Multiple Instance Learning With Random Forest for Event Logs Analysis and Predictive Maintenance in Ship Electric Propulsion System

Azzeddine Bakdi, Nicolay Bjørlo Kristensen, and Morten Stakkeland

***Abstract*—In this article, a novel weakly supervised machine learning approach is proposed for intelligent predictive maintenance (IPdM). It employs balanced random forest and multiple instance learning based on event logs from ships' electric propulsion systems. The objectives are predicting failure likelihood, time to failure, and explainable predictions to ensure timely crew intervention. The IPdM approach uncovers, then learns, and classifies *sequences of* events that represent early causes or symptoms to forecast imminent failures. In particular, this article contributes effective solutions to irregular, imbalanced, and unlabeled data issues where conventional methods become obsolete. First, the events occur at irregular intervals; they include alarms, warnings, and operational information collected across multiple units and control systems. Second, the datasets exhibit extreme imbalance due to few failures and multiple failure modes; this entails biased predictions. Third, the training datasets are weakly labeled; only the failure timestamp is known without any expert input on prior causes or early symptoms. Temporal random indexing is proposed to transform textual log messages into a numerical lower dimensional space where timeseries analyses are applicable. Balanced random-forest models are developed for unbiased classification and regression. The overall approach learns recursively the ungiven data labels while training the base learners. The IPdM approach is validated through millions of events of multithousand types collected from two years of seagoing vessels. It successfully forecasts actual propulsion failures and performs better when compared with contemporary methods.**

***Index Terms*—Alarm system, balanced random forest (B-RF), event-driven predictive maintenance, event logs, failure prediction, imbalance, inexact labeling, inverter, liquified natural gas (LNG) carriers, multiple instance learning (MIL), ship electric propulsion system, temporal random indexing (TRI), time to failure (t2f), weakly supervised learning.**

Manuscript received 8 November 2021; accepted 4 January 2022. Date of publication 19 January 2022; date of current version 9 September 2022. This work was supported by Norwegian Research Council Research-Based Innovation Center BigInsight under Project 237718. Paper no. TII-21-4968. *(Corresponding author: Azzeddine Bakdi.)*

Azzeddine Bakdi and Nicolay Bjørlo Kristensen are with the Department of Mathematics, University of Oslo, 0851 Oslo, Norway (e-mail: bkdaznsun@gmail.com; nicolay@dsl.no).

Morten Stakkeland is with the Department of Mathematics, University of Oslo, 0851 Oslo, Norway, and also with ABB AS, 1360 Fornebu, Norway (e-mail: mortsta@math.uio.no).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2022.3144177.

Digital Object Identifier 10.1109/TII.2022.3144177

I. Introduction

INTELLIGENT predictive maintenance (IPdM) has emerged among the most attractive applications of Industry 4.0. It exploits the industrial Internet of Things with advanced big data analytics and machine learning (ML) algorithms to ensure intelligent, reliable, and safe operations. In shipping 4.0, ship automation, autonomy, digitalization, safety, and maritime electrification are central objectives. IPdM, as proposed in this article, has the capability of reducing the number of unscheduled interruptions in the electrical propulsion systems (EPS); hence, it reduces downtime and improves the safety of critical operations. The rising levels of system automation and integration increase complexity and room for random nonaging-related failures, where the traditional maintenance strategies are becoming obsolete [1]. Moreover, failure forecasting becomes an important component for future autonomous vessels, allowing operators to optimally schedule maintenance operations and procurement of parts. Fortunately, the rapid digitalization is a key enabler of IPdM; yet, there are still open challenges that must be addressed in order to elevate intelligent IPdM capabilities beyond its hype cycle.

The past decades have witnessed the academic and industrial transformation of maintenance strategies from reactive and preventive maintenance (PvM) to predictive maintenance (PdM) strategies [2], [3]. However, the state-of-the-art PdM includes mainly fault detection [4] and diagnosis [5]. Advanced IPdM strategies involve condition-based maintenance [6], such as aging, prognostics of deterioration, and state-of-health estimation [7]. Such PdM algorithms are classified into four groups. Physics-based models, e.g., digital twins that are difficult to establish for large systems. Knowledge-based expert systems, including fuzzy logic [6], are simple and explainable but less accurate. Data-driven algorithms use statistical methods, e.g., PCA [5] and ML, such as SVM [8] and recurrent neural networks [9]. Their combinations form hybrid methods.

This article considers IPdM algorithms that can 1) forecast the failure in advance, 2) predict *time to failure* (t2f), and 3) give *explainable predictions* to enable fast intervention. First, emphasis should be put on multiunit systems IPdM [10], taking many dependencies into account. Such systems include modern ships and particularly liquified natural gas (LNG) carriers, up to 300 m long. Second, IPdM is crucial in modern ships, which form the vital blue economy, where over 90% of the worlds'

merchandise is transported by sea [11]. Shipping accidents cause catastrophic fatalities, economic losses, and damage to the environment [11]. Third, the abovementioned condition-based IPdM methods ignore volatile stochastic conditions, events, and operators' activities. They consequently miss-predict short-term failures that are related to random events instead of steady aging or deterioration. Fourth, the article extends the concept of maintenance beyond the traditional definition. The ultimate target is to improve safety during critical operations and avoid unscheduled interruptions in a complex system. Hence, the main goal is to perform timely prediction of failures in the EPS [1), 2), and 3)], allowing the crew to eventually perform mitigative actions, e.g., changing the operation. The needed maintenance may be as simple as restarting the system, although more traditional maintenance may be needed in some cases. This approach can potentially reduce the risk of midsea propulsion loss, as in the Viking Sky incident [12]. It provides sufficient time to *fail safe* and prevent navigational accidents, as in the case of Bulk India grounding [13].

This article, hence, focuses on event-driven IPdM, particularly in LNG carriers' EPSs. Numerous normal and abnormal events are recorded on board the vessels; these are usually trivial and unlikely to cause any harm. Yet, special combinations of events might propagate to other units, triggering patterns or *sequences of events* that cause a failure; e.g., {valve controller anomaly, cooling-water unit malfunction, generator blackout, propulsion failure, and grounding accident} in [13].

This work also provides novel comprehensive solutions to tackle three major IPdM challenges; this consequently contributes to increase IPdM accuracy, especially to attain objectives [1), 2), and 3)]. These targeted issues include the irregular event timestamps and inconsistent event messages, the extreme event-data imbalance due to few failure events and multiple failure modes, and more importantly, the lack of sufficient fully labeled data to train and test IPdM models.

The rest of this article is organized as follows. Section II provides the literature review on maritime IPdM practices, knowledge gaps, common challenges, and the contributed solutions. Section III provides an overview of the electric propulsion system, numerical figures on data quality and quantity issues, and the designed event-data processing techniques. Section IV derives and describes the implementation of balanced random forest (B-RF) and the overall weakly supervised IPdM approach. Section V discusses the results of real-world failures prediction, PdM explainability, and comparisons with closely related methods. Finally, Section VI concludes this article.

II. Literature Review

Sea-based equipment are frequently subject to both failures and defects as well as needless repairs. The current maritime maintenance culture involves three paradigms, which impose opposite advantages and limitations. Traditionally, the *reactive maintenance* paradigm is referred to as the "corrective" or "curative" approach and it is widely common in the maritime industry, as reviewed in [14]. *Ship's planned maintenance systems* are based on periodically scheduled replacements and repairs according to requirements established by the manufacturers and classification societies. It has been in use for a long period of time [15]. In *PvM*, the crew is responsible for keeping the equipment in good conditions through frequent inspections, conditional analyses, and preventive measures. This approach is active in the maritime industry, especially for mechanical systems [16].

Optimal maintenance strategies idealistically aim at three targets: preventing unplanned failures, subduing unnecessary maintenance-related costs and downtime, and exploiting the full lifespan of machinery. PdM aims at predicting when the system is likely to fail and determining the required maintenance actions. Prescriptive maintenance extends this scope to identify root causes and determine recommendations to reduce operational risks. In particular, it is worth mentioning the general lack of intelligent PdM methods in the maritime literature [17]. This article contributes here by developing PdM methods for the vital propulsion system.

Zonta *et al.* [3] reviewed the PdM state-of-the-art and highlighted multiple challenges that face applied PdM methods. Similarly, the survey in [18] considered PdM as "still in its infancy" and related this to the following three open issues:

1) multisource data quality;
2) identification/prediction accuracy;
3) the combination of artificial intelligence (AI) methods with maintenance scheduling.

With the rich and fast-evolving literature of statistics and ML, the limited accuracy 2) is primarily attributed to the data quantity and quality limitations 1). However, explainable AI (XAI) is essential for addressing IPdM issue 3) [19]. Motivated by these challenges, the presented article contributes to improving IPdM accuracy; unconventional ML methods are presented herein to cope with inevitable PdM data issues and to provide interpretable failure predictions.

Notwithstanding the imperative role of alarm systems in operation safety [20], the full exploitation of their massive logged events is insufficiently explored in the PdM literature. This shortage is first related to the event-data complexity. The events are textual in nature, have multiple attributes, originate from multiple equipment, occur irregularly, and exhibit parallel and sequential correlations. Second, the timeseries analyses literature is much richer compared with event- and time-to-event analyses [21]. Hence, very few event-driven PdM methods [22]–[24] are reported in the literature, and they are limited to single-unit applications. Random indexing (RI) is a text mining method [23] that is widely applied to natural language processing. Despite its success in processing textual data, the method neither incorporates the time dimension nor the time correlation between textual inputs. Given the significance of the latter properties in event logs analyses, temporal RI (TRI) is developed in this work. It contributes effective event-data processing, after which most of the advanced timeseries analyses are applicable.

A second major issue is related to imbalanced data, which poses a serious challenge for ML methods; it also leads to biased models and predictions. In general, data imbalance refers to

the skewed distribution of data points over existing classes; its common challenges are reviewed in [25]. Common remedies for this issue include weighted or cost-sensitive methods, oversampling minority classes, and undersampling majority classes. In IPdM context, the training data is extremely imbalanced, including both between-class and within-class imbalances. Failure events occur infrequently and imply insufficient training data examples. Besides, the failure-cause samples are extremely minor compared with the remaining samples. Furthermore, the failure-cause samples are scattered in multiple regions due to multiple failure modes and near-failure cases. Failure causes are difficult to learn and predict. The original random forest (RF) is selected as a base learner for many advantages. RF is useful for high-dimensional data; it is robust to noise features; it provides explainable predictions from interpretable features. It is also robust to small data imbalance. However, a balancing strategy is implemented to further improve RF performance against extreme imbalance.

The most challenging IPdM issue is, due to unlabeled training data [3], [18], a dominant requirement for predictive methods. Yet, the lack of sufficient labeled data—as highlighted in [24]—remains an unsolved issue in the PdM literature. Conventional ML methods are classified into three categories. Clustering techniques [26] are used for unsupervised learning from unlabeled data, but their applications are limited to finding similarities between samples. On the contrary, supervised learning requires fully labeled data to construct predictive models. Semisupervised learning [27] is based on labeled data for one representative class to train models that detect out-of-class samples as in change point and fault detection. This includes the basic RF applications for anomaly detection and classification [28], [29]. Recently, weakly supervised learning approaches [30] have emerged as potential solutions to construct predictive models that are highly desired in practice, yet, from leveraging inexactly and/or incompletely labeled data. However, these techniques remain unexploited in the PdM literature despite the strongly connected problem. This article proposes a weakly supervised ML approach for IPdM. The aim is to learn the ungiven data labels coordinately while tuning the parameters of the base learners. This forms a nonconvex optimization problem; it is solved as a multiple instance learning (MIL) process using deterministic annealing (DA).

III. Electric Propulsion and Event Logs

This article proposes a new event-driven approach for intelligent IPdM in EPSs of LNG carriers. Historical event datasets are collected from hybrid diesel–electric vessels during ordinary seagoing operations. Feature-engineering techniques are first developed to transfer event-log files into a consistent numerical form in a lower dimensional space.

A. Ship EPS System

Various designs of EPSs [31] depend on ship size, desired speed, loads, and performance. EPSs are superior over conventional propulsion systems in terms of increased reliability, functionality, efficient space and energy use, reduced emissions, and possibility to use energy storage devices for transient efficiency or fully electric battery propulsion.

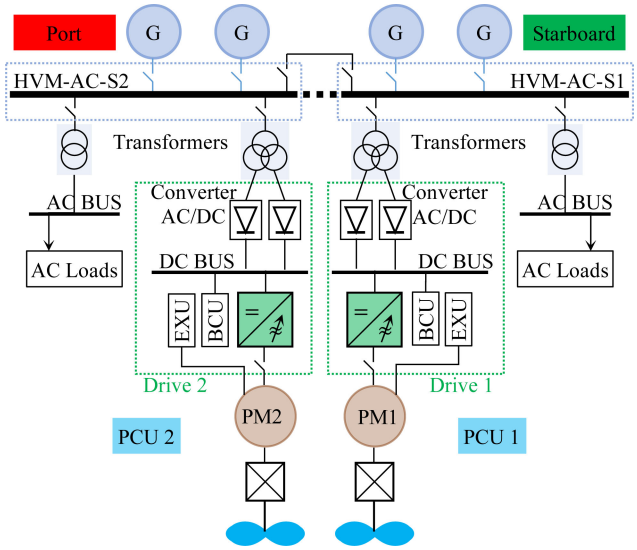

Fig. 1. General overview of an electric propulsion system and power plant.

A typical structure of a propulsion and power plant—in the vessels under this study—is illustrated in the simplified diagram of Fig. 1. The power plant is powered by the generators (G), which convert fuel energy into electric energy in the form of alternating current (ac); they are connected to the system through the high voltage main ac switchboard. For reliability and performance, four generators power the vessels, and the main switchboard is symmetrically divided into two sections. Power transformers step up or down the voltage and current as necessary. The direct current (dc) bus is connected through ac to dc rectifiers. The frequency-controlled drives include the inverters as the main controllers of the variable-speed propulsion motors connected to the propellers through the gearbox. Battery control unit, braking unit, and excitation units are part of the system. Overall, the propulsion control unit is the master controller that maintains a smooth and efficient vessel operation. Protection and auxiliary units, such as cooling-water units, and other loads are not shown in Fig. 1, but they are monitored by the collected data. This IPdM design uses event data from all monitored units, and noise features are filtered at the end.

As shown in Fig. 1, the inverters are central units in the system. This work considers *trips* as the failures of concern that should be timely forecasted and prevented. This failure signifies an unplanned shut down of the inverter unit during operation. Such a failure causes propulsion loss, which may result in severe hazards, as in the incidents [12], [13].

B. Event Data

Around two years of historical data are collected from various vessels. This article fully describes an ith event by the octet

$$E_i = (t_i,\ D_i,\ U_i, V_i, M_i, S_i, A_i, L_i) \tag{1}$$

where

E recorded event;
i event index or event number;
t timestamp when E was recorded;
D drive where E was detected, e.g., port or starboard;
U unit where E is detected, e.g., "cooling-water unit;"
V severity, $V \in \{Alarm,\ Warning,\ Information\}$;
M event name, e.g., "over temperature;"
S activation status, $S \in \{Active,\ Inactive\}$;
A acknowledgment status, $A \in \{Acknowledged,\ Unacknowledged\}$;
L level, $L \in \{Low, High,\ LL,\ HH,\ \ldots\}$.

The events are involuntary and occur at irregular timestamps while the surface ship operates at diverse conditions. They result from various units and locations on board the ship. The events are mainly classified into three severity categories. An *alarm* event signifies an abnormal state of a unit that requires the operator's intervention to restore the unit to the normal state. *Warning* events indicate an abnormal state, but they do not require the operator's intervention and the unit can continue to function. *Informational events* report the completion of task execution, new setpoints, and crew activities. Each event reports a categorical message that describes the happening.

Particularly, alarm events are reported as *active* when monitored states become out of safety limits, whereas alarms are reported as *inactive* when the unit condition is back within the control limits. The alarms activate as unacknowledged and they are reported again after they are acknowledged by the operators. The timespans between activation, deactivation, and acknowledging give significant insight in event-log analysis.

The events are also associated with categorical levels. In particular, alarm levels represent how far the state is outside the safety bounds, the level or risk, or intervention priority. Human–machine interface is important in alarm management to avoid alarm flooding (or overloading) [32]. However, the presented event logs analysis uses all categories of events.

The events are transformed into a single attribute called *event type* that uniquely combines all the above discrete attributes

$$E_i = \left(t_i,\ U^j\right) \quad \forall i\ =\ 1,\ \ldots, N_E,\ j \in \{1,\ \ldots,\ N_U\} \tag{2}$$

where event type is a categorical variable that takes N_U different discrete values U^j. i and j are, respectively, the *event index* and the *event type index*; i is ordinal while j is nominal. N_E and N_U are, respectively, the number of all events and the number of event types from one vessel; they are, respectively, proportional to sample size and dimensionality; they are, respectively, in the orders of 10^5–10^6 and 10^3–10^4 in various EPSs.

Using the event representation in (2), historical events and failures are illustrated in Fig. 2 for one of the studied vessels. These data include few thousands of alarm events; the rest are informational and warnings. The data sample is huge, yet it is not straightforward to train failure predictive models. Many observations, as shown in Fig. 2, reflect the following typical IPdM challenges.

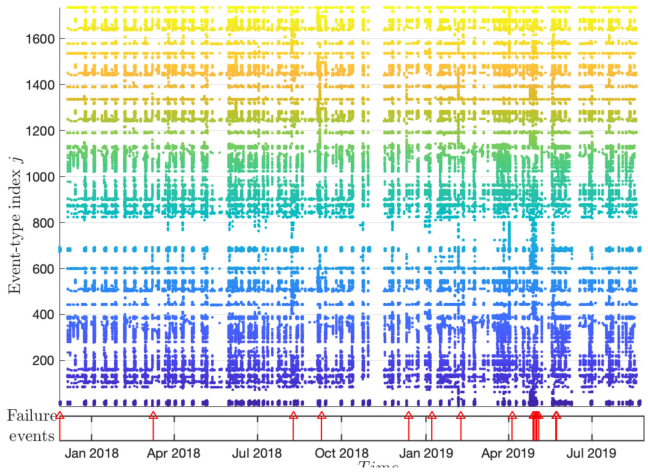

Fig. 2. Visualization of historical event data and failure events collected from one LNG carrier vessel.

1) High-dimensional data and spatiotemporal patterns.
2) Variation in both prefailure and normal operation contexts.
3) Failures make a minority class with extreme between-class imbalance.
4) Multiple failure modes cause within-class imbalance.
5) Only sequences of events can explain a failure perdition.
6) Unlabeled data without expert knowledge to forecast a failure.
7) Unknown event/unit/moment that caused or witnessed symptoms before a failure.
8) Event timestamps are irregular.

Finally, event logs are inconsistent in dimension and in linguistic representation across a fleet of ships equipped with the same EPS technology. Log files include long messages designed for operators and inspectors; these are inconsistent due to varying numbers and types of equipment, different alarm dictionaries and abbreviations, and even unsimilar alarm philosophies. Consequently, practically identical types of events might have different linguistic representations. Since manual data matching is impractical, feature engineering is needed to represent irregular and consistent data in a form that allows efficient training of generalizable IPdM models.

C. Feature Engineering

Two feature-engineering approaches are designed using an extended version of RI combined with time and window aggregation to tackle the above issues of inconsistent, high-dimensional, and irregular data. The ith *time window* W_i represents a momentary operation context of time length T

$$\begin{cases} W_i = (T_i,\ N_i)\ , \quad T_i = t_0\ + i \times T, \quad i\ =\ 1,\ \ldots,\ N_W \\ N_W = \left\lceil \frac{\left(t_{N_E} - t_0\right)}{T} \right\rceil\ ; \quad N_i \in \mathbb{N}^{N_U \times 1}\ \text{s.t.} \\ N_{i,\,j} = \sum\limits_{T_i - T < t_n \le T_i} \mathbb{I}\left(E_n = \left(t_n,\ U^j\right)\right) \end{cases} \tag{3}$$

where T_i are the regular time intervals and N_i is the count of event types' occurrences within a window of duration T. $\mathbb{I}$ is the indicator function that returns 1 if its argument condition is true and 0 otherwise. The time windows approach was used in [22] to discretize raw signals into events using the concept of rising and falling edges. The full *sequence* of events S_i is modeled using rolling windows, which aggregate L_S successive time windows using aggregation functions $\mathcal{F}^k$

$$S_i = \left(T_i,\ \left[\mathcal{F}^k\left(W\right)\right]\right)\ ,\quad W\ = [W_{i-L_S+1}, \ldots, W_n, \ldots, W_i]$$
$$\mathcal{F}^k : \mathbb{N}^{N_U \times L_S} \rightarrow \mathbb{R}^{N_U \times 1},\quad k\ =\ 1,\ 2,\ 3$$
$$\begin{cases} \mathcal{F}^1\ (W) = \max\left(W\right) \\ \mathcal{F}^2\ (W) = \min\left(W\right) \\ \mathcal{F}^3\ (W) = \text{mean}\left(W\right) \end{cases}. \tag{4}$$

These functions are chosen because momentary and total absence or overload of a particular event type can be a failure-indicator feature. In this article, t_i, T_i, and $L_s \times T$ are, respectively, in the order of milliseconds, minutes, and hours; whereas the objective is to predict failures few days in advance.

Unfortunately, S_i resides in a $3N_U$-dimensional space without any measure of similarity of event types. The natural language processing technique of RI [23] is used analogically to word-space model and word context representation. This work extends RI into TRI that represents event types, event contexts, and event sequences in a consistent lower dimensional space. First, each event type U^j is assigned a unique *random index vector* I^j of dimension $d \ll N_U$ whose kth elements are sparsely populated as

$$I_k^j = \begin{cases} -1, & probability \frac{p}{2} \\ 0, & probability\ 1-p \\ 1, & probability \frac{p}{2} \end{cases} \tag{5}$$

where p is the population rate, set around 10%.

Conventional RI measures the word context in an expression defined by a number of adjacent words. In TRI, the jth *event type context* C^j is measured in a time window of duration $2T_{\text{RI}}$

$$\forall j \text{ initialize } C^j\ =\ 0 \in \mathbb{N}^{d\times 1}\ \ \forall i\ =\ 1,\ \ldots,\ N_E \text{ do}$$
$$\begin{cases} \exists k \text{ s.t. } E_i = \left(t_i,\ U^k\right)\ , \text{ update } C^k \\ C^k \leftarrow C^k + \sum_{\substack{t_i - T_{\text{RI}} \le t_n \le t_i + T_{\text{RI}} \\ n \ne i}} I^m \\ m \text{ s.t. } E_n = (t_n,\ U^m) \end{cases} \tag{6}$$

where $C\ = [C^1,\ \ldots,\ C^k,\ \ldots,\ C^{N_U}]\ \in \mathbb{N}^{d\times N_U}$ is the *co-occurrence matrix*, and it is sparse due to skewed event types' occurrence rate. Combined with (3) and (4), TRI represents time windows RIW and sequences RIS using their contexts CW and CS through an exponential weighting factor α

$$W_i = (T_i,\ N_i)\ \ \Rightarrow \text{RI}\ W_i = (T_i,\ CW_i)\ ,\ CW_i \in \mathbb{N}^{d\times 1}$$
$$C\ W_i = \sum_n C^m\ \text{ s.t. }\ E_n = (t_n,\ U^m)\ \ \&\ \ T_{i-1} < t_n \le T_i \tag{7}$$

$$\text{RI}S_i = (T_i,\ CS_i)\ ,\ CS_i \in \mathbb{R}^{d\times 1}$$

TABLE I
DIMENSIONS OF RAW AND PROCESSED HISTORICAL EVENT DATA

V	TS	NS	N_U	N_E	NFE	N	NFS	NFST
1	638	16	1,737	339,521	36	91,562	24	6
2	690	36	3,232	443,624	40	99,142	30	14
3	697	26	3,331	1,187,118	48	100,141	17	4
4	675	24	2,308	580,914	14	96,946	12	5

Columns stand for: V-Vessel number; TS-Timespan in days; NS-Number of raw data sources (log files); N_U-Number of unique event types; N_E-Number of recorded events; NFE-Number of failure events in raw data; N-Number of data samples after feature engineering; NFS-Number of failure samples; NFST-Number of failure samples in the training data subset.

$$CS_i = \sum_{n=n_0}^{i} \beta\ e^{\alpha(n-n_0)}\ CW_n,\ \text{s.t.}$$
$$n_0 =\ i - L_S + 1;\ \frac{1}{\beta} = \sum_{k=0}^{L_S-1} e^{\alpha\ k}. \tag{8}$$

The conventional RI word-space model is useful for text compression, word meaning extraction, and similarity determination. TRI inherits these advantages in IPdM data for substantial dimensionality reduction and measuring event type similarity. Event types that occur in the same operation context have similar TRI context vectors regardless of their different textual messages. TRI is an incremental method, and it can be transferred to another system or updated for new event types efficiently without changing the dimensionality d.

The processed data samples are represented as $X \in \mathbb{R}^{N\times M}$, where N is the number of samples and $M\ =\ 3N_U + d$ is the total number of features. In the following, samples represent event sequences, and they are categorized as follows.

1) Failure samples contain a failure event to be forecasted.
2) Infected samples reflect causes or early symptoms before an emerging failure; the main goal is to learn and predict them.
3) Normal samples represent problem-free operation context.

A failure sample can be an infected sample that predicts another failure in the future, but not the same failure it contains. Table I describes historical data from four LNG carriers equipped with EPSs; it reflects imbalance, dimensionality, and the multiship data heterogeneity.

IV. WEAKLY SUPERVISED IPDM MODELS

Using the engineered features, failure likelihood and t2f predictions are casted into classification and regression problems. The models are based on RF for their interpretability, robustness to noise features and data imbalance, and good performance under high dimensions. B-RF and MIL extensions are introduced in MIL-B-RF to tackle two main IPdM issues: unlabeled and extremely imbalanced data.

Given unknown infected samples, the developed weakly supervised learning approach learns the unknown actual *class labels* $c_i = 1, \ldots, C$ for all samples $X_i \in \mathbb{R}^{1\times M},\ i\ = 1, \ldots, N$

while training the model that predicts future *failure likelihood*, then a second model to predict *t2f* $y_i \in \mathbb{R}$.

A. Balanced Random Forest

RF is an ensemble of classification or regression trees, which recursively model the dataset by splitting samples into M-dimensional subsets called nodes. RF uses a training dataset of pairs $(X_i,\ c_i)$ or $(X_i,\ y_i)$ of samples with known responses. Each decision tree is grown by minimizing a cost function, defined over *node* $\mathcal{D}$ as the squared error [33] for a regression problem

$$\mathcal{D} \subseteq \{(X_i, y_i)\},\ \text{cost}\ (\mathcal{D}) = \sum_i (y_i - \bar{y})^2$$
$$\bar{y} = \frac{1}{|\mathcal{D}|} \sum_i y_i \tag{9}$$

where $\bar{y}$ is the *mean response* in $\mathcal{D}$ and $|\mathcal{D}|$ is the cardinality of $\mathcal{D}$; or the Gini impurity index [33] for a classification problem

$$\mathcal{D} \subseteq \{(X_i,\ c_i)\},\ \text{cost}\ (\mathcal{D}) = \sum_{c=1}^{C} \hat{\pi}_c\,(1 - \hat{\pi}_c)$$
$$\hat{\pi}_c = \frac{1}{|\mathcal{D}|} \sum_i \mathbb{I}\,(\,c_i = \,c)\ \forall c = 1, \ldots,\ C \tag{10}$$

where $\hat{\pi}_c$ is the *class-conditional probability*.

A leaf node $\mathcal{D}$ is worth splitting into left and right nodes $\mathcal{D}_L$ and $\mathcal{D}_R$ if the cost reduction gain $\Delta\mathcal{G}$ is significant

$$\Delta\mathcal{G}\ =\ \text{cost}\,(\mathcal{D}) - \left(\frac{|\mathcal{D}_L|}{|\mathcal{D}|}\text{cost}\,(\mathcal{D}_L) + \frac{|\mathcal{D}_R|}{|\mathcal{D}|}\text{cost}\,(\mathcal{D}_R)\right) \tag{11}$$

by finding the optimal splitting feature index f^* and its optimal splitting value v^* s.t. $X^{f^*} = v^*$

$$(f^*,\ v^*) = \arg \min_{f \in \{1,\ \ldots, M\}} \min_{v \in R_j} \Big(\text{cost}\left(\{(X_i,\ r_i)\,|X_i^f \le v\}\right) + \text{cost}\left(\{(X_i,\ r_i)\,|X_i^f > v\}\right)\Big) \tag{12}$$

where r_i is the response (y_i or c_i) and R_j is the set of values that the feature X^j can take. The tree can be grown until a stopping criterion, such as $\Delta\mathcal{G}$ threshold, tree depth, or $|\mathcal{D}|$ lower limit.

The *j*th *feature importance* $\mathcal{I}_j(\mathcal{T})$ of a trained tree $\mathcal{T}$ is easily determined as the total cost reduction gain from all $\mathcal{N}$ internal nodes where the *j*th feature was used for splitting is as follows:

$$\mathcal{I}_j\ (\mathcal{T}) = \sum_{n \in \mathcal{N}} \Delta\mathcal{G}_n \mathbb{I}\,(\,f_n^* = \,j) \tag{13}$$

and it yields a crucial explain ability advantage in IPdM. The *prediction confidence* of X_i by the b^{th} trained classification tree $\mathcal{T}^b$ is $\hat{\pi}_c$ of the leaf node where X_i ended in the model, denoted as $\hat{\pi}_c^b\ (X_i) = \hat{\pi}_{c,i}^b$. The *classification confidence* is defined as

$$\hat{f}_{c,i}^b = \mathcal{T}_c^b\ \ (X_i) = \hat{\pi}_{c,i}^b\ - \frac{1}{C}\ \forall i, c, b. \tag{14}$$

The predicted response of X_i by the *b*th regression tree $\mathcal{T}^b$ is $\bar{y}$ of the leaf node where X_i resides, denoted as $\hat{y}_i^b = \mathcal{T}^b\ (X_i)$.

RF algorithm aggregates those predictions from N_T trained trees in a model $F(X)$. The responses $\hat{y}_i$ or $\hat{\pi}_{c,i}$ and $\hat{c}_i$ of X_i and the normalized total feature impotence $\mathcal{I}_j(F)$ are determined as

$$\begin{cases} \hat{y}_i = \ F\ (X_i) = \frac{1}{N_T} \sum\limits_{b\,=\,1}^{N_T} \hat{y}_i^b \\ \hat{\pi}_{c,i} = F_c\ \ (X_i) = \frac{1}{N_T} \sum\limits_{b\,=\,1}^{N_T} \hat{\pi}_{c,i}^b \\ \hat{c}_i = \arg \max\limits_{c=1,\ldots,\,C} F_c\,(X_i) \end{cases} \tag{15}$$
$$\mathcal{J}_j^{'}\ (F) = \frac{1}{N_T} \sum_{b\,=\,1}^{N_T} \mathcal{J}_j\left(\mathcal{J}^b\right).\mathcal{J}_j\ \ (F) = \frac{\mathcal{J}_j^{'}(F)}{\sum_{j=1}^{M} \mathcal{J}_j^{'}(F)}.$$

This bagging process increases robustness to overfitting and noise features; Hastie *et al.* [33] showed that the RF model variance σ_F^2 is lower than the individual trees variance σ^2

$$\sigma_F^2 = \ \rho\ \sigma^2 + \frac{1-\rho}{N_T}\sigma^2 \tag{16}$$

and it is minimum for zero pairwise correlation of the trees ρ. $\rho = 0$ is not possible if the features are dependent. To reduce ρ, RF uses *bootstrap aggregation* by training the trees using randomly selected $m < M$ features and $n < N$ samples; Hastie *et al.* [33] suggested $m^* = \sqrt{M}$. This process is further modified to avoid biased models trained through extremely imbalanced data. In B-RF, $n/2$ bootstrapped samples are first drawn randomly from the minority class, then $n/2$ samples are sampled using the same process from the other class. This B-RF method is simple and robust given unknown changing imbalance rate during MIL.

B. MIL-B-RF Approach

Sample class c_i and label notations are defined for IPdM as

$$(X_i,\ c_i): \begin{cases} c_i = \ 1 \Leftrightarrow X_i \text{ is Negative} \equiv \text{Normal sample} \\ c_i = \ 2 \Leftrightarrow \begin{cases} X_i \text{ is Positive} \equiv \text{Infected sample} \\ y_i > 0 \text{ is time to failure} \end{cases} \end{cases}. \tag{17}$$

Recall that c_i is, in fact, unknown. This unlabeled data form a weakly supervised learning problem [30]. In *accurate supervision* frameworks, the *actual class label probability*

$$p_{i,c} = \ p(\ c_i = \ c|X_i)\ \forall c, i \tag{18}$$

is determined using crowdsourcing, expert knowledge, or labeling functions. These are unavailable for this IPdM problem, that is, formulated as an *inexact supervision* problem where $\tilde{p}_{i,c}$ are the optimization variables to be learned. In MIL, unlabeled samples are grouped into N_B disjoint sets, called *Bags* $B^j = \{(X_i^j,\ c_i^j = ?)\}$ assigned some known *bag labels* c^j yet it is desired to classify the samples and not the bags.

Given partial knowledge, each EPS failure's causes or symptoms should have occurred during $\aleph \times L_S \times T$ (few days) before the failure, i.e., at least one *witness sample* from these $\aleph$ samples (few hundreds) should predict that failure. This defines *positive bags* (denoted B^+) with positive *bag labels*. $B^j \equiv B^+$ with $c^j = \ 2$ is a positive bag preceding the *j*th failure

$$\forall\ E_f = (t_f, U \in \{\text{Failures}\})\ ,\ \exists\ T_{n-1} < t_f \le T_n$$

$$\text{New} \begin{cases} \text{Positive bag } B^j \equiv B^+, \quad c^j = 2 \\ B^j = \{(X_k, c_k = ?) \,|n - \aleph < k \leq n-1\} \\ y_k = t_f - T_k > 0 \\ B^j = \left\{ \left(X_i^j,\ c_i^j = ? \right) | i = 1, \ldots, \aleph \right\} \end{cases} . \quad (19)$$

In the case of short time between failures, B^+ are reduced and contain $\aleph^j \leq \aleph$ samples. The remaining training samples are grouped in *negative bags* B^- : $\forall X_n \notin B^j$ s.t. $c^j = 2$, $X_n \in B^q$ s.t. $c^q = 1$. Not all samples in B^+ are positive, most are negative and similar to samples in B^-; all samples must satisfy three *bags conditions*

$$\begin{cases} \forall j \ s.t. \ c^j = 1 \ \sum_i \mathbb{I}\left(c_i^j = 2 \right) = 0 \\ \forall j \ s.t. \ c^j = 2 \ \sum_i \mathbb{I}\left(c_i^j = 2 \right) > 0 \\ \forall i \ \sum_c \tilde{p}_{i,c} = 1 \end{cases} . \quad (20)$$

MIL is formulated as [34]

$$(\{\tilde{c}\}, \mathcal{F}^*) = \arg \min_{\{c\},\, \mathcal{F}(.)} \mathcal{L}\left(F_c\left(X\right)\right) \ \text{s.t.} \begin{cases} B^+ : \forall c^j = 2, \sum_i \mathbb{I}\left(\tilde{c}_i^j = 2 \right) \geq 1 \\ B^- : \forall c^j = 1, \sum_i \mathbb{I}\left(\tilde{c}_i^j = 1 \right) = 0 \end{cases} . \quad (21)$$

Here, MIL aims at training an optimal B-RF predictive model $\mathcal{F}$, $\mathcal{F}\left(X_i\right) = \hat{c}_i$, while learning the unknown actual labels $\tilde{c}$ for all training samples by minimizing a *total loss function* $\mathcal{L}$ defined over F_c such that bag conditions (20) are satisfied. The loss ℓ is selected as negative total *classification margin* $\mathcal{M}$ [35]

$$\begin{aligned} \mathcal{L}\ \left(F_c\left(X\right)\right) &= \sum_{j=1}^{N_B} \sum_{i=1}^{\aleph^j} \ell\left(X_i^j, \tilde{c}_i^j \right) \\ \mathcal{M}\ \left(X_i^j, \tilde{c}_i^j \right) &= F_{\tilde{c}_i^j}\ \left(X_i^j \right) \\ &\quad - \max_{\substack{c=1,\,\ldots,\,C \\ c \neq \tilde{c}_i^j}} F_c\left(X_i^j \right) = -\ell\left(X_i^j, \tilde{c}_i^j \right). \end{aligned} \quad (22)$$

Combined with (15), notice that $\mathcal{M}>0$ for correct prediction $\hat{c}_i^j = \tilde{c}_i^j$, i.e., $\hat{c}_i^j = \mathcal{F}(X_i^j)$ matches the uncovered label $\tilde{c}_i^j$; $\mathcal{M} = 1$ if $\mathcal{F}$ has total confidence in $\hat{c}_i^j$; $\mathcal{M}$ is negative for wrong predictions, and $\mathcal{M} = -1$ for total confidence in the wrong label. Hence, this design aims at clear separation besides correct predictions. Notice that (21) is an integer programming problem since $\{\tilde{c}\}$ are integers, the problem is reformulated into finding the actual class probabilities $\tilde{p}_{i,c}^j = p(\ \tilde{c}_i^j = c|X_i^j)\ \forall i, j, c$ defined over a space of distributions $\mathcal{P}$.

Optimizing for both $(\tilde{p}, \mathcal{F}^*)$ is a nonconvex problem that can be solved using DA [34] by adding a weighted entropy term to $\mathcal{L}$ and solving iteratively for

$$\tilde{p} = \arg \min_{p \in \mathcal{P}} \mathcal{L}_{\text{DA}}\left(F\left(X\right), p\right)$$

$$\mathcal{L}_{\text{DA}}\ \left(F\left(X\right), p\right) = \mathbb{E}_{\mathcal{P}}\ \left(\ell\left(F\left(X\right)\right)\right) - \tau\ \mathcal{H}\left(\mathcal{P}\right)$$

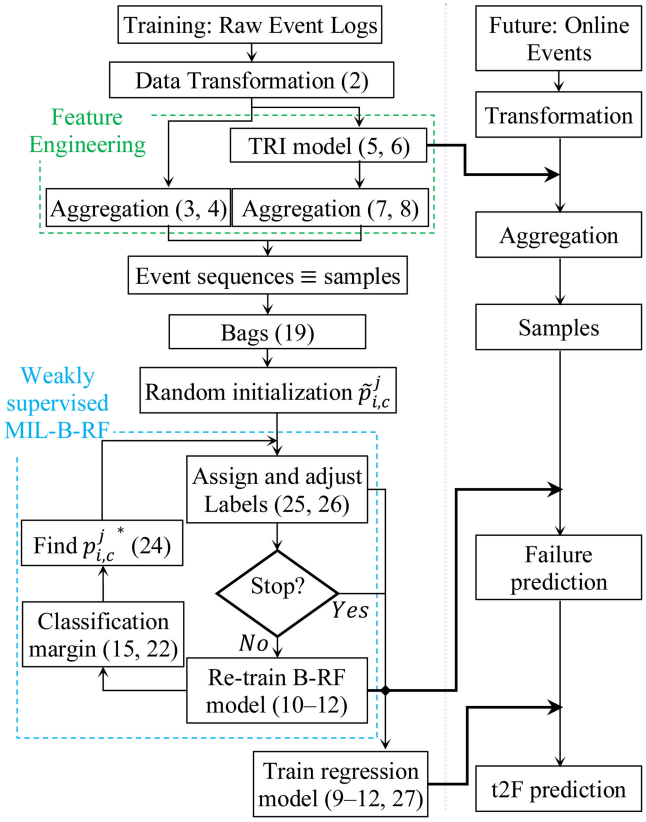

Fig. 3. General overview of the MIL-B-RF-based EPS IPdM approach.

$$= \sum_{j=1}^{N_B} \sum_{i=1}^{\aleph^j} \sum_{c=1}^{C} p_{i,c}^j\ \ell\left(X_i^j, c \right) + \tau \sum_{j=1}^{N_B} \sum_{i=1}^{\aleph^j} \sum_{c=1}^{C} p_{i,c}^j \log\left(p_{i,c}^j \right) \quad (23)$$

given the constraints in (20). $\tau\ (r) = e^{-\gamma\, r}$ is a cooling function over iterations r with parameter $\gamma > 0$. Initial larger τ values encourage MIL *exploration*, whereas final smallest values enforce *exploitation*. At one iteration, (23) solves as [35]

$$\frac{\partial \mathcal{L}_{\text{DA}}\left(F\left(X\right), p^*\right)}{\partial p} = 0 \ \Rightarrow \ p_{i,c}^j{*} = e^{\frac{\mathcal{M}\left(X_i^j, c\right) - \tau}{\tau}} \quad (24)$$

optimal parameters $p_{i,c}^j{*}$ are then adjusted to $\tilde{p}_{i,c}^j$ to satisfy (20)

$$\begin{cases} \forall\ c^j = 1, \ \tilde{p}_{i,1}^j = 1\ and\ \tilde{p}_{i,2}^j = 0 \\ \forall\ c^j = 2, \ \tilde{p}_{i,c}^j = \frac{p_{i,c}^{j\ *}}{\sum_c p_{i,c}^{j\ *}} \end{cases} . \quad (25)$$

Labels $\tilde{c}_i^j$ are assigned to samples X_i^j using optimized adjusted probability distribution $\tilde{p}_{i,c}^j$ followed by label correction

$$\begin{aligned} &\forall\ c^j = 2, \text{ if } \ \sum_i \mathbb{I}\left(\tilde{c}_i^j = 2 \right) = 0 \\ &\text{then } \ \tilde{c}_{i^*}^j = 2 \text{ s.t. } \ i^* = \arg\max_i \tilde{p}_{i,c}^j. \end{aligned} \quad (26)$$

As shown in Fig. 3, this MIL process is repeated iteratively by training new weakly supervised B-RF models and learning

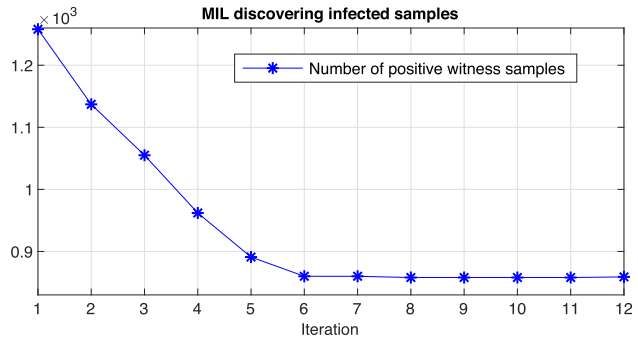

Fig. 4. MIL-B-RF convergence: Learning the unknown actual class labels.

the unknown actual labels. Once there are no changes in label uncovering, the final failure prediction model is returned. The t2F regression model is trained using only positive prediction samples and h most important features

$$R = \{(X = [x_{i,j}],\ y_i)\} \quad \text{s.t.} \begin{cases} \hat{c}_i = 2 \\ \mathcal{I}_j(\mathcal{F}) \geq \mathcal{I}^{50} \end{cases} \tag{27}$$

where y_i is the t2F (19) and $\mathcal{I}^h$ is the hth highest total Gini feature importance (15). $h \ll M$ to avoid noise features and a high-dimensional problem due to very few positive samples.

V. Results and Discussion

Historical EPS event datasets (see Table I) span around two years; the first-year dataset is used for training, whereas the second is interpreted as the future data and used to test forecasting real-world failures. There is no information about the units, events, or time slots that indicate failure causes or symptoms. IPdM models are tested with no knowledge about the failure time, and the latter is only used for assessment. The time windows are constructed over $T = 10$ min to represent momentary operation context; the rolling sequences span 1 h, $L_S = 6$. The random index vector has a population rate of $p = 0.1$ and a size $d = 100 \ll N_U$, which grants a substantial dimensionality reduction without loss of information; the event type context is measured over $T_{\text{RI}} = 10$ s; the exponential TRI weighting factor is selected as $\alpha = 1$. With insufficient data, feature-engineering parameters are initialized reasonable values and tuned in several trials due to the lack of extra tuning dataset.

The weakly labeled training data have a maximum positive bag size of three days, i.e., a witness sample forecasting an emerging failure is a candidate among 427 unlabeled prefailure samples. The iterative MIL cooling function is controlled with $\gamma = 1/5$. This resulted in the convergence, as shown in Fig. 4.

The MIL-B-RF algorithm converges relatively slower than MIL in [34], but it is stable and identifies actual class labels with high confidence, i.e., $p^j_{i,c}*$ are very close to 0 or 1 at the last iteration.

MIL-B-RF performance is shown in Fig. 5. Most samples are correctly flagged as negative with the highest confidence $F_2(X_i) \approx -0.5 \equiv F_1(X_i) \approx 0.5$. The default *classification confidence limit* $\text{CL} = 0.4$ (90% prediction confidence) is used to flag positive samples that forecast failures, $F_2(X_i) > 0.4$.

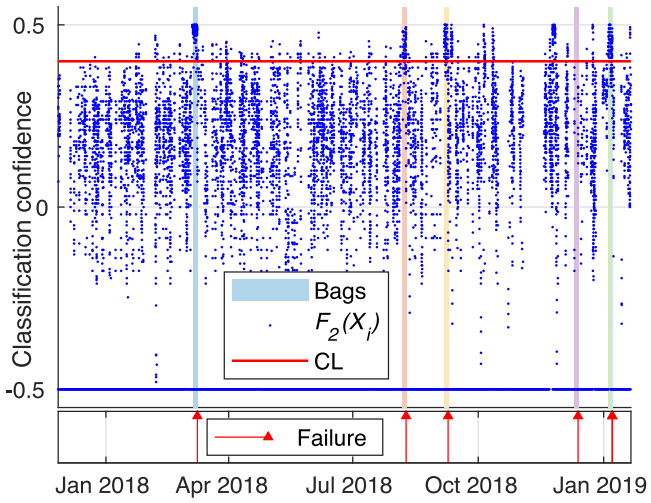

Fig. 5. MIL-B-RF uncovering actual class labels in the training dataset. Bags are three days before failure, $F_{c=2}(X_i)$ is the classification confidence in an infected sample, $F_2(X_i) > \text{CL}$ signifies a positive prediction.

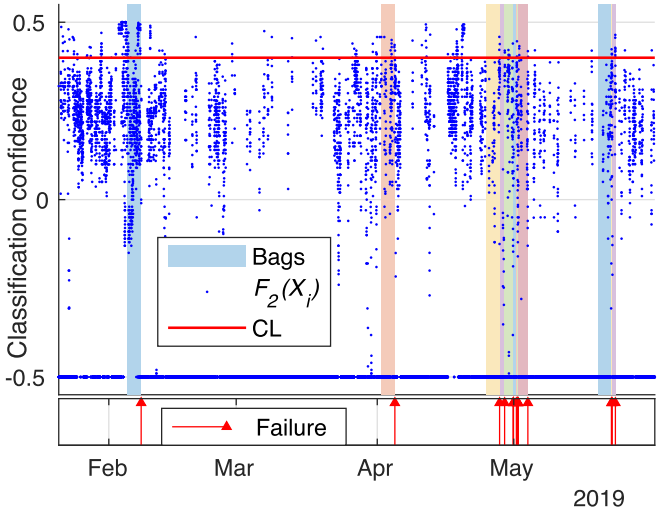

Fig. 6. MIL-B-RF real-world failure prediction in a test dataset from ship 1.

The trained MIL-B-RF is tested on an independent dataset where actual failures are predicted successfully, as shown in Fig. 6. False positive rate (FPR) is extremely low; it is attributed to near-failure conditions and PvM strategies; it can be reduced by tuning CL or designing a moving filter. This, however, requires a separate hyperparameter tuning dataset. Performance results are summarized in Table II for the four vessels; the first raw lists the results for negative samples, while the last row counts only failures that are forecasted by at least one witness sample in the bag (427 samples per bag); most failures are predicted successfully with nearly 100% true positive rate (TPR).

TABLE II
FAILURE PREDICTION PERFORMANCE ACROSS FOUR VESSELS

	Vessel 1		Vessel 2		Vessel 3		Vessel 4	
	$\hat{c}=1$	$\hat{c}=2$	$\hat{c}=1$	$\hat{c}=2$	$\hat{c}=1$	$\hat{c}=2$	$\hat{c}=1$	$\hat{c}=2$
$X_i^j, c_i^j = 1$	28,333	281	12,603	227	19,144	553	30,912	2,081
$B^+: c^j = 2$	1	17	0	16	0	13	0	7

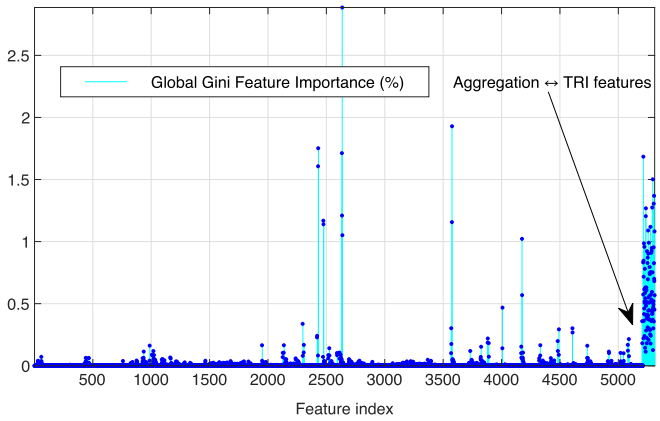

Fig. 7. MIL-B-RF total Gini feature importance; right-most 100 features are the TRI-engineered features.

TABLE III
PERFORMANCE COMPARISON OF MIL-B-RF MODELS TRAINED ON SUBSETS OF FEATURES FOR VESSEL 1

	$\mathcal{F}_0$: all features		$\mathcal{F}_1$: all aggr. features		$\mathcal{F}_2$: TRI features		$\mathcal{F}_3$: *max* aggr. features	
	$\hat{c}=1$	$\hat{c}=2$	$\hat{c}=1$	$\hat{c}=2$	$\hat{c}=1$	$\hat{c}=2$	$\hat{c}=1$	$\hat{c}=2$
$X_i^j, c_i^j = 1$	28,333	281	28,364	250	28,444	170	28,319	295
$B^+: c^j = 2$	1	17	6	12	5	13	5	13

The total Gini feature importance (15) grants MIL-B-RF an interpretability advantage. Indeed, Fig. 7 shows the few features that mostly explain the model predictions. More remarkably, the 100 TRI-engineered features contributed 60.7% cumulative feature importance compared with the remaining 5211 aggregation features, but the latter are more explainable than the TRI features, which are unexplainable due to their nearly uniform importance and untraceable due to their encoded form.

The most important aggregation features are from the max operation function (4). Hence, Table III compares the testing performance of four MIL-B-RF models trained on subsets of the features. $\mathcal{F}_0$, $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_3$ use, respectively, all features, only aggregation features, only TRI features, and only max operation features. TRI outperforms simple aggregation again.

Although it is most important to predict the failure likelihood and explain the predictions, t2F (27) indicates the available time bonus for intervention. Fig. 8 depicts the test performance results for t2F prediction error $\Delta t2F$ normalized to the bag time length. The mean error is close to zero; the error variance is attributed to very few infected samples used for training.

Finally, Table IV compares the presented MIL-B-RF results with three closely related event-driven PdM methods [22]–[24].

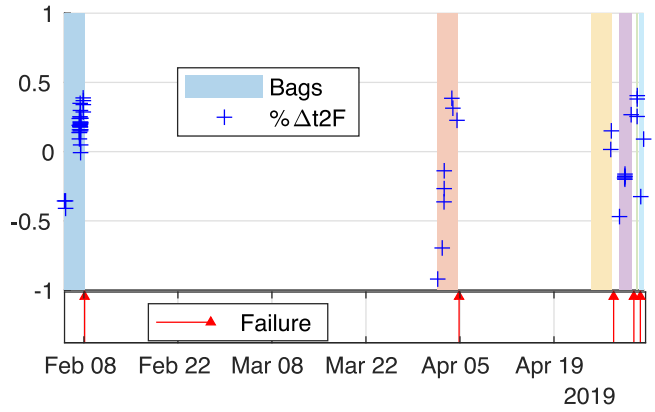

Fig. 8. T2f prediction error of true positives in the test dataset.

TABLE IV
MIL-B-RF PERFORMANCE COMPARISON WITH CONTEMPORARY WORKS

Approach	Application	Performance
MIL-B-RF	Multi-unit EPS in LNG carrier	$1\% \leq \text{FPR} \leq 6.73\%$ $94.4\% \leq \text{TPR}$
[22] IPdM	Milling machine equipment	$51\% \leq \text{TPR} \leq 69\%$
[23] IPdM	Cold forming press equipment	$0.48 \leq F_1 \text{ Score} \leq 0.61$
[24] IPdM	Anonymous company	$2\% \leq \text{FPR} \leq 48\%$ $17\% \leq \text{TPR} \leq 93\%$

Due to the limited IPdM reports in the maritime literature, the selected comparisons include results of various applications. However, these methods have been designed and tuned for failure prediction based on event data; hence, their best results are comparable. First, all the compared methods have shown some sort of success in predicting some failures based on raw or synthetic event data. Compared with MIL-B-RF, the other methods did not focus on time correlation between the events, and they did not address the issues of extreme data imbalance and weak data labels. Their performance is, therefore, limited. The results of the article presented in [22] reflect insufficient failure prediction sensitivity and bias in the approach since the best TPR is below 70%. This observation is in line with the results of the article presented in [24], showing a relatively low F_1 score (higher is better up to 1). This bias is reduced in [23] as the failure prediction rate increased up to 93%. However, the latter came at the cost of huge variance where TPR can be as low as 17% and FPR can be as high as 48%. On the contrary, the presented MIL-B-RF IPdM approach predicts failures with high TPR$\geq$94.4%, low FPR$\leq$6.73%, and minimum variance.

VI. CONCLUSION

In this article, we presented a novel IPdM approach, driven by event sequence analysis, and applied to electric propulsion systems of large ships. It contributed to 1) predicting failure likelihood, 2) predicting t2f, and 3) providing explainable predictions. Objectives 1) and 2) were casted into weakly supervised classification and regression problems. In this regard, a threefold

solution was derived to effectively remedy major limitations of event-driven techniques.

TRI was proposed to map the irregular textual logs into a consistent numerical array form. The developed technique resulted in substantial dimensionality reduction and contributed to the dominant failure predictors. In conjunction, event aggregation techniques were developed for explainable PdM to track failure sources.

Notwithstanding the extremely imbalanced data, unbiased models were developed using a balancing strategy. Despite the lack of hand-labeled data, the overall approach was successfully designed to recursively discover the unknown actual class labels of infected samples at high confidence. It trained unbiased base learners that correctly forecast failures within extremely minor classes. In comparison, the presented approach outperformed contemporary methods.

Finally, most of the actual failures were forecasted successfully—by at least one witness sample—within three days preceding the failure. These results were significantly important in the maritime sector to mitigate and reduce the likelihood of propulsion loss during critical maneuvers. This, in turn, prevents hazardous accidents, economic losses, and downtime while it promotes the safety of equipment and personnel.

Future works may consider hyperparameter tuning from larger datasets, sequential (deep) learning to avoid feature-engineering information loss, and prescriptive maintenance.

References

[1] M. S. Bouguerra, A. Gainaru, and F. Cappello, "Failure prediction: What to do with unpredicted failures?," in *28th IEEE Int. Parallel Distrib. Process. Symp.*, 2014.

[2] L. Silvestri, A. Forcina, V. Introna, A. Santolamazza, and V. Cesarotti, "Maintenance transformation through industry 4.0 technologies: A systematic literature review," *Comput. Ind.*, vol. 123, 2020, Art. no. 103335.

[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, 2020, Art. no. 106889.

[4] A. V. Malawade, N. D. Costa, D. Muthirayan, P. P. Khargonekar, and M. A. A. Faruque, "Neuroscience-inspired algorithms for the predictive maintenance of manufacturing systems," *IEEE Trans. Ind. Informat.*, vol. 17, no. 12, pp. 7980–7990, Dec. 2021.

[5] W. Yu, T. Dillon, F. Mostafa, W. Rahayu, and Y. Liu, "A global manufacturing big data ecosystem for fault detection in predictive maintenance," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 183–192, Jan. 2020.

[6] M. Karakose and O. Yaman, "Complex fuzzy system based predictive maintenance approach in railways," *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 6023–6032, Sep. 2020.

[7] W. Zhang, D. Yang, Y. Xu, X. Huang, J. Zhang, and M. Gidlund, "DeepHealth: A self-attention based method for instant intelligent predictive maintenance in industrial internet of things," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5461–5473, Aug. 2021.

[8] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, "Machine learning for predictive maintenance: A multiple classifier approach," *IEEE Trans. Ind. Informat.*, vol. 11, no. 3, pp. 812–820, Jun. 2015.

[9] Q. Wang, S. Bu, and Z. He, "Achieving predictive and proactive maintenance for high-speed railway power equipment with LSTM-RNN," *IEEE Trans. Ind. Informat.*, vol. 16, no. 10, pp. 6509–6517, Oct. 2020.

[10] B. de Jonge and P. A. Scarf, "A review on maintenance optimization," *Eur. J. Oper. Res.*, vol. 285, no. 3, pp. 805–824, 2020.

[11] A. Bakdi, I. K. Glad, E. Vanem, and Ø. Engelhardtsen, "AIS-based multiple vessel collision and grounding risk identification based on adaptive safety domain," *J. Mar. Sci. Eng.*, vol. 8, no. 1, 2020, Art. no. 5.

[12] M. Ibrion, N. Paltrinieri, and A. R. Nejad, "Learning from failures in cruise ship industry: The blackout of Viking sky in Hustadvika, Norway," *Eng. Failure Anal.*, vol. 125, 2021, Art. no. 105355.

[13] ATSB Transport Safety Report, "Grounding of bulk carrier Bulk India," Aust. Transp. Saf. Bur., Dampier, WA, Australia, Sep. 11, 2020.

[14] S. Drago, A. A. Purcărea, A. Cotorcea, N. Florin, and D. Coofre, "Naval maintenance. From corrective maintenance to condition monitoring and IoT. Future trends set by latest IMO amendments and autonomous ships," in *Proc. Int. Sci. Conf. SEA-CONF*, 2021, pp. 323–340.

[15] L. Stazić, T. Stanivuk, and V. Mihanović, "Testing of the evaluation methodology for ship's planned maintenance system database," *J. Appl. Eng. Sci.*, vol. 17, no. 3, pp. 273–279, 2019.

[16] D. Kimera and F. N. Nangolo, "Maintenance practices and parameters for marine mechanical systems: A review," *J. Qual. Maintenance Eng.*, vol. 26, no. 3, pp. 459–488, 2020.

[17] Z. H. Munim, M. Dushenko, V. J. Jimenez, M. H. Shakil, and M. Imset, "Big data and artificial intelligence in the maritime industry: A bibliometric review and future research directions," *Maritime Policy Manage.*, vol. 47, no. 5, pp. 577–597, 2020.

[18] Y. Ran, X. Zhou, P. Lin, Y. Wen, and R. Deng, "A survey of predictive maintenance: Systems, purposes and approaches," 2019. [Online]. Available: https://arxiv.org/abs/1912.07383

[19] B. Hrnjica and S. Softic, "Explainable AI in manufacturing: A predictive maintenance case study," in *Advances in Production Management Systems Towards Smart and Digital Manufacturing*, B. Lalic, V. Majstorovic, U. Marjanovic, G. von Cieminski, and D. Romero, Eds. Cham, Switzerland: Springer, 2020, pp. 66–73.

[20] P. Goel, E. N. Pistikopoulos, M. S. Mannan, and A. Datta, "A data-driven alarm and event management framework," *J. Loss Prevention Process Ind.*, vol. 62, 2019, Art. no. 103959.

[21] F. Dama and C. Sinoquet, "Time series analysis and modeling to forecast: A survey," 2021. [Online]. Available: https://arxiv.org/abs/2104.00164

[22] C. Gutschi, N. Furian, J. Suschnigg, D. Neubacher, and S. Voessner, "Log-based predictive maintenance in discrete parts manufacturing," *Procedia CIRP*, vol. 79, pp. 528–533, 2019.

[23] I. Fronza, A. Sillitti, G. Succi, M. Terho, and J. Vlasenko, "Failure prediction based on log files using random indexing and support vector machines," *J. Syst. Softw.*, vol. 86, no. 1, pp. 2–11, 2013.

[24] A. Naskos, G. Kougka, T. Toliopoulos, A. Gounaris, C. Vamvalis, and D. Caljouw, "Event-based predictive maintenance on top of sensor data in a real industry 4.0 case study," in *Machine Learning and Knowledge Discovery in Databases*, P. Cellier and K. Driessens, Eds. New York, NY, USA: Springer, 2020, pp. 345–356.

[25] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Comput. Surv.*, vol. 52, no. 4, 2020, Art. no. 79.

[26] A. Saxena *et al.*, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[27] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[28] Y. Luo, K. Li, Y. Li, D. Cai, C. Zhao, and Q. Meng, "Three-layer Bayesian network for classification of complex power quality disturbances," *IEEE Trans. Ind. Informat.*, vol. 14, no. 9, pp. 3997–4006, Sep. 2018.

[29] Z. Chai and C. Zhao, "Enhanced random forest with concurrent analysis of static and dynamic nodes for industrial fault classification," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 54–66, Jan. 2020.

[30] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, 2018.

[31] C. Nuchturee, T. Li, and H. Xia, "Energy efficiency of integrated electric propulsion for ships—A review," *Renewable Sustain. Energy Rev.*, vol. 134, 2020, Art. no. 110145.

[32] J. Wang, F. Yang, T. Chen, and S. L. Shah, "An overview of industrial alarm systems: Main causes for alarm overloading, research status, and open problems," *IEEE Trans. Autom. Sci. Eng.*, vol. 13, no. 2, pp. 1045–1061, Apr. 2016.

[33] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2017.

[34] H. Bischof, A. Saffari, and C. Leistner, "MIForests: Multiple-instance learning with randomized trees," in *European Conference on Computer Vision*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Germany: Springer, 2010, pp. 29–42.

[35] Y. Xia, Q. Zhu, and W. Wei, "Weakly supervised random forest for multi label image clustering and segmentation," in *Proc. 5th ACM Int. Conf. Multimedia Retrieval*, Shanghai, China, 2015, pp. 227–233.

Azzeddine Bakdi received the Ph.D. degree in control engineering from the University of Boumerdes, Boumerdes, Algeria, in 2018.

He is currently a Postdoctoral Research Fellow with BigInsight Research-Based Innovation Center, Department of Mathematics, University of Oslo, Oslo, Norway. His research interests include computational engineering and particularly in statistics and data science for intelligent, safe, and pollution-free maritime applications.

Nicolay Bjørlo Kristensen received the M.S. degree in data science from the University of Oslo, Oslo, Norway, in 2021.

His research interests include the topics of predictive maintenance and signal and image processing/analysis.

Morten Stakkeland received the Ph.D. degree in engineering cybernetics from the Norwegian University of Science and Technology, Trondheim, Norway, in 2009.

He is currently an Associate Professor II with the Department of Mathematics, Statistics and Data Science Research Group, University of Oslo, Oslo, Norway. His main position is with ABB Marine and Ports, where he is developing the next generation of digital services utilizing advanced analytics and statistics.