE$	Root Mean Square Error
$RMSprop$	Root Mean Square Propagation
RNN	Recurrent Neural Network
RTF	Run-to-Failure
RUL	Remaining Useful Life
$t-SNE$	t-Distributed Stochastic Neighbor Embedding
VAE	Variational Auto-Encoder
VI	Visual Inspection

List of Publications

Journal Papers (Published)

- **JP1**: Sagar Jose, Khanh T. P. Nguyen, Kamal Medjaher, Ryad Zemouri, Mélanie Lévesque, and Antoine Tahan. "Fault detection and diagnostics in the context of sparse multimodal data and expert knowledge assistance: Application to hydrogenerators." Computers in Industry 151 (2023): 103983.

- **JP2**: Sagar Jose, Khanh T. P. Nguyen, Kamal Medjaher, Ryad Zemouri, Mélanie Lévesque, and Antoine Tahan. "Advancing multimodal diagnostics: Integrating industrial textual data and domain knowledge with large language models." Expert Systems with Applications 255 (2024): 124603.

Journal Papers (Submitted)

- **JP3**: Sagar Jose, Khanh T. P. Nguyen, and Kamal Medjaher. "Enhancing Industrial Prognostic Accuracy in Noisy and Missing Data Context: Assessing Multimodal Learning Performance." (Submitted to Journal of Intelligent Manufacturing (May 2024)).

- **JP4**: Sagar Jose, Khanh T. P. Nguyen, Kamal Medjaher, Ryad Zemouri, Mélanie Lévesque and Antoine Tahan. "A modular deep learning methodology for multi-fault machine health diagnostics from sparse and imbalanced multimodal data." (Submitted to Neurocomputing (August 2024)).

- **JP5**: Sagar Jose, Ryad Zemouri, Khanh T. P. Nguyen, Kamal Medjaher, Mélanie Lévesque and Antoine Tahan. "Prognostics of complex machinery with sparse multilabel multimodal run-to-failure data: A graph neural network approach." (Submitted to Advanced Engineering informatics (July 2024)).

Conference Papers

- **CP1**: Sagar Jose, Raymond Houe Ngouna, Khanh T. P. Nguyen, and Kamal Medjaher. "Solving time alignment issue of multimodal data for accurate prognostics with

CNN-Transformer-LSTM network." In 2022 8th International Conference on Control, Decision and Information Technologies (CoDIT), vol. 1, pp. 280-285. IEEE, 2022.

- **CP2**: Sagar Jose, Ryad Zemouri, Mélanie Lévesque, Khanh T. P. Nguyen, Antoine Tahan, and Kamal Medjaher. "Informed machine learning for image-data-driven diagnostics of hydrogenerators." IFAC-PapersOnLine 56, no. 2 (2023): 11912-11917.
- **CP3**: Duc An Nguyen, Sagar Jose, Khanh T. P. Nguyen, and Kamal Medjaher. "Explainable multimodal learning for predictive maintenance of steam generators." In PHM Society Asia-Pacific Conference, vol. 4, no. 1. 2023.
- **CP4**: Sagar Jose, Khanh T. P. Nguyen, Kamal Medjaher, Ryad Zemouri, Mélanie Lévesque, and Antoine Tahan. "Bridging expert knowledge and sensor measurements for machine fault quantification with large language models." In 2024 IEEE International Conference on Advanced Intelligent Mechatronics (AIM), pp. 530-535. IEEE, 2024.
- **CP5**: Sagar Jose, Khanh T. P. Nguyen, Kamal Medjaher, Ryad Zemouri, Mélanie Lévesque, and Antoine Tahan. "From Fragments to Futures: Construction of Synthetic Run-to-Failure Trajectories for Fault State Prognostics." (**Best paper award winner** at the 2024 Prognostics and System Health Management Conference (PHM 2024)).

Book Chapters

- **BC1**: Sagar Jose, Khanh T. P. Nguyen, and Kamal Medjaher. "Multimodal Machine Learning in Prognostics and Health Management of Manufacturing Systems." In Artificial Intelligence for Smart Manufacturing: Methods, Applications, and Challenges, pp. 167-197. Cham: Springer International Publishing, 2023.

Chapter 1

Introduction

Contents

1.1 Preamble

Since the day the prehistoric man crafted his first tools from stone, we have sought ways to prevent our creations from breaking down. Today, in the age of digital intelligence, we are rapidly revolutionizing how we optimize the maintenance of our modern machines. Maintenance is one industrial operation with significant potential for cost reduction. With computation capabilities advancing rapidly, using observed history of a machine to predict its future health evolution is gaining prominence in this field of research. However, the prerequisite of having extensive observed historical data is a bottleneck that hampers progress. This thesis aims to bridge the gap between data-driven machine health forecasting and the limitations of industrial data by leveraging multiple sources and types of data that have largely been overlooked. This includes photographs, text, and other types of data, focusing on supplementing the shortage of information from any single data source. To mitigate the data shortage, we develop methodologies for model training techniques that efficiently use multiple types of data and incorporate domain knowledge. By addressing these industrial data challenges, this thesis seeks to fill the gap between state-of-the-art academic research and practical industrial applications.

In particular, this thesis focuses on addressing the lack of sufficient data for effective predictive maintenance. Data-driven predictive models depend heavily on historical data to forecast future machine health accurately. However, in many industrial settings, the

availability of such comprehensive data is limited, posing a significant challenge to developing dependable predictive models.

To overcome this challenge, this research explores the integration of various types of data to enrich the information available for machine health monitoring. Traditional predictive maintenance models primarily rely on sensor signals. However, valuable insights can be gained by incorporating additional data sources such as photographs, text reports, and other relevant information. By leveraging these diverse data types, we aim to provide a more comprehensive view of machine health.

The rationale behind this approach can be understood by considering how human intelligence operates on information gained through the five senses — sight, hearing, touch, smell, and taste. In contrast, neural network-based artificial intelligence models commonly found in predictive maintenance studies typically utilize only sensor signals. This narrow focus limits the capabilities of the models. Even in the few studies that utilize multiple data sources, the interactions among these sources are not well-studied. Consequently, models make predictions based on an incomplete view of machine health, thereby missing out on many insights that could be gleaned from a holistic approach.

In the literature, using varied data sources is gaining traction in fields such as medicine and robotics. However, its adoption in predictive maintenance has been slow. This underutilization potentially limits industries from realizing significant maintenance cost savings and leaves a wealth of domain knowledge and data untapped. Developing methodologies to harness and integrate these diverse data sources could significantly enhance predictive maintenance practices.

In this chapter, the background and context of the thesis will be presented first, in section 1.2. This will be followed by a discussion of the research motivation and objectives in section 1.3. The scope and boundaries of the thesis will then be outlined. Finally, the chapter will conclude in section 1.4 with an overview of the thesis structure.

1.2 Context and Background

From wheels to airplanes to expansive manufacturing plants, human inventions have continually sought to transcend the limitations imposed by nature. However, these creations inevitably degrade and break down over time. Throughout history, as humans have developed new technologies, efforts have simultaneously been made to slow their natural deterioration. As our dependence on these structures and machines increases, ensuring their reliability, availability, and safety becomes increasingly critical.

Since perpetual self-healing systems are still beyond our reach, it becomes necessary to maintain our machines. Initially, maintenance in the industry was corrective — performed after a failure. This was unacceptable for critical systems, such as aircraft, where failure must be prevented at all costs. Consequently, regular maintenance became the norm, with scheduled interventions regardless of the actual condition of the machine. While thorough, this strategy leads to unnecessary expenses.

The field then advanced to condition-based maintenance (CBM), where the actual health condition of the machine is monitored, and maintenance is performed when predefined conditions are met during operations. This approach conducts maintenance before failure, yet only when necessary, optimizing resource use.

However, there is still room for improvement. Rather than waiting for degradation to reach a threshold before performing maintenance, predicting the future health state of a machine based on current observations can offer greater flexibility in resource allocation and maintenance planning. This extension of CBM, called predictive maintenance (PM), involves evaluating the current health state (diagnostics) and forecasting future health (prognostics). CBM or PM implementation is based on information or indicators provided by the Prognostics and Health Management (PHM) algorithms. The terminology is heavily inspired by medical sciences, reflecting numerous parallels to that discipline.

Prognostics can be classified into three groups: physical modeling, data-driven methods, and hybrid approaches. As machines grow more complex, physical modeling becomes intractable, steering the field toward data-driven approaches. Data-driven PHM — particularly machine health prognostics — based on condition monitoring data demands intensive computation. Imagine an industrial machine equipped with an array of sensors, continuously generating streams of data. These sensors measure various parameters such as vibration, temperature, and pressure, producing long sequences of data over time. Making sense of these measurements, discerning patterns and anomalies, evaluating the current health condition, and predicting future states is not a trivial task. For large machines with complex degradation mechanisms, this analysis exceeds the capacity of traditional signal processing methods. Indeed, the challenge is not just the volume of data, but the complexity in identifying features hidden in its patterns. Each sensor captures a different aspect of the machine's operation, and these readings can vary significantly under different conditions. Consequently, traditional methods are insufficient for this task. The task of integrating these disparate data, identifying anomalous symptoms, and making accurate predictions demands assistance. This brings us to the other pillar of this thesis — artificial intelligence (AI).

Alan M. Turing introduced machine intelligence in the 1950s (TURING (1950)), and the term "Artificial Intelligence" was coined at the 1956 Dartmouth Conference. Early AI research focused on symbolic methods but faced setbacks, leading to AI winters. The

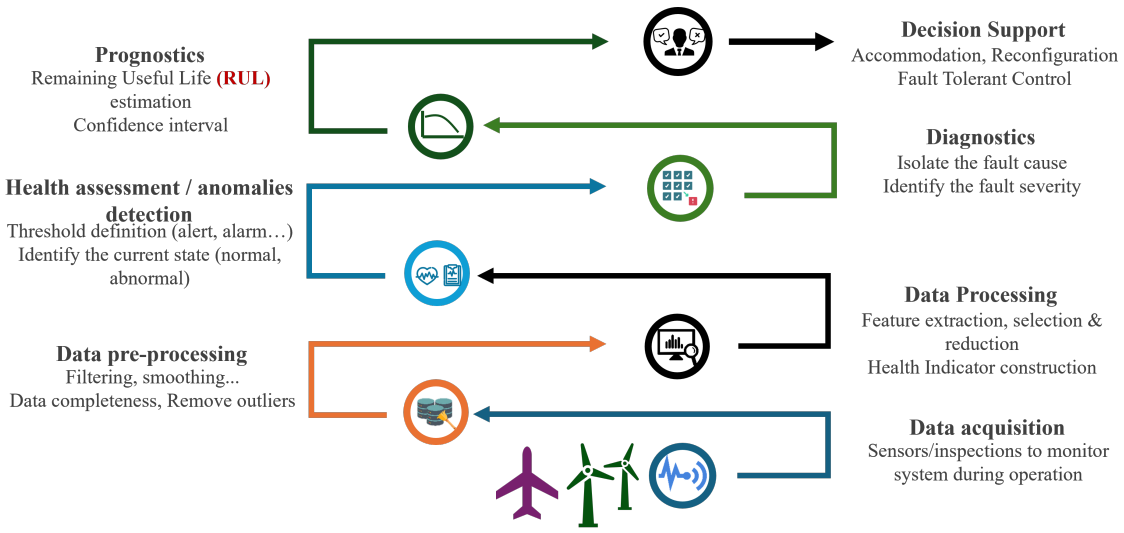


Figure 1.1: Data-driven PHM pipeline, presented as PHM cycle in Gouriveau *et al.* (2016a).

1980s marked a resurgence, with Japan's Fifth Generation Computer Project and the backpropagation algorithm Rumelhart *et al.* (1986). The 1990s boom, driven by internet data and computational power, led to the rise of deep learning. Today, AI applications span natural language processing, computer vision, autonomous vehicles, and more. Multimodal AI, which integrates multiple types of data (text, images, sound) to create more robust and versatile AI systems, is a rapidly advancing field with applications in healthcare and other critical fields.

While the PHM community has embraced some AI tools, data limitations continue to cage progress to simulations and testbench environments. Factors contributing to this gap between literature and industrial applications include the slow investment in data collection, the novelty of deep learning architectures for image and text modalities, the computational resources required, and the hesitation of domain experts to rely on black-box models. Additionally, many state-of-the-art methods remain untested on real-world data.

Generally, data acquisition in the data-driven PHM pipeline (Figure 1.1) is predominated by one-dimensional sensor signals like vibration and temperature. However, interpreting these long sequences of high-frequency data to predict future trends is not intuitive for humans. Indeed, for most tasks in the real world, the human brain relies on multiple modalities of information. Sight and sound are often needed for most of our comprehension of the world. Figure 1.2 shows an illustrated comparison between multimodal data received by the human brain through the five senses and industrial condition monitoring data processed by an artificial neural network.

In the industry, vast amounts of domain knowledge are buried in manuals, maintenance

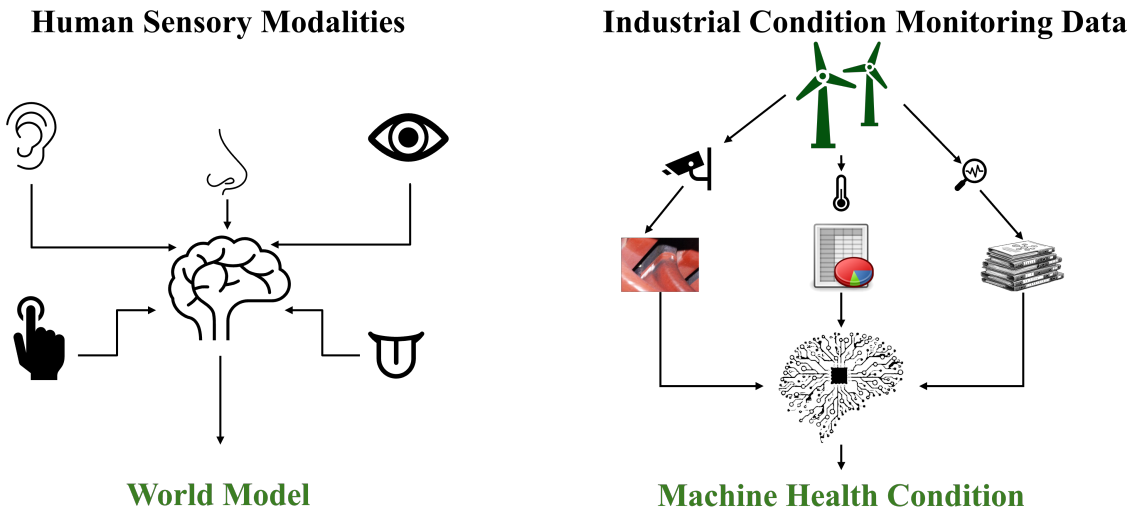


Figure 1.2: Data modalities processed by human brain versus industrial data.

reports, and other texts. Visual inspection by reliability personnel is a practice still carried out in many industries. For human decision-makers, it is natural to consider photographs, descriptions, and numerical observations together to make informed maintenance decisions. Each modality complements and reinforces the others, providing a more complete picture. There is a gap between how humans think and how data-driven models approach machine health prognostics, and bridging this gap could also potentially provide a route for PHM research to cope with the data shortage from industry.

With the latest advancements in deep learning, it is time for the PHM community to explore the use of multimodal data. Understudied modalities in PHM include images and text, and the challenges of integrating these with other diverse data sources must be addressed. The following section will identify the specific challenges to be overcome to achieve this, synthesize the objectives that drive this thesis, and clarify the boundaries within which this research project will be conducted.

1.3 Research Issues and Main Contributions

The current state of the art (as will be presented in Chapter 2) is lacking in multimodal data studies within PHM. In the few studies that even tangentially address this direction, there is no mention of the practical challenges of industrial multimodal data, and the studies remain within simulated datasets that are clean and balanced, far from real data conditions. This is mostly because the industry does not typically follow rigorous multimodal data collection practices conducive to training deep learning models. The general practice of

data scientists, when faced with one modality of high sparsity, is often to discard that modality entirely.

Indeed, it is highly challenging to perform prognostics tasks without sufficient data in the absence of physical models. Yet, the literature rarely seeks to apply the available sparse data, such as through the incorporation of the knowledge of domain experts who already analyze samples of visual inspection images to augment model training on image data. Moreover, black-box deep learning model predictions are difficult for domain experts to interpret. Additionally, training a health forecasting model requires run-to-failure trajectories from each machine, which are unattainable for expensive, operational machines in the real world. Consequently, prognostics remains a challenging task.

From an eagle's eye view, this thesis aims to address the aforementioned gaps in the current state of the art by developing methodologies to perform PHM tasks from multimodal data, integrating domain knowledge and cross-modal interactive learning, addressing multimodal data challenges of industrial condition monitoring data, exploiting untouched expertise repositories such as industrial texts, and performing end-to-end prognostics under the constraints of high data scarcity often seen in industrial data. The developed methodologies must be applied to real industrial data and validated by domain experts while maintaining generalizability and reproducibility. Achieving these ambitious objectives requires addressing the following critical challenges:

1. **Limitation of data quality and availability**: Multimodal datasets obtained from industrial condition monitoring often suffer from issues such as missing data and noise. These limitations can significantly impact the performance of prognostic models. It is crucial to develop techniques that can robustly handle incomplete and noisy datasets to maintain reliability of predictions under poor data conditions.

2. **Variable sparsity of multimodal condition monitoring data**: Industrial datasets are often sparse and irregular, with certain data modalities being more frequently recorded than others. This variability poses a challenge for training robust models. Methods need to be developed to ensure that predictions are unbiased regardless of the absence or dominance of certain condition monitoring data.

3. **Label imbalance**: Multimodal monitoring data often exhibits high label imbalance, especially in the context of multi-fault diagnostics for complex machines. Handling this imbalance is crucial to avoid biased predictions and ensure that the models can accurately diagnose a wide range of faults.

4. **Agreement with expert knowledge**: Incorporating domain expertise and subjective elements present in certain PHM tasks is essential for improving model reliability and acceptance. This requires exploring methods to integrate expert knowledge, particularly the knowledge accumulated in textual form within the industry, into the AI

models. This integration can help create more interpretable models that align with human understanding and reasoning.

5. **Lack of run-to-failure data**: Performing machine health prognostics without complete run-to-failure data is a significant challenge in industrial PHM. Most expensive and critical machines do not have comprehensive failure histories due to maintenance interventions.

Implications for industry and academia

Addressing the aforementioned challenges, this thesis will not only advance the theoretical understanding of multimodal learning and predictive maintenance but also provide practical solutions for improving the reliability and safety of industrial machines.

Contributions to academia

1. *Bridging the gap between theory and practice:* The thesis aims to advance the state of the art in PHM by bridging the gap between simulated data conditions and real-world industrial data. It will provide a framework for developing predictive models that are applicable in real industrial settings.

2. *Multimodal data integration:* This study will highlight the importance of multimodal data integration and demonstrate effective techniques for combining different data types. This will pave the way for further research in multimodal learning techniques and their applications in PHM.

3. *Incorporation of domain expertise:* By incorporating domain expertise into predictive models, this research will enhance the interpretability and acceptance of AI models in the industry. This will foster greater collaboration between data scientists and industry practitioners.

4. *Advancing predictive maintenance models:* The resilience of models to poor condition monitoring data and their ability to provide comprehensive health state evaluations will set a new benchmark for predictive maintenance. This will promote the development of more accurate and reliable prognostic models.

5. *Promotion of interdisciplinary research:* The methodologies developed will serve as a foundation for interdisciplinary research, combining AI, data science, and domain-specific knowledge. This will encourage more collaborative efforts to address complex problems in PHM.

Contributions to industry

1. *Enhanced reliability and safety:* The developed methodologies aim to provide robust tools to improve the reliability, availability, and safety of critical machinery. By leveraging multimodal data sources, industries can achieve more accurate and comprehensive machine health assessments.

2. *Encouragement of rigorous data collection:* The adoption of these methodologies could encourage industries to invest in more rigorous data collection practices. This will involve integrating diverse data types, such as sensor readings, images, and textual reports, into their maintenance strategies.

3. *Utilization of limited data:* The techniques developed must demonstrate the value of utilizing available data, even when sample sizes are small or data is incomplete. This will help industries make better use of their existing data resources rather than discarding potentially useful information due to perceived limitations.

4. *Improved maintenance planning:* The ability to integrate multimodal data and expert knowledge could lead to more effective maintenance planning and resource allocation. This will help in predicting future states of machinery more accurately, thereby reducing downtime and maintenance costs.

Scope and limitations of the thesis

The scope of this thesis will remain within industrial machine health diagnostics and prognostics. The applications could include energy systems, transport systems, and other large and expensive machines with complex degradation mechanisms. The contributions of this thesis will be suited to systems where multimodal condition monitoring data can be collected. In terms of scientific approaches, only data-driven methods will be studied, and physics-based models are not addressed due to the impracticality of designing such models for the scale of application systems of interest. Detailed analysis of well-studied signals in the literature, such as vibration and temperature, will not be revisited. Instead, the focus will be on learning from multiple data modalities together, with detailed analysis centered on less-studied modalities in the field, such as images and text.

Because these areas are not extensively addressed in the literature, this thesis faces several challenges due to the novelty of the questions addressed and the scarcity of comparable studies in the literature. Although multiple methods are compared throughout the thesis, no existing work directly addresses the specific questions explored here. Furthermore, the methodologies developed are applied to industrial data that are not publicly available at the time of writing.

It is important to note that the development and results of this thesis were achieved under certain resource constraints. While the use of larger pre-trained models and greater computational resources could potentially improve quantitative metrics, the methodologies

themselves are designed to be independent of these limitations. Additionally, this thesis focuses on diagnostics and prognostics within the PHM discipline and does not explore maintenance optimization based on prognostic information.

1.4 Thesis Outline

In Chapter 1, we presented the overall introduction to the thesis, the background, objectives, and its scope.

Chapter 2 presents the overview of the literature on multimodal learning, and the existing works on data-driven PHM using multimodal data. This chapter identifies the literature gap and positions the thesis.

Chapter 3 presents the first exploration of multimodal learning within the PHM context, studying in depth the impact of missing and noisy data. The chapter also develops a methodology to make a model resilient to poor data quality conditions.

Chapter 4 develops a diagnostics methodology from multimodal data and applies it to industrial data from a fleet of hydrogenerators, addressing several challenges such as alignment, missing data, and sparsity. This chapter also presents a methodology for incorporating expert knowledge into the model design pipeline to mitigate data shortage, as well as learning from text data to account for subjectivity in health-level quantification.

Chapter 5 addresses prognostics in the industry from the context of multimodal data. Specifically, it addresses the challenge of huge data scarcity and class imbalance. This chapter also presents a graph-based method to perform health state prognostics without any run-to-failure data.

Finally, Chapter 6 will summarize the thesis, conclude, and discuss perspectives for the research community to carry forward the work.

Chapter 2

Literature Review and Research Positioning

Contents

"We can only see a short distance ahead, but we can see plenty there that needs to be done."

— Computing Machinery and Intelligence by Alan M. Turing (TURING (1950)).

2.1 Introduction

In this chapter, we introduce the preliminaries and background needed to motivate and appreciate the development in the subsequent chapters of the thesis. We start with a brief introduction to prognostics and health management (PHM) in the industrial context, focusing on data-driven approaches. In section 2.2, we present some well-known benchmark datasets in this domain and highlight the current field focus on unimodal sensor signal analysis. Exploring alternative data sources leads to an introduction to multimodal data and the field of multimodal learning in section 2.3. We then review existing literature within the PHM domain that addresses the exploitation and challenges of industrial multimodal data in section 2.4, divided into studies on diagnostics, prognostics, and maintenance optimization. The literature discussed in these sections leads to identifying specific gaps in knowledge to be explored. The positioning of this thesis to existing literature is presented in section 2.5, and the conclusion of this chapter in section 2.6.

2.2 Brief Introduction to Data-driven Industrial PHM

Maintenance in the industrial context comprises the actions taken during the life cycle of a production system to allow it to continue performing its intended function. Maintenance activity can be performed correctively (after failure) or preventively (before failure). Preventive maintenance can be either periodic or condition-based (Gouriveau *et al.* (2016b)). A subdomain within condition-based maintenance (CBM), called predictive maintenance, is the main focus of this thesis. Successful implementation of a predictive maintenance policy allows for reducing maintenance costs and increasing the availability and reliability of manufacturing systems. However, the effectiveness of predictive maintenance is significantly dependent on the development of a reliable process for prognostics of system health evolution.

Early studies of prognostics are based on developing a physics model of system degradation processes, known as physics-based prognostics (Kim *et al.* (2017)). Although physics-based approaches can offer precise long-term remaining useful life (RUL) predictions, developing physics-of-failure models for real systems is highly challenging. This difficulty is exacerbated by the increasing complexity of manufacturing systems, particularly within the context of Industry 4.0 (Zio (2022)).

With the advancement of sensor technologies and data acquisition systems, data-driven approaches to PHM have gained prominence. The advanced sensing techniques and data analysis tools emerging from machine learning and deep learning disciplines enable the rise of data-driven prognostics, offering an alternative solution to overcome the limitations

of model-based prognostics. Data-driven PHM leverages the condition monitoring data generated by modern industrial systems to develop predictive models for machine health diagnostics and prognostics. These models utilize statistical, machine learning, and deep learning techniques to analyze sensor data, identify patterns, and predict the future health states of the machines.

Definition

Definition 2.1 (Diagnostics and Prognostics):
***Diagnostics**: The process of detecting, isolating, and analyzing anomalies or faults in machines after their occurrence. The primary goal of diagnostics is to understand the current health state of equipment and pinpoint the exact nature and location of any issues. These are often grouped together as fault detection and diagnostics (FDD).*
***Prognostics**: The process of predicting the future health state of machines and equipment, including the estimation of the time to failure (TTF) or remaining useful life (RUL). The objective of prognostics is to provide actionable insights that allow for proactive maintenance planning, thereby preventing unexpected breakdowns and optimizing maintenance schedules.*
While diagnostics focuses on identifying existing problems, prognostics aims to foresee potential future issues. Together, diagnostics and prognostics form the foundation of a robust predictive maintenance framework, enhancing the reliability, availability, and safety of industrial systems.

In literature, data-driven prognostics is a rapidly developing field (Xu *et al.* (2019)) with numerous bench-marking datasets being published (Jia *et al.* (2018)). The most common data used for data-driven PHM in industry include vibration, temperature (Zhao *et al.* (2017), Falk *et al.* (2021), Yan *et al.* (2017)), electric current (Tian *et al.* (2014), Hendrickx *et al.* (2020)), sound (Lu *et al.* (2018) Lu *et al.* (2017)), pressure (Zhao *et al.* (2017)), speed (Pittino *et al.* (2020), Schlechtingen and Santos (2011)), and voltage signals (Bzymek (2017)). Indeed, the shift from model-based to data-driven prognostics marks a significant advancement in PHM. However, leveraging the full potential of data-driven approaches requires addressing the limitations of unimodal data and exploring the integration of multimodal data sources.

Table 2.1 synthesizes the benchmark data sets which are highly cited in PHM literature. One can see that most of the datasets contain unimodal data, i.e., one-dimensional numeric data. These benchmark datasets have played a crucial role in advancing data-driven PHM. Nevertheless, their reliance on unimodal data underscores the need to investigate how multimodal data can be utilized to gain a more comprehensive understanding of machine health. In the manufacturing industry, other data such as visual inspection photographs, operator reports, and more information are collected and stored without being exploited

for training diagnostics or prognostics models. Such data can become a valuable additional source to improve the performance of PHM models (Yang *et al.* (2021)).

In fact, multimodal data are widely used in the healthcare industry (Tekin *et al.* (2015), Yoon *et al.* (2016), Rahimi *et al.* (2016)). For example, integrating medical imaging, patient records, and genetic data has significantly improved diagnostic accuracy and personalized treatment plans. A detailed overview of the application of multimodal data in healthcare is given by Cai *et al.* (2019). The success of multimodal data integration in healthcare demonstrates its potential to address similar challenges in PHM. Exploring multimodal data not only aims to fill the gaps left by unimodal data but also seeks to provide additional layers of information that can be critical when traditional sensor data is sparse or incomplete.

In this view, it is necessary to investigate the question of how to seek and utilize supplementary sources of manufacturing information to complement the machine health indicators obtained from sensors and therefore improve PHM performance. The potential for using multimodal data for PHM purposes, such as RUL prediction, was demonstrated by the study by Yang *et al.* (2021). However, the complexity of these data in terms of structure, as well as the requirements of high computation resources, pose considerable challenges that need to be solved before these data can be exploited to support the RUL prediction. This study demonstrates the value added to PHM tasks by multimodal data, but uses simulated data. Indeed, the constructed signal curve images and repeated simple texts in this dataset does not capture the real complexity of industrial multimodal condition monitoring data.

Although both academia and industry pay some attention to mining multimodal data for improving PHM performance, this avenue of research is still in its infancy. Before reviewing the few works that use multimodal data for PHM, the next section presents a detailed introduction to the concepts of multimodality and an overview of multimodal learning research. The concepts presented in the next section are essential for reviewing existing related works, identifying gaps in the literature, and formulating the research plan for this thesis.

Table 2.1: Benchmark datasets for PHM.

Dataset Name	Data type	Purpose
CWRU Bearing Dataset (Case Western Reserve University (2021))	• Drive end accelerometer data • Fan end accelerometer data • Base accelerometer data	• Motor bearing condition assessment • Fault diagnosis
Tennessee Eastman Process Dataset (Chen (2019))	• Reactor Pressure • Reactor Level • Reactor Temperature • Stripper Level • Stripper Pressure • and other measurements	• Fault detection
SEU Bearing Dataset (Shao (2022))	• Vibration signals • Fault positions	• Fault detection
NASA Bearing Dataset (Lee *et al.* (2007))	• Vibration signals	• Anomaly detection • RUL prediction
PHM2012 Data Challenge Dataset (Nectoux *et al.* (2012))	• Vibration signals • Temperature	• RUL prediction

Continued on next page

Table 2.1 continued from previous page

Dataset Name	Data type	Purpose
Airbus Helicopter Accelerometer Dataset (Sas (2020))	• Vibration signals	• Anomaly detection • Fault detection
AMPERE Dataset (Soualhi *et al.* (2023))	• Speed • Current • Voltage • Vibration	• Motor system monitoring • Fault detection • Fault diagnostics
Numenta Anomaly Benchmark (Ahmad *et al.* (2017))	• Artificially generated numerical data	• Anomaly detection
NASA Turbofan Dataset (CMAPPS) (Saxena and Goebel (2008))	• Total temperature at fan inlet • Total temperature at LPC outlet • Total temperature at HPC outlet • Total temperature at LPT outlet • Pressure at fan inlet • And other numerical data	• Anomaly detection

2.3 Introduction to Multimodality

This section introduces multimodal data to the reader and aims to highlight how it differs from unimodal data in its form and treatment. First, key terminologies in the domain are presented. Then, the rest of this section progresses through an overview of the evolution of multimodal learning and its key challenges. Finally, the latest developments and techniques for multimodal learning with deep learning networks and foundation models are discussed, setting up the necessary background for the subsequent chapters.

2.3.1 Modality in datasets

The word modality has multiple definitions. The first comes from the word 'mode', which refers to the point of maximum frequency in a distribution. The term multimodal in this space refers to a population distribution that has multiple local maxima in the probability density function (Silverman (1981)). Another definition refers to the way information is perceived and understood (Leahy and Sweller (2011)). This definition of modality is more relevant to our study. This is illustrated by the way our human brain receives information from the world and processes it into an understanding of a scenario (Norris (2019)). In detail, we perceive the world through our five senses (sight, hearing, taste, smell, and touch). These are the sensory modalities.

Further, in the context of computing, modality of data refers to the structure in which a computer program receives the data and the way the data are processed to gain knowledge (Lahat *et al.* (2015)). In computing, the most common modalities are vision, audition, language, proprioception, haptics, and so on.

Multimodality refers, in the context of information and data, to the existence of multiple modalities in the same set of data (Caesar *et al.* (2020), Chen *et al.* (2015)). A key concept to understand here is that a dataset is called multimodal when it contains information on multiple modalities to describe features of the same function.

An example comes from the study of communicative behaviors. In-person communication between people consists of three types of communicative behaviors: verbal, vocal, and visual. It is important to understand that even as a person is speaking verbally, information can be conveyed at the same time through vocal expression such as intonation, laughter, etc. (Tsiourti *et al.* (2017)). Visual information such as gestures, body language, and expressions add to the information. Also to be noted is that within the verbal modality are features such as the lexicon (choice of words), the choice of grammatical structures, and so on.

This leads to an important idea: multiple modalities of data can serve one of two purposes. It can either reinforce the information conveyed through one modality, or it can provide complementary information.

Definition

Definition 2.2 (Multimodal Data):
***Datum**: A single piece of factual information, typically an observation or measurement collected through various means. It represents the most granular level of data, such as a single temperature reading or a specific timestamped event. Data is the plural of datum.*
***Dataset**: An organized collection of data, typically structured in a way that facilitates analysis. A dataset may consist of multiple attributes or features collected over time and stored in formats such as tables, databases, or spreadsheets.*
***Modality**: A specific type or channel of data, representing a particular method of capturing information. Modalities can include various forms such as sensor signals, images, text, audio, and video.*
***Multimodal Dataset**: A dataset that contains data points of multiple modalities. These modalities can include a combination of sensor signals, images, text, audio, and other types of data.*
In this thesis, the term multimodal data refers to a multimodal dataset.

There is much overlap between the usage of the terms multimodal and multimedia. Multimedia data is data including media data types such as text, images, video, audio, drawings, and so on. Multimodal data can include also non-media data such as proprioception, point clouds, etc. In summary, multimedia data can be considered a subset of multimodal data.

Heterogeneous data refers to data that differ in some property. Among the possible differences, one is structural heterogeneity. For our purpose of studying data processing in PHM, structurally heterogeneous data can be considered the same as multimodal data. However, the two terms multimodal data and heterogeneous data are different in some particular contexts. For example, if two sets of data have the same structural representation format, but differ in their population distribution, the term "heterogeneous" is more relevant than the term "multimodal". Particularly, data coming from two sensors, e.g., temperature and pressure, that have similar numerical structures, are considered unimodal data.

The definitions and approaches to understanding the term multimodality have been compiled by Parcalabescu *et al.* (2021). As seen so far, the definitions of multimodal data and multimodality are not conclusive in the literature yet. However, in this study, the definition of multimodal data as structurally heterogeneous data is adequate.

2.3.2 Multimodal learning

Multimodal learning is defined as an activity of extracting useful knowledge from multimodal data while giving due consideration to cross-modal influences. Learning from multiple modalities is important because the information in the real world often involves more than one modality. In fact, in a dataset containing data from different modalities, one modality could carry information that is not available from the other modality. An example is an image of a city with its caption mentioning the name of the city (Srivastava and Salakhutdinov (2012)). Without the textual information, the name of the city could be hard or impossible to deduce from the image alone.

This section aims to give an overview of multimodal machine learning. It begins with a brief look at the historical evolution of multimodal learning in subsection 2.3.3, and the impact of multimodal data in different scientific fields, particularly life sciences and robotics.

2.3.3 Evolution of multimodal machine learning

In the literature, the evolution of multimodal learning is seen to be chronologically separated into four time periods (Morency *et al.* (2022), Baltrušaitis *et al.* (2018)): 1970 - 1980, 1980 - 2000, 2000 - 2010, and after 2010. This is shown in Figure 2.1. There is fifth, very recent trend that could be considered a new age, but is still nascent.

As shown in the figure, the field saw a groundbreaking shift around 2010 from multimodal research to multimodal machine/deep learning. As further discussion and later chapters will heavily rely on artificial neural network-based deep learning methods, a formal definition is given in Definition 2.3. For a more detailed background of deep learning, readers are directed to LeCun *et al.* (2015). With an understanding of neurons and learning algorithms, it is relevant to know that the transition of multimodal learning research to these tools was due in large part to the following factors. Firstly, the creation and free sharing of new large-scale multimodal datasets. The easy availability of cheap data storage and ease of sharing data through the internet contributed to this. Faster computers and GPU (Graphical Processing Unit) development enabled researchers and developers to implement deep neural networks and train them on large datasets. These two factors - availability of data and high computing capacity - have been touted as the reasons for the renewal of neural architectures in general (Toosi *et al.* (2021)). The third reason is that very high dimensional data such as vision and language could now be represented in a uniform neural encoding in the form of vectors. In the case of vision, the success of convolutional networks in representing features (LeCun *et al.* (1999)) was an influential milestone for deep learning. Then, vision and language being two modalities around which

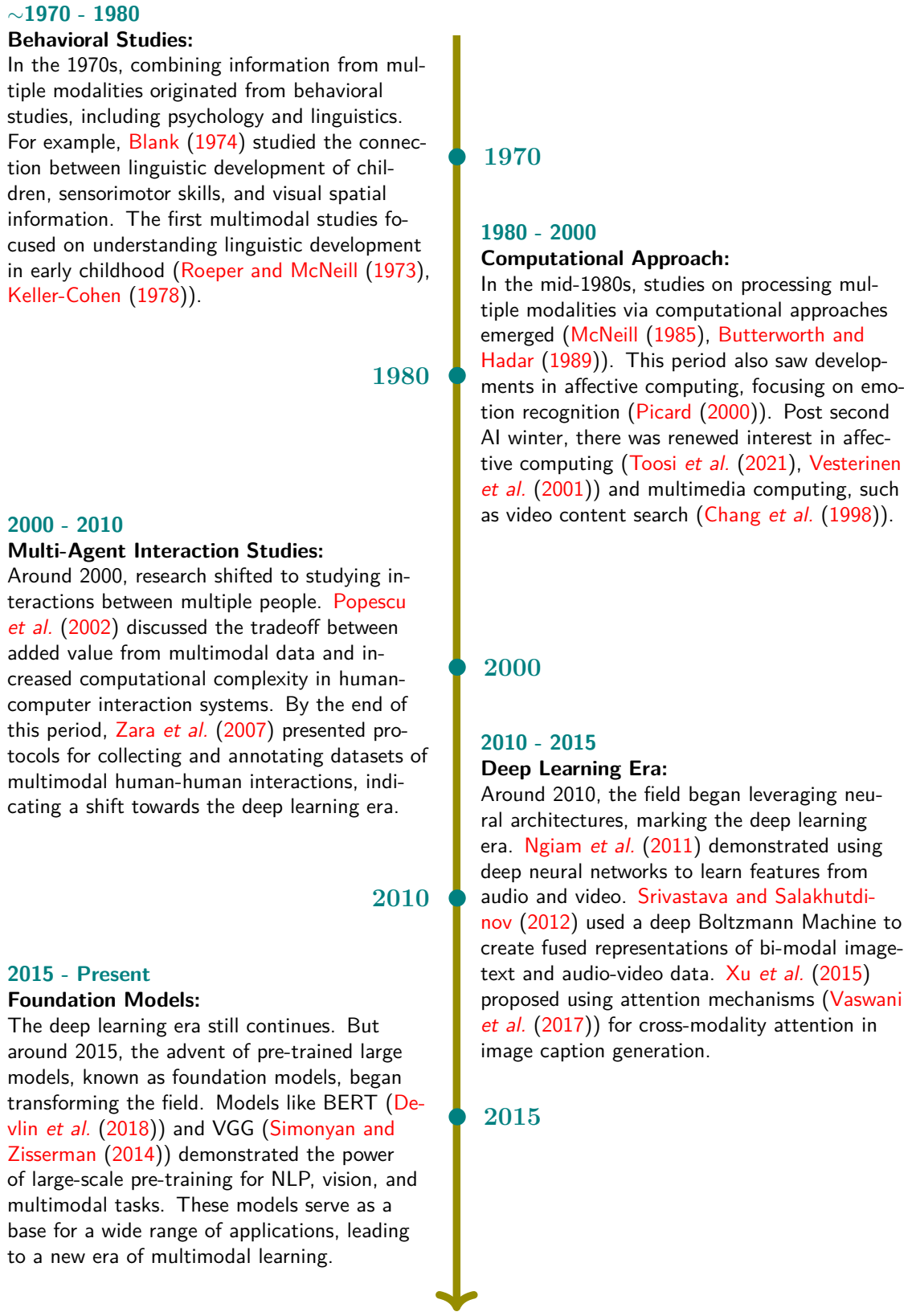


Figure 2.1: Timeline of the evolution of multimodal learning

a large part of multimodal research was oriented, these advancements were key to the rise in research trend on multimodal deep learning.

Definition

Definition 2.3 (Artificial Neural Network):
Neuron: *An artificial neuron is a mathematical function that takes a vector of inputs* $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ *and produces an output* y*. Formally, a neuron computes:*

$$y = \sigma(\mathbf{w} \cdot \mathbf{x}^T + b) \tag{2.1}$$

where $\mathbf{w} = [w_1, w_2, \ldots, w_n]$ *is the weight vector,* b *is the bias, and* σ *is an activation function such as sigmoid, ReLU, or tanh.*
Neural Network: *An artificial neural network is a composition of neurons arranged in layers. If* L *denotes the number of layers, the output of layer* l *is* $\mathbf{h}^{(l)}$*. For input* $\mathbf{x}$*, the layers compute:*

$$\mathbf{h}^{(1)} = \sigma(\mathbf{W}^{(1)}\mathbf{x}^T + \mathbf{b}^{(1)}) \tag{2.2}$$

$$\mathbf{h}^{(l)} = \sigma(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}), \quad \textit{for} \quad l = 2, \ldots, L-1 \tag{2.3}$$

$$\mathbf{y} = \mathbf{h}^{(L)} = \sigma(\mathbf{W}^{(L)}\mathbf{h}^{(L-1)} + \mathbf{b}^{(L)}) \tag{2.4}$$

where $\mathbf{W}^{(l)}$ *and* $\mathbf{b}^{(l)}$ *are the weight matrix and bias vector of the* l*-th layer, respectively.*
Deep Learning: *Deep learning involves neural networks with at least one hidden layer (typically more) between the input and output layers. The term "deep" refers to the depth of the network, i.e., the number of layers.*
Training: *Training or learning in neural networks involves finding the optimal weights* $\mathbf{W}^{(l)}$ *and biases* $\mathbf{b}^{(l)}$ *that minimize a loss function* $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}})$*, where* $\hat{\mathbf{y}}$ *is the predicted output and* $\mathbf{y}$ *is the true output. This is achieved through iterative optimization techniques such as gradient descent:*

$$\mathbf{W}^{(l)} \leftarrow \mathbf{W}^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(l)}} \tag{2.5}$$

$$\mathbf{b}^{(l)} \leftarrow \mathbf{b}^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(l)}} \tag{2.6}$$

where η *is the learning rate, which controls the step size in the optimization process.*

As the scientific development of multimodal learning evolved, the interest and therefore the production of research volume in this field also rose rapidly. In the rest of this section, we will look quantitatively at the development of multimodal learning. As shown in Figure 2.2, with the remarkable success of deep learning methods in the field of computer vision and natural language processing, the interest in the field of multimodal learning has risen rapidly in recent years.

When investigating the fields where most of the work in multimodal learning is done, one can see that a large part of it is in computer science and AI research. Looking at the distribution shown in Figure 2.3, it can be noted that other than computer science and AI domains, a significant share of the work is done in medical science and related fields. It can be inferred that most of the work in computer science and AI would involve the development of algorithms and tools for working on multimodal data, whereas research on medical sciences and robotics would be application-oriented. In the next subsections, we take a global look at the existing works in those fields that apply multimodal data to solve their specific problems. This study is done with a view to examining their potential to solve domain-specific challenges in the PHM field.

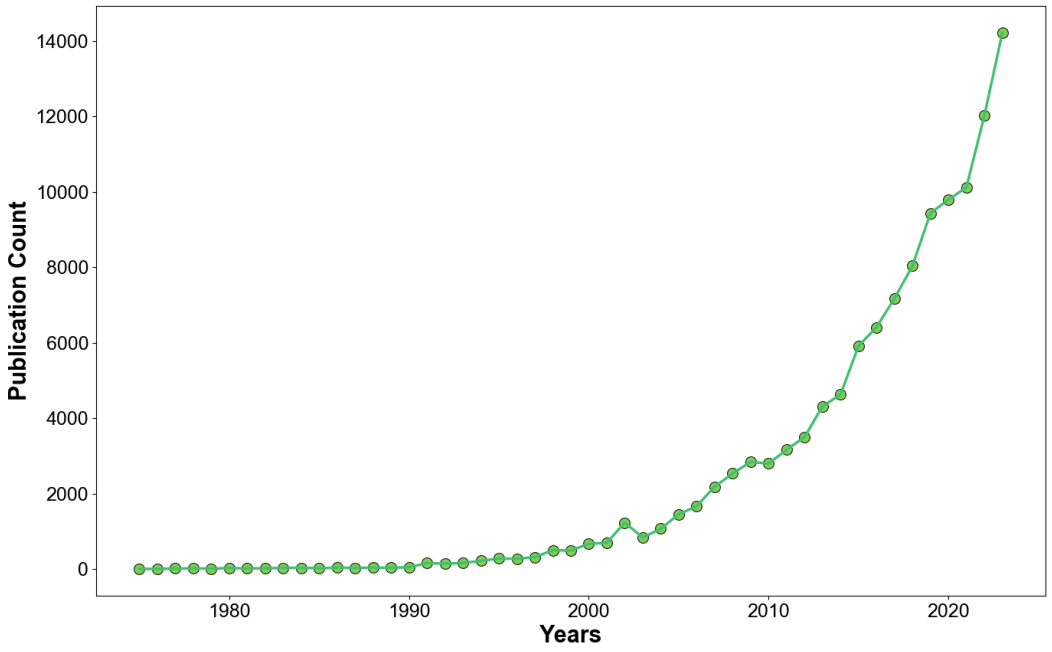


Figure 2.2: Trend of publications on multimodal learning over the years.

2.3.3.1 Multimodal learning in life sciences, medicine, and related fields

From an intuitive point of view, certain analogies can be drawn between healthcare and industrial maintenance. In healthcare, the health state of a human being is observed and treatments are administered as and when necessary to prolong his or her life in the best condition. This is similar to the health management activity of machine systems in industry. Therefore, observations made from studying the application of multimodal data

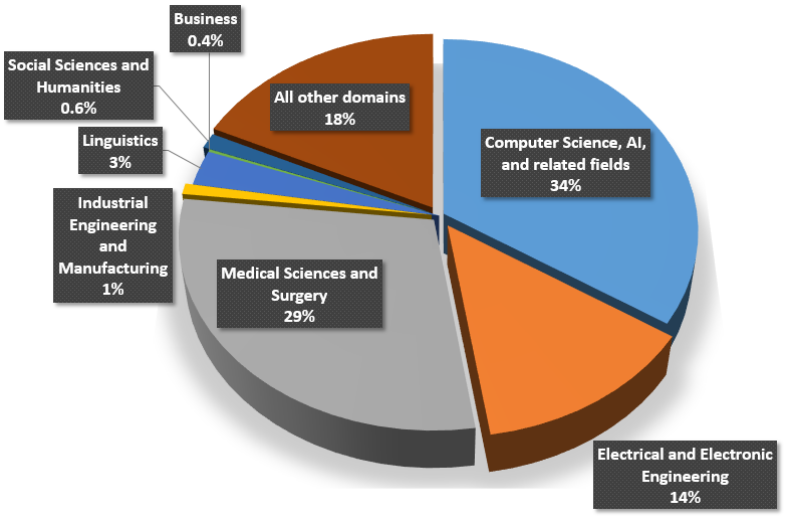


Figure 2.3: Main domains where multimodal research is conducted.

in healthcare could potentially be exploited to apply multimodal data in PHM.

The use of multimodal data along with machine learning techniques is gaining importance in life sciences, medical science, psychology, and other related fields. The associated science is progressing at a rapid pace, with several reviews published every year. Some of the notable papers include Stoyanov *et al.* (2018), Huang *et al.* (2020), Heiliger *et al.* (2022) and Behrad and Abadeh (2022).

Multimodal data in these disciplines not only include medical imaging, data from various scans such as computed tomography (CT), positron emission tomography (PET), magnetic resonance imaging (MRI), and so on, but also omics data, clinical data such as various measurements, demographic information, real-time signals such as electrocardiogram (ECG), electroencephalogram (EEG), etc. Depending on the type of disease or condition, other types of data can also be available.

Spasov *et al.* (2018), Yala *et al.* (2019) and Yoo *et al.* (2019) used convolutional neural networks (CNNs) for medical images and fused the learned features with clinical records to identify a medical condition. Spasov *et al.* (2018) and Yala *et al.* (2019) used simple concatenation to fuse the multimodal data. Yoo *et al.* (2019) reported duplicating the clinical information to solve the dimensionality difference problem between features from images and clinical records. In Huang *et al.* (2020), the authors synthesized fusion techniques used according to the characteristics of the problem to solve. The findings compiled in Table 2.2 are not limited to medical data and can be potentially used for industrial maintenance as well.

Cao *et al.* (2020) discussed the use of Auto-GAN to synthesize data and address the

Comparison of fusion strategies			
Scenario	Early Fusion	Feature fusion	Late fusion
Prediction without all modalities	×	×	✓
Feature level interaction	✓	✓	×
Cross-modal compatible feature extraction	×	✓	×
Training on sparse data	×	×	✓
Training on only one model	✓	✓	×
Ease of model design	✓	×	×
Input concatenation at different abstraction levels	×	✓	×

Table 2.2: Comparison of early, feature level, and late fusion. Adapted from Huang *et al.* (2020).

problem of data sparsity. Li *et al.* (2020) introduced a GAN for retinal disease diagnosis with multimodal images. Hervella *et al.* (2019) used a U-Net for retinal vessel segmentation using multimodal data. Chen *et al.* (2019) used an attention-based method for prognosis of breast cancer from omics and clinical data. This work is particularly interesting to our study because it discusses the design of an architecture for prognostics of the future health state of the system under study.

Maghdid *et al.* (2020) introduced transfer learning with X-ray and CT images from a network trained on pneumonia data to detect COVID-19. Lassau *et al.* (2021) used a deep learning model to extract features from CT images, and then concatenated them with lab tests and other clinical data to input to a logistic regression model for predicting case severity of COVID-19 patients. A notable observation to be made here is that deep learning methods are hard to replace when image modality is involved. A detailed overview of deep learning architectures that have been used with multimodal data in medicine is given in Behrad and Abadeh (2022). Wang *et al.* (2018) introduced TieNet in which radiology images are converted to language-embedded reports by converting the image modality to text. A prerequisite of these methods is the availability of data from multiple modalities.

One significant advantage provided by the comparative maturity of multimodal study in a field such as medicine is the availability of real-world datasets. Notable datasets include:

- MIMIC-CXR dataset Johnson *et al.* (2019) containing 227,835 imaging studies for 65,379 patients along with free-text radiology reports.
- PADCHEST dataset Bustos *et al.* (2020) containing chest X-rays with multi-label annotated reports.

- ImageCLEF challenges Abacha *et al.* (2019) datasets containing image and text for multimodal information retrieval.

From this section, it can be concluded that the use of multimodal data is thriving in life sciences and medical fields, and this is one of the drivers of multimodal deep learning research. The techniques identified in this field could be adapted to other fields, particularly PHM, with promising results.

As discussed at the beginning of this section, healthcare, and industrial maintenance can be compared in certain aspects. Therefore, the increasing use of multimodal data in healthcare provides a promising perspective of multimodal learning in industrial health management.

2.3.3.2 Multimodal learning in robotics, affective computing and other domains

Multimodal data are particularly important for human-robot interaction, where the visual, auditory, language, and proprioception modalities, at the least, have to be combined. Even though the scale of this field cannot be compared to the medical science domain, the scientific advancements made here are significant. A comprehensive review has been made by Spezialetti *et al.* (2020). This study, which focuses on emotion recognition for human-robot interaction, is closely tied to affective computing. Data such as thermal facial images and brain activity signals were studied.

In literature, several works demonstrate the use of CNN-type networks on image data. Barros *et al.* (2015) used a cross-channel CNN to extract features from face expression and body motion data. Álvarez-Sánchez JR (2020) also used a CNN variant for emotion recognition from facial images, EEG, Galvanic Skin Response (GSR), and blood pressure.

Robot manipulation task failures are studied by Inceoglu *et al.* (2021), where the authors present a multimodal dataset comprising RGB images, depth images, and audio from robots. The dataset is then used to train a multimodal neural network to detect incomplete or failed task scenarios. The network design is particularly inspiring for fault detection tasks in the PHM domain. In the network structure proposed by the authors, called FINO-Net, data from comparable modalities such as RGB and depth images are stacked on top of each other and input to the same convolutional path. A separate path for audio data begins with a log mel spectrogram rendering of the audio data which converts the audio into mel frequency spectral coefficient representation. Features from this representation are input into a convolution block. The separate paths are later fused with a dense layer. This philosophy of designing individual paths suited for the treatment of each data modality and fusing the features near the end decision level can be adapted to

PHM purposes, as will be shown in later chapters.

This section has demonstrated that the use of multimodal data is widespread in other fields, with a wide range of techniques applied to solve several challenges. The advancement of neural network-based multimodal learning applications suggests that it is high time for the PHM community to turn their attention to exploring this direction. In the next section, some of the challenges of working with multimodal data will be discussed.

2.3.4 Challenges of multimodal machine learning

In this section, we look at the core technical and scientific challenges that arise when we attempt to perform machine learning or deep learning on multimodal data. According to the studies by Baltrušaitis *et al.* (2018), Gao *et al.* (2020) and Gaw *et al.* (2021), one can cite five principal challenges of multimodal learning: representation, alignment, fusion, co-learning, and translation. Among them, representation and alignment are crucial challenges that need to be solved in many tasks involving multimodal deep learning. The other three challenges are not common to all multimodal deep learning problems but depend on the particular problem addressed.

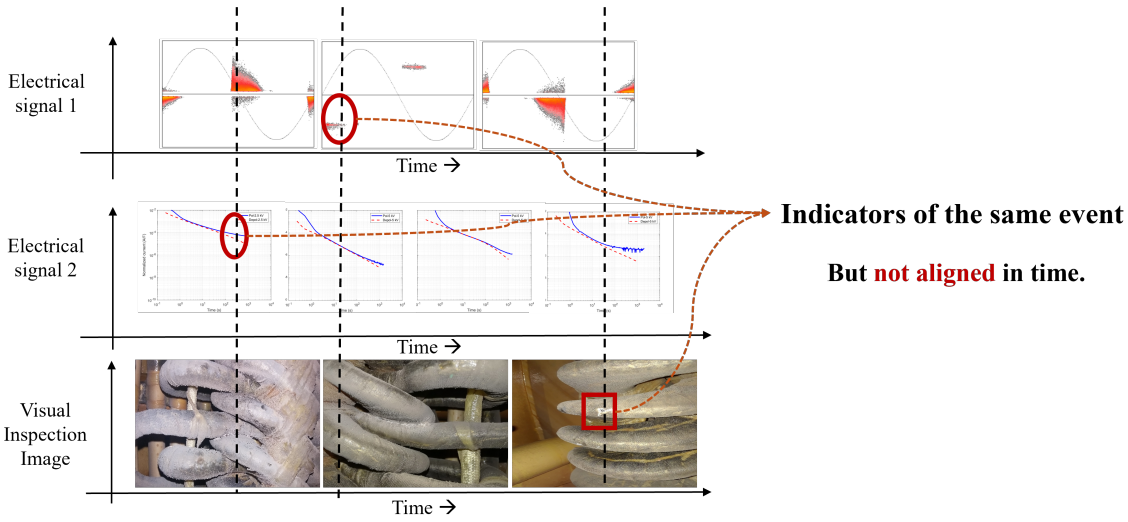


Figure 2.4: Illustration of time alignment issue in multimodal condition monitoring data.

- **Representation:** This challenge involves joining data from multiple modalities into a uniform representation space. The two main approaches are:
  - **Joint representation:** Transforming all modalities into a single combined

representation. For example, the bimodal deep belief network by Ngiam *et al.* (2011) enables arithmetic operations on image and text representations.

- **Coordinated representation:** Each modality is transformed separately, and a coordination spectrum (from strong to weak) is defined. Notable examples include the deep Boltzmann Machine for image captioning (Srivastava and Salakhutdinov (2012)) and audio-visual emotion recognition (Lu *et al.* (2018)).

Representation is a tradeoff problem, balancing information loss from each modality while exploiting complementarity and redundancy.

- **Alignment:** This involves identifying direct relations between elements of different modalities, crucial for temporal data due to synchronization challenges. While several factors contribute to this, the differences in sample collection rate, sequence length, and so on between data from different modalities are crucial. An illustration of the time alignment issue in condition monitoring data is depicted in Figure 2.4. Alignment resolution can be:

  - **Explicit alignment:** Directly finding correspondences using techniques like deep canonical time warping (Trigeorgis *et al.* (2016)), useful for tasks like event reconstruction from partial video, text, and audio descriptions.
  - **Implicit alignment:** Achieving latent alignment as an intermediate step, often using context information methods like attention mechanisms (Vaswani *et al.* (2017)). In PHM, alignment should typically be solved implicitly while training a model to tackle a PHM task such as fault detection.

- **Fusion:** The challenge is to combine different modalities to infer higher-level information, such as emotion from a video. In the context of PHM, this could be inferring the health state of a machine from multimodal condition monitoring data. Key considerations include:

  - **Level of fusion:** The challenge is to determine at what level to fuse one modality with another. Taking the example of image and text, the raw information level for an image is a pixel, and for text, is a word. However, fusing at the level of pixels and words does not necessarily produce useful features. In the case of images, useful features emerge after multiple levels of convolution and pooling. The difficulty lies in learning at what level a feature is mature or insightful enough to benefit from the information from another modality.
  - **Fusion techniques:** Include model-agnostic approaches (D'mello and Kory (2015)) and model-based approaches like deep neural networks (Ngiam *et al.* (2011)), kernel-based methods (Liu *et al.* (2013)), and graphical methods (Lafferty *et al.* (2001)).

- **Co-learning:** This involves transferring learning between modalities, ensuring models can perform tasks with one modality at test time even if trained on multiple modalities. This is important for industrial PHM because, in many cases, a model may be able to learn useful information from one modality when it is trained, but that modality may not be available at use time, due to sensor failure or other constraints. Approaches include:
  - **Strong and weak pairing:** Defining close or loose relationships between datasets.
  - **Joint representation learning:** Learning representations through cyclic translations between modalities, as demonstrated by Pham *et al.* (2019).
- **Translation:** This is the task of converting data from one modality to another, such as generating image captions. Approaches include:
  - **Example-based translation:** Mapping from source to target examples, like nearest neighbor methods.
  - **Model-based translation:** Learning rules to generate translations from examples. An example is forecasting human poses from language descriptions (Ahuja and Morency (2019)).

For industrial machine fault diagnostics and health state prognostics, all the challenges except translation are highly relevant. Thus representation, alignment, co-learning and fusion will be addressed later on.

2.3.5 Tools and techniques used in multimodal deep learning

This section summarizes the most commonly used tools and techniques in multimodal deep learning to solve the challenges described above. The classification (Figure 2.5) is based on the principle behind the network mechanism.

- CNN variants: CNNs are among the best architectures to work on image data. The architecture consists of a convolution operation followed by a pooling and, usually, a fully connected layer. CNNs can capture spatially correlated features in an image, and also from any data that can be represented as an image (Zhang and Wallace (2016)).

- U-Net: The U-Net architecture was primarily developed for image segmentation (Ronneberger *et al.* (2015a)). It consists of a convolutional "shrinking" path where an

image is shrunk to a small dimension, followed by a deconvolutional "expanding" path where the small dimension feature representation is expanded back to the original large dimension. U-Net has been used in multimodal applications in medicine where segmentation is required (Hervella *et al.* (2019)).

- ResNets: ResNet stands for a residual network (Szegedy *et al.* (2017)). It was originally developed to allow neural networks to be very deep without causing vanishing or exploding gradient problems. This was done by adding a residual or skip connection, which is a parallel connection between one layer and the layer after the next one, skipping the layer in between. ResNets have since been used to create many pretrained networks, which have then been used for transfer learning with multimodal data (Zhang and Shi (2020)).

- RNN variants: RNNs are a class of neural networks that work best to model sequence data. Therefore, this is particularly useful for PHM, where the majority of data are temporal sequences (Pascanu *et al.* (2013)). The basic principle of RNN is to parse items in an input series one after the other while updating a hidden state that stores the history of what it has seen before. In this way, RNNs succeed in capturing the sequential relationship in the data.
  However, RNNs are not very good at capturing long-sequence data. Gated Recurrent Unit (GRU) (Chung *et al.* (2014)) and Long Short Term Memory (LSTM) (Staudemeyer and Morris (2019)) are extensions of RNN which overcome some of its disadvantages.

- Attention based: Attention mechanism was originally implemented to solve the natural language translation problem (Vaswani *et al.* (2017)). It is based on the premise of human visual attention. To put it simply, attention mechanism assigns weights to the data depending on how important a piece of data is to the task at hand. Attention based methods, particularly transformer attention, and variants, have succeeded in achieving state-of-the-art performance in both image and text-related tasks. The multimodal-BERT by Khare *et al.* (2021) and the Perceiver model by Jaegle *et al.* (2021) are notable examples. Transformer-based models are widely used as foundation models, which will be discussed in the next section.

- GANs: Generative Adversarial Network was created as a new method of training a generative network (Goodfellow *et al.* (2014)). A GAN consists of two neural networks, called the generator and the discriminator. The generator trains to generate new data that are similar to the training dataset. The discriminator trains to distinguish the 'fake' data produced by the generator. As both generator and discriminator

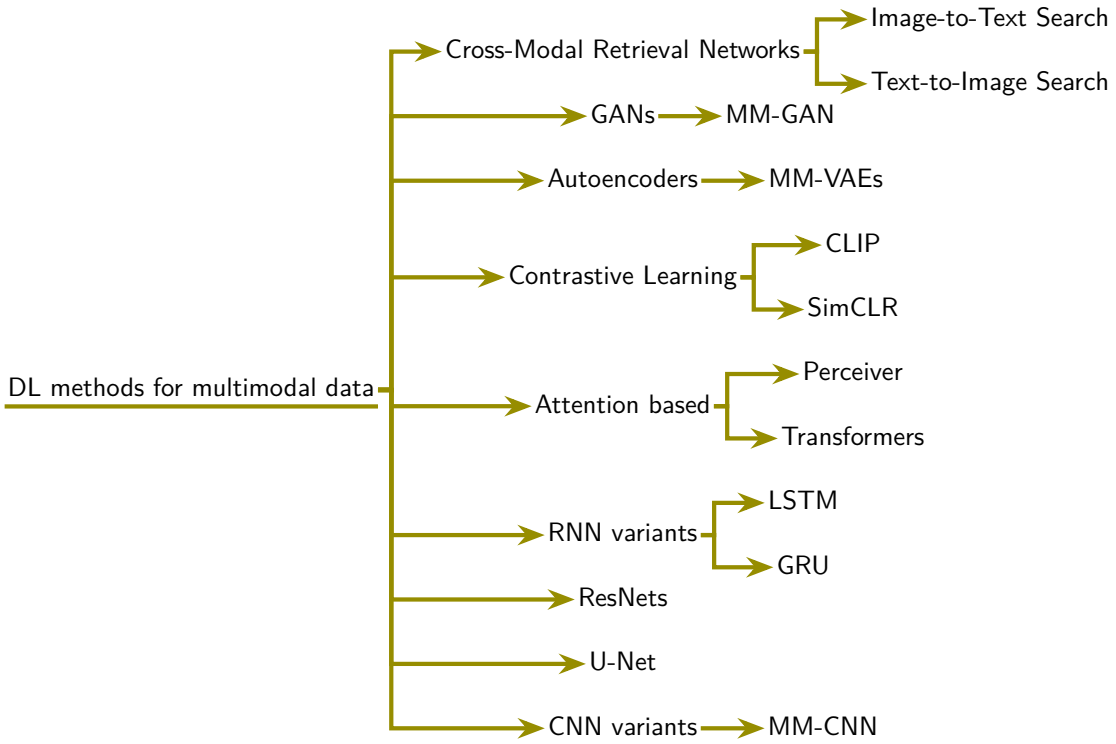


Figure 2.5: Deep learning methods for multimodal data.

are trained, the output from the generator begins to more accurately resemble the original dataset, thereby generating believable data. Variants of GANs have been used in modality translation tasks and for data synthesis (Cao *et al.* (2020)).

- Contrastive Learning: Contrastive learning is a self-supervised learning technique that learns representations by contrasting positive pairs against negative pairs. In multimodal contexts, it aligns representations from different modalities, such as images and text, by maximizing the agreement between corresponding pairs, helping to learn a joint embedding space. Chen *et al.* (2020a) presented the SIMCLR framework for contrastive learning, later extended to multimodal scenarios by Yuan *et al.* (2021), Zolfaghari *et al.* (2021) and Hager *et al.* (2023). Taleb *et al.* (2022) have applied this method for medical imaging with genetic data.

- Autoencoders: Autoencoders learn efficient codings of input data by reconstructing the input from a compressed representation. For multimodal data, they are extended to learn a shared latent space that captures correlations between modalities like images and text, enabling cross-modal tasks such as generating text from images. Cohen Kalafut *et al.* (2023) developed an autoencoder method for joint multimodal

imputation and embedding applied to single-cell genetic data in human and mouse brains.

- Cross-Modal Retrieval Networks: These networks facilitate retrieval tasks across different modalities by projecting data into a common embedding space. They enable tasks like text-to-image and image-to-text search by ensuring semantically similar items from different modalities are close in the shared space, allowing for efficient retrieval based on relevance. The state-of-the-art of these models are transformer-based, but can be considered as a specialized application of representation learning. Deep visual-semantic embedding model (DeViSE) by Frome *et al.* (2013), universal image-text representation (UNITER) by Chen *et al.* (2020b) and improved visual-semantic embeddings (VSE++) by Faghri *et al.* (2017) are examples of cross-modal retrieval networks.

Among the listed methods, transformer-based models have demonstrated remarkable scalability in terms of parameter counts and the ability to leverage vast training corpora, leading to impressive feature extraction capabilities. Consequently, training large models has become a pivotal direction in both research and industry. In the following section, we delve into the potential of this emerging technique for multimodal learning by exploring the concept of foundation models.

2.3.6 Foundation models for multimodal learning

Pre-training large neural networks on large datasets for domain adaptation in a downstream task is developing into a new paradigm for deep learning. As these large models form the foundation for a neural network structure, they are called foundation models. They serve as preliminary feature extractors in a deep learning pipeline. In recent years, a large number of foundation models have been developed and published, making high-scale deep learning on multimodal data possible. While foundation models on multimodal datasets are available (CLIP and Flamingo), another method is to use unimodal foundation models to train a multimodal application model. An introductory overview of pre-trained models trained on image, language, and multimodal data is given in Figure 2.6.

Foundation models for image data have been pivotal in advancing the field of computer vision. These models are pre-trained on large-scale datasets like ImageNet, which contains over 14 million annotated images across 1,000 categories. Training on ImageNet (or other large image datasets) allows these models to learn rich and diverse feature representations, making them well-suited for adaptation to various downstream tasks.

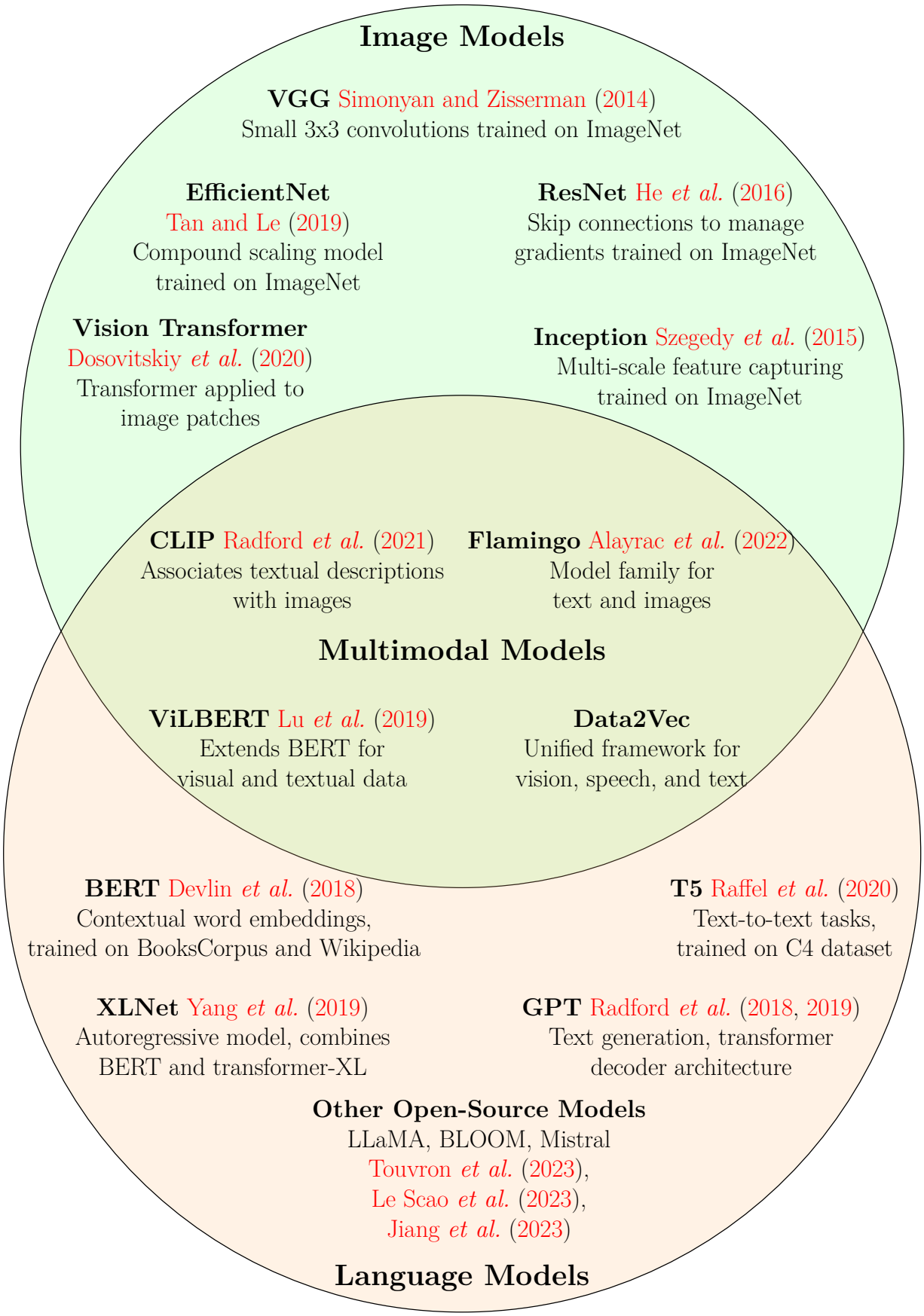


Figure 2.6: Foundation models for multimodal learning

Foundation models for language data have revolutionized natural language processing (NLP) by providing powerful pre-trained models that can be fine-tuned for a variety of downstream tasks. These models are typically trained on large-scale text corpora, such as Wikipedia, Common Crawl, and BooksCorpus, which provide diverse linguistic features and contexts. Training on these datasets allows the models to learn rich language representations that can be adapted to tasks like text classification, translation, and question answering.

Multimodal foundation models are designed to integrate and process information from multiple data modalities, such as text, images, and audio. These models leverage the strengths of individual modalities to provide more comprehensive and context-aware representations. By training on large multimodal datasets, these models can handle complex tasks that require understanding and reasoning across different types of data.

The assortment of models shown in Figure 2.6 highlights that the field of training large deep learning models is rapidly advancing. While not comprehensive, it offers a snapshot of the advancements and possibilities in multimodal learning. It can be seen that there are many pre-trained models available for adaptation to PHM tasks. With pre-trained foundation models established, we will next explore techniques for adapting them to downstream tasks.

2.3.6.1 Techniques for pre-trained model adaptation

Adapting pre-trained foundation models to downstream tasks involves fine-tuning and modifying the models to meet the specific requirements of the target application. This process leverages the rich feature representations learned during pre-training and adapts them for more specialized tasks. Below are some common techniques for adapting pre-trained models:

- **Fine-tuning**: Fine-tuning involves training the pre-trained model on a smaller task-specific dataset. This process adjusts the weights of the model to better fit the new data while retaining the general knowledge acquired during pre-training (Howard and Ruder (2018)). Fine-tuning is particularly effective when the pre-trained model is well-aligned with the downstream task.

- **Feature extraction**: In this approach, the pre-trained model is used as a fixed feature extractor. The pre-trained model processes the input data to generate feature representations, which are then fed into a separate classifier or regressor tailored to the specific task (Zheng *et al.* (2020)). This technique is useful when computational resources are limited or when the task-specific dataset is small.

- **Knowledge distillation**: Knowledge distillation transfers the knowledge from a large pre-trained model (teacher) to a smaller and more efficient model (student). The student model is trained to replicate the behavior of the teacher model, making it suitable for deployment in resource-constrained environments while retaining much of the teacher model's performance (Zhou *et al.* (2024)).

Takeaway

Learning from multimodal data using deep learning has significantly advanced in recent years, with widespread adoption in fields such as medicine. Techniques to extract useful information from images and text have reached a high level of maturity. The availability of pre-trained foundation models for these modalities presents an opportunity to utilize these data in many fields, even in domains such as PHM where these data are scarce.

While there are more techniques for exploiting a pre-trained model for a domain-specific task, fine-tuning and feature extraction are of particular interest to this thesis. As industrial data typically suffers from high sparsity, the method of using a foundation model for initial feature extraction followed by domain and task-specific adapter module training will be used in the rest of this thesis. A large language model will be fine-tuned on industrial text data in Chapter 4. Now that an overview of the current advancement in multimodal learning is established, the next section looks at the existing studies in data-driven PHM using multimodal data.

2.4 Data-driven PHM with Multimodal Data

In this section, we look at the existing studies in PHM that use multimodal data. This section is divided into subsections based on PHM purposes such as fault detection and diagnostics, prognostics, and maintenance optimization. To the best of our knowledge, the studies referenced in this section are all the existing works that use multimodal data to solve any PHM task.

2.4.1 Multimodal machine learning in fault detection and diagnostics

Fault detection and diagnostics (FDD) is typically performed after a fault has occurred in the system. Fault detection involves finding if an anomaly occurred, fault isolation aims

to identify where exactly the fault occurred, and diagnostics involves analyzing why it happened.

Table 2.3 presents the existing works that use multimodal data to solve fault detection and diagnostics. One can see that almost all studies investigate a combination of numerical and image data for multi-modal learning, and propose a deep learning model to fuse data for fault detection, isolation, and diagnostics.

Problem	Method/tool	Data	Application
Data fusion for fault diagnosis Yuan *et al.* (2018)	M-CNN	Vibration, IR images	Rotor system
Data fusion for fault diagnosis Yuan *et al.* (2018)	M-ResNet-DCA	Vibration, IR images	Rotor system
Network fault isolation Kao *et al.* (2019)	LSTM	Network metrics, customer complaints	IPTV network
Data fusion for fault diagnosis Ma *et al.* (2014)	RBM-AE	Electric signals, images	Power transformers and circuit breakers
Data fusion for diagnosis Zhou *et al.* (2021)	DNN, AE, CNN	Vibration, image of vibration signal	Bearing platform
Data fusion for fault detection Marei and Li (2021)	MLP, CNN, GRU	Temperature, Operation details	Plastic molding
Data fusion for fault detection Mian *et al.* (2022)	NCA, RA, SVM	Vibration, Thermal images	Bearings

Table 2.3: Research on multimodal learning for fault detection and diagnostics.

Reviewing the studies in Table 2.3, it is observed that the integration of numerical and image data has shown promise in enhancing fault detection and diagnostics. However, these works primarily focus on combining two types of data, leaving a gap in exploring more diverse multimodal data sources. This indicates the necessity for developing method-

ologies that can effectively fuse various data types to improve fault detection accuracy and robustness.

2.4.2 Multimodal machine learning in prognostics

Prognostics is the activity of projecting the health state of a machine or system into the future. This projection is used to anticipate failures and take proactive actions as needed. Prediction of the RUL of a machine is one of the key activities in prognostics. Table 2.4 presents the existing works that use multimodal data for prognostics. We observe that all deep learning-based works use CNN in combination with other architectures. It should also be mentioned that the study by Zhang *et al.* (2021) does not use deep learning methods, but instead explores visualization and clustering of maintenance data to support preventive maintenance.

Problem	Method/tool	Data	Application
Data fusion for RUL prediction Yang *et al.* (2021)	CNN, MLP	Simulated data: Inspection records, signal images, maintenance history	Steam generator
Data fusion for RUL prediction Marei and Li (2021)	CNN-LSTM, ResNet-28	Process parameters, tool images	Machining
Data visualization Zhang *et al.* (2021)	ccPCA, ccMCA, UMAP + DBSCAN	Network metrics, operation parameters, machine status description	Maintenance log analysis
Wear condition prognosis Wang *et al.* (2019b)	CNN, RNN	Process parameters, tool images	Cutting tool

Table 2.4: Research on multimodal learning for prognostics.

Table 2.4 highlights the use of multimodal data in prognostics, particularly in predicting the RUL of machinery. While CNN-based models show promise in RUL prediction, the reliance on simulated data highlights the challenge of applying these methods to real-world scenarios. There is a clear need to develop and validate models that can handle the complexity and variability of industrial multimodal data for more reliable prognostics.

2.4.3 Multimodal machine learning for maintenance optimization

Maintenance optimization includes activities related to decision support based on prognostic information resulting from previous steps in the PHM pipeline. This goes beyond studying the health state of the system and extends to specifying what actions can be taken for optimal maintenance. Table 2.5 presents the existing works that use multimodal data for prescriptive maintenance. Digital twins and associated methods appear to be the dominant methods when observing the table. The papers presented in Table 2.5 define frameworks for the maintenance process and specify suitable techniques for each part of the respective framework. However, it should be noted in advance that the papers do not present case studies with failure prediction mechanisms. On the contrary, the methods are only recommended as a potentially suitable solution for the prediction task in the framework presented in the papers.

Problem	Method/tool	Data	Application
Decision support framework design Ansari *et al.* (2019)	Digital shadow	Maintenance records, machine parameters	Cyber-physical production systems
Failure prediction for decision support Ansari *et al.* (2020)	Dynamic Bayesian Networks	Maintenance records, machine parameters	Cyber-physical production systems
Maintenance framework design Zacharaki *et al.* (2021)	Digital twin, rule-based model	(Only concept presented in the paper)	Refurbishment of industrial equipment

Table 2.5: Research on multimodal learning for maintenance optimization.

Examining the studies in Table 2.5, it is clear that frameworks like digital twin and dynamic Bayesian networks are pivotal in maintenance optimization. However, these frameworks often lack practical validation with real-world data. This gap suggests a significant opportunity to develop and validate comprehensive maintenance optimization strategies that incorporate multimodal data for better decision support.

To the best of our knowledge, no other works have addressed the challenges and explored the opportunities of using multimodal data for industrial diagnostics and prognos-

tics. In summary, the reviewed studies reveal a promising yet underexplored potential of multimodal data in PHM. The identified gaps underscore the need for developing robust methodologies that can effectively leverage diverse data sources. The next section consolidates the observations made in this chapter to position the objectives of this thesis in the literature.

Takeaway

Multimodal data adoption in industrial PHM is a very nascent field with little focus in the literature given to this so far. The few works present deal with a maximum of two modalities at once, and do not go into detail about overcoming the challenges of multimodal learning. Therefore, there is a significant lack of studies addressing the challenges of real industrial multimodal condition monitoring data.

2.5 Research Positioning

According to the literature review, it was observed that most of the datasets used in the data-driven PHM domain are unimodal sensor signals such as vibration, temperature, or other one-dimensional numerical data. In contrast, other domains such as medical science, robotics, and affective computing have shown significant adoption of multiple data modalities and applied deep learning methods, leveraging their interconnections to solve some domain challenges. Real-world applications indicate that neural network-based multimodal learning is now mature enough for the PHM community to explore as a solution to data scarcity in industrial condition monitoring. Recent advancements in large neural network models for feature extraction offer potential solutions to the fundamental challenges that have previously limited the field.

Multimodal data offers a promising solution to the limitations of unimodal data by providing complementary and reinforcing information. For instance, visual inspection photographs, operator reports, and other textual data can supplement traditional sensor data, offering a more comprehensive view of machine health. This is particularly important in industrial settings where sensor data may be scarce or incomplete. By leveraging multimodal data, we can enhance the robustness and accuracy of PHM models, ultimately improving maintenance decisions and reducing operational costs.

In this context, the thesis aims to tackle several critical challenges in the domain of PHM by exploring the potential of multimodal data integration and advanced ML techniques. The following research questions underscore the originality and novelty of this study:

1. *How can multimodal data improve the accuracy and robustness of PHM models compared to traditional unimodal approaches?*
   To explore this, the thesis will investigate the integration of various data modalities, such as sensor signals, visual inspection photographs, and textual operator reports. The goal is to develop and evaluate models that leverage these multimodal inputs to enhance the comprehensiveness and predictive performance of machine health assessments.

2. *What techniques can be developed to handle missing and noisy data in multimodal industrial condition monitoring datasets?*
   This question will be addressed by studying the impact of data quality issues on model performance and developing robust techniques to handle these issues. The focus will be on creating models that maintain high accuracy and reliability even when some data modalities are missing or corrupted.

3. *What methodologies can be developed to ensure the temporal alignment and joint representation of multimodal data in PHM tasks?*
   The research will focus on developing techniques for synchronizing and jointly representing data from different modalities, with an emphasis on temporal alignment methods that ensure coherent integration of time-series data with other types of information.

4. *How can domain knowledge be effectively incorporated into multimodal PHM models to enhance interpretability and acceptance by industry experts?*
   The research will develop methodologies that integrate domain expertise into the model design and training process, as well as leverage textual knowledge from industry experts. Ensuring that model outputs are interpretable and align with existing domain knowledge will be a key objective.

5. *What strategies can be employed to address the class imbalance in multimodal condition monitoring data, especially in the context of multi-fault diagnostics?*
   To tackle this, the thesis will investigate techniques such as advanced sampling methods, distribution aware training, and cost-sensitive learning. The aim is to develop models capable of accurately diagnosing multiple faults in highly imbalanced datasets.

6. *How can machine health prognostics be performed without extensive run-to-failure data for expensive and complex machines?*
   This challenge will be addressed by exploring alternative approaches to prognostics that do not rely on complete run-to-failure data. Potential solutions include transfer learning, few-shot learning, and the use of synthetic or simulated data to augment limited historical data.

7. *How can the results of multimodal PHM models be validated on real industrial datasets to ensure their generalizability and reproducibility?*
   This question will be addressed by applying the developed methodologies to real industrial datasets and validating the results with domain experts. Ensuring that the models are generalizable to different industrial settings and reproducible across various datasets will be a crucial part of this research.

In summary, this thesis positions itself at the intersection of multimodal learning and industrial PHM, aiming to bridge the gap between academic research and practical industrial applications.

The scientific development plan to address each of the aforementioned challenges is given below, and the structural outline of the thesis is also illustrated in Figure 2.7.

The first contribution (Chapter 3) presents an initial exploration of multimodal data using a simulation dataset. We develop a **methodology to analyze the impact of missing and noisy data using a simulation dataset. A crossmodal attention-based co-learning technique** is presented to increase resilience to these conditions.

The second contribution (Chapter 4) focuses on **development of a new methodology for fault detection and health index calculation under sparse multimodal data.** This builds upon the first contribution and applies it to a real industrial dataset of hydrogenerators. Data modalities including image, text, electrical signals, and other time series will be processed. **Time alignment issue due to different condition monitoring rates of multimodal data will be addressed.** A methodology will be developed to use expert knowledge-assisted feature extraction for mitigating data shortage within each modality. This methodology will also highlight fidelity to domain knowledge without demanding physics-of-failure modeling, as well as **account for human expert subjectivity in health estimation through industrial text mining.**

Building upon the techniques and methodologies developed through the course of the thesis, the final chapter (Chapter 5) presents an **end-to-end methodology for prognostics using multimodal condition monitoring data without necessitating run-to-failure (RTF) trajectories from the history of any machine.** This will be accomplished in a two-step methodology. Firstly, we will develop a **modular deep learning framework to perform diagnostics under significant class imbalance and sparsity**, while also managing resource constraints. Secondly, we will use the features from this diagnostics model to **construct graph-format RTF data and train a prognostics model**. The end-to-end prognostics methodology presented in this chapter is applied to an industrial dataset of hydrogenerators and validated by domain experts. This final contribution represents the culmination of all the techniques and methodologies developed throughout this thesis, resulting in a prognostic methodology that effectively leverages

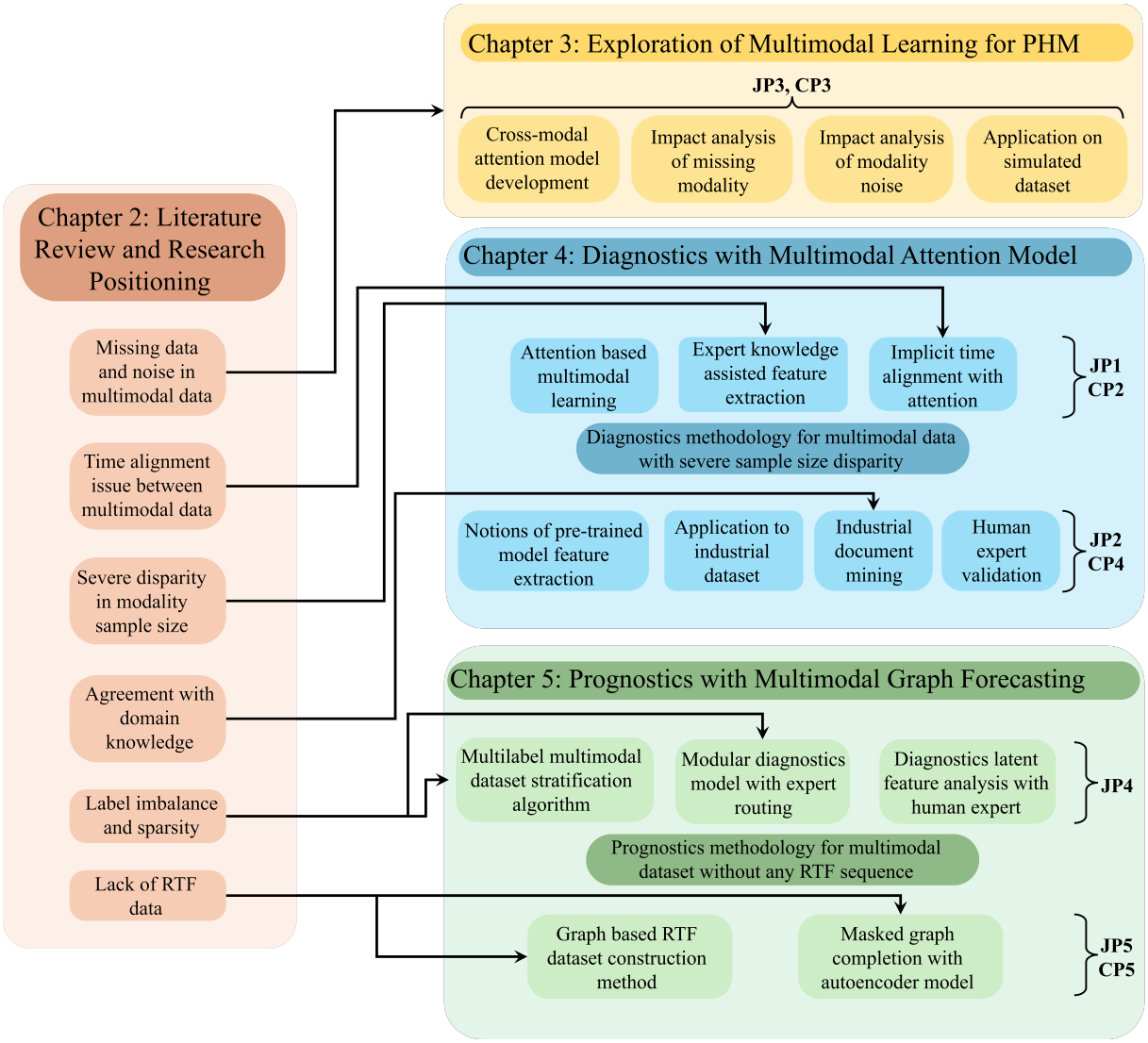


Figure 2.7: Schematic outline of thesis plan in relation to literature gaps identified in this chapter.

multimodal learning to address the various challenges arising from industrial data scarcity.

2.6 Conclusion

In this chapter, an overview of data-driven PHM was presented. We delved into the concept of multimodal data, its history, and the significant challenges associated with it. This was followed by an exploration of the tools and latest advancements used to process and analyze multimodal data.

We then reviewed existing works in the PHM domain that utilize multimodal data, highlighting the progress made and the limitations that still exist. This examination allowed us to identify critical gaps in the literature, which serve as the research motivation and direction for this thesis.

In the next chapter, we will begin addressing these research gaps. We will develop an initial methodology for working with multimodal data using a simulated dataset. This methodology will serve as a foundation throughout the thesis. The challenges of missing modality and modality-specific noise will be investigated. Also, we will develop and propose techniques to mitigate these issues.

Chapter 3

Exploring Multimodal Learning in Industrial PHM

Contents

"The concept of "bad data" and poor data quality has been around nearly as long as humans have existed, albeit in different forms. With Captain Robert Falcon Scott and other early Antarctic explorers, poor data quality (or rather, data-uninformed decision making) led them to inaccurately forecast where and how long it would take to get to the South Pole, their target destination."

— Data Quality Fundamentals by Barr Moses, Lior Gavish, and Molly Vorwerck (Moses *et al.* (2022)).

3.1 Introduction

In Chapter 2, we presented an overview of multimodal data applications and challenges in data-driven PHM. By combining information from multiple modalities, a more comprehensive overview of the system's health can be obtained, enabling the detection of subtle changes in its behavior that might not be observable from a single modality. Nonetheless, the efficient exploitation of multimodal data is encumbered by significant challenges, particularly in managing missing data and modality-specific noise, aspects still underexplored within the PHM domain.

Research across various domains has proposed strategies for tackling these challenges. For instance, Lee *et al.* (2019) investigated the impact of missing data in semiconductor manufacturing, while Liu *et al.* (2021) and Ma *et al.* (2021) studied the effects of noise and missing modalities in emotion recognition and multimodal learning, respectively. Other studies, such as those by Nagulapati *et al.* (2021) and Akrim *et al.* (2023), have focused on how training data size affects prediction accuracy. One can see that while the literature extensively investigates missing data effects in various contexts, there is a notable absence of studies addressing different levels of missing data modalities within the PHM domain. Similarly, despite the consideration of noise conditions in some studies, no exploration of diverse noise levels in a multimodal PHM dataset has been identified.

This chapter aims to bridge this gap by examining the integration of diverse multimodal data sources, including maintenance records, imagery, and technical reports. We utilize a simulated multimodal dataset on steam generator prognostics, which includes numerical, textual, and image data. Our exploration focuses on the effectiveness of cross-modal transformer attention layers in handling missing modalities and noise, and how these factors influence feature learning and model performance.

This initial exploration lays the groundwork for developing robust multimodal learning frameworks tailored for industrial applications, where data quality and completeness are paramount. The insights gained here will be crucial for future models throughout this thesis, providing foundational knowledge on the key challenges and potential strategies in multimodal PHM. A graphical overview of the chapter's structure and methodologies is presented in Figure 3.1.

The rest of this chapter is organized as follows. In section 3.2, an attention based multimodal learning method is presented. Section 3.3 presents the dataset and the simulation of missing and noisy data conditions, setting the parameters for the analyses. Section 3.4 introduces the multimodal neural network structure and the experiments analysing performance variation under a range of simulated data conditions. Finally, section 3.5 summarizes and concludes the chapter.

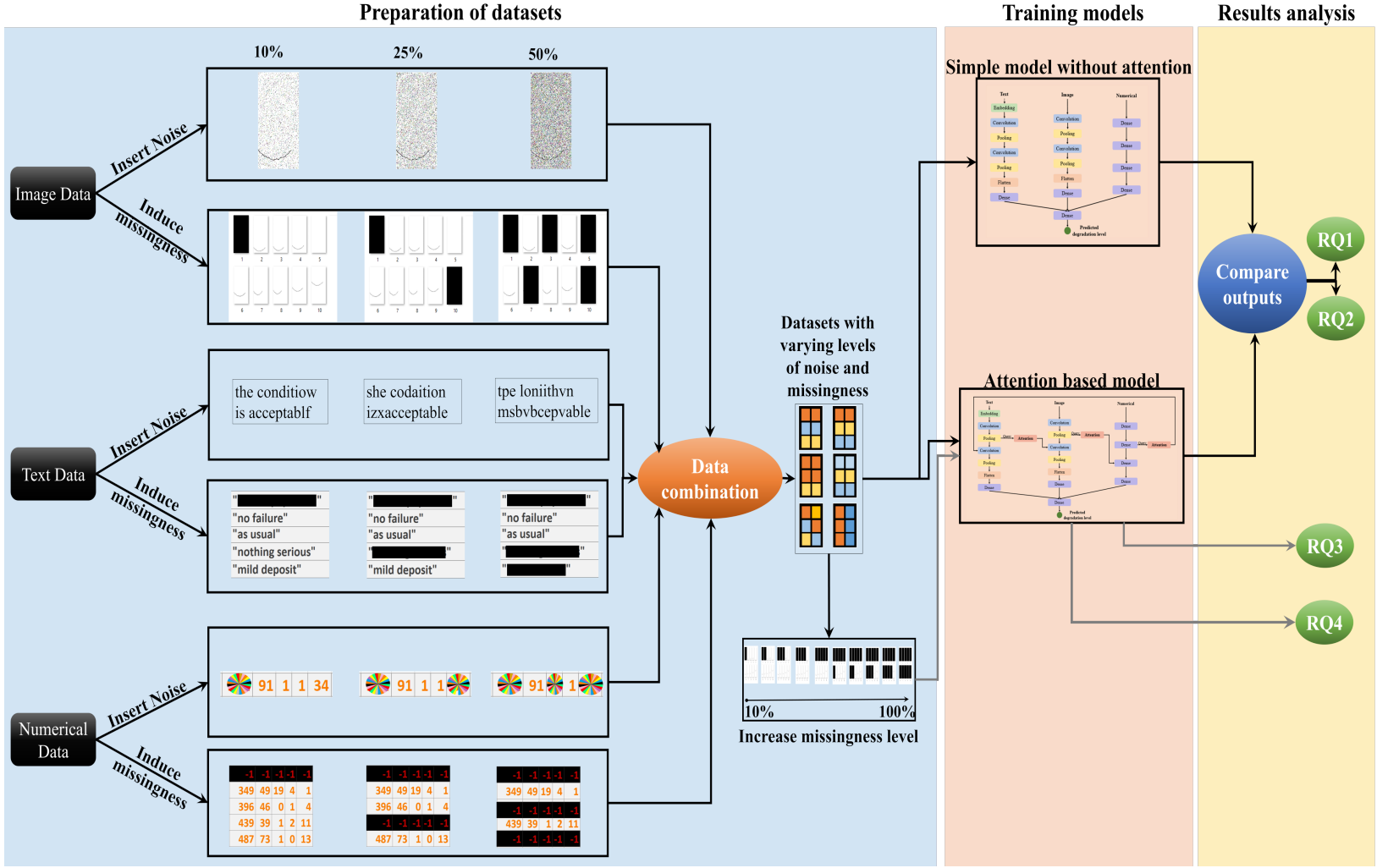


Figure 3.1: Schematic overview of Chapter 3. The investigation is conducted in three stages: The first stage is the insertion of noise or induction of missingness from the three data modalities at various levels. Then these are combined in different combinations to create a set of datasets. The second stage involves training the two deep-learning models on all of the datasets. The final stage involves comparing the outputs of the models on each dataset and studying the results to conclude.

3.2 Cross-modal Context Passing with Attention

In this chapter, we aim to establish a general principle for designing multimodal neural network architectures that can serve for all the industrial datasets we will work with in later chapters. From an engineering perspective, the key challenge of multimodal data are the disparity between dimensions. From a deep-learning point of view, the challenge is extracting features from diverse kinds of data. Considering that realistic data conditions in the industry would be far from dense and balanced, a multimodal learning model for PHM needs to be robust to low data quality and exploit maximum information from sparse data. To arrive at such a model, we need to first build a feature extraction pipeline for each of the data, and then implement a multimodal interaction mechanism that can compensate for the data limitations. The first part is straightforward, simply involving designing a network with layer mechanisms suited to treat each modality, e.g., convolutional layers for images.

Building a multimodal network in which interactions between modalities can compensate for limitations within individual modalities is, by far, a more challenging problem. The transformer attention mechanism (Vaswani *et al.* (2017)) has demonstrated strong performance in various tasks involving multimodal data. It enables the model to selectively focus on relevant features across different modalities by computing attention weights based on the relationships between keys, queries, and values.

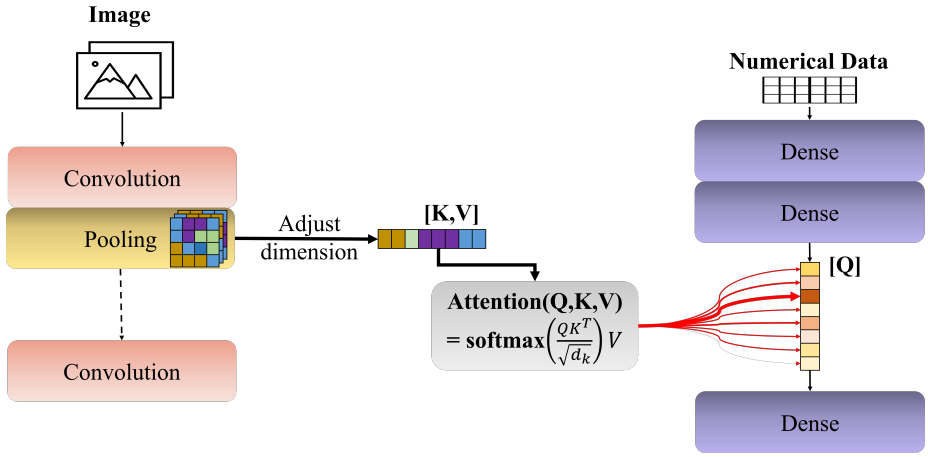


Figure 3.2: A simplified illustration of a crossmodal attention layer from image branch to numerical branch. The query (Q) comes from the numerical features looking to be enhanced by additional context. Key (K) and value (V) are derived from the image features. The attention score determines how much each image feature (value) should contribute to the final output that goes into the numerical processing stream.

Here, we explore how transformer attention can enhance cross-modal information trans-

fer, specifically by improving feature extraction from each modality using insights gained from others. For instance, considering an attention layer from image to numerical data, the mechanism operates by assigning weights to features rather than modifying the features themselves. By leveraging contextual information from the images, the attention mechanism enhances the representation of the numerical data based on the computed attention weights.

Definition

Definition 3.1 (Attention)**:**
The attention layer takes in three inputs - the key K, the query Q, and the value V. Query represents the data or features seeking contextual enhancement. The key is used to compute compatibility or relevance scores against the query. The value contains the data that is aggregated based on these scores to enrich or augment the query.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The attention score estimates the pertinence of each key component in relation to the query.

In the context of image-to-numerical attention, image features would act as both the key and the value. The query, representing numerical data, seeks contextual enhancement from the image features. The attention layer calculates a similarity score between the numerical data's hidden state (query) and the sequence of image features (key). This score denotes the significance of each numerical data component in relation to the image. These attention weights are utilized to compute the final weighted representation of the numerical features, illustrated in Figure 3.2.

Implementing crossmodal attention in multimodal learning models faces the significant challenge of dimension mismatches among different modalities. This mismatch complicates matrix multiplication as the key dimensions must align with those of the value vector. To overcome this, several strategies have been developed.

- Simplification with 1D convolution filters: This method replaces the traditional key-query-value mechanism but often fails to capture complex intermodal relationships.
- Reshaping data blocks: Matching the dimensions of one modality's data block with another preserves spatial information but may lose some misaligned features.
- Flattening and dense layer integration: Flattening output from a modal block and using a dense layer to match dimensions across modalities sometimes results in spatial information loss but effectively aligns the dimensions for necessary computations.

Each approach has its merits and drawbacks, depending on the specific characteristics of the dataset. Our research found that integrating flattened outputs with a dense layer effectively manages dimensional discrepancies. With this cross-modal context passing technique established, we will now test the multimodal model on a specific dataset and evaluate its performance across different data quality conditions.

3.3 Research Questions

Multimodal condition monitoring inherently involves data sparsity due to varying data collection frequencies across modalities. In practical multimodal PHM systems, missing data is a common challenge, further compounded by the diversity of sensors — including human input for textual data — which introduces varying noise levels. This chapter examines the effects of different noise intensities and data absence on multimodal learning. A primary focus is on assessing the efficacy of transformer attention mechanisms in enhancing feature extraction across modalities and their resilience under suboptimal data conditions. The study addresses the following research questions to deepen our understanding of how multimodal learning models perform under the diverse data quality conditions typically encountered in industrial settings:

1. How do varying levels of missing and noisy data in each modality (text, image, and numerical) impact the performance of multimodal learning models for industrial prognostics?

2. Can the incorporation of crossmodal attention mechanisms improve the robustness and performance of multimodal learning models in case of missing and noisy data?

While we observe the performance of a model under various data conditions, managing these with imputation or noise correction techniques is not attempted in this chapter. Rather, the objective is to study the ability of cross-modal attention layers to learn from the data as it is. Next, the case study dataset will be presented in section 3.3.1. Then, simulation of poor data conditions will be outlined in section 3.3.2. Design of datasets by combination of different levels of data missingness and noise will be presented in section 3.3.3 and dataset preparation by splitting into train and test subsets will be shown in section 3.3.4.

3.3.1 Steam generator dataset

To address the aforementioned research questions, a suitable multimodal dataset is essential. Given the exploratory nature of this study on multimodal data for PHM, we have opted for a simulated dataset specifically designed for steam generator prognostics, as detailed in Yang *et al.* (2021). This dataset provides a controlled environment to investigate the dynamics of multimodal learning within an industrial context.

The dataset comprises 50 degradation trajectories from 50 steam generators and includes both perfect and imperfect maintenance interventions. Each trajectory comprises approximately 150 observations, consisting of image, text, and numerical data as inputs, with machine degradation level as the target (see Fig. 3.3b).

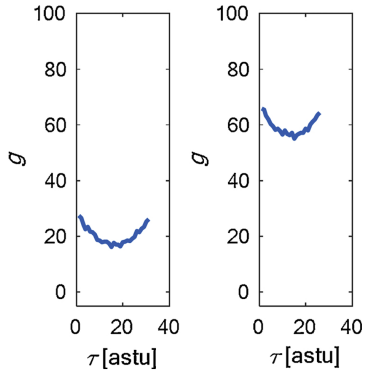
- Image data: Snapshots of wide range level (WRL) signals from the steam generators (not actual camera images). These are indicators of the level of sediment deposited in the steam generator, which corresponds to the degradation level. Figure 3.3a (left) shows a smaller difference between pressure levels at the top and bottom through a time interval, indicating less deposit and thus, less degradation than 3.3a (right).
- Text data: Brief notes by maintenance technicians, such as *"Middle level condition. The mechanical cleaning is done. Now SG has little deposits"*.
- Numerical data: Time, time from the last maintenance, number of mechanical cleanings, and others.
- Prediction target: An arbitrary degradation unit in the range 0 to 100 (Figure 3.3b).

For more detailed information on the dataset, interested readers can refer to the paper Yang *et al.* (2021).

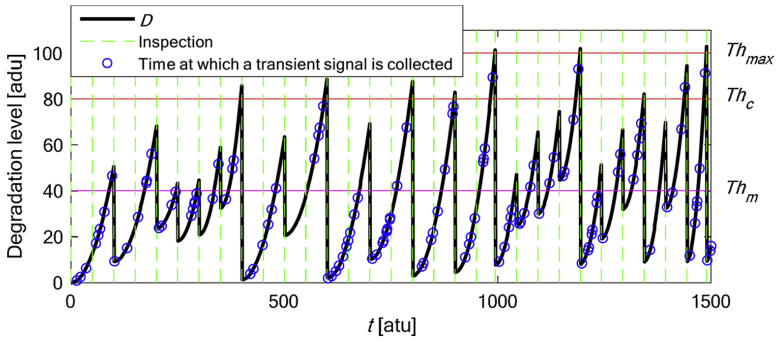
Despite being a simulated dataset, it is appropriate for this initial exploratory study as it offers a controlled experimental environment that is essential for systematically examining the effects of noise and missing data. This controlled setting is vital for comprehensively understanding how multimodal learning techniques perform under various conditions. Moreover, the simulated dataset enables the investigation of a broad spectrum of scenarios and data attributes that are often not accessible in real industrial data.

3.3.2 Simulation of noisy and missing data conditions

Within the existing literature, various types of noise have been examined for each modality. Given the specific research questions addressed in this chapter, and without generality loss,



(a) Samples of image data (adapted from Yang *et al.* (2021)). This shows the wide range level (WRL) signal, indicating sediment deposit.



(b) Machine degradation trajectory (adapted from Yang *et al.* (2021)). This figure shows one steam generator's simulated health state evolution. The black line tracks the degradation level (from 0 to 100). The degradation increases with time until maintenance interventions restore the machine to a completely or partially healthy state.

Figure 3.3: Illustration of image data and prediction target in the dataset

uniform random noise is investigated for images, character-random swaps for text, and random changes within the value range for numerical data. The simulation of missingness involves the complete removal of samples. Particularly, to create the noisy and missing datasets, the following steps are undertaken:

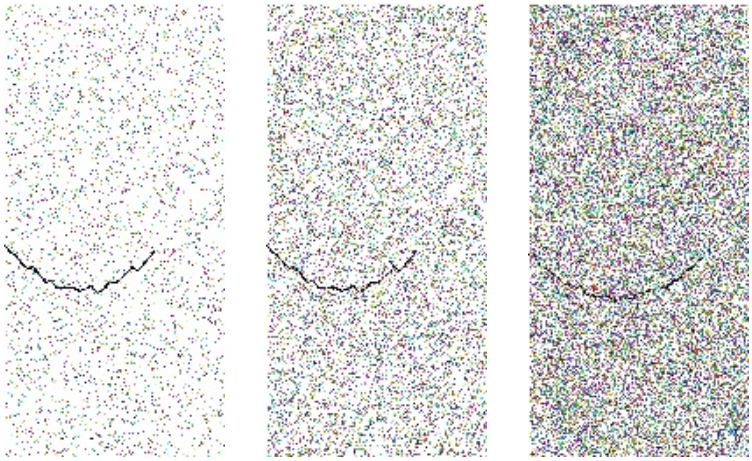
1. Adding noise: While adding noise at a certain percentage (0%, 10%, 25%, and 50%), every sample in the dataset was altered to match that noise level.

    - Images: Noise is added to the images by randomly changing pixel values (Figure 3.4). For example, to simulate a 50% noise level, 50% of the pixels in every image in the dataset would be randomly altered.
    - Text: For text data, noise is introduced by randomizing a percentage of text characters in every row. For instance, if the noise level is set at 10%, 10% of the characters in each text sample would be randomized. At a 25% noise level, the text becomes nearly unreadable for humans.
    - Numerical: For numerical data, a percentage of the numerical channels (of the five in the dataset) will be randomized. For example, at a 10% noise level, 10% of the numerical channels would have randomized values within the range of its possible values.

2. Simulating missingness: While simulating missingness, the corresponding percentage (0%, 10%, 25%, and 50%) of samples are entirely removed from the dataset.

- Images: If the image missingness level is set at 25%, then randomly selected 25% of the images in the dataset would be removed.
- Text: If the text missingness level is set at 25%, then all characters in a random selection of 25% of samples would be replaced by white spaces.
- Numerical: If the numerical data missingness level is set at 25%, then 25% of the rows would have their numerical values set to -1 (-1 is not a valid value in this dataset).

All randomization is done by the Python pseudorandom generator based on the Mersenne Twister (Matsumoto and Nishimura (1998)).



(a) Image data sample with 10% noisy pixels.

(b) Image data sample with 25% noisy pixels.

(c) Image data sample with 50% noisy pixels.

Figure 3.4: Illustration of simulating noise in image data.

3.3.3 Noise and missingness combinations

To systematically evaluate the performance of the proposed multimodal learning models under various data quality conditions, a range of datasets with different combinations of noise and missingness levels for each modality were generated. Let D represent a dataset, where D can be defined as:

$$D = \{M_i \times N_j \mid i, j \in \{1, 2, \ldots, n\}, n \text{ is the number of modalities}\}$$

In this study, $n = 3$, represents the three modalities: image, text, and numerical data. For each modality, the noise and missingness levels are independently considered. Let

$P = \{p_1, p_2, \ldots, p_k\}$ be the set of level configurations for both noise and missingness. In this case, $P = \{0, 10, 25, 50\}$ with a cardinality $|\mathrm{P}| = 4$. Then, M_i and N_j are defined as follows:

$$M_i = \{m \mid m \in P\} \text{ represents the missingness levels for modality } i.$$

$$N_j = \{n \mid n \in P\} \text{ represents the noise levels for modality } j.$$

By considering noise and missingness levels of each modality independently, the study explores the impact of $|P|^2$ unique combinations for each modality. In the current study, $|P|^2 = 16$, which represents the combinations spanning from no noise or missingness (0% for both) to the highest level of noise and missingness (50% for both). This results in a total of $(|P|^2)^n = 4096$ distinct datasets, comprehensively covering the possible scenarios of data quality issues across the modalities.

3.3.4 Training, validation and test data

To address the challenge of limited run-to-failure trajectories in industrial settings, we only use 25 trajectories for model training and the remaining 25 are set aside for testing. The initial 25 trajectories are divided into 5 folds for 5-fold cross-validation training.

In the coming sections, models will be trained under a range of missing and noisy data conditions. Each of these will be tested on two variations of the test set. The first is a "good quality" test set, where no missing or noisy conditions are simulated. The second is a "poor quality" test set, where missingness and noise are simulated at the same levels as the corresponding training set. This enables two distinct observations of a model. First, the results on the good-quality test set demonstrate how the model performs when presented with high-quality samples after being trained on poor-quality data. The predictions on the poor quality set, on the other hand, indicate the model performance under the same data quality conditions it was trained on.

All the steps until this point are illustrated in the first part of the overall schema shown in Figure 3.1. The next subsection will present the model development and training steps.

3.4 Multimodal Learning with Cross-modal Attention

The datasets created so far represent a wide range of poor data conditions across different modalities in a multimodal prognostics dataset. In this subsection, we will first establish the neural network structures to be trained on these datasets, and then compare their

performances. Subsections 3.4.1 and 3.4.2 will outline the neural network structures and establish the performance baselines. Then, subsections 3.4.3 and 3.4.4 will discuss the comparative analyses of the models under missing and noisy dataset conditions.

3.4.1 Unimodal model design

Initially, we preprocess each data modality in the dataset by developing unimodal neural networks. This involves constructing three distinct neural network modules, specifically tailored for processing image, text, and numerical data respectively.

For images, the unnecessary x and y-axis markers are cropped to keep only the useful information including WRL curves. These images are then processed with a simple convolutional neural network (Albawi *et al.* (2017)) consisting of convolutional and pooling layers. Convolutional networks are well suited to image tasks because of their capacity to capture local features and spatial relationships in the data. The output is a 1x1 dense node that predicts the degradation level based only on the image data. The architecture is shown in Figure 3.5a.

The text data first undergoes some preprocessing such as lowercasing, tokenizing, and padding. Then, an embedding layer is used to create a vector representation of the preprocessing text. This is followed by a convolutional network similar to the image path (see Figure 3.5b). In fact, convolutional networks have proven effective in capturing local and global semantic information in the text. Specifically, three 1D convolution layers were used for the text path.

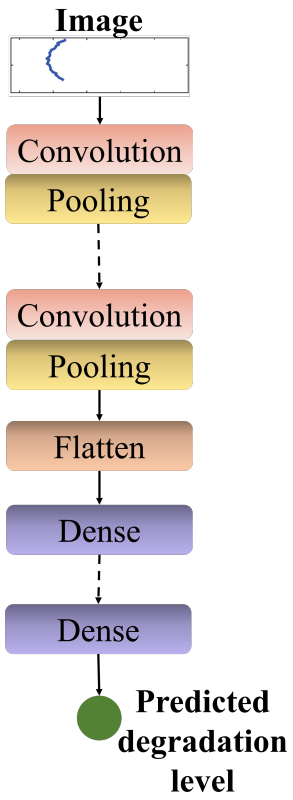
The numerical data is processed by a fully connected network with a series of dense layers (Figure 3.5c), suitable for processing structured numerical data due to its capacity to learn nonlinear patterns.

The results of the unimodal models trained on 25 trajectories are given in Table 3.1. One can see that among the three modalities, unimodal learning based on numerical data provides the best results.

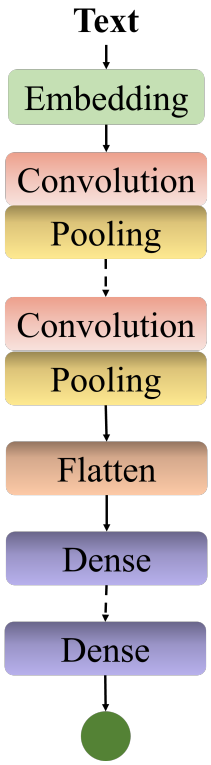
Table 3.1: Performances of unimodal learning.

Data/Model	MAE	MSE
Image	14.59	355.02
Text	18.56	635.10
Numerical	12.83	176.66

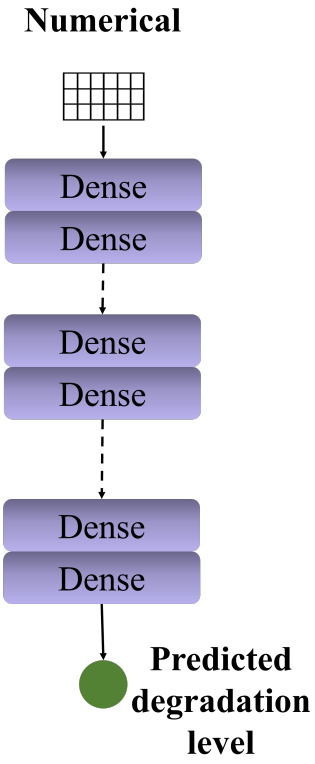
Now that the unimodal baselines are established, we can use these as building blocks



(a) CNN architecture to process image data.



(b) CNN architecture to process text data.



(c) Fully connected network to process numerical data.

Figure 3.5: Unimodal network structures that form the branches of the multimodal architecture.

to build bimodal architectures, and finally construct the optimal multimodal learning architecture for three data modalities.

3.4.2 Multimodal architecture design

The previously mentioned unimodal structures — for image, text, and numerical data — serve as foundational components for constructing the multimodal baseline model. A straightforward approach involves merging the outputs of these three unimodal modules at their final layer, a method exemplified in Yang *et al.* (2021).

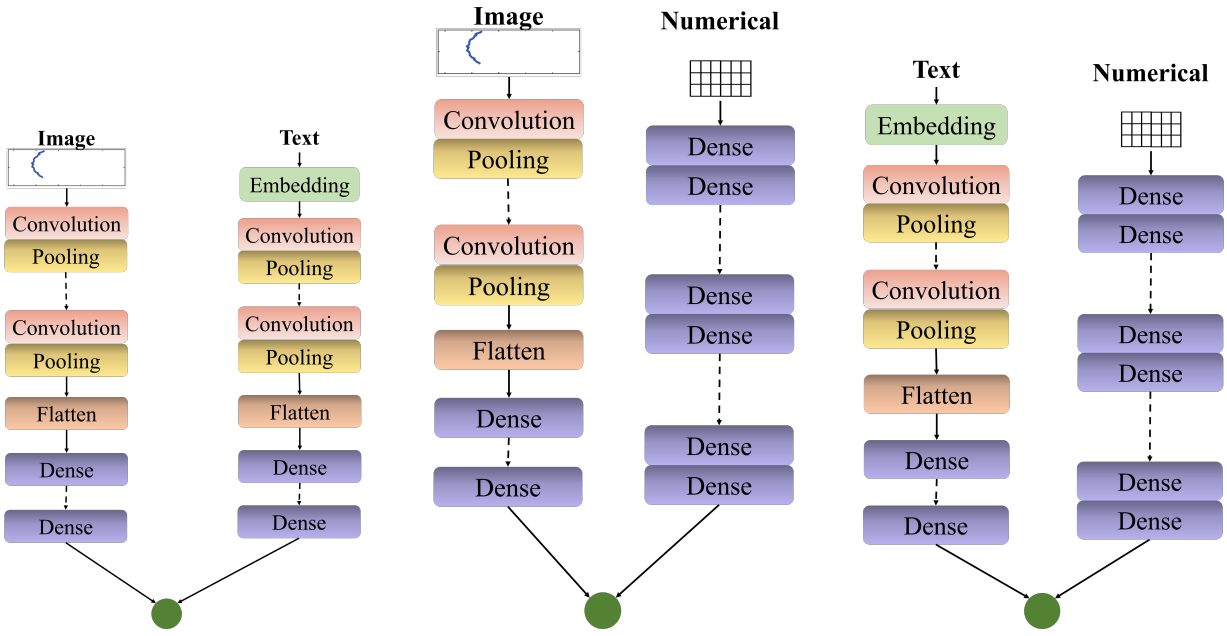
Initially, we integrate pairs of modalities to develop three bimodal models, as depicted in Figure 3.6. Subsequently, all three modalities are combined to form a straightforward trimodal model, illustrated in Figure 3.7. This architecture, which primarily utilizes late

fusion of the unimodal networks, is called the "simple model". The performance evaluation for these models is presented in Table 3.2.

One can see that multimodal learning utilizing three modalities yields the most favorable outcomes. However, the models in this study were trained solely on 25 trajectories, whereas Yang *et al.* (2021) trained their models using 40 trajectories. Consequently, a direct comparison between the results of this study and those of Yang et al. (2021) is not feasible due to the disparity in the training dataset size. However, the simple trimodal model presented in Table 3.2 shares the same architecture as described in Yang *et al.* (2021). Consequently, while a direct comparison with the findings from Yang *et al.* (2021) is not feasible, all subsequent model comparisons in this study will juxtapose the architecture suggested in the referenced paper with a crossmodal attention network.

Table 3.2: Performance of simple models (without attention mechanism).

Data/Model	MAE	MSE
Image + Text	14.04	197.47
Image + Numerical	12.69	169.28
Text + Numerical	17.91	494.21
Image + Text + Numerical	11.36	179.87



(a) Text and image (b) Image and numerical (c) Text and numerical

Figure 3.6: Simple 2-modal network structures without attention mechanism.

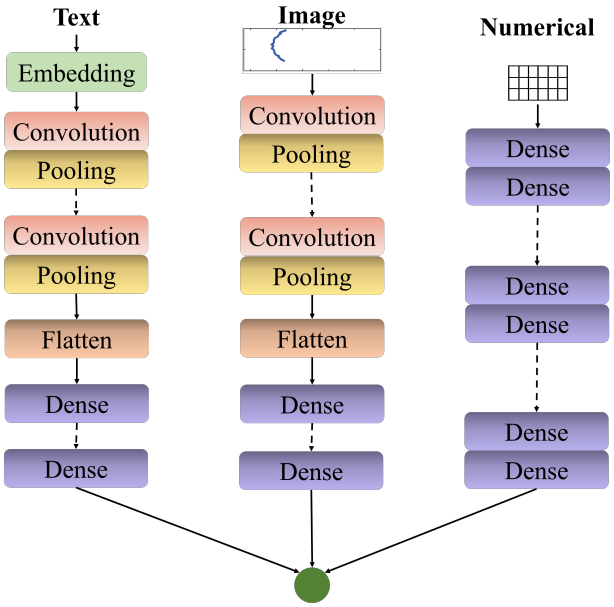


Figure 3.7: Simple 3-modal network without attention mechanism. The general model design until this step follows the original dataset paper by Yang *et al.* (2021).

Building on the previous discussion of bimodal networks, implementing crossmodal attention-based information transfer is critical. A pivotal aspect of this implementation involves configuring the attention layers appropriately. In this context, we investigate various configurations, specifically testing alternating attention directions between modalities. The results of these experiments are presented in Table 3.3.

Table 3.3: Performances of different attention-based models with 2 modalities.

Data	Attention	MAE	MSE
Image + Text	Text to image	12.77	194.95
Image + Text	Image to text	13.67	198.17
Image + Numerical	Image to numerical	10.89	128.71
Image + Numerical	Numerical to image	11.64	134.85
Text + Numerical	Numerical to text	15.40	299.74
Text + Numerical	Text to numerical	17.31	381.41

The results in Table 3.3 provide a starting point for the attention configuration for the 3-modal setup. The critical challenge in designing a neural network with cross-modal attention layers is the placement of the attention layer. The naive but expensive design choice is to attend in both directions. While this can seem to be exhaustive, bidirectional attention is not always the best case, because some data may be unsuited to provide context for other data. Indeed, determining which data modality contains suitable features

to inform the learning from another is a crucial step. Ideally, this can be deduced from domain knowledge of the degradation modes and the condition monitoring data.

Given the lack of established domain knowledge to prioritize information quality across modalities, the performance of 2-modal attention configurations can reveal which data modality enhances feature extraction from others. For instance, the results shown in Table 3.3 indicate that in this dataset, text data more effectively guides the feature extraction process for images than vice versa.

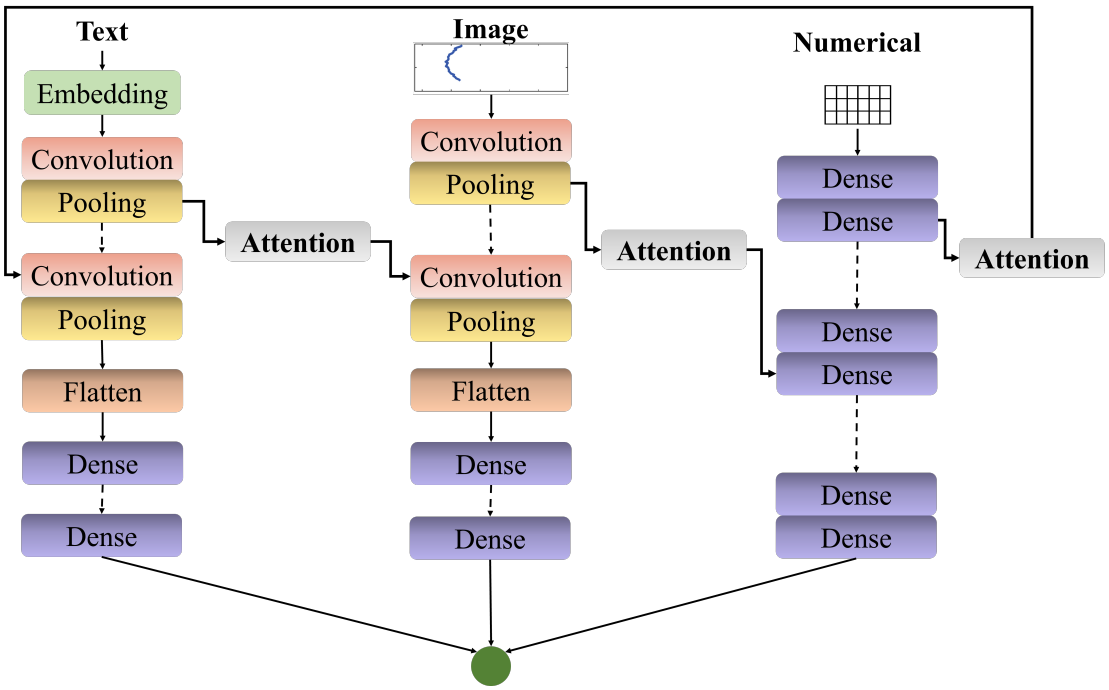


Figure 3.8: Attention model on text, image, and numerical data. This is the new model proposed in this work. The crossmodal attention layers are implemented with transformer attention (Vaswani *et al.* (2017)).

Expanding upon these insights, we proceed to evaluate configurations of attention layers within a 3-modal framework. Here, we explore four distinct cross-modal attention combinations to assess their efficacy:

1. Attention from text to image, from image to numerical, from numerical to text;
2. Attention from image to text, text to numerical, numerical to image;
3. Attention from image to text, and from image to numerical;
4. Attention in both directions between all three modalities.

Table 3.4: Performance of different attention-based models with 3 modalities.

Attention order	MAE	MSE
Text to image, image to num, num to text	6.91	82.76
Text to num, num to image, image to text	18.64	701.62
Image to text, image to num	15.08	473.19
Bidirectional attention between all modalities	20.12	764.67

The results of the multimodal learning models using the four different combinations are reported in Table 3.4. It is observed that the configuration employing the attention mechanism from text to image, image to numerical, and numerical to text yields the best performance. This is consistent with the configurations found in the 2-modal tests done in Table 3.3. The resulting 3-modal attention configuration, depicted in Figure 3.8, will be further examined throughout the remainder of this study as the representative architecture for multimodal learning incorporating attention, in comparison to the baseline multimodal model without attention.

Key findings

Cross-modal attention layer placement should flow from the more informative modality to a less informative one. No universally optimal order applies to all datasets. In an industrial setting, this order may be known to the domain experts. If not, unimodal models and performance comparisons of bimodal models with different attention configurations can establish this order. If the experiments do not conclusively determine a natural hierarchy among the modalities, implementing a bidirectional attention mechanism, though potentially costly, represents a prudent fallback strategy. This approach ensures comprehensive data integration, maximizing the learning potential across all modalities.

3.4.3 Investigation of multimodal learning performance in missing data context

This section begins the comparative evaluation of the simple model (Figure 3.7) and the attention model (Figure 3.8) on two test sets. The "good test data" is the test set without any simulated poor data quality conditions, whereas the other test set simulates the same conditions as the training set of the corresponding model.

Table 3.5 presents a comparison between the attention-based and simple models, trained

Table 3.5: Comparison of attention model and simple model when missing data.

Missing level in training data	**MAE on good test data**		**MAE on test data with missing**	
	Attention	Simple	Attention	Simple
Image 10%	8.69	12.02	10.8	27.05
Image 25%	10.81	13.2	11.57	28.28
Image 50%	11.53	14.54	13.24	29.65
Text 10%	10.59	12.74	11.38	22.56
Text 25%	10.37	12.88	10.84	23.6
Text 50%	10.01	13.53	10.95	24.93
Numerical 10%	9.36	11.25	11.02	21.8
Numerical 25%	9.5	12.81	11.88	23
Numerical 50%	10.33	15.72	12.79	28.1
Image 10%, Text 10%, Numerical 10%	9.96	14.24	13.44	32.98
Image 25%, Text 25%, Numerical 25%	10.12	19.1	14.62	33.06
Image 50%, Text 50%, Numerical 50%	10.42	32.68	17.41	37.07

under varying conditions of data missingness (10%, 25%, and 50% of the training data). Particularly, it displays the mean absolute error (MAE) values of both models on a good-quality test set without any missing data, as well as on a low-quality test set where data missingness is simulated at the same levels as the training set.

Furthermore, Figures 3.9, 3.10, 3.11, and 3.12 illustrate the missingness levels ranging from 0% to 100% for each modality (image, text, and numerical data) within the training set. Each figure consists of four line charts representing the performance (MAE value) of the attention-based model on the good-quality test set, the attention-based model on the low-quality test set, the simple model on the good-quality test set, and the simple model on the low-quality test set.

A cursory examination reveals that the attention-based model consistently outperforms the simple model across all figures. Additionally, within each figure, the performance of the attention-based model at 0% missing data aligns with an MAE value of 6.91 (as indicated in Table 3.4), while the performance of the simple model at 0% missing data corresponds to an MAE value of 11.36 (as indicated in Table 3.2).

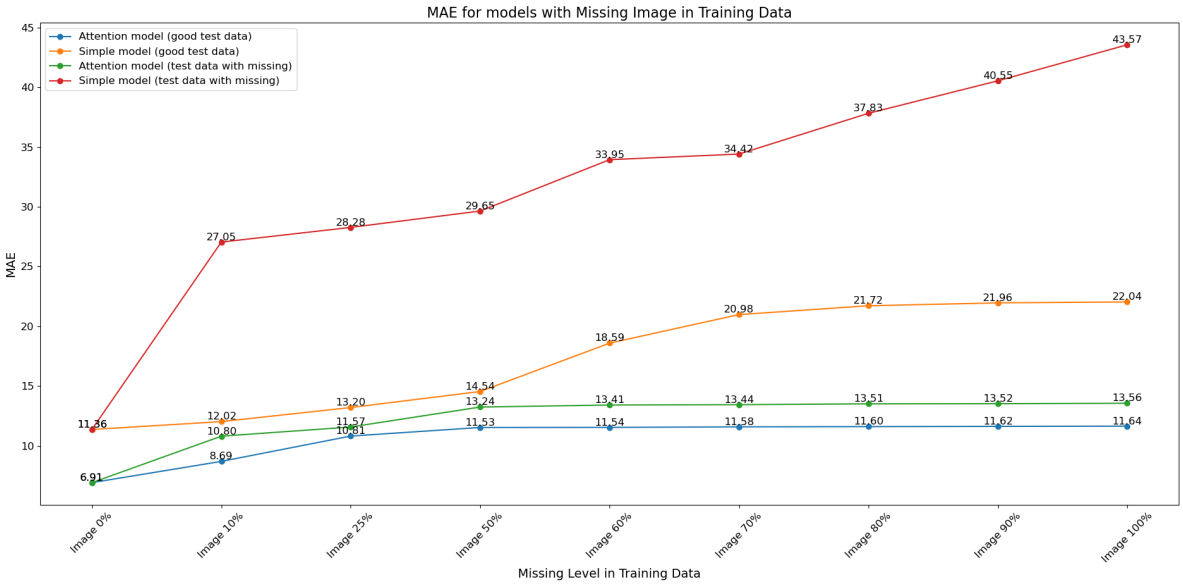


Figure 3.9: Comparison of attention and simple models trained on dataset with missing image. Each point in the figure represents a model trained on a dataset with a different percentage of missing images.

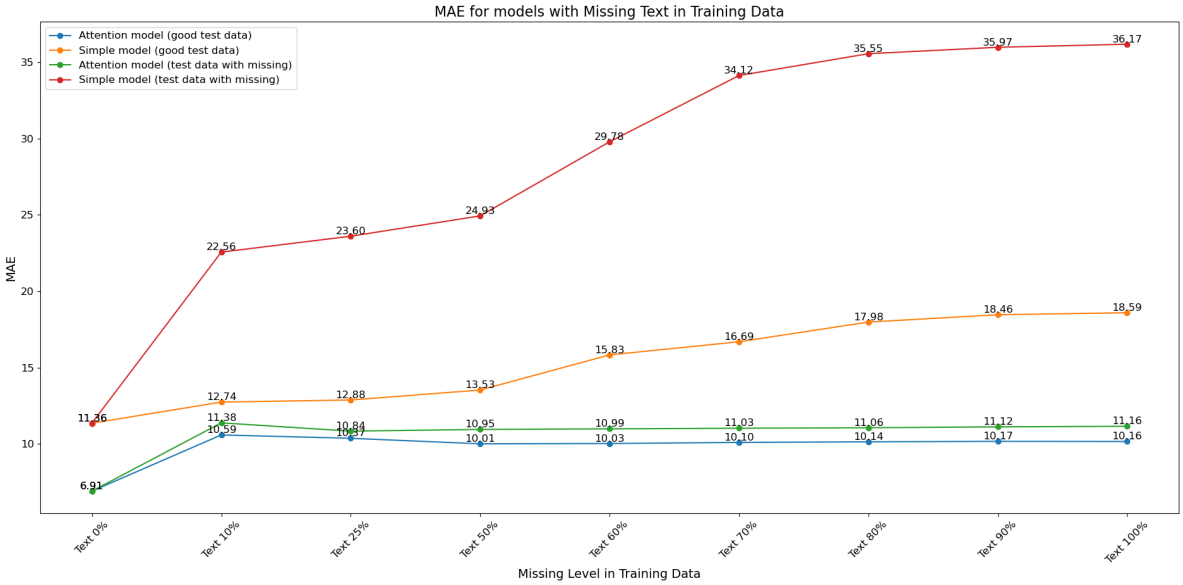


Figure 3.10: Comparison of attention and simple models trained on the dataset with missing text.

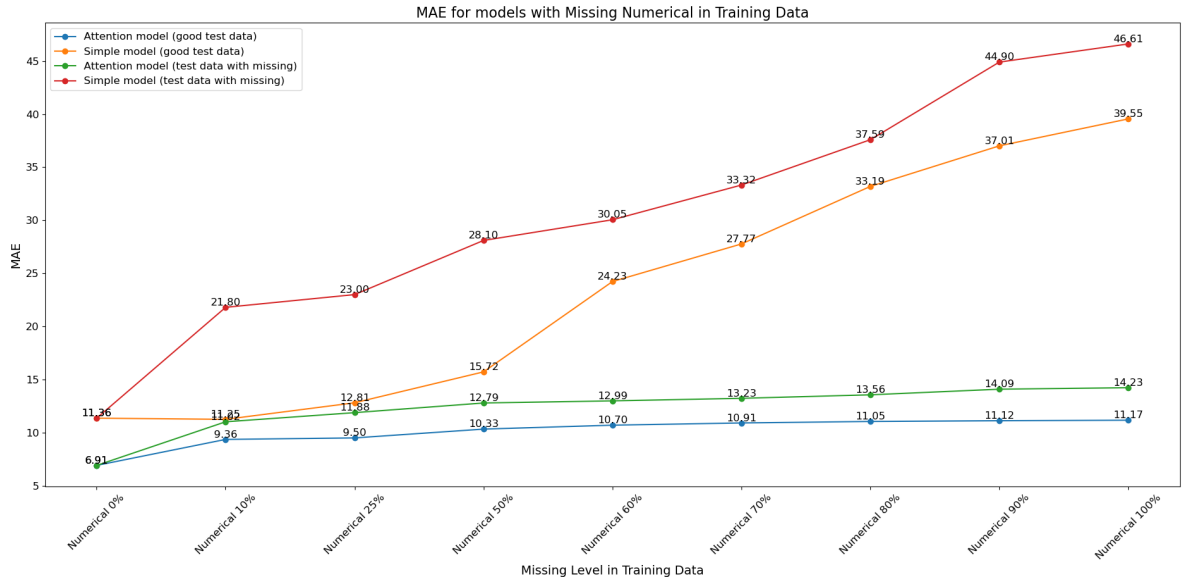


Figure 3.11: Comparison of attention and simple models trained on the dataset with missing numerical data.

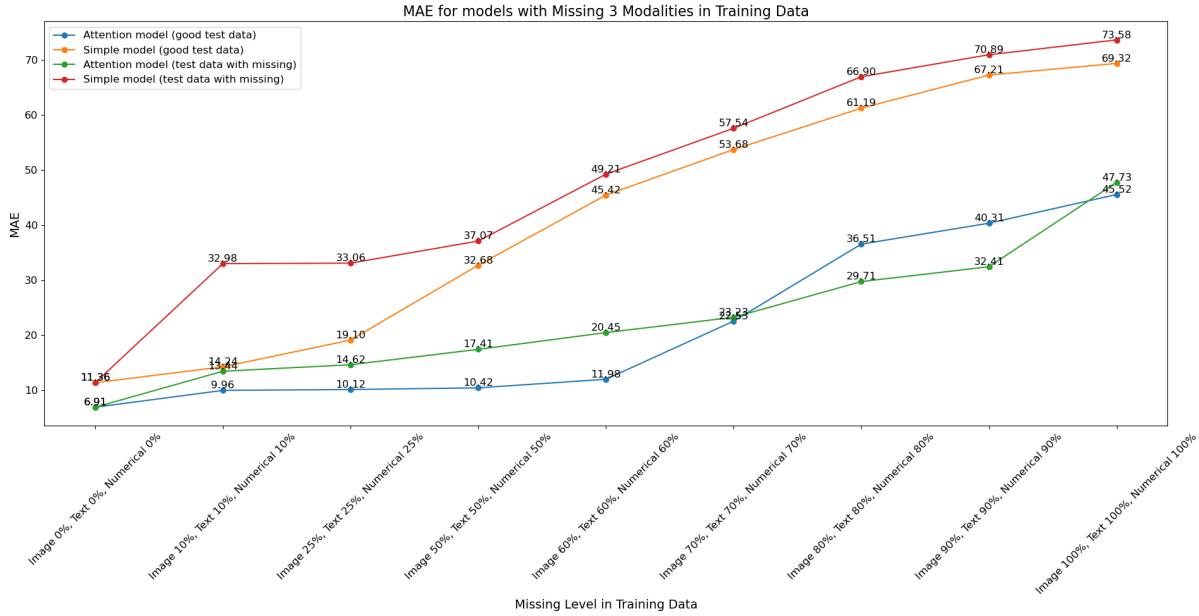


Figure 3.12: Comparison of attention and simple models trained on the dataset with all data missing at different percentages. The attention model mitigates performance degradation when at least 50% training data is available, above which the performance degrades at a rate comparable to the non-attention model.

Table 3.6: Comparison of attention model and simple model when data are noisy.

Noise level in training data	MAE on good test data		MAE on test data with noise	
	Attention	Simple	Attention	Simple
Image 10%	7.52	13.18	9.91	24.13
Image 25%	8.26	13.6	10.16	28.52
Image 50%	11.08	14.73	10.96	30.51
Text 10%	9.47	11.83	11.54	24.04
Text 25%	11.05	12.69	12.31	25.59
Text 50%	11.6	14.44	12.61	27.17
Numerical 10%	8.36	12.32	10.6	23.66
Numerical 25%	9.93	12.53	13.12	30.21
Numerical 50%	11.29	13.16	12.83	30.88
Image 10%, Text 10%, Numerical 10%	9.21	12.94	9.71	27.08
Image 25%, Text 25%, Numerical 25%	10.27	13.39	11.03	31.89
Image 50%, Text 50%, Numerical 50%	13.05	17.33	11.91	47.12

3.4.4 Investigation of multimodal learning performance in noisy data context

Table 3.6 shows the performance comparison of the attention and simple models trained under different conditions of data noise, where image, text, and numerical data contain noise at 10%, 25%, and 50% in the training data.

Figure 3.13 illustrates the performance comparison between the attention model and the simple model trained under the noise conditions detailed in Table 3.6, when evaluated on a clean test set (free of noise). The y-axis represents the mean absolute error (MAE), while the x-axis delineates the various noise conditions. Given the independence of noise levels across different modalities, the graph displays distinct lines for each experiment. The orange line illustrates the performance degradation of the attention model with escalating noise in text data, whereas the brown line tracks the performance of the simple model. Notably, the attention model exhibits a sharper increase in error, moving from 10% to 25% text noise, compared to the more gradual escalation seen in the simple model. Conversely, as text noise intensifies from 25% to 50%, the attention model demonstrates a relatively stable performance, outperforming the simple model.

Figure 3.14 compares the performances of the attention model and the simple model on the test set having noise data at the same levels as the training set. It can be seen that, in all cases, the attention model has a more stable performance compared to the simple model. This is most apparent in the case where there is noise in all three modalities, as

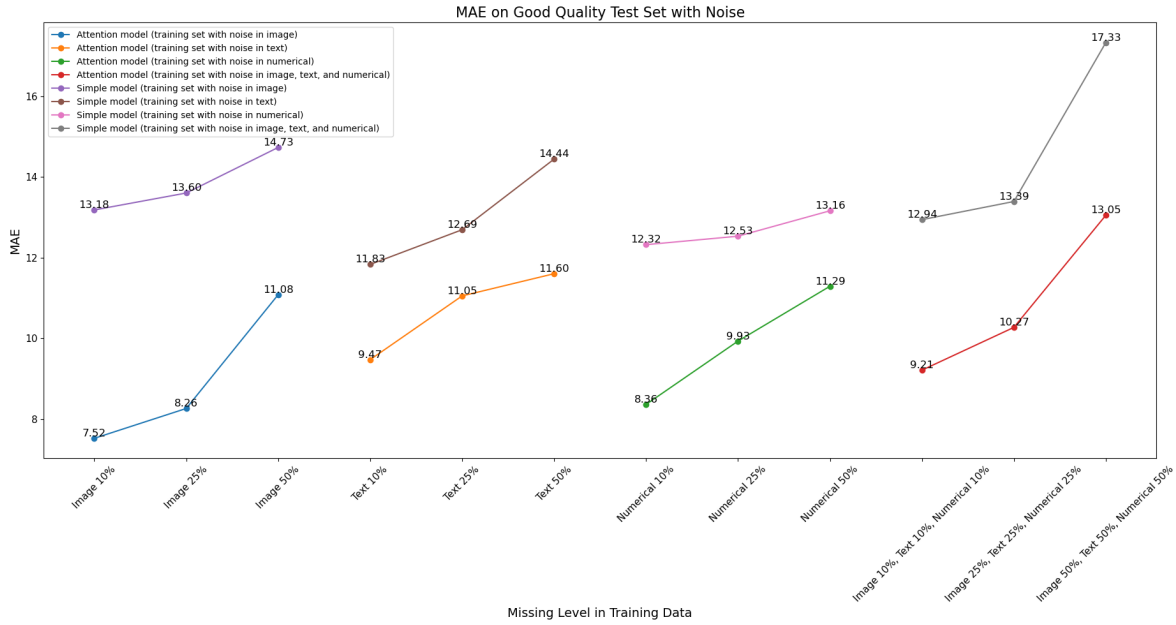


Figure 3.13: Comparison of attention and simple models trained on the dataset with noise and tested on the dataset without noise.

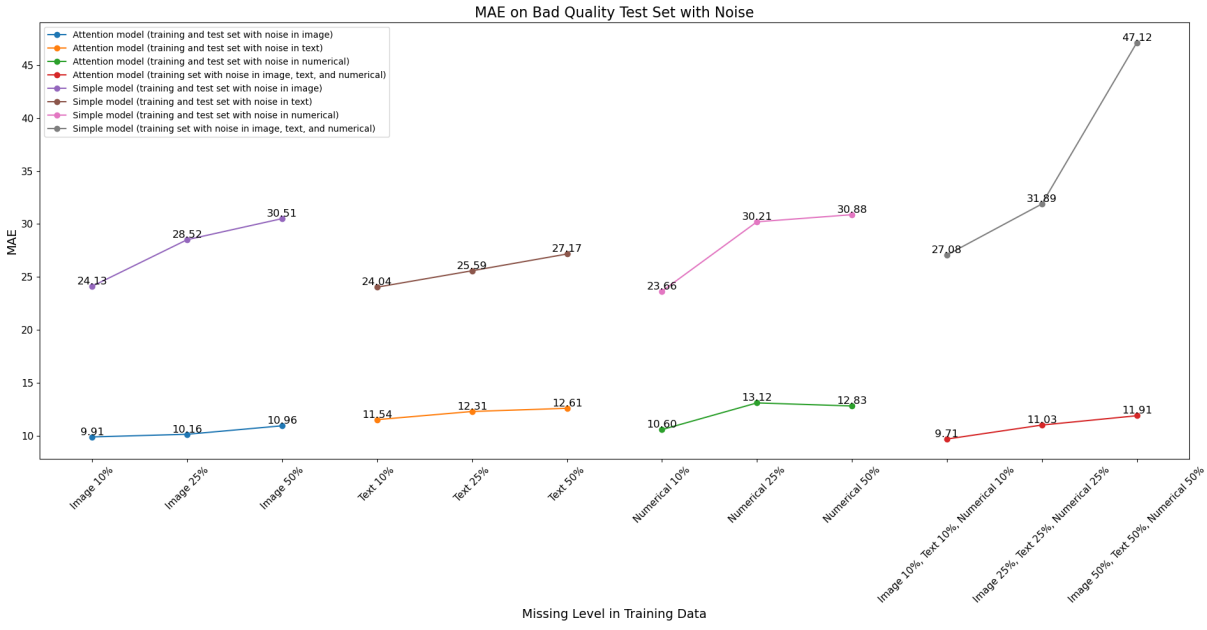


Figure 3.14: Comparison of attention and simple models trained and tested on the test set with the same noise level as the corresponding training set.

shown by the red and gray lines in Figure 3.14. This shows that in the cases where the training set contains noise, the attention model has learned to manage noisy samples better than the simple model.

Key findings

1. How do varying levels of missing and noisy data in each modality (text, image, and numerical) impact the performance of multimodal learning models for industrial prognostics?
    - The attention model sustains prediction performance on high-quality test sets even when trained on data with missingness levels up to 50% on all modalities.
    - The performance of the attention model does not decline too much while increasing noise in all three modalities to 50%, but the error of the simple model nearly doubles.
2. Can the incorporation of crossmodal attention mechanisms improve the robustness and performance of multimodal learning models in the presence of missing and noisy data?
    - Across the full range of noise and missingness levels considered in this study, the attention model has consistently outperformed the simple model.
    - The attention model manages to maintain the rate of performance decline at low levels until around 50% data is missing.

3.5 Conclusion

In this chapter, we have laid the groundwork for understanding and managing multimodal PHM data, focusing on the challenges posed by noise and missing data. Through simulated scenarios, we explored the impacts of these factors on the performance of multimodal learning models, specifically comparing a simple model based on dense layer connection and an advanced attention-based model.

Key findings from this research demonstrate that the attention model is capable of maintaining high prediction performance, even when confronted with significant levels of missing or noisy data across all modalities. Remarkably, this model demonstrates robust-

ness against increasing noise, maintaining its performance much better than the simple model, which relies on dense layer connections and shows a near doubling of error under similar conditions. Moreover, the consistent superiority of the attention model across all levels of data imperfection tested highlights the efficacy of crossmodal attention mechanisms in enhancing model robustness and performance in the face of missing and noisy data.

Building upon these insights, the next chapter will extend these principles to develop a refined methodology for diagnostics utilizing sparsely available multimodal data. We will apply the proposed methodology to a real-world dataset, aiming to validate and potentially enhance the robustness and applicability of our models. This progression ensures a seamless transition from the theoretical explorations of this chapter to practical, industry-focused applications, setting the stage for advanced implementations later in this thesis.

Chapter 4

Diagnostics from Sparse Multimodal Data

Contents

"When used jointly, two modalities normally yield additional information or even a new level of meaning. In many situations, a certain modality allows us to convey information in a way that is sometimes impossible when using another modality. For instance, it is not possible to exactly denote the date of birth in an image without using any textual information; on the other hand, it is not possible to exactly describe a human face or a plant's shape using textual information...

— Ralph Ewerth et al. "Computational approaches for the interpretation of image-text relations" in the book *Empirical Multimodality Research: Methods, Evaluations, Implications.* (Ewerth *et al.* (2021)).

4.1 Introduction and Context

In Chapter 3, we conducted an initial exploration into multimodal learning for PHM using a simulated dataset, which demonstrated that integrating attention layers to facilitate cross-modal interactions enhances the robustness of multimodal learning under suboptimal data quality conditions. This chapter progresses from simulated environments to practical applications, employing fault detection and diagnostics (FDD) on real industrial multimodal data.

In most industrial settings, comprehensive and clean data are unobtainable, often leading to a gap between the anticipated and actual performance of data-driven models (Omri *et al.* (2019)). Data collection frequencies usually differ among condition monitoring tools, which complicates continuous monitoring due to issues such as sparse data, time alignment conflicts, and inadequate training datasets. When data availability and modeling techniques are insufficient, leveraging domain knowledge may provide a viable solution. While various studies such as Kokel *et al.* (2020), Altendorf *et al.* (2012), Atoui and Cohen (2021), Yucesan and Viana (2021) and Liu *et al.* (2015) study different concepts of integrating domain knowledge into a fault detection model, no study considers the abstract knowledge of domain experts or their logical thought process for FDD.

To address all the aforementioned challenges, an expert knowledge-assisted multimodal learning methodology is proposed in this chapter to build a data-driven solution for industrial FDD. The proposal builds on the findings from Chapter 3 to achieve robustness to missing modalities, and will also address multimodal time alignment. We will test this on a dataset from a hydroelectric power generator fleet.

This chapter is organized as follows: Section 4.2 presents the proposed methodology for fault detection and diagnostics. Then section 4.3, introduces the industrial context and applies the methodology to the case study data. Then, section 4.4 presents the results of the application and discusses the ablation study. This completes the first part of the chapter. Section 4.5 extends the methodology by incorporating human knowledge from industrial text documents and inspection notes to enhance degradation level quantification and the outcomes of this process. Finally, section 4.6 summarizes and concludes this chapter.

4.2 Methodology for Fault Detection and Diagnostics

This section develops a methodology that uses expert knowledge and multimodal learning to handle data sparsity and other challenges of industrial automatic system FDD. Particularly, it addresses three following specific challenges: (1) Time alignment issues arise because condition monitoring (CM) data are collected at different times; (2) Some CM tools, such as visual inspection, provide only limited samples due to infrequent data collection; (3) Incomplete expert knowledge about health-indicating features in certain CM data results in variable reliability across different CM tools.

The overall diagram of the proposed methodology is given in Figure 4.1. Its first principle is to translate the human expert's knowledge into an automated digital process. The expert establishes rules linking features extracted from the data to machine health states. Since the expert has already identified useful features, the deep learning model requires less data for training. The second principle is that multiple data modalities can work together for effective diagnostics, but this is more efficient if good features are pre-extracted from each modality. Based on these principles, the methodology is divided into three phases:

1. **Knowledge formalization phase.** This phase formalizes expert knowledge in condition monitoring. Initially, it involves analyzing the relationship between system behavior and the features of each CM data type, aiming to define and interpret the necessary features for extraction. This understanding is encapsulated in knowledge graphs. Subsequently, the task division step determines the appropriate CM tools for specific problems and identifies relevant characteristics for each tool. The insights gained guide the feature engineering process, detailed in the subsequent phase, for various CM measurements.

2. **Knowledge-assisted feature extraction phase.** The feature extraction phase leverages expert knowledge to identify critical features for effective CM. Expert knowledge credibility varies - it may be complete and reliable or incomplete and ambiguous. In scenarios with complete knowledge, the methodology automates the expert's logical process using deep learning models, enabling neural networks to extract specified features from the measurements taken by a tool. Conversely, with ambiguous knowledge, the approach involves creating a pretext task to derive intermediate rather than final features, as illustrated in the upcoming case study (section 4.3.3).
   Models are tailored and trained based on the context to extract pertinent features from each CM data type. Post-training, models undergo validation by human experts. Unsuccessful validations lead to model redesign or refinement. Successfully extracted features then serve as inputs for the multimodal learning phase.

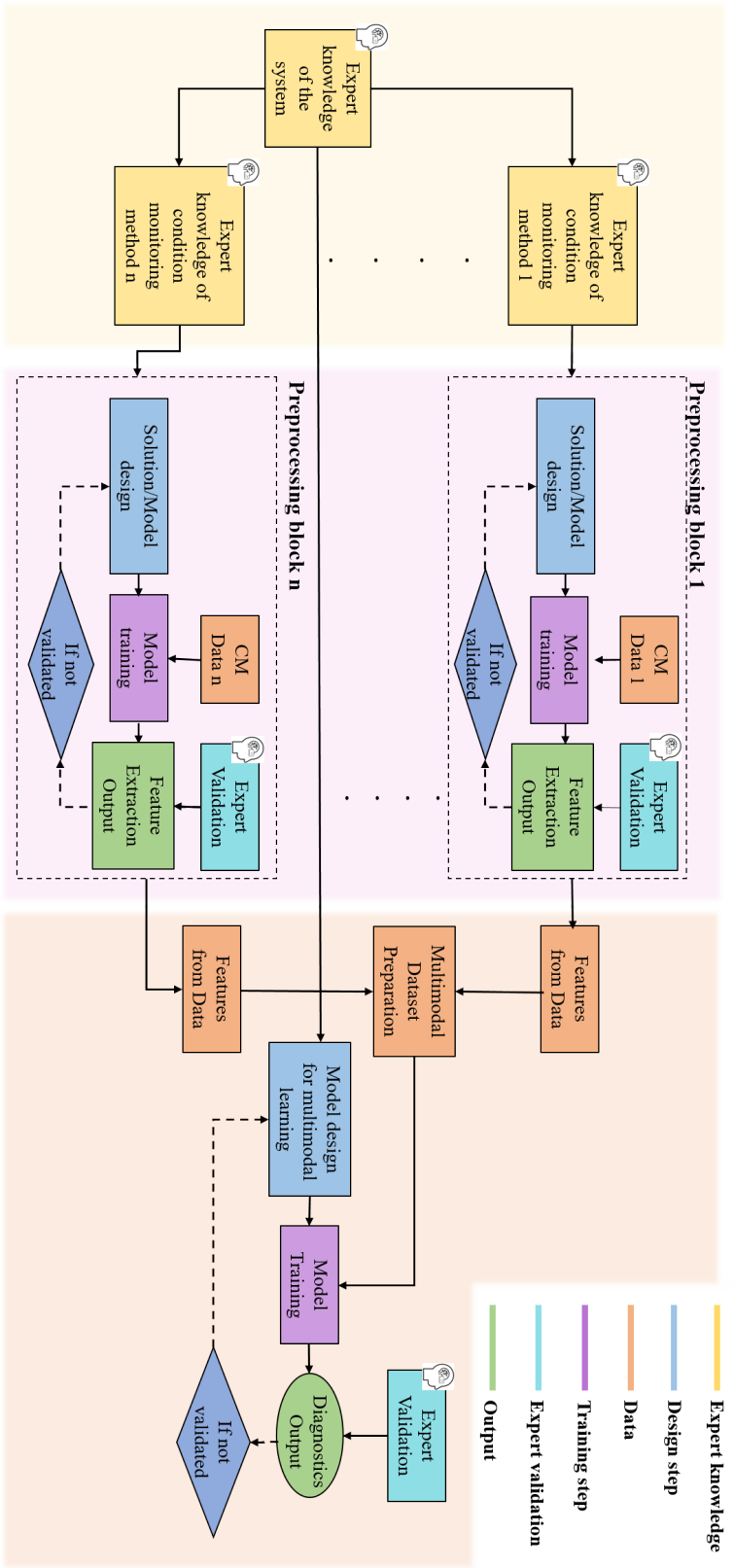


Figure 4.1: Overview of the FDD methodology

3. **Multimodal learning phase.** This phase integrates features extracted from various data modalities to produce FDD outputs. The design of a multimodal learning model incorporates domain expertise on the dynamics between data from multiple sources. For instance, experts may recognize how anomalies manifest differently in temperature versus vibration sensors and understand the typical time misalignment patterns where one signal precedes another in indicating faults.
   Drawing on insights from Chapter 3, an attention mechanism is utilized to capture these intermodal relationships, with the placement of crossmodal attention layers being crucial. This placement, while initially based on unimodal and bimodal model performances, benefits significantly from domain expertise in an industrial setting. Experts can prioritize data modalities based on information quality, guiding the attention from higher to lower priority data.
   Following the design, the model undergoes training using the previously extracted features. The final output of the model indicates the physical degradation state of the machine.

4.3 Application to a Hydrogenerator Fleet

This section presents an industrial dataset from a hydrogenerator fleet and applies the proposed methodology to this data. Subsection 4.3.1 describes the context and dataset of the case study. Subsections 4.3.2 and 4.3.3 present in detail the knowledge formalization and feature extraction steps while subsection 4.3.4 discuss the multimodal model design.

4.3.1 Description of the hydrogenerator case study

In this application, two types of physical degradation states related to partial discharge (PD) occurring within the stator of hydrogenerators were the target diagnostic outputs. Figure 4.2 shows an example of a failure propagation graph of the stator, where each node represents a discrete physical degradation state. An edge represents the transition of the system from one physical degradation state to another following the evolution of the degradation process. The end nodes represent failures in a particular failure mode. In this study, two physical degradation states in this graph are studied. The two states differ by their location (context) in relation to the components of the stator and are related to two types of PD sources. The state code-named E7 denotes partial discharge activity happening between the bars in the overhang portion of the stator winding and is also referred to as Gap PD. The physical degradation state E2A is related to partial discharge activity occurring on the stator bar at the exit of the magnetic core. The data are collected from 26 hydrogenerators.

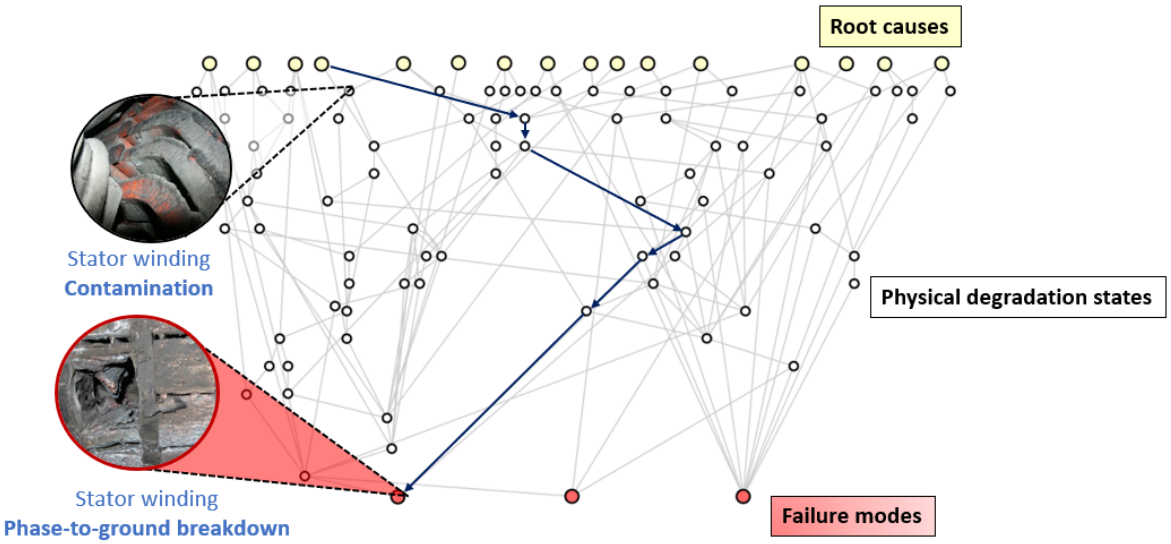


Figure 4.2: Illustration of knowledge graph showing failure propagation. Adapted from Blancke *et al.* (2018).

The stator degradation is monitored through visual inspection (V.I), phase-resolved partial discharge (PRPD), partial discharge analyzer (PDA), ozone, and temperature. Some examples of visual inspection images, PRPD, and PDA measurements corresponding to the states E7 and E2A are given in Figures 4.3, 4.4, and 4.5 respectively. PD activity induces insulation degradation which can be observed upon visual inspection by the presence of a white powder. In Figure 4.3, the corresponding degradation products for both physical degradation states are marked by white circles. In addition, for a more detailed understanding of PRPD and PDA, one can consult the paper by Hudon and Belec (2005). It should be noted that features related to the physical degradation state E7 can be easily extracted from both PRPD and PDA measurements. In the case of E2A, features related to this physical degradation state are the same as those of another physical degradation state related to PD activity on bars inside the magnetic iron core (also called slot discharge). The only way to be certain that E2A is active is to validate PRPD and PDA measurements by visual inspection of bars at the exit of the magnetic iron core. Ozone and temperature measurements are simple numerical data.

Here, the **first challenge of the time alignment issue** manifests as follows. Visual inspections of a hydrogenerator in a power plant may occur every six years or less, while PRPD and PDA measurements are taken more frequently. This is illustrated in Figure 4.6. The data collection frequency follows the order of PDA, PRPD, ozone, temperature, and then visual inspection images. Moreover, measurements of different types are not taken simultaneously but rather separated by months or even years. Note that in the figure, only

Figure 4.3: Visual Inspection images showing both physical states.

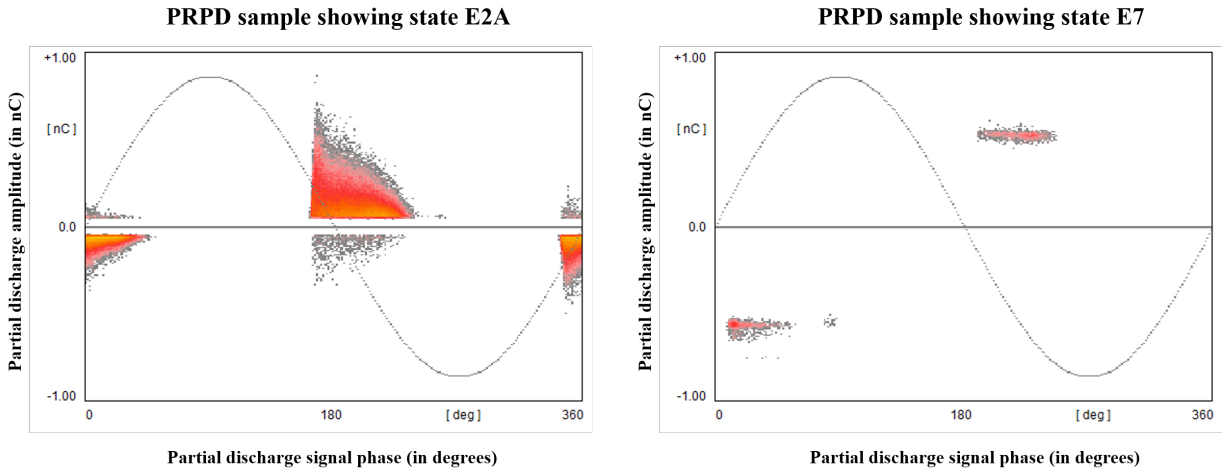


Figure 4.4: Visualization of PRPD samples. Adapted from Hudon and Belec (2005).

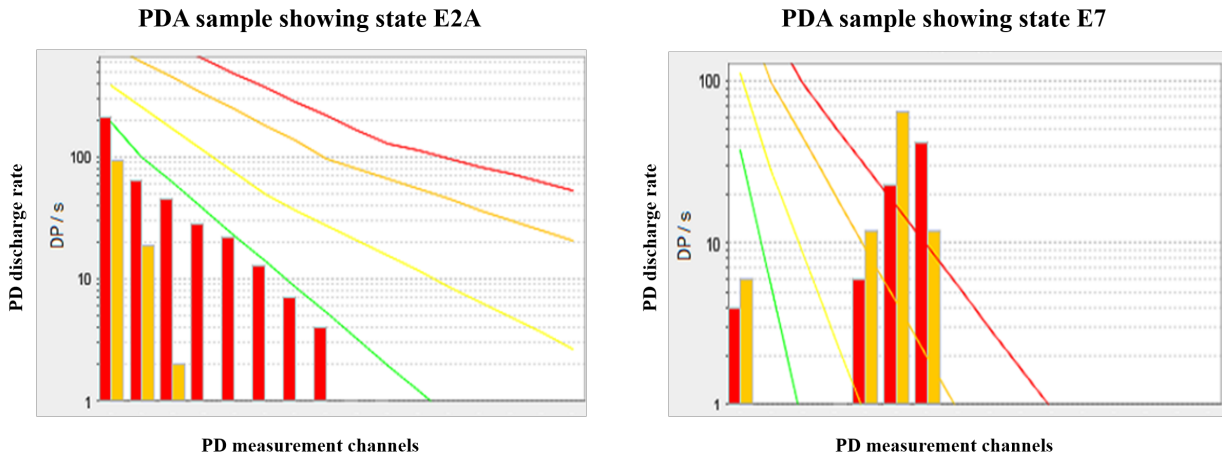


Figure 4.5: Visualization of PDA samples. Adapted from Hudon and Belec (2005).

samples related to the two physical degradation states in the scope of this study are shown.

The **second challenge of insufficient data volume** is especially true for visual

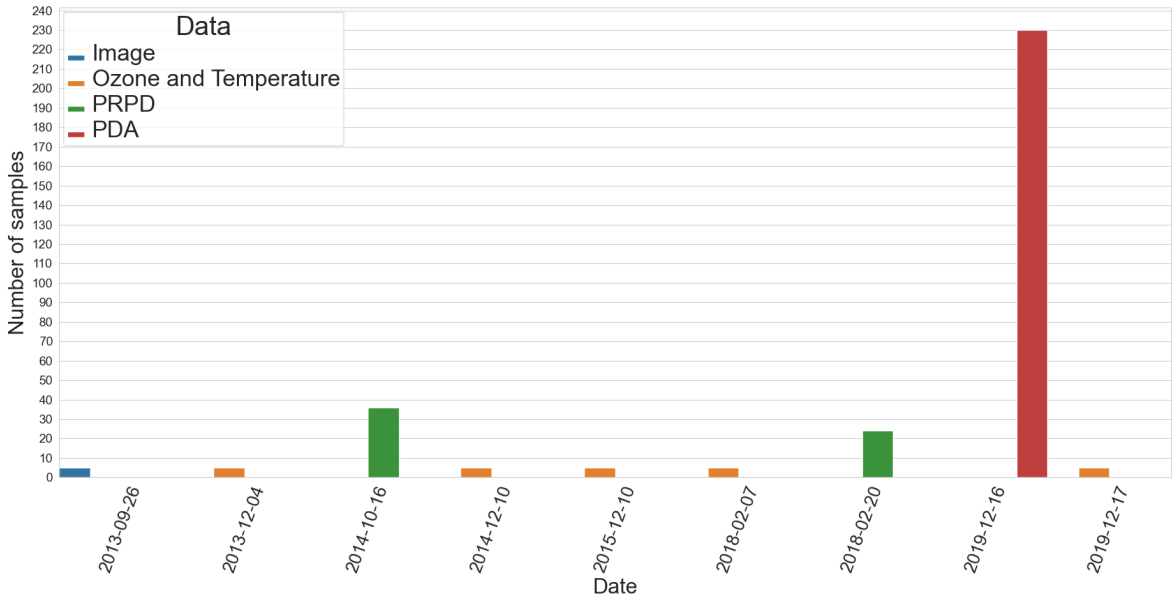


Figure 4.6: Data availability view showing V.I, PRPD, PDA, ozone and temperature data for one generator.

inspection data, where images are collected only once every few years. In fact, there are only 86 images in total (concerning the two physical degradation states in the scope of this study), and these come from 26 hydrogenerators. These hydrogenerators have varying visual characteristics such as color, as seen in Figure 4.3. Thus, it is infeasible to directly train the diagnostic function based on images without the assistance of expert knowledge about how to recognize the symptoms associated with a given physical degradation state.

The **third challenge is related to the limitations of knowledge about the features to be extracted from the data**. This is especially true for PRPD and PDA, where features related to the physical degradation state E2A are the same as those of another physical degradation state in the failure propagation graph. This limitation will be further discussed in Section 4.3.3. In contrast, with a visual inspection image, the diagnostics can be certain. Furthermore, there is no concrete knowledge of how to monitor the condition with ozone and temperature data.

A full view of the customization of the proposed methodology to the case study is shown in Figure 4.7. As the hydrogenerator degradation is a slow process (in the order of years), the measurements taken months apart could indicate the same degradation state. This information as well as the informativeness order of the data sources guide the adaptation of the knowledge formalization, feature extraction, and multimodal learning phases to the case study. Implementation of these phases will be detailed in the remainder of this section.

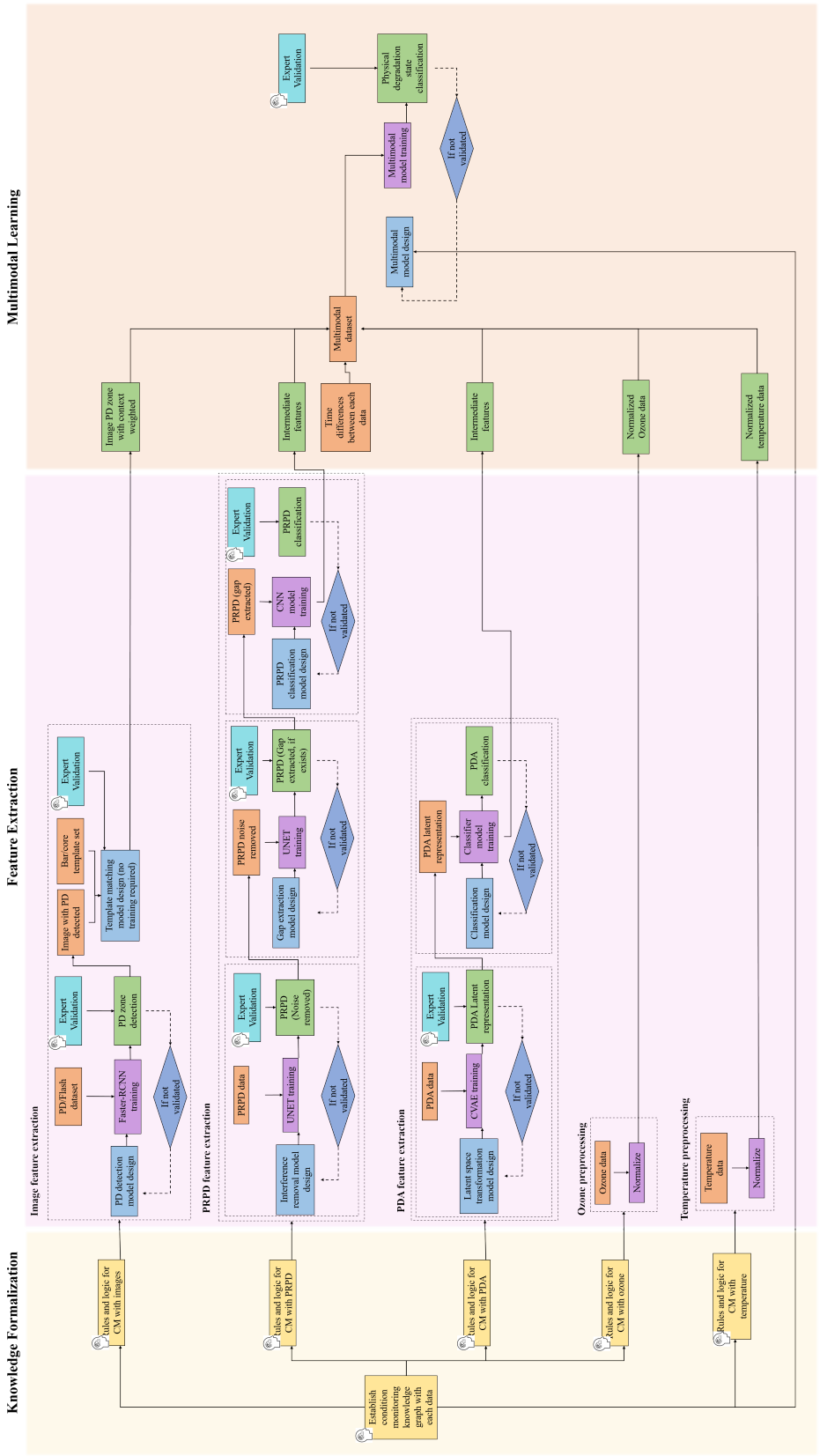


Figure 4.7: Application of methodology to the case study.

4.3.2 Knowledge formalization of hydrogenerator fault detection

In this phase, the first step is to represent the knowledge of the human expert about the stator states derived from each of the CM measurements. This involves knowledge regarding the relevant features associated with each kind of CM data, as well as the FDD rules based on the presence or absence of these features. A representation of this knowledge is illustrated in Figure 4.8. Considering this figure, one can see that a visual inspection image sample may contain degradation products induced by partial discharge, which can be seen as a white powder or rust. One type of partial discharge can appear between bars of the stator, indicating the state E7 and the other type of partial discharge studied herein can appear on the bar at the exit of the magnetic core, indicating the state E2A.

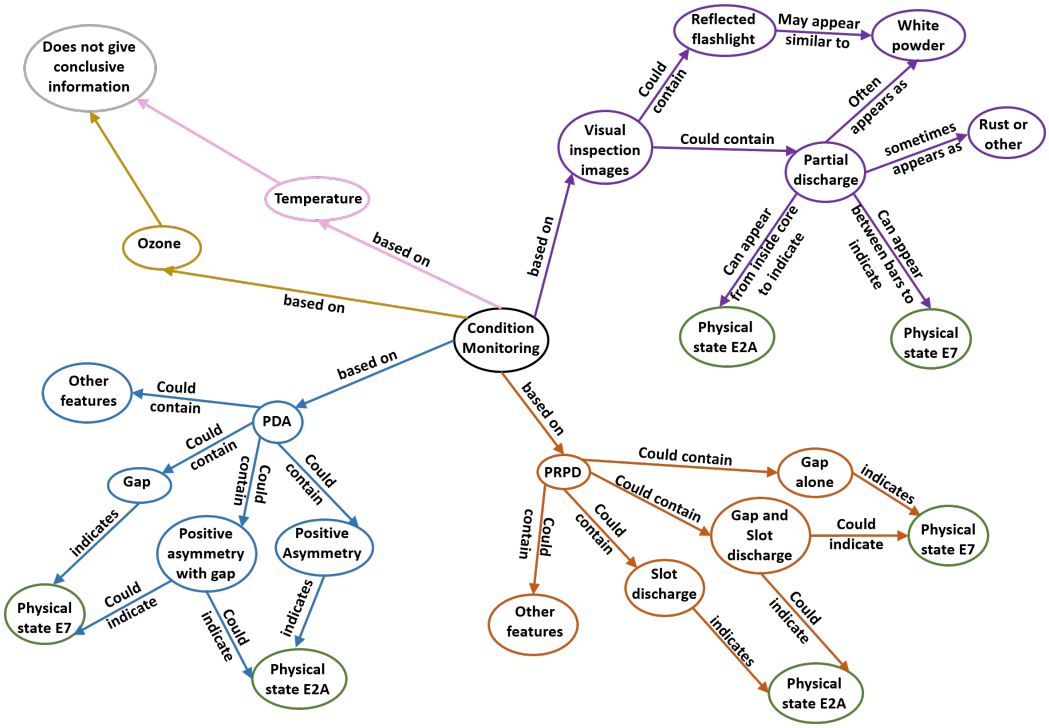


Figure 4.8: Knowledge graph of condition monitoring based on all available tools.

The second step of this phase seeks to differentiate the global knowledge graph into distinct branches in accordance with certain CM tasks. Each branch begins at the condition monitoring node, followed by the CM tool, symbolized by a unique color, and ends at the nodes that characterize the hydrogenerator physical degradation states. For instance, condition monitoring utilizing visual inspection image data is represented by purple nodes and edges. CM utilizing PRPD is indicated by orange, and PDA by blue. FDD is not feasible utilizing ozone or temperature singly, but this is also symbolized by distinct branches in the knowledge graph. In the following subsection, this knowledge will be used to facilitate the knowledge-assisted feature extraction phase.

4.3.3 Knowledge-assisted feature extraction models

In the feature extraction phase, a preprocessing pipeline was made for each of the CM data. Most relevant features were extracted from images despite the small data volume by following the logical process of the human expert. For PRPD and PDA, the known features are not perfect indicators, so these are only used as targets to create a pretext task to transform the data into features.

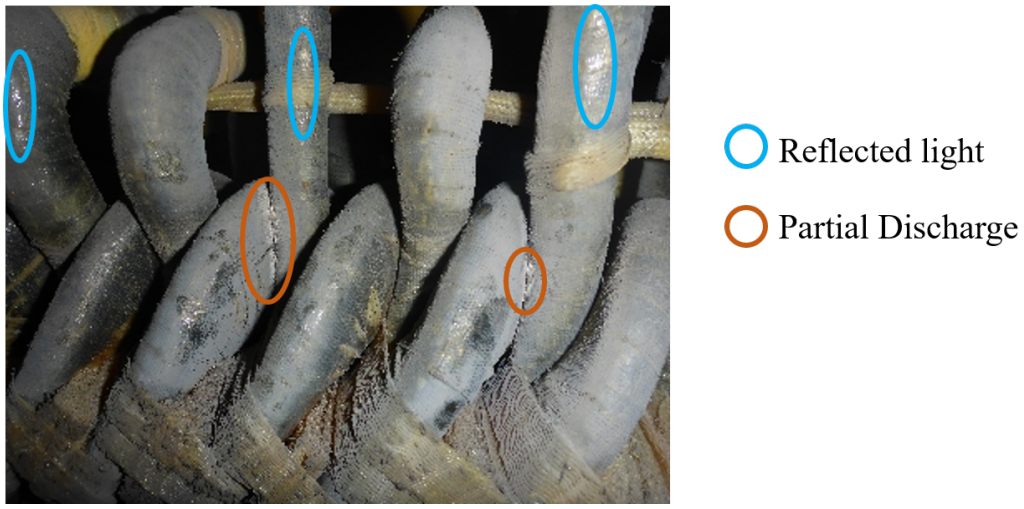


Figure 4.9: Visual inspection image showing a partial discharge degradation products and reflection of light.

4.3.3.1 Feature extraction from images

Of all the CM data available in the study, visual inspection is the most reliable as the features indicating the physical states are already known based on expert knowledge. Considering the branch of condition monitoring based on images in the knowledge graph (Figure 4.8), the identification of the physical degradation states E2A and E7 can be done in two steps. The first step is to detect the presence of partial discharge degradation products from an image, as shown in Figure 4.9. The second step is to infer the context of the partial discharge, such as whether it originates in between bars or on bars at the exit of the magnetic core.

Step 1: Detection of anomaly zone from image
One of the main challenges to accurately detecting partial discharge degradation residue in an image is that it could look very similar to a reflected flashlight. Indeed, most of the time, partial discharge degradation residue appear as white powder, and the visual

properties look similar to a light reflection.

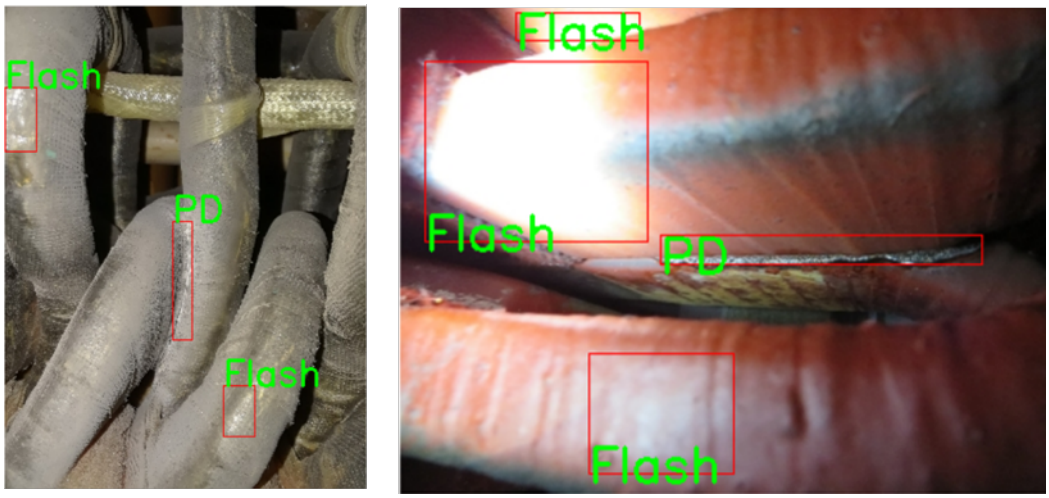


Figure 4.10: Visualization of the dataset created to train the PD detection model.

To solve this task, the problem was reframed as an object detection task. An object detection model will be trained to detect two objects, named PD and Flash, as shown in Figure 4.10. For each image, all instances of partial discharge degradation products are identified by the expert. Then, an entry is made in the dataset (for training the object detection model) including the coordinates of the bounding box around the partial discharge degradation product and a label 'PD' indicating that the bounding box represents an instance of white powder due to partial discharge. The same process was done for all instances of a reflected flashlight in the image, with the assigned label 'Flash'. For example, the training image in Figure 4.10 (left) will have three entries in the dataset, one for the partial discharge and two for the light reflection.

Object detection models always have a tension between speed and accuracy. Here, the slow nature of degradation allows choosing a more accurate model over a fast one. Therefore, a Faster-RCNN (by Ren *et al.* (2015)) based on VGG16 is chosen to train the object detector 'PD' vs 'Flash'. (Discussion of the Faster-RCNN architecture is beyond the scope of this work. Interested readers can refer Ren *et al.* (2015)). The performance of the trained detector is then validated by the human expert. The bounding box drawn by the expert is considered as the true one and is compared to the predicted box given by the trained detector.

Results of detection of anomaly zones from images

Figure 4.11 illustrates the performance of the Faster-RCNN trained to detect partial discharge degradation. In each of the subfigures, it can be observed that the model has effectively learned to locate the degradation zone. However, it should be noted that the

bounding boxes generated by the model are not fully enclosing the degradation region.

Figure 4.11: Some results of PD and reflected flashlight detected by the Faster-RCNN model.

The validity of the predicted bounding boxes is verified against the bounding boxes annotated by the human expert. The mean average precision (AP), defined as the area under the precision-recall curve, for partial discharge degradation, is shown in Figure 4.12. In this case study, AP attains 72% for the partial discharge (PD) class with an Intersection over Union (IoU) threshold of 0.5.

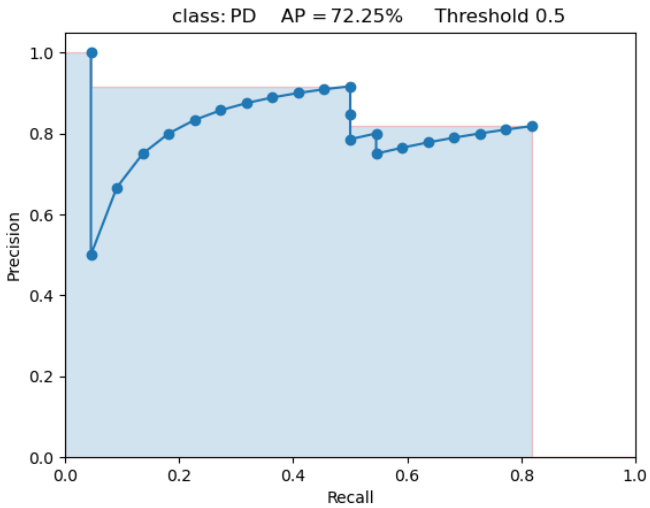


Figure 4.12: Plotting the precision against recall for each of the test images.

Each data point in Figure 4.12 presents the precision and recall values of a single test image. In this scenario, precision refers to the number of correctly classified fault types (physical degradation states) out of all the identified types. Recall gives the number of identified faults out of all existing faults. The precision-recall curve illustrates the tradeoff between minimizing wrong predictions and maximizing the number of correct predictions.

In a given visual inspection image, multiple instances of the target class (PD) may exist. The precision and recall for that image are calculated by considering all instances of the target class and comparing them to the predictions made by the model. For example, the first point in Figure 4.12 has a precision of 1 and a low recall of approximately 0.05. This implies that the test image contains multiple true instances of PD, however, only a small number of these instances were detected by the model, resulting in the low recall. Also, all of the predicted bounding boxes have an IoU greater than 0.5 with a true bounding box, meaning that none of the predictions made by the model were incorrect.

Step 2: Inference of the context of PD degradation

Once the degradation is detected, the next step is to determine its context: whether the PD exists between two bars (state E7) or inside the stator core (state E2A). As these two regions have distinct template characteristics, the identification of state E7 and state E2A can be equated to a template-matching task.

To do this, two sets of context templates are made by simply cropping the regions of bars and the ones of cores from the training set, as shown in Figure 4.13.

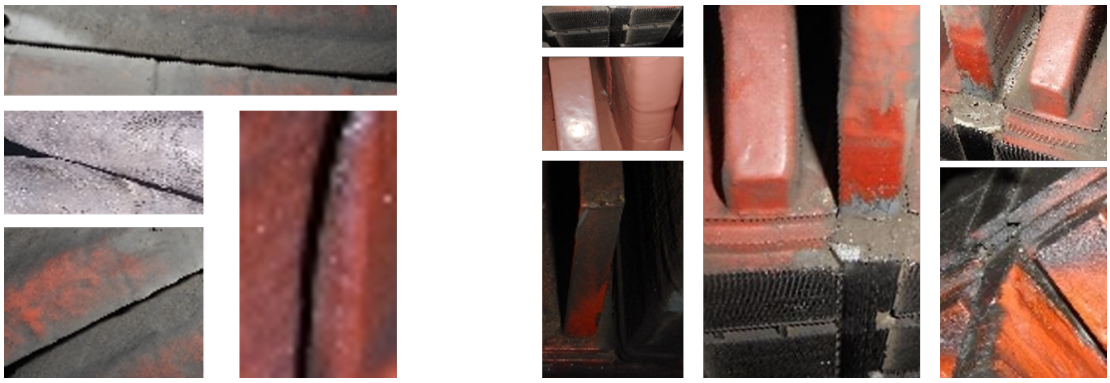


Figure 4.13: Templates of bars (left) and templates of cores (right).

The template matching pipeline is shown in Figure 4.14. This is a no-training, zero-shot method built upon two VGG16 models. For each test image, the PD bounding box is predicted from the previous step. This box is expanded as shown in Figure 4.14. The expanded box is cropped and given as input to the first VGG16 model. One of the templates is the input for the other VGG16. These two inputs are passed through the first three blocks of the VGG16 models, and a cosine similarity between the output matrices is calculated.

A test image is matched in this manner with all the templates in the repository. The matrix giving the highest similarity score is kept. This step is validated by visually (manually) verifying that only the correct context is matched and that a match is never missed.

At the output, in each image, the regions having similarities with the templates will be kept as they are, whereas the rest will be faded to black. This enables the identification of regions that should be given particular attention by the multimodal learning model in the next step. Consequently, the multimodal learning model would not necessitate a large image dataset for training its feature extraction step.

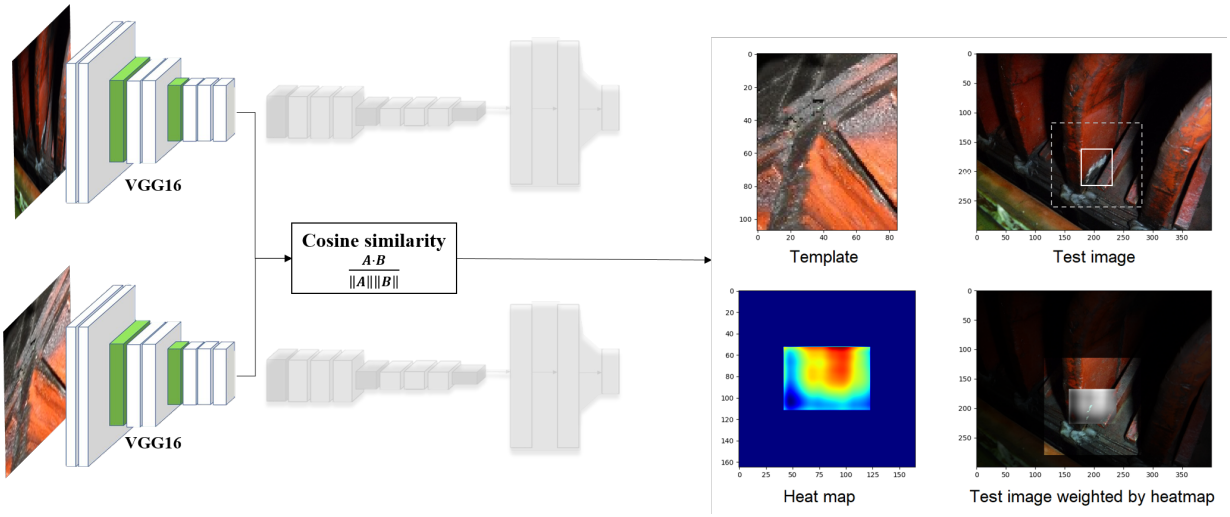


Figure 4.14: Template matching using partial VGG16 and similarity score.

Result of inference of PD context from images

In this study, the targeted context of the degradation can either be the appearance of partial discharge between bars, or on bar at the exit of the magnetic core. This context is identified by the template matching model shown in Figure 4.14, and some results of the template matching are shown in Figures 4.15 and 4.16.

In the first experiment (Figure 4.15), a template of a core is matched against a test image of a core, from a very different perspective. The bottom left subfigure shows the heat map visualization of the similarity score between the template and the region surrounding the PD degradation. To the right of the heat map, the similarity is overlaid on the test image, showing that a similarity is detected between the template and the region of the image surrounding the PD. As the template of a core was matched with the test image, it can be inferred that the context of this observation pertains to the stator core. In the second experiment, Figure 4.16, a template of a bar is matched against a test image showing a degradation between bars. As expected, a similarity is detected.

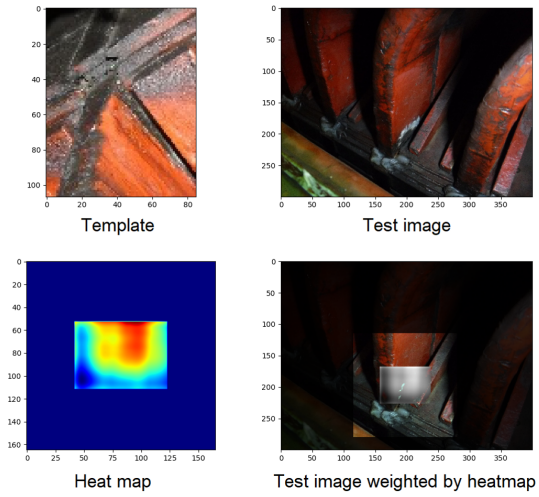


Figure 4.15: Template matching results: Core exit PD.

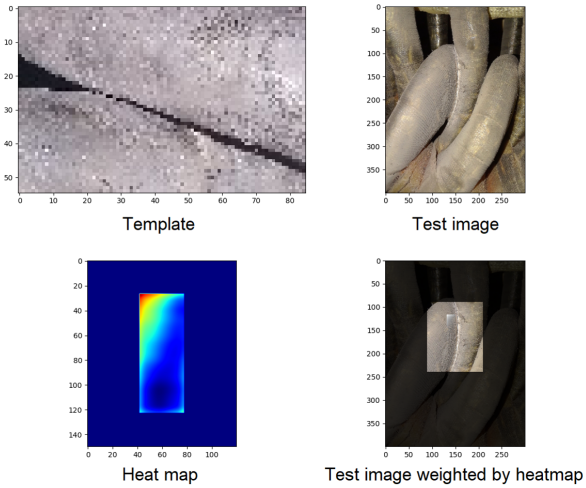


Figure 4.16: Template matching results: Inter-bar PD

4.3.3.2 Feature extraction from PRPD

PRPD measurements are collected more frequently than visual inspection images in the dataset. As shown in Figure 4.8, PRPD signals may possess multiple characteristics, only a few of which are indicative of the physical states under investigation. Additionally, there is some ambiguity when inferring physical degradation states based on PRPD characteristics. For instance, when a PRPD measurement shows both gap and slot discharges, it could indicate either a physical state E7, a physical state E2A, or both.

To preprocess the PRPD, a dataset of PRPD, including those in other physical states, has been collected. This dataset is used to train a model for extracting the relevant features presented in the knowledge graph. Firstly, a U-Net (see Ronneberger *et al.* (2015b)) is employed to remove interference from the signal and another U-Net to extract the gap related to E7, if present. Secondly, a convolutional neural network (CNN) indicates to which class the PRPD measurement belongs as multiple PD sources can be simultaneously active, i.e., slot discharge, gap PD, internal PD, delamination PD with copper conductors, or corona discharge at the junction between the semiconducting and grading coating. The structure of the 2U-Net-CNN model is shown in Figure 4.17.

Although the 2U-Net-CNN model is optimized to achieve the best classification, its outputs are not perfect indicators of the physical degradation states. Consequently, the output of the intermediate layer (before the last three layers of the network) is extracted and utilized to train the multimodal learning model in the next step. These feature vectors, which contain useful characteristics of PRPD measurements, could enable the multimodal model to learn more effective information from PRPD measurements when given additional

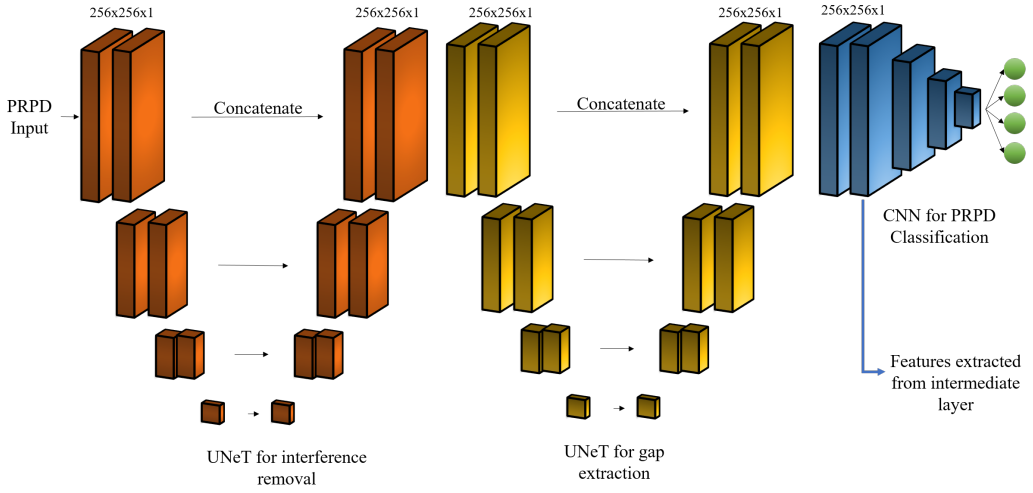


Figure 4.17: Feature extraction from PRPD.

information from other data modalities.

Results of feature extraction from PRPD

The results of the first two U-Net modules, which aim to remove interference and extract gaps from PRPD measurements, are manually validated by the industrial expert. Besides, the performance of the CNN module, which is used to classify the features of PRPD measurements into various classes, is evaluated through the metrics presented in Table 4.1. As observed from the table, the model demonstrates a satisfactory capability in correctly categorizing the PRPD features into the classes of "slot discharge" and "internal". However, its performance in the remaining classes is not good. Considering this, the intermediate features extracted before the classifier output will be used as inputs for the multimodal learning model. This approach will enable the multimodal learning model to extract additional useful information from the PRPD measurements while being guided by more reliable CM measurements, such as visual inspections.

Table 4.1: Classification report for PRPD classifier.

Class↓/Metric→	Precision	Recall	F1-Score	Accuracy
Slot discharge	87.2%	83.85%	85.89%	92.6%
Delamination	52.63%	40.0%	45.45%	90.4%
Internal	79.42%	82.4%	80.88%	79.2%
Corona discharge	56.67%	64.15%	60.18%	91.0%

4.3.3.3 Feature extraction from PDA

PDA measurements are available at a larger volume than PRPD and visual inspections (see Figure 4.6). However, there is ambiguity when inferring the physical degradation state E2A based on PDA, as shown in Figure 4.8. Consequently, similar to the case of PRPD, the known PDA features alone cannot be fully relied on.

To preprocess the PDA, a dataset of PDA including those in other physical states is collected. A convolutional variational autoencoder (see Figure 4.18) is trained to indicate to which feature class a PDA sample belongs, i.e., negative asymmetry, positive asymmetry, symmetry, negative asymmetry with gap, positive asymmetry with gap, symmetry with gap, or gap. However, similar to the case of PRPD, these classes are not reliable indicators of the physical degradation states by themselves. Consequently, the intermediate features, which are extracted at the output of the expanded intermediate layer from the classifier, will be used to train the multimodal learning model in the next step.

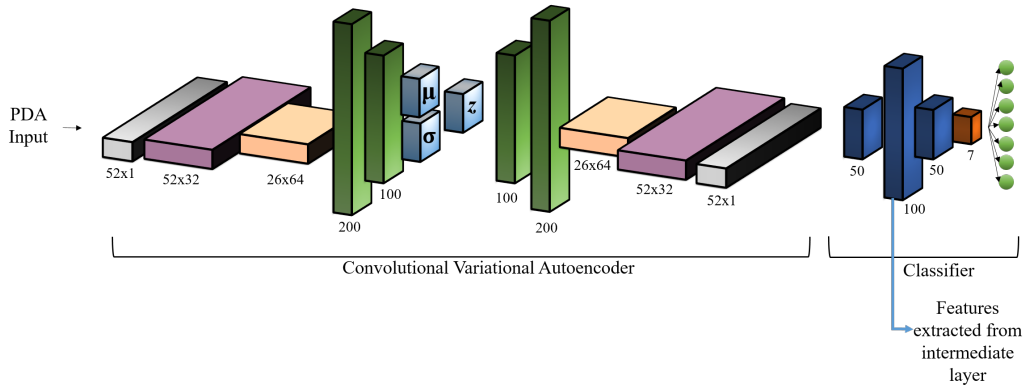


Figure 4.18: Feature extraction from PDA. Adapted from Zemouri *et al.* (2019).

Results of feature extraction from PDA

A deep convolutional variational autoencoder classifier was trained to classify the PDA signals into seven feature classes. Among them, only the feature classes "gap" and "positive asymmetry with gap" are relevant to the investigated degradation states (E2A and E7). This model attains an accuracy of 90% for all classes. A detailed explanation of the results can be seen in Zemouri *et al.* (2019).

4.3.3.4 Preprocessing ozone and temperature data

Ozone and temperature, both of which are simple numerical data, cannot directly indicate the physical degradation state of the machine, as far as expert knowledge extends. Consequently, these data do not require any feature extraction step, but rather can simply be normalized and used in conjunction with other data.

So far, all the steps worked with one data modality at a time. Further steps require the formalization of a multimodal dataset, and then design of a model to learn from this dataset.

4.3.4 Multimodal diagnostics model for two degradation states

In this phase, the outputs of the feature extraction phase will be injected into a multimodal learning model. This model exploits useful information from all data modalities to learn the mapping function between the observations and the physical degradation states (E2A and E7) of the hydrogenerators. This involves defining a multimodal dataset and assignment of appropriate target labels, and then designing a neural network architecture for training on this dataset.

4.3.4.1 Multimodal Dataset Formalization

There are three tasks in the formalization of the multimodal dataset: (1) formalizing the data samples by grouping the measurements taken within a time window as one sample; (2) accounting for time alignment issues; and (3) assigning true labels to these samples.

As shown in Figure 4.6, different CM data are collected at different times and at different sampling frequencies. To create multimodal data samples, given that the degradation rate of the hydrogenerators is too slow compared to the sampling rates of all CM modalities, we propose to group the observations from different CM measurements within a time window corresponding to three years. This time window is based on the domain knowledge that a machine is likely to remain in one degradation state for this long.

Figure 4.19 illustrates the preparation of samples within the multimodal dataset, highlighting those with complete and partially missing data. It illustrates the construction of samples based on measurements taken within a specified time window: a solid black rectangle signifies a sample with complete data taken simultaneously, while a dashed black polygon indicates a sample with temporally distinct measurements. Samples with missing data, such as the one represented by a dashed red polygon containing only PDA, ozone,

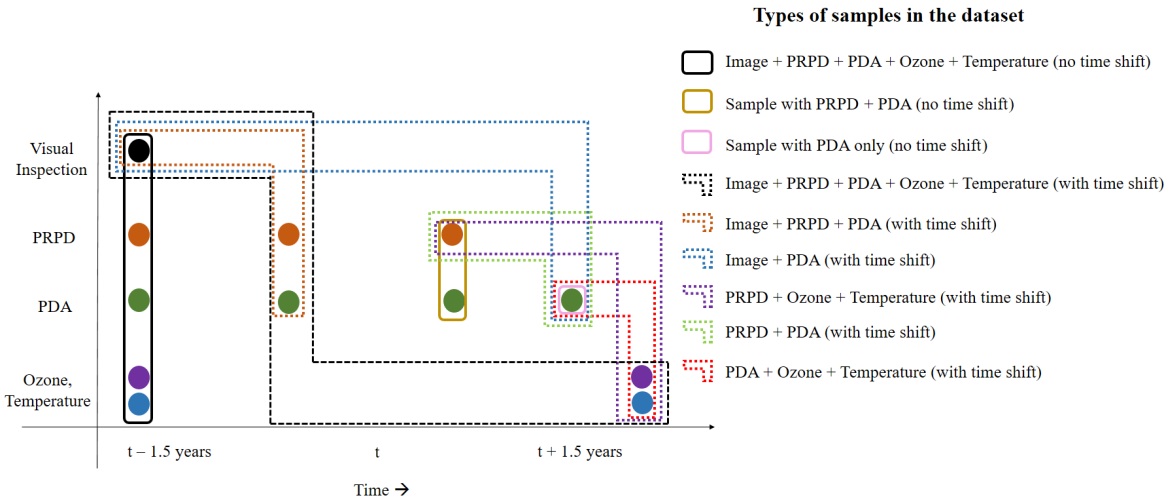


Figure 4.19: Illustration of the different samples in the multimodal dataset.

and temperature data, employ a zero matrix for imputation — substituting missing images with a pure black image of equivalent size.

Definition of a multimodal dataset sample

Definition 4.1 (Dataset Sample)**:**
A sample in the multimodal dataset comprises data points from various modalities, $m_k^{(i)}(t_k^i)$, collected at specific times and organized within a predetermined temporal structure. Here, $m^{(i)}$ indicates a modality from the set $\mathcal{M}$,

$$\mathcal{M} = \{m^{(1)}, m^{(2)}, ..., m^{(n)}\} \equiv \{m^{(i)}\}_{i=1}^{n}$$

where i denotes the priority of a modality as determined by expert judgment. Each k^{th} sample from the i^{th} modality, recorded at time t_k^i, is captured within a sliding time window $W_t = [t - \Delta t, t + \Delta t]$. This window facilitates the synchronization of data across modalities, ensuring that each sample provides a coherent snapshot of the system's state at a similar time point, despite variations in data collection frequencies or delays among modalities.
The temporal difference T_h between samples of different modalities is calculated using Algorithm 1, promoting comprehensive temporal alignment within the dataset. Each sample includes condition monitoring measurements from the set $\mathcal{M}$ and a vector T_h.

Algorithm 1 Algorithm to calculate time difference vector.

$h \leftarrow 1$
for $i = 1$ to $n - 1$ **do**
  for $j = i$ to n **do**
    if $((m_k^{(i)}(t_k^i)$ not missing) and $(m_k^{(j)}(t_k^j)$ not missing)) **then**
      $T_h = |t_k^i - t_k^j|$
    else
      $T_h = 0$
    $h \leftarrow h + 1$

Label assignment based on informativeness rank of data modalities

Definition 4.2 (Target Label)**:**
The ground truth label for each sample in a multimodal dataset is determined based on the informativeness rank of the available data modalities. Given a set of modalities $D_k = \{d_1, d_2, ..., d_n\}$ *for each sample* k*, and an informativeness ranking* $\mathcal{I} = \{i_1, i_2, ..., i_n\}$ *where* $i_1 > i_2 > ... > i_n$*, the label is assigned as follows:*
Define $C = \{c_1, c_2, ..., c_m\}$ *as the set of all classes for which the degradation state is to be determined. Each modality* d_i *within the sample* D_k *is checked sequentially from the highest to the lowest rank based on* $\mathcal{I}$*. The label for each class* c_j *in the sample is assigned based on the first modality that is present, according to the rule:*

$$Label(c_j) = \begin{cases} 1 & \text{if } d_i \text{ indicates a positive degradation state for } c_j \\ 0 & \text{otherwise} \end{cases}$$

for all c_j *where* $c_j \in C$
This method ensures that the label of each sample accurately reflects the most reliable data, adhering to the expert-defined informativeness of the modalities. It assigns a label of '1' to any class c_j *where the first available modality* d_i *indicates an active degradation state, and a label of '0' for all other classes, maintaining consistency with the highest fidelity data representation of the monitored physical state.*

Each dataset sample as defined by Definition 4.1 contains time differences given in terms of months. For example, if a visual inspection is taken in June 2020 and a PRPD in January 2020, $t(V.I) - t(PRPD)$ equals 5.

In this case study, based on Definition 4.2, there are three possibilities for the assignment of the true labels for an active physical degradation state for each sample:

1. If all data are present (image + PRPD + PDA + ozone + temperature), it assigns a true label based on visual inspection information.

2. If only the image is missing (PRPD + PDA + ozone + temperature), it assigns a true label based on PRPD measurements.

3. If both the image and PRPD are missing (PDA + ozone + temperature), it assigns a true label based on PDA measurements.

Thus, for each sample, a true label (0 or 1) is based to indicate if either or both of the physical degradation states are active. If both are active, the true label will be 1 for both classes. Since both labels can be true at the same time, this is a multilabel dataset. Once the labels are assigned to all samples, the multimodal dataset is ready to use.

4.3.4.2 Multimodal model design

The multimodal learning model, shown in Figure 4.20, uses the following input to predict the physical degradation states E2A and E7 of the hydrogenerators:

1. Relevant region of visual inspection image extracted from feature extraction phase;

2. Intermediate features from PRPD classifier after passing through interference removal and gap extraction;

3. Intermediate features from PDA classifier after passing through a latent space transformation;

4. Normalized ozone;

5. Normalized temperature;

6. Vector of time differences between CM measurements.

The proposed multimodal learning model integrates convolution blocks, dense layers, and attention mechanisms. Convolution blocks process image and image-like data, whereas dense layers manage numerical data. Features are flattened at the model's conclusion and linked to a dense layer. Connections between layers are depicted with solid arrows and dashed arrows between blocks of the same type indicate repetitions, such as multiple convolution blocks in the path handling preprocessed visual inspection data.

Three of the attention layers are crossmodal information passing layers, which take into account the influences between different data modalities, while the fourth is used to address the time alignment issue. The crossmodal attention connections are designed based on the industrial expert's knowledge of the data reliability order. For example, the

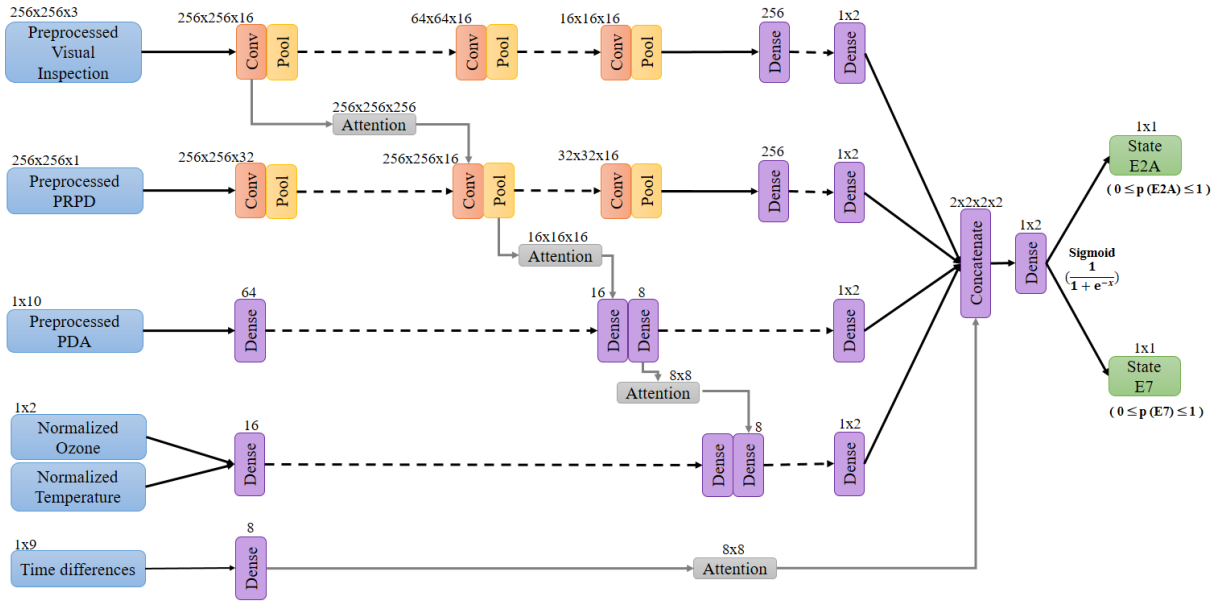


Figure 4.20: Multimodal model structure.

features extracted from the visual inspections are the most reliable ones for indicating the symptoms related to the physical degradation states of the hydrogenerator. Therefore, an attention layer is connected from the first block of the image-learning-path to the PRPD-learning-path.

This attention layer will guide the multimodal learning model to seek the most relevant features from PRPD data by assigning the appropriate weights. If visual inspections are missing, the attention layer will assign equal weights to all features of the PRPD, meaning that there will be no effect from the image-learning-path to the PRPD-learning-path.

Similarly, given the superior reliability of PRPD indicators in comparison to those of PDA, the second attention layer leverages information derived from PRPD features to guide the PDA learning process. It starts from the PRPD-learning-path layer after the one that receives information from the image-learning-path, to take into account also the information from visual inspection. Next, the third attention layer, designed with a similar principle, guides the ozone- and temperature-learning-path based on the valuable information obtained from the PDA learning path, as well as incorporating information inherited from other more reliable CM measurements.

Next, unlike the first three attention layers, the fourth one, starting from the time difference vector, aims to solve the time alignment problem. This attention will learn the assignment of weights to each modality of data based on the time differences between them. For example, if the time difference between the PDA and the other data is high,

this attention may give low weight to the PDA data.

Finally, the output layer uses the sigmoid activation function to perform the multilabel classification task. It provides two outputs, one for each physical state. The output values range from 0 to 1. An output of 0 indicates that the hydrogenerator is not in the corresponding physical state, and vice versa. In contrast to softmax, a sigmoid is more suitable because it can predict a value of 1 for multiple classes.

Takeaways

Key design principles for multimodal data-driven FDD models

Integration of expert knowledge: The methodology harnesses expert insights to formalize and prioritize fault detection features, optimizing the feature extraction process and enhancing model reliability in sparse data scenarios.
Synergy of multimodal data: By strategically combining different modalities of data, the methodology is designed to address challenges such as time misalignment and data sparsity, leading to more robust fault detection and diagnostics.
Iterative validation and refinement: Continuous expert validation and iterative refinements are critical, ensuring that the models remain aligned with practical, real-world applications and effectively capture the nuances of physical system degradation.

4.4 Diagnostics Results

The features extracted from the previous phase are used to train the multimodal learning model for the detection and diagnostics of two physical degradation states included in the failure propagation graph of hydrogenerators (E7 and E2A). Each degradation state has two possible values, with a value of 0 indicating that the physical degradation state is inactive and 1 indicating that the physical degradation state is active. The confusion matrix for the output classes on the test samples is presented in Figure 4.21. One can see that all degradation states (E7, E2A, and E7 & E2A) are successfully detected. In addition, the model exhibits near-perfect performance for the state E7. This outcome is expected since the gap, which serves as the indicator for state E7 in PRPD and PDA, is relatively easy to detect.

The results on the full test set of 962 samples can be seen in Figure 4.22. In this figure, the predictions form roughly three clusters as expected. Perfect predictions would cluster all the points near (0,1), (1,0), and (1,1) according to the corresponding true class. It can

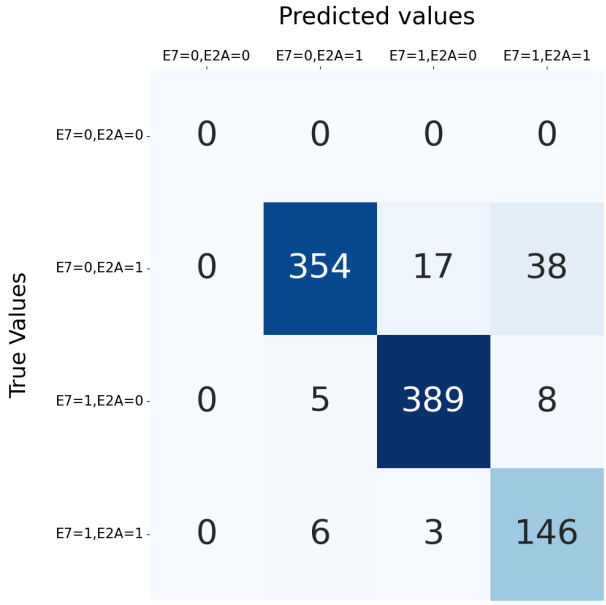


Figure 4.21: Confusion matrix for the test set on proposed model (trained on preprocessed data).

be observed that the predictions for the class where both states are active seem tightly grouped. This is not the case for the other two classes.

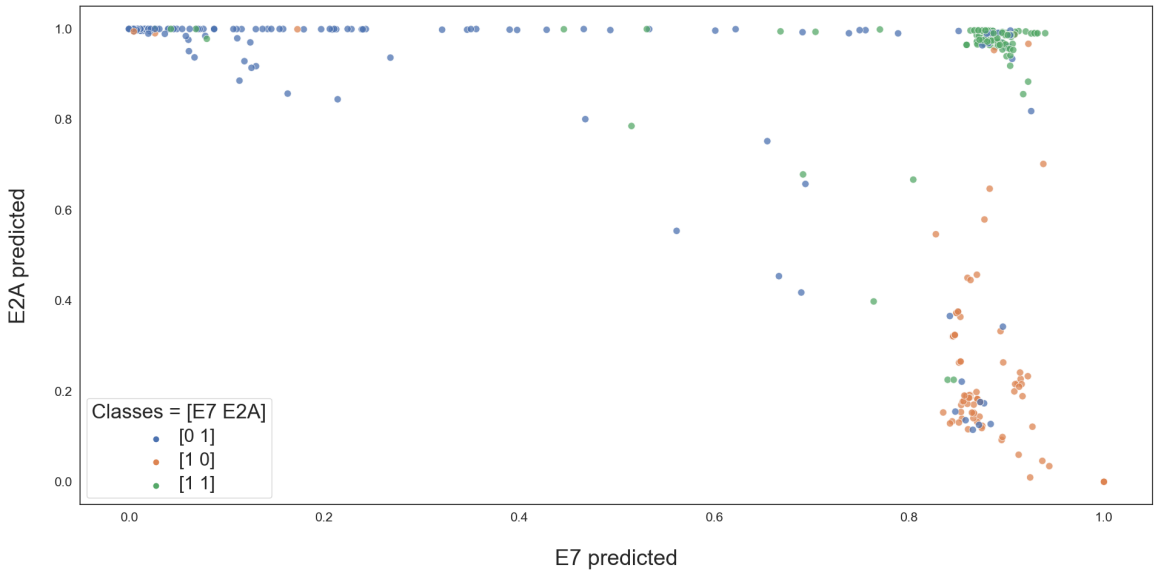


Figure 4.22: Full results of the main model showing prediction clusters.

A subset of 100 results from the test set is shown in Figure 4.23. Here, the individual prediction errors can be seen clearly. One can see that each prediction yields a value ranging

from 0 to 1. It is worth mentioning that if the difference between the true and predicted value is less than 0.5, the prediction can be rounded to the correct result. Contrarily, if the difference is greater than 0.5, the prediction is considered incorrect. As presented in Figure 4.23, the majority of the instances of degradation state E7 are accurately identified.

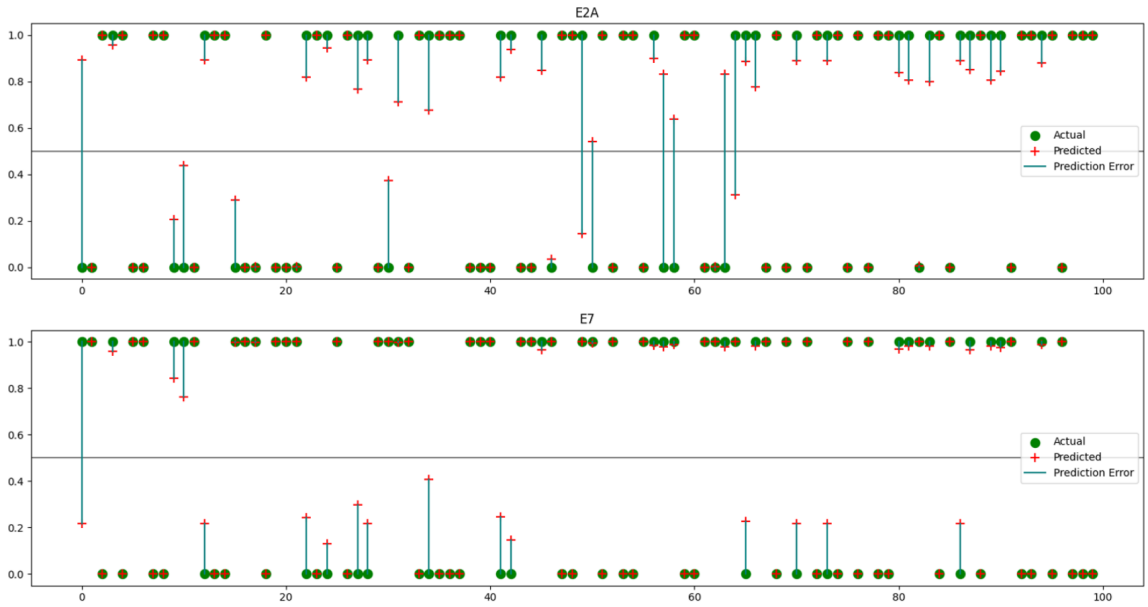


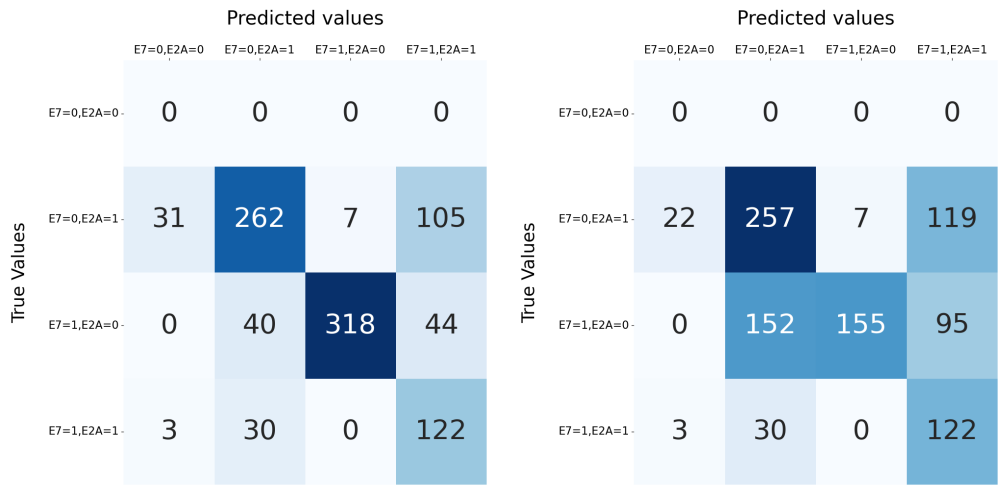
Figure 4.23: Results of prediction on 100 test samples.

The rest of this section will will cover the ablation study of the proposed methodology, focusing on (1) the design of the feature extraction phase informed by expert knowledge, and (2) the use of attention mechanisms in the multimodal learning approach to direct the training of certain CM modalities using reliable information from others. The importance of knowledge-assisted feature extraction and attention mechanisms will be explored in subsection 4.4.1. Additionally, the methodology's efficacy in making accurate predictions with sparse or missing CM data types will be demonstrated in subsection 4.4.2.

4.4.1 Role of knowledge-assisted feature extraction and attention layers

To investigate the role of the knowledge-assisted feature extraction phase and the attention mechanisms, a comparison was made of the performance of the proposed methodology with its simple versions: (1) with the knowledge-assisted feature extraction phase but without the attention mechanisms, called Model A, and (2) without the knowledge-assisted feature extraction phase or the attention mechanisms, called Model B.

Comparison with model A: In this experiment, a simple version of the proposed multimodal learning model is created by removing all the attention connections. The obtained results are shown in Figure 4.24a. Compared to Figure 4.21, one can see that due to the lack of attention mechanisms, model A cannot detect all degradation states. For illustration, 31 observations of the degradation state E2A are wrongly recognized as a healthy state, and 105 instances of the state E2A are misclassified as a combined defect (E7 and E2A). These findings emphasize the significance of the attention connections between different modalities in the proposed methodology.



(a) Confusion matrix of the model A (simple version without attention mechanisms).

(b) Confusion matrix of the model B (simple version without attention mechanisms and knowledge-assisted feature extraction phase)

Figure 4.24: Results of model without attention and model without attention or feature extraction.

Comparison with model B: In this experiment, the simple version (Model B) without the attention connections and without the knowledge-assisted feature extraction phase was trained on the raw data. The results are presented in Figure 4.24b. It is evident that the predictions made by Model B are not much better than a statistical average. However, it is noteworthy that model B performs better in detecting the physical degradation state E2A than the physical degradation state E7. This observation is in contrast to the performance of the proposed methodology, as well as the expert-based expectation that the main indicator of the physical degradation state E7 (gap in PRPD or PDA) would be easier to detect. This indicates that Model B, without the guidance of expert knowledge or domain-specific preprocessing, is unable to extract useful features indicating the physical degradation state E7 from the CM measurements. These results emphasize the significance of the knowledge-assisted feature extraction phase in the proposed methodology.

A comprehensive comparison of the performance of the proposed methodology with models A and B is presented in Table 4.2. Evaluation metrics such as precision, recall, f1-score, and accuracy highlight the superiority of the proposed methodology over models A and B. The results indicate that the performance order is as follows: (1) the proposed methodology, (2) model A (simple version without attention mechanism but with the knowledge-assisted feature extraction phase), and (3) model B (simple version without attention mechanism and without the knowledge-assisted feature extraction phase).

Table 4.2: Classification reports for the different models compared to the proposed model.

Model	Class	Precision	Recall	F1-Score	Accuracy
Model B	E7=0, E2A=1	0.59	0.63	0.61	0.56
	E7=1, E2A=0	0.96	0.39	0.55	
	E7=1, E2A=1	0.36	0.79	0.50	
Model A	E7=0, E2A=1	0.79	0.65	0.71	0.73
	E7=1, E2A=0	0.98	0.79	0.87	
	E7=1, E2A=1	0.45	0.79	0.57	
Proposed	E7=0, E2A=1	0.97	0.87	0.92	0.92
	E7=1, E2A=0	0.95	0.97	0.96	
	E7=1, E2A=1	0.78	0.94	0.85	

4.4.2 Performance of the proposed framework under sparse data context

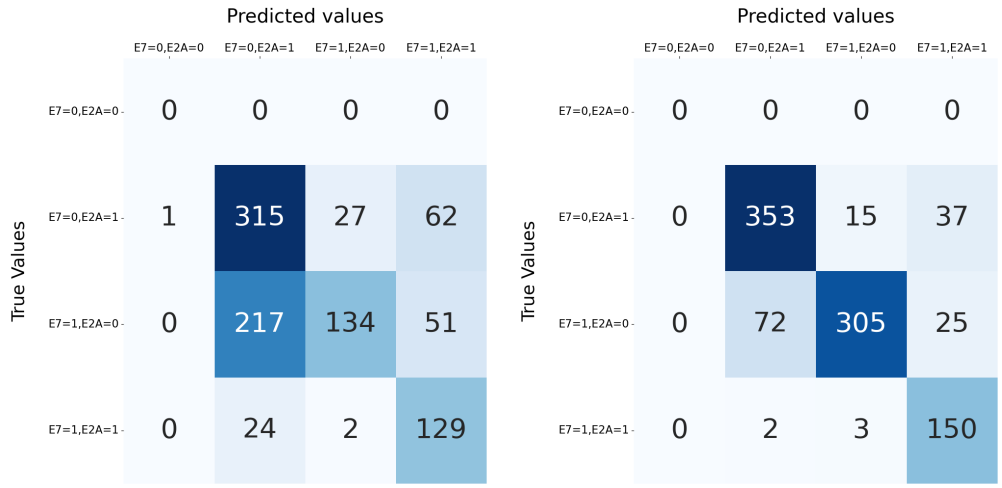
In this section, a series of experiments are conducted to observe the performance of the proposed model under sparse data context, when missing one or more data modalities. Precisely, the three following experiments are conducted:

1. Visual inspection images are missing from test samples;
2. Images and PRPD data are missing from test samples;
3. Images, PRPD, and PDA are missing from test samples.

These experiments can represent possible real conditions in the industry, where some CM tools may not be available once the model is deployed for FDD in real time.

In the first experiment, all images from the test set are removed. Precisely, the image in each sample is replaced with a pure black image. The confusion matrix of the model A is shown in Figure 4.25a, and that of the proposed model in Figure 4.25b. It can be seen that even when visual inspections are missing, the proposed methodology demonstrates

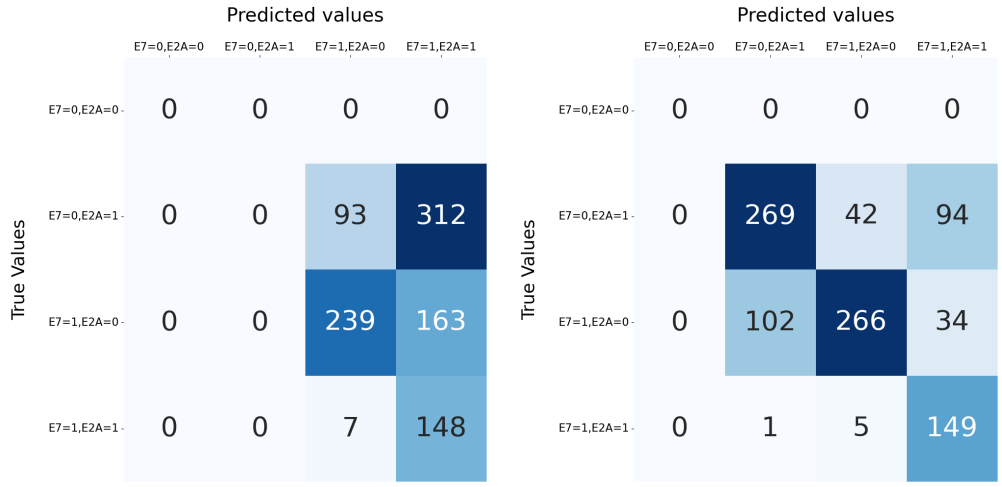
remarkable resilience and retains its accuracy in detecting all degradation states. The misclassification results are within acceptable limits, highlighting the robustness of the proposed methodology.



(a) Confusion matrix of the model A on test set without image data.

(b) Confusion matrix of the proposed model on test set without image data.

Figure 4.25: Results of model A and proposed model on test set without image data



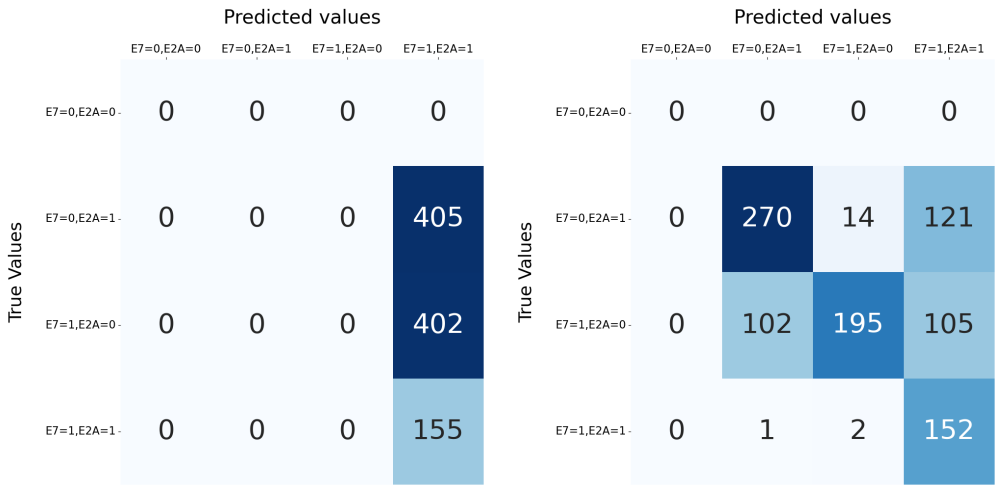
(a) Confusion matrix of the model A on test set without image and PRPD data.

(b) Confusion matrix of the proposed model on test set without image and PRPD data.

Figure 4.26: Results of model A and proposed model on test set without image and PRPD data

In the second experiment, the images are replaced by black images, and all PRPD data are replaced by a zero matrix of the same dimension. In this case, the model is forced to

make predictions based only on PDA, ozone, and temperature data. The confusion matrix of the predictions made by model A is shown in Figure 4.26a and that of the proposed model in Figure 4.26b. One can observe that the proposed methodology works much better than model A. In particular, considering Figure 4.26b, when images are missing, 102 instances of degradation state E7 are misclassified as degradation state E2A. Despite this, the proposed methodology still successfully detects all anomalies (such as degradation states E7, E2A, and E7 & E2A).



(a) Confusion matrix of the model A on the test set without image, PRPD, and PDA data.

(b) Confusion matrix of the proposed model on the test set without image, PRPD, and PDA data.

Figure 4.27: Results of model A and proposed model on test set without image, PRPD, and PDA data

In the third experiment, images, PRPD, and PDA are all removed from each sample in the test set. In this experiment, the model predicts based on only ozone and temperature data. The confusion matrix of the predictions made by model A is shown in Figure 4.27a and that of the proposed model in Figure 4.27b. Given that human experts and model A are unable to make reliable FDD conclusions based only on ozone and temperature data, the proposed methodology still successfully detects all anomalies. Moreover, it indicates the combined degradation states (E7 & E2A) nearly perfectly. Through this series of experiments, the superiority of the proposed model, and especially the attention mechanism in the context of sparse data were demonstrated.

A comprehensive view of the performances of model A and the proposed methodology under the effect of sparse data issue is presented in Table 4.3. One can observe that in all cases the proposed methodology works better than model A (multimodal learning without attention mechanism) and performs quite well when only images are missing. Although the performance of the proposed methodology decreases as more modalities of input are

missing, in the absence of most CM data types (images, PRPD, and PDA) it is still better than what can be achieved by human experts and by model A.

Indeed, the attention mechanism in the proposed multimodal learning model allows exploiting useful information from ozone and temperature measurements for FDD of hydrogenerators according to the guidance of other CM data modality learning paths. Therefore, even when certain CM data modalities are missing, the proposed methodology can still use the learned features from the remaining CM data to predict the active physical degradation states of hydrogenerators. These results highlight the robustness of the proposed methodology when missing one or a few CM data types.

Table 4.3: Classification reports for predictions made on the test set with partially missing data.

Model	Missing Data	Class	Precision	Recall	F1-Score	Accuracy
Model A	Image	E7=0, E2A=1	0.57	0.78	0.66	0.60
		E7=1, E2A=0	0.82	0.33	0.47	
		E7=1, E2A=1	0.53	0.83	0.65	
Proposed	Image	E7=0, E2A=1	0.83	0.87	0.85	0.84
		E7=1, E2A=0	0.94	0.76	0.84	
		E7=1, E2A=1	0.71	0.97	0.82	
Model A	Image, PRPD	E7=0, E2A=1	0.00	0.00	0.00	0.40
		E7=1, E2A=0	0.71	0.59	0.65	
		E7=1, E2A=1	0.24	0.95	0.38	
Proposed	Image, PRPD	E7=0, E2A=1	0.72	0.66	0.69	0.71
		E7=1, E2A=0	0.85	0.66	0.74	
		E7=1, E2A=1	0.54	0.96	0.69	
Model A	Image, PRPD, PDA	E7=0, E2A=1	0.00	0.00	0.00	0.16
		E7=1, E2A=0	0.00	0.00	0.00	
		E7=1, E2A=1	0.16	1.00	0.28	
Proposed	Image, PRPD, PDA	E7=0, E2A=1	0.72	0.67	0.69	0.64
		E7=1, E2A=0	0.92	0.49	0.64	
		E7=1, E2A=1	0.40	0.98	0.57	

Thus far in this chapter, a methodology using knowledge-assisted feature extraction and multimodal learning was proposed to perform fault detection and diagnostics of industrial systems in the context of sparse data. The performance of the proposed methodology was investigated in a real industrial case study of hydrogenerators. The obtained results highlight its effectiveness in overcoming the challenges of (1) time alignment between different types of CM data; (2) limited samples of some CM tools; and (3) different certainty levels of expert knowledge about the hydrogenerator physical degradation states derived from CM measurements. Particularly, the knowledge-assisted feature extraction phase in the proposed methodology plays a crucial role in exploiting more valuable information

from all CM modalities. In addition, the multimodal learning approach relying on the attention mechanism allows the feature-learning paths from certain CM modalities to be guided by other more reliable CM indicators. This mechanism enhances the efficiency of multimodal learning even when reliable CM indicators are not available. As a result, the proposed methodology can still predict the degradation states of hydrogenerators with an acceptable degree of accuracy, even when the reliable CM indicators are missing. These findings demonstrate the robustness of the proposed methodology in handling the sparse data issue.

The model presented in this section was designed for classifying CM measurements into degradation types, focusing solely on data from CM tools. Another key task in FDD is evaluating the intensity or risk level of degradation while a machine is in a given physical state. This evaluation is influenced by human judgment, introducing subjectivity into the calculation and providing an opportunity to include text data. The next section extends the model to quantify risk levels, enhanced by incorporating textual remarks made by personnel during inspections.

4.5 Extension of Methodology to Incorporate Text Data

In the previous section, we addressed the classification of degradation types from condition monitoring measurements. This section extends the methodology to quantify degradation levels using the same data, aiming to compute a health index (HI) value that reflects machine health from 0 to 100. While industry experts have established rule-based methods for this purpose, the subjective judgments of inspection personnel significantly influence the variables and parameters of the established algorithms.

To address this subjectivity, we integrate text data from inspection remarks and domain-specific documents into our multimodal diagnostics framework. This integration is designed to enhance the accuracy of HI calculations by incorporating the nuanced insights that textual data provide about machine health.

The health index starts at 100, indicating the onset of a new degradation state, and decreases as the machine's condition worsens, reaching 0 as it transitions to a more severe state. By leveraging both quantitative machine data and qualitative expert insights from text, the enhanced model not only classifies physical degradation states but also dynamically estimates the degradation level, bridging the gap between objective data and subjective expert evaluations in the diagnostic process.

The rest of this section is organized as follows. Subsection 4.5.1 outlines the text data

and its necessary preprocessing steps. Subsection 4.5.2 describes the methodology for incorporating text data into the classification model to enhance HI calculation. Finally, Section 4.5.3 details ablation experiments and presents the results to validate the components of the proposed method.

4.5.1 Technical text preprocessing

There are two primary sources of text data utilized in quantifying the health index. The first consists of structured documents such as proprietary guidelines, maintenance instructions, and industry standards that serve as a knowledge base for technicians. The second source is comprised of inspection notes and remarks by personnel, which often include brief, variably formatted descriptions with challenges such as domain-specific jargon, abbreviations, colloquial expressions, and noise-like typographical errors. These informal texts pose difficulties for standard NLP algorithms due to their lack of standardization and informal content.

Comme le groupe [redacted], le stator et les pôles du rotor sont très très sales. La saleté peut constituer un exemple à montrer aux gens lorsque l'on veut démontrer qu'un alternateur est sale. Il est donc impossible de visualiser s'il y a des points d'échauffement au stator et au rotor. L'isolation sur les connexions du rotor commencent à s'effilocher mais rien d'alarmant. Les collets d'entrefer et de jantes semblent encore solides. Il y a du gliptall qui se décolle sur les pôles, il y a énormément de saleté sur les pôles mais tout semble OK...

... À cet endroit j'ai pu voir 2 cales de descendu à midi et une autre à environ à 10hre. Elles ont été replacées mais je pense qu'elles vont redescendre rapidement. La descente de cale indique que le calage commence peut être à être "lousse".

Figure 4.28: Technician's remarks on a visual inspection including technical jargon and colloquial French

Figure 4.28 shows an example of an inspection note (in French) while the authors' English translation is provided in Figure 4.29. The photographs corresponding to this text from the visual inspection are given in Figure 4.30. As can be seen from these samples, the inspection notes provide essential insights into visual assessments, indicating their potential to bridge the gap from measurement to rating calculation. However, these notes, often in conversational French riddled with typos, grammatical errors, technical jargon, and colloquialisms, present significant challenges for traditional NLP methods. For example, the samples show that the text is written in colloquial French, as these notes are primarily

written for technicians to communicate with each other. The text contains common typos and grammatical errors, along with technical terms and jargon, like “noon, 10, 9, and 7 o’clock", used to describe positions on the circular stator. Additionally, technicians occasionally write words such as “lousse" in quotes to indicate a French-accented pronunciation of the English word “loose".

```
Like the [redacted] group, the stator and rotor poles are very very
dirty. Dirt can be an example to show to people when we want to
demonstrate that a generator is dirty. It is therefore impossible to
visualize if there are heating points on the stator and rotor. The
insulation on the rotor connections is starting to fray but nothing
alarming. Air gap collars and rims still seem solid. There is a
gliptall that is peeling off on the poles. There is a lot
of dirt on the poles but everything seems OK...

... At that place I could see 2 shims went down at noon and another at
around 10 o’clock. They have been replaced, but I think they will slip
back down quickly. The slipping of the shims indicates that the stall
is beginning maybe to be "loose".
```

Figure 4.29: English translation of technician’s remarks from Figure 4.28.

Figure 4.30: Photographs taken during a visual inspection. The photos show a high contamination level.

Conventional pre-Large Language Model (LLM) NLP methods struggle to effectively interpret texts laden with technical jargon and colloquial language. In contrast, LLMs, trained on a diverse array of textual data, offer a viable solution for such complex text analysis. In response, we propose fine-tuning an LLM on domain-specific knowledge to better contextualize this text.

Prior to deploying Large Language Models (LLMs), text data must undergo specific preparation steps that accommodate the processing of technical, non-English texts across various industrial settings, as demonstrated in our case study. The preparation processes differ for maintenance remarks and knowledge base documents due to their distinct structures and functions. Maintenance remarks serve as queries or prompts for LLMs, while knowledge base documents are formatted specifically for fine-tuning LLMs. These steps, while tailored for this case study, are designed to be broadly applicable to a range of industrial text processing applications.

1. **Text cleaning and formatting.** It is crucial to manage accented and non-standard punctuation in special characters carefully to prevent data loss and misinterpretation, with decisions on retention or removal based on the embedding model's capabilities. Additionally, addressing encoding issues from computerized maintenance management systems (CMMS) that cause garbled text and errors is essential for accurate data processing.
2. **Handling language-specific requirements.** Most text preprocessing methods are designed for languages using Latin characters, requiring different approaches for non-Latin scripts. Preprocessing colloquial French text, as in this study, introduces challenges such as the essential nature of special characters and accents (e.g., é, è, ê, ë) that cannot be discarded without losing meaning. General language tools often struggle with non-standardized industrial terminology, and while pre-trained language models typically favor English, French-specific models lack sufficient depth for industrial applications. Translating French to English can lead to significant information loss and errors, as most models are not trained on domain-specific jargon.

```
<p>2019-01-24 | [redacted] | Modifications à linspection
enregistrer par: [redacted].Les modifications suivantes ont étés
apportées :-
Développantes :
    MODIFIER Barres (bobines) - Présence de débris.  Cote précédente
    de 1,0.
    MODIFIER Barres (bobines) - Fissuration.  Cote précédente de 1,0.
    MODIFIER Barres (bobines) - Contamination (saleté). Cote précédente
    de 3,0.
Raison :Note: Groupe Horizontal ... Haut = Aval  &  Bas = Amont
Beaucoup d¿huile et saleté sur le côté amont sur le bobinage et
les pôles et barres
</p>
```

Figure 4.31: Text data from a CMMS, highlights data cleaning challenges such as markup tags and formatting issues.

Figure 4.31 demonstrates how non-standard punctuation, such as "¿" instead of an apostrophe, and typing errors from form field entries introduce noise that disrupts machine learning model processes like tokenization and parsing. Despite these textual issues, the note "Beaucoup d'huile et saleté sur le côté amont sur le bobinage et les pôles et barres' (English: A lot of oil and dirt on the upstream side of the winding and the poles and bars) provides essential degradation indicators. To mitigate these challenges, a preprocessing pipeline is needed that includes tokenizers for special characters, charset normalization, named entity recognition for removing irrelevant names, and bilingual embeddings to address language discrepancies between French and English-centric models.

3. **Document preparation and text extraction for fine-tuning.** While the previous steps are necessary to prepare the remarks for use as queries, the preparation of domain knowledge texts for fine-tuning requires some different steps. This involves converting relevant documents, such as norms and guidelines, into text files and removing unnecessary characters like bullet points. Extraction from formats like PDFs is a well-documented and straightforward process. However, effective cleaning and formatting are crucial as they organize the data into meaningful segments, facilitating the efficiency of subsequent processing steps.

4. **Chunking and tokenization.** The text must first be segmented into appropriately sized chunks to ensure semantic context is preserved for effective modeling. While maintenance remarks, typically brief, are easily split into chunks, determining the optimal chunk size for longer documents during fine-tuning depends on the model's sequence limits. Ablation experiments (Section 4.5.3) will later assess the impact of varying chunk sizes across different models using both rule-based methods and machine learning. After chunking, the text is then tokenized into discrete units, or "tokens", using techniques such as rule-based methods, finite-state transducers, or subword tokenization like byte-pair encoding (BPE). These tokens become the inputs for the embedding model.

4.5.2 Health index calculation methodology

In this section, we present a methodology to augment the diagnostics model with text data to enable health index calculation. It consists of three main steps:

1. Use domain knowledge texts to fine-tune LLM.
2. Use fine-tuned LLM to embed short text.
3. Use embedded short text to weight the fused inputs.

4.5.2.1 Step 1: Choosing and fine-tuning an LLM on domain knowledge

The experts observe the photographs from visual inspection and identify the degradation level based on their accumulated domain knowledge. In industry, such knowledge is often formalized in texts such as guidelines, standards, and so on. In our case study, we have access to documents such as ISO diagnostics standards, inspection guidelines, and detailed degradation severity calculation process explanations available within the industry. A large language model is trained on diverse text sources encompassing conversational, pedagogical, and other literary styles, making them approximate generalist humans. Just as a human would need to study and absorb domain-specific texts to become proficient, fine-tuning a language model to a particular domain is akin to this process. Therefore, the initial step involves selecting an LLM and fine-tuning it using the available domain-specific texts and documents.

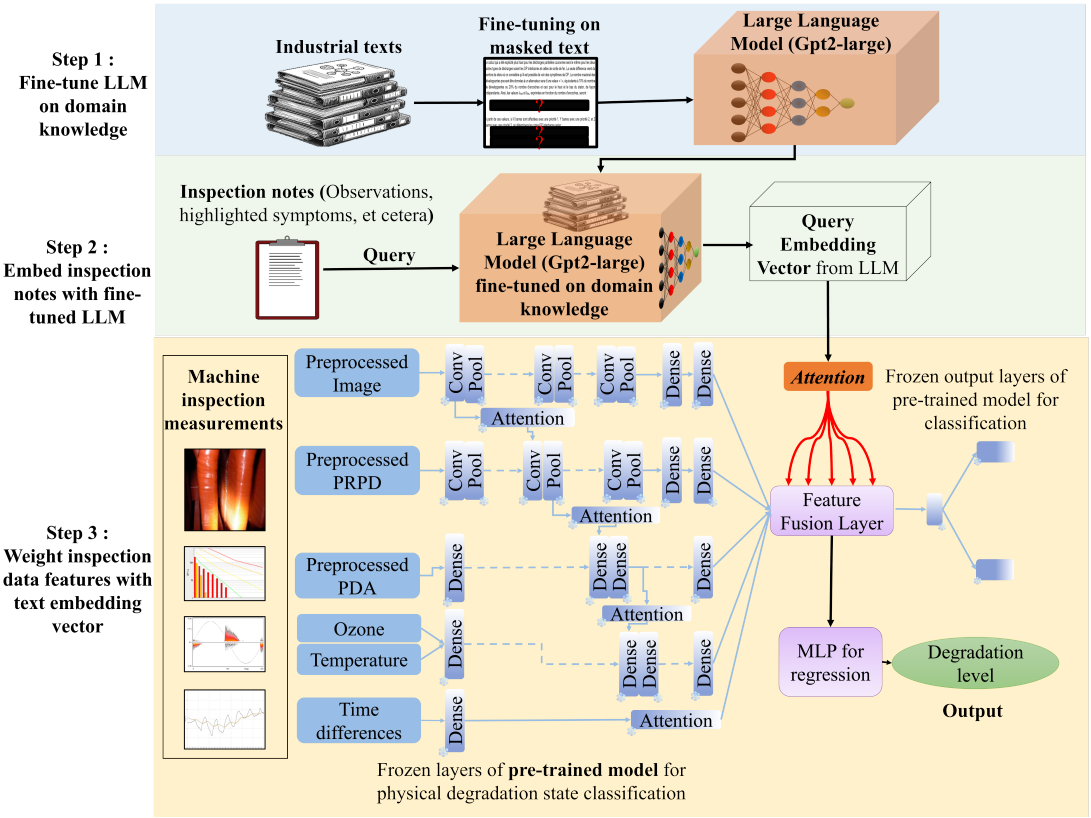


Figure 4.32: Overview of the proposed method to improve the performance of a machine degradation level calculation model with text data. The method involves fine-tuning an LLM on the industrial text documents, using the fine-tuned LLM to embed the notes written by technicians on an inspection of the machine and then using the embedded inspection notes to attention-weight the inspection measurements, and passing this to an MLP for computing the machine's degradation level.

When selecting an LLM, specific criteria must be considered. Firstly, the LLM should possess language capability in the industry's operating language. For industries operating predominantly in English, numerous English LLMs are accessible. However, options are more limited for languages other than English, although open-source multilingual LLMs are available for languages like French. It is important to note that if an open-source LLM is not available for the language relevant to the application case, implementing this methodology may prove to be excessively challenging. While training an LLM is an option, it may incur unjustifiable costs. Expanding the language capabilities of a multilingual LLM is also a possibility, albeit beyond the scope of this study.

Once the language is accounted for, the rest of the choice is to strike a balance between performance and resource constraints. While models such as Mistral-7B (Jiang *et al.* (2023)), Llama-2 (Touvron *et al.* (2023)), and others demonstrate excellent capabilities, they are beyond the computational resource limits. Thus, gpt2-large was chosen (Radford *et al.* (2019), Ethayarajh (2019)) .

Masked language modeling (Sinha *et al.* (2021)) is a technique typically used in model pre-training. It is a self-supervised training method to familiarize a model on a text corpus. In simple terms, the training is done by masking certain parts of the text and training the model to predict the masked content. While typically applied during pre-training and followed by task-specific training, this study employs it to fine-tune an already trained LLM with domain-specific texts.

The fine-tuned model is expected to demonstrate expertise in the subject matter covered by the texts. However, it remains uncertain whether this expertise genuinely signifies a profound comprehension or simply results from the model's ability to predict the most likely next word. This distinction lies beyond the scope of our investigation. The primary goal is to measure any improvements or potential declines in the performance of the LLM when it is fine-tuned using pertinent text data. This training method operates by covering up certain portions of the input data randomly and then tasking the model with predicting the hidden part.

4.5.2.2 Step 2: Embedding inspection notes using fine-tuned LLM

In this step, the goal is to get a vector representation for the inspection notes that is as close to the real meaning within the context as possible. Thus, a domain-specific knowledge fine-tuned LLM is more adept at incorporating short text within a suitable context compared to a language model lacking such expertise or utilizing generic embeddings.

LLMs are typically configured for a task, most commonly language generation. In this step, the layers optimized for such outputs are ignored. Instead, the inspection note is

input to the LLM as a query, and the input embedding created by the LLM is extracted as a matrix.

This step is similar to the retrieval step of retrieval augmented generation (RAG) (Cai *et al.* (2022)). However, it differs in that it does not necessitate the retrieval of context from domain knowledge. In the subsequent section, we will elucidate the rationale behind the omission of retrieval in this methodology. However, once the query embedding is obtained, retrieving pertinent context becomes straightforward. During the project's developmental phase, text retrieval related to queries was experimentally examined as a rudimentary validation of the embedding.

There is no exact method to test the quality of the query embedding. For domain-specific use cases, embedding benchmarks are not available. Therefore, the quality of the embedding is transitively evaluated by the accuracy of the degradation level calculation in the next step.

4.5.2.3 Step 3: Use embedded short text to weight monitoring data.

The final step aims to quantify the degradation level of machinery based on inspection data, emphasizing the strategic use of inspection notes. Typically, these notes are not direct measurements but expert observations summarizing key insights from various data sources. Personnel drafting these notes possess a deep understanding of the degradation evaluation process, making these texts crucial for accurate degradation level assessments.

Our method innovatively integrates these inspection notes into the model, not merely as another data source but as a means to enhance feature weighting derived from inspection data. Trained technicians author these notes, and they are subsequently processed through a LLM that has been fine-tuned with industry-specific knowledge. This process involves creating vector representations of the inspection notes. These vectors then affect the weighting of significant features in the diagnostic model we previously developed in section 4.3.4. By utilizing the detailed insights from the inspection notes, this model allows refining the diagnostic outputs.

For multimodal data-driven diagnostics, our proposed architecture incorporates feature extraction for each data source, leading to a fusion layer. This layer, selected based on domain knowledge and positioned near the output, optimally integrates features from the condition monitoring data. The inspection notes, vectorized by a domain-knowledge fine-tuned LLM, enhance this layer by weighting the most relevant features for degradation level assessment. This setup allows the fine-tuned LLM to act as a bridge linking raw data features and expert insights from inspection notes, effectively enhancing diagnostic accuracy as depicted in Figure 4.32.

4.5.3 Health index calculation results

The proposed methodology contains three main elements: 1) text embedding with an LLM, 2) fine-tuning the embedding model on domain knowledge, and 3) the text input mode for the feature fusion (direct or weighted by the attention). To evaluate the efficacy of each element in our methodology, we conducted a series of ablation experiments. These contain comparisons between a small Word2Vec-like embedding model (*FrWac2Vec* (Fauconnier (2015))) and Gpt2-large, fine-tuned and without fine-tuning, and direct and attention weight input modes for both. The architectural setups for all the experiments are shown in Appendix D (Figures D.1 and D.2). The different experiments are listed as follows:

1. **No text input:** Modified the output layer of the existing diagnostics model to perform a regression task targeting the degradation level. This modification involves retraining only the model's final output layer. This initial experiment sets a baseline for estimating degradation intensity using only the quantitative data collected during inspections, without any textual annotations from technicians (Figure D.1).

2. ***FrWac2Vec*(text) + no fine-tuning + direct input:** Embedded the text data (technician's remarks) using a small off-the-shelf model for French text embedding model (*FrWac2Vec*) and added it as an additional input. Only the technicians' notes were used, excluding other text data, e.g., guidelines for technicians, forming the second baseline This experiment explores the minimum performance enhancements from adding text remarks (Figure D.2a).

3. ***FrWac2Vec*(text) + no fine-tuning + attention weight input:** The notes are embedded using *FrWac2Vec* without any fine-tuning. Instead of providing the text as a direct input, it is used to weigh the features derived from other inputs. This experiment explores the assumption that the text primarily offers observations related to other measurements (Figure D.2b).

4. ***FrWac2Vec*(text) + fine-tuning + direct input:** The small embedding model *FrWac2Vec* is first fine-tuned on industrial text documents such as guidelines and standards. The inspection notes are embedded using this fine-tuned model and provided as a direct input. This explores the value of providing context for embedding the inspection notes (Figure D.2c).

5. ***FrWac2Vec*(text) + fine-tuning + attention weight input:** This experiment combines both fine-tuning the small embedding model *FrWac2Vec* and using the embedded inspection notes to attention weight other measurements (Figure D.2d). This concludes the experiments with the small embedding model.

6. **Gpt2-large(text) + no fine-tuning + direct input:** Here, the first embedded inspection notes with an LLM (Gpt2-large) is attempted for the first time. The LLM is used without any fine-tuning, and the embedded notes are added directly as input. Given that most open-source LLMs are trained on diverse, general text data from the internet, this study aims to determine whether an LLM trained on such a broad corpus can enhance the extraction of valuable information from industrial text data (Figure D.2c).

7. **Gpt2-large(text) + no fine-tuning + attention weight input:** Here, the inspection notes embedded by Gpt2-large is used to weight the other data features (Figure D.2b).

8. **Gpt2-large(text) + fine-tuning + direct input:** In this experiment, Gpt2-large is fine-tuned on internal company documents including standards, norms, and guidelines. The aim is to examine the effects of fine-tuning an LLM to specific contextual needs. The inspection notes embedded by the fine-tuned LLM is then introduced as an additional data source (Figure D.2c).

9. **(Proposed method) Gpt2-large(text) + fine-tuning + attention weight input:** This final setup brings together all the elements of the proposed methodology. The LLM (Gpt2-large) is fine-tuned on the documented domain knowledge, the inspection notes are embedded using this fine-tuned LLM, and this is used to weight other condition monitoring data features (Figure D.2d, 4.32).

Table 4.4: Results comparison.

Experiment	Embedding	Text input mode	MAE
1	Baseline: No text data	–	44.2
2	frWac	Direct	31.1
3	frWac	Weight	27.4
4	Fine tuned frWac	Direct	26.1
5	Fine tuned frWac	Weight	24.4
6	Gpt2-large	Direct	15.7
7	Gpt2-large	Weight	10.1
8	Fine tuned Gpt2-large	Direct	14.6
Proposed	**Fine tuned Gpt2-large**	**Weight**	**4.2**

The results on the test set are synthesized in Table 4.4, which displays the performance of nine setups in estimating the health index, a numerical value ranging from 0 to 100. The mean absolute error (MAE), a suitable metric for this range, is reported in the table's final column. The initial attempt to estimate the HI without using text resulted in a mean absolute error of 44.2, indicating limited utility of the baseline model. This could be

because the subjective element plays a significant role in HI calculation, and monitoring measurements alone cannot account for this. Table 4.4 shows that the proposed method achieves an MAE of around 4, which demonstrates a significant improvement over the baseline. The prediction plots and comparative analysis of results from the ablation study are given in Appendix D.

Key findings

The experiments with text yield the following observations:

- Text data improves diagnostics tasks with an inherent subjectivity, such as degradation level estimation.
- It is more effective to use text data as a separate entity than other condition monitoring measurements. Using text data as an observation on the measurements rather than an additional measurement yields better results.
- Embedding inspection notes within proper context returns significant improvements to performance. Fine-tuning an LLM on domain knowledge documents is an effective way to develop a suitable embedding model.

4.6 Conclusion

In this chapter, we first developed a methodology to perform fault detection from a multimodal dataset with challenges such as sparsity, time alignment conflicts, and varying data collection rates. The methodology leveraged and improved upon the multimodal learning techniques introduced in Chapter 3. We applied the methodology to a real-world dataset and, through several ablation experiments, it demonstrated its effectiveness in handling realistic data conditions. Then, we extended the model to perform health index calculation by incorporating inspection notes, assisted by a large language model fine-tuned on industrial text documents. This model performed remarkably well on a task with a high level of human judgment inherent to it.

Findings from this chapter show that the cross-modal attention methodology is highly robust to missing data conditions in industrial data and can help mitigate time alignment conflicts. This chapter also demonstrated that aligning the diagnostics pipeline with domain knowledge can improve the multimodal model's capacity to handle varying monitoring rates and data sparsity. Furthermore, it highlighted the value of treating text and other monitoring data differently, showing that using textual observations to weight other

data features can improve diagnostic model performance. All results and findings were validated in close consultation with a domain expert.

This chapter synthesized techniques and approaches that will recur throughout the thesis. Firstly, the development of the diagnostics model followed an approach of creating standalone neural networks optimized for small tasks, introducing modularity to the multimodal methodology. For instance, the object detection FRCNN model for detecting partial discharge from images can later be replaced by a better or larger model. Secondly, as demonstrated by the feature extraction models, extracting near-output layer features from a model trained on a pretext task is a valuable tactic when handling multimodal data where individual modality treatment is non-trivial and may require dedicated expert models. Thirdly, extending the classification model to perform a regression task for health index calculation by freezing pre-trained layers is a simple yet effective technique that facilitates fine-grained modifications within a broader framework.

Building on the methodology and techniques developed in this chapter, the next chapter will expand the diagnostics to a wider scope, aiming to perform prognostics. We will develop a methodology for forecasting future health states using multimodal data with severe imbalance and sparsity and apply it to a real-world dataset. The next chapter will bring together all the principles and techniques developed throughout the thesis in a capstone project on prognostics.

Chapter 5

Prognostics with Multimodal Graph Forecasting

Contents

> *"Data-driven predictions can succeed — and they can fail. It is when we deny our role in the process that the odds of failure rise. Before we demand more of our data, we need to demand more of ourselves.*
>
> — Nate Silver, in *The Signal and the Noise.* (Silver (2012)).

5.1 Introduction

In Chapter 4, we developed a methodology for fault detection and diagnostics from multimodal condition monitoring data and applied it to an industrial dataset from hydrogenerators, addressing data sparsity, time alignment conflicts, and varying data collection rates. We also extended the methodology to incorporate inspection notes to perform health index calculations.

The previous chapter presented and discussed the various challenges of industrial data within the realm of diagnostics. While challenging in terms of training a model to extract proper features, diagnostics is a task to explain the past, which means it is possible to validate a model against reality. Prognostics, which deals with the future, is a much more challenging task in the industrial context. In this chapter, the challenge of forecasting future health evolution given the current and historical health records of a machine will be addressed. An end-to-end methodology from multimodal condition monitoring data to future health forecasts will be developed. This will be applied to the data from hydrogenerators, with the scope extended to include more fault types.

The main objective of the methodology is to perform prognostics even without any run-to-failure (RTF) data, which is a challenge that arises in industrial scenarios due to many practical reasons. The methodology developed in this chapter will address this by first doing diagnostics classification on the condition monitoring data, extracting and projecting the diagnostics features on a 2D space, and tracing the degradation of machines on this plane to create RTF sequences from many machines. In the course of realizing this, the methodology will also need to tackle several additional challenges to obtain accurate diagnostics. Finally, RTF sequences generated in the form of graphs can be used to train a graph prediction model that gives machine health prognostics.

The rest of this chapter is organized as follows. First, section 5.2 will present some literature and motivate the need to develop a methodology to address the lack of RTF data in the industry. Then, in section 5.3, we will present the methodology for machine health prognostics that overcomes this limitation. Section 5.4 will present the extended industrial context and apply the methodology. Section 5.5 presents the prognostics results and section 5.6 concludes the chapter.

5.2 Motivation and Context

The PHM cycle involves data acquisition, data processing, detection, diagnostics, prognostics, decision, and finally intervention (Medjaher *et al.* (2013)). In this sequence, although the activity leading up to diagnostics is far from trivial, the challenges associated with prognostics are even harder to surmount (Soleimani *et al.* (2021), Zio (2022)). Indeed, forecasting a machine's health relies on the assumption that historical data can inform future evolution, requiring sufficient data to represent all evolution possibilities. Ideally, multiple sequences of condition monitoring data from deployment to breakdown (RTF data) would be available, covering all degradation mechanisms and trends. However, in practice, such data is rare. Decision makers rarely allow expensive or critical machines to run to failure without maintenance, and condition monitoring usually does not focus on acquiring a comprehensive dataset for model training.

The PHM literature is rapidly evolving, yet most new methods are tested on clean, simulated datasets, leading to a gap between research outcomes and practical applicability in industrial settings. Industrial data often fail to meet the requirements for advanced machine learning techniques. While deep learning has proven viable for forecasting machine health states with sufficient data (Wen *et al.* (2022)), practical constraints frequently hinder the collection of such datasets. High-value machinery is rarely run to failure, and condition monitoring is sporadically conducted due to costs, which delays data collection post-maintenance and complicates impact assessments.

Recognizing the challenges in obtaining complete RTF data, Kim *et al.* (2020) introduced the DAPROG model, which uses dynamic time-warping for data augmentation from similar systems. Wang *et al.* (2008), Wang (2010) proposed trajectory similarity methods to align current machine data with historical trajectories. However, these techniques require some existing RTF data and are limited in settings with no available trajectories, particularly for high-cost or safety-critical machinery.

Observing that the acquisition of uninterrupted RTF trajectory data is impractical, this study proposes a methodology to construct a comprehensive prognostics dataset by integrating RTF trajectory fragments from various machines. This approach aims to make the application of neural network methods in industrial prognostics more feasible. Using a feature transfer method to address the issue of imbalanced diagnostic data has been proposed by Lu and Yin (2021), and Islam *et al.* (2023) has demonstrated improvements in regression tasks through the analysis of intermediate features. However, to our knowledge, this is the first attempt to transfer in-domain RTF trajectory fragments by analyzing feature space proximity, providing a novel contribution to the field.

To realize this trajectory construction method, it is necessary to perform diagnostics

classification on condition monitoring samples. Some challenges of diagnostics were addressed in Chapter 4. However, the classification task in the previous chapter targeted two health states of similar risk levels without the long-tail data distribution common in prognostics. Prognostics data are often label-imbalanced since machines are usually maintained in healthy states, making faulty state samples rare. Conversely, for long-lifetime machines, early-state measurement tools might have been developed later, distorting data distribution. Beyond the classification challenges in Chapter 4, prognostics must address label imbalance in multimodal multilabel datasets within the PHM domain. This chapter will also tackle multiple diagnostic and prognostic challenges. We develop a robust prognostics methodology that relies on highly accurate diagnostics models. Once diagnostics training is optimized, its features can help develop an RTF dataset for efficient prognostics.

Addressing the limitation of RTF data availability through feature space continuity analysis, this study generates RTF data within a graph structure. Consequently, this enables the application of graph neural network (GNN)-based prognostics modeling methods. In the literature, Li *et al.* (2022b) reviewed GNN methods for diagnostics and prognostics, highlighting graph generation methods based on sample similarity measures. GNN's potential in modeling non-Euclidean relationships is emphasized. Xu *et al.* (2024) proposed a knowledge-integrated GNN for performance prognostics. Ding *et al.* (2022) introduced an encoder-decoder meta-learning method for limited and variable-length data prognostics. Ding *et al.* (2024) developed a few-shot prognostics method using spatio-temporal sequences. Li *et al.* (2021) proposed a hierarchical GNN considering spatial sensor relationships, while Li *et al.* (2022a) used graph features for sensor interaction explanations. Zhang *et al.* (2023) and Wei *et al.* (2023) employed graph attention networks (GAT) and graph convolutional networks (GCN) for multi-sensor feature learning, respectively. However, our review revealed no method that considers GNN-based prognostics with physical degradation states as graph nodes and state transition times as its edges.

In this light, the proposed methodology aims to achieve the following objectives:

- During the diagnostic phase:

    1. Tackle the challenge of imbalanced industrial condition monitoring data, which often results in model bias.
    2. Address the prohibitive computational costs traditionally associated with training neural network models for reliable diagnostics from multimodal data.
    3. Enable the addition of new output classes without retraining the entire model, overcoming a common scalability issue.
    4. Mitigate the limitations of an entirely opaque "black box" model by preserving interpretability within the large model framework.

- During the prognostic phase:

    1. Develop a diagnostics feature similarity-based method to construct comprehensive RTF trajectories from fragmented monitoring data collected across different machines. This methodology would leverage similarities in feature spaces to interpolate missing data segments, enabling more complete datasets for training prognostic models.
    2. Propose a graph structure for modeling the evolution of degradation states, with physical significance embedded into the nodes and edges. This structured representation would facilitate a more intuitive understanding of machine deterioration processes and their interactions at different stages of machine life cycles.
    3. Present a graph-masked autoencoder architecture tailored for machine health forecasting. This architecture would apply graph neural networks to predict future health states by learning from the constructed RTF trajectories, ensuring effective generalization from partial to full system states.

5.3 Methodology for Prognostics Using Incomplete Run-To-Failure Data

This section presents a methodology to perform machine health prognostics from incomplete run-to-failure data. The goal is to compile a prognostics dataset by integrating partial trajectories derived from condition monitoring across a fleet of machines. Additionally, the methodology seeks to demonstrate the feasibility of constructing a prognostics dataset without a complete end-to-end trajectory from any single machine. The methodology requires some assumptions to be held, listed as follows:

- The symptoms of the machine can be discretized into physical degradation states that are distinguishable from each other, even though multiple states may be active at any time.
- The collective condition monitoring data from the machine fleet contains at least one activation of each fault type.
- There is sufficient domain knowledge available to prepare a diagnostics dataset for supervised training, with a tolerance for uncertainty.

Provided these criteria are met, the presented methodology can address the lack of complete RTF sequence from any one machine and facilitate prognostics, as illustrated in

Figure 5.1. It shows a two-part methodology, where the first part involves understanding the discrete health states that a machine undergoes through its degradation and developing a neural network model to classify condition monitoring data to active fault (health) states - that is, diagnostics. As the first step involves understanding the condition monitoring data for diagnostic classification, the methodology starts with a domain study. This informs the inputs, structure, and outputs of a diagnostics model and assists in the preparation of a dataset for training a diagnostics model. The diagnostics model must also be trained under severe class imbalance and data sparsity.

The second part involves extracting the intermediate features from the diagnostics model and projecting on a 2D space, analyzing the feature plot with an expert to validate its coherence to degradation evolution. Then, it constructs graph format RTF trajectories by connecting fragments of RTF from different machines based on feature proximity in this space. This would produce an RTF dataset for the whole fleet, which can be used to train any forecasting model. The rest of this section explains the steps in detail.

5.3.1 Domain study and preliminary data analysis

This section outlines the first step of the proposed methodology. To support diagnostics, domain study and data analysis include three crucial exploratory data analyses and stratification of multilabel multimodal data. This should be done in line with expert-informed identification of the physical degradation states that the machine could evolve through, the associated condition monitoring data, and the symptoms indicating each state. For prognostics, this step involves establishing the failure propagation mechanism in the form of a graph from the initial condition to failure modes.

5.3.1.1 Expert-assisted domain study

The initial step in developing a data-driven diagnostics model involves understanding the degradation processes of the machinery, necessitating collaboration with domain experts. This partnership is crucial as industrial condition monitoring data are seldom pre-labeled or structured optimally for model training, often featuring ambiguous health labels (Pei *et al.* (2022)). Consequently, the expertise of industrial professionals becomes indispensable in assigning accurate class labels and identifying specific degradation types from varied data sources, thereby significantly informing the model design, as described in Chapter 4.

Additionally, the diagnostic value of data from different sensors often unequally contributes to identifying each type of degradation. Some sensors may exhibit more pronounced symptoms for certain degradation modes than others. Recognizing the most

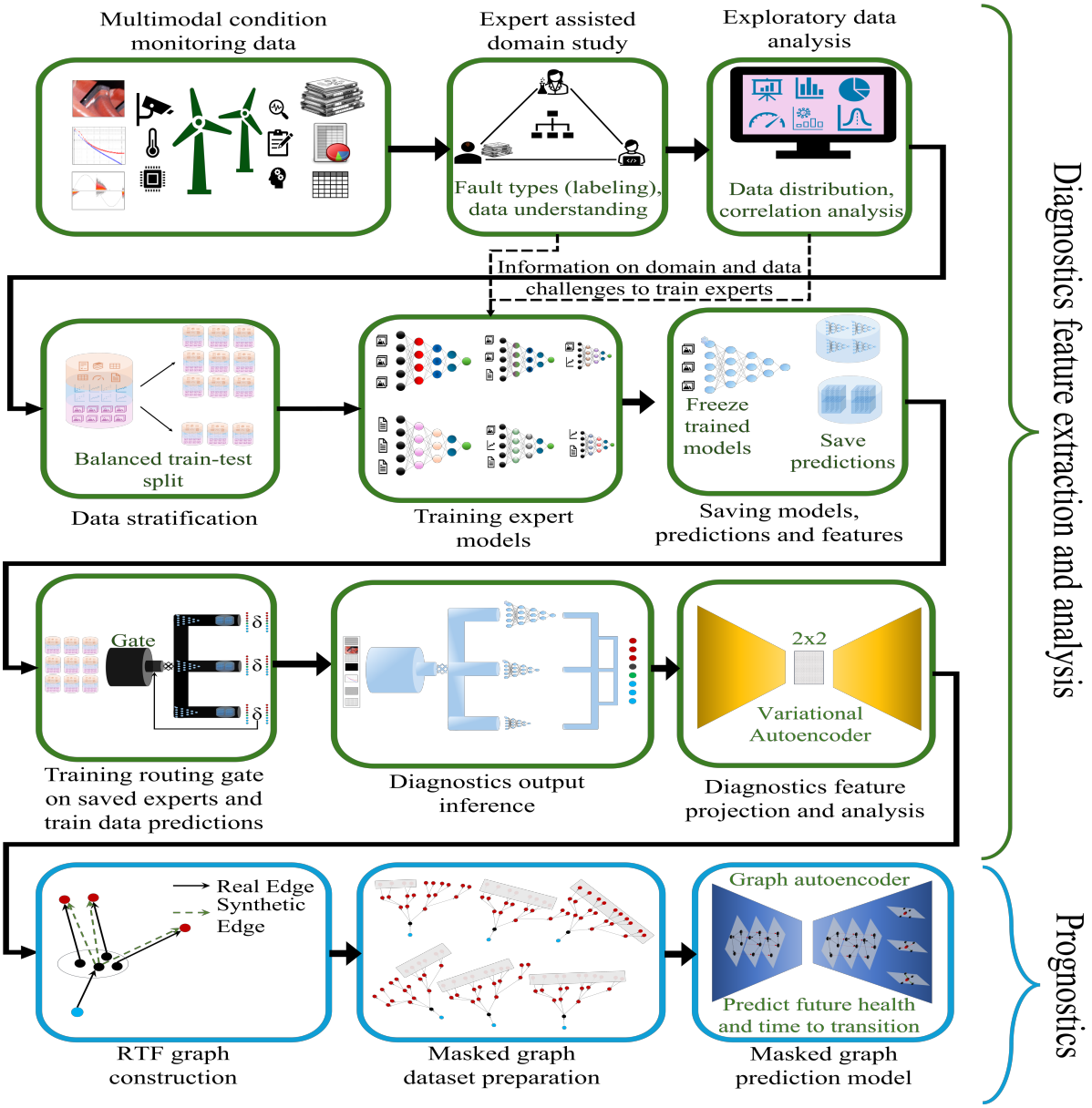


Figure 5.1: Prognostics methodology from condition monitoring data to future health prediction. The first part involves diagnostics and feature extraction, and the second part involves RTF dataset construction from the diagnostics feature space.

expressive sensors is vital for subsequent analytical phases, especially in devising a stratification method that ensures an equitable distribution of information quality across data strata, which is in turn necessary to address imbalances.

5.3.1.2 Exploratory data analysis

Once the input data are understood and target labels are defined, exploratory data analysis (EDA) is conducted to uncover the dataset's characteristics, identifying key learning bottlenecks. This analysis, critical within data science, provides the foundation for the subsequent phases of the methodology (Camizuli and Carranza (2018)).

The EDA process initiates with an examination of label distribution, crucial for detecting data imbalances that could skew the model's performance towards more frequent degradation types, potentially neglecting rarer but critical faults (Zhang *et al.* (2022)). This analysis of label distribution and imbalances informs the necessity for resampling or training adjustments to ensure comprehensive class representation.

Further analysis includes studying label correlations within a multi-label dataset, which reveals interactions between different fault types and helps understand complex fault interdependencies in industrial systems (Tarekegn *et al.* (2021)). This step is vital for refining data stratification strategies for dataset splitting, ensuring the training and validation reflect real-world fault co-occurrence scenarios. A label co-occurrence calculation method is given in Appendix A (algorithm 5).

Lastly, dimension reduction techniques like PCA and t-SNE are employed not to reduce computational costs but to understand the data distribution space. Visualizing data in reduced dimensions helps identify clusters and patterns, informing model design in later stages. These insights from EDA directly guide the data stratification process, crucial for the effective training of the diagnostics model.

5.3.1.3 Stratification of multilabel multimodal data

This section introduces a new stratification algorithm designed for the complexities of multimodal multilabel datasets. Stratification involves dividing a dataset into subsets, called strata, that reflect the original distributions to ensure that training, cross-validation, and testing sets accurately represent the full dataset. This is a crucial step to handle imbalances in the dataset.

Addressing the challenge of multimodal datasets, where modalities vary in informativeness for each class label, this algorithm ensures class informativeness across strata.

Although multilabel dataset stratification is discussed in existing literature (Sechidis *et al.* (2011)), the distribution of modalities within strata remains underexplored. The proposed algorithm balances label distribution and aligns modality distributions using the Earth Mover's Distance (EMD) (Panaretos and Zemel (2019)) for modality alignment, coupled with a label presence scoring system for equitable dataset partitioning. The comprehensive stratification strategy is detailed in Algorithms 2 and 3.

Algorithm 2 Multilabel Multimodal Dataset Stratification: Part 1

Initialize desired counts for each subset and label-modality pair based on D, I, and DMD.
Initialize weights $w_1 = 0.5$ and $w_2 = 0.5$.
for $j \leftarrow 1$ **to** k **do**
    $c_j \leftarrow |D| \times r_j$ ▷ Desired number of examples in each subset S_j
for each label λ_i in L **do**
    $D_i \leftarrow \{(x, Y) \in D : \lambda_i \in Y\}$ ▷ Instances with label λ_i
    for $j \leftarrow 1$ **to** k **do**
        $c_j^i \leftarrow |D_i| \times r_j$ ▷ Desired number of examples of label λ_i in subset S_j
end Part 1. ▷ Proceed to Part 2 for the detailed stratification process

Algorithm 2 begins by establishing target numbers for each subset based on the total dataset (D), the modality informativeness table (I), and the desired modality distribution (DMD). Weights w_1 and w_2, both initialized to 0.5, equally balance label presence and modality distribution alignment in the stratification process. These weights can be adjusted if the dataset shows greater label imbalance or modality sparsity, enhancing w_1 or w_2 accordingly to accommodate specific disparities. For each subset S_j, the algorithm calculates c_j, the desired number of instances, by multiplying the total instances by the proportion r_j. This is also done for each label λ_i within the dataset to determine c_j^i, the target number of instances per label in subset S_j, ensuring representation mirrors overall dataset composition.

In Algorithm 3, the first step involves calculating the current modality distribution (CMD) in each stratum for every subset and label, setting a baseline for alignment with the DMD. Subsequently, the EMD between CMD and DMD for each label and subset is calculated. This measure, quantifying the alignment effort, guides the optimal placement of instances by minimizing this distance.

The proposed stratification process optimizes two criteria: label representation and modality distribution alignment. Thus, a combined score for each candidate instance in every subset is calculated by integrating normalized Label Presence Score ($NormLPS$) and normalized EMD score ($NormEMDS$), weighted by w_1 and w_2. The c_j^i value from the first part helps derive LPS, with normalization scaled between [0,1] based on data

Algorithm 3 Multilabel Multimodal Dataset Stratification: Part 2

while $|D| > 0$ **do**
  Calculate $CMD_{\lambda_i,j}$ for each label λ_i in each subset S_j.
  for each label λ_i in L **do**
    Distribution score $(CMD, DMD) = EMD(CMD_{\lambda_i,j}, DMD_{\lambda_i})$
  Select label λ_l with the highest need based on the lowest EMD scores and the fewest remaining examples.
  $D_l \leftarrow \{(x, Y) \in D : \lambda_l \in Y\}$
  for each $(x, Y) \in D_l$ **do**
    for each subset S_j **do**
      Calculate normalized $NormLPS_{x,j}$ and $NormEMDS_{x,j}$.
      $CombinedScore_{x,j} = w_1 \cdot NormLPS_{x,j} + w_2 \cdot (1 - NormEMDS_{x,j})$
    $M \leftarrow \arg\max_j(CombinedScore_{x,j})$ ▷ Subset with best score for this example
    if $|M| = 1$ **then**
      $m \leftarrow M$
    else
      $M' \leftarrow \arg\max_{j \in M}(c_j)$ ▷ Further prioritize by largest number of desired examples
      if $|M'| = 1$ **then**
        $m \leftarrow M'$
      else
        $m \leftarrow$ randomElementOf(M')
    $S_m \leftarrow S_m \cup \{(x, Y)\}$ ▷ Add the example to the selected subset
    $D \leftarrow D \setminus \{(x, Y)\}$ ▷ Remove the example from the original dataset
    Update $CMD_{\lambda_i,m}$ for S_m, adjust desired counts for λ_i and modality distribution.
return $S_1, ..., S_k$

available at each iteration step. This combined score, ensuring that both label diversity and modality alignment are considered during instance placement, is calculated by:

$$CombinedScore_{x,j} = w_1 \cdot NormLPS_{x,j} + w_2 \cdot (1 - NormEMDS_{x,j}) \tag{5.1}$$

Finally, instances are allocated to the subset where they achieve the highest combined score. Following each allocation, the CMD for each affected subset is updated to ensure decisions reflect the most current distribution information.

This stratification process repeats until all instances are distributed, forming strata $(S_1, S_2...S_k)$ that preserve the original dataset's label diversity and modality characteristics. The algorithm strives to optimize label representation and modality distribution within each stratum. Once the dataset is stratified, some strata are designated as the test set,

while the remainder form the training and validation sets. With the completion of data preparation, the methodology progresses to the development of the diagnostics model.

5.3.2 Diagnostics with mixture of experts architecture

This section involves the development of a mixture of experts-based architecture to address the diagnostics challenges presented in Section 5.1. It includes four steps from training expert models to feature aggregation and inference.

5.3.2.1 Expert models

The diagnostics strategy involves developing a classification model by identifying the necessary expert models, informed by EDA results from Section 5.3.1.2. The analysis reveals data imbalances and correlated labels, pinpointing data bottlenecks where specific expert training is essential. For instance, overrepresented classes may bias the model towards false positives, while underrepresented ones may lead to false negatives. Addressing these issues requires training distinct expert models for severely imbalanced classes and for those with strong label correlations.

Definition

Definition 5.1 (Expert Model)**:**
An ***Expert Model*** E_i *is a neural network parameterized by weights* ϕ_i*, designed to focus on a specific task* τ_j*. Tasks* τ_j *may involve distinguishing between highly correlated classes or specializing in a single data modality. Multiple expert models* $\{E_i\}$ *can be specialized on the same task* τ_j*.*
Each expert model is trained on a data subset curated for its respective task τ_j*. All expert models* $\{E_i\}$ *follow a multi-branch, late-fusion structure with modality-specific feature extraction paths and maintain uniform input and output dimensions to streamline model training and integration.*
Each expert model E_i *undergoes individual training until convergence, followed by model validation using a subset of its training data.*

Each dataset for training is formatted uniformly, with input columns for all modalities and target columns for all classes, regardless of an expert's specific focus. This standardization simplifies subsequent steps and ensures consistency in dataset handling. For instances lacking specific modalities, null matrices are used as placeholders. Neural network layers are tailored to the data corresponding to each task, potentially utilizing pre-trained models for feature extraction. This ensures uniform input and output dimensions across all

experts, streamlining model training and integration. Training durations vary based on dataset size, data complexity, and neural network dimensions, ensuring each expert model is finely tuned for its diagnostic task.

5.3.2.2 Generating and storing predictions

Upon training completion, loading each expert model into memory to generate predictions across the entire training dataset is a critical step that significantly impacts computational resource costs. These predictions are saved to create a uniform input-prediction dataset, which is used to train the gate in the next step. This approach ensures data format standardization across all models, as each expert is designed with consistent input and output dimensions.

Additionally, it is advisable to save the feature vectors from the last layers of each model, typically from the fusion layer which has smaller dimensions, thus ensuring efficient storage use. While generating and storing predictions, extracting and saving features from these final layers is a straightforward process that enhances the dataset's utility for later stages.

5.3.2.3 Routing gate

A routing gate is a classification model that analyzes a data sample to identify potential challenges and accordingly chooses the experts most likely to accurately classify the sample.

Definition

Definition 5.2 (Routing Gate)**:**
A ***Routing Gate*** $g(.)$ *with parameters* ρ *is a classification model that selects the most suitable subset of experts* $\{E_i\}$ *for a given data instance. The gate assigns a score* α_i *to each expert* E_i *from the entire set of experts* $M = \{E_1, E_2, \ldots, E_{|M|}\}$*, where* $|M|$ *represents the cardinality of* M*. The output* $\boldsymbol{\alpha} = g(x, \rho)$ *is a continuous probability distribution where* $\boldsymbol{\alpha} \in [0, 1]^{|M|}$ *and* $\sum_i \alpha_i = 1$*.*

Although straightforward in concept, training the gate to effectively handle data imbalance and optimize expert selection is complex. It involves training the gate to output probabilities indicating each expert's likelihood of correct classification, taking into account label imbalances that may bias the gate's decisions towards certain experts.

To train the gate robustly, the stratification strategy described in Section 5.3.1.3 is utilized. It provides strata reflecting the original dataset's label and modality distribution.

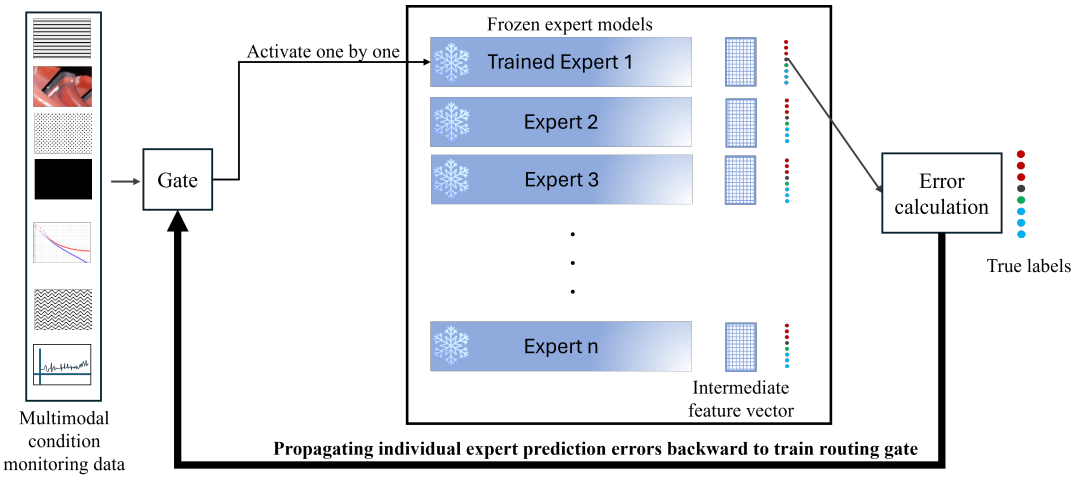


Figure 5.2: Training the gate to select from pre-trained experts for each data sample.

The gate undergoes k-fold cross-validation across these strata, allowing it to learn to select the appropriate experts accurately. Training the gate involves calculating discrepancies between the experts' predictions and the true labels for each sample. This training is efficient as it leverages pre-stored expert predictions, eliminating the need to reload expert models into memory, thus speeding up the process. An illustration of the gate training pipeline is given in Figure 5.2.

Various neural network architectures can be used as the gate, from simple feed-forward networks to more complex transformer-based models. Given their efficacy in handling multimodal data, transformer architectures with self-attention layers and a softmax output layer are recommended for the gate. The softmax output layer is necessary for assigning probabilities α_i to each expert's likelihood of making accurate predictions. As there is a potential overlap between the tasks of expert models, softmax is preferable over sigmoid function. The structural design of gate architecture consisting of a transformer stack followed by a softmax layer is given in Appendix B.

5.3.2.4 Feature aggregation and inference

After training the gate and setting the maximum number of expert models to be selected for each inference based on resource constraints, the next step is to train the feature aggregation module. This module aims to combine the features of the selected expert models effectively.

The feature aggregation module training is illustrated in Figure 5.3. First, each sample in the training set undergoes inference through the gate, which assigns probabilities to the experts. Activating all experts for each example significantly increases the computational

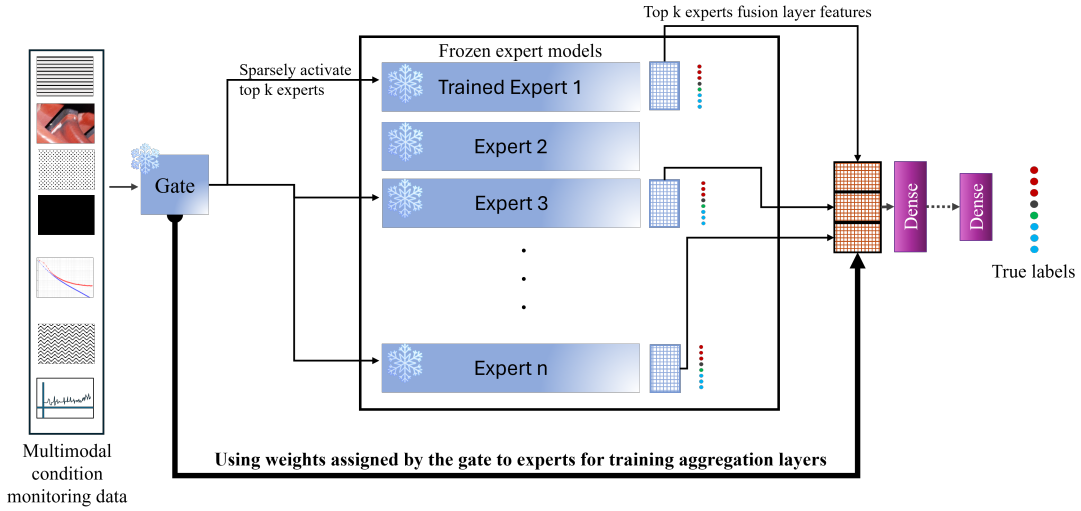


Figure 5.3: Illustration of training the feature aggregation and weight transformation of features collected from top k experts chosen by the gate. Here, the gate is already trained, and the learning happens in the aggregation layer. This is the last training step.

cost. To circumvent this, only a top-k of $|M|$ experts are activated, where $1 < k < |M|$. The output representations of the k active experts are averaged according to the respective routing weights, whose sum is re-normalised to 1. the new probability assigned to each expert i among the selected top-k is :

$$\alpha_i' = \frac{\alpha_i}{\sum_1^k \alpha_k} \tag{5.2}$$

These probabilities are normalized to ensure they sum to 1. The fusion layer features from each selected model are then retrieved from the saved files (Section 5.3.2.2). Subsequently, each feature vector f_i obtained from the expert E_i is multiplied by the weight α_i' assigned to the corresponding expert by the gate, reflecting the expert's confidence or expected relevance for the sample. Thus, the new output of the top-k experts is:

$$f_i' = \alpha_i'.f_i \tag{5.3}$$

This process, known as weight transformation or weighted sum, is common in ensemble learning and a mixture of expert models (Iqball and Wani (2023), Gong *et al.* (2023)).

Following the weight transformation, the transformed features f_i' are aggregated by the aggregation function φ. This can be achieved by summing the weighted feature vectors to create a single feature vector $\mathbf{f}$ as follows:

$$\mathbf{f} = \oplus f_i'|_{i \in top-k} \tag{5.4}$$

where $\oplus$ is the element-wise sum operator. Alternatively, the transformed features can be passed through dense layers or other transformations, such as self-attention layers, before reaching an output layer with sigmoid activation for multilabel classification. This approach enables the model to dynamically adjust the influence of each expert based on the input data, enhancing flexibility and potentially improving model accuracy.

Highlights

Modular deep learning based diagnostics: In the first part of the prognostics methodology, the objective is to perform diagnostics classification on condition monitoring data samples. Several challenges of multimodal multilabel data imbalances need to be addressed. A stratification algorithm, specialized expert models, and an expert selection gate work together to address these.

5.3.3 Diagnostics feature space analysis

The diagnostics model would take the raw condition monitoring data as input and classify the characteristics extracted from the data among the defined physical degradation states. The layers close to the output layer of this neural network should contain features that are most optimized to represent the physical degradation symptoms active in the input data. These features must be visualized and analyzed with the following steps:

1. Select a layer close to the output layer of the diagnostics model. This can be called the feature layer.

2. Input the training data samples to the diagnostics model, and save the feature layer output vector for each sample.

3. Train and save an encoder-decoder model (such as variational autoencoder (VAE) (Zemouri *et al.* (2019)) on the feature vector data.

4. Plot the 2D latent space of the encoder-decoder model.

Each point in the obtained plot will represent the physical degradation state of a machine at a time. The rest of the methodology depends on the quality of these features, that is, the accuracy of the diagnostics model. Therefore, steps 5.3.2 and 5.3.3 must be repeated until the visual analysis of the latent space meets the following requirements:

- R1: The physical degradation states are clustered, and there is a clear separation between the clusters.

- R2: There is a directional component in the arrangement of the degradation in terms of their degradation severity, such that the low-risk states are far from high-risk states, and the intermediate states are in between them.

- R3: Points very close to each other or coinciding on a 2D projection should be similar in terms of degradation risk level.

In the above list of requirements, R1 is primarily dependent on diagnostics model accuracy. So if the performance metrics of the model are high, the clusters should form in any dimension reduction method. It is expected that there is some information loss in the dimension reduction, but this is a necessary step for the methodology. An encoder-decoder model that can be saved and restored is preferred over stochastic methods such as t-SNE, for reusability. If R1 is met, then R2 usually depends only on the 2D projection of the feature space.

Finally, R3 is the most important requirement which forms the foundational assumption on which the methodology is based. The similarity of degradation risk level between the two data samples must be validated by domain experts. To ensure an unbiased assessment, it is advisable to conduct a blinded randomized experiment. In this experiment, domain experts are presented with pairs of data samples, which may vary in type and represent either similar or distinct symptoms of degradation. Selection of these pairs from the feature space plot should be strategic, encompassing pairs that are distantly separated, closely adjacent, and overlapping. Experts are then tasked with evaluating the relative risk levels of these pairs without access to the feature space plot. This method isolates the experts' evaluations from visual biases potentially introduced by the plot's arrangement, focusing solely on the data's inherent characteristics.

For complex machinery characterized by multimodal data and multiple fault mechanisms, the task of analysis presents significant challenges for experts, and a degree of ambiguity in the outcomes is inevitable. However, the diagnostics model and feature space must be refined until the consensus can be reached that points close to each other are closer in terms of relative risk level than points that are farther from each other on the feature space plot. In essence, an assumption is drawn that there is some relation (not necessarily a linear relation) between the Euclidean distance between two points on the 2D plot of diagnostics features and the difference in degradation level. Once a feature space is obtained that is consistent with this requirement, the next step involves tracing real degradation paths in the feature space.

5.3.4 Construction of RTF sequence graphs

Once the feature plot meets the required validation criteria, the observations on this plot can be used to construct RTF sequences for each machine by connecting partial RTF from different machines. This is done in two steps. The first step is to trace real degradation paths from the monitoring history of each machine, and the second step is to construct RTF sequences. These will be explained in the following subsections.

5.3.4.1 Graph dataset generation step 1: Graph edges from the same machine

This stage aims to identify transitions from a lower-risk state to a higher-risk state. Transitioning between states requires confirmation that the machine could realistically evolve from the initial state to the subsequent one within the timeframe separating the two inspections. This critical step demands meticulous execution and a profound understanding of fault propagation mechanisms, as well as the symptoms manifested in the condition monitoring data. During consecutive inspections, the observed degradation at a later stage might arise from a different cause than previously detected. Consequently, observing sequential degradations with escalating risk levels does not necessarily imply that the latter state evolved directly from the former. Thus, it is imperative to establish informed and knowledge-based rules to define transitions between two condition monitoring data samples accurately.

Once the transitions are established, these can form the first set of edges in a graph dataset. These are designated such that each node is represented by the diagnostics feature vector corresponding to that sample. An edge between two nodes is simply embedded with the time between the two inspections that produced the start and end node data samples. This time value can be normalized to a reasonable size, depending on the speed of degradation. For a machine that degrades from installation to failure in one year, a time unit of days seems reasonable, whereas for a long-lasting machine with slow degradation in the order of decades, months or even years could be a reasonable time unit.

At the end of this step, the output is a set of directed graph edges between points on the feature space, where each edge represents a real degradation from a lower-risk physical state to a higher-risk physical state, in one machine. For most machines, the irregularity of condition monitoring, maintenance interventions with indeterminate impact on health state, and undetected degradation states could result in disconnected edges. This set of disconnected directed edges can be called ‘real edges’. The next step involves connecting these disconnecting edges by creating synthetic edges.

5.3.4.2 Graph dataset generation step 2: Synthetic graph edge generation

The objective here is to take the disconnected edges from the previous step and connect them to form graph structures representing the uninterrupted health state evolution of machines from start to failure. This step depends on the feature space continuity analysis from section 5.3.3. The hypothesis posits that two points nearby within Euclidean space on the feature space plot can be considered equivalent. To operationalize this, it is necessary to define a radius threshold that demarcates the extent to which degradation conditions must be similar to justify a synthetic connection. This radius is determined based on the dataset characteristics and the analytical methods discussed in Section 5.3.3. A smaller radius may be suitable for datasets with numerous tightly clustered points, while a larger radius may be necessary for more sparsely distributed data, acknowledging that the radius size is directly proportional to the prediction uncertainty in subsequent analysis steps.

Algorithm 4 Synthetic Run-to-Failure Trajectory Generation

Input: Historical inspection data D, radius threshold r, generations g
Output: Synthetic health evolution graph G
Initialize G with real transitions from D
for each machine in D **do**
    Sort inspection data by date
    Identify pairs of inspections (u, v) indicating deterioration
    Add real edges (u, v) to G
for $i = 1$ to g **do**
    for each real edge (u_p, v_p) in G **do**
        for each real edge (u_k, v_k) in G **do**
            if distance$(u_k, v_p) < r$ **then**
                Create synthetic edge (v_p, v_k)
                Add edge feature as distance(u_k, v_p)
                Add (v_p, v_k) to G
return G

The procedure for establishing synthetic edges based on the defined radius and the set of real edges is detailed in Algorithm 4. Each node represents the degradation state of a machine at a point in time, based on diagnostics done on the condition monitoring data collected at that time. The node feature is embedded with the feature vector extracted from the diagnostics model, the same as the feature projected onto the 2D plane. The edge feature is defined by the temporal interval between two nodes, specifically the time elapsed between the inspections producing the condition monitoring data for the start and end nodes. The algorithm systematically processes each real transition edge across all

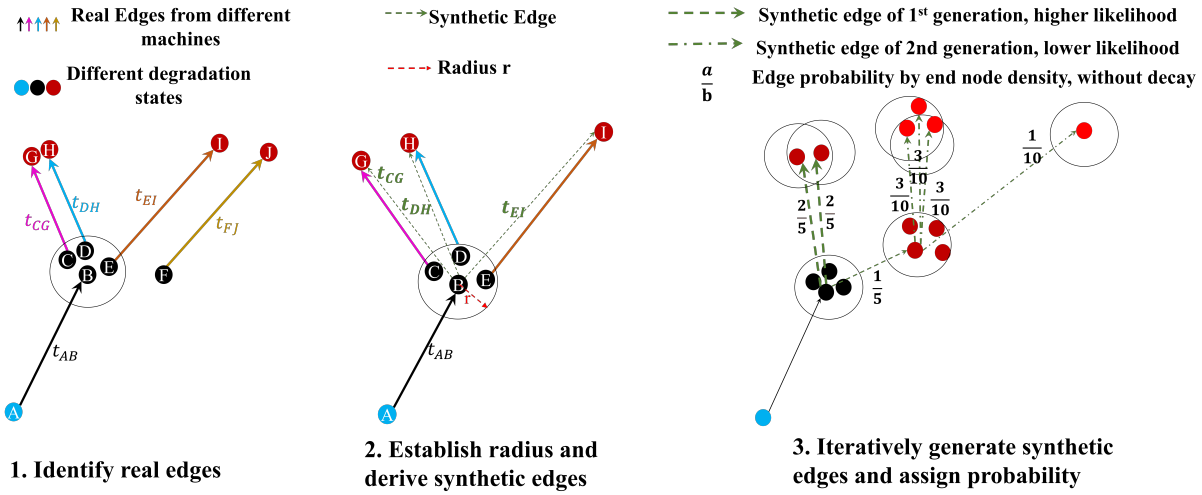


Figure 5.4: Diagrammatic illustration of the edge generation process. The probability assignment algorithm is given in Appendix A, section A.3.

machines. For a given real edge, if another real edge exists within the specified radius of the first edge's end node (regardless of the originating machine), a new 'synthetic edge' is formed connecting the end node of the first real edge to the end node of the second, as illustrated in Figure 5.4. The time difference between the nodes of the second real edge is then attributed to the synthetic edge. For example, if machine M1 transitions from state A to B over time t_{AB} and machine M2 transitions from state C to G over time t_{CG}, and states B and C are proximate, it is hypothesized that M1 could similarly transition from B to G in time t_{CG}. This approach facilitates the synthesis of potential degradation pathways across different machines based on their spatial and temporal proximity within the feature space.

The aforementioned inference presupposes that the degradation trajectories of machines M1 and M2 are sufficiently analogous. It also relies on the premise that the operational conditions influencing M2 degradation progression from state C to G are comparable to those that would affect M1, suggesting that M1 would exhibit similar behavior under identical conditions. If the knowledge to group machines with similar degradation trends based on machine design parameters, operating and load conditions, and so on is known, it is recommended to create synthetic edges from machines grouped by degradation trends and speeds. If such knowledge is not known, the only choice is to connect the edge from all machines based only on feature proximity. If the dataset encompasses multiple machines exhibiting varied degradation trends, it is feasible to discern these patterns solely from the data, without the necessity for domain expertise. Nevertheless, in cases of sparse data, applying knowledge-based grouping can partially compensate for the lack of comprehensive information.

Additionally, strategic feature engineering can enhance the model's capability to categorize machines based on the available data, effectively bridging gaps in the dataset. This feature engineering is designed to assign a likelihood measure to the synthetic edges, an adaptation necessary particularly when the dataset lacks comprehensiveness. The algorithm and logic of likelihood measure assignment are described in Appendix A (Algorithm 7).

The output of this step is a set of connected, directed graphs, where each graph corresponds to the possible evolution paths of each machine, starting from the first inspection of that machine. Therefore, this approach generates comprehensive, continuous RTF trajectory data for an entire fleet of machines, even in the absence of complete RTF data for any individual machine.

5.3.5 Masked graph dataset preparation

This step involves processing the graph structures obtained in the previous step to prepare a graph dataset for modeling. The objective of this dataset will be to train a model that takes a part of the graph as input and predicts the full graph. This means the model needs to predict the evolution of the health state of a machine given the first part of its RTF trajectory.

As these are directed graphs, the ordering of the nodes is important. Furthermore, this ordering directly affects the computational complexity in the modeling step. Both depth-first search (DFS) and breadth-first search (BFS) ordering are applicable, as both are valid methods of tracing the fault propagation from start to failure. However, BFS reduces computational complexity in graph generation tasks. Thus, the first step is to order each graph by BFS. Then, these can be stored as graph structures for graph-based deep learning, based on the many libraries available (Wang *et al.* (2019a)).

Then, for training, a graph dataset needs to be generated based on the requirements of the model. The following datasets are created:

- A partially masked graph dataset by masking the last nodes of the graph at different masking rates.
- A randomly masked graph dataset with a random mask variation of the same algorithm.

At the end of this step, a dataset is prepared to model a graph neural network for predicting future health states from historical health state evolution data.

5.3.6 Graph neural network health forecasting model

In this application, the health evolution of machines is modeled as graph structures where each node represents a degradation state, and an edge represents the time to evolve from one state to the next. As such, graph neural networks are a promising tool to examine.

Once the graph dataset is generated, any number of graph modeling techniques can be applied to predict the future health states based on a partial graph. Two broad categories of applicable graph techniques are graph reconstruction methods such as masked graph autoencoders (Hou *et al.* (2022)), and generative graph approaches such as GraphRNN (You *et al.* (2018)) or graph diffusion methods (Chamberlain *et al.* (2021)).

This step completes the methodology. In the next section, the proposed methodology will be applied to an extended case study of the hydrogenerator fleet presented in Chapter 4.

Highlights

In the second part of the prognostics methodology, the features from diagnostics classification were projected on a 2D plane. On this plane, the differences between individual machines were found to not exist anymore. Thus, two points coinciding on this feature space plot could be considered to be at an equivalent point in their paths from a healthy state to failure and could be substituted for one another. This allows us to construct full RTF sequences by connecting fragments of RTF trajectories from different machines, solving the data limitation.

5.4 Data and Application

This section presents a real-world industrial case study that applies the proposed methodology to a fleet of hydrogenerators, high-value machines with intricate fault propagation mechanisms. Initially, the problem statement and specifications of the data are introduced. Subsequently, the methodology is systematically applied to this dataset, demonstrating each step in practice.

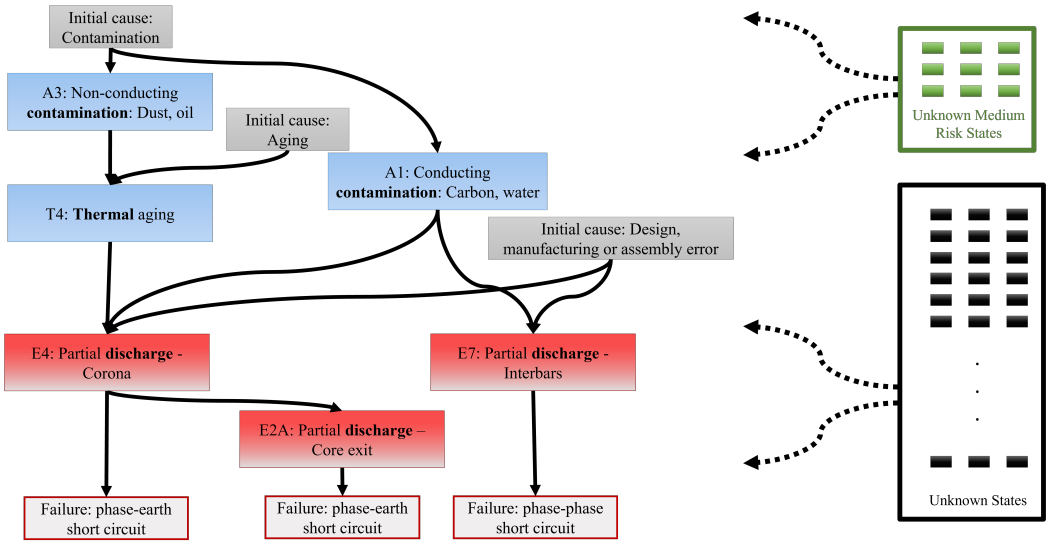


Figure 5.5: Subset of the expert knowledge-based fault propagation graph in scope. For full graph details, see Blancke *et al.* (2018). The medium-risk states (green) and unknown states (black) are connected to several of the studied states (blue and red), but their transitions are not studied in detail.

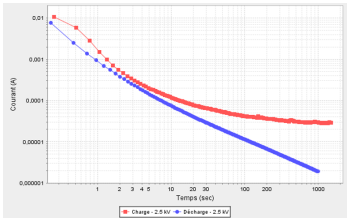
5.4.1 Domain study of hydrogenerator fault propagation mechanisms

In Chapter 4, only two fault types (physical degradation states) of the hydrogenerator fault propagation graph (Blancke *et al.* (2018)) were considered. In this chapter, this scope will be extended to a larger subset of this graph, depicted in Figure 5.5.

Figure 5.5 illustrates the initial causes of hydrogenerator faults in gray blocks, which are not physical states. The figure highlights three low-risk physical degradation states in blue, labeled A3, A1, and T4, and three very high-risk states in red, labeled E4, E7, and E2A, which are close to failure. These six states constitute the primary focus of this study. That means, the data and manifested symptoms of these states are studied in detail, and the diagnostics model is optimized for these states. The three blocks outlined in red represent failure modes and are not physical degradation states. On the side, a set of states are shown in green and black, both of which show the physical states that are part of the fault propagation mechanisms of the hydrogenerators, but outside the scope of this study. The scope was narrowed for practical reasons during this phase of the study. However, the condition monitoring samples from all of those states are used for the training model. The difference is that all samples of out-of-scope states that are understood to be comparatively less risky are together labeled “Unknown Medium Risk States”, and all the other states are labeled “Unknown States”. As the states in scope are only a small subset of the entire

graph, there are far more samples of unknown than the known categories.

As shown in Figure 5.5, the hydrogenerators should start their degradation in the low-risk states (A1, A3), and degrade to reach the high-risk states (E4, E7, E2A). However, tracing the condition monitoring history of the machines shows that the first recorded inspection of most machines starts in the high-risk states, then the low-risk states are detected at a later inspection, and then the high-risk states again. This seems counter to the knowledge-based fault propagation path and suggests that the machines started in the high-risk states. But in fact, the tools to detect the low-risk states were developed much later in the operation of the machines, which resulted in this data. Practically, this means that there are very few machines where the transition time from a low to high-risk state was recorded, and the majority of the dataset records transitions from the medium to high-risk states. This is the inverse of the long-tail distribution often found in PHM datasets (PHM data often contain more samples in healthy states and fewer in degradation states). Yet, in terms of RTF data requirements for prognostics modeling, this is as much of an obstacle as the more common long-tail data distribution.



(a) PDC data sample

(b) Thermal degradation

(c) Contamination

Figure 5.6: Data samples for states A1, A3 and T4.

In addition to the condition monitoring data modalities introduced in Chapter 4, a time series data modality called polarizing and depolarizing currents (PDC) is included. PDC is the only known tool that can accurately distinguish conducting (A1) and non-conducting (A3) contamination (Figure 5.6a). These states also have visual inspection images (5.6c). State T4 has only 17 visual inspection photographs in total (5.6b). Then, tabular data from the industrial maintenance management software are also included.

The correlation between the eight target degradation states and monitoring data types is detailed in Table 5.1. Initial discussions with domain experts were crucial to understand the range of degradation states and their symptom manifestation across different monitoring tools. Despite the inherent complexity and diversity of these modes, which introduce some uncertainty, this foundational step also involved associating a risk level with each state based on its proximity to failure, thereby establishing the groundwork for the model's development.

The dataset preparation involves using condition monitoring data as inputs and degra-

Table 5.1: Degradation states (class labels) and condition monitoring data sources (inputs)

State	Description	Risk level	Image	Text	PDC	PRPD	PDA	Ozone	Temperature	Categorical
A1	Conducting contamination (water, carbon)	Low	✓	✓	✓	x	x	x	x	✓
A3	Non-conducting contamination (oil, dust)	Low	✓	✓	✓	x	x	x	x	✓
T4	Thermal aging	High	✓	✓	x	x	x	x	x	✓
E4	Corona partial discharge	High	✓	✓	x	✓	✓	✓	✓	✓
E7	Partial discharge between bars	High	✓	✓	x	✓	✓	✓	✓	✓
E2A	Partial discharge at core exit	High	✓	✓	x	✓	✓	✓	✓	✓
Unknown medium-risk states	All the degradation states of the machine where indicative features are not known (outside the scope of this study), but the states are considered medium risk.	Medium	✓	✓	x					
Unknown states	All the degradation states whose feature analysis is outside scope of this study.		✓	✓	x	✓	✓	✓	✓	✓

dation states as labels, with each sample representing inspection data collected over a three-year window, as discussed in Chapter 4. This forms a multimodal dataset designed for multilabel classification. The next step involves analyzing the characteristics of this dataset.

5.4.2 Exploratory data analysis for hydrogenerators

Exploratory data analysis involves studies such as label distribution analysis, label correlation study, and sample similarity analysis. Figure 5.7 illustrates the results of the label distribution analysis in this multi-label dataset, with the x-axis representing label combinations. The graph displays the top ten combinations, highlighting that the "Unknown state" category comprises the highest number of samples. Notably, high-risk states are also well-represented. Conversely, low-risk states, including the thermal aging state T4, are significantly underrepresented, with only 17 instances of T4 in the dataset, posing additional challenges for analysis and modeling. Figure 5.8 displays label correlations, indicating positive relationships among the three partial discharge states, particularly between E4 and E2A. The "Unknown state" also shows positive correlations with these discharge states, and similarly, the two contamination states are correlated.

It is crucial to recognize the inherent ambiguity in class labels due to the limitations of data sources. For instance, visual inspection photographs identifying contamination cannot differentiate between conducting elements (like water, carbon) and non-conducting elements (like oil, dust); thus, contamination presence is generally noted without specification, leading to a label of 'True' for both A1 and A3. Such nuances in label meaning imply that multiple 'True' labels in a data instance suggest the potential activity of these states, not their definite presence. This also applies to E4 and E2A, as discussed in Chapter 4.

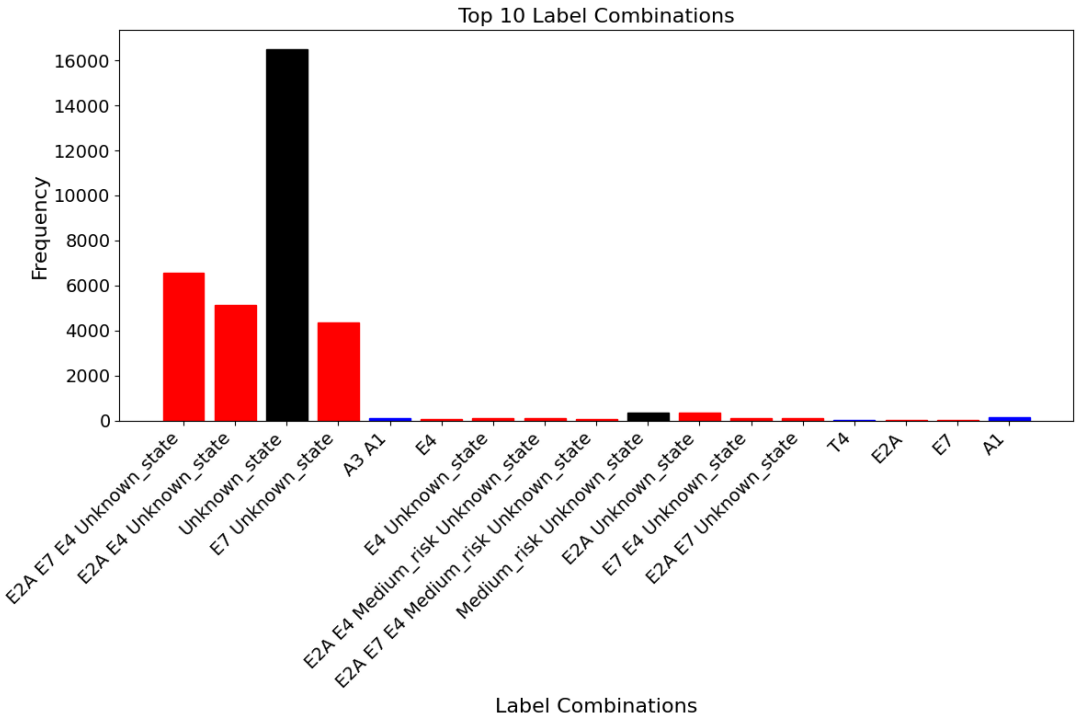


Figure 5.7: Distribution of the condition monitoring data by fault type. The labels show combined fault found in a three-year window.

Finally, the sample similarity analysis using t-SNE did not produce useful results. As anticipated, data from each modality clustered separately without any overlap, indicating that all samples of each data type were grouped, independent of the target class. Consequently, this analysis did not provide any actionable insights to facilitate classification.

5.4.2.1 Stratification

The EDA results in the previous section (5.4.2) have provided insights into data challenges and identified key bottlenecks, which will guide the design of expert models. Before proceeding, it is crucial to split the data into training and testing sets, ensuring each contains representative samples of all label combinations and modality-class mappings. A random dataset split may result in skewed representation, such as a training set lacking instances of underrepresented classes or modalities, thereby failing to address the label and modality balance essential for multimodal datasets. Stratifying the data into strata that maintain both label and modality distributions ensures that subsets, including test, training, and validation sets, are comprehensive and representative of the entire dataset.

To initiate the stratification process, the number of subsets (strata) must first be estab-

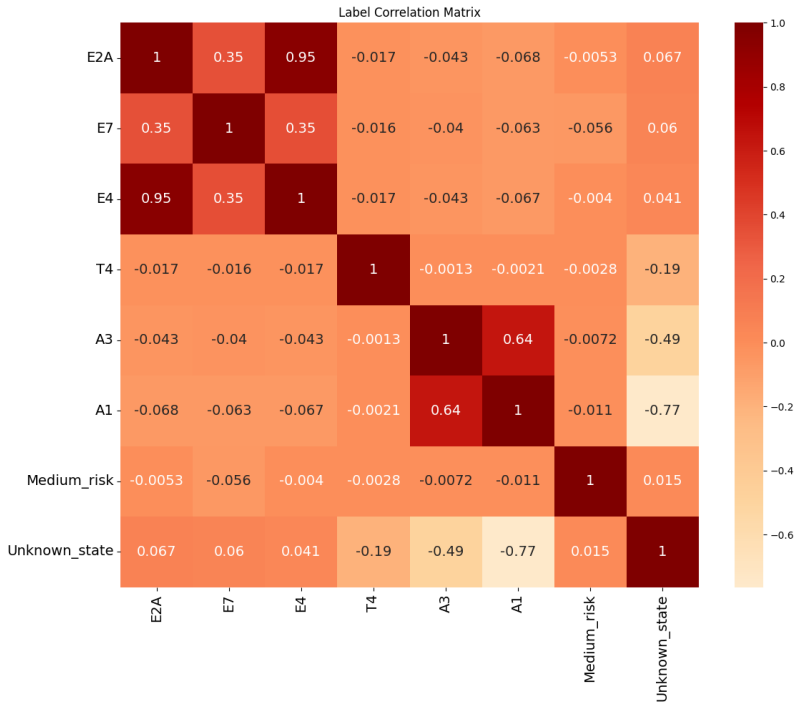


Figure 5.8: Label correlation of the output classes. High correlation between the three partial discharge types (E4, E2A and E7), as well as the two contamination states (A1 and A3).

lished, limited by the size of the smallest class to avoid resampling issues. In this dataset, the smallest class is T4, with seventeen data points, leading to the creation of ten subsets; two are allocated for testing, and the remaining for training and validation.

Table 5.2 displays the modality informativeness for each class, ranking image data as most informative for the three partial discharge classes (E4, E2A, and E7) due to clear visual distinctions of symptoms in photographs. However, for A1 and A3, PDC data proves more useful, while for T4, images are the sole data source. The informativeness of other classes and modalities, such as text, categorical data, ozone, and temperature, remains unquantified and their utility is considered similar, though their exact ranking involves some uncertainty.

Following these preparations, the data is stratified using Algorithms 2 and 3. These strata will be utilized for training the gate in Section 5.4.4 after training the expert models, as detailed in the next section.

Table 5.2: Modality informativeness table showing the comparative information quality of each data source for the target classes.

Informativeness rank	A1	A3	T4	E4	E2A	E7	Unknown medium	Unknown
1	PDC	PDC	Image	Image	Image	Image		
2	Image	Image		PRPD	PRPD	PRPD		
3				PDA	PDA	PDA		
4				Text-categorical	Text-categorical	Text-categorical		
5				Ozone/ Tempera-ture	Ozone/ Tempera-ture	Ozone/ Tempera-ture		

5.4.3 Expert models for hydrogenerator data challenges

Table 5.3 enumerates all expert models examined in this study following the analysis results presented in the previous sections. Each expert can classify the 8 target classes, albeit with varying accuracy across data subsets.

Expert model No. 2, for example, is an extension of the classification model trained in Chapter 4 to include one more type of partial discharge. The detailed architectural design and description of the most important expert models are given in Appendix B.

After all the experts are trained, their predictions on the entire training set are stored for training the expert selector routing gate module in the next step.

5.4.4 Gate and aggregation

To initiate gate training, expert predictions on the entire training dataset are generated and stored alongside fusion layer features. Training occurs within a k-fold (here, k=4) cross-validation setup using strata from Section 5.4.2.1. This dataset incorporates balanced strata from the training set along with expert predictions, ensuring the gate's robustness to dataset bias and imbalance. Figure 5.2 illustrates gate training, where expert models remain frozen, with predictions loaded from a file, ensuring the independence of gate training from the expert modules training.

Following gate training, the module is frozen, and the assigned probabilities for each expert per data instance are stored for subsequent use. These probabilities are utilized to weight, transform, and aggregate expert features.

Table 5.3: Expert models.

Expert Model	Input Modality	Task
1	Image + Text.	Image expert for all classes.
2	Image + Text + PRPD + PDA + Oz/ Temp/ Tabular.	Partial discharge (E4, E2A, E7) classification.
3	PDC.	Classify A1 vs A3.
4	Samples where unknown is present with another class.	Separate unknown class from others to avoid universal true prediction.
5	Samples where unknown is present with unknown medium.	Separate unknown class from unknown states with medium risk.
6	Rows without unknown or unknown medium.	Train an expert without data bias from the unknown states.
7	PDC + Image + Text (rows where unknown medium is present, no image with known partial discharge).	Separate A1 and A3 from unknown medium, which have positive label correlation.
8	PDC + Contamination images	Contamination classes expert.
9	No PDC or contamination images.	Separate unknown states from partial discharge states, which have positive label correlation.
10	Thermal aging + Contamination images.	Low risk classes expert.
11	Text only.	All classes, but from inspection notes only. There are some instances where text is present but no other data.
12	PDA only.	Separate partial discharges with only PDA.
13	PRPD only.	Separate partial discharges with only PRPD.
14	Partial discharge Image + PRPD.	Separate partial discharges with Image + PRPD.
15	Partial discharge Image + PDA.	Separate partial discharges with Image + PDA.
16	Partial discharge PRPD + PDA.	Separate partial discharges with PRPD + PDA.

5.4.4.1 Feature aggregation

In this step, the features extracted by the top k experts are aggregated and transformed. While it is feasible to conduct this step concurrently with gate training, the methodology opts to fully train the gate first. The gate's model selection is then utilized to train the weight transformation.

Using the same strata employed for gate training, the initial step determines the number of experts to load into memory simultaneously based on the available memory (16 GB Nvidia RTX A4000) and the size of the largest expert models. In this study, the gate selects the top $k = 4$ experts for each sample in the training dataset. During training, experts do not need to be loaded into memory; the fusion layer features for all experts are stored in a file (see Section 5.3.2.2). These features, output by the modality branch fusion layer present in all experts, ensure consistency across designs.

The feature vectors from the top k experts are transformed using weights provided by the gate, scaled to sum to one. If the memory resources allow loading all experts at once, this step may be redundant. However, lacking such resources, this process weights the features based on each expert's contribution to the final output. The transformed features are aggregated and fed into dense connections, leading to a sigmoid layer for the final multilabel classification output. Only the weight transformation, dense layers, and output layer require training in this step, making it relatively fast compared to training the experts.

5.4.4.2 Diagnostics inference

The inference process is performed on the test set, which has not been used for any of the training steps thus far. This process is illustrated in Figure 5.9. During inference, a test sample is input to the gate, which assigns probabilities to each expert. The top k experts are then selected and loaded into memory. Each expert processes the data, and features from their fusion layer are extracted. Note that each expert model can make its own classification, but this output is not used. Indeed, the process flow for each expert ends at the fusion layer. The feature matrices are then transformed using gate weights, aligning them according to each expert's expected contribution. These transformed features are densely connected to subsequent layers, with the final output layer activated by sigmoid. This inference strategy maximizes available memory without adding time costs due to sequential expert activation.

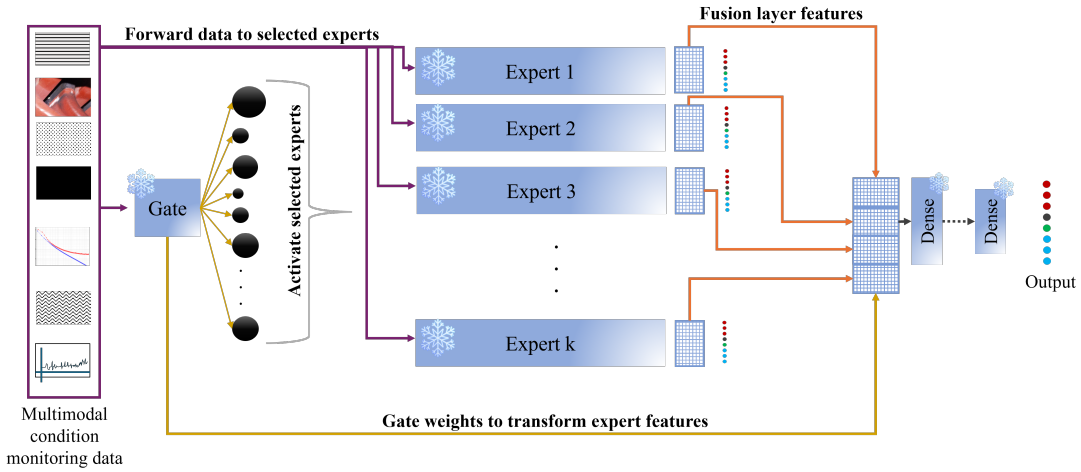


Figure 5.9: Data flow pipeline at inference time, after all training steps are complete. The sample is forwarded to the gate, which activates k experts. The sample is then forwarded to the experts. The experts each extract features from the data, and the fusion layer features of the experts are retrieved. These are transformed by the gate weights and aggregated. The aggregated features are processed and densely connected to the output layer.

5.4.5 Results for hydrogenerator diagnostics

Thus far, a modular DL-based classification model has been trained for diagnostics. The test set inference results of this model were compared with several variations by altering the number of activated experts from 1 to 4 and adjusting the expert fusion feature dimensions to 8, 16, and 32. The best model with 4 active experts and a feature size of 32 achieved an exact match ratio of 88%. Precision, recall, and F1 score are all over 90%, and good scores were achieved on other metrics such as Hamming loss, log loss, and Jaccard score. The full result analysis and ablation studies conducted on the diagnostics model architecture are given in Appendix C. As the quantitative metrics of the diagnostics model are satisfactory, we arrive at the last part of the methodology (Figure 5.1). The next step is to analyze the intermediate features of this model.

5.4.6 Diagnostics feature analysis and expert validation

The feature space obtained from the diagnostics model is used to train a VAE, and the latent space is plotted in a 2D scatter plot. This is shown in Figure 5.10. This figure was obtained after several iterations of optimizing the diagnostics model and feature dimension reduction. The features shown in the figure met the requirements R1, R2, and R3 described in Section 5.3.3.

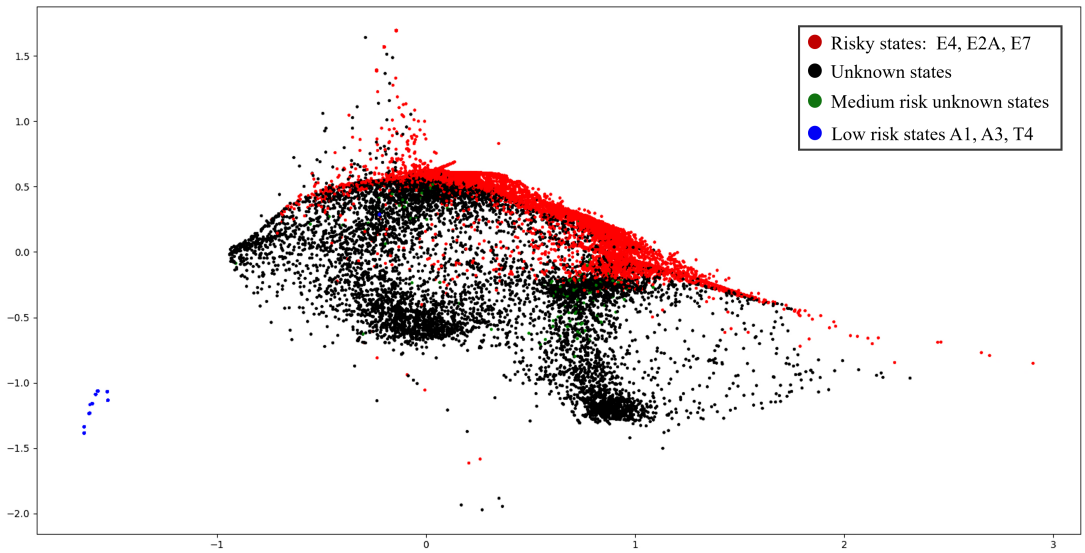


Figure 5.10: Visualisation of 2-D representation of the training data features of diagnostics model using variational autoencoder

The states representing the low-risk states are in the bottom left corner of the space, far from the states in red, which represent the high-risk states. It should be noted that the condition monitoring samples were labeled such that if in a time window of three years a set of samples showed the degradation states (E4, E2A, or E7), it was considered high-risk and colored red.

In this feature space, a single point originates from features extracted from a multimodal data sample and represents a combination of physical degradation states. The consistency of the 2D space was validated with an expert. Such validation is a lengthy and challenging task, as it is quite difficult to take two samples of different modalities (image and PRPD, for example) which could belong to different machines, and compare them. This step was conducted until sufficient conclusions were drawn to support the assumptions in the methodology. However, further refinement and validation of the features could always improve the model. Once this step is complete, the next step is to prepare the dataset.

5.4.7 RTF sequence generation and masked graph dataset

This step involves creating a synthetic edge and a graph dataset based on the algorithms in subsections 5.3.4.1, 5.3.4.2 and 5.3.5.

The radius parameter, defining the local maximal threshold for the neighborhood of nodes, is set to incrementally increase from $1e-8$. This allows an exploration of varying

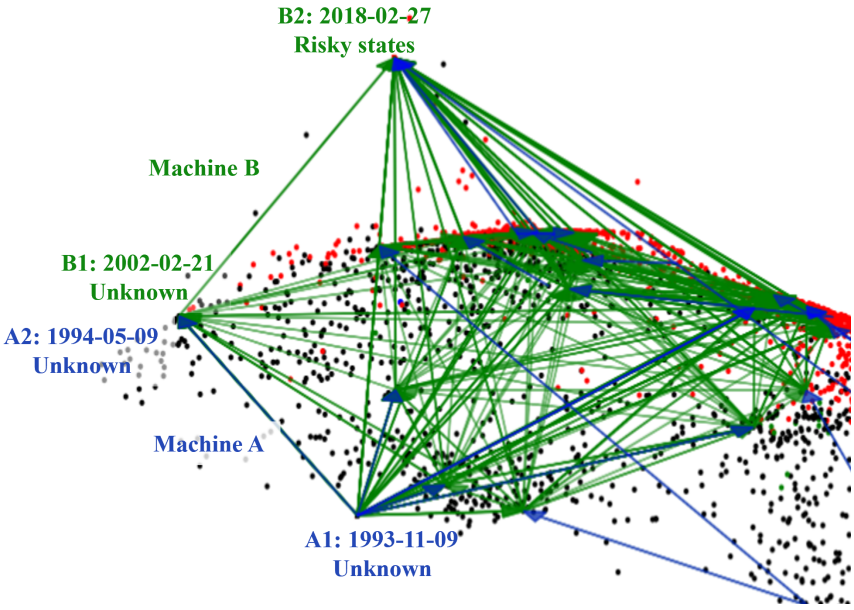


Figure 5.11: Creating synthetic edges for a single machine (image cropped for visibility). The blue edges highlight the evolutions of one machine A. The green edges are all from different machines. The figure highlights a constructed sequence for A by connecting with a partial trajectory of machine B.

scales of proximity among the diagnostic features in the 2D space. The decay factor for moderating the influence of synthetic edges across generations is set at 0.9. This value ensures that each subsequent generation of synthetic edges contributes slightly less to the transition density, simulating a natural attenuation of influence over distance and generational depth. This does not come into play, as the iterative generation was capped at 1, due to the highly dense feature space. Additionally, transition density values for outgoing edges from a node are normalized to 1, thus converting raw transition densities into a probabilistic framework that reflects the likelihood of transitioning from one state to another within the modeled system. These parameter values are crucial for enabling the synthetic edge generation algorithm to robustly accommodate the uncertainties in predicting machine degradation paths and speeds, while also adapting effectively to the dataset's limitations.

A dataset of directed graphs is then constructed using the Deep Graph Library (DGL). To create this dataset, the actual feature vector from the diagnostics model is used to embed the nodes as a tensor. Similarly, the time differences between states in terms of several months embed the edge feature. Following the graph construction, the nodes within each directed graph are then ordered by BFS search to capture the sequential progression of degradation. Then, the graph needs to be masked to create a training dataset (A masking algorithm is presented in Appendix A). A masking rate, varied across a predefined range,

is applied to both node features and edges from the end of the directed graphs to simulate partial evolution scenarios. Graphs masked from 10% to 90% mask rates are included in the dataset. This dataset is then used in the next step for model training.

5.4.8 Graph neural network for prognostics modeling

The dataset constructed in the previous step can be used in many ways to implement machine health forecasting. While recurrent networks or diffusion models are potential alternatives, the main objective of this section is not to present the strongest neural network, but rather to validate the utility of the proposed RTF data construction method. Thus, an autoencoder architecture is chosen for its one-step forecast capability as opposed to the complexity of a recurrent model.

In this step, a Graph Masked Autoencoder (GraphMAE) (Hou *et al.* (2022)) model is designed, and tailored to predict the machine degradation processes from initial condition monitoring data. This involves crafting an enhanced graph MAE model that leverages the representational capabilities of Graph Attention Networks (GATs) to encode and decode the states and transitions inherent in the degradation pathways.

The model comprises an encoder-decoder architecture where both components employ graph attention layers. The encoder uses GATs to aggregate feature information from neighboring nodes, allowing it to learn a rich representation of the node features and their local graph topology. Specifically, a series of graph attentional convolution layers (Veličković *et al.* (2017)) allows for progressive refinement of these representations. The decoder mirrors this structure to reconstruct the original node features from the encoded representations. The structure contains three modules for reconstructing the masked nodes, predicting if an edge exists from each of the previous nodes to the newly constructed node, and for constructing the edge feature if an edge exists. The BFS ordering reduces the computational complexity of the edge existence prediction to the last layer. The illustration of the model to reconstruct the masked graph is shown in Figure 5.12.

A multi-task loss function is devised to facilitate the learning process, addressing three tasks: node feature reconstruction, edge existence prediction, and edge feature prediction. The node feature reconstruction is a regression task to the actual feature vectors, so this is trained with a mean square error (MSE) as the loss. The edge existence prediction is a classification between the presence or absence of edges, thus it uses a binary cross-entropy loss. For the module to predict edge features, another MSE loss is applied as this is also a regression. These loss components are dynamically weighted, allowing the model to balance its focus between tasks based on their relative learning progress (Guo *et al.* (2018)).

Training the model involves iterating over a dataset of masked graphs, where the model

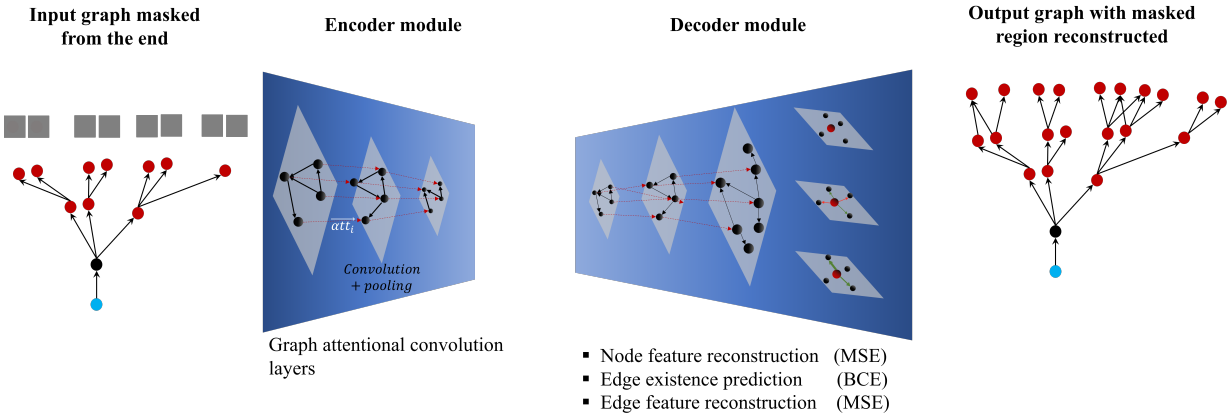


Figure 5.12: Illustration of a masked auto-encoder with the modules for node reconstruction, edge existence prediction, and edge feature reconstruction.

is optimized to predict node features and the presence of edges, with the option to also predict attributes of these edges. The multi-task optimization to balance the training of these tasks uses dynamic task weighting based on recent loss on each task. Each of the three tasks begins with an equal weight. After n epochs, the weights of the tasks are updated as follows:

$$\text{Task Weight} = \frac{\text{Recent Task Loss}}{\text{Total Loss}}$$

This allows the model to focus its parameter optimization on the tasks it finds difficult, leading to balanced learning on all three tasks.

5.5 Prognostics Results

The model is tested with partial graphs for completion, which means that it is given the first state or first few state transitions in the history of a machine and demanded to forecast the future evolution, both the future states and transition times. The node feature reconstruction module has an MSE of 0.007 on the test set. This means that the model is quite adept at predicting the future states given the evolution history of a machine. This is not very surprising, as fault propagation mechanisms are quite well-known in terms of possible future paths. Predicting the time to future states is a much more difficult task.

The result of the edge feature reconstruction module, which is responsible for predicting the time, is given in Table 5.4. The table compares the MSE and RMSE of the edge feature reconstruction module on three node embedding size variations of the model trained by masking graphs from the end and a random masking approach. The model trained on the largest node size gets the best performance, an RMSE of around 5. As the time is in

Table 5.4: Results showing the time prediction error of models trained on different conditions

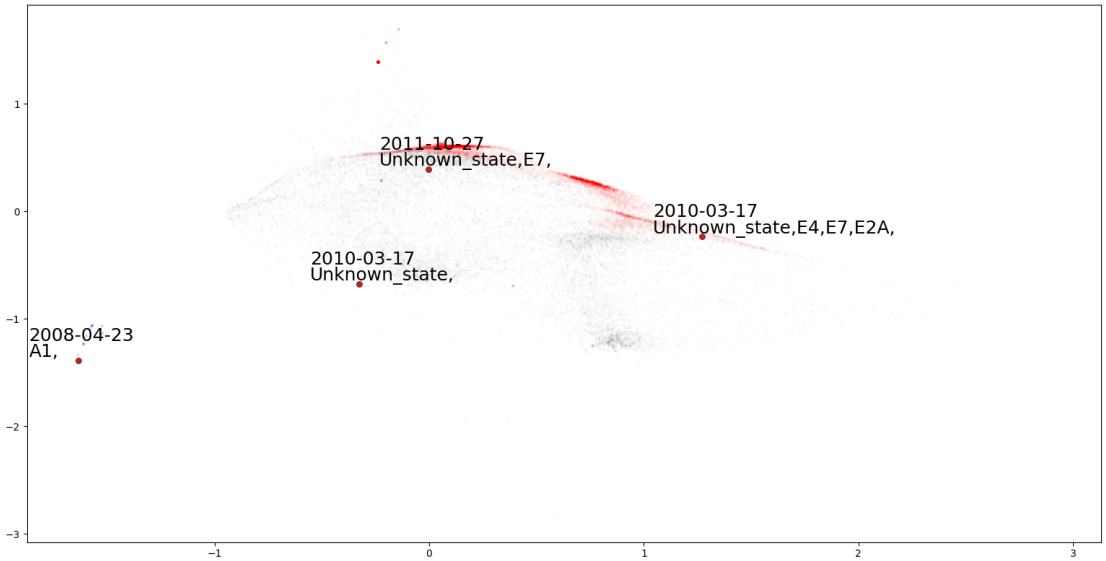
Model	Graph node feature size	MSE: Edge feature reconstruction error	RMSE
Graph MAE - end mask	8	254.4	$\sim$ 15.95 months
Graph MAE - end mask	16	144.8	$\sim$ 12.03 months
Graph MAE - end mask	32	23.55	$\sim$ 4.85 months
Graph MAE - random mask	32	26.77	$\sim$ 5.17 months

the unit of months, this amounts to a prediction error of 5 months. Considering that the lifetime of the machine is over 70 years and the degradation is a very slow process in the order of years, an error of 5 months is reasonable. Moreover, it was shown that a more expressive node feature can improve the prediction of time to evolve to a future state.

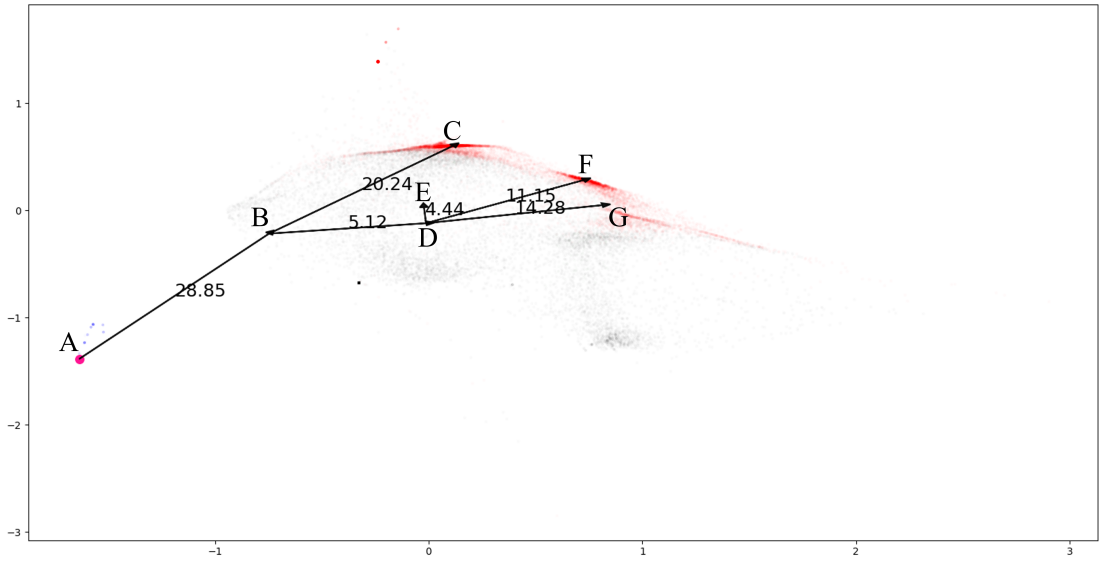
The masked autoencoder trained on randomly masked graphs performs slightly worse than masking from the end alone. This was tested on the same test set, which only contains graphs masked from the end, for consistent comparison between models. However, training on the random masked graphs does not seem to bring an advantage over masking only from the end.

An example of the model prediction is shown in Figure 5.13. It presents the inspection data of a machine M_i on the top and the predicted output on the bottom. This is a typical example of real data conditions in the industry, where inspection history is highly sparse. The model was given only the first node as input and predicted the future health state and time to the predicted states. One single condition monitoring data sample was input to the diagnostics model, and the intermediate feature from this model was input to the GNN. The output of the GNN is a vector of node features in the same dimension as its input, along with a binary edge existence vector and an edge feature vector. The true data plot (Figure 5.13a) was obtained by feeding the diagnostics features to the VAE to obtain the 2D coordinates and overlaying the highlighted points on the entire diagnostics feature space plot. The prediction plot (Figure 5.13b) was obtained by feeding the predicted node feature vectors to the VAE and overlaying them on the feature space plot.

The path $A \rightarrow B \rightarrow C$ predicted by the model is quite close to the actual inspection data. Given a machine in a health state A1 at time t, it is possible to evolve to the state E7 at ($t+ \sim 5$ years). The prediction is within possible limits, as validated by the domain expert. However, the other paths to states D, E, and G do not fall within the time shown by the inspection data.



(a) Actual inspection dates and identified fault states of machine M_i. Points from the test machine are highlighted by reducing the feature plot opacity.



(b) Predictions for machine M_i with only one inspection sample as input. (Coordinates obtained by encoding the GNN output with VAE.)

Figure 5.13: Masked graph autoencoder predictions on one machine.

However, further analysis shows that it may not be possible for the machine M_i to evolve to the two other states from the initial state in two years, suggesting that these two inspection samples may have a different original cause than point A. It should be noted that while the predictions made by the model can be evaluated quantitatively against a test set, the analysis of each prediction by case is a difficult process due to the ambiguity about the degradation speeds of machines. However, the results are quite promising considering the data limitations of the case study.

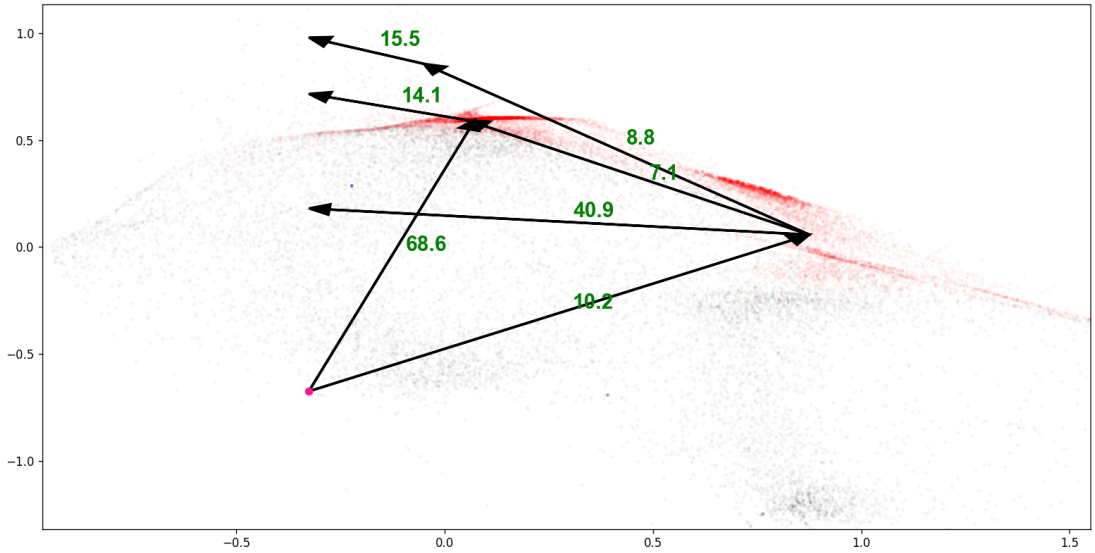


Figure 5.14: Predictions on machine M_j with only one input sample.

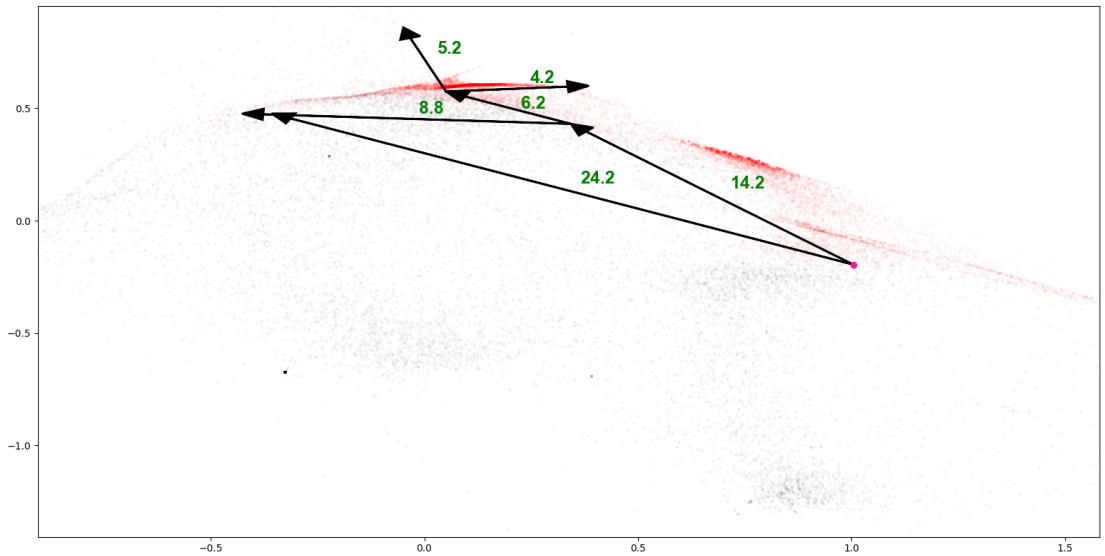


Figure 5.15: Predictions on machine M_k with only one input sample.

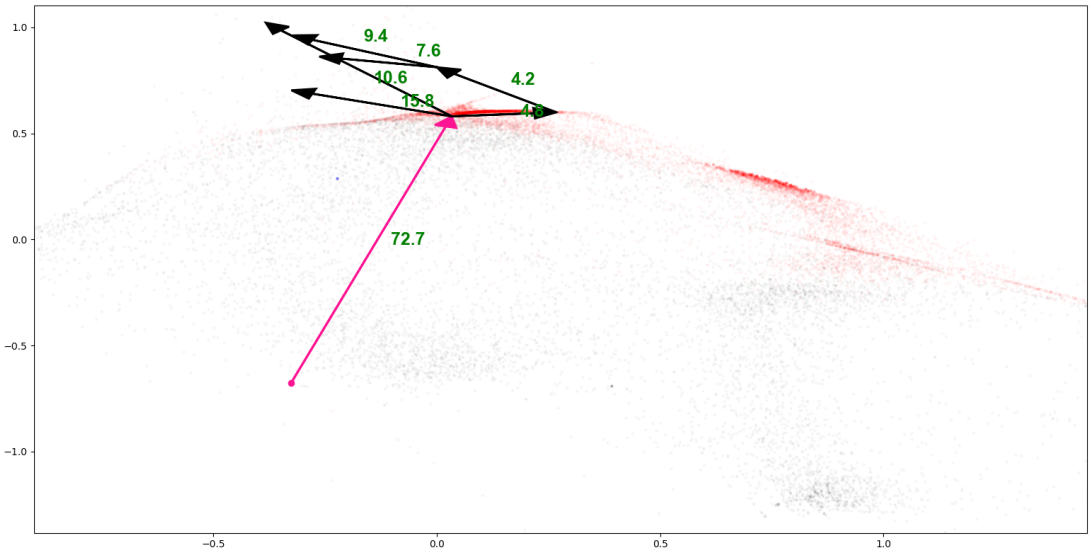


Figure 5.16: Predictions on machine M_j with two input samples and an edge.

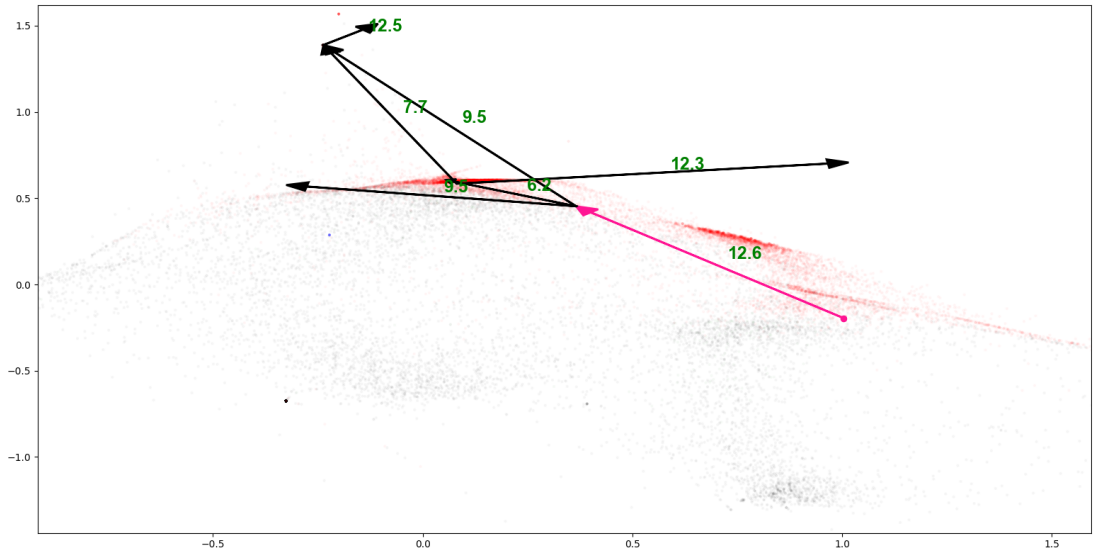


Figure 5.17: Predictions on machine M_k with two input samples and an edge.

Figures 5.14 to 5.17 illustrate additional results from the model predictions involving two machines M_j and M_k. Initially, the model predicts using only a single inspection sample as input (Figures 5.14 and 5.15). These predictions are then compared with outcomes derived from M_j and M_k when two samples and the corresponding edge are provided as input (Figures 5.16 and 5.17), for the respective machines.

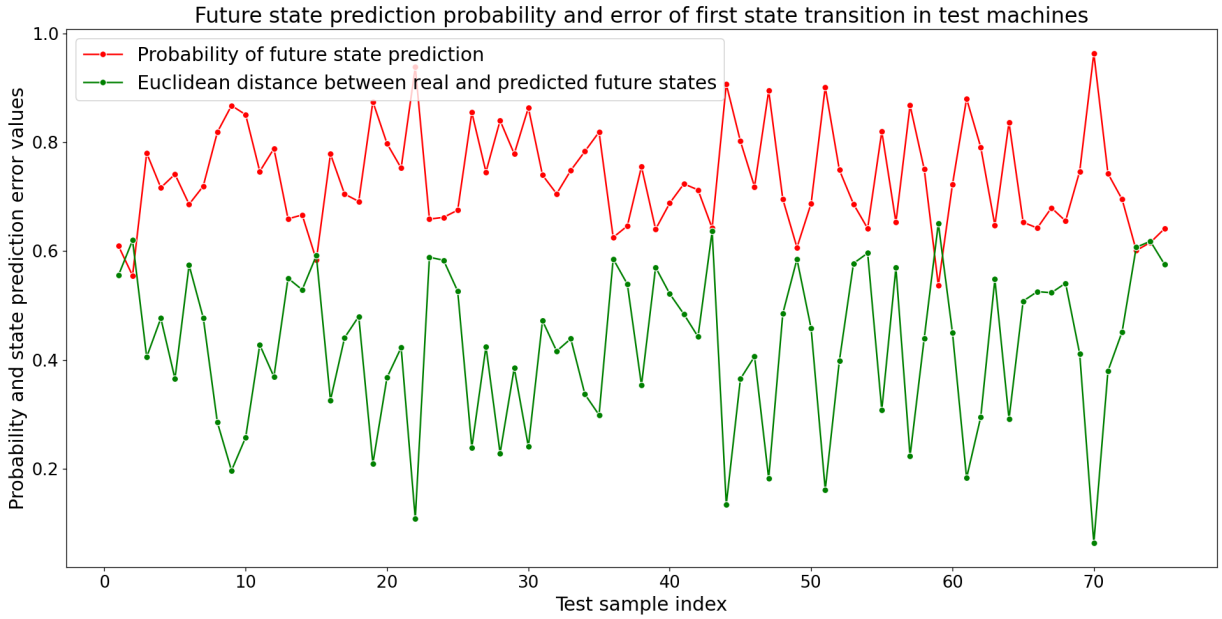


Figure 5.18: Euclidean distance between predicted and actual future state feature vector of the first transition for all machine units in the test set is shown in the green line graph. The probability of the corresponding predictions are given by the red points.

Figure 5.18 shows a comparison between the prediction error of future states and corresponding prediction probabilities. For all the machines in the test set, only the first transition is considered in this plot, as the real degradation is only known for one transition edge (after which the model is based on transitions from other machines). For each machine, the model typically predicts more than one possible future state when starting from a given initial state. In this plot, the predicted future state with the highest probability is considered for each machine. The predicted probability is given by the red points, and the difference between actual future state to the predicted state (calculated as Euclidean distance between the corresponding feature vectors) are shown in the green points. It can be observed that when the predicted probability of a future state is high, the corresponding predicted future state is close to the real observed state, and vice versa.

Figure 5.19 shows the same state prediction error values in the green points as Figure 5.18. However, in this figure, these are compared with the state transition time prediction errors, shown in blue points. Each blue point represents the difference between actual transition time to the true future state and the predicted transition time to the predicted

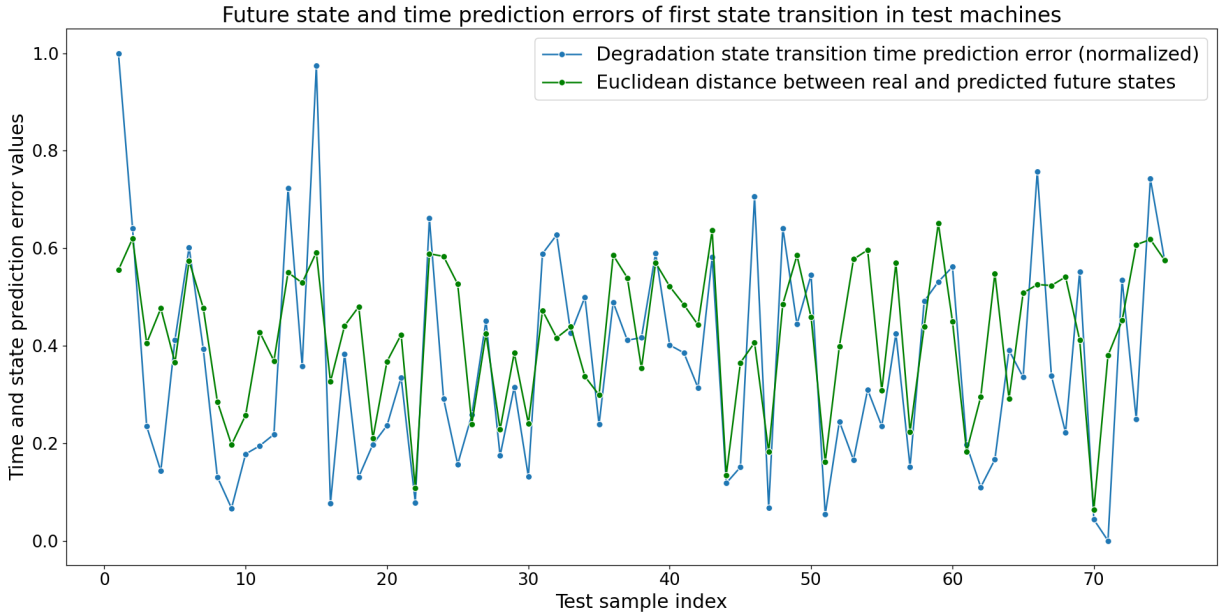


Figure 5.19: Euclidean distance between predicted and actual future state feature vector of the first transition for all machine units in the test set is shown in the green line graph (same as Figure 5.18. The prediction errors of the actual time to transition and the predicted time to transition (normalized for readability) are given by the blue points.

future state. This difference value is normalized within the range of transition time errors of the test set predictions by min-max scaling (Equation 5.5). This is done to scale the time prediction error (which is originally in the number of months) to a comparable scale as the state prediction errors, for efficient visualization. In most cases, when the predicted state is closer to the real future state, the predicted transition time is also closer to the actual transition time. However, it can be observed from the figure that there are several cases when the results deviate from this. Prediction of transition time to a future degradation state remains a more challenging task, even when the future state itself can be predicted reliably.

$$\begin{aligned} \epsilon &= |t_p - t_t| \\ \epsilon_{\text{norm}} &= \frac{\epsilon - \min(\epsilon)}{\max(\epsilon) - \min(\epsilon)} \end{aligned} \tag{5.5}$$

where ϵ represents the absolute error between the predicted transition time t_p and the true transition time t_t, and ϵ_{norm} denotes the normalized time prediction error, scaling ϵ to a [0,1] range aligned to the minimum and maximum errors among the test set predictions.

The predictions starting from a single node are expected to align closely with those that utilize more initial information. However, when the model begins with only one node,

the resulting predictions are broader, indicating multiple potential futures. Figures 5.16 and 5.17 display the actual observed future up to one edge, which falls within the range of the model predictions from the initial node. This consistency across predictions reinforces confidence in the model's reliability. More perspectives on the results are discussed in Appendix A.3.0.1.

5.6 Conclusion

In this chapter, we developed a methodology to perform machine health prognostics in the absence of any RTF trajectories. We accomplished this by first performing diagnostics, by analyzing the intermediate features from the classification model, and by connecting partial trajectories from different machines based on feature proximity in this space. The main contribution of the methodology is in relaxing the data requirement from complete trajectories to partial trajectories.

To achieve this objective, the methodology imposed strict requirements on diagnostics accuracy. This involved solving several challenges such as class imbalance and sparsity in a multimodal, multilabel classification dataset scenario. To solve these, we presented a modular deep learning architecture with specialized experts, a routing gate, and a new dataset stratification algorithm, which all work together to solve the many data challenges.

The proposed methodology was applied to a real-world dataset from a hydrogenerator fleet. The diagnostics model obtained consistently high scores across several metrics and held up to expert validation. The features from this model enabled the construction of an RTF dataset in graph format, and a masked graph reconstruction autoencoder model was trained to predict future health given the initial health monitoring data of a machine. The results obtained are in agreement with expert validation.

The complete methodology developed in this chapter built upon the cross-modal attention architecture introduced in Chapter 3, expanded the diagnostics methodology in Chapter 4, and integrated them all together to form an end-to-end prognostics solution that addresses a range of industrial and scientific challenges.

This chapter concludes the contributions presented in this thesis. The general conclusions and reflections will be presented in the next chapter.

Chapter 6

Conclusion

Contents

This research aimed to explore multimodal data-driven PHM techniques to enhance the reliability and efficiency of industrial machines. This study has successfully developed and validated several predictive models that leverage various data modalities to support maintenance decision tasks.

The full schematic summary of the thesis is illustrated in Figure 6.1. In the rest of this chapter, we will recall the objectives of this thesis, recap and highlight the key findings, discuss the limitations of this work, and recommend directions for future research.

6.1 Recall Research Problems and Objectives

At the outset, several key challenges were identified within the field of PHM for industrial machines:

- Addressing the issue of missing and noisy data in industrial multimodal datasets.
- Developing robust models that can handle sparse and irregular datasets.

- Incorporating domain expertise and subjective elements to enhance model reliability.
- Managing high class imbalance in multimodal data for accurate diagnostics.
- Creating methods for health prognostics without complete run-to-failure data.

The main objective was to develop methodologies and algorithms that could integrate multiple data modalities and leverage domain knowledge to create robust predictive models applicable to real industrial settings.

6.2 Summary of Key Contributions

This research has made significant strides in addressing the following challenges.

- In Chapter 2, we reviewed the literature on data-driven PHM and noted that most datasets consisted of unimodal sensor signals. Additionally, many benchmark datasets were simulated, revealing a gap between state-of-the-art advancements and practical applicability. We identified the lack of industrial data as a contributing factor and proposed multimodal data as a potential solution. Our investigation into multimodal learning and its applications in fields such as medicine, along with recent developments in foundation models, indicated that multimodal learning is sufficiently mature for application in PHM. A review of PHM studies utilizing multimodal data underscored the need for further exploration in this area. Consequently, we formulated several research questions to guide this thesis.
- In Chapter 3, we conducted a first exploration into multimodal learning in PHM using a simulated dataset. Identifying missing data and noise as key challenges, we developed a cross-modal attention-based multimodal learning method. After performing a comparative analysis across a wide range of missing and noisy data conditions, we concluded that the proposed attention-based learning technique could mitigate data limitations significantly. This forms the first contribution and is used throughout the rest of the development of the thesis.
- In Chapter 4, we made two contributions. First, we proposed a multimodal diagnostics methodology that addressed severe data sparsity in certain modalities by incorporating domain knowledge into the design process of a specialized unimodal feature extraction pipeline. Existing foundation models were leveraged to support this. Time alignment issue between modalities was also addressed with a time-differences vector to attention weight the data features. The robustness of the proposed methodology

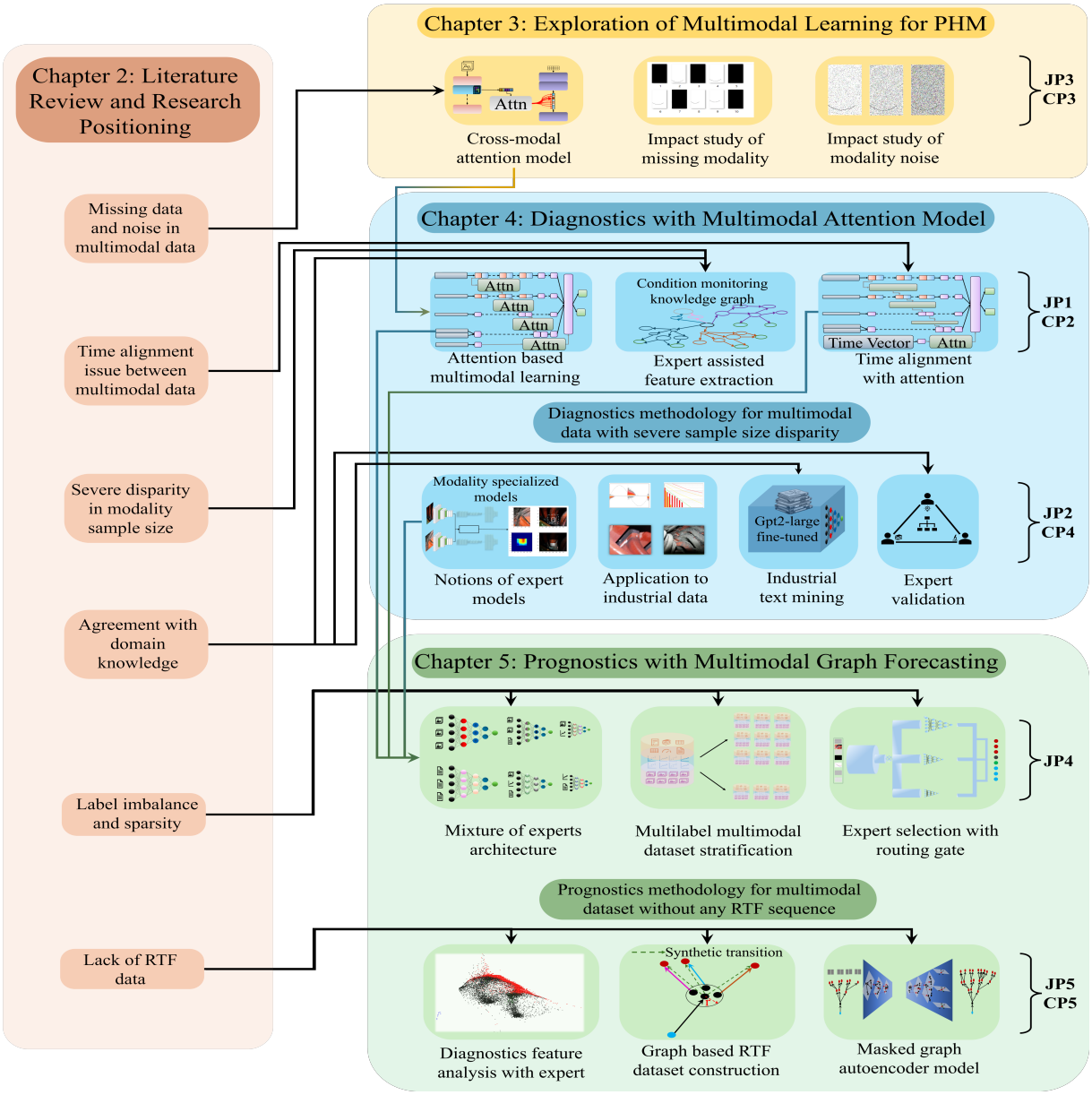


Figure 6.1: Schematic summary of thesis objectives and contributions. The black arrows represent the connection between a research objective and a contribution, whereas the colored arrows traces the scientific development through chapters.

on sparse and irregular data was validated on an industrial dataset of hydrogenerators. Second, the subjectivity in certain PHM tasks such as health index estimation was addressed by using a large language model fine-tuned on industrial text documents and inspection notes, thus ensuring the data-driven model is aligned with human expert observation of the condition monitoring measurements. The findings of this chapter laid the groundwork to extend the research to prognostics.

- Finally, in Chapter 5, we explored forecasting of machine health into the future. We proposed a methodology to address a critical challenge that hinders the application of state-of-the-art techniques to the industry: lack of run-to-failure data. The proposed methodology heavily depends on the accuracy of diagnostics, and novel contributions were made to address label imbalances and multimodal information distribution to prevent model bias. The methodology takes the features of a multimodal data-based diagnostics classification model and projects it on a 2D plane. With expert validation that coinciding points on the 2D space can be exchanged, we constructed RTF trajectories by connecting partial trajectories from multiple machines. This let us develop an RTF dataset in graph format, which was used to train a graph prediction model for prognostics. This final contribution is a promising solution to the challenge of data scarcity in the industry, and ties together all the techniques and principles developed throughout the thesis in an end-to-end prognostics methodology from data to prediction.

The contributions presented in the thesis are methodologies that can be adapted, and their implementation steps were demonstrated on an industrial dataset from a hydrogenerator fleet. To the best of our knowledge, this thesis presents the first work in PHM literature to address these challenges and demonstrate the findings on a real-world dataset, where the results are corroborated by industry experts. Several publications have been generated throughout this research, highlighting the developments presented in this thesis.

Journal Papers (Published)

- **JP1**: Sagar Jose, Khanh T. P. Nguyen, Kamal Medjaher, Ryad Zemouri, Mélanie Lévesque, and Antoine Tahan. “Fault detection and diagnostics in the context of sparse multimodal data and expert knowledge assistance: Application to hydrogenerators.” Computers in Industry 151 (2023): 103983.
- **JP2**: Sagar Jose, Khanh T. P. Nguyen, Kamal Medjaher, Ryad Zemouri, Mélanie Lévesque, and Antoine Tahan. “Advancing multimodal diagnostics: Integrating industrial textual data and domain knowledge with large language models.” Expert Systems with Applications 255 (2024): 124603.

Journal Papers (Submitted)

- **JP3**: Sagar Jose, Khanh T. P. Nguyen, and Kamal Medjaher. "Enhancing Industrial Prognostic Accuracy in Noisy and Missing Data Context: Assessing Multimodal Learning Performance." (Submitted to Journal of Intelligent Manufacturing (May 2024)).
- **JP4**: Sagar Jose, Khanh T. P. Nguyen, Kamal Medjaher, Ryad Zemouri, Mélanie Lévesque and Antoine Tahan. "A modular deep learning methodology for multi-fault machine health diagnostics from sparse and imbalanced multimodal data." (Submitted to Neurocomputing (August 2024)).
- **JP5**: Sagar Jose, Ryad Zemouri, Khanh T. P. Nguyen, Kamal Medjaher, Mélanie Lévesque and Antoine Tahan. "Prognostics of complex machinery with sparse multilabel multimodal run-to-failure data: A graph neural network approach." (Submitted to Advanced Engineering informatics (July 2024)).

Conference Papers

- **CP1**: Sagar Jose, Raymond Houe Ngouna, Khanh T. P. Nguyen, and Kamal Medjaher. "Solving time alignment issue of multimodal data for accurate prognostics with CNN-Transformer-LSTM network." In 2022 8th International Conference on Control, Decision and Information Technologies (CoDIT), vol. 1, pp. 280-285. IEEE, 2022.
- **CP2**: Sagar Jose, Ryad Zemouri, Mélanie Lévesque, Khanh T. P. Nguyen, Antoine Tahan, and Kamal Medjaher. "Informed machine learning for image-data-driven diagnostics of hydrogenerators." IFAC-PapersOnLine 56, no. 2 (2023): 11912-11917.
- **CP3**: Duc An Nguyen, Sagar Jose, Khanh T. P. Nguyen, and Kamal Medjaher. "Explainable multimodal learning for predictive maintenance of steam generators." In PHM Society Asia-Pacific Conference, vol. 4, no. 1. 2023.
- **CP4**: Sagar Jose, Khanh T. P. Nguyen, Kamal Medjaher, Ryad Zemouri, Mélanie Lévesque, and Antoine Tahan. "Bridging expert knowledge and sensor measurements for machine fault quantification with large language models." In 2024 IEEE International Conference on Advanced Intelligent Mechatronics (AIM), pp. 530-535. IEEE, 2024.
- **CP5**: Sagar Jose, Khanh T. P. Nguyen, Kamal Medjaher, Ryad Zemouri, Mélanie Lévesque, and Antoine Tahan. "From Fragments to Futures: Construction of Synthetic Run-to-Failure Trajectories for Fault State Prognostics." (**Best paper award**

winner at the 2024 Prognostics and System Health Management Conference (PHM 2024)).

Book Chapters

- **BC1**: Sagar Jose, Khanh T. P. Nguyen, and Kamal Medjaher. "Multimodal Machine Learning in Prognostics and Health Management of Manufacturing Systems." In Artificial Intelligence for Smart Manufacturing: Methods, Applications, and Challenges, pp. 167-197. Cham: Springer International Publishing, 2023.

6.3 Discussion of Findings

Our findings support the hypothesis that multimodal data could be a solution to address the several challenges faced by PHM research and the industrial community. Multiple methodologies were developed to address the challenges on the path to operationalizing this data, incorporating domain expertise, and leveraging historical maintenance records and foundation models with multimodal learning techniques. This research aligns with existing literature, demonstrating that the effectiveness of multimodal approaches in other domains is replicable in the field of PHM, and can be extended by introducing novel fusion techniques and data shortage mitigation strategies.

The implications of these findings are several, suggesting that industries can significantly reduce downtime and maintenance costs by using multimodal condition monitoring data to improve their predictive maintenance models. This research potentially opens up several research avenues previously unexplored due to data scarcity, which is perhaps the most critical roadblock in data-driven PHM research.

Our key findings, derived from extensive experiments and validations, underscore the potential of multimodal data based approaches in enhancing the reliability and efficiency of PHM processes. In detail, the findings from Chapter 3 demonstrate that the proposed multimodal model significantly outperforms simpler models, particularly when data imperfections are present. The resilience of the attention mechanism to increasing noise and missing data highlights its efficacy in real-world industrial settings where data quality is often compromised. This robustness is crucial for practical applications, ensuring reliable maintenance decisions even with imperfect data inputs.

In Chapter 4, our results demonstrate that the integration of domain knowledge proved highly beneficial, pointing to the need for more interdisciplinary approaches that combine machine learning with expert insights. The use of textual observations to weight other data features proved particularly effective, demonstrating the value of treating different

data types according to their unique characteristics. This integration of domain expertise ensures that the diagnostics model aligns closely with real-world operational conditions, enhancing its applicability and reliability.

Chapter 5 presents a groundbreaking solution for machine health prognostics in the absence of complete RTF data. The results demonstrate the efficacy of the proposed novel prognostics methodology, paving the way for refinement and application in various industrial contexts. This approach has the potential to lead to more generalized solutions for predictive maintenance across multiple sectors.

6.3.1 Implications for theoretical research

This thesis significantly contributes to the theoretical understanding of PHM, particularly in multimodal data integration. The methodologies developed herein not only advance the state of the art but also provide a robust framework for future research in this domain.

Specifically, it introduces innovative methodologies for integrating diverse condition-monitoring data sources. By demonstrating how different data modalities — such as sensor signals, visual inspection images, textual observations, and other monitoring data — can be effectively combined, this work enhances the **theoretical understanding of multimodal data fusion**. The cross-modal attention-based model, in particular, showcases the power of attention mechanisms in handling missing and noisy data. This contributes to the broader field of machine learning by providing a blueprint for building robust predictive models in environments where data quality is a critical limiting factor.

Another key theoretical contribution of this thesis is **the integration of domain expertise into the neural network design pipeline**. The diagnostics models rely on a generic foundation model for the first level of feature extraction, followed by a nuanced feature extraction round, in a pipeline that resembles a human expert building domain knowledge on top of a natural world model. This approach highlights the value of leveraging accumulated industry knowledge to guide the feature extraction process. By incorporating expert insights, the proposed methodologies ensure that the models are not only data-driven but also aligned with practical operational conditions. This fusion of domain knowledge with advanced machine learning techniques opens new avenues for interdisciplinary research, encouraging collaborations between machine learning researchers and industry experts.

Furthermore, the developed methodologies allow for **addressing one of the most pressing challenges in prognostics: the lack of complete RTF data**. By demonstrating the feasibility of creating RTF datasets from incomplete data, this work provides a valuable theoretical framework for addressing data limitations in other research contexts.

From a broader perspective, this research **contributes to the theoretical foundations of machine learning by illustrating how advanced architectures, such as attention mechanisms, foundation models and modular designs, can be applied to real-world problems**. The novel dataset stratification algorithm and the use of a routing gate are particularly noteworthy contributions that can inspire further theoretical explorations in model optimization and data handling.

6.3.2 Implications for industrial application

This thesis offers significant practical contributions that can transform maintenance practices in industrial settings, potentially enhancing operational efficiency and reducing costs. Specifically, the methodologies and models developed in this thesis can be **easily applied in industrial environments**. By effectively integrating multimodal data sources, these approaches can improve the detection, diagnostics and prognostics of potential failures, thereby reducing downtime and associated costs. This practical application can lead to more efficient maintenance schedules, better resource allocation and, ultimately, a more streamlined operational process.

Another significant implication of this research is **the encouragement it provides to data scientists and industrial researchers to utilize all available data sources**. Traditionally, data types with too few samples have been discarded. However, the multimodal learning techniques and design methodologies proposed in this thesis demonstrate the value of integrating diverse data modalities, including industrial documents and expert knowledge. This meticulous approach can uncover insights that might be missed when relying solely on more abundant data types, leading to more robust and informed maintenance decisions.

For large-scale prognostics projects, the modular deep learning approach introduced in Chapter 5 presents a **scalable and efficient development strategy**. This approach allows different teams to work on specific modules independently, adhering to general structural guidelines without needing to conform to rigid model training specifications. This flexibility can reduce the complexity and stress associated with wide-scope development projects. Additionally, the modular approach supports iterative scaling, which is crucial under resource constraints often encountered in practical industrial applications. This makes it feasible to gradually expand and refine predictive maintenance systems as more data and resources become available.

Additionally, the integration of domain expertise into the model design process **fosters interdisciplinary collaboration within the enterprise**. By combining the insights of industry experts with advanced machine learning techniques, the proposed methodologies

ensure that predictive models are both practically relevant and theoretically sound. This collaboration can lead to more effective solutions that are closely aligned with the specific needs and challenges of industrial operations.

Ultimately, the practical contributions of this thesis have the **potential to significantly enhance maintenance strategies in industrial settings**. By providing a robust framework for predictive maintenance that leverages all available data sources and integrates domain expertise, this research can lead to more accurate diagnostics, timely interventions, and optimized maintenance schedules. This not only improves the reliability and efficiency of industrial machines but also contributes to overall operational excellence.

6.4 Limitations

While this thesis has made significant contributions to the field of PHM, there remains considerable room for improvement. Below are some notable limitations, the reasoning behind them, and perspectives on their implications.

- **Dependence on supervised labels for diagnostics:** The diagnostics models in both chapters 4 and 5 are trained on labels assigned to condition monitoring data samples from domain experts. In Chapter 4, this is a strength of the methodology when it comes to visual inspection samples where domain experts have absolute certainty on the labeling. But even in this chapter, the data for which distinctive features are not well known had to be labeled with multiple "possible" types of degradation. In the later chapter, this was exacerbated by the expansion of the scope. Many data samples had to be labeled "Unknown", primarily due to the impracticality of developing dedicated feature extraction pipelines based on the existing domain knowledge about corresponding data. This is a limiting factor of the proposed methodology. While the proposed expert knowledge-based feature extraction pipeline (Chapter 4) improves model agreement with experts, this necessitates detailed study of each data modality while scaling up to more degradation types. This is counter to the "black box" feature extraction promise of deep learning. Therefore, the insistence of the methodology on incorporating expert knowledge is at once its strength and a limiting factor.

- **Assumption of degradation trend transferability:** In this first proof-of-concept, the partial RTF fragments from all available machines were used for constructing the RTF dataset. Only the feature proximity in the 2D projection was used as a criterion, and it was assumed that a trend observed on one machine could be transposed to another as long as their diagnostics features met the criteria listed in Chapter 5.

Indeed, the feature space supports this assumption to some extent, as not all samples with the same degradation type are clustered together. This suggests that additional factors, such as machine characteristics, may also influence the clustering. Furthermore, provisions are made in the synthetic edge generation algorithm to account for this uncertainty in the form of a likelihood factor. Therefore, it is possible that this grouping can be learned from data alone - the methodology makes provisions for this requirement. Yet, it would be the best case if domain knowledge could be used to form a rule-based framework for grouping certain machines where their observed trajectory can certainly be transferred between them. In this project, no such rules were obtained in the case of hydrogenerators. For practitioners replicating the methodology in other applications, it is recommended to consider this while generating RTF data.

- **Validation challenges at the limit of expert knowledge:** Indeed, depending on expert knowledge is a strength during the initial phases. However, this gets more challenging as the data-driven model pushes the limits of existing domain knowledge, especially when it comes to the validation of model predictions. Two instances in particular are notable. First, the validation of 2D projection of diagnostics features. For this step, it is challenging for human experts to compare relative degradation severity of two condition monitoring data samples of different modalities, from different machines. For example, asking an expert to analyze a visual inspection photograph from one machine and an electromagnetic reading from another machine, and then provide a comparative analysis, is a time-consuming process. Furthermore, one has to account for human elements such as decision fatigue, anchoring bias, confirmation bias, recency bias, framing effect, and many other factors. To account for many of these, we set up this validation step as a randomized blinded trial, where the expert was only given samples and no other information such as the model predictions on them. The validation was tackled in multiple sessions to avoid decision fatigue. However, the time-consuming nature of this step is a necessary concession due to its importance in the methodology.

  Secondly, after the model has made future health forecasts for each machine, validation becomes even more difficult. For machines that have never run to failure in reality, validating the timeline predicted by the prognostics model is a difficult task for human experts. We fully acknowledge that there are limitations to the validation of the prognostic results. Numerical quantification of model metrics on a test set is only somewhat useful when the test set is not a known absolute ground truth of machine health evolution. However, the reality is that in the absence of data or concrete knowledge, an abstract validation from experts is the best form of validation that can be obtained. For the industrial case study, it may be necessary to let these machines run for some time to see if the predictions align with future observations.

- **Distribution shift:** The study proposes a prognostics methodology based on observed partial health evolution segments from multiple machines. However, the assumption is made that the machines in the dataset collectively represent the full range of possible evolution paths. That is, it does not account for natural distribution shifts. That said, this is suitable for the case study of hydrogenerators, where the data come from several decades of monitoring, and operating conditions are not expected to shift significantly. For other applications where distribution shifts can be expected, the scalability of the modular diagnostics model implies that adapting the model to new distributions will be significantly less costly than training a new model. Thus, while distribution shifts are not explicitly considered, the methodology is designed to support smooth adaption to shifts during lifelong learning.

- **Resource constraints:** The models developed in this model use several large foundation models in the feature extraction phase. It is acknowledged that larger and more sophisticated foundation models are rapidly being released for images, text, and other modalities, and using the later models may lead to better model performances. Even during development time, the largest available models of the day were not used, as this project had to balance between resource availability and results.

6.5 Recommendations for Future Research

This thesis presented an initial proof-of-concept of the potential of multimodal data to address several challenges in the PHM field. The findings reveal both limitations to address and new avenues to explore, some of which are listed below.

- **Short time segment predictions to enable real-world validation:** The graph masked autoencoder prognostics model in Chapter 5 predicts the future health evolution of a machine in one shot. This indeed renders the prediction difficult to validate. Instead, a conditional model can be trained to generate future possibilities at a given time t in the future, or a generative model can be trained to predict edges of a shorter time segment. For example, a graph RNN can be trained to predict edges of one one-month time period. Then, for validation, it may not be difficult for the industry to let the machine run for a few months and evaluate model predictions against observed reality. This would be an absolute validation and provide solid grounds for model refinement.

- **Uncertainty quantification of each methodology steps:** The model accumulates uncertainty at each step of the methodology, where these could be potentially quantified for better-informed decision making. In particular, the uncertainty of the

supervised classification model, the 2D feature projection, the radius-based RTF sequence generation step, and the prognostics prediction uncertainties must be quantified in a cascading uncertainty model. Researchers interested in replicating and extending the proposed methodology are invited to incorporate uncertainty quantification at relevant steps.

- **Multimodal foundation models:** The proposed models relied on a unimodal feature extraction step before multimodal learning. While this was a deliberate decision, the design choice came at the cost of an opportunity to explore multimodal foundation models for feature extraction. Therefore, future research might consider multimodal learning from the very first steps by using a multimodal foundation model for simultaneous feature extraction from multiple data modalities.
- **Cost benefit analysis:** The proposed methodology involves collecting multiple modalities of data - which could be expensive, or not, depending on the application. It also involves incorporating human experts in the development pipeline, which implies time cost. Finally, there are resource costs involved in fine-tuning large image and text models, as well as an array of expert models. While the costs may be trivial for certain applications - such as hydro power generation centers - it may be different for medium-scale enterprises. Therefore, a cost-benefit analysis of the proposed methodology may be of value to industrial practitioners.

6.6 Closing Statement

Reflecting on the research journey, it becomes clear that the integration of multimodal data and domain knowledge is not just a theoretical pursuit but a practical necessity in modern industrial maintenance. This research underscores the importance of interdisciplinary approaches in solving complex industrial problems, highlighting the convergence of data science, engineering, and domain expertise. The landscape of data-driven multimodal learning has changed significantly in recent years. Yet, there is much that remains to be done, and the future is as exciting as it is promising.

Every scientific endeavor lays another stone on the road between our present and a future where science has crafted a better day for humanity. In the pursuit to make the vision of self-monitoring, self-regulating, self-healing machines a reality, we hope that those who walk this path after us may find this little cobblestone we have laid to be steady and true, giving them a firm footing from which to leap far beyond our imaginations.

Appendix A

Algorithms

This appendix presents several algorithms relevant for reproducibility of the methodologies presented in this thesis.

A.1 Multilabel co-occurrence calculation

The algorithm to calculate the label co-occurrence matrix during exploratory data analysis of multilabel classification scenarios (Chapter 5, section 5.4.2) is presented in Algorithm 5.

Algorithm 5 Calculate Label Co-occurrence Matrix

Input: DataFrame df, Class Columns C
Output: Co-occurrence Matrix M
procedure CalculateCooccurrence(df, C)
    Initialize M with zeros, size $|C| \times |C|$
    for each record r in df **do**
        Extract label subset L where $r_L = 1$
        for each pair (i, j) in L **do**
            $M_{ij} \leftarrow M_{ij} + 1$
            $M_{ji} \leftarrow M_{ji} + 1$ ▷ Matrix is symmetric
    return M

A.2 Graph masking

An algorithm to mask directed graphs from the end at a given masking rate is used in Chapter 5, section 5.3.5. This is given in Algorithm 6.

Algorithm 6 Graph Masking Algorithm

$maskRates \leftarrow [0.1, 0.2, 0.3, \ldots, 0.9]$
for each graph G in train set **do**
  for each $rate$ in $maskRates$ **do**
    $totalNodes \leftarrow$ NumberOfNodes(G)
    $totalEdges \leftarrow$ NumberOfEdges(G)
    $nm \leftarrow \lceil rate \cdot totalNodes \rceil$
    $em \leftarrow \lceil rate \cdot totalEdges \rceil$
    $nodesToMask \leftarrow$ GetLastNodes(G, nm)
    $edgesToMask \leftarrow$ GetLastEdges(G, em)
    $G_{masked} \leftarrow$ MaskNodesAndEdges($G,$)
  ($nodesToMask, edgesToMask$)
    SaveGraph($G_{masked}, rate$)

A.3 Synthetic graph edge probability assignment

In Chapter 5, section 5.3.4.2, a methodology was presented to construct synthetic RTF trajectory fragments by analyzing feature proximity on diagnostics feature space. While the graph model can learn the likelihood of these synthetic edges representing realistic transitions with sufficient data, it is also possible to expedite this learning with engineering this probability as an edge feature. This process is described below.

The synthetic edge generation is run recursively for a fixed number of generations over a range of radii. This entails that a synthetic edge generated in one iteration of the algorithm is treated as a 'real' edge in a subsequent iteration, with the search radius expanded accordingly. Throughout this process, a generation count is maintained for each edge, which is used to apply a decaying factor to the edges, thereby adjusting their influence in the analysis. Consequently, a synthetic edge derived directly from a real edge is assigned a higher likelihood than an edge formed from another synthetic edge. Each generation signifies a degree of derivation from the original dataset, enabling the exploration of further potential states and transitions.

The likelihood value is a measure of how likely an edge is to exist. This measure is calculated based on transition density (the number of edges near the end node of a real edge) and the proximity of these edges' terminal nodes to each other. This metric serves to reduce uncertainty regarding degradation trends. Additionally, a decay factor is incorporated into this value, providing a means to address uncertainties about whether the proximity of features necessarily implies similar outcomes.

If the dataset includes a broad spectrum of machines, conditions, and transition scenarios and accurately represents the distribution of transition times and their probabilities,

Algorithm 7 Assigning Probabilities to Synthetic Transitions

Input: synthetic_edges_df, data_sorted, radii
Output: synthetic_edges_df with probabilities

procedure CalculateTransitionDensity
    $decay\_factor \leftarrow 0.9$
    for each synthetic_edge in synthetic_edges_df **do**
        $r \leftarrow$ radius from synthetic_edge
        $end_node \leftarrow$end node from data_sorted
        Initialize $K \leftarrow 0$
        for each edge in synthetic_edges_df with same start and r **do**
            Calculate distance $start\_node$ of edge to end_node of synthetic_edge
            if distance $\leq r$ **then**
                $K \leftarrow K + 1$
        $generation \leftarrow$ generation of synthetic_edge
        $adjusted\_K \leftarrow K \times (decay\_factor)^{generation}$
        Assign $adjusted\_K$ to synthetic_edge

procedure NormalizeTransitionDensity
    for each start_node_index group in synthetic_edges_df **do**
        Calculate total K for the group
        if total $K > 0$ **then**
            Normalize K values in the group
        else
            Assign 0 to normalized K in the group

then the model can directly learn the underlying patterns from the data. When the data is not comprehensive enough, incorporating density-based probability adjustments into the time feature assignment can be viewed as a form of explicit feature engineering to help the model.

A.3.0.1 Perspectives on RTF data generation algorithm

Based on the results obtained in the application of the methodology (Chapter 5, section 5.5), several observations are made. First, if the model is exposed to a diverse and comprehensive dataset of graph edges that includes actual transition times between machine states, then the model would learn to model the different speeds of the machines from the data alone. In a prognostics project, the data usually has to contain several instances of full RTF trajectories. The contribution of the methodology is in relaxing this requirement to partial trajectories. If the data contains sufficient instances of partial transitions, the

constructed dataset will be enough to let the model learn the different speed characteristics.

Secondly, in the case that the data does not contain sufficient examples, the model further proposes another explicitly engineered density measure to encourage this learning. The algorithm to assign probabilities to synthetic transitions within a network of machine diagnostics seeks to address the inherent uncertainty in predicting the degradation paths of different machines based solely on feature proximity in a visualized 2D space. By leveraging the concept of transition density, particularly through the use of a decaying factor that adjusts based on the generation of synthetic edges and the aggregation of similar transitions within a local neighborhood, this approach introduces a nuanced layer of probabilistic reasoning to the generation of synthetic edges.

This probabilistic weighting accounts for the varying likelihoods that machines with diagnostics features in close proximity might follow similar degradation trajectories. The critical insight here is the recognition that while spatial closeness in feature space suggests potential similarity in machine behavior, differences in design, usage, and operational conditions could lead to divergent outcomes. The potential discrepancies in degradation speeds are addressed through the probabilistic weighting of these synthetic transitions. The calculation of transition densities serves as a mechanism to modulate the influence of each synthetic transition based on its proximity to the original data and the density of similar transitions within its local neighborhood. By modulating the influence of synthetic edges with the calculated transition densities and normalizing these within groups, the algorithm effectively simulates a spectrum of possible outcomes. High-density transitions suggest a greater consensus among the data that certain paths are more plausible, thereby reinforcing these connections' validity in the synthetic model. Conversely, lower-density transitions, which reflect more unique or less commonly observed paths, are weighted accordingly, capturing the uncertainty and diversity in machine degradation patterns.

Thus, the algorithms (4 and 7) does more than merely extrapolate from existing data; it constructs a probabilistic model that mirrors the real-world complexity and uncertainty of machine degradation. The incorporation of time differences as edge features in the real and synthetic transitions further enriches this model, allowing for an implicit consideration of degradation speeds. By embedding the temporal progression through edges, the method introduces a dynamic element that reflects the temporal dimension of machine degradation. The probability assignments thus not only account for the similarity in degradation paths but also embed consideration of varying degradation speeds. This approach allows for a more nuanced and realistic simulation of potential degradation evolution pathways, providing a valuable tool for predictive maintenance and operational optimization.

In essence, while the method starts with an assumption of similarity based on feature proximity, its probabilistic treatment of synthetic transitions, coupled with the integration of temporal dynamics, offers a detailed approach to capturing the complex, uncertain na-

ture of machine degradation across different operational contexts. This approach provides a foundation for more informed predictive maintenance strategies by acknowledging and quantifying the inherent uncertainty in extrapolating from known data points to forecast future machine states and their timing. The methodology is successful in the application to an industrial scenario with severe data limitations.

Appendix B

Model Designs for Modular Architecture

In Chapter 5, several neural network models were trained to constitute the modular deep learning ecosystem used for diagnostics. Below, we present the architecture and description of some of the most activated expert models. In addition, other model designs such as the routing gate architecture will also be detailed.

B.1 Expert model for all images

Figure B.1 illustrates an expert model specialized in classifying image data. Given that image data relate to all target classes, having an image expert capable of classifying main class groups (partial discharge, contamination, thermal, unknown) is crucial. As image data often come with accompanying notes in this dataset, the model exclusively processes image-text pairs. Its architecture is designed to extract features from both text and image. For text, GPT-2 (Radford *et al.* (2019)) serves as the foundation model, fine-tuned on industry-specific domain knowledge documents, which then embeds the inspection notes associated with visual inspection images, forming one branch of the neural network. For images, the model uses VGG16 as the foundation model, followed by a faster-RCNN optimized for detecting partial discharge (Jose *et al.* (2023b)). Next, a CNN classifier learns features to detect partial discharge residues, contamination, thermal degradation, or none. Textual observations from inspection personnel accompany the visual inspection, used to weigh image features with an attention layer, ensuring careful processing of both modalities. For other data, inputs are replaced with zero matrices, forming a feed-forward connection. Branches are fused via concatenation and then forwarded to an output layer. Class T4 receives a higher weight due to limited samples, critical for accurate classification since it's only detectable from images. Thus, data imbalance handling occurs at the expert level.

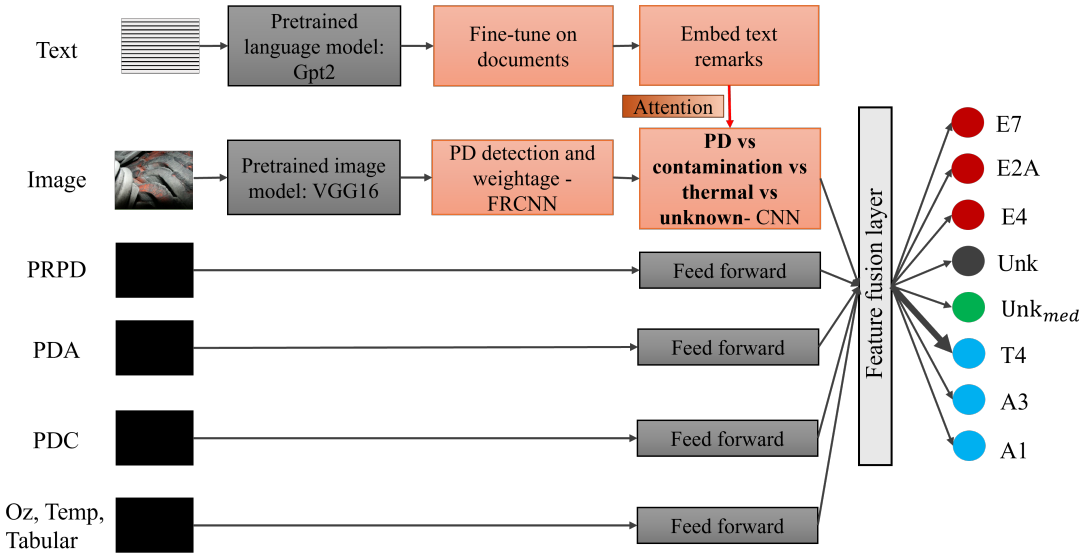


Figure B.1: Expert model trained only on image+text data modalities. For training this model, the absent modalities are represented with a zero vector.

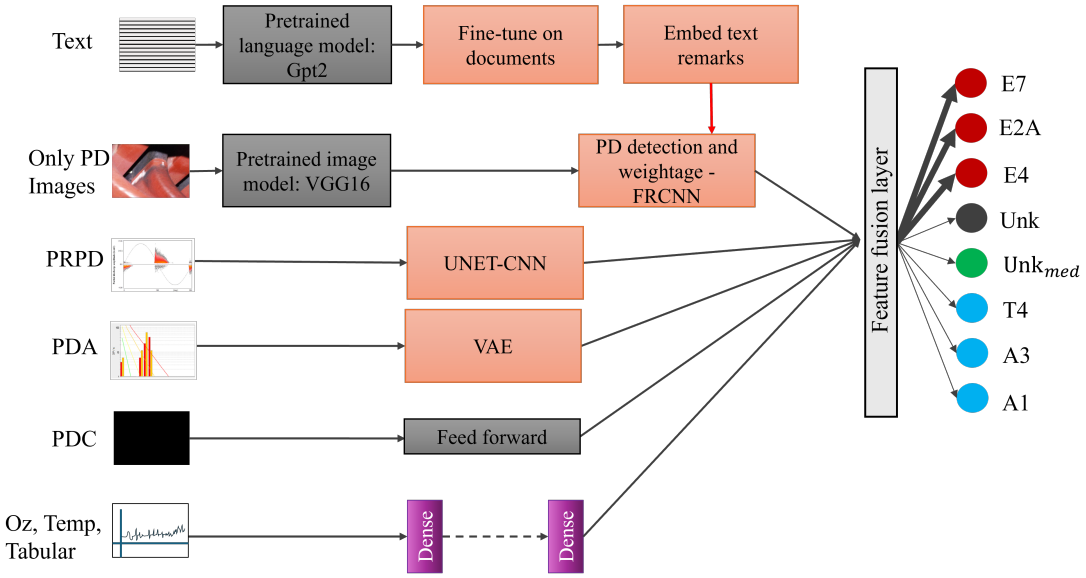


Figure B.2: Expert model to distinguish between the three types of partial discharge states, which show a high correlation in the dataset. This is an extension of the model trained in Chapter 4.

B.2 Expert model for all partial discharge types

The second expert model, depicted in Figure B.2, focuses on distinguishing between three partial discharge states: E4, E2A, and E7. These states are identified through visual inspection images, PRPD, PDA, ozone, temperature, and other tabular data. Data instances with only PDC or unknown states are excluded. The dataset for training the expert undergoes a carefully designed feature extraction pipeline for each data type (for details, see Jose *et al.* (2023a)). At the output layer, E4, E2A, and E7 states are assigned high weights.

B.3 Expert model for PDC data

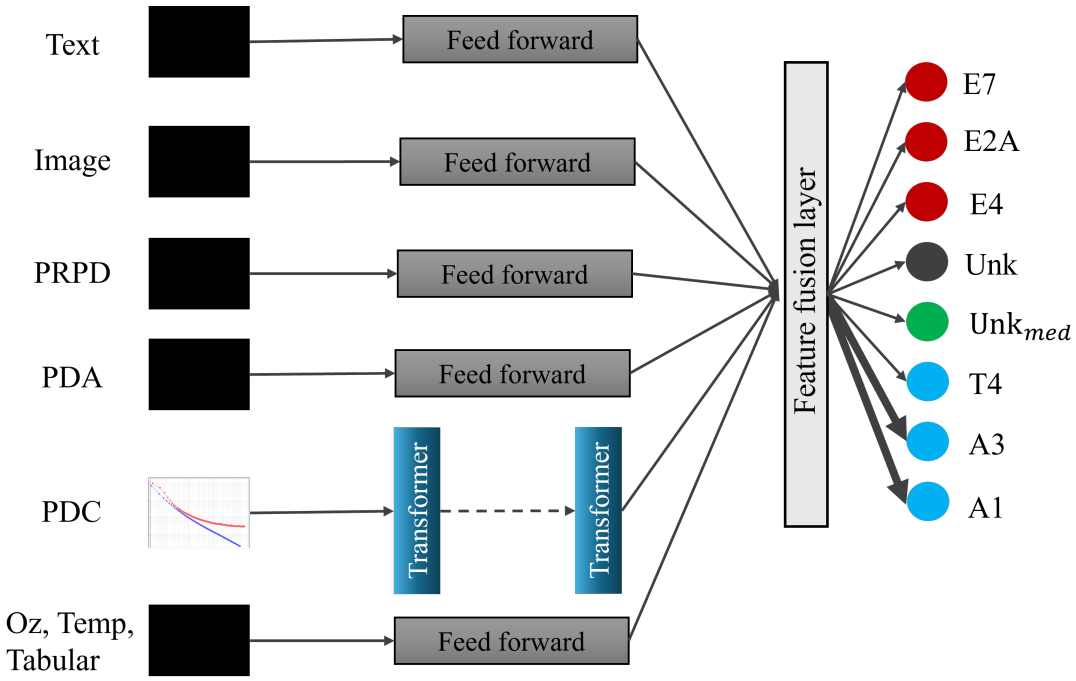


Figure B.3: Expert model to distinguish between conducting and non-conducting contamination from PDC only.

Figure B.3 illustrates another expert model trained solely on polarizing and depolarizing currents (PDC) data. PDC uniquely distinguishes between conducting and non-conducting contamination. While contamination presence can be detected from photographs, distinguishing between different types (e.g., oil and dust vs. water and carbon) is challenging. PDC data, consisting of two-time series, detect conduction based on end-to-end value differences. Stacking transformer layers suffices for feature extraction. With approximately 500 PDC samples compared to around 100 contamination images, this model efficiently

classifies contamination types from PDC-only instances. The model assigns higher weights to outputs A1 and A3.

B.4 Gate architecture

In Chapter 5, section 5.3.2.3, the idea of a routing gate to select most suitable expert models for each incoming multimodal data samples was introduced. In the case study, a gate architecture design based on a transformer stack was used. The architecture is given below:

Given the multimodal data inputs $x_1, x_2, \ldots, x_n$, where each x_i is a data sample from modality i, the concatenated input vector is prepared for input to transformer layers as:

$$X_{\text{concat}} = [x_1, x_2, \ldots, x_n]$$

$$X_{\text{flat}} = \text{Flatten}(X_{\text{concat}})$$

$$X_{\text{embed}} = W_{\text{embed}} X_{\text{flat}} + b_{\text{embed}}$$

where W_{embed} and b_{embed} are the parameters of the dense layer, transforming the flattened input into a fixed-size embedding suitable for processing by the transformer.

The transformer stack processes the embedded input through several layers, each consisting of multi-head self-attention and position-wise feedforward networks. Let T_k represent the k-th transformer layer in a stack of L layers.

- W_i^Q, W_i^K, W_i^V are the weight matrices for the queries, keys, and values respectively for each head i. - W^O is the output weight matrix that projects the concatenated results of all attention heads back to the transformer's model dimension, d_{model}.

Multi-Head Self-Attention (MHSA) is computed as:

$$\text{MHSA}(H^{(k-1)}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^O$$

where each head head_i is computed as:

$$\text{head}_i = \text{Attention}(H^{(k-1)}W_i^Q, H^{(k-1)}W_i^K, H^{(k-1)}W_i^V)$$

and Attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\text{Concatenated Output} = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)$$
$$\text{Projected Output} = \text{Concatenated Output} \cdot W^O$$

The output of the MHSA, now appropriately dimensioned, is further processed by the position-wise feedforward network FFN within each transformer layer:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Each transformer layer T_k applies these functions to the output of the previous layer $H^{(k-1)}$ (or X_{embed} for $k = 1$):

$$H^{(k)} = \text{FFN}(\text{MHSA}(H^{(k-1)}))$$

The output from the transformer, $H^{(L)}$, is passed through a softmax layer to calculate the probability distribution over the experts:

$$P = \text{softmax}(W_p H^{(L)} + b_p)$$

where W_p and b_p are the trainable parameters of the softmax layer, mapping the transformer output to a distribution over experts.

While there are many ways to implement a routing gate module, this design is intended to serve as a useful guideline for practitioners interested in reproducing the methodology.

Appendix C

Ablation Studies on Diagnostics Model

In Chapter 5, we designed a modular deep learning architecture consisting of a number of task-specialized expert neural network models trained to optimize a subtask of the overall diagnostics problem, with a routing gate trained to select appropriate experts for each incoming test sample. Below, we present and discuss the results of the proposed methodology on the test set, roughly 20% of the full dataset. Considering 8 output classes related to 8 degradation types investigated in this study, each having values of either 0 (degradation type absent) or 1 (degradation type present), there exist 256 possible label combinations, e.g., E2A=0, E7=0, E4=0, T4=0, A1=0, A3=0, Unknown medium=0, Unknown=0.

To evaluate the performance comprehensively across all label combinations, various metrics are employed, including exact match ratio, Hamming loss, Jaccard score (micro average, weighted average, and sample-wise average), average precision, average recall, average F1-score, and log loss. The details of these metrics can be consulted in Sorower (2010), Park and Read (2019).

In the ablation study, we systematically explore the impact of varying model architectures on performance by adjusting the number of active experts and feature dimensions across twelve distinct configurations. First, a non-modular classification model is simply trained on the training set. This will be compared with the most basic configuration of modular approach. The ablation tests then alter the number of experts from 1 to 4 and adjust the expert fusion feature dimensions to 8, 16, and 32, while maintaining a consistent branch fusion layer size of 32 across all configurations. Each model employs a structured layer sequence that transitions from the fusion dimension to intermediate layers of 16 and 8 dimensions, culminating in the output layer.

Performance metrics, including Hamming loss, log loss, and Jaccard score, as well as precision, recall, F1-score, and exact match ratio, are detailed in Tables C.1 and C.2, providing a comprehensive evaluation of the effectiveness of each configuration. The tables show that even with only one active expert and small feature size configuration, the performance is significantly better than a non-modular approach. A visualization of the scores are also shown in Figure C.1, where the improvement in each metric by increasing

the active experts and feature size is clearly shown. The qualitative metrics from different configurations are discussed henceforth.

Table C.1: Hamming loss, log loss and Jaccard score of model configurations.

Active experts	Feature size	Hamming loss	Log loss	Jaccard score (micro)	Jaccard score (weighted)	Jaccard score (sample wise)
0 (Non-modular)	-	0.46	16.43	0.31	0.41	0.53
1	8	0.26	9.19	0.52	0.62	0.73
1	16	0.25	8.94	0.53	0.63	0.74
1	32	0.22	7.96	0.57	0.66	0.76
2	8	0.79	6.75	0.62	0.7	0.8
2	16	0.18	6.51	0.63	0.71	0.8
2	32	0.17	6.06	0.65	0.73	0.82
3	8	0.13	4.86	0.7	0.78	0.85
3	16	0.13	4.83	0.7	0.78	0.85
3	32	0.12	4.26	0.73	0.8	0.87
4	8	0.1	3.72	0.76	0.82	0.88
4	16	0.09	3.41	0.78	0.83	0.89
4	32	0.06	2.3	0.84	0.88	0.92

Table C.2: Average precision, recall, F1-score and exact match ratio.

Active experts	Feature size	Precision (micro)	Precision (weighted)	Precision (sample wise)	Recall (micro)	Recall (weighted)	Recall (sample wise)	F1 Score (micro)	F1 Score (weighted)	F1 Score (sample wise)	Exact match ratio
0 (Non-modular)	-	0.41	0.67	0.54	0.55	0.55	0.54	0.47	0.57	0.44	0.49
1	8	0.63	0.79	0.74	0.73	0.76	0.74	0.68	0.75	0.74	0.69
1	16	0.64	0.8	0.74	0.74	0.74	0.75	0.69	0.76	0.74	0.7
1	32	0.67	0.81	0.77	0.77	0.77	0.77	0.72	0.78	0.77	0.72
2	8	0.72	0.84	0.8	0.8	0.8	0.8	0.76	0.81	0.8	0.76
2	16	0.73	0.84	0.81	0.81	0.81	0.81	0.77	0.82	0.81	0.76
2	32	0.74	0.85	0.82	0.82	0.82	0.82	0.78	0.83	0.82	0.78
3	8	0.79	0.87	0.86	0.86	0.86	0.86	0.82	0.86	0.86	0.81
3	16	0.79	0.88	0.86	0.85	0.85	0.86	0.82	0.86	0.86	0.81
3	32	0.81	0.89	0.87	0.87	0.87	0.87	0.84	0.87	0.87	0.82
4	8	0.83	0.9	0.89	0.89	0.89	0.89	0.86	0.89	0.89	0.84
4	16	0.85	0.91	0.9	0.89	0.89	0.9	0.87	0.9	0.9	0.85
4	32	0.9	0.93	0.93	0.92	0.92	0.93	0.91	0.93	0.93	0.88

Table C.1 examines the impact of varying the number of active experts and feature sizes across model configurations based on three key performance metrics: Hamming loss, Log loss and Jaccard score.

- **Hamming loss:** There is a consistent decrease in Hamming loss when both the number of active experts and feature sizes increase. This trend suggests that models with higher complexity featuring more experts and larger features are more effective in minimizing incorrect label predictions per sample.
- **Log loss:** Reflecting the confidence and accuracy of model predictions, log loss also shows a substantial decrease with increasing model complexity. The simplest model (1 expert, feature size 8) recorded a log loss of 9.19, indicating lower prediction confidence. Conversely, the most complex one (4 experts, feature size 32) demonstrated a significant improvement, with a log loss of 2.3, suggesting enhanced prediction accuracy and confidence.
- **Jaccard score:** Improvements in the Jaccard score (measured micro, weighted, and sample-wise) positively correlate with increases in the number of experts and feature sizes. Notably, the sample-wise Jaccard score highlights the enhanced capability of more complex models to accurately predict label sets on an individual sample basis, thereby improving relevance in predictions across the dataset.

Table C.2 reveals distinct patterns in performance variation across different model configurations, emphasizing the effects of varying the number of active experts and feature sizes measured by several metrics.

- **Recall metrics:** The three average recall metrics (micro, weighted, and sample-wise) remain consistent across all model sizes. This consistency suggests that the models' capability to identify relevant labels is stable, irrespective of model complexity.
- **Micro-averaged precision:** This metric is consistently lower compared to weighted and sample-wise averages, indicating a higher proportion of false positives across all labels, treating each class equally regardless of frequency.
- **Weighted-average precision:** Higher values suggest that the models perform well on more frequent labels, which dominate the calculation due to their higher weights.
- **Sample-Average Precision:** Indicates effective prediction of correct positive labels on a per-sample basis.
  The obtained results indicate that increasing the number of activated experts and feature size contributes to a reduction in false positives, as evidenced by the smaller disparity among the three precision metrics in more complex models. Additionally, the models tend to overpredict labels, particularly with fewer active experts, which affects the micro-averaged precision but not as much as the other precision metrics.
- **F1 score:** One can see a uniform improvement in micro, weighted, and sample-wise F1-scores with increased complexity in the models, characterized by more active

experts and larger feature sizes. The weighted F1-score, which takes into account the frequency of labels, typically shows a higher score than the micro-average for models with fewer experts and smaller feature sizes. This suggests that these models are better at correctly predicting more frequent labels as opposed to rare ones.

- **Exact match ratio:** Improvements in the exact match ratio are noted as the number of experts and feature size increase, signifying better overall performance in predicting precise label combinations.

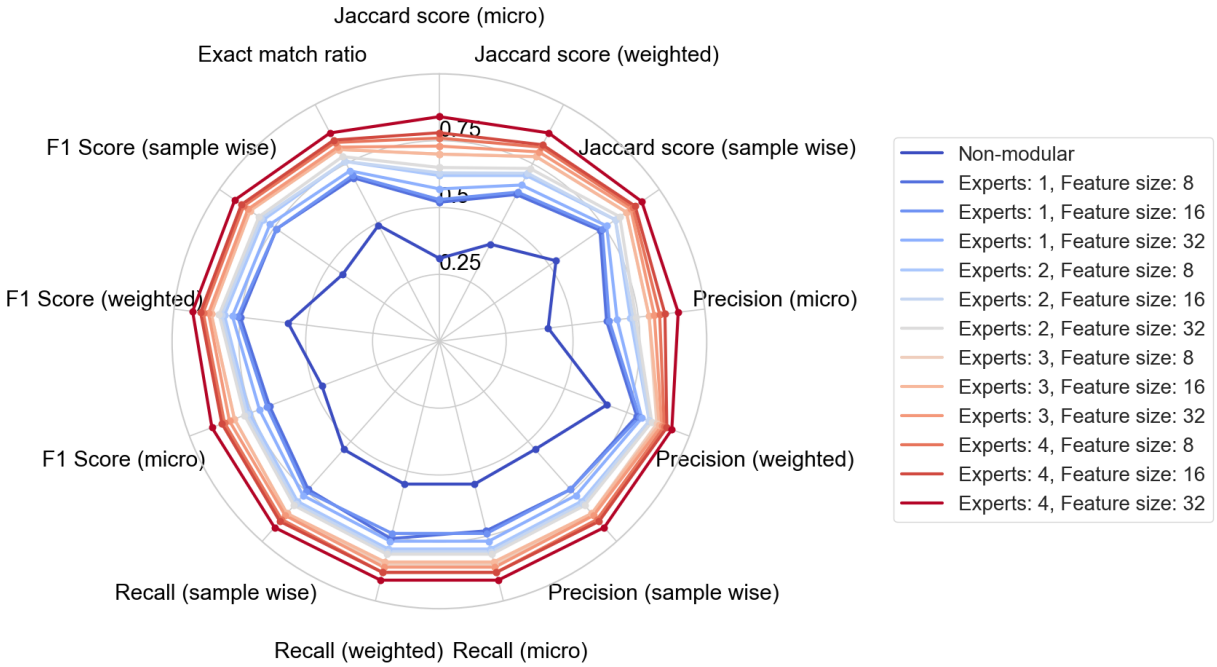


Figure C.1: Radar chart comparing the metrics (Jaccard score, precision, recall, F1 Score and exact match ratio) clearly show the performance difference between non-modular and modular approach, as well as the improvement on increasing the number of active experts and feature size.

The aforementioned ablation study substantiates that enhancing the number of experts and the dimensionality of features within an ensemble model markedly improves performance. Notably, the most sophisticated configurations — employing four experts with a feature size of 32 — exhibit pronounced synergistic effects, enabling more detailed data representations through the aggregation of varied expert insights. Such ensemble methods excel in discerning diverse data patterns, which is instrumental in the superior generalization to novel data. Moreover, increased complexity within this well-regularized framework paradoxically acts as a safeguard against overfitting by promoting the learning of broad patterns over narrow data-specific memorization.

When applied to imbalanced datasets, the advanced model approaches ideal performance metrics. The empirical evidence points to a significant descent in log loss from 9.19 to 2.3, a climb in the exact match ratio to 0.88, and a peak in Jaccard scores at 0.92. These metrics attest to the ability of complex model structures to counterbalance dataset imbalances effectively, thereby optimizing performance across all pivotal metrics.

Appendix D

Ablation Studies on Health Index Calculation with Text Data

In Chapter 4 section 4.5, we presented an extension to the diagnostics model to calculate degradation level (health index) by incorporating text data. This appendix presents the details of the ablation experiments done to validate the methodology and the results. The experiments are listed as follows:

1. **No text input:** Modified the output layer of the existing diagnostics model to perform a regression task targeting the degradation level. This modification involves retraining only the model's final output layer. This initial experiment sets a baseline for estimating degradation intensity using only the quantitative data collected during inspections, without any textual annotations from technicians (Figure D.1).

2. ***FrWac2Vec*(text) + no fine-tuning + direct input:** Embedded the text data (technician's remarks) using a small off-the-shelf model for French text embedding model (*FrWac2Vec*) and added it as an additional input. Only the technicians' notes were used, excluding other text data, e.g., guidelines for technicians, forming the second baseline This experiment explores the minimum performance enhancements from adding text remarks (Figure D.2a).

3. ***FrWac2Vec*(text) + no fine-tuning + attention weight input:** The notes are embedded using *FrWac2Vec* without any fine-tuning. Instead of providing the text as a direct input, it is used to weigh the features derived from other inputs. This experiment explores the assumption that the text primarily offers observations related to other measurements (Figure D.2b).

4. ***FrWac2Vec*(text) + fine-tuning + direct input:** The small embedding model *FrWac2Vec* is first fine-tuned on industrial text documents such as guidelines and standards. The inspection notes are embedded using this fine-tuned model and provided as a direct input. This explores the value of providing context for embedding the inspection notes (Figure D.2c).

5. ***FrWac2Vec*(text) + fine-tuning + attention weight input:** This experiment combines both fine-tuning the small embedding model *FrWac2Vec* and using the embedded inspection notes to attention weight other measurements (Figure D.2d). This concludes the experiments with the small embedding model.

6. **Gpt2-large(text) + no fine-tuning + direct input:** Here, the first embedded inspection notes with an LLM (Gpt2-large) is attempted for the first time. The LLM is used without any fine-tuning, and the embedded notes are added directly as input. Given that most open-source LLMs are trained on diverse, general text data from the internet, this study aims to determine whether an LLM trained on such a broad corpus can enhance the extraction of valuable information from industrial text data (Figure D.2c).

7. **Gpt2-large(text) + no fine-tuning + attention weight input:** Here, the inspection notes embedded by Gpt2-large is used to weight the other data features (Figure D.2b).

8. **Gpt2-large(text) + fine-tuning + direct input:** In this experiment, Gpt2-large is fine-tuned on internal company documents including standards, norms, and guidelines. The aim is to examine the effects of fine-tuning an LLM to specific contextual needs. The inspection notes embedded by the fine-tuned LLM is then introduced as an additional data source (Figure D.2c).

9. **(Proposed method) Gpt2-large(text) + fine-tuning + attention weight input:** This final setup brings together all the elements of the proposed methodology. The LLM (Gpt2-large) is fine-tuned on the documented domain knowledge, the inspection notes are embedded using this fine-tuned LLM, and this is used to weight other condition monitoring data features (Figure D.2d, 4.32).

In Experiment 2, where inspection notes are incorporated as an input modality using a simple off-the-shelf French language embedding model, there is a significant reduction in the MAE by $\approx 30\%$ compared to the baseline model's performance.

Experiment 3 explores an alternative use of text by employing inspection notes to weigh other input data, rather than serving as a direct input modality as in Experiment 2. This approach, which aligns with the practical application of inspection notes, results in a performance improvement of $\approx 12\%$ compared to Experiment 2.

Experiment 4 evaluates whether using industrial guideline documents, which technicians rely on as a knowledge base, can enhance diagnostic performance when integrated into the text embedding model. It fine-tunes the simple embedding model (*FrWac2Vec*) using documents such as inspection guidelines and standards. The inspection notes, once

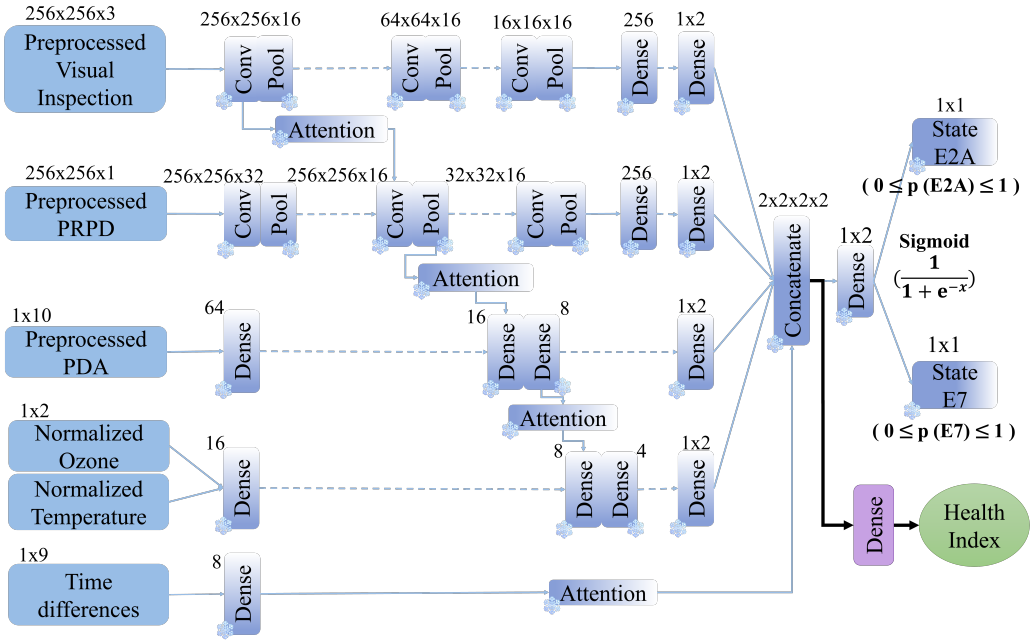
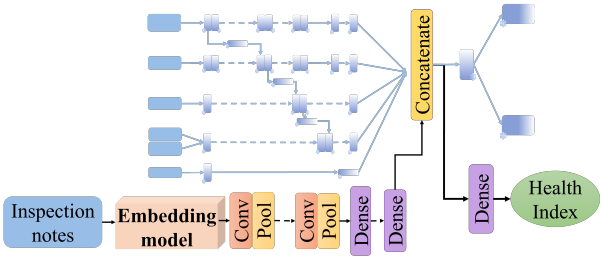


Figure D.1: Experiment 1 - Modification of output layer to perform regression. All the layers from the previously trained classification model (see Jose *et al.* (2023a)) is frozen. New output layer is added from the fusion layer, to predict degradation level from condition monitoring data. No text data is used in this setup.

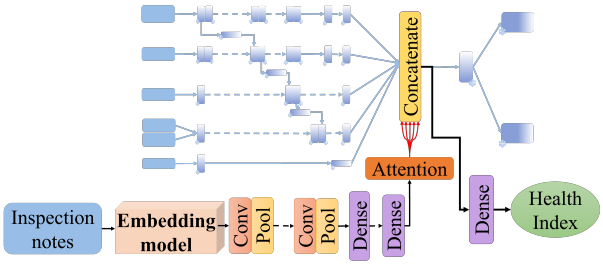
embedded in this fine-tuned model, are input directly into the diagnostics. Compared to Experiment 3, this method resulted in a modest performance improvement of $\approx 4.75\%$.

Experiment 5 modifies the approach of Experiment 4 by using the inspection notes to weigh other inputs, achieving a modest performance improvement of $\approx 6.5\%$. Both Experiments 4 and 5 illustrate the limitations of using an embedding model trained on a small text corpus, even with fine-tuning. This approach proves inadequate for assimilating the domain knowledge embedded in the texts.

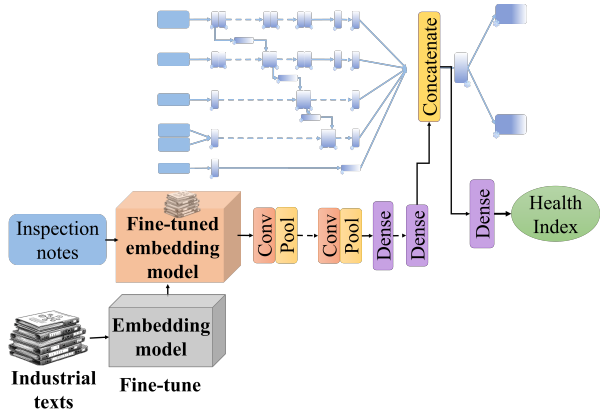
Experiment 6 explores the effectiveness of using LLMs to process industrial text. In this setup, an LLM, without any fine-tuning, embeds the inspection notes which are then directly input into the model. This approach significantly enhances performance, reducing the error by more than 35% compared to Experiment 5. Then, experiment 7 integrates the concepts of utilizing large language models to embed inspection notes and using these embeddings to weigh other inputs. This strategy results in a significant error reduction, with a decrease of $\approx 35\%$ compared to Experiment 6. Experiment 8 studies using LLM to embed the inspection notes after fine-tuning the LLM with domain-specific texts. In this configuration, the fine-tuned embeddings are directly input into the diagnostics model. This method results in an increased MAE of 14.6.



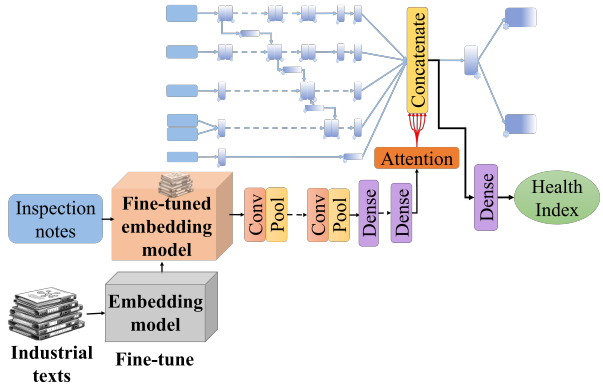
(a) Inspection notes input directly as a data source via an off-the-shelf text embedding model (Experiments 2, 6).



(b) Inspection notes embedded with an off-the-shelf text embedding model weights other data features with attention (Experiments 4, 7).



(c) Inspection notes input directly via a text embedding model fine-tuned on domain knowledge (Experiments 3, 8).

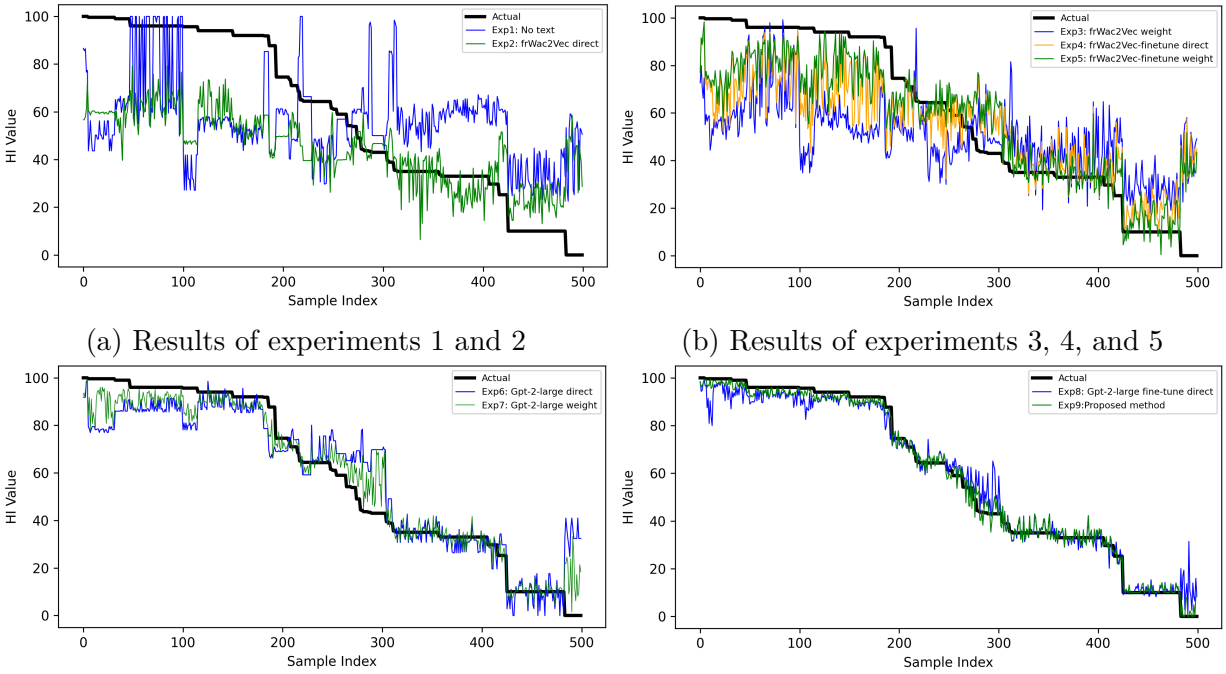


(d) Inspection notes embedded with a fine-tuned text embedding model weights other data features with attention (Experiments 5, 9).

Figure D.2: All experiment setups showing different arrangements of using text data to augment the model. The experiments differ by the way text features are merged with other modality features (direct or attention weight on other data), the embedding model (small model *FrWac2Vec* or LLM Gpt2-large), and whether the embedding model is finetuned on the domain knowledge or not.

Finally, the proposed method (Experiment 9) synthesizes all the principles discussed in this study. It employs inspection notes embedded by a domain-knowledge fine-tuned LLM to weigh the other inputs, resulting in a significant reduction in error to 4.2.

Figure D.3 shows the performance of nine experiment models across 500 test samples. It comprises ten line plots representing the variance in performance among the ablation models. The alignment between the green line (representing the proposed model's predictions) and the black line (representing actual data) underscores the impressive accuracy of the proposed method. It serves as compelling evidence of the model's effectiveness, consistent with the most robust ground truth available, namely, human expert evaluation.



(a) Results of experiments 1 and 2

(b) Results of experiments 3, 4, and 5

(c) Results of experiments 6 and 7

(d) Results of experiments 8 and 9 (proposed).

Figure D.3: Comparison of test predictions on 500 samples.

Indeed, through the integration of inspection notes enriched by a domain-knowledge fine-tuned LLM, the proposed method not only emphasizes the significance of incorporating expert knowledge into predictive models but also demonstrates its capacity to minimize error margins, as depicted in the graph.

Bibliography

Abacha, A. B., Hasan, S. A., Datla, V. V., Liu, J., Demner-Fushman, D., and Müller, H. (2019). Vqa-med: Overview of the medical visual question answering task at imageclef 2019. *CLEF (Working Notes)*, 2. *Cited in page 24*

Ahmad, S., Lavin, A., Purdy, S., and Agha, Z. (2017). Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 262:134–147. *Cited in page 15*

Ahuja, C. and Morency, L.-P. (2019). In *Language2Pose: Natural Language Grounded Pose Forecasting*, pages 719–728. *Cited in page 27*

Akrim, A., Gogu, C., Vingerhoeds, R., and Salaün, M. (2023). Self-supervised learning for data scarcity in a fatigue damage prognostic problem. *Engineering Applications of Artificial Intelligence*, 120:105837. *Cited in page 43*

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., *et al.* (2022). Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*. *Cited in page 31*

Albawi, S., Mohammed, T. A., and Al-Zawi, S. (2017). Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*, pages 1–6. Ieee. *Cited in page 52*

Altendorf, E. E., Restificar, A. C., and Dietterich, T. G. (2012). Learning from sparse data by exploiting monotonicity constraints. *arXiv preprint arXiv:1207.1364*. *Cited in page 66*

Álvarez-Sánchez JR, V.-C. M. (2020). Ferrández-vicente jm fernández e affective robot story-telling human-robot interaction: exploratory real-time emotion estimation analysis using facial expressions and physiological signals. *IEEE Access*, 8:134051–134066. *Cited in page 24*

Ansari, F., Glawar, R., and Nemeth, T. (2019). Prima: a prescriptive maintenance model for cyber-physical production systems. *International Journal of Computer Integrated Manufacturing*, 32(4-5):482–503. *Cited in page 36*

Ansari, F., Glawar, R., and Sihn, W. (2020). In Beyerer, J., Maier, A., and Niggemann, O., editors, *Prescriptive Maintenance of CPPS by Integrating Multimodal Data with Dynamic Bayesian Networks*, Technologien für die intelligente Automation, pages 1–8, Berlin, Heidelberg. Springer. *Cited in page 36*

Atoui, M. A. and Cohen, A. (2021). Coupling data-driven and model-based methods to improve fault diagnosis. *Computers in Industry*, 128:103401. *Cited in page 66*

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443. *Cited in pages 18 and 25*

Barros, P., Weber, C., and Wermter, S. (2015). In *Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction*, pages 582–587. *Cited in page 24*

Behrad, F. and Abadeh, M. S. (2022). An overview of deep learning methods for multimodal medical data mining. *Expert Systems with Applications*, page 117006. *Cited in pages 22 and 23*

Blancke, O., Tahan, A., Komljenovic, D., Amyot, N., Lévesque, M., and Hudon, C. (2018). A holistic multi-failure mode prognosis approach for complex equipment. *Reliability Engineering & System Safety*, 180:136–151. *Cited in pages viii, x, 70, and 129*

Blank, M. (1974). Cognitive functions of language in the preschool years. *Developmental Psychology*, 10(2):229. *Cited in page 19*

Bustos, A., Pertusa, A., Salinas, J.-M., and de la Iglesia-Vayá, M. (2020). Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797. *Cited in page 23*

Butterworth, B. and Hadar, U. (1989). Gesture, speech, and computational stages: a reply to McNeill. *Psychological Review*, 96(1). *Cited in page 19*

Bzymek, A. (2017). Application of selected method of anomaly detection in signals acquired during welding process monitoring. *International Journal of Materials and Product Technology*, 54(4):249–258. *Cited in page 12*

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. (2020). nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631. *Cited in page 16*

Cai, D., Wang, Y., Liu, L., and Shi, S. (2022). Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3417–3419. *Cited in page 103*

Cai, Q., Wang, H., Li, Z., and Liu, X. (2019). A survey on multimodal data-driven smart healthcare systems: Approaches and applications. *IEEE Access*, 7:133583–133599. event: IEEE Access. *Cited in page 13*

Camizuli, E. and Carranza, E. J. (2018). Exploratory data analysis (eda). *The encyclopedia of archaeological sciences*, pages 1–7. *Cited in page 115*

Cao, B., Zhang, H., Wang, N., Gao, X., and Shen, D. (2020). Auto-gan: Self-supervised collaborative learning for medical image synthesis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):10486–10493. number: 07. *Cited in pages 22 and 29*

Case Western Reserve University (2021). Download a data file | case school of engineering | case western reserve university. [Online; accessed 2022-05-23]. *Cited in page 14*

Chamberlain, B., Rowbottom, J., Gorinova, M. I., Bronstein, M., Webb, S., and Rossi, E. (2021). Grand: Graph neural diffusion. In *International Conference on Machine Learning*, pages 1407–1418. PMLR. *Cited in page 128*

Chang, S.-F., Chen, W., Meng, H. J., Sundaram, H., and Zhong, D. (1998). A fully automated content-based video search engine supporting spatiotemporal queries. *IEEE transactions on circuits and systems for video technology*, 8(5):602–615. *Cited in page 19*

Chen, C., Jafari, R., and Kehtarnavaz, N. (2015). Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*, pages 168–172. IEEE. *Cited in page 16*

Chen, H., Gao, M., Zhang, Y., Liang, W., and Zou, X. (2019). Attention-based multi-nmf deep neural network with multimodality data for breast cancer prognosis model. *BioMed Research International*, 2019:e9523719. publisher: Hindawi. *Cited in page 23*

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020a). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR. *Cited in page 29*

Chen, X. (2019). Tennessee eastman simulation dataset. *Cited in page 14*

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. (2020b). Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer. *Cited in page 30*

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*. *Cited in page 28*

Cohen Kalafut, N., Huang, X., and Wang, D. (2023). Joint variational autoencoders for multimodal imputation and embedding. *Nature Machine Intelligence*, 5(6):631–642. *Cited in page 29*

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. *Cited in pages 19 and 31*

Ding, P., Jia, M., Ding, Y., Cao, Y., and Zhao, X. (2022). Intelligent machinery health prognostics under variable operation conditions with limited and variable-length data. *Advanced Engineering Informatics*, 53:101691. *Cited in page 111*

Ding, P., Xia, J., Zhao, X., and Jia, M. (2024). Graph structure few-shot prognostics for machinery remaining useful life prediction under variable operating conditions. *Advanced Engineering Informatics*, 60:102360. *Cited in page 111*

D'mello, S. K. and Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3):1–36. *Cited in page 26*

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., *et al.* (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. *Cited in page 31*

Ethayarajh, K. (2019). How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*. *Cited in page 102*

Ewerth, R., Otto, C., Müller-Budack, E., Pflaeging, J., Wildfeuer, J., Bateman, J. A., and Gruyter, D. (2021). Computational approaches for the interpretation of image-text relations. *Empirical Multimodality Research: Methods, Evaluations, Implications*, pages 109–138. *Cited in page 66*

Faghri, F., Fleet, D. J., Kiros, J. R., and Fidler, S. (2017). Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*. *Cited in page 30*

Falk, C., Sand, V. D. R., Corasaniti, S., and Reiff-Stephan, J. (2021). A comparison study of data-driven anomaly detection approaches for industrial chillers. *TH Wildau Engineering and Natural Sciences Proceedings*, 1. [Online; accessed 2022-05-23]. *Cited in page 12*

Fauconnier, J.-P. (2015). French word embeddings. *Cited in page 104*

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26. *Cited in page 30*

Gao, J., Li, P., Chen, Z., and Zhang, J. (2020). A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864. *Cited in page 25*

Gaw, N., Yousefi, S., and Gahrooei, M. R. (2021). Multimodal data fusion for systems improvement: A review. *IISE Transactions*, pages 1–19. *Cited in page 25*

Gong, J., Chen, Z., Ma, C., Xiao, Z., Wang, H., Tang, G., Liu, L., Xu, S., Long, B., and Jiang, Y. (2023). Attention weighted mixture of experts with contrastive learning for personalized ranking in e-commerce. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3222–3234. IEEE. *Cited in page 121*

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. volume 27. Curran Associates, Inc. [Online; accessed 2022-06-01]. *Cited in page 28*

Gouriveau, R., Medjaher, K., and Zerhouni, N. (2016a). *From prognostics and health systems management to predictive maintenance 1: Monitoring and prognostics*. John Wiley & Sons. *Cited in pages vii and 4*

Gouriveau, R., Medjaher, K., and Zerhouni, N. (2016b). From prognostics and health systems management to predictive maintenance 1: Monitoring and prognostics, volume 4. *Cited in page 11*

Guo, M., Haque, A., Huang, D.-A., Yeung, S., and Fei-Fei, L. (2018). Dynamic task prioritization for multitask learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 270–287. *Cited in page 140*

Hager, P., Menten, M. J., and Rueckert, D. (2023). Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23924–23935. *Cited in page 29*

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. *Cited in page 31*

Heiliger, L., Sekuboyina, A., Menze, B., Egger, J., and Kleesiek, J. (2022). Beyond medical imaging - a review of multimodal deep learning in radiology. publisher: TechRxiv. *Cited in page 22*

Hendrickx, K., Meert, W., Mollet, Y., Gyselinck, J., Cornelis, B., Gryllias, K., and Davis, J. (2020). A general anomaly detection framework for fleet-based condition monitoring of machines. *Mechanical Systems and Signal Processing*, 139:106585. arXiv:1912.12941 [cs, eess, stat]. *Cited in page 12*

Hervella, Á. S., Rouco, J., Novo, J., and Ortega, M. (2019). Self-supervised deep learning for retinal vessel segmentation using automatically generated labels from multimodal data. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE. *Cited in pages 23 and 28*

Hou, Z., Liu, X., Cen, Y., Dong, Y., Yang, H., Wang, C., and Tang, J. (2022). Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 594–604. *Cited in pages 128 and 140*

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*. *Cited in page 32*

Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., and Lungren, M. P. (2020). Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3(1):1–9. *Cited in pages xiii, 22, and 23*

Hudon, C. and Belec, M. (2005). Partial discharge signal interpretation for generator diagnostics. *IEEE Transactions on Dielectrics and Electrical Insulation*, 12(2):297–319. *Cited in pages viii, 70, and 71*

Inceoglu, A., Aksoy, E. E., Ak, A. C., and Sariel, S. (2021). Fino-net: A deep multimodal sensor fusion framework for manipulation failure detection. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6841–6847. IEEE. *Cited in page 24*

Iqball, T. and Wani, M. A. (2023). Weighted ensemble model for image classification. *International Journal of Information Technology*, 15(2):557–564. *Cited in page 121*

Islam, M. T., Zhou, Z., Ren, H., Khuzani, M. B., Kapp, D., Zou, J., Tian, L., Liao, J. C., and Xing, L. (2023). Revealing hidden patterns in deep neural network feature space continuum via manifold learning. *Nature Communications*, 14(1):8506. *Cited in page 110*

Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., and Carreira, J. (2021). Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR. *Cited in page 28*

Jia, X., Huang, B., Feng, J., Cai, H., and Lee, J. (2018). *A Review of PHM Data Competitions from 2008 to 2017*. *Cited in page 12*

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., *et al.* (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*. *Cited in pages 31 and 102*

Johnson, A., Pollard, T., Mark, R., Berkowitz, S., and Horng, S. (2019). Mimic-cxr database. *PhysioNet10*, 13026:C2JT1Q. *Cited in page 23*

Jose, S., Nguyen, K. T., Medjaher, K., Zemouri, R., Lévesque, M., and Tahan, A. (2023a). Fault detection and diagnostics in the context of sparse multimodal data and expert knowledge assistance: Application to hydrogenerators. *Computers in Industry*, 151:103983. *Cited in pages xii, 168, and 178*

Jose, S., Zemouri, R., Lévesque, M., Nguyen, K. T., Tahan, A., and Medjaher, K. (2023b). Informed machine learning for image-data-driven diagnostics of hydrogenerators. *IFAC-PapersOnLine*, 56(2):11912–11917. *Cited in page 166*

Kao, H.-Y., Wang, Y.-Y., Huang, C.-M., and Hsu, C.-P. (2019). Heterogeneous data ensemble learning in end-to-end diagnosis for iptv. pages 1–6. 2019 20th Asia-Pacific Network Operations and Management Symposium (APNOMS). ISSN: 2576-8565. *Cited in page 34*

Keller-Cohen, D. (1978). Context in child language. *Annual Review of Anthropology*, 7:453–482. *Cited in page 19*

Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U. D., and Jawahar, C. (2021). Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1033–1036. IEEE. *Cited in page 28*

Kim, N.-H., An, D., and Choi, J.-H. (2017). Prognostics and health management of engineering systems. *Switzerland: Springer International Publishing*. *Cited in page 11*

Kim, S., Kim, N. H., and Choi, J.-H. (2020). Prediction of remaining useful life by data augmentation technique based on dynamic time warping. *Mechanical Systems and Signal Processing*, 136:106486. *Cited in page 110*

Kokel, H., Odom, P., Yang, S., and Natarajan, S. (2020). A unified framework for knowledge intensive gradient boosting: Leveraging human experts for noisy sparse domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4460–4468. *Cited in page 66*

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Cited in page 26*

Lahat, D., Adali, T., and Jutten, C. (2015). Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477. *Cited in page 16*

Lassau, N., Ammari, S., Chouzenoux, E., Gortais, H., Herent, P., Devilder, M., Soliman, S., Meyrignac, O., Talabard, M.-P., Lamarque, J.-P., *et al.* (2021). Integrating deep learning ct-scan model, biological and clinical variables to predict severity of covid-19 patients. *Nature communications*, 12(1):1–11. *Cited in page 23*

Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., *et al.* (2023). Bloom: A 176b-parameter open-access multilingual language model. *Cited in page 31*

Leahy, W. and Sweller, J. (2011). Cognitive load theory, modality of presentation and the transient information effect. *Applied Cognitive Psychology*, 25(6):943–951. *Cited in page 16*

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444. *Cited in page 18*

LeCun, Y., Haffner, P., Bottou, L., and Bengio, Y. (1999). Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pages 319–345. Springer. *Cited in page 18*

Lee, D.-H., Yang, J.-K., Lee, C.-H., and Kim, K.-J. (2019). A data-driven approach to selection of critical process steps in the semiconductor manufacturing process considering missing and imbalanced data. *Journal of Manufacturing Systems*, 52:146–156. *Cited in page 43*

Lee, J., Qiu, H., Yu, G., and Lin, J. (2007). Rexnord technical services, bearing data set, ims, university of cincinnati. nasa ames prognostics data repository. *NASA Ames, Moffett Field, CA*. *Cited in page 14*

Li, T., Sun, C., Li, S., Wang, Z., Chen, X., and Yan, R. (2022a). Explainable graph wavelet denoising network for intelligent fault diagnosis. *IEEE Transactions on Neural Networks and Learning Systems*. *Cited in page 111*

Li, T., Zhao, Z., Sun, C., Yan, R., and Chen, X. (2021). Hierarchical attention graph convolutional network to fuse multi-sensor signals for remaining useful life prediction. *Reliability Engineering & System Safety*, 215:107878. *Cited in page 111*

Li, T., Zhou, Z., Li, S., Sun, C., Yan, R., and Chen, X. (2022b). The emerging graph neural networks for intelligent fault diagnostics and prognostics: A guideline and a benchmark study. *Mechanical Systems and Signal Processing*, 168:108653. *Cited in page 111*

Li, X., Jia, M., Islam, M. T., Yu, L., and Xing, L. (2020). Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. *IEEE Transactions on Medical Imaging*, 39(12):4023–4033. *Cited in page 23*

Liu, F., Zhou, L., Shen, C., and Yin, J. (2013). Multiple kernel learning in the primal for multimodal alzheimer's disease classification. *IEEE journal of biomedical and health informatics*, 18(3):984–990. *Cited in page 26*

Liu, W., Qiu, J.-L., Zheng, W.-L., and Lu, B.-L. (2021). Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729. *Cited in page 43*

Liu, Z., Liu, Y., Zhang, D., Cai, B., and Zheng, C. (2015). Fault diagnosis for a solar assisted heat pump system under incomplete data and expert knowledge. *Energy*, 87:41–48. *Cited in page 66*

Lu, G., Liu, J., and Yan, P. (2018). Graph-based structural change detection for rotating machinery monitoring. *Mechanical Systems and Signal Processing*, 99:73–82. *Cited in pages 12 and 26*

Lu, G., Zhou, Y., Lu, C., and Li, X. (2017). A novel framework of change-point detection for machine monitoring. *Mechanical Systems and Signal Processing*, C(83):533–548. *Cited in page 12*

Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32. *Cited in page 31*

Lu, N. and Yin, T. (2021). Transferable common feature space mining for fault diagnosis with imbalanced data. *Mechanical systems and signal processing*, 156:107645. *Cited in page 110*

Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., and Peng, X. (2021). Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310. *Cited in page 43*

Ma, Y., Guo, Z., Su, J., Chen, Y., Du, X., Yang, Y., Li, C., Lin, Y., and Geng, Y. (2014). Deep learning for fault diagnosis based on multi-sourced heterogeneous data. pages 740–745. 2014 International Conference on Power System Technology. *Cited in page 34*

Maghdid, H. S., Asaad, A. T., Ghafoor, K. Z., Sadiq, A. S., and Khan, M. K. (2020). Diagnosing covid-19 pneumonia from x-ray and ct images using deep learning and transfer learning algorithms. Technical report. *Cited in page 23*

Marei, M. and Li, W. (2021). Cutting tool prognostics enabled by hybrid cnn-lstm with transfer learning. *The International Journal of Advanced Manufacturing Technology*. *Cited in pages 34 and 35*

Matsumoto, M. and Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 8(1):3–30. *Cited in page 50*

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review*, 92(3):350–371. publisher-place: US publisher: American Psychological Association. *Cited in page 19*

Medjaher, K., Zerhouni, N., and Baklouti, J. (2013). Data-driven prognostics based on health indicator construction: Application to pronostia's data. In *2013 European Control Conference (ECC)*, pages 1451–1456. IEEE. *Cited in page 110*

Mian, T., Choudhary, A., and Fatima, S. (2022). A sensor fusion based approach for bearing fault diagnosis of rotating machine. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 236(5):661–675. *Cited in page 34*

Morency, L.-P., Liang, P. P., and Zadeh, A. (2022). Tutorial on multimodal machine learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 33–38, Seattle, United States. Association for Computational Linguistics. *Cited in page 18*

Moses, B., Gavish, L., and Vorwerck, M. (2022). *Data Quality Fundamentals*. " O'Reilly Media, Inc.". *Cited in page 42*

Nagulapati, V. M., Lee, H., Jung, D., Brigljevic, B., Choi, Y., and Lim, H. (2021). Capacity estimation of batteries: Influence of training dataset size and diversity on data driven prognostic models. *Reliability Engineering & System Safety*, 216:108048. *Cited in page 43*

Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Morello, B. C., Zerhouni, N., and Varnier, C. (2012). "PRONOSTIA: An experimental platform for bearings accelerated degradation tests". In *IEEE International Conference on Prognostics and Health Management, PHM'12., Denver, Colorado.* *Cited in page 14*

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *ICML*. *Cited in pages 19 and 26*

Norris, S. (2019). *Systematically working with multimodal data: Research methods in multimodal discourse analysis*. John Wiley & Sons. *Cited in page 16*

Omri, N. O., Al Masry, Z., Giampiccolo, S., Mairot, N., and Zerhouni, N. (2019). Data management requirements for phm implementation in smes. In *2019 Prognostics and System Health Management Conference (PHM-Paris)*, pages 232–238. IEEE. *Cited in page 66*

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6:405–431. *Cited in page 116*

Parcalabescu, L., Trost, N., and Frank, A. (2021). What is multimodality? *arXiv:2103.06304 [cs]*. arXiv: 2103.06304. *Cited in page 17*

Park, L. A. and Read, J. (2019). A blended metric for multi-label optimisation and evaluation. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pages 719–734. Springer. *Cited in page 171*

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *arXiv:1211.5063 [cs]*. arXiv: 1211.5063. *Cited in page 28*

Pei, H., Si, X.-S., Hu, C., Li, T., He, C., and Pang, Z. (2022). Bayesian deep-learning-based prognostic model for equipment without label data related to lifetime. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(1):504–517. *Cited in page 113*

Pham, H., Liang, P., Manzini, T., Morency, L.-P., and Poczos, B. (2019). Found in translation: Learning robust joint representations by cyclic translations between modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6892–6899. *Cited in page 27*

Picard, R. W. (2000). *Affective Computing*. MIT Press. Google-Books-ID: GaVncRTcb1gC. *Cited in page 19*

Pittino, F., Puggl, M., Moldaschl, T., and Hirschl, C. (2020). Automatic anomaly detection on in-production manufacturing machines using statistical learning methods. *Sensors*, 20(8):2344. number: 8 publisher: Multidisciplinary Digital Publishing Institute. *Cited in page 12*

Popescu, G. V., Burdea, G. C., and Trefftz, H. (2002). Multimodal interaction modeling. In *Handbook of Virtual Environments*, pages 475–494. CRC Press. *Cited in page 19*

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv:2103.00020 [cs]*. *Cited in page 31*

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., *et al.* (2018). Improving language understanding by generative pre-training. *Cited in page 31*

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., *et al.* (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9. *Cited in pages 31, 102, and 166*

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67. *Cited in page 31*

Rahimi, S. A., Jamshidi, A., Ruiz, A., and Aït-Kadi, D. (2016). A new dynamic integrated framework for surgical patients' prioritization considering risks and uncertainties. *Decision Support Systems*, 88:112–120. *Cited in page 13*

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28. *Cited in page 76*

Roeper, T. and McNeill, D. (1973). Review of child language. *Annual Review of Anthropology*, 2:127–137. *Cited in page 19*

Ronneberger, O., Fischer, P., and Brox, T. (2015a). U-net: Convolutional networks for biomedical image segmentation. *arXiv:1505.04597 [cs]*. arXiv: 1505.04597. *Cited in page 27*

Ronneberger, O., Fischer, P., and Brox, T. (2015b). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer. *Cited in page 80*

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536. *Cited in page 4*

Sas, A. (2020). Airbus helicopter accelerometer dataset. Accepted: 2020-05-19T12:16:26Z publisher: ETH Zurich type: dataset. *Cited in page 15*

Saxena, A. and Goebel, K. (2008). Turbofan engine degradation simulation data set. *NASA Ames Prognostics Data Repository*, pages 1551–3203. *Cited in page 15*

Schlechtingen, M. and Santos, I. (2011). Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection. *Mechanical Systems and Signal Processing*, 25:1849–1875. *Cited in page 12*

Sechidis, K., Tsoumakas, G., and Vlahavas, I. (2011). On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*, pages 145–158. Springer. *Cited in page 116*

Shao, S. (2022). *Mechanical-datasets*. original-date: 2018-01-16T19:12:43Z. *Cited in page 14*

Silver, N. (2012). *The signal and the noise: Why so many predictions fail-but some don't.* Penguin. *Cited in page 109*

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(1):97–99. *Cited in page 16*

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. *Cited in pages 19 and 31*

Sinha, K., Jia, R., Hupkes, D., Pineau, J., Williams, A., and Kiela, D. (2021). Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *Cited in page 102*

Soleimani, M., Campean, F., and Neagu, D. (2021). Diagnostics and prognostics for complex systems: A review of methods and challenges. *Quality and Reliability Engineering International*, 37(8):3746–3778. *Cited in page 110*

Sorower, M. S. (2010). A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 18(1):25. *Cited in page 171*

Soualhi, M., Soualhi, A., Nguyen, T.-P. K., Medjaher, K., Clerc, G., and Razik, H. (2023). Ampere: Detection and diagnostics of rotor and stator faults in rotating machines. *Cited in page 15*

Spasov, S. E., Passamonti, L., Duggento, A., Lio, P., and Toschi, N. (2018). A multimodal convolutional neural network framework for the prediction of alzheimer's disease. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2018:1271–1274. PMID: 30440622. *Cited in page 22*

Spezialetti, M., Placidi, G., and Rossi, S. (2020). Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI*, 7. [Online; accessed 2022-06-01]. *Cited in page 24*

Srivastava, N. and Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. *Advances in Neural Information Processing Systems*, 25. [Online; accessed 2021-05-31]. *Cited in pages 18, 19, and 26*

Staudemeyer, R. C. and Morris, E. R. (2019). Understanding lstm – a tutorial into long short-term memory recurrent neural networks. *arXiv:1909.09586 [cs]*. *Cited in page 28*

Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J. M. R., Bradley, A., Papa, J. P., Belagiannis, V., *et al.* (2018). *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings*, volume 11045. Springer. *Cited in page 22*

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*. *Cited in page 28*

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9. *Cited in page 31*

Taleb, A., Kirchler, M., Monti, R., and Lippert, C. (2022). Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20908–20921. *Cited in page 29*

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR. *Cited in page 31*

Tarekegn, A. N., Giacobini, M., and Michalak, K. (2021). A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965. *Cited in page 115*

Tekin, C., Atan, O., and Van Der Schaar, M. (2015). Discover the expert: Context-adaptive expert selection for medical diagnosis. *IEEE Transactions on Emerging Topics in Computing*, 3(2):220–234. *Cited in page 13*

Tian, J., Azarian, M. H., and Pecht, M. (2014). Anomaly detection using self-organizing maps-based k-nearest neighbor algorithm. *PHM Society European Conference*, 2(1). number: 1. *Cited in page 12*

Toosi, A., Bottino, A. G., Saboury, B., Siegel, E., and Rahmim, A. (2021). A brief history of AI: how to prevent another winter (a critical review). *PET clinics*, 16(4):449–469. *Cited in pages 18 and 19*

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., *et al.* (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*. *Cited in pages 31 and 102*

Trigeorgis, G., Nicolaou, M., Zafeiriou, S., and Schuller, B. (2016). Deep canonical time warping. pages 5110–5118. *Cited in page 26*

Tsiourti, C., Weiss, A., Wac, K., and Vincze, M. (2017). Designing emotionally expressive robots: a comparative study on the perception of communication modalities. In *Proceedings of the 5th international conference on human agent interaction*, pages 213–222. *Cited in page 16*

TURING, A. (1950). 1-computing machinery and intelligence. *MIND*, 59(236):433. *Cited in pages 3 and 10*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. *Cited in pages viii, 19, 26, 28, 45, and 56*

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*. *Cited in page 140*

Vesterinen, E. *et al.* (2001). Affective computing. In *Digital Media Research Seminar, Helsinki*. Citeseer. *Cited in page 19*

Wang, M., Zheng, D., Ye, Z., Gan, Q., Li, M., Song, X., Zhou, J., Ma, C., Yu, L., Gai, Y., Xiao, T., He, T., Karypis, G., Li, J., and Zhang, Z. (2019a). Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*. *Cited in page 127*

Wang, P., Liu, Z., Gao, R. X., and Guo, Y. (2019b). Heterogeneous data-driven hybrid machine learning for tool condition prognosis. *CIRP Annals*, 68(1):455–458. *Cited in page 35*

Wang, T. (2010). *Trajectory similarity based prediction for remaining useful life estimation*. University of Cincinnati. *Cited in page 110*

Wang, T., Yu, J., Siegel, D., and Lee, J. (2008). A similarity-based prognostics approach for remaining useful life estimation of engineered systems. In *2008 international conference on prognostics and health management*, pages 1–6. IEEE. *Cited in page 110*

Wang, X., Peng, Y., Lu, L., Lu, Z., and Summers, R. M. (2018). Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. Technical report. DOI: 10.48550/arXiv.1801.04334. *Cited in page 23*

Wei, Y., Wu, D., and Terpenny, J. (2023). Bearing remaining useful life prediction using self-adaptive graph convolutional networks with self-attention mechanism. *Mechanical Systems and Signal Processing*, 188:110010. *Cited in page 111*

Wen, Y., Rahman, M. F., Xu, H., and Tseng, T.-L. B. (2022). Recent advances and trends of predictive maintenance from data-driven machine prognostics perspective. *Measurement*, 187:110276. *Cited in page 110*

Xu, G., Liu, M., Wang, J., Ma, Y., Wang, J., Li, F., and Shen, W. (2019). Data-driven fault diagnostics and prognostics for predictive maintenance: A brief overview. pages 103–108. 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE). ISSN: 2161-8089. *Cited in page 12*

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR. *Cited in page 19*

Xu, Q., Gao, P., Wang, J., Zhang, J., Ip, A., and Zhang, C. (2024). Akgnn-pc: An assembly knowledge graph neural network model with predictive value calibration module for refrigeration compressor performance prediction with assembly error propagation and data imbalance scenarios. *Advanced Engineering Informatics*, 60:102403. *Cited in page 111*

Yala, A., Lehman, C., Schuster, T., Portnoi, T., and Barzilay, R. (2019). A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, 292(1):60–66. PMID: 31063083. *Cited in page 22*

Yan, K., Ji, Z., and Shen, W. (2017). Online fault detection methods for chillers combining extended kalman filter and recursive one-class SVM. *Neurocomputing*, 228:205–212. *Cited in page 12*

Yang, Z., Baraldi, P., and Zio, E. (2021). A multi-branch deep neural network model for failure prognostics based on multimodal data. *Journal of Manufacturing Systems*, 59:42–50. *Cited in pages viii, 13, 35, 48, 49, 53, 54, and 55*

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32. *Cited in page 31*

Yoo, Y., Tang, L. Y., Li, D. K., Metz, L., Kolind, S., Traboulsee, A. L., and Tam, R. C. (2019). Deep learning of brain lesion patterns and user-defined clinical and mri features for predicting conversion to multiple sclerosis from clinically isolated syndrome. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 7(3):250–259. *Cited in page 22*

Yoon, J., Davtyan, C., and van der Schaar, M. (2016). Discovery and clinical decision support for personalized healthcare. *IEEE journal of biomedical and health informatics*, 21(4):1133–1145. *Cited in page 13*

You, J., Ying, R., Ren, X., Hamilton, W., and Leskovec, J. (2018). Graphrnn: Generating realistic graphs with deep auto-regressive models. In *International conference on machine learning*, pages 5708–5717. PMLR. *Cited in page 128*

Yuan, X., Lin, Z., Kuen, J., Zhang, J., Wang, Y., Maire, M., Kale, A., and Faieta, B. (2021). Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004. *Cited in page 29*

Yuan, Z., Zhang, L., and Duan, L. (2018). A novel fusion diagnosis method for rotor system fault based on deep learning and multi-sourced heterogeneous monitoring data. *Measurement Science and Technology*, 29(11):115005. publisher: IOP Publishing. *Cited in page 34*

Yucesan, Y. A. and Viana, F. A. (2021). Hybrid physics-informed neural networks for main bearing fatigue prognosis with visual grease inspection. *Computers in Industry*, 125:103386. *Cited in page 66*

Zacharaki, A., Vafeiadis, T., Kolokas, N., Vaxevani, A., Xu, Y., Peschl, M., Ioannidis, D., and Tzovaras, D. (2021). Reclaim: Toward a new era of refurbishment and remanufacturing of industrial equipment. *Frontiers in Artificial Intelligence*, 3:570562. *Cited in page 36*

Zara, A., Maffiolo, V., Martin, J. C., and Devillers, L. (2007). Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics. In *International Conference on Affective Computing and Intelligent Interaction*, pages 464–475. Springer. *Cited in page 19*

Zemouri, R., Levesque, M., Amyot, N., Hudon, C., Kokoko, O., and Tahan, S. A. (2019). Deep convolutional variational autoencoder as a 2d-visualization tool for partial discharge source classification in hydrogenerators. *IEEE Access*, 8:5438–5454. *Cited in pages ix, 82, and 122*

Zhang, T., Chen, J., Li, F., Zhang, K., Lv, H., He, S., and Xu, E. (2022). Intelligent fault diagnosis of machines with small & imbalanced data: A state-of-the-art review and possible extensions. *ISA transactions*, 119:152–171. *Cited in page 115*

Zhang, T. and Shi, M. (2020). Multi-modal neuroimaging feature fusion for diagnosis of alzheimer's disease. *Journal of Neuroscience Methods*, 341:108795. *Cited in page 28*

Zhang, X., Fujiwara, T., Chandrasegaran, S., Brundage, M., Sexton, T., Dima, A., and Ma, K.-L. (2021). A visual analytics approach for the diagnosis of heterogeneous and multidimensional machine maintenance data. *Cited in page 35*

Zhang, X., Zhang, X., Liu, J., Wu, B., and Hu, Y. (2023). Graph features dynamic fusion learning driven by multi-head attention for large rotating machinery fault diagnosis with multi-sensor data. *Engineering Applications of Artificial Intelligence*, 125:106601. *Cited in page 111*

Zhang, Y. and Wallace, B. (2016). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. Technical report. DOI: 10.48550/arXiv.1510.03820. *Cited in page 27*

Zhao, P., Kurihara, M., Tanaka, J., Noda, T., Chikuma, S., and Suzuki, T. (2017). Advanced correlation-based anomaly detection method for predictive maintenance. *2017 IEEE International Conference on Prognostics and Health Management (ICPHM)*. *Cited in page 12*

Zheng, Y., Huang, D., Liu, S., and Wang, Y. (2020). Cross-domain object detection through coarse-to-fine feature adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13766–13775. *Cited in page 32*

Zhou, A., Wang, J., Wang, Y.-X., and Wang, H. (2024). Distilling out-of-distribution robustness from vision-language foundation models. *Advances in Neural Information Processing Systems*, 36. *Cited in page 33*

Zhou, F., Yang, S., He, Y., Chen, D., and Wen, C. (2021). Fault diagnosis based on deep learning by extracting inherent common feature of multi-source heterogeneous data. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 235(10):1858–1872. publisher: IMECHE. *Cited in page 34*

Zio, E. (2022). Prognostics and health management (phm): Where are we and where do we (need to) go in theory and practice. *Reliability Engineering & System Safety*, 218:108119. *Cited in pages 11 and 110*

Zolfaghari, M., Zhu, Y., Gehler, P., and Brox, T. (2021). Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1450–1459. *Cited in page 29*

Multimodal Learning Strategies for Industrial Machine Health Diagnostics and Prognostics under Data Scarcity

Abstract

Prognostics and Health Management (PHM) with data-driven techniques is heavily dependent upon the availability of extensive and high-quality datasets, a requirement often challenging to fulfill in industrial condition monitoring environments. This discrepancy creates a significant gap between state-of-the-art PHM methodologies and their practical application in real-world scenarios. The prevailing focus in data-driven PHM research on unimodal datasets highlights the potential of multimodal data to bridge this gap. This thesis explores the integration of multimodal data to advance PHM models for industrial machines. It systematically addresses pivotal challenges such as data missingness and noise, sparse and irregular datasets, class imbalance, and the scarcity of run-to-failure data. The research develops innovative methodologies that incorporate multiple data modalities and harness domain-specific expertise to create robust predictive models. The primary contributions of this research include:

1. **Cross-modal attention-based learning:** A new multimodal learning method is designed to mitigate the limitations posed by missing and noisy data. It allows integrating information across multiple modalities, thereby enhancing the accuracy and robustness of predictive models.
2. **Expert-knowledge-assisted multimodal diagnostics methodology:** This methodology combines domain expertise with multimodal learning to enable comprehensive diagnostics, thereby improving fault detection and classification in industrial machinery.
3. **Graph-based prognostics approach:** This innovative approach constructs run-to-failure trajectories from incomplete data using graph-based techniques, offering a significant advancement in failure prognostics.

These methodologies were rigorously validated using both simulation and industrial dataset of hydrogenerators, demonstrating significant improvements in PHM and predictive maintenance capabilities. The results underscore the potential of multimodal data to significantly enhance the reliability and efficiency of PHM methods and algorithms. This thesis proposes a comprehensive framework for leveraging diverse data sources and domain expertise, promising to transform maintenance strategies and reducing operational costs across various industries. The findings pave the way for future research and practical

implementations, positioning multimodal data integration as a pivotal advancement in the field of PHM.

Keywords: Prognostics and health management (PHM); Diagnostics; Run-to-failure data; Prognostics; Predictive maintenance; Data-driven techniques; Multimodal learning; Machine learning; Deep learning; Cross-modal attention; Graph neural networks; Industrial applications.

Stratégies d'apprentissage multimodal pour le diagnostic et le pronostic de la santé des machines industrielles dans un contexte de manque de données

Résumé

Les approches de Pronostic et gestion de la santé des systèmes (Prognostics and Health Management : PHM) guidées par les données sont fortement dépendantes de la disponibilité et de la qualité d'historiques de défaillances, une exigence souvent difficile à satisfaire dans les systèmes de surveillance en conditions industrielles. Cette divergence crée un écart significatif entre les méthodologies de PHM et leur application pratique sur des systèmes réels. L'accent prédominant mis sur les ensembles de données unimodales dans les travaux de recherche en PHM basée sur les données met en lumière le potentiel des données multimodales pour combler cet écart. Cette thèse explore l'intégration des données multimodales afin d'améliorer les méthodes et les algorithmes de PHM appliqués aux machines industrielles. Elle aborde de manière systématique des défis cruciaux tels que l'absence de données, les données bruitées, les données clairsemées et irrégulières, le déséquilibre des classes et la rareté des données de fonctionnement jusqu'à la défaillance. Elle propose des méthodologies innovantes qui intègrent plusieurs modalités de données et exploitent l'expertise spécifique au domaine pour créer des modèles prédictifs robustes.

Les contributions principales de la thèse se déclinent comme suit :

1. **Apprentissage basé sur l'attention intermodale:** une nouvelle méthode d'apprentissage multimodal conçue pour atténuer les limites posées par les données manquantes et bruitées. Elle permet d'intégrer des informations provenant de multiples modalités, améliorant ainsi la précision et la robustesse des modèles prédictifs.
2. **Méthodologie de diagnostic multimodal assisté par les connaissances d'experts:** cette méthodologie combine l'expertise du domaine avec l'apprentissage multimodal pour permettre des diagnostics complets, améliorant ainsi la détection et la classification des défauts dans les machines industrielles.
3. **Approche de pronostic basée sur des graphes:** cette approche innovante construit des trajectoires de fonctionnement jusqu'à la défaillance à partir de données incomplètes en utilisant des techniques basées sur les graphes, offrant une avancée significative dans le domaine du pronostic de défaillances.

Ces méthodologies ont été rigoureusement validées sur des données de simulation ainsi que sur des données industrielles provenant d'hydro-générateurs, démontrant des amélio-

rations significatives des algorithmes de PHM et de maintenance prédictive. Les résultats soulignent le potentiel des données multimodales pour améliorer considérablement la fiabilité et l'efficacité des modèles de PHM.

Cette thèse apporte un cadre complet pour exploiter diverses sources de données et l'expertise du domaine, promettant de transformer les stratégies de maintenance et de réduire les coûts opérationnels dans diverses industries. Les résultats ouvrent la voie à des recherches futures et à des applications pratiques, positionnant l'intégration des données multimodales comme une avancée essentielle dans le domaine du PHM.

Mots clés: Pronostic et gestion de la santé des systèmes (PHM) ; Diagnostic ; Données de défaillances ; Pronostic ; Maintenance prédictive; Méthodes guidées par des données ; Apprentissage multimodal ; Apprentissage automatique ; Apprentissage profond ; Attention intermodale ; Réseaux de neurones graphiques ; Applications industrielles.

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Université de Toulouse

Titre : Stratégies d'apprentissage multimodal pour le diagnostic et le pronostic de la santé des machines industrielles dans un contexte de manque de données

Mots clés : Pronostic et gestion de la santé des systèmes (PHM), Méthodes guidées par des données, Maintenance prédictive, Apprentissage profond, Apprentissage multimodal, Diagnostic et pronostic

Résumé : Les approches de Pronostic et gestion de la santé des systèmes (Prognostics and Health Management : PHM) guidées par les données sont fortement dépendantes de la disponibilité et de la qualité d'historiques de défaillances, une exigence souvent difficile à satisfaire dans les systèmes de surveillance en conditions industrielles. Cette divergence crée un écart significatif entre les méthodologies de PHM et leur application pratique sur des systèmes réels. L'accent prédominant mis sur les ensembles de données unimodales dans les travaux de recherche en PHM basée sur les données met en lumière le potentiel des données multimodales pour combler cet écart.
Cette thèse explore l'intégration des données multimodales afin d'améliorer les méthodes et les algorithmes de PHM appliqués aux machines industrielles. Elle aborde de manière systématique des défis cruciaux tels que l'absence de données, les données bruitées, les données clairsemées et irrégulières, le déséquilibre des classes et la rareté des données de fonctionnement jusqu'à la défaillance. Elle propose des méthodologies innovantes qui intègrent plusieurs modalités de données et exploitent l'expertise spécifique au domaine pour créer des modèles prédictifs robustes.
Les contributions principales de la thèse se déclinent comme suit :
1. Apprentissage basé sur l'attention intermodale : une nouvelle méthode d'apprentissage multimodal conçue pour atténuer les limites posées par les données manquantes et bruitées. Elle permet d'intégrer des informations provenant de multiples modalités, améliorant ainsi la précision et la robustesse des modèles prédictifs.
2. Méthodologie de diagnostic multimodal assisté par les connaissances d'experts : cette méthodologie combine l'expertise du domaine avec l'apprentissage multimodal pour permettre des diagnostics complets, améliorant ainsi la détection et la classification des défauts dans les machines industrielles.
3. Approche de pronostic basée sur des graphes : cette approche innovante construit des trajectoires de fonctionnement jusqu'à la défaillance à partir de données incomplètes en utilisant des techniques basées sur les graphes, offrant une avancée significative dans le domaine du pronostic de défaillances.
Ces méthodologies ont été rigoureusement validées sur des données de simulation ainsi que sur des données industrielles provenant d'hydro-générateurs, démontrant des améliorations significatives des algorithmes de PHM et de maintenance prédictive. Les résultats soulignent le potentiel des données multimodales pour améliorer considérablement la fiabilité et l'efficacité des modèles de PHM.
Cette thèse apporte un cadre complet pour exploiter diverses sources de données et l'expertise du domaine, promettant de transformer les stratégies de maintenance et de réduire les coûts opérationnels dans diverses industries. Les résultats ouvrent la voie à des recherches futures et à des applications pratiques, positionnant l'intégration des données multimodales comme une avancée essentielle dans le domaine du PHM.

Title: Multimodal learning strategies for industrial machine health diagnostics and prognostics under data scarcity

Key words: Prognostics and health management (PHM), Data-driven techniques, Predictive maintenance, Deep learning, Multimodal learning, Diagnostics and prognostics

Abstract: Prognostics and Health Management (PHM) with data-driven techniques is heavily dependent upon the availability of extensive and high-quality datasets, a requirement often challenging to fulfill in industrial condition monitoring environments. This discrepancy creates a significant gap between state-of-the-art PHM methodologies and their practical application in real-world scenarios. The prevailing focus in data-driven PHM research on unimodal datasets highlights the potential of multimodal data to bridge this gap.
This thesis explores the integration of multimodal data to advance PHM models for industrial machines. It systematically addresses pivotal challenges such as data missingness and noise, sparse and irregular datasets, class imbalance, and the scarcity of run-to-failure data. The research develops innovative methodologies that incorporate multiple data modalities and harness domain-specific expertise to create robust predictive models.
The primary contributions of this research include:
1. Cross-modal attention-based learning: A new multimodal learning method is designed to mitigate the limitations posed by missing and noisy data. It allows integrating information across multiple modalities, thereby enhancing the accuracy and robustness of predictive models.
2. Expert-knowledge-assisted multimodal diagnostics methodology: This methodology combines domain expertise with multimodal learning to enable comprehensive diagnostics, thereby improving fault detection and classification in industrial machinery.
3. Graph-based prognostics approach: This innovative approach constructs run-to-failure trajectories from incomplete data using graph-based techniques, offering a significant advancement in failure prognostics.
These methodologies were rigorously validated using both simulation and industrial dataset of hydrogenerators, demonstrating significant improvements in PHM and predictive maintenance capabilities. The results underscore the potential of multimodal data to significantly enhance the reliability and efficiency of PHM methods and algorithms.
This thesis proposes a comprehensive framework for leveraging diverse data sources and domain expertise, promising to transform maintenance strategies and reducing operational costs across various industries. The findings pave the way for future research and practical implementations, positioning multimodal data integration as a pivotal advancement in the field of PHM.