MAINTAG - Multi-Agent-based Predictive Maintenance Dataset Tagging System

Oghuz BEKTASH[1], Dorian JOUBAUD[2], Chetan. S. Kulkarni[3], and Sylvain KUBLER[4]

[1,2,4] *SnT, University of Luxembourg, 6 Rue Richard Coudenhove-Kalergi, L-1359 Luxembourg, Luxembourg*
oguz.bektas@uni.lu
dorian.joubaud@uni.lu
sylvain.kubler@uni.lu

[3] *NASA Ames Research Center (KBR, Inc), Moffett Field, CA 94043, USA*
chetan.s.kulkarni@nasa.gov

ABSTRACT

With the ongoing digitization of global activities, the number of predictive maintenance datasets has been steadily growing. These datasets are often manually classified *in literature review papers* to assess their relevance for predictive maintenance applications. However, this manual approach is increasingly unsustainable, as it is time-intensive and prone to errors. The accelerating pace at which new datasets emerge in both scientific and industrial contexts makes this problem even worse. To overcome these challenges, there is a growing need for automated solutions to curate, analyze, and categorize (tag) datasets in the literature. To this end, we propose and evaluate `MAINTAG` (Multi-Agent-based Predictive Maintenance Dataset Tagging System), a novel multi-agent system designed to automate the classification of predictive maintenance datasets. MAINTAG is compatible with any criteria-based taxonomy and is assessed by benchmarking its tagging accuracy against recent state-of-the-art literature.

MAINTAG uses multiple AI agents built on large language models. Different agents handle different parts of the classification process. One agent identifies the application domain. Another determines the task type. A third figures out the supervision approach. The last one classifies the learning algorithm. Each agent uses GPT-based models to read through dataset documentation. They provide their classifications along with confidence scores. We evaluate MAINTAG on historical PHM challenge datasets (2008-2017). Results matched expert classifications with high correspondence across most categories. This shows that automated tagging can be as reliable as human experts. Our approach offers a practical way to manage the flood of new datasets in predictive maintenance. It maintains quality while keeping up with the rapid pace of predictive maintenance data creation.

1. INTRODUCTION

Rapid growth of Industry 4.0 has expanded the volume and diversity of datasets available for predictive maintenance (PdM). Operational data now comes from various domains such as vibration and temperature measurements in manufacturing to flight-recorder streams in aerospace. All of these information provides comprehensive inputs for Prognostics and Health Management (PHM) systems. This expansion is also visible in a broader trend in CM/PdM research due to the increasing system complexity, adoption of AI/ML, IoT, and high-speed communication technologies like 5G/6G. Accordingly, there is an exponential growth in both academic and industrial efforts (Nguyen et al., 2021; Pimenov et al., 2023).

Historically, maintenance strategies began with reactive repairs (A. Heng et al., 2009) and scheduled preventive tasks (A. S. Y. Heng, 2009; Bohlin et al., 2010). The field then shifted toward to data-driven strategies like Condition-Based Maintenance (CBM) and PdM. These frameworks make use of real-time sensor data to reduce unnecessary interventions (Jardine et al., 2006). PdM extends them with predictive analytics for early failure detection (Brotherton et al., 2000; Byington et al., 2008). Still, a core challenge is using complex condition monitoring (CM) data for fault detection, diagnostics, and prognostics (Tsui et al., 2019). In the past, synthetic data offered some value for maintenance applications (N. H. Eklund, 2006). However, for real world cases, standard datasets remain essential for reliable model development (Zhao et al., 2021). Until recently, the lack of publicly accessible datasets has been blocking progress (Sarker et al., 2022; Ramasso & Saxena, 2014). As a result, early studies highlighted the need for shared databases to enable benchmarking and facilitate innovation (Kans & Ingwald, 2008; Simões et al., 2011; Uusipaavalniemi & Juga, 2008).

However, despite growing dataset availability, dataset selection remains largely manual. It still relies on time-consuming literature reviews and surveys that struggle to keep pace with

the increasing rate of new releases across public and proprietary sources. In response to these, we introduce MAINTAG, a multi-agent system for automated dataset tagging in PHM applications. MAINTAG features modular agents for tasks such as term extraction, semantic indexing, web-based enrichment, and domain-specific evaluation based on PHM frameworks.

2. Related Work

PHM field keeps expanding through the integration of data-driven methodologies. The field also reshapes how complex systems monitor health, predict failures, and optimize maintenance strategies. Annual challenge datasets released through the annual PHM Competitions have have a notable impact in this progress. However, the manual curation and classification of these datasets —along with many others on platforms such as Kaggle, Zenodo, and similar repositories— remain labor-intensive and error-prone, despite repeated calls for scalable and intelligent alternatives.

2.1. PHM Data Competitions

PHM data challenges from 2008 to 2017 demonstrated a various applications in PHM tasks, moving from RUL prediction to diagnostic and regression problems. PHM 2008 challenge established a baseline with multivariate time-series data for RUL prediction in turbofan engines (Saxena & Goebel, 2008). In 2009, the new challenge shifted to fault detection and magnitude estimation for information about bearing geometry (N. Eklund & Bechhoefer, 2009). 2010 introduced a RUL estimation task multi-sensor inputs for high-speed CNC milling machine cutters(N. Eklund et al., 2010).

Following challenges targeted fault detection (2011, wind anemometer data) (N. Eklund & Kessler, 2011), and RUL prediction under accelerated degradation (2012, rolling bearings) (Nectoux et al., 2012). In 2013 and 2014, the challenges explored diagnostics and health assessment & fault detection tasks, respectively (N. Eklund & Kessler, 2013; Garvey & Wigny, 2014).

In 2015, the focus returned to fault classification and prognosis across multiple industrial plants using labeled operational data (Rosca et al., 2015), followed by 2016's health state tracking task of components within a wafer chemical-mechanical planarisation system (Propes & Rosca, 2016). Finally, 2017 emphasized on the combination of physics-based modeling and statistical approaches for prediction (Propes et al., 2017).

These benchmark datasets show methodological variations. Tasks include time series based RUL estimation to complex, multi-dimensional tasks involving event logs, hybrid models, and semi-supervised settings. Together, they provide a reliable data source for automated systems like MAINTAG to standardize dataset classification in the PHM domain.

These datasets have been reviewed by several influential studies. Jia et al. (2018) provided one of the most comprehensive state of art across the 2008–2017 data challenges. Similarly, Huang et al. (2017) provided an in-depth examination of data characteristics, challenge objectives, and algorithmic frameworks. Su & Lee (2023) further extended these by proposing an extended work based on open-source PHM challenges from 2018–2023. They also identify common limitations in PHM data challenge competitions by emphasizing data-related and model-related issues. PHM challenges. Now, there is a growing number of benchmark datasets emerging from community-driven platforms such as *Kaggle* and *Zenodo*. This further diversifies data modalities and experimental setups. These sources introduce new challenges in standardization, benchmarking consistency, and cross-domain generalizability.

2.2. MAINTAG: Toward Automated Dataset Tagging

Our work addresses these challenges through the development of **MAINTAG**, a multi-agent system designed to automate the tagging and classification of predictive maintenance datasets. MAINTAG supports any criteria-based taxonomy and is evaluated against the historical PHM challenges summarized above. Unlike earlier efforts, MAINTAG:

- Divides tagging into sub-tasks with parallel agents;
- Computes interpretable confidence scores for each tagging decision;
- Benchmarks classification performance against human experts and literature reviews.

Through this system, we show that automated tagging is both feasible and accurate. This offers an automated alternative to traditional literature-based curation.

Beyond PHM surveys, researchers in related domains have also explored automated metadata tagging and ontology-driven labeling. For instance, Mishra et al. Mishra et al. (2020) proposed a unified architecture for tagging Building Automation System metadata. Similarly, Lutz et al. Lutz et al. (2023) applied text classification to extract KPIs from unstructured wind turbine work orders. In the scientific data community, Gonçalves et al. Gonçalves et al. (2019) aligned biomedical metadata fields with ontologies using clustering and embeddings, and Dumschott et al. Dumschott et al. (2023) demonstrated how ontologies enhance FAIRness in plant research datasets. These works demonstrate the widespread impact of ontology-aware curation. However, MAINTAG is the first to apply such principles for PHM dataset tagging.

Our work is the first to bridge the review-based synthesis of Jia et al. (2018) with an automated solution that directly tackles the challenges of dataset curation and classification in PHM. While they provided a foundational framework for evaluating PHM datasets by categorizing datasets by system type, task, supervision, and learning method (see Table 1), their approach relied entirely on manual effort. Our work advances this by operationalization of their evaluation crite-

Dataset	System Type	PHM Task	Supervision Type	Learning Algorithm
PHM2008	Aircraft Engine	Prognosis	Supervised	Classification, Regression, Time Series
PHM2009	Gearbox	Fault Detection & Diagnosis	Supervised	Clustering
PHM2010	Milling Cutter	Assessment	Supervised	Regression
PHM2011	Anemometer	Fault Detection	Unsupervised	Anomaly Detection (Statistics and Residual- and Distance-Based)
PHM2012	Bearing	Prognosis	Supervised	Regression & Time Series
PHM2013	Unknown	Diagnosis	Supervised	Classification
PHM2014	Unknown	Assessment & Fault Detection	Unsupervised	Anomaly Detection / Statistics
PHM2015	Power Plant	Fault Detection & Diagnosis	Supervised	Classification
PHM2016	CMP	Other	—	—
PHM2017	Bogie	Fault Detection & Diagnosis	Supervised	Anomaly Detection (Residual- and Distance-Based)

Table 1. PHM Datasets Classified by System, Task, Supervision Type and Learning Algorithm Jia et al. (2018)

ria within a modular multi-agent system. These agents handle tasks such as terminology extraction, semantic indexing, and relevance scoring across diverse data sources. In doing so, we eliminate much of the manual burden. While large language models have been applied in generic classification tasks, MAINTAG is novel in operationalizing dataset taxonomy through a multi-agent pipeline tailored to PHM. In contrast to adjacent metadata-tagging efforts in domains, our approach is the first to target PHM datasets with explainable, confidence-aware aggregation. This integration points an important shift from static, manually maintained reviews to a dynamic and automated tagging framework capable of evolving with the predictive maintenance data landscape.

3. Methodology and MAINTAG System Design

This section provides the design of the MAINTAG system and details the research architecture, data sources, and computational procedures for automated dataset tagging. MAINTAG leverages a multi-agent architecture capable of classifying PHM datasets based on predefined taxonomic attributes using AI agents. The system consists of several specialized agent. Each powered by different GPT model variants depending on their computational requirements. MAINTAG orchestrator agent uses GPT-4 for coordination. On the other hand, sub-agents employ GPT-4-mini for efficiency in focused classification tasks.

3.1. Research Design

We formalize the task of PHM dataset tagging within a multi-agent framework and also provide interpretable outputs and confidence-aware aggregation that can be directly applied in practice. The research follows a methodology comprising three phases: system conceptualization, multi-agent framework development, and evaluation against expert baselines and historical PHM datasets.

MAINTAG is built upon a hierarchical agent-based system. Each agent is specialized in extracting and reasoning over one of the four core tagging indicators (see Table: 2):

- **Domain Type (D)**: Identifies the application field: Aerospace (A), Energy (E), Transportation (T), Manufacturing (M), Semiconductors (S).
- **Usage Type (U)**: Categorizes the dataset by its primary analytical function — Fault Detection (FD), Diagnosis, Assessment, Prognosis.
- **Supervision Type (S)**: Determines if learning is supervised (S) or unsupervised (U).
- **Algorithm Type (A)**: Tags the dominant modeling paradigms: Regression, Classification (C), Time Series (T), etc.

Each agent returns a decision as a categorical label with an associated confidence score $p_i \in [0, 1]$, where $i \in \{D, U, S, A\}$. The final structured output is:

$$\text{Tag}_{\text{MAINTAG}} = \{(i, \hat{y}_i, p_i)\}_{i=1}^{4} \tag{1}$$

where $\hat{y}_i$ is the predicted label for indicator i, and p_i is the confidence score derived from softmax-normalized heuristics and LLM scoring.

Indicator	Description and Options
Domain Type	Describes the broad industrial or operational context from which the dataset originates. Helps determine its relevance to real-world use cases and facilitates domain-aware benchmarking. • **Aerospace and Aviation (A)** – e.g., aircraft engines, UAVs, satellites • **Energy and Power Systems (E)** – e.g., fuel cells, turbines, generators • **Transportation and Mobility (T)** – e.g., railway bogies, vehicle suspension • **Manufacturing and Industrial Machinery (M)** – e.g., milling machines, gearboxes • **Electronics and Semiconductors (S)** – e.g., CMP, circuit systems
Usage	Captures the dataset's primary objective within PHM workflows. This classification is key for selecting datasets aligned with specific modeling goals. • **Detection** – Identify whether a fault has occurred (binary outcome) • **Diagnosis** – Determine the specific root cause of a failure • **Assessment** – Quantify current health or risk state • **Prognosis** – Predict future degradation or remaining useful life (RUL)
Nature of Supervision	Indicates the level of label availability in the training data, which constrains the type of applicable learning algorithms. • **Supervised** – Fully labeled data for fault types or degradation levels • **Unsupervised** – No labels; typically used for anomaly detection or clustering
Learning Algorithm	Describes the analytical technique or machine learning paradigm used to model the dataset. Useful for benchmarking, model selection, and reproducibility. • **Regression** – Predict continuous outcomes (e.g., RUL, wear) • **Classification** – Categorize conditions (e.g., fault/no fault, fault types) • **Time Series Prediction** – Forecast future sensor values or trends

Table 2. Expanded definitions and options for key indicators used in MAINTAG.

3.2. Data Collection

We evaluated MAINTAG on all publicly available PHM challenge datasets from 2008–2017 (see Section 2.1). These datasets come from diverse domains and represent real-world diagnostics and prognostics tasks, including RUL prediction, anomaly detection, and component-level classification.

The system also incorporates contextual metadata extracted via a WebSearchAgent to supplement sparse datasets. Each dataset was parsed for:

- Metadata (sensor types, time granularity, labels)
- Task structure (prediction vs. classification)
- Domain clues (nomenclature, source links)

Expert labels from Table: 1 served as the ground truth for benchmarking model performance.

3.3. Analysis Methods

MAINTAG's core processing pipeline consists of three layers:

1. **Intent Parsing**: An OrganizerAgent receives input metadata or user queries and routes the request to domain-specific agents.
2. **Inference and Classification**: Each indicator agent performs a rule-guided LLM inference using prompt engineering templates. The agents return a prediction $\hat{y}_i$, a justification string, and a confidence score p_i.
3. **Aggregation and Output**: The Runner aggregates agent outputs into a structured JSON result. If discrepancies or conflicts arise, a resolver heuristic recalculates scores using confidence-weighted majority voting.

Mathematically, the aggregation step uses a decision function:

$$\hat{y}_i = \arg\max_{c \in C_i} p(c \mid \text{agent}_i, \text{input})$$

where C_i is the class set for indicator i, and $p(c)$ is estimated from the LLM's logit distribution, normalized using a temperature-controlled softmax:

$$p(c) = \frac{\exp(z_c/\tau)}{\sum_{c' \in C_i} \exp(z_{c'}/\tau)}$$

where τ is a tunable temperature parameter (default $\tau = 1$) and z_c is the logit score from the agent's response structure.

Evaluation metrics included:

- Accuracy against expert-annotated labels for all 10 PHM datasets.
- Cross-consistency between domain type and algorithm choice (e.g., RUL tasks requiring regression in aerospace).

- Execution time and agent agreement levels (confidence entropy).

Table 3. MAINTAG accuracy per indicator across PHM challenges (2008–2017).

Indicator	Accuracy
Domain Type	0.80 (8/10)
Usage Type	0.90 (9/10)
Supervision	0.90 (9/10)
Algorithm	0.60 (6/10)

4. Evaluation and Findings

We evaluated MAINTAG's classification fidelity in relation to the benchmark taxonomies articulated by Jia et al. Jia et al. (2018), see (Figure). Our findings are structured around taxonomy alignment, agent confidence, and observed mismatches.

4.1. Key Findings

MAINTAG performed multi-label classification across all 10 major PHM data challenges (2008–2017), with most predicted tags aligning with the high-level taxonomy proposed by Jia et al. Jia et al. (2018), and at least one meaningful match per dataset. The classification was based on four indicators: Domain Type (DT), Usage Type (U), Nature of Supervision (NS), and Learning Algorithm (LA). Table 4 summarizes MAINTAG's automated decisions.

Key findings are:

- **Domain Identification (DT)** was mostly consistent across all datasets, with agreement in 8/10 cases. As requested, MAINTAG mapped turbofan engines to aerospace, gearboxes to manufacturing, and CMP systems to semiconductor manufacturing. The different cases were PHM13 and PHM-4, respectively. Both datasets lack detailed descriptions due to confidentiality, but they reference a common issue in industrial remote monitoring and diagnostics. As a result, Maintag labeled them under the "Manufacturing and Industrial Machinery" domain.
- **Usage Type (U)** was correctly predicted in 9/10 datasets, with nuanced interpretation for dual-purpose challenges (e.g., *PHM17*, which involved both diagnostics and prognostics data evaluation). The only differing case was PHM-2010, which was labeled as assessment in Table 1, yet its description clearly refers to remaining useful life estimation. Therefore, Maintag assigned it a prognosis tag.
- **Supervision Type (NS)** agreement was high in labeled datasets (9/10)
- **Learning Algorithm (LA)** predictions followed challenge structures, with mapping to regression for RUL tasks and classification or time series. The performance dropped for datasets lacking explicit labels. This indicator showed slightly lower performance, which was expected, as such learning algorithms are less frequently mentioned in the README files.

5. Insights and Discussion

MAINTAG delivered strong results when tested on past PHM datasets. The system matched human experts mostly in Domain Type tasks. It also performed well in Algorithm Type and Usage Type categories. The most striking finding was how closely it matched expert decisions. This happened without any direct training from humans. The system worked purely from text descriptions and data patterns.

The reduced agreement in Supervision Type highlights an inherent ambiguity in challenge documentation, which occasionally omits labeling strategies or blends methodologies (e.g., hybrid physical-statistical models). MAINTAG's confidence scoring, however, provides a mechanism for users to identify low-certainty outputs, which is a critical improvement over deterministic classifiers.

The deployment of MAINTAG can significantly streamline the dataset curation process in both academic and industrial settings. Potential implications include:

- **Faster Research Progress**: Scientists can find the right datasets for their work without ant explicit effort. Clear tags can make scientific data searching much practical than before.
- **Standardization**: MAINTAG can help different labs and contests use the same classification methods. This can cut down on confusing or conflicting dataset descriptions by different researchers.
- **Building Block in PHM**: Automated tagging can set the stage for smarter PHM workflows. Real-world factories need systems that can handle data, label it, and put models to work without delays.
  Also, MAINTAG lays the groundwork for benchmarking future datasets released beyond 2017 and supports alignment with FAIR data principles (Findable, Accessible, Interoperable, Reusable).
  While promising, the current version of MAINTAG has limitations discussed below, that we plan to address in future work):
  - **Dataset Sparsity**: MAINTAG depends on sufficient descriptive text. In cases where minimal documentation exists, its inference capabilities degrade significantly.
  - **Ontology Rigidity**: MAINTAG currently assumes a static four-indicator taxonomy. Extending to multi-label or hierarchical ontologies would enhance flexibility but introduces complexity.
  - **Absence of Continuous Learning**: The system does not currently adapt or fine-tune based on human feedback, though this could be integrated in future reinforcement learning phases.

	Class.	Rationale	Conf.
PHM08	DT: A	The dataset involves turbofan engines, which are core components of aircraft. The prognostics challenge focuses on predicting the Remaining Useful Life (RUL) of these engines.	H
	U: P	The primary aim is to estimate the Remaining Useful Life (RUL) of the engines, which is a prognostic task.	H
	NS: S	The dataset provides a training set with labeled RUL values, enabling the training of models in a supervised manner.	H
	LA: T	The data consists of multivariate time series from engine sensors, ideal for time series prediction models to forecast future performance and remaining life.	H
PHM09	DT: M	The dataset focuses on fault detection and magnitude estimation in gear systems, which are key components in industrial machinery.	M
	U: FD,D	The competition was centered around detecting faults and estimating their magnitude, requiring participants to identify fault types and locations.	H
	NS: S,U	The dataset began as unlabeled, but was later complemented with labels, allowing for both unsupervised and supervised learning approaches.	M
	LA: C,T	Algorithms were developed to classify fault types and predict fault progression over time, utilizing both classification and time series prediction models.	M
PHM10	DT: M	The challenge focuses on CNC milling machine cutters.	M
	U: P	The task is to estimate the Remaining Useful Life (RUL) of the equipment.	M
	NS: S	Training and test data are provided, allowing for model development and validation.	M
	LA: T	Estimating RUL involves predicting future states based on historical sensor data.	M
PHM11	DT: E	The challenge involves fault detection in anemometers, which are key in the wind power industry.	H
	U: FD	The main task is anemometer fault detection, indicating the usage is focused on detecting faults.	H
	NS: NA	The description does not specify if the data is labeled, so the nature of supervision remains unclear.	–
	LA: NA	Without specific details on the learning algorithm used, such as supervised models like classification or unsupervised like clustering, it remains unknown.	–
PHM12	DT: M	The challenge focuses on the life estimation of bearings, which is a critical component in industrial machinery.	M
	U: P	The primary focus of the challenge is on the estimation of the remaining useful life (RUL) of bearings, indicating a focus on prognosis.	H
	NS: S	The challenge included labeled data (run-to-failure datasets) provided to participants for building models, which implies a supervised learning context.	H
	LA: T	Participants were tasked with estimating the RUL, which involves time series predictions from operational and failure data.	H
PHM13	DT: M	The task involves maintenance action recommendation in an industrial context.	H
	U: D	The focus is on recognizing confirmed issues and avoiding false alarms (nuisance cases).	M
	NS: S	Requires labeled data for recommending problem types vs. nuisance cases, based on typical methodologies.	M
	LA: C	Teams used methods like Bayesian approach, decision trees, and ensembles to classify cases into problems or non-problems.	H
PHM14	DT: M	The task involves industrial remote monitoring and diagnostics of assets.	H
	U: FD,A	The task focuses on monitoring to identify assets as high or low risk of failure.	M
	NS: S,U	Mixed approaches may be involved in segmenting health scores into risk categories.	M
	LA: Cl,R	Health score generation likely involves both classification and regression techniques.	H
PHM15	DT: M	The dataset focuses on fault detection and prognostics within industrial plant settings.	M
	U: D,P	The dataset aims to detect faults and predict future failure events in plant operations.	H
	NS: S	The task involves predicting missing faults from provided training data, indicating supervised learning.	H
	LA: T,C	The use of time series data to predict future faults aligns with time series prediction, and identifying fault types aligns with classification.	H
PHM16	DT: M	The task involves predicting removal rates in CMP tools, typically used in manufacturing processes, particularly in semiconductors.	H
	U: NA	The classification for usage could not be determined based on the given information.	–
	NS: NA	The supervision type isn't clear from the description provided.	–
	LA: NA	The specific learning algorithms were not detailed in the available information.	–
PHM17	DT: T	The dataset focuses on tracking the health state of components within a train car, necessitating diagnostics specific to transportation systems.	H
	U: D,P	The task involves predicting faulty regimes (Prognosis) and detecting faults (Detection) in train components.	H
	NS: S	The dataset likely involves labeled training data to predict and diagnose faults.	H
	LA: S	The task involves predicting health states and faults using statistical and physics-based modeling.	H

Table 4. PHM Dataset Classification Summary (Acronyms for Indicators; rationales unchanged) NA:Not Available

6. CONCLUSION

This study introduces MAINTAG as an automated system designed to automatically apply descriptive tags to datasets in PHM. Essentially, MAINTAG is a multi-agent system consists of several independent AI "agents". Each agent is designed as an expert in a specific task and they allows the system to categorize datasets as intended.

The paper proposed a two-part evaluation to validate the model performance. First, it was tested against a collection of historical PHM data challenges. Second, we used the expert taxonomy from Jia et al. (2018) as our standard. This comparison showed how closely MAINTAG's classifications matched human expert decisions. Specifically, it achieved a high degree of alignment when identifying key dataset characteristics, such as the domain , the usage context , and nature of supervision.

While MAINTAG has proven to be an effective approach for expert-aligned classification, its architecture can provide a foundation for several advancements.Our future work will focus on evolving MAINTAG from a static tagging system into a more dynamic one. Therefore, the future research intends to explore

- **Incorporation of Semantic Retrieval**: Use of embedding-based similarity measures to classify datasets with minimal metadata.a
- **Online Learning**: Incorporation of user feedback and expert corrections into a loop to refine agent outputs.
- **Ontology Expansion**: Transition from fixed tags to hierarchical structures to support cross-domain mapping.

REFERENCES

Bohlin, M., Doganay, K., Kreuger, P., Steinert, R., & Warja, M. (2010). Searching for gas turbine maintenance schedules. *AI Magazine*, *31*(1), 21–36.

Brotherton, T., Jahns, G., Jacobs, J., & Wroblewski, D. (2000). Prognosis of faults in gas turbine engines. In *2000 ieee aerospace conference. proceedings (cat. no. 00th8484)* (Vol. 6, pp. 163–171).

Byington, C. S., Watson, M. J., & Bharadwaj, S. P. (2008). Automated health management for gas turbine engine accessory system components. In *2008 ieee aerospace conference* (pp. 1–12).

Dumschott, K., Dörpholz, H., Laporte, M.-A., Brilhaus, D., Schrader, A., Usadel, B., ... Kranz, A. (2023). Ontologies for increasing the fairness of plant research data. *Frontiers in Plant Science*, *14*, 1279694.

Eklund, N., & Bechhoefer, E. (2009). *PHM09 Challenge Data Set.* https://www.phmsociety.org/competition/phm/09.

Eklund, N., & Kessler, S. (2011). *Phm society 2011 anemometer dataset.* https://phmsociety.org/phm_competition/2011-phm-society-conference-data-challenge/.

Eklund, N., & Kessler, S. (2013). *Phm society 2013 maintenance log dataset.* https://phmsociety.org/conference/annual-conference-of-the-phm-society/annual-conference-of-the-prognostics-and-health-management-society-2013/phm-data-challenge/.

Eklund, N., Li, X., Bechhoefer, E., & Menon, P. (2010). *2010 phm society conference data challenge.* https://phmsociety.org/phm_competition/2010-phm-society-conference-data-challenge/.

Eklund, N. H. (2006). Using synthetic data to train an accurate real-world fault detection system. In *The proceedings of the multiconference on" computational engineering in systems applications"* (Vol. 1, pp. 483–488).

Garvey, D., & Wigny, R. (2014). *Phm society 2014 asset risk dataset.* https://phmsociety.org/conference/annual-conference-of-the-phm-society/annual-conference-of-the-prognostics-and-health-management-society-2014/phm-data-challenge-2/.

Gonçalves, R. S., Kamdar, M. R., & Musen, M. A. (2019). Aligning biomedical metadata with ontologies using clustering and embeddings. In *European semantic web conference* (pp. 146–161).

Heng, A., Zhang, S., Tan, A. C., & Mathew, J. (2009). Rotating machinery prognostics: State of the art, challenges and opportunities. *Mechanical systems and signal processing*, *23*(3), 724–739.

Heng, A. S. Y. (2009). *Intelligent prognostics of machinery health utilising suspended condition monitoring data* (Unpublished doctoral dissertation). Queensland University of Technology.

Huang, B., Di, Y., Jin, C., & Lee, J. (2017). Review of data-driven prognostics and health management techniques: lessions learned from phm data challenge competitions. *Machine Failure Prevention Technology*, *2017*, 1–17.

Jardine, A. K., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical systems and signal processing*, *20*(7), 1483–1510.

Jia, X., Huang, B., Feng, J., Cai, H., & Lee, J. (2018). A review of phm data competitions from 2008 to 2017: Methodologies and analytics. In *Annual conference of the prognostics and health management society.*

Kans, M., & Ingwald, A. (2008). Common database for cost-effective improvement of maintenance per-

formance. *International journal of production economics*, *113*(2), 734–747.

Lutz, M.-A., Schäfermeier, B., Sexton, R., Sharp, M., Dima, A., Faulstich, S., & Aluri, J. M. (2023). Kpi extraction from maintenance work orders—a comparison of expert labeling, text classification and ai-assisted tagging for computing failure rates of wind turbines. *Energies*, *16*(24), 7937.

Mishra, S., Glaws, A., Cutler, D., Frank, S., Azam, M., Mohammadi, F., & Venne, J.-S. (2020). Unified architecture for data-driven metadata tagging of building automation systems. *Automation in Construction*, *120*, 103411.

Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Chebel-Morello, B., Zerhouni, N., & Varnier, C. (2012). Pronostia: An experimental platform for bearings accelerated degradation tests. In *Ieee international conference on prognostics and health management, phm'12.* (pp. 1–8).

Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., Niyato, D., ... Poor, H. V. (2021). 6g internet of things: A comprehensive survey. *IEEE Internet of Things Journal*, *9*(1), 359–383.

Pimenov, D. Y., Bustillo, A., Wojciechowski, S., Sharma, V. S., Gupta, M. K., & Kuntoğlu, M. (2023). Artificial intelligence systems for tool condition monitoring in machining: Analysis and critical review. *Journal of Intelligent Manufacturing*, *34*(5), 2079–2121.

Propes, N., Girstmair, B., & Rosca, J. (2017). *Phm society 2017 bogie dataset.* https://phmsociety.org/conference/annual-conference-of-the-phm-society/annual-conference-of-the-prognostics-and-health-management-society-2017/phm-data-challenge-5/.

Propes, N., & Rosca, J. (2016). *Phm society 2016 cmp dataset.* https://phmsociety.org/wp-content/uploads/2016/05/PHM16DataChallengeCFP.pdf.

Ramasso, E., & Saxena, A. (2014). Performance benchmarking and analysis of prognostic methods for cmapss datasets. *International Journal of Prognostics and Health Management*, *5*(2), 1–15.

Rosca, J., Williard, N., Eklund, N., & Song, Z. (2015). *Phm society 2015 power plant dataset.* https://www.phmsociety.org/events/conference/phm/15/data-challenge.

Sarker, S., Arefin, M. S., Kowsher, M., Bhuiyan, T., Dhar, P. K., & Kwon, O.-J. (2022). A comprehensive review on big data for industries: challenges and opportunities. *IEEE Access*, *11*, 744–769.

Saxena, A., & Goebel, K. (2008). *PHM08 Challenge Data Set.* https://www.nasa.gov/content/prognostics-center-of-excellence-data-set-repository. (NASA Prognostics Data Repository, NASA Ames Research Center, Moffett Field, CA)

Simões, J. M., Gomes, C. F., & Yasin, M. M. (2011). A literature review of maintenance performance measurement: A conceptual framework and directions for future research. *Journal of Quality in Maintenance Engineering*, *17*(2), 116–137.

Su, H., & Lee, J. (2023). Machine learning approaches for diagnostics and prognostics of industrial systems using open source data from phm data challenges: a review. *arXiv preprint arXiv:2312.16810*.

Tsui, K. L., Zhao, Y., & Wang, D. (2019). Big data opportunities: System health monitoring and management. *IEEE Access*, *7*, 68853–68867.

Uusipaavalniemi, S., & Juga, J. (2008). Information integration in maintenance services. *International Journal of Productivity and Performance Management*, *58*(1), 92–110.

Zhao, Z., Wu, J., Li, T., Sun, C., Yan, R., & Chen, X. (2021). Challenges and opportunities of ai-enabled monitoring, diagnosis & prognosis: A review. *Chinese Journal of Mechanical Engineering*, *34*(1), 56.

ACKNOWLEDGMENT

(University of Luxembourg authors): This research was funded in part by the Luxembourg National Research Fund (FNR), grant reference [5G BRIDGES/2023-Phase 2/IS/19113706/5G-ARTEMIS].