IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 June 2023, accepted 9 July 2023, date of publication 13 July 2023, date of current version 24 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3295113

SURVEY

Publicly Available Datasets for Predictive Maintenance in the Energy Sector: A Review

EDA JOVICIC[1], (Member, IEEE), DARIA PRIMORAC[2], MARKO CUPIC[1], (Member, IEEE), AND ALAN JOVIC[1], (Member, IEEE)

[1]Faculty of Electrical Engineering and Computing, University of Zagreb, 10000 Zagreb, Croatia
[2]Independent Researcher, 8000 Zurich, Switzerland

Corresponding author: Alan Jovic (alan.jovic@fer.hr)

This work was supported by the European Regional Development Fund in the Republic of Croatia under the Operational Programme Competitiveness and Cohesion 2014–2020, "Platform for Intelligent and Energy Efficient Control of Industrial IoT Devices," under Grant KK.01.2.1.02.0001.

ABSTRACT Predictive maintenance (PdM) uses statistical and machine learning methods to detect and predict the onset of faults. PdM is often used in industrial IoT settings in the energy sector, where research works usually consider specific types of faults depending on the application. However, since PdM is mainly data-driven and needs to work in real time, the public availability of datasets is required in order to build efficient and effective models applicable across multiple domains. Unlike methods, the publicly available datasets obtained from sensors in the energy sector have not been properly reviewed or categorized. In this work, we consider five subsectors of the energy sector: wind, solar, oil & gas, diesel & thermal, and electrical power grid. We provide a detailed description of the properties of the publicly available PdM datasets in these subsectors. The review of the datasets is conducted on a number of scientific and commercial repositories: IEEE DataPort, UCI Machine Learning Repository, Kaggle, EDP, and Mendeley Data. The datasets are graded into three categories according to objective criteria. We also provide references to significant related research work that uses the considered datasets. The observed challenges in using the datasets in this field are thoroughly discussed. We find that there is a troublesome scarcity of publicly available datasets in the energy sector, more so of data coming from real, non-simulated sources. Three datasets, 3W (oil & gas), EDP-WT (wind), and OREC (wind) stand out as highly valuable for researchers in this field.

INDEX TERMS Datasets, deep learning, machine learning, predictive maintenance (PdM), energy sector.

I. INTRODUCTION

Predictive maintenance (PdM) aims to successfully estimate the period in which in-service equipment maintenance should be performed to avoid its potential failure and the associated consequences. PdM encompasses many data-driven methods, mostly from statistics and machine learning, in order to achieve the goal of efficient fault detection and prediction [1]. The advantages of using PdM include: 1) expenditure savings through the lower cost of maintenance due to knowing in advance when to buy a certain spare part or piece of equipment, 2) energy savings through optimizing the exploitation of non-renewable sources (oil & gas) and renewable sources, and 3) a safer work environment achieved by detecting a fault and stopping production before failure happens. According to a study that considered 268 European companies from various sectors, PdM decreased costs by 12%, improved availability by 9%, extended the lifetime of an aging asset by 20% and reduced safety, health, environmental and quality risks by 14% [2].

PdM is broadly considered to be an important part of Industry 4.0 [3], [4], [5]. It has been a part of industrial development for many years, mostly in the context of sensor networks [6], [7] and later, the Internet of Things (IoT) [8], [9]. It is often applied to industrial IoT (IIoT) signal data acquired under various levels of control [10], [11]. Fig. 1 depicts the number of published research papers in the field of PdM in the industrial context from 2013 to 2022 according

The associate editor coordinating the review of this manuscript and approving it for publication was M. Shamim Kaiser.

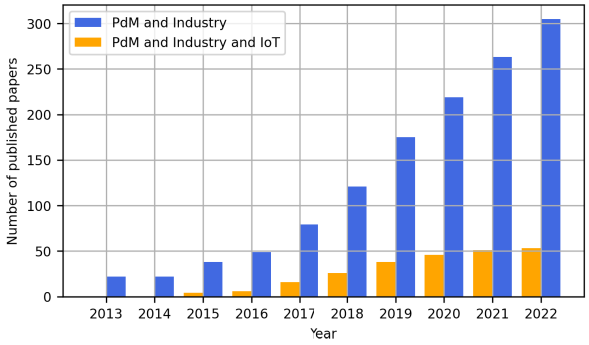

FIGURE 1. Published papers in the field of predictive maintenance in an industrial context, compared with an explicit mention of IoT, from Web of Science Core Collection database.

to the Web of Science Core Collection database. The number of papers that explicitly mention IoT with PdM is also shown. It can be observed that the field is becoming increasingly popular and that the use of IoT in PdM is growing steadily.

Still, there are some challenges for PdM that can be observed from the related work, such as: 1) the methods are data-driven [12], which means that without adequate data the models are unsuccessful; 2) the industry itself is mostly closed, therefore, proprietary data, especially labeled data, are usually not shared, which limits the research works; 3) the optimal PdM models require close collaboration between the academic sector and industry due to the significant domain knowledge and artificial intelligence requirements posed to such models, which is not often the case.

Currently, from the perspective of data science, researchers observe a lack of available high-quality datasets in the field of PdM that would allow them to construct wide-spread and applicable PdM models [13]. This problem seems to be pronounced in the IoT-supported energy sector, which leads to suboptimal smart solutions and increased costs and impacts on the environment. Namely, the research work in the energy sector mainly focuses on a specific application and a specific dataset [14], [15]. The developed models are not applicable to other PdM datasets from the same subsector or from different subsectors due to significant differences in the datasets' properties. This statement holds even when recent approaches from deep learning, such as transfer learning, are used [16] because the datasets are too specific.

As stated earlier, most PdM approaches are data-driven, meaning that they are dependent on provided data. Due to the importance of high-quality data availability topic, in this paper, we present a detailed survey of the PdM datasets in the energy sector. We show that there is indeed a scarcity of publicly available PdM datasets in this field, especially the non-simulated ones. The contributions of this review work are the following:

- a detailed description of the properties of the publicly available predictive maintenance datasets in the IoT-supported energy sector,
- a reference list of significant research works that use the considered datasets along with applied methods and PdM tasks,
- a categorization (rating) of the datasets based on a small set of objective criteria,
- a generalized data-driven pipeline for predictive maintenance,
- a discussion of the observed challenges related to the datasets in this field.

Unlike other research that focuses on various applicable machine learning methods and constructed models, the focus of this work is on the elaboration of the datasets that can be used by researchers to build high-impact and versatile PdM machine learning models in the energy sector. We consider that this kind of work is important for future considerations, methods and datasets alike, in this growing field.

The remainder of this paper is structured as follows. In Section II, we review the related work in the field of PdM. In Section III, we explain the precise methodology of the review. Section IV deals with a detailed description of the reviewed datasets, their properties and the corresponding research work. Section V presents a data-driven pipeline for PdM. We discuss the encountered challenges in Section VI and conclude the paper in Section VII.

II. RELATED WORK

Numerous research works in the field of predictive maintenance are focused on maintenance methods, the application of machine learning (ML) and deep learning (DL) methods, as well as types of failures in specific application domains. However, hardly any attention was given to the review of datasets on PdM.

Data-driven methods are the most common choice for achieving PdM. The authors in [17] classify six ML and DL algorithms in specific industrial applications and compare five performance metrics for each classification algorithm. Furthermore, they list the most common challenges in practice, which also include the challenges of data acquisition. Similarly, the paper [4] gives a review of not only methods but also architecture and provides a list of 13 crucial targets for PdM in Industry 4.0 that are applicable to Small and Medium Enterprises (SME). The authors observed that most papers focused on increasing the remaining useful life (RUL) of the system or detecting anomalous events, while the most common choice of technology for applying PdM is ML and DL. In comparison, the paper [18] gives a deeper insight into common ML methods applied to four types of industrial maintenance approaches: corrective, preventive, condition-based, and PdM. Likewise, in [1], the authors focus on four mainstream DL-based methods for Intelligent Predictive Maintenance (IPdM). In there, all methods were compared in terms of data characteristics, model performance, and application scenarios.

Nowadays, PdM often utilizes the IoT for real-time data acquisition in order to efficiently prepare the maintenance works. The application of IoT in PdM was surveyed in a recent paper [19] using scientometric analysis to point out the most common keywords, cited authors, contributing countries, and cited journals. The authors also presented

the applications of PdM in industries, including the energy sector, along with the main benefits of PdM, which are safety, security, reliability, and efficiency. Contrary to the considered benefits, the authors in [20] presented the main challenges in the IoT-Enabled PdM. Those challenges include the selection of the components that would benefit most from PdM, the proper design of the IoT infrastructure, the development of the algorithms and methods, and lastly, the exploitation of IoT-enabled monitoring to really ensure that PdM brings added value. One of the innovative uses of IIoT as the main tool is a new paradigm called Hybrid Self-Corrective Maintenance (Hybrid-SCM) [21]. The paradigm was proposed after reviewing PdM industry case studies. The Hybrid-SCM combines Condition-based Maintenance (CbM) with Self-Corrective Maintenance (SCM) to create a subsystem that can learn about its condition by itself and take corrective actions when necessary.

Other research papers focus on specific use-cases of PdM, such as PdM for motors [22], [23], wind turbines [14], [24], [25], hydraulic cylinders [26], power transformers [15], etc. In these papers, the focus is on the applications of different data analysis methods in the corresponding fields. We will now review some of these papers.

As summarized in [22], different approaches can be followed for applying PdM in motors. All of them included ML methods, where the most common one was Random Forest (RF), combined with different motor features, e.g., vibration, acoustical and speed oscillations, motor current, etc. A similar review was done for induction motors [23], with emphasis on the types of faults. The most common faults were bearing faults, which also had the best detection accuracy. In wind turbines (WT) for the energy sector, SCADA systems usually collect data that can be used for power prediction, fault detection, optimal control settings, performance evaluation, and necessary maintenance. Since the number of offshore wind turbines has been growing, research has focused on PdM and fault diagnosis techniques focused on windings and insulation failures [25]. Furthermore, in [14], the emphasis was on predicting the RUL for offshore wind turbine power converters. The authors reviewed the existing methods and proposed a novel methodology using a digital twin framework for implementing PdM. Understanding the types of failures and how often they might occur in wind turbines showed how the growing size of WT generators brings new maintenance problems, even though they seem to be more robust and reliable [24]. On the other hand, hydraulic cylinders are widely used in different industries as mechanical actuators, and they are affected by a variety of factors, such as fluid contamination, fluid leakage, worn piston rods, or internal corrosion. In [26], the results of using various sensors to diagnose these faults are reviewed. Another approach to using RUL was shown in [15], where winding hot-spot temperature usually determined the remaining life of the power transformer, an important unit used in electric power generating stations.

Another field of research in PdM is the use of various signal processing and analysis techniques, such as Fast Fourier Transform, Wavelet Transform, and Artificial Neural Network (ANN). This line of work is important because PdM is mainly based on industrial time series, which are affected by noise and artifacts. In a recent work [27], a review of processing methods was done for current, vibration, and acoustic signal analysis for PdM. Paper [28] focused only on the vibration signal analysis due to its low cost and better results compared to others, especially when using features with a higher dimension, as opposed to stationary signal processing techniques, such as the ones described in [29].

To our knowledge, there is only one paper that gives an overview of the datasets available for a specific industry. This paper [30] presented a summary of datasets from wind industry-related resources. The study also offered a review of research papers that made use of these datasets. The research topics range from evaluating wind potential to predicting wind speed and the consequent output power. The listed datasets were grouped into three domains: open datasets of wind turbine capacity and wind farm projects, wind resources, and wind farm monitoring. While interesting for the wind subsector, this study did not consider a broader perspective on the available PdM datasets in the energy sector. Nevertheless, the paper was motivational for this study.

Our review of the literature did not find any research work that would provide a more general overview of the available datasets for PdM. Review papers in the field of PdM are focused on the review of maintenance methods, signal processing methods, and types of failures in a specific application.

III. REVIEW METHODOLOGY

An overview of the review methodology for selecting datasets is depicted in Fig. 2. The datasets were reviewed in a number of scientific and commercial repositories, namely IEEE DataPort, UCI Machine Learning Repository, Kaggle, EDP, and Mendeley Data. The keywords used for searching the datasets were: fault, fault detection, predictive maintenance, oil well, oil and gas, wind turbine, wind power, solar power, photovoltaic farm, boiler, diesel engines, electric grid, electric power, and transmission line.

A. INCLUSION CRITERIA

A dataset had to meet certain criteria in order to be included in the review: 1) the dataset had to have logs containing information about faults, anomalies, or maintenance so that it could be used for PdM; 2) the dataset had to be related to the energy sector and its five considered subsectors: wind, solar, oil & gas, diesel and thermal power, electrical power grid; 3) the dataset had to be IoT-related and not be for general industrial use (such as hydraulics, gearboxes, bearings, etc.) due to the significant expansion of the published scientific papers related to IoT; 4) the data had to be in the format of a time series (images or graphs were not considered). Only

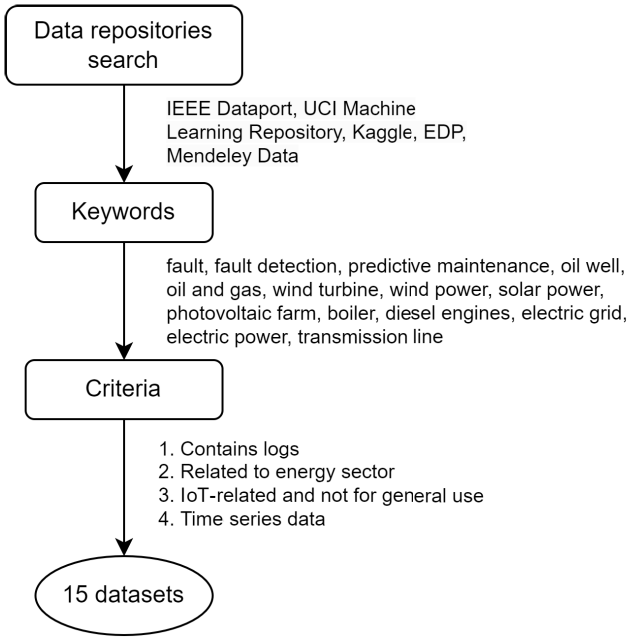

FIGURE 2. Review methodology for selecting datasets.

the datasets that satisfy all of the criteria are selected and reviewed in this paper.

B. DATASET GRADING

For evaluating the quality of the considered datasets for PdM, we devised a grading system, where each dataset was graded as class I, II, or III, with class I being the highest grade and class III being the lowest grade. Several simple and objective criteria were used for grading the datasets, as follows. A class I graded dataset can not contain data from simulations. Also, for accessibility reasons, if the dataset was not open-access, but required significant effort (aside from registration), it can not be graded as class I. The grade of the dataset is lowered if it has a relatively small number of instances (less than 1,000), predictive variables (only 1), or fault events (5 or fewer). The motivation for using these thresholds is that ML and DL models may not work properly if the above criteria are not met. If a dataset has missing or limited documentation describing the data, or if the licence information is not available, the dataset is graded as class III. The grading criteria are summarized in Table 1.

To sum up, only datasets that contain data from a real source, have more or equal to 1,000 instances, more than 1 predictive variable, more than 5 faults, have complete documentation and licence information and easy access to the data, can be graded as class I datasets. Simulated datasets, or datasets with a number of instances, variables or faults under the threshold, or datasets that have a harder access to data (aside from the registration process) are graded as class II. Datasets that are both simulated and have a number of instances, variables or faults under the threshold, or hard access, or missing documentation and licence information are graded as the lowest class: class III. To promote future datasets, the criteria were intentionally devised in such a way that a newly published high-quality dataset that is not yet cited in relevant literature can still be graded as class I.

TABLE 1. Criteria expressions for class affiliation.

Class	Data source, Nº ins, Nº vars, Nº faults, Access, Doc & licence
I	Real AND $\geq$1000 AND $>$1 AND $>$5 AND Easy AND Complete
II	Simulated OR $<$1000 OR $=$1 OR $\leq$5 OR Hard AND Complete
III	{[Simulated AND ($<$1000 OR $=$1 OR $\leq$5)] OR Hard} OR Missing

IV. DATASETS

After taking into account all of the criteria in the inspected repositories, 15 datasets were selected and are considered in this review.

Table 2 shows the main properties of the obtained datasets. For each dataset, the type of data is listed. For wind turbines, the datasets are usually in SCADA system format together with meteorological (MET) data. On the other hand, other data are collected from a variety of sensors (e.g., temperature, pressure, vibration, voltage, and current sensors) that are specific to each dataset. We also list the usage licences. Most licences are CC-BY, which allows users to freely distribute, remix, adapt, and more, as long as the creator is properly credited. This type of licence also allows commercial use of datasets. Whether a dataset contains documentation describing the data is noted in the Doc column, as is information about whether the dataset is simulated in the Sim column. To better understand the datasets from a data science viewpoint, the table also shows the granularity of the data, the number of instances (Nº ins), the number of variables (Nº vars), the number of faults (Nº faults) and the name of the repository where the dataset is provided. The last column contains the grades (I, II or III) for each dataset.

Table 3 shows the properties of 26 published journal and conference papers from the Web of Science Core Collection (WoSCC) database where research work was done on one of the considered datasets. The statistical, ML and DL methods used in each paper were listed. The last described property is the type of PdM task done. Exploratory data clustering searches for feature clusters correlated to faulty states. Fault prediction refers to predicting a fault in advance, while fault detection tries to detect whether a fault is happening in a specific instance. Fault type classification focuses on differentiating types of faults, while predicting decay state focuses on estimating the actual decay state of a specific component so proper decisions can be made later. Not all the datasets had published papers that reported their use, i.e., only 10 of

TABLE 2. The overview of the main properties of the obtained datasets.

Dataset	Data	Licence	Doc/Sim	Granularity	№ ins	№ vars	№ faults	Repository	Grade
Wind Power									
EDP-WT [31]	MET, SCADA	CC BY-SA	Y / N	10 min	$5.2 \cdot 10^5$	121	28	EDP Open Data	I
WT-IIoT [33]	SCADA	not defined	N / N	10 min	49,029	64	553	Kaggle	III
OREC [34]	MET, SCADA	documented	Y / N	1 s & 10 min	$1.5 \cdot 10^8$ & $2.5 \cdot 10^5$	603 & 711	N/A	dataPOD	I
Wind turbine PMSG [35]	current	ODbL	Y / Y	1 kHz	$4.2 \cdot 10^5$	2	40	Kaggle	II
Solar Power									
Fault Detection in PV Farms [36]	current, voltage	CC BY-SA	Y / Y	-	700	31	575	Kaggle	III
GPVS-Fault [37]	current, voltage	CC BY	Y / Y	10 kHz	$2.2 \cdot 10^6$	14	16	Mendeley Data	II
Oil & Gas Industry									
3W [32]	temperature, pressure	CC BY	Y / N	1 Hz	$5.0 \cdot 10^7$	8	1,387	Kaggle	I
Valhall OP [10]	sensor data	documented	Y / N	sensor dependent	$\sim 1 \cdot 10^8$	live stream	N/A	Kaggle	II
Diesel and Thermal Power Plants									
Naval Propulsion Plant [38], [39]	sensor data	CC BY-NC	Y / Y	-	11,934 & $3.5 \cdot 10^5$	18 & 30	11,934 & $3.5 \cdot 10^5$	UCI ML Repository	II
3500-DEFault [40]	vibration, pressure	CC BY	Y / Y	15 kHz	3,500	97	3,250	Mendeley Data	II
EDP - Inside a Boiler [41]	sensor data	CC BY-SA	Y / N	5 min	$5.8 \cdot 10^5$	114	2	EDP Open Data	II
Simulated Boiler Fault Data [42]	sensor data	CC BY	Y / Y	-	27,280	7	25,520	IEEE DataPort	II
Electrical power grid									
Transmission line faults [43]	current, voltage	CC0: Public Domain	Y / Y	1 kHz	$1.2 \cdot 10^4$ & 7861	11	5,496	Kaggle	II
Transformer and PAR transients [44]	current, voltage	CC BY	Y / Y	10 kHz	$7.3 \cdot 10^7$	4	$7.3 \cdot 10^7$	IEEE DataPort	II
Transients in ISPARs [45]	current	CC BY	Y / Y	-	60,552	4	60,552	IEEE DataPort	II

the 15 datasets included in this review were cited in published articles indexed in WoSCC. In the selected articles, the most commonly used statistical method was principal component analysis (PCA). Among the ML methods, RF was used in 50% of the articles studied. After RF, the second and third most frequently used methods were Support Vector Machine (SVM) and Decision Tree (DT). Most of the published papers used more than one ML method, either in an ensemble or for cross-comparison of models. Less than 50% of the included papers used both statistical and ML methods. DL methods (e.g., CNN, LSTM) were used only occasionally, despite their recent popularity in the AI community.

The EDP-WT dataset [31] has the most papers (six of them), and the second dataset, 3W [32], has five published papers in WoSCC. In both cases, a high number of papers correlates with the highest grading of the dataset, class I. The Naval Propulsion Plant dataset (1&2) has the same number of published papers as dataset 3W, although in this case the papers were authored by only three different research groups that used similar ML methods.

More detailed properties of the considered datasets are described below.

A. EDP-WT

The EDP Wind Turbine dataset is part of the EDP-WT Failure Detection Challenge, which evaluated predictive capabilities for detecting early failures in WTs. The components monitored are the gearbox, generator, generator bearing,

TABLE 3. Datasets in articles.

Dataset	Published work	Statistical methods	ML methods	PdM task
EDP-WT	Menezes et al. 2020 [30]	PCA	-	Exploratory data clustering
	Garan et al. 2022 [46]	high correlation filter, PCA, ICA	DT, feature selection	Fault prediction, RUL
	Udo and Muhammad 2021 [47]	Statistical Process Control (SPC)	LSTM, MLR, XGBoost	Fault prediction
	Latiffianty et al. 2022 [48]	CUSUM	LoMST	Fault detection
	Tidriri et al. 2021 [49]	-	DT, RF, MLP, SVC	Fault prediction and type classification
	de Sa et al. 2020 [50]	-	Soft Label SVM, Hard Label RF, NSGA-II	Fault detection
OREC	Chatterjee and Dethlefs 2020a [51]	-	LSTM, XGBoost	Fault prediction
	Chatterjee and Dethlefs 2020b [52]	-	Attention-based CNN, deconfounder	Fault prediction
Fault Detection in PV Farms	Ghoneim et al. 2021 [36]	-	RF, LR, NB, AdaBoost, CN2 rule induction	Fault detection and type classification
GPVS-Fault	Bakdi et al. 2021 [53]	KLD, PCA, KDE	-	Fault detection
	Wali and Khan 2022 [54]	-	RF, SHAP	Fault type classification
3W	Marins et al. 2021 [55]	PCA	RF	Fault detection and type classification
	Li et al. 2021 [56]	Rademacher complexity	TF-IDF, knowledge graph embedding	Fault prediction
	Soriano-Vargas et al. 2021 [57]	z-score	Isolation Forest	Fault detection
	Turan and Jäschke 2021 [58]	t-test, PCA, LDA, QDA	DT, LR, SVM, RF, AdaBoost	Fault prediction and type classification
	Carvalho et al. 2021 [59]	z-score, LDA, QDA	1NN, GNB, ZR, RF	Fault detection
Naval Propulsion Plant (1&2)	Coraddu et al. 2015 [60]	-	SVR, RLS	Predicting decay state
	Coraddu et al. 2016 [38]	-	SVR, RLS	Predicting decay state
	Cipollini et al. 2018a [61]	-	ANN, RLS, SVR, RF, RRE, KNN	Predicting decay state
	Cipollini et al. 2018b [62]	-	ANN, RLS, SVR, RF, RRE, KNN, GKNN, OCSVM	Predicting decay state
	Tan et al. 2021 [63]	-	Isolation Forest	Predicting decay state
Simulated Boiler Fault Data	Shohet et al. 2020 [64]	-	KNN, DT, RF, SVM	Fault detection
Transmission line faults	Jamil et al. 2015 [65]	correlation coefficient	ANN	Fault detection and type classification
Transformer and PAR transients	Bera et al. 2018 [66]	t-test	Extremely Randomized Trees, RF, MLP, LR, SVM	Fault type classification
	Bera et al. 2020 [67]	-	SVM, RF, DT, GBC	Fault detection and type classification
Transients in ISPARs	Bera and Isik 2021 [68]	-	Minimum Redundancy Maximum Relevance, RF, XGBoost, NB, SVM, NN, KNN	Fault detection and type classification

transformer, and hydraulic group [31]. The inspection, repair, and replacement costs of these parts can be found in the documentation and are used to assess the savings and predictive power of the algorithm. According to the ranking of the proposed methods on the challenge, the highest savings achieved was 76,000€.

The dataset includes 2 years of SCADA data (10-minute period) from 5 WTs (Wind Farm 1) located in the West African Gulf of Guinea and the meteorological mast, including fault detections and event logs. These datasets are also already split into training and test datasets (80/20 ratio). The training dataset is from 2016 (a full year) and the test dataset consists of nine months of data, from January 2017 to September 2017, with a minimal number of records missing. All datasets are only available to registered users of the EDP Open Data Platform where the process of registering is simple.

Some of the SCADA signals (81) are the average temperatures of various WT components (e.g., gearbox, oil in hydraulic group, and generator bearing), average rotor speed, total active power, average nacelle direction, etc. Some of the signals from the meteorological mast (40) are maximum, minimum, and average wind speed, wind direction, pressure, and humidity from multiple sensors. The dataset contains 28 faults altogether. On the other hand, the WT logs show various remarks for the five WTs over the course of two years. Examples of ∼100 different remarks include: pause pressed on a keyboard, hot generator, pause over RCS, high wind speed, oil leakage in the hub, etc. However, there is no documentation available for all the values (codes) in the log records. These logs may carry some useful information about the data, but the information is not necessary for application of PdM, so the lack of documentation does not lower the grade of the dataset.

B. WT-IIoT

WT-IIoT dataset comes from a single unknown WT. The dataset is available on the Kaggle domain and originates from the Microsoft Azure Predictive Maintenance Template [33]. The data could be used to classify fault modes based on various SCADA components. The problem is that while the fault log and maintenance documentation are provided (tables *fault_data* and *status_data*), the dataset lacks a full description of the variables and status/fault codes. Licence information is also not included. Out of the provided data, the status data have the longest recorded time span from January 2014 to December 2015. The shortest is SCADA data from April 2014 to April 2015. There are some missing values, such as when SCADA timestamps don't match fault timestamps, simply because at certain times there are no faults happening.

Some of the 64 SCADA variables are: reactive power, blade angle, nacel position, and temperatures of the various systems (e.g., bearing, rotor, and stator). The meteorological variables are limited to minimum, maximum, and average wind speed. The *status_data* table contains 9 variables, the most useful variable for applying PdM being the status text describing the current operating state of WT. Other variables include main status, sub status, full status, T, service, fault message, and value 0, which could be useful if additional documentation were provided to describe each categorical value. In the fault data, there were five types of faults represented through 553 fault events, even though what they represent is unknown.

C. OREC

OREC (Offshore Renewable Energy Catapult) is the UK's leading technology innovation and research centre for offshore renewable energy. OREC's Levenmouth Demonstration Turbine (LDT) is located off the coast of Fife in Scotland and is the world's most advanced open access offshore wind turbine (7 MW, Samsung) dedicated to research and development. The OREC's data collection consists of a meteorological and a LDT SCADA dataset provided at 1-second and 10-minute intervals, respectively [34]. The collection can be searched through the *POD (Platform for Operational Data)* service, where a small data sample can also be retrieved. To gain full access to the data, a POD registration is required, where the customer must specify how the data will be used. The terms and conditions agreement is extensive (24 pages) and differs for each dataset.

The large number of available variables can be filtered by functional groups (cooling system, machinery enclosure, met mast, general, alarms, etc.), measured variables (temperature, pressure, rotation, etc.), units, and data types. After specifying the variables and time interval, a small fee is charged to cover the cost of data retrieval, depending on the size or complexity of the query. It is charged only after the responsible person has received the query.

All data are available as of January 2017 and are updated every month, except for LDT Substation data, which are available as of September 2017. Therefore, the estimated number of the meteorological mast and SCADA entries is $1.5 \cdot 10^8$ and $2.5 \cdot 10^5$ for 1-s and 10-min data, respectively. Since data access is limited, the number of missing values is unknown. Some of the SCADA signals (573) are: yaw brake pressure, yaw motor temperature state, mainframe and hub temperatures, generator export energy, and rotor speed. Some of the signals from the meteorological mast (14 for 1-Hz measurements and 66 for 10-minute measurement intervals) are: wind speed at different heights, wind direction, pressure, and temperatures at different heights. The alarm log contains ten columns, some of which include: the time the alarm started and ended, the downtime, the reference number indicating the event code, and the source of the stoppage. The alarm log could be used as a target variable when applying PdM. Some of the SCADA signals from the LDT substation (16 for 1 Hz measurements and 72 for 10-minute measurement intervals) include: power factor, reactive power, voltage, and current. A detailed description of the SCADA, meteorological and all other variables is available on site. In addition, documentation is included with the SCADA records explaining the WT

status codes and abbreviation entries. Event codes and their descriptions are also available on the LDT Alarm Log Record page.

D. WIND TURBINE PMSG

Permanent Magnet Synchronous-based Generator (PMSG) are commonly used as wind generators in wind turbines. They recurrently interrupt their operation due to stator faults. These faults usually occur between turns (turn-to-turn fault) or between a turn and the machine housing (turn-to-ground fault) [69]. The most common type of internal fault is the leakage current of the coils through connections caused by faults in the insulation of the components. The objective of the dataset is to enable evaluation of the effects of the fault severity due to different positions of the fault in the stator coil and the number of turns [35].

The dataset comes from simulating a mathematical model of PMSG using Simulink/MATLAB. There are several parameters varied for the simulation:

1) types of operation: normal, turn-to-turn, turn-to-ground
2) generator load [%]
3) switching frequency of the power converter [kHz]
4) percent of faulty turns [%]
5) fault resistance: range from 0 (total insulation break) to ∞ (no faulty).

The dataset contains instances representing current signals from normal stators and faulty stators in the range of 1% to 10% of faulty turns of type 1 and type 2. The data were saved in .mat format, with the file name containing event information (switching frequency, type of fault, and fault resistance). Each file contains two variables: I_s measures current values, and *tempo* measures time steps. The dataset includes 42 events, 40 of which are of the faulty type and 2 of the normal type. Since this is a simulated dataset, there are no missing data.

E. FAULT DETECTION IN PV FARMS

This dataset consists of measurements on a simulated 250-kW PV system created using Simulink/MATLAB [36]. The purpose of the dataset is to evaluate the effects of faults happening on various locations and various conditions on PV systems. The PV system consists of 88 parallel strings, each including seven series modules. Each module has 128 cells, a maximum power of 414.801 W at 72.9 V, a current of 5.69 A, an open-circuit voltage of 85.3 V, and a short-circuit current of 6.09 A.

The data were split into training (600 instances) and test (100 instances) datasets. There are four defined states:

1) free-of-fault (16.67%)
2) string fault, tested on string 1 (25.50%)
3) string to ground fault, tested on string 1 (24.83%)
4) string-to-string fault, tested between strings 1 and 2 (33%)

There are 12 attributes, 6 of which are currents measured by 2 ammeters at the top and bottom of the strings 1, 2, and 3 during simulation, total average DC voltage, total average DC power, total average current, temperature, radiation, and class. For each current measurement, an average, maximum, minimum, and variance value were extracted, giving a total of 30 features. The temperature, radiation, and fault resistance measurements ranged from 10Â°C to 35Â°C, 100 W^2 to 1,000 W^2, and 1 Ω to 2,000 Ω. The total simulation time was 0.4 s, and it was assumed that a fault occurs at 0.2 s. In the training dataset, all measurements were made after the fault occurred in the period from 0.2 s to 0.4 s.

F. GPVS-FAULTS

The Grid-connected Photovoltaic System Faults (GPVS-Faults) dataset is the result of laboratory experiments on faults in a PV microgrid application [37]. The data were obtained from sensor measurements and a virtual Phasor Measurement Unit (PMU). Experiments ran for approximately 10 to 15 seconds, with faults manually inserted halfway through the experiments. Each experiment was run in two modes: Limited Power Mode (IPPT) and Maximum Power Mode (MPPT).

There are 16 data files corresponding to seven types of faults (inverter fault (F1), feedback sensor fault (F2), grid anomaly (F3), PV array mismatch (F4, F5), MPPT/IPPT controller fault (F6), and boost converter fault (F7)) and one fault-free experiment. The faults have different severities and occur at different locations. The data files are available in both .mat and .csv formats. Each data file contains 14 variables: time, PV array current and voltage measurements, DC voltage measurements, phase A, B, and C current and voltage measurements, and positive-sequence estimated current and voltage magnitude and frequency. A more detailed description of the individual faults and the experimental setup can be found in the documentation [53].

Unlike other simulated experiments, the exact timestamp of fault occurrence is not known, the high-frequency measurements are noisy, and there are temperature and insolation disturbances and variations during and between scenarios. Different modes (MPPT or IPPT) have adverse effects on detecting low-magnitude faults. The challenge is to detect the faults before they cause total failure.

G. 3W

The 3W dataset consists of data collected by the Brazilian company Petrobras on naturally flowing offshore wells [32]. The goal of this dataset is to evaluate the effects of different types of events using eight process variables. The name 3W was chosen because the dataset is composed of instances from **3** different sources (real, simulated, and hand-drawn) that contain adverse events in oil **W**ells. The eight types of events are:

1) Abrupt Increase of Basic Sediment and Water (BSW)
2) Spurious Closure of the Downhole Safety Valve (DHSV)
3) Sever Slugging
4) Flow Instability
5) Rapid Productivity Loss
6) Quick Restriction in the Production Choke (PCK)

7) Scaling in Production Choke (PCK)
8) Hydrate in production line

The adverse events are characterized by eight process variables: pressure at the permanent downhole gauge (P-PDG), pressure (P-TPT) and temperature at the temperature/pressure transducer (T-TPT), pressure upstream of production choke (P-MON-CKP), temperature downstream of production choke (T-JUS-CKP), gas lift flow rate (QGL), pressure (P-JUS-CKGL) and temperature downstream of gas lift choke (T-JUS-CKGL).

The data were collected from 21 different wells, with the oldest event occurring in April 2012 and the most recent in June 2018. The data are divided into folders according to the event type, and there are a total of 1,984 instances with a total of 15,872 variables. There are two types of instance labeling. Each instance was labeled with a single code representing one of the adverse events or normal operation, a total of nine different codes. The second labeling was done at the observation level, so that each instance has up to three periods: normal, faulty transient, and faulty steady state. Only the following units were used: Pascal [Pa], standard cubic meters per second [m^3/s], and degrees Celsius [°C]. The source of each instance was specified in the name of the file. All instances were obtained at a fixed sampling rate of 1 Hz. There are several difficulties with the actual data described in the documentation: missing variables (31.17%), frozen variables (9.67%) and unlabeled observations (0.01%). Some events are less frequent than others, so for some adverse events, most of the available data are simulated and hand-drawn events. Although there are difficulties with the data, the dataset is not degraded to a lower class due to the large amount of high quality data and the well-written documentation. Rather, this dataset stands out among all others as the only one that has labeled transition periods that facilitate the implementation of PdM by using domain knowledge specific to each type of failure.

H. VALHALL OP

The data were collected from a single compressor on the Aker BP's oil platform (OP), located in the North Sea in the Valhall field [10]. The dataset includes time series data, maintenance history, process and instrumentation diagrams for Valhall's first (of four) stage natural gas compressor and associated process equipment: a first stage suction cooler, a first stage suction scrubber, and first stage discharge coolers. The gas compressor is used to compress and treat the gas to meet the required export pressure and specifications. Only the first stage compressor was selected because it is a subsystem with clearly defined boundaries.

The dataset is available as part of the Open Industrial Data Project with the goal of analyzing changes in the provided time series data due to maintenance history. The project is the result of a collaboration between Aker BP, one of Europe's largest independent oil companies, and Cognite, a Nordic software company. A live data stream is provided on a subscription basis and is free of charge. Full access is available to registered users via Cognite's Asset Data Insight, a web-based visualization tool for analyzing, monitoring and planning data, and an API key. SDK installation is required to retrieve data from multiple data sources and make it available as one complete dataset. The terms of use are specified in a document available upon registration. Depending on the variable and time interval selected, some values may be missing. In reviewing the data, it was found that the majority of the data prior to 2013 has significant gaps. Some examples of time series that can be accessed with Asset Data Insight from Cognite are valve position, valve temperature, compressor suction pressure, compressor discharge flow, and motor vibrations.

I. NAVAL PROPULSION PLANT (1 & 2)

These two datasets include data necessary for applying predictive maintenance for naval propulsion systems, specifically to gas turbines. The data were simulated using a numerical simulator of a naval vessel (frigate) in Simulink/MATLAB.

The first data edition is characterized by a gas turbine (GT) propulsion plant [38]. The various blocks that make up the complete simulator (propeller, hull, GT, gear box, and controller) have been developed and fine-tuned on several similar real propulsion plants in their previous works [70], [71]. Measurements of 16 features that indirectly represent the state of the system subjected to performance decay have been acquired and stored in the dataset through the parameter space. Some of these features are lever position, ship speed, GT shaft torque, GT and gas generator rate of revolutions, GT compressor inlet and outlet air temperature and pressure, and fuel flow. In addition, the degradation coefficients of the compressor and turbine are also calculated. Each possible degradation state can then be described by a combination of the compressor degradation coefficient, the turbine degradation coefficient, and the ship's speed (which is a linear function of lever position). The range of compressor and turbine degradation was sampled with a uniform grid of 0.001 precision to achieve good granularity of representation. The ship speed was investigated by sampling the range of possible speed from 3 knots to 27 knots with a granularity of representation of 3 knots.

The second dataset edition is characterized by a COmbined Diesel ELectric And Gas (CODLAG) propulsion plant [39]. The blocks describing the behavior of the main components of the system are the GT, the GT compressor, the hull and the propeller. Each entry contains a 25-feature vector and additional five degradation coefficients: propeller thrust and torque decay state coefficients, GT compressor and turbine state coefficients, and hull decay state coefficient. Both datasets are simulated and there are no missing values. The stated licence for the second edition is CC BY-NC, meaning that any commercial use is prohibited.

J. 3500-DEFAULT

The objective of this dataset is to diagnose diesel engine faults and support predictive maintenance, which was achieved

by analyzing the variation of cylinder pressure curves and crankshaft torsional vibration response [40]. The engine chosen as a case study is the MWM Acteon 6.12TCE diesel engine with a four-stroke cycle. The database includes a total of 3500 different fault scenarios for 4 different operating conditions: normal (no fault, 250 scenarios), pressure reduction in the intake manifold (250 scenarios), compression ratio reduction in the cylinders (1500 scenarios), and reduction in the amount of fuel injected into the cylinders (1500 scenarios). The dataset is simulated and does not contain missing values. In all scenarios, the engine rotation frequency was set at 2500 RPM because it had the lowest joint error rate in the estimation of the mean and maximum pressures of the combustion cycle between the experimental data (according to the data provided by the manufacturer) and the simulated data during the validation phase of the thermodynamic and dynamic models. However, this high rotation frequency is not characteristic of naval ships

The dataset has 97 variables. The first 84 columns correspond to a feature vector. The last 13 columns refer to the severity (up to 50%) of the engine operating variables. The adopted feature vector was selected from the thermodynamic model and obtained from the processing of signals such as pressure and temperature inside the cylinder and the torsional vibration of the engine flywheel. The vector was created by estimating the mean and maximum pressure values from the six pressure cylinder signals (12 variables) and obtaining spectral information from the torsional vibration curves (72 variables). The cylinder and vibration signals were simulated at a sampling frequency of 15 kHz for 1.008 s, giving a total of 15120 samples for each channel signal. The spectral variables include the first 24 harmonics (the first 24 half orders of the engine) of torsional spectrum frequency, amplitude, and phase.

K. EDP - INSIDE A BOILER

The EDP Boiler Dataset contains three years of data recorded at two boilers (X and Y) used in a thermal power plant as part of an EDP challenge to predict the onset of slag formation [41]. Therefore, previously detected slagging events with different corresponding intensity levels are included. There is a minimum number of missing values. Years are anonymized (e.g., ''xxx0'' is the first year of data, ''xxx1'' is the second year of data, etc.). This dataset is only available to registered users of the EDP Open Data Platform.

The dataset includes 141 variables, the description of which is included in the dataset documentation. Some of the variables are boiler furnace pressure, drum temperature, reheated steam temperature at inlet and outlet, main steam pressure, and so on. The documentation states that the data were recorded every minute. However, the records indicate a sampling frequency of 5 minutes. The task of the EDP challenge was to successfully predict the next slagging event. A total of 2 slagging events are labeled. The EDP challenge is scored based on the total predicted savings and costs caused by true and false positive predictions. Savings from true positive predictions can be as high as 350,000 €, depending on how early the slagging event was predicted.

L. SIMULATED BOILER FAULT DATA

The dataset consists of data simulated for the Viessmann Vitorond 200 Gas Fired Boiler VD2 Series 380 using Simulink/MATLAB based on the Simscape boiler model [42]. The purpose of the dataset is to evaluate three types of faulty states under varying conditions. The data consist of five simulated variables:

1) fuel flow rate [kg/s],
2) ambient air condition [K],
3) heating hot water return temperature [K],
4) heating hot water supply temperature [K],
5) boiler loop flow rate [kg/s].

The variables are usually monitored by building automation systems (BAS). The model was validated by replicating the test conditions of *ANSI/AHRI Standard 1500 - Performance Rating of Commercial Space Heating Boilers*, comparing the outputs with published manufacturer data. The data were split into one normal state and three faulty states: excess air (15-50%), fouling of the heat exchanger (1-46%), and scaling of the water-side heat exchanger element (1-46%). All faults were simulated with a step size of 5%. Different iterations were performed by changing the gas fuel rate (1-4 kg/s), water mass flow rate (3-12.5 kg/s), and combustion temperature (283-303 K). A total of 27,281 simulations were performed using a factorial sampling method. Since the dataset was simulated, there are no inconsistencies or missing data. An *IEEE DataPort* subscription is required for full access to the dataset.

M. TRANSMISSION LINE FAULTS

Transmission lines are the most important part of the power grid. In this dataset, fault detection on transmission lines is studied because quick detection and classification of faults can help keep the power grid stable [43]. The complete dataset is located on Kaggle, under the title Electrical Fault Detection and Classification. The dataset comes from measurements made on a power system simulated with Simulink/MATLAB. The system consists of two 400 kV generators located at each end of the transmission line. The length of the transmission line is 300 km and the model was simulated for various types of faults at different locations along the transmission line length with different values of the fault resistance.

There are two datasets present, each having six input variables: three voltages of the respective three phases and three currents of the respective three phases. Both are normalized with respect to the pre-fault values of the voltages and currents, respectively. The first dataset, *detect_dataset.csv*, deals with the fault detection problem and contains 12,001 instances. All instances are binary labeled, with 0 representing the *No-fault* state and 1 representing the *Fault is present* state. The second dataset, *classData.csv*, addresses the problem of classifying the type of fault. Instead of one, there are four output variables, each corresponding to the

fault condition of each of the three phases, and one output for the ground line. The output is either 0 or 1 and represents the absence or presence of a fault on the corresponding line A, B, C, or G (where A, B, and C represent the respective three phases of the transmission system and G represents ground). There are ten possible fault types, but only five of them are present in the dataset, along with a no-fault state. This dataset contains 7861 instances. Both datasets contain 5,496 fault events. Since the dataset was simulated, there are no inconsistencies or missing data.

N. TRANSFORMER AND PAR TRANSIENTS

This dataset consists of 3-phase differential currents from internal faults and six other transient cases in a 5-bus interconnected system for phase angle controllers (PAR) and power transformers [44]. PARs are a special class of transformers used to control active power flow in parallel transmission lines. In systems using parallel transmission lines, detecting and classifying the type and location of faults is important to enable a timely reaction and contain the failure locally. These types of faults cannot be predicted in advance, so this dataset can be used to implement supervisory system control. The dataset was created using PSCAD/EMTDC software to simulate power transformers and indirect balanced phase angle regulators (ISPAR) with the same voltages at the transmit and receive ends and with two transformer units. By varying different system parameters, 100,908 transient cases are simulated. The simulation was performed for three types of internal faults and six types of transient disturbances. The internal faults are:

1) power transformer internal faults (36720 files),
2) ISPAR series transformer internal faults (36720 files),
3) ISPAR exciting transformer internal faults (14688 files),

where each has 11 types of faults defined (e.g., phase A to Ground, Phase AB to Ground, Phase AB,…) together with turn-turn fault and winding-winding fault case. The six types of transient disturbances are:

1) capacitor switching: 180 files,
2) external faults with current transformer (CT) saturation: 7920 files,
3) ferroresonance: 720 files,
4) magnetizing inrush: 1800 files,
5) non-linear load switching: 360 files,
6) sympathetic inrush: 1800 files.

Each text file has 726 rows and 4 columns: time, phase A of the differential current, phase B of the differential current, and phase C of the differential current. In each text file, time starts at 0.05 s and ends at 0.1225 s. The internal faults and transients occur 15.0 s after the start of each simulation case for the internal faults and transients. In the text files, there are 3-phase differential current samples from 15.0 s (0.05 s) to 15.0725 s (0.1225 s), forming 726 rows. An *IEEE DataPort* subscription is required for full access to the dataset.

O. TRANSIENTS IN ISPARS

This dataset consists of simulated data for internal faults and transient cases for ISPARs [45]. Similarly to the dataset Transformer and PAR transients, these types of faults cannot be predicted in advance, but classifying the type and location of faults in a timely manner is important to minimize the effect on the whole network. One possible implementation of this dataset is supervisory system control. The data were simulated using PSCAD/EMTDC software. The internal faults are simulated on the primary and secondary sides of the exciting and series units. They include the faults that occur inside the enclosure and at the locations of CTs. Basic internal faults include short circuits and phase faults, turn-turn faults, and winding-winding faults. Transients include magnetizing inrush, sympathetic inrush, external faults with CT saturation and overexcitation conditions.

The variable inputs to the simulations were the percentage of turns shorted, fault resistance, faulty unit, fault type, fault inception time, phase shift: forward & backward, and PAR tap positions. This resulted in a total of 60,552 fault cases, of which 46,872 were internal faults and 13,680 were transient faults. The dataset consists of 12 files. Files *fault_location* contain measurements of phase A, B, and C for the internal fault cases. Each row represents one cycle and consists of 167 samples. File *fault_location_target* contains information on the location of the fault (series or exciting unit). The files called *transients* contain measurements for phases A, B, and C for the transient fault cases. The type of transient fault is marked in the file *fault_transient_target*. Since the dataset was simulated, no inconsistencies or missing values are present. The dataset also includes files *fault_transient* for which no documentation is provided.

The total run time of the simulation is 10.2 s, the switching time is 10.0 s, and the duration of faults is 0.05 s (3 cycles). An *IEEE DataPort* subscription is required for full access to the data.

V. DATA-DRIVEN PIPELINE FOR PREDICTIVE MAINTENANCE

As mentioned earlier, data-driven approaches are the most common choice for achieving efficient PdM. A typical data-driven PdM pipeline for the reviewed datasets is shown in Fig. 3. Any data-driven method starts with the acquisition of data from a repository. The next step is data preprocessing, where the data are processed and transformed so that they can be efficiently processed by a statistical, ML or DL model. Common preprocessing activities include data transformation (normalization), data cleaning (removing noise, outliers, missing or frozen variables), and data reduction [72].

The following step is feature extraction. In some cases, the dataset consists of already extracted features (e.g., EDP-WT [31]) such as minimum, maximum, mean, and standard deviation of the signal. In other cases, the dataset contains time series coming from different sensors (e.g., 3W [32]). After feature extraction, a recommended step is feature selection, which is used to reduce the dimensionality of the dataset

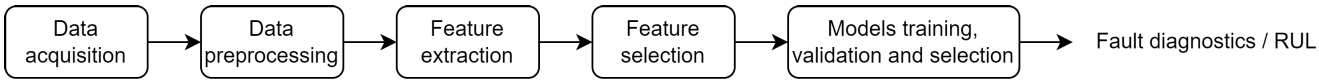

FIGURE 3. Common data-driven pipeline for PdM.

by selecting the most relevant features and eliminating the redundant and irrelevant ones. The final step is the training of one or more models (model development) and validation (evaluation of model performance). The best model is selected based on its performance metrics related to the goal. The goal is to predict/detect/classify faulty conditions or predict the RUL of a system or a machine part.

After getting the information about fault diagnosis or RUL, maintenance can be scheduled, and/or spare parts can be ordered. Scheduling maintenance according to these predictions reduces the number of unnecessary equipment checks, extends the lifetime of aging assets, and resolves equipment problems before faults occur, thus increasing the safety of the work environment [2], [46].

VI. DISCUSSION

The main challenge in this study was the small number of datasets with maintenance logs or recorded faults that are publicly available. Without maintenance logs or recorded faults, it is impossible to implement PdM because the existing approaches are mainly data-driven. Most of the datasets originating from the energy industry focus on predicting the generated power output, such as the Yalova WT dataset [73] and the La Houte Bourne Wind Farm [74], or the profitability of a plant, such as the Russian Oil Well dataset [75]. Some datasets were made with the intent of gathering information that can help address maintenance and energy management problems, such as the Sunlab Faro dataset [76], where labs were set up to test the performance and reliability of PV modules under different weather conditions and installation techniques, and the Container Vessel dataset [77], where data could be used to incorporate decision strategies to reduce human intervention to improve the shipboard energy management system in real time. However, none of these datasets contained any type of anomaly, fault, or maintenance logs, making them useless for the PdM problem.

The datasets included in this review can be divided into two groups depending on the source of the data: simulated or real-life. More than 50% of the datasets contained data simulated using MATLAB or collected in an artificial environment (e.g., an experimental setup). These datasets could be useful in obtaining a proof-of-concept, but depending on the setup, the results could vary significantly from real-life data. In Table 4, a comparison of real-life and simulated datasets is given in terms of advantages and disadvantages for PdM.

The listed advantages and disadvantages of datasets, together with other remarks on good practices, are further discussed along several topics: quality of data, data labeling, data collection, and evaluating PdM solutions. Afterwards, the results of the new grading system are analyzed and guidelines for creating new PdM datasets are given. In the end, we discuss the direction of future research.

TABLE 4. Comparison of real-life and simulated datasets.

Real-life dataset	Simulated dataset
Advantages	
• real measured data	• balanced dataset • varying conditions of faults
Disadvantages	
• small number of faults • imbalanced dataset • data need preprocessing • inconsistent data labeling	• simulation does not adequately describe the real system • small number of variables

A. QUALITY OF DATA

The main concern when analyzing datasets is data quality. When working with real-life datasets, many problems can occur during the process of measuring and recording data, such as broken sensors or inconsistent labeling. Most real-life datasets need to go through preprocessing (removing noise, missing or frozen variables, and outliers), which is not the case with simulated datasets.

Another point in data quality is the dataset size, such as the number of events, number of variables, and number of instances. When applying PdM, it is important to have a good number of faulty events that are the focus of observation. The problem is that real-life datasets mostly contain a small number of fault events, due to the nature of machines that rarely fail. For example, the EDP - Inside a Boiler dataset [41] contains only 2 faulty events, whereas the EDP-WT [31] contains 28 of them. This problem also leads to imbalanced datasets, where the simulated datasets usually have an equal number of instances with faults and normal states. To address this problem, data would need to be collected from real sources over longer periods of time, not just months, but years. Such a long period of data collection presents several challenges, such as storage, speed of data analysis, and the cost of constantly updating and annotating the data.

Simulated datasets, aside from usually being balanced, also have the possibility to vary different conditions that can lead to a fault. In this way, the behaviour of the system can be studied more thoroughly, allowing the construction of a better PdM model. On the other hand, the main problem with simulated datasets is that they can never describe a real system completely accurately. When building a simulation model, the focus is usually on simulating specific parts of the system or processes, not the whole system. This is one of the reasons why most of the simulated datasets have a small number of variables, such as Transformer and PAR transients (4 variables) [44], Transients in ISPARs (4 variables) [45], and Wind Turbine PMSG (2 variables) [35].

Considering all the listed advantages and disadvantages, the 3W dataset [32] stands out in several ways. This dataset contained more faults (1387 instances) than other similar datasets, and it was also the only dataset that contained data from three different sources: real (428 instances), simulated (939 instances), and hand-drawn (20 instances). Nevertheless, some faulty events were more prevalent than others, which can lead to the aforementioned class imbalance problems in machine learning PdM models [78].

Depending on the quality of the data, some of the listed problems can render a dataset unusable in the worst case.

B. DATA LABELING

Almost all of the datasets included just two types of data labels: normal and faulty. The 3W dataset is unique in being the only dataset that includes not only data from a faulty state, but also labels of the transition periods for each type of fault. Knowing the transition period for different types of faults is important because it allows for the prediction of faults and the detection of anomalies before they occur. For many types of faults in different systems, there is a lack of domain knowledge that could indicate how early a faulty state can be detected. Another important problem is the time period over which a prediction of a faulty state is considered useful. For example, if a correct prediction is made minutes or hours rather than days or months before a fault event occurs, does that make a difference in terms of reducing maintenance costs? In the EDP-WT Dataset Challenge [31], the authors set a fixed transition period of 60 days before a fault event. If the fault event was correctly predicted between 2 and 60 days before the fault event, it was considered a true positive, meaning savings were achieved. The problem with a fixed transition period is that the dataset includes many types of faults occurring on five different components (gearbox, generator, bearings, transformer, hydraulic group), which means that each fault is unique and should be treated differently.

C. DATA COLLECTION

The data for all real-life datasets were gathered using a variety of sensors integrated into the systems under consideration. The most common sensors measured temperature, pressure, current, voltage, and vibration. We note that although all of the sensors can be integrated into a system in an IoT setting, the datasets considered do not include information about the protocols used. Among the non-simulated datasets, the Valhall OP dataset [10] stands out as the only dataset whose data are being directly live-streamed via Open Industrial Data (OID) project. Free access to the data is the result of a collaboration between Aker BP and Cognite, whose goal is to accelerate innovation in data-intensive fields. On the other hand, three out of four datasets in the wind subsector use data gathered from SCADA systems. If we compare IoT and SCADA systems, we could say that IoT is a natural extension and evolution of SCADA, with one of the common concepts being machine-to-machine (M2M) communication [79].

D. EVALUATION OF PdM SOLUTIONS

To evaluate ML models used for PdM in the obtained dataset cases, authors typically use common metrics such as accuracy, precision, and recall [52], [55], [58], [80]. The EDP-WT dataset [31] and EDP - Inside a Boiler dataset [41] stand out as the only datasets that include costs for replacement, repair, and inspection for each component. The authors also included a formula for calculating total prediction savings to evaluate solutions for PdM. In this way, the benefits of using PdM models are clearly visible [48], [81].

E. GRADING RESULTS AND GUIDELINES FOR NEW PdM DATASETS

Finally, after analyzing all criteria, only three datasets in Table 2 were graded as class I (EDP-WT, OREC, and 3W). In contrast, two datasets were graded as class III because of various problems, such as missing documentation or license information, or a relatively small number of instances or faulty events. These problems make it difficult to implement PdM. Among the datasets that were graded as class II, many are of good quality and can be used for PdM, but the datasets are simulated.

When creating a new dataset for the PdM use case, the focus should be on getting real-life data, because simulated datasets can never accurately describe the real system. The most important point is recording a larger number of fault events, which can be done by gathering data for a longer period of time on one system or gathering data from multiple instances of the same system. The data should be preprocessed (frozen or missing variables removed) and the data labeling should be done consistently. Documentation describing variables, types of faults, and system components must be provided. If possible, it is recommended to include the cost of specific faults or maintenance, to better evaluate different PdM solutions. Expert knowledge of specific faults is encouraged and can be included in data labeling, such as adding labels for transition states, not labeling only faulty or normal states.

F. FUTURE RESEARCH

Some of the datasets obtained through search, but not elaborated in this review, were datasets that did not include time series, instead, they consisted of images. Those datasets are: PV cell anomaly detection dataset [82] that contains infrared images of PV cells with different types of anomalies, Vibration time-frequency images of wind turbine planetary gearboxes [83] that contains a total of 160 vibration time-frequency maps, and Frequency occurrence plots for motor fault diagnosis based on image recognition [84] that has 150 three-second sampling motor current signals.

In addition to the inspected PdM datasets that are specific to the energy sector, there are also PdM datasets that are for general use, meaning the focus of the dataset is on a specific part of a machine that is widely used in different types of industries, such as Intelligent bearing fault diagnosis dataset [85], Gearbox Fault Diagnosis [86], Composed fault

dataset (COMFAULDA) [87], and Microsoft Azure Predictive Maintenance dataset [88]. While the focus of this work was on energy sector datasets, for further research, some of these datasets could also be considered.

VII. CONCLUSION

Predictive maintenance is an important part of Industry 4.0 and has the potential to improve maintenance processes and reduce costs and environmental impact in the energy sector. One of the keys to applying PdM to the energy industry is the availability of high-quality datasets that would allow researchers to build models with broader applicability. In this paper, existing datasets and their properties were examined, their advantages and limitations were highlighted, and an objective grading was proposed. A total of 15 datasets were included and described for five subsectors of the energy sector: wind, solar, oil & gas, diesel & thermal, and electrical power.

Less than half of the datasets received had data that came from real sources. The datasets included data collected by a variety of sensors, with temperature, pressure, current, and voltage being the most common. The Valhall OP dataset [10] went a step further than other datasets in that it provided a live stream of data that was freely available. Whereas non-simulated datasets are the most valuable, they often contain a small number of faults because machines rarely experience them. Unlike other datasets that use two labels for data: normal and faulty, the 3W dataset [32] stands out as the only one that introduces a new label, the transition period. Knowing the transition period of a fault can help understand how early a fault can be detected.

A grading system was devised to evaluate the quality of each dataset. According to the criteria, two datasets were graded as class III, ten datasets were graded as class II, and only three datasets (EDP-WT, OREC, and 3W) were graded as class I, standing out as highly valuable for PdM research in the energy sector.

From our review of the field, it can be concluded that many more high quality datasets need to be made available to achieve a wider dissemination of effective predictive maintenance models. Future work will include exploration of other datasets mentioned earlier in the discussion that either contain PdM data that are not specific to the energy sector or contain other data formats, such as images instead of time series.

ACKNOWLEDGMENT

The authors are grateful to Igor Stancin, Eugen Vusak, Bojana Dalbelo Basic, and Vladimir Spisic for their support and discussions during the preparation of the manuscript.

REFERENCES

[1] H. Wang, W. Zhang, D. Yang, and Y. Xiang, "Deep-learning-enabled predictive maintenance in industrial Internet of Things: Methods, applications, and challenges," *IEEE Syst. J.*, vol. 17, no. 2, pp. 2602–2615, Jun. 2023.
[2] PWC. (Sep. 2018). *Predictive Maintenance 4.0, Beyond the Hype: PdM 4.0 Delivers Results*. Accessed: Jun. 7, 2023. [Online]. Available: https://www.pwc.be/en/documents/20180926-pdm40-beyond-the-hype-report.pdf
[3] T. Zonta, C. A. da Costa, R. da Rosa Righi, M. J. de Lima, E. S. da Trindade, and G. P. Li, "Predictive maintenance in the industry 4.0: A systematic literature review," *Comput. Ind. Eng.*, vol. 150, Dec. 2020, Art. no. 106889.
[4] V. Rastogi, S. Srivastava, M. Mishra, and R. Thukral, "Predictive maintenance for SME in industry 4.0," in *Proc. Global Smart Ind. Conf. (GloSIC)*, Nov. 2020, pp. 382–390.
[5] S. K. Jagatheesaperumal, M. Rahouti, K. Ahmad, A. Al-Fuqaha, and M. Guizani, "The duo of artificial intelligence and big data for industry 4.0: Applications, techniques, challenges, and future research directions," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 12861–12885, Aug. 2022.
[6] L. Krishnamurthy, R. Adler, P. Buonadonna, J. Chhabra, M. Flanigan, N. Kushalnagar, L. Nachman, and M. Yarvis, "Design and deployment of industrial sensor networks: Experiences from a semiconductor plant and the North Sea," in *Proc. 3rd Int. Conf. Embedded Networked Sensor Syst.*, New York, NY, USA, Nov. 2005, pp. 64–75.
[7] J. Eriksson, F. Österlind, N. Finne, N. Tsiftes, A. Dunkels, T. Voigt, R. Sauter, and P. J. Marrón, "COOJA/MSPSim: Interoperability testing for wireless sensor networks," in *Proc. 2nd Int. ICST Conf. Simulation Tools Techn.*, 2009, pp. 1–7, Art. no. 27.
[8] X. Xu, T. Chen, and M. Minami, "Intelligent fault prediction system based on Internet of Things," *Comput. Math. Appl.*, vol. 64, no. 5, pp. 833–839, Sep. 2012.
[9] A. Kanawaday and A. Sane, "Machine learning for predictive maintenance of industrial machines using IoT sensor data," in *Proc. 8th IEEE Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Nov. 2017, pp. 87–90.
[10] AkerBP. (2022). *Open Industrial Data*. Accessed: May 19, 2022. [Online]. Available: https://openindustrialdata.com/data/
[11] R.-I. Chang, C.-Y. Lee, and Y.-H. Hung, "Cloud-based analytics module for predictive maintenance of the textile manufacturing process," *Appl. Sci.*, vol. 11, no. 21, p. 9945, Oct. 2021.
[12] J. Gama, R. P. Ribeiro, and B. Veloso, "Data-driven predictive maintenance," *IEEE Intell. Syst.*, vol. 37, no. 4, pp. 27–29, Jul. 2022.
[13] B. Veloso, J. Gama, R. P. Ribeiro, and P. M. Pereira, "A benchmark dataset for predictive maintenance," 2022, *arXiv:2207.05466*.
[14] K. Sivalingam, M. Sepulveda, M. Spring, and P. Davies, "A review and methodology development for remaining useful life prediction of offshore fixed and floating wind turbine power converter with digital twin technology perspective," in *Proc. 2nd Int. Conf. Green Energy Appl. (ICGEA)*, Mar. 2018, pp. 197–204.
[15] J. I. Aizpurua, S. D. J. McArthur, B. G. Stewart, B. Lambert, J. G. Cross, and V. M. Catterson, "Adaptive power transformer lifetime predictions through machine learning and uncertainty modeling in nuclear power plants," *IEEE Trans. Ind. Electron.*, vol. 66, no. 6, pp. 4726–4737, Jun. 2019.
[16] T. Han, C. Liu, W. Yang, and D. Jiang, "Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application," *ISA Trans.*, vol. 97, pp. 269–281, Feb. 2020.
[17] W. Zhang, D. Yang, and H. Wang, "Data-driven methods for predictive maintenance of industrial equipment: A survey," *IEEE Syst. J.*, vol. 13, no. 3, pp. 2213–2227, Sep. 2019.
[18] O. Merkt, "On the use of predictive models for improving the quality of industrial maintenance: An analytical literature review of maintenance strategies," in *Proc. Federated Conf. Comput. Sci. Inf. Syst. (FedCSIS)*, Sep. 2019, pp. 693–704.
[19] Z. J. Khan, "Predictive maintenance and Internet of Things," in *Proc. Int. Conf. Comput., Electron. Electr. Eng.*, Oct. 2021, pp. 1–5.
[20] M. Compare, P. Baraldi, and E. Zio, "Challenges to IoT-enabled predictive maintenance for industry 4.0," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4585–4597, May 2020.
[21] A. Prajapati, R. Arno, N. Dowling, and W. Moylan, "Enhancing reliability of power systems through IIoT—Survey and proposal," in *Proc. IEEE/IAS 55th Ind. Commercial Power Syst. Tech. Conf.*, May 2019, pp. 1–7.
[22] A. A. Manjare and B. G. Patil, "A review: Condition based techniques and predictive maintenance for motor," in *Proc. Int. Conf. Artif. Intell. Smart Syst. (ICAIS)*, Mar. 2021, pp. 807–813.
[23] N. A. R. Maciejewski, A. E. Treml, and R. A. Flauzino, "A systematic review of fault detection and diagnosis methods for induction motors," in *Proc. FORTEI-Int. Conf. Electr. Eng. (FORTEI-ICEE)*, Sep. 2020, pp. 86–90.
[24] M. J. Kabir, A. M. T. Oo, and M. Rabbani, "A brief review on offshore wind turbine fault detection and recent development in condition monitoring based maintenance system," in *Proc. Australas. Universities Power Eng. Conf. (AUPEC)*, Sep. 2015, pp. 1–7.

[25] K. Alewine and W. Chen, "A review of electrical winding failures in wind turbine generators," in *Proc. Electr. Insul. Conf. (EIC)*, Jun. 2011, pp. 392–397.

[26] V. V. Shanbhag, T. J. J. Meyer, L. W. Caspers, and R. Schlanbusch, "Failure monitoring and predictive maintenance of hydraulic cylinder—State-of-the-art review," *IEEE/ASME Trans. Mechatronics*, vol. 26, no. 6, pp. 3087–3103, Dec. 2021.

[27] E. Babu, J. Francis, E. Thomas, R. Cherian, and S. S. Sunandhan, "Review on various signal processing techniques for predictive maintenance," in *Proc. 2nd Int. Conf. Power Electron. IoT Appl. Renew. Energy Control (PARC)*, Jan. 2022, pp. 1–8.

[28] Vanraj, D. Goyal, A. Saini, S. S. Dhami, and B. S. Pabla, "Intelligent predictive maintenance of dynamic systems using condition monitoring and signal processing techniques—A review," in *Proc. Int. Conf. Adv. Comput., Commun., Autom. (ICACCA)*, Apr. 2016, pp. 1–6.

[29] P. R and P. Ramachandran, "An overview of predictive maintenance for industrial machine using vibration analysis," in *Proc. Innov. Power Adv. Comput. Technol. (i-PACT)*, Nov. 2021, pp. 1–7.

[30] D. Menezes, M. Mendes, J. A. Almeida, and T. Farinha, "Wind farm and resource datasets: A comprehensive survey and overview," *Energies*, vol. 13, no. 18, p. 4702, Sep. 2020.

[31] EDP Group. (2018). *Wind Turbine Failure Detection*. Accessed: Apr. 24, 2022. [Online]. Available: https://opendata.edp.com/explore/?refine.keyword=Wind+Challenge&sort=modified

[32] R. E. V. Vargas, C. J. Munaro, P. M. Ciarelli, A. G. Medeiros, B. G. D. Amaral, D. C. Barrionuevo, J. C. D. D. Araújo, J. L. Ribeiro, and L. P. Magalhães, "A realistic and public dataset with rare undesirable real events in oil wells," *J. Petroleum Sci. Eng.*, vol. 181, Oct. 2019, Art. no. 106223.

[33] W. Soontronchai. (Oct. 19, 2019). *IIOT Data of Wind Turbine*. Accessed: May 15, 2022. [Online]. Available: https://www.kaggle.com/datasets/wasuratme96/iiot-data-of-wind-turbine

[34] ORE Catapult. (2022). *OREC Data Collection*. Accessed: Apr. 25, 2022. [Online]. Available: https://pod.ore.catapult.org.uk/data-collections

[35] B. A. Sá. (Apr. 17, 2020). *Wind Turbine PMSG—Short-Circuit Fault*. Accessed: Sep. 10, 2022. [Online]. Available: https://www.kaggle.com/datasets/brunoadonis/wind-turbine-pmsg-short-circuit-fault-mcsa

[36] S. S. M. Ghoneim, A. E. Rashed, and N. I. Elkalashy, "Fault detection algorithms for achieving service continuity in photovoltaic farms," *Intell. Autom. Soft Comput.*, vol. 29, no. 3, pp. 467–479, 2021.

[37] A. Bakdi, A. Guichi, S. Mekhilef, and W. Bounoua. (2020). *GPVS-Faults: Experimental Data for Fault Scenarios in Grid-Connected PV Systems Under MPPT and IPPT Modes*. [Online]. Available: https://data.mendeley.com/datasets/n76t439f65/1

[38] A. Coraddu, L. Oneto, A. Ghio, S. Savio, D. Anguita, and M. Figari, "Machine learning approaches for improving condition-based maintenance of naval propulsion plants," *Proc. Inst. Mech. Eng., M, J. Eng. Maritime Environ.*, vol. 230, no. 1, pp. 136–153, Feb. 2016.

[39] A. Coraddu, L. Oneto, F. Cipollini, and D. Anguita. (Jan. 17, 2017). *Hull, Propeller and Gas Turbine Efficiency Decay: Data Analysis With Minimal Feedback*. [Online]. Available: https://pureportal.strath.ac.uk/en/datasets/hull-propeller-and-gas-turbine-efficiency-decay-data-analysis-wit

[40] D. Pestana, "Diesel engine faults features dataset (3500-default)," Mendeley Data, V1, 2020. Accessed: Jul. 17, 2023. [Online]. Available: https://data.mendeley.com/datasets/k22zxz29kr/1, doi: 10.17632/k22zxz29kr.1.

[41] EDP Group. (2020). *Inside a Boiler—Slagging Prediction*. Accessed: Apr. 24, 2022. [Online]. Available: https://opendata.edp.com/explore/?refine.keyword=Boiler&sort=modified

[42] R. Shohet, M. Kandil, and J. McArthur, "Simulated boiler data for fault detection and classification," IEEE DataPort, 2019. Accessed: Jul. 17, 2023. [Online]. Available: https://ieee-dataport.org/open-access/simulated-boiler-data-fault-detection-and-classification, doi: 10.21227/awav-bn36.

[43] E. S. Prakash. (May 26, 2021). *Electrical Fault Detection and Classification*. Accessed: Sep. 15, 2022. [Online]. Available: https://www.kaggle.com/datasets/esathyaprakash/electrical-fault-detection-and-classification

[44] P. K. Bera, C. Isik, and V. Kumar, "Transients and faults in power transformers and phase angle regulators (DATASET)," IEEE DataPort, 2020. Accessed: Jul. 17, 2023. [Online]. Available: https://ieee-dataport.org/open-access/transients-and-faults-power-transformers-and-phase-angle-regulators-dataset, doi: 10.21227/1d1w-q940.

[45] P. Bera and C. Isik, "Data: Transients in indirect symmetrical phase shift transformers," IEEE DataPort, 2020. Accessed: Jul. 17, 2023. [Online]. Available: https://ieee-dataport.org/open-access/data-transients-indirect-symmetrical-phase-shift-transformers, doi: 10.21227/d8fv-6257.

[46] M. Garan, K. Tidriri, and I. Kovalenko, "A data-centric machine learning methodology: Application on predictive maintenance of wind turbines," *Energies*, vol. 15, no. 3, p. 826, Jan. 2022.

[47] W. Udo and Y. Muhammad, "Data-driven predictive maintenance of wind turbine based on SCADA data," *IEEE Access*, vol. 9, pp. 162370–162388, 2021.

[48] E. Latiffianti, S. Sheng, and Y. Ding, "Wind turbine gearbox failure detection through cumulative sum of multivariate time series data," *Frontiers Energy Res.*, vol. 10, pp. 1–12, May 2022.

[49] K. Tidriri, A. Braydi, and H. Kazmi, "Data-driven decision-making methodology for prognostic and health management of wind turbines," in *Proc. Austral. New Zealand Control Conf. (ANZCC)*, Nov. 2021, pp. 104–109.

[50] F. P. G. de Sá, D. N. Brandão, E. Ogasawara, R. D. C. Coutinho, and R. F. Toso, "Wind turbine fault detection: A semi-supervised learning approach with automatic evolutionary feature selection," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Jul. 2020, pp. 323–328.

[51] J. Chatterjee and N. Dethlefs, "Deep learning with knowledge transfer for explainable anomaly prediction in wind turbines," *Wind Energy*, vol. 23, no. 8, pp. 1693–1710, Aug. 2020.

[52] J. Chatterjee and N. Dethlefs, "Temporal causal inference in wind turbine SCADA data using deep learning for explainable AI," *J. Phys., Conf. Ser.*, vol. 1618, Sep. 2020, Art. no. 022022.

[53] A. Bakdi, W. Bounoua, A. Guichi, and S. Mekhilef, "Real-time fault detection in PV systems under MPPT using PMU and high-frequency multi-sensor data through online PCA-KDE-based multivariate KL divergence," *Int. J. Electr. Power Energy Syst.*, vol. 125, Feb. 2021, Art. no. 106457.

[54] S. Wali and I. Khan, "Explainable signature-based machine learning approach for identification of faults in grid-connected photovoltaic systems," in *Proc. IEEE Texas Power Energy Conf. (TPEC)*, Feb. 2022, pp. 1–6.

[55] M. A. Marins, B. D. Barros, I. H. Santos, D. C. Barrionuevo, R. E. V. Vargas, T. de M. Prego, A. A. de Lima, M. L. R. de Campos, E. A. B. da Silva, and S. L. Netto, "Fault detection and classification in oil wells and production/service lines using random forest," *J. Petroleum Sci. Eng.*, vol. 197, Feb. 2021, Art. no. 107879.

[56] Y. Li, T. Ge, and C. Chen, "Data stream event prediction based on timing knowledge and state transitions," *Proc. VLDB Endowment*, vol. 13, no. 10, pp. 1779–1792, Mar. 2021.

[57] A. Soriano-Vargas, R. Werneck, R. Moura, P. M. Júnior, R. Prates, M. Castro, M. Gonçalves, M. Hossain, M. Zampieri, A. Ferreira, A. Davólio, B. Hamann, D. J. Schiozer, and A. Rocha, "A visual analytics approach to anomaly detection in hydrocarbon reservoir time series data," *J. Petroleum Sci. Eng.*, vol. 206, Nov. 2021, Art. no. 108988.

[58] E. M. Turan and J. Jäschke, "Classification of undesirable events in oil well operation," in *Proc. 23rd Int. Conf. Process Control (PC)*, Jun. 2021, pp. 157–162.

[59] B. G. Carvalho, R. E. Vaz Vargas, R. M. Salgado, C. J. Munaro, and F. M. Varejão, "Flow instability detection in offshore oil wells with multivariate time series machine learning classifiers," in *Proc. IEEE 30th Int. Symp. Ind. Electron. (ISIE)*, Jun. 2021, pp. 1–6.

[60] A. Coraddu, L. Oneto, A. Ghio, S. Savio, M. Figari, and D. Anguita, "Machine learning for wear forecasting of naval assets for condition-based maintenance applications," in *Proc. Int. Conf. Electr. Syst. Aircr., Railway, Ship Propuls. Road Vehicles (ESARS)*, Mar. 2015, pp. 1–5.

[61] F. Cipollini, L. Oneto, A. Coraddu, A. J. Murphy, and D. Anguita, "Condition-based maintenance of naval propulsion systems with supervised data analysis," *Ocean Eng.*, vol. 149, pp. 268–278, Feb. 2018.

[62] F. Cipollini, L. Oneto, A. Coraddu, A. J. Murphy, and D. Anguita, "Condition-based maintenance of naval propulsion systems: Data analysis with minimal feedback," *Rel. Eng. Syst. Saf.*, vol. 177, pp. 12–23, Sep. 2018.

[63] Y. Tan, C. Niu, H. Tian, Y. Lin, and J. Zhang, "Decay detection of a marine gas turbine with contaminated data based on isolation forest approach," *Ships Offshore Struct.*, vol. 16, no. 5, pp. 546–556, May 2021.

[64] R. Shohet, M. S. Kandil, Y. Wang, and J. J. McArthur, "Fault detection for non-condensing boilers using simulated building automation system sensor data," *Adv. Eng. Informat.*, vol. 46, Oct. 2020, Art. no. 101176.

[65] M. Jamil, S. K. Sharma, and R. Singh, "Fault detection and classification in electrical power transmission system using artificial neural network," *SpringerPlus*, vol. 4, no. 1, pp. 1–13, Dec. 2015.

[66] P. K. Bera, R. Kumar, and C. Isik, "Identification of internal faults in indirect symmetrical phase shift transformers using ensemble learning," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2018, pp. 1–6.
[67] P. K. Bera, C. Isik, and V. Kumar, "Discrimination of internal faults and other transients in an interconnected system with power transformers and phase angle regulators," *IEEE Syst. J.*, vol. 15, no. 3, pp. 3450–3461, Sep. 2021.
[68] P. K. Bera and C. Isik, "A data mining based protection and classification of transients for two-core symmetric phase angle regulators," *IEEE Access*, vol. 9, pp. 72937–72948, 2021.
[69] B. A. Sá, C. M. V. Barros, C. A. Siebra, and L. S. Barros, "A multilayer perceptron-based approach for stator fault detection in permanent magnet wind generators," in *Proc. IEEE PES Innov. Smart Grid Technol. Conf.-Latin Amer.*, Sep. 2019, pp. 1–6.
[70] M. Altosole, G. Benvenuto, and U. Campora, "Numerical modelling of the engines governors of a codlag propulsion plant," in *Proc. Int. Conf. Mar. Sci. Technol.*, 2010, pp. 1–17.
[71] M. Altosole, G. Benvenuto, M. Figari, and U. Campora, "Real-time simulation of a COGAG naval ship propulsion system," *Proc. Inst. Mech. Eng., M, J. Eng. Maritime Environ.*, vol. 223, no. 1, pp. 47–62, Mar. 2009.
[72] T. P. Carvalho, F. A. A. M. N. Soares, R. Vita, R. D. P. Francisco, J. P. Basto, and S. G. S. Alcalá, "A systematic literature review of machine learning methods applied to predictive maintenance," *Comput. Ind. Eng.*, vol. 137, Nov. 2019, Art. no. 106024.
[73] H. D. Nguyen. (Mar. 1, 2021). *Yalova Wind Turbine Dataset*. Accessed: May 15, 2022. [Online]. Available: https://www.kaggle.com/code/winternguyen/wind-power-curve-modeling
[74] ENGIE Group. (Oct. 9, 2019). *La Houte Bourne Wind Farm*. Accessed: May 15, 2022. [Online]. Available: https://opendata-renewables.engie.com/explore/
[75] R. Zalevskikh. (Feb. 3, 2021). *Oil Well Operation Parameters (2013–2021)*. Accessed: May 15, 2022. [Online]. Available: https://www.kaggle.com/datasets/ruslanzalevskikh/oil-well
[76] ENGIE Group. (2018). *Sunlab Faro*. Accessed: Apr. 24, 2022. [Online]. Available: https://opendata.edp.com/explore/?refine.keyword=Sunlab&sort=modified
[77] C. Chin and T. Chan, "Container vessel data," IEEE DataPort, 2019. Accessed: Jul. 17, 2023. [Online]. Available: https://ieee-dataport.org/documents/container-vessel-data, doi: 10.21227/r1p7-6z94.
[78] S. Sridhar and S. Sanagavarapu, "Handling data imbalance in predictive maintenance for machines using SMOTE-based oversampling," in *Proc. 13th Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Sep. 2021, pp. 44–49.
[79] R. Hunzinger, "SCADA fundamentals and applications in the IoT," in *Internet of Things and Data Analytics Handbook*, H. Geng, Ed. Hoboken, NJ, USA: Wiley, 2017, pp. 283–293.
[80] D. Pestana-Viana, R. H. R. Gutiérrez, A. A. de Lima, F. L. E. Silva, L. Vaz, T. M. de Prego, and U. A. Monteiro, "Application of machine learning in diesel engines fault identification," in *Proc. 10th Int. Conf. Rotor Dyn.*, K. L. Cavalca and H. I. Weber, Eds. Cham, Switzerland: Springer, 2019, pp. 74–89.
[81] J. Eriksson, "Machine learning for predictive maintenance on wind turbines: Using SCADA data and the Apache Hadoop ecosystem," M.S. thesis, Linköping Univ., Linköping, Sweden, 2020.
[82] B. Su, Z. Zhou, and H. Chen, "Photovoltaic cell anomaly detection dataset," IEEE DataPort, 2022. Accessed: Jul. 17, 2023. [Online]. Available: https://ieee-dataport.org/documents/photovoltaic-cell-anomaly-detection-dataset, doi: 10.21227/pz6t-3s77.
[83] P. Zhang, "Vibration time-frequency images of planetary gearboxes," IEEE DataPort, 2022. Accessed: Jul. 17, 2023. [Online]. Available: https://ieee-dataport.org/documents/vibration-time-frequency-images-planetary-gearboxes, doi: 10.21227/0zxx-m405.
[84] E. J. Piedad, Y.-M. Hsueh, C.-C. Kuo, and H.-C. Chang, "Frequency occurrence plots for motor fault diagnosis based on image recognition," IEEE DataPort, 2019. Accessed: Jul. 17, 2023. [Online]. Available: https://ieee-dataport.org/documents/frequency-occurrence-plots-motor-fault-diagnosis-based-image-recognition, doi: 10.21227/77da-c563.
[85] S. N. Chegini, P. Amini, B. Ahmadi, A. Bagheri, and I. Amirmostofian, "Intelligent bearing fault diagnosis using swarm decomposition method and new hybrid particle swarm optimization algorithm," *Soft Comput.*, vol. 26, no. 3, pp. 1475–1497, Feb. 2022.
[86] B. R. Japon. (Feb. 28, 2021). *Gearbox Fault Diagnosis*. Accessed: Jun. 2, 2022. [Online]. Available: https://www.kaggle.com/datasets/brjapon/gearbox-fault-diagnosis
[87] D. Martins, D. Pestana-Viana, A. Lima, D. Hadadd, R. Homero, and L. Vaz, "Composed fault dataset (COMFAULDA)," IEEE DataPort, 2022. [Online]. Available: https://ieee-dataport.org/documents/composed-fault-dataset-comfaulda, doi: 10.21227/89ye-ap56.
[88] A. Biswas. (Oct. 15, 2020). *Microsoft Azure Predictive Maintenance*. Accessed: Jun. 2, 2022. [Online]. Available: https://www.kaggle.com/datasets/arnabbiswas1/microsoft-azure-predictive-maintenance

EDA JOVICIC (Member, IEEE) received the M.Sc. degree in electronic and computer engineering from the Faculty of Electrical Engineering and Computing (FER), University of Zagreb, Zagreb, Croatia, in 2021.

She is currently a Researcher with the FER, University of Zagreb. Her research interests include biomedical engineering and machine learning with applications.

DARIA PRIMORAC received the M.Sc. degree in computational physics from the University of Split, Split, Croatia, in 2015, and the Ph.D. degree in relativistic astrophysics from the Sapienza University of Rome, Rome, Italy, in 2020.

She was with the Research Institute in Astrophysics and Planetology, Toulouse, France, in 2014. She was with the Institute for Space Astrophysics and Planetology, INAF, Rome, from 2019 to 2021. She was a Researcher with the Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia, from 2021 to 2022. She is currently an Independent Researcher residing in Zurich, Switzerland. She has coauthored ten articles in international journals. Her current research interests include machine learning applications in predictive maintenance.

MARKO CUPIC (Member, IEEE) received the M.Sc. and Ph.D. degrees in computer science from the Faculty of Electrical Engineering and Computing (FER), University of Zagreb, Zagreb, Croatia, in 2006 and 2011, respectively.

Since 2003, he has been with FER, University of Zagreb, where he is currently an Associate Professor of computer science. He has authored or coauthored more than 30 refereed articles in international publications. His current research interests include computer supported education, computer graphics, computer games development, optimization algorithms, machine learning with applications, programming languages, and software engineering.

Dr. Cupic is a member of ACM.

ALAN JOVIC (Member, IEEE) received the Dipl.Ing. and Ph.D. degrees in computer science from the Faculty of Electrical Engineering and Computing (FER), University of Zagreb, Zagreb, Croatia, in 2006 and 2012, respectively.

From 2006 to 2007, he was an Expert Associate with the Rudjer Boskovic Institute, Zagreb. Since 2007, he has been with FER, University of Zagreb, where he is currently an Associate Professor of computer science. He has authored or coauthored more than 70 refereed articles in international publications. His research interests include data science, machine learning, biomedical engineering, and software engineering.

Dr. Jovic received several national and international awards and acknowledgements for his work. He is a member of EMBS and MIPRO.