Anomaly Detection and Breakdown Diagnosis for Condition Monitoring of Marine Engines

Nhu Khue Vuong[1], Sateesh Babu Giduthuri[1], Gen Liang Lim[1], Terrence Tan[2], Savitha Ramasamy[1]
[1]*Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), Singapore*
[2]*PSA International (PSA), Singapore*
{vuongnk, Giduthuri_Sateesh_Babu, Lim_Gen_Liang}@i2r.a-star.edu.sg, terrent@globalpsa.com, ramasamysa@i2r.a-star.edu.sg

***Abstract*— Marine vessels are complex interconnected systems and maintaining the health of individual components of the system increases uptime, boosts efficiency and safety of vessel operations. With recent transition from preventive to predictive maintenance of engineering assets and the prevalence of IoT sensors embedded within these assets, effective condition monitoring of engineering assets in marine vessels is now a reality. This paper aims at developing a solution for condition monitoring and diagnosis of potential breakdowns in the main engines of marine vessels using sensor data. Specifically, we analyze irregularly sampled multi-variate time series data originating from multiple sensors onboard the vessel engine to develop an autoencoder-based anomaly detection model for effective condition monitoring of the engines. In addition to the anomaly detection model, we devise a hierarchical framework to diagnose the potential cause of breakdown. The model is trained on data obtained from engines of two vessels. We train the model using historical time-series data corresponding to the vessel's condition and operational profile over a month. The model is validated using historical time-series data collected over a year of vessel operations. The performance study of our model demonstrates its ability to predict breakdowns in advance with an average F-1 score of 85.3%, and average 11-14 days in advance of the actual reported breakdown dates. This proposed solution can be a promising tool for the condition monitoring and diagnostics of marine vessel engines.**

Keywords*— *main engine breakdowns, marine vessels predictive maintenance, autoencoders

I. INRODUCTION

Marine vessel is a complex system that comprises of numerous subsystems and components. The main engine is the most critical subsystem responsible for supplying the propulsive force for the vessel. A breakdown of the main engine results in increased downtime which leads to significant revenue loss due to vessel's inability to operate. Conventionally, maintenance of marine engines is facilitated through scheduled preventive maintenance [1]. As estimating the optimal maintenance schedule is cumbersome [2], there is a need for predictive maintenance that enables proactive maintenance only as and when required.

Predictive maintenance solutions can be facilitated through prognosis (classification), remaining useful life prediction (regression) or anomaly detection (unsupervised representation). Although there is a huge volume of data generated through sensors in the vessels, the maintenance records are often noisy owing to manual inputs. This results in noisy labels that affect the performance of supervised approaches adversely. Hence, unsupervised anomaly detection approach towards predictive maintenance of marine engines [3] is the most viable choice.

Though statistical approaches such as auto associative kernel regression [4], dynamical linear models and sequential testing [5] were initially explored, these methods do not suit well for data with dynamically varying sensor measurements due to the uncertain physical operating environment of marine vessels. Clustering based approaches [6] including self-organizing maps, spectral clustering, k-means clustering, density-based clustering, and mixture of Gaussian models face the fallacy that it is uncertain to ascertain if the clusters correspond to another operational condition or anomaly. Moreover, when sensor data set involves high dimensional spaces, clustering-based approaches suffers from low performance. Although an ensemble of clustering approaches is scalable to high dimensional and large-scale sensor data [7], they still face the former issue. Physics-based models and unsupervised deep learning algorithms are possible candidates for predictive maintenance applications of marine engines. However, physics-based models require the knowledge of physical equations and are not easily trainable and transferable across subsystems. On the other hand, unsupervised deep learning algorithms are based on representational learning for time-series sequence prediction and automatic feature extraction towards detecting anomalies [8, 9]. Thus, unsupervised deep learning-based anomaly detection methods are more suitable for anomaly detection in marine engines.

In addition to detecting anomalies, it is highly desirable to also detect the root cause of the anomalies to facilitate diagnosis. Such root cause prediction can be performed independent of or in tandem with the anomaly detection model. Recently, independent models for anomaly detection of marine engines and its root cause prediction have been developed in [7, 10]. However, this method suffers from the following setbacks: (1) it does not scale well especially when dealing with components represented by many sensor parameters, (2) as the anomaly detection and root cause prediction methods are independent, they are often not synchronized.

In this paper, we propose a data-driven unsupervised anomaly detection approach for the predictive maintenance of marine vessels, which are operational in real-world. The contribution of this study is three-fold. First, we develop an anomaly detection approach based on autoencoders for predictive maintenance of marine vessel engine using historical time-series data collected over a year of vessel operations. Second, we devise a reconstruction error based diagnosis of anomalies to identify sensors that attribute to the detected anomaly and then the engine breakdown. Finally, we evaluate the anomaly detection model against real-world operational data of the engines. The data is inherently noisy because the breakdowns reported in the maintenance logs are generally delayed depending on the date of maintenance activity. Hence, the data poses inherent challenges to the modeling and evaluation. We address these challenges in our study and the experimental results show that the proposed model can predict breakdowns in advance of the actual reported breakdowns with an average F1 score of 85.3% and is able to diagnose the sensors causing anomalies.

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00044

II. MATERIALS AND MOTIVATIONS

A. Materials

The datasets used in this study are retrieved from 2 different vessels over a year of vessel operations. Each of the vessels has two main engines: starboard (SB) and portside (PS). These are four-stroke diesel engines with maximum rotation per minute of 750. These engines are equipped with sensors that have different sampling frequencies. The available sensors on each vessel and their corresponding sampling frequencies are shown in Table 1. From the Table, it can be observed that each of the main engine of V1 has 14 sensors and each of the main engine of V2 has 16 sensors. Thus, we include 28 parameters from V1 and 32 parameters from V2 in our analysis. The duration of data used in the analysis depends on the operation of the individual vessels. Accordingly, we use data from Jun 2021 to Jul 2022 for V1 and data from Dec 2021 to Jul 2022 for V2.

TABLE I. MAIN ENGINE SENSORS OF EACH VESSEL

Description of Main Engine Sensors	Frequency	V1	V2
Boost air pressure (BAP)	60	√	√
Boost air temperature (BAT)	60	√	√
Fuel oil flow (FOF)	2	√	√
Fuel oil used (FOU)	60	√	√
Freshwater pressure (FWP)	60	√	√
Fuel pressure (FP)	2	√	√
Pressure of lubircant oil (PLO)	60	√	√
Temperature of lubricant oil (TLO)	60	√	√
Engine rotation per minute (RPM)	3	√	√
Running hours (RH)	60	√	√
Exhaust gas temperature of turbocharger (TCEGT)	4	√	√
Pressure of lubricant oil in turbocharger (TCPLO)	60	√	√
RPM of turbocharger (TCRPM)	60	√	√
Average exhaust gas of 6 cylinders (EGC)	2	√	√
Freshwater temperature (FWT)	60		√
Seawater pressure (SWP)	60		√

We have a total of 2,275,737,602 raw observations (number of samples X number of sensors) from these vessels in the raw datasets. Subsequently, we synchronize all sensors' sampling frequencies to 60Hz and resample the data every minute (one-minute interval). Detailed counts of the raw datasets and the processed datasets for each vessel can be found in Table 2. The processed and resampled datasets are obtained based on the preprocessing procedure detailed in Section IIIA.

TABLE II. COUNTS OF RAW AND PROCESSED DATASETS

Vessel Label	Collection Period (months)	Raw Datasets		Processed and Resampled Datasets	
		No. of Samples	*No. of Sensors*	*No. of Samples*	*No. of Sensors*
V1	14	36,682,342	38	420,572	28
V2	8	20,995,443	42	247,448	32

Compared with other studies [10-12], the datasets used in our work are larger in volume and longer in terms of collection period. In [10], Kim et al. used the datasets with approximate 22.5 million observations collected for 10 months. Other similar studies [11-12] had shorter data collection periods (3 to 10 months) and fewer sensors.

B. Motivations

We present the reported breakdowns in the main engines of V1 and V2 along with their approximate timeline in Fig. 1. In total, there were 27 breakdowns occurred from Jun 2021 to Jul 2022 in the main engines of V1 and 12 breakdowns between Jan 2022 and Jul 2022 in the main engines of V2. The recorded dates of breakdowns shown in Fig. 1 were retrieved from the maintenance logs when engine breakdowns were addressed, and correction actions took place. Therefore, there could be some time gap between the actual breakdown event date till the recorded date in the maintenance log. This poses the following challenges in both data labelling and model evaluation: (1) as the exact date of breakdown is unknown, labels derived from these maintenance logs for training supervised models will be noisy; (2) some breakdowns remain unreported, hence training supervised models using the data with erroneous labels will result in inaccurate models; (3) as breakdowns may occur within shorter duration, it is challenging to ascertain and relate a precursor or anomaly to individual breakdowns because they could either indicate precursors to the earlier reported breakdown or could also be early precursors for subsequent breakdowns.

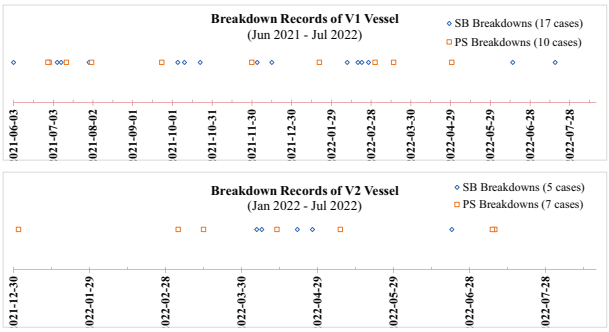

Fig. 1. Breakdown Records of V1 and V2's Main Engines (PS & SB)

From Table 2 and Fig. 1, we have approximately 58 million raw data samples in total and only 39 breakdown events occurred during 8-14 months of vessels' operation. The breakdowns can be due to diverse reasons and can be attributed to multiple interrelated subsystems that are unknown to us. More importantly, the amount of breakdown or breakdown events is exceptionally less, making the data highly imbalanced. All the above-mentioned challenges are a severe bottleneck to the development of a supervised classification based predictive maintenance method for marine engines.

Hence, we propose an unsupervised anomaly detection model towards predictive maintenance of V1 and V2. In doing so, we aim to train an autoencoder model on data during healthy operation of the engines and detect anomalous data samples based on the reconstruction error generated from the trained model. Theoretically, we can train one autoencoder for each engine of each vessel; thus, training four autoencoders in total for the four engines in the two vessels. However, this approach results in multiple models and the number of models will grow with the size of the marine fleet. This also adds to the computational cost of training and storage cost of saving the models for inference in the deployment environment. In the next Section, we propose an end-to-end anomaly detection method to address all the above-mentioned challenges and apply this to the use case in consideration in Section IV.

III. METHOD

Our proposed method to the predictive maintenance application of marine vessels' main engines in real-world encompasses four key steps shown in Fig 2. We first remove samples with noisy parameters and re-sample parameters to standardize the sensing or sampling frequency. Next, we select data samples for training and testing/validation and normalize the dataset for training autoencoder model. This is followed by training an autoencoder model for anomaly detection and root cause diagnosis. Finally, we evaluate the performance of the model against the noisy ground truth labels derived from maintenance reports. In doing so, we estimate the length of the prediction window for each reported breakdown.

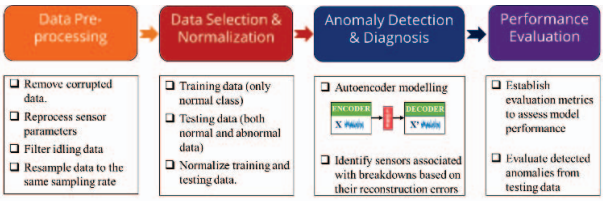

Fig. 2. Overall Framework for Predictive Maintenance of Marine Engines

A. Data Preprocessing

We preprocess the raw data to obtain the clean data for modeling. First, erroneous values such as negative values and duplicated timestamps are removed. Next, we process RH (Running Hours) and FOU (Fuel oil used) parameters (or sensors) because these original parameters are accumulated over time. The difference between the first and last value of the RH parameter for each day is estimated to represent the running hours per day. The FOU is calculated by differentiating the fuel oil used between two consecutive timestamps. Third, we take the average exhaust gas cylinder of 6 cylinders in each SB and PS engine and represent the average value by a single parameter named EGC (Exhaust Gas Cylinders) because we observe that these 6 exhaust gas cylinders are highly correlated to each other. As a result, the processed dataset has 10 sensors less compared with the raw dataset (see Table 1). After that, we filter samples where the RPM is below 375 or above 800 for both vessels because according to the domain expert and their sea trial reports, the engine should be operating at an RPM in between 375 and 800. In the last step, we resample the processed datasets by averaging samples within a one-minute interval to standardize sampling frequencies of all engine sensors. Having synchronized sampling frequencies is necessary for providing online or real-time condition monitoring once the developed models are deployed in operation. The size of the final processed dataset for both vessels are mentioned in Table 2.

B. Data Selection and Normalization

As our approach is to build autoencoders to model engine data under healthy operating conditions, it is crucial to select appropriate training and testing data. From the condition and operational profile of both vessels (in Table 3), we find that there was no breakdown reported in the month of May 2022 for both SB and PS engines of V1 vessel. Similarly, no breakdown was reported in the month of Feb 2022 for the SB and PS engines of V2. Therefore, it is highly probable that the engines were in good working conditions during these periods. As such, data in these months are selected to train autoencoders to model engines under normal or healthy operating conditions. Consequently, we use data in other months to validate the performance of the trained autoencoders.

TABLE III. SUMMARY OF BREAKDOWNS FOR BOTH VESSELS

Vessel	Engine	Jun 21	Jul 21	Aug 21	Sep 21	Oct 21	Nov 21	Dec 21	Jan 22	Feb 22	Mar 22	Apr 22	May 22	Jun 22	Jul 22
V1	SB	2	4	2	0	3	0	3	0	5	0	2	0	1	1
	PS	2	4	2	1	0	1	0	1	0	2	1	0	0	0
V2	SB	Vessel not in operation yet							0	0	0	3	0	1	0
	PS								1	0	4	1	1	0	2

Both training and testing or validation data of each vessel are normalized using min-max scaling from the training data.

C. Anomaly Detection & Diagnosis of Anomalies

Fig. 3 depicts the overall framework for detection and diagnosis of anomalies using an autoencoder network based on time series data from sensors of marine engines.

1) Autoencoder Modelling for Anomaly Detection

The autoencoder is trained on N time series samples of n engine sensors. The training samples are later fed into the trained autoencoder to reconstruct the healthy training data. For each training sample, we then take the reconstruction error (RE) which is the absolute difference of the reconstructed sample and the original sample. Based on the distribution of the RE of all samples in the training dataset, we extract the 99th percentile to determine the threshold τh to distinguish between normal and anomalous sample. The 99th percentile threshold is commonly used in the literature for using autoencoders to detect anomalies [13]. Since the autoencoder is trained on healthy data, its RE will be high for anomalous samples that do not come from the same distribution as the healthy ones. Therefore, if the RE of samples in the test data set exceeds the threshold τh, the sample is identified as anomaly. Otherwise, the test sample is considered healthy.

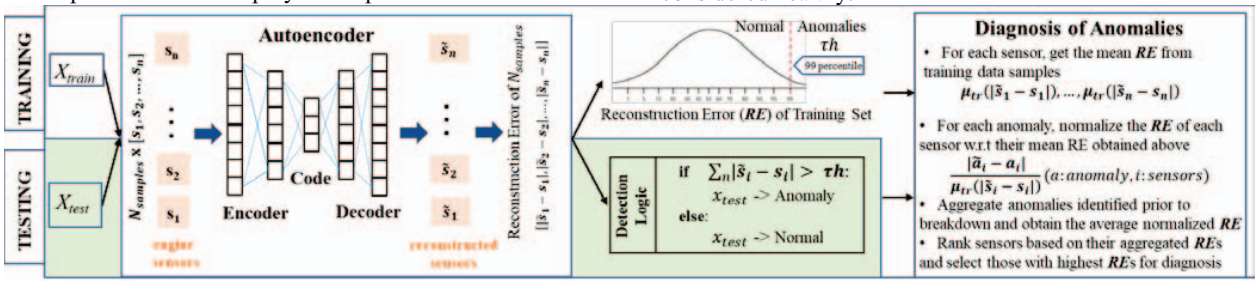

Fig. 3. Overall Framework for Detection and Diangosis of Anomalies

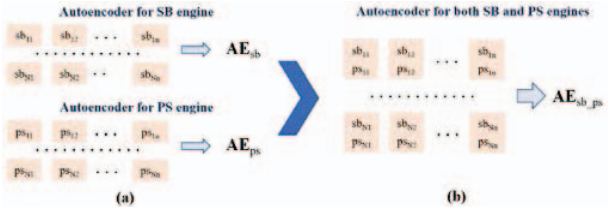

Fig. 4. Multiple Autoencoders (a) versus Single Autoencoder (b) Approach

We explore two different modes of training the autoencoder as shown in Fig 4. Each marine vessel has two independent engines SB and PS. Therefore, we train separate autoencoders for each engine of each vessel (Fig. 4a) in the first mode (multiple autoencoders). Specifically, individual autoencoders for the SB and PS engines are trained to detect anomalies in SB and PS engines, respectively. This approach is not effective as we need to train as many autoencoders as the type of engines. On the other hand, since both SB and PS engines have the same set of sensors sampled at the same frequency, we propose to use a single autoencoder to detect anomalies in both SB and PS engines (Fig. 4b) in the second mode (single autoencoder). To build a single autoencoder for both engines or multiple engines of the same vessel, we stack SB and PS sensor data consecutively. We report the performance results of both modes and compare their merits in the later Section.

2) Diagnosis of Anomalies

Detection of anomalies will provide early alarm of impending breakdowns to enable timely intervention or maintenance activity. However, it is also crucial to know the root cause indicator which is the potential cause or contributor (e.g. sensors) to the detected anomaly so that it expedites diagnosis before maintenance.

Algorithm 1 Algorithm for diagnosis of anomalies

Input: n sensors, A anomalies, N training samples, B breakdowns

Output: K sensors for each breakdown b

for each sensor $s_i (i = \overline{1, n})$ **do**:
  compute $\boldsymbol{\mu_{tr}}(|\tilde{\boldsymbol{s}}_i - \boldsymbol{s}_i|) = \sum_{j=1}^{N} |\tilde{\boldsymbol{s}}_{ij} - \boldsymbol{s}_{ij}|/N$
for each anomaly a **do**:
  normalize RE $\hat{a}_i = |\tilde{a}_i - a_i|/\boldsymbol{\mu_{tr}}(|\tilde{\boldsymbol{s}}_i - \boldsymbol{s}_i|)$
for each breakdown b **do**:
  identify anomalies belong to breakdown b
  for each day prior to the recorded breakdown date **do**:
    aggregate anomalies detected and average the normalized RE for each sensor $s_i (i = \overline{1, n})$
    rank sensors based on average normalized RE and select the top K sensors

The root cause indicators are estimated based on Algorithm 1. For each identified anomaly, we divide the RE of each sensor by its average RE from the training samples (or normalized reconstruction error). This step is to normalize each sensor's abnormality to their normal or healthy condition and give an indication of how abnormal the sensor is compared to its normal operating state. For each day prior to the breakdown, we aggregate all identified anomalies on that day and take the mean value of normalized reconstruction errors for each sensor. Then we plot the heatmap to show the abnormality of each sensor on each day prior to the breakdown. The top K sensors with brightest color (i.e. more abnormality) selected on each day will be identified as sensors attributed to the breakdown.

D. Performance Evaluation

Majority of previous works using machine learning models to detect anomalies or faults in marine main engines [7, 10, 11, 12, 14] only report the number of anomalies detected and/or cluster them into similar groups of breakdowns but lack of assessment on the relevance of detected anomalies against known breakdowns or ground truths. As such, no accuracy performance was reported in those works. As we have the approximate recorded time of occurrence for each breakdown, we evaluate the anomalies detected by the model and estimate their performance accuracy against the ground truth labels derived from the approximate time of occurrence from the maintenance reports. Thus, we report the True Positive (TP), False Negative (FN), and False Positive (FP) to evaluate each breakdown. We also report the Precision ($\frac{TP}{TP+FP}$), Recall ($\frac{TP}{TP+FN}$), and F1-measure ($\frac{2*Precision*Recall}{Precision+Recall}$) to assess the performance of the anomaly detection models. As described in Section IIB, the recorded time of breakdown may have discrepancies with the actual time of breakdown occurrence. Therefore, we propose to use a time-window parameter (δ days) to evaluate the model performance by limiting the duration within which to search for anomalies. That means only anomalies detected within δ-day before the recorded date are relevant to that breakdown. In case there is a previous breakdown within the δ-day window of current breakdown, only anomalies detected between the previous breakdown and current breakdown are considered related to the current breakdown. Fig. 5 illustrates the above idea using examples when the δ-day is set to 10. In Fig. 5a, the previous breakdown occurred on 15 Jun 2021 which is more than 10 days apart the current breakdown on 29 Jun 2021. Hence, if δ-day is set 10, only anomalies found from 19 Jun 2021 to 29 Jun 2021 are considered relevant to the breakdown on 29 Jun 2021. If there are anomalies found within that period (19 to 29 Jun 2021), it will be considered as a TP case. If no anomalies found within that period, it will be treated as a FN case. In Fig. 5b, since the previous breakdown (25 Jun 2021) was only 4 days apart the current breakdown (29 Jun 2021) which is shorter than the δ-day window of 10 days, hence only anomalies found from 25 to 29 Jun 2021 are considered relevant to the breakdown on 29 Jun 2021. Anomalies found beyond that period will be considered for other breakdowns or FP cases. For FP cases, if anomalies are detected within a period of δ-day but there is no reported breakdown within that period, it will be counted as a FP case. Using the δ-day parameter to assess each breakdown in the above manner will prevent inflating the number of TPs and reduce the possibility of deflating the number of FPs in reporting. In our experiments, we perform assessments with different values of δ-day and report the models' performance accordingly.

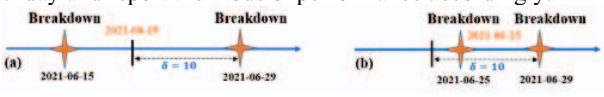

Fig. 5. Examples to Evaluate Model Performance

IV. EXPERIMENTAL RESULTS

A. Experimental Settings

TABLE IV. SUMMARY OF TRAINING AND TESTING DATA SIZE

Vessel	Total Samples	Training	Testing
V1	420,572	36446	384,126
V2	247,448	8438	239,010

TABLE V. AUTOENCODER ARCHITECTURE

Section	Layer	V1's Autoencoders		V2's Autoencoders	
		Shape	*Parameters*	*Shape*	*Parameters*
Input	Input	1x14	0	1x16	0
Encoder	Dense	1x64	960	1x64	1088
	Dense	1x32	2080	1x32	2080
	Dense	1x16	528	1x16	528
Code	Dense	1x8	136	1x8	136
Decoder	Dense	1x8	72	1x8	72
	Dense	1x32	288	1x32	288
	Dense	1x14	462	1x16	528

Table 4 summarizes the training and testing samples for both vessels. Table 5 summarizes the autoencoder architecture used in our experiments. For comparison purposes, the same architecture is applied for both multiple autoencoders approach and single autoencoder approach. We use Adam optimizer with the learning rate of 0.001 and mean absolute error as the loss function for training. All networks are trained for 3000 epochs with a batch size of 512. All experiments were conducted on the Ubuntu 20.04.6 LTS, Intel(R) Core(TM) i7-9700K CPU @ 3.60GHz, Nvidia GeForce RTX 2080 Ti, with the framework of Python 3.8.12.

For δ-day parameter, we take the days difference (delta) between 2 consecutive breakdowns of the same engine (SB or PS) for both vessels and plot their histogram in Fig. 6. It is observed that majority of breakdowns are apart by 46 days. Therefore, we evaluate the models' performance with 5 different δ-day values including 10, 15, 20, 30, and 45.

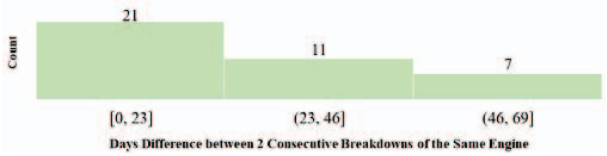

Fig. 6. Histogram of Days Difference between 2 Consecutive Breakdowns

Tables 6 and 7 summarize the performance results of single autoencoder and multiple autoencoder approaches for each vessel.

TABLE VI. PERFORMANCE RESULTS ON V1 VESSEL

Vessel Engine	δ-day	Total Cases	Single Autoencoder Approach						Multiple Autoencoder Approach					
			TP	*FN*	*FP*	*Pre (%)*	*Rec (%)*	*F1 (%)*	*TP*	*FN*	*FP*	*Pre (%)*	*Rec (%)*	*F1 (%)*
V1 SB	10	17	17	0.1	2	89.4	99.4	**94.1**	17	0.1	2	89.4	99.4	**94.1**
	15	17	17	0.1	1	94.4	99.4	**96.8**	17	0.1	1	94.4	99.4	**96.8**
	20	17	17	0.1	1	94.4	99.4	**96.8**	17	0.1	1	94.4	99.4	**96.8**
	30	17	17	0.1	1	94.4	99.4	**96.8**	17	0.1	1	94.4	99.4	**96.8**
	45	17	17	0.1	1	94.4	99.4	**96.8**	17	0.1	1	94.4	99.4	**96.8**
V1 PS	10	10	10	0	7	58.9	100	74.1	9.9	0.1	6.6	60.3	99	**74.8**
	15	10	10	0	5	66.8	100	80.1	9.9	0.1	4.7	68.1	99	**80.6**
	20	10	10	0	4.2	70.5	100	82.7	10	0	4.1	71	100	**83.0**
	30	10	10	0	3	77.1	100	87.0	10	0	2.8	78.4	100	**87.8**
	45	10	10	0	2.8	78.2	100	87.7	10	0	2.6	79.5	100	**88.5**

TABLE VII. PERFORMANCE RESULTS ON V2 VESSEL

Vessel Engine	δ-day	Total Cases	Single Autoencoder Approach						Multiple Autoencoder Approach					
			TP	*FN*	*FP*	*Pre (%)*	*Rec (%)*	*F1 (%)*	*TP*	*FN*	*FP*	*Pre (%)*	*Rec (%)*	*F1 (%)*
V2 SB	10	5	5	0	5	50.0	100	**66.7**	5	0	5	50.0	100	**66.7**
	15	5	5	0	3	62.5	100	**76.9**	5	0	3	62.5	100	**76.9**
	20	5	5	0	3	62.5	100	**76.9**	5	0	3	62.5	100	**76.9**
	30	5	5	0	2	71.4	100	**83.3**	5	0	2	71.4	100	**83.3**
	45	5	5	0	1	83.3	100	**90.9**	5	0	1	83.3	100	**90.9**
V2 PS	10	7	6.5	0.5	2.7	72.0	92.9	80.2	7	0	3	70.0	100	**82.4**
	15	7	6.5	0.5	1.8	79.2	92.9	84.8	7	0	2	77.8	100	**87.5**
	20	7	6.5	0.5	1.8	79.2	92.9	84.8	7	0	2	77.8	100	**87.5**
	30	7	6.5	0.5	0.9	88.2	92.9	90.0	7	0	1	87.5	100	**93.3**
	45	7	6.9	0.1	0.9	88.8	98.6	93.2	7	0	1	87.5	100	**93.3**

To avoid biasness and account for random weight initialization of neural networks, we perform 10 repetitions of training and testing for each approach and average the results from 10 repetitions for reporting. For example, of the 10 repetitions of experiments, only 1 experiment produces a FN case, then the average FN of 10 trials will be 0.1 (=1/10). We can see that the performance of single autoencoder approach is comparable to the one of multiple autoencoder approaches for both vessels. The multiple autoencoder approach achieves slightly better recall and F1 scores for detection of anomalies in PS engine. There is almost no change in the number of TPs as δ-day changes. This shows that the models can detect anomalies prior to reported breakdowns regardless of δ -day values experimented. Meanwhile, as δ-day value increases, the performance of both approaches also increases due to the decline in the number of FPs. This further substantiates the role of δ-day parameter in preventing deflating the number of FPs. Based on the results in Tables 6 and 7, we choose 20 as the optimal value of δ-day for reporting and analysis because it achieves the best TPs and balanced FPs for both engines of 2 vessels and more than half of breakdowns occurred within 23 days apart as shown in Fig. 6. On average, the precision, recall, and F1-measure of the single autoencoder approach are 76.6%, 98.1%, and 85.3%. Fig. 7 plots the number of anomalies detected for each repetition of 10 experiments. The right axes and orange curves in Fig. 7 show how early (i.e. days in advance) those anomalies were detected prior to engine breakdowns. On average, pre-cursors appear 13.3 and 11.5 days prior to breakdowns in engines of V1 and V2 respectively.

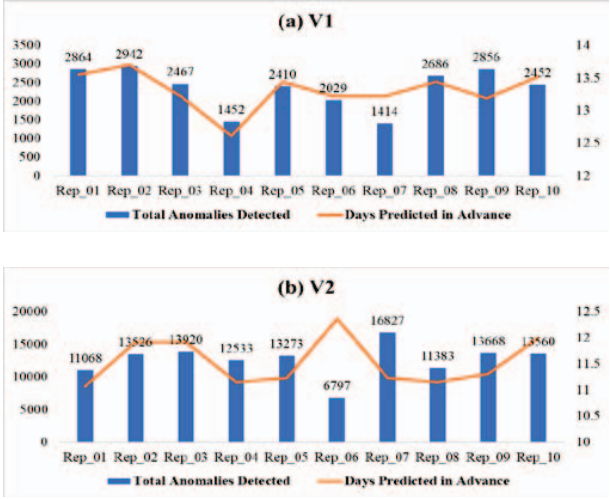

Fig. 7. Total Anomalies Detected and Days Predicted in Advance

B. Root Cause Prediction

Fig. 8 shows an example of the heatmap used to identify sensors that are detected as potential causes of the detected anomaly events. The x-axis represents the number of the days prior to the breakdown date which was 17 Mar 2022 for V1. The y-axis numbered from 0 to 13 represents the 14 sensors of V1 vessel PS engine which are shown in the legend of the plot. Based on the heatmap, FP and FOF sensors are the predominant contributors of most anomalies.

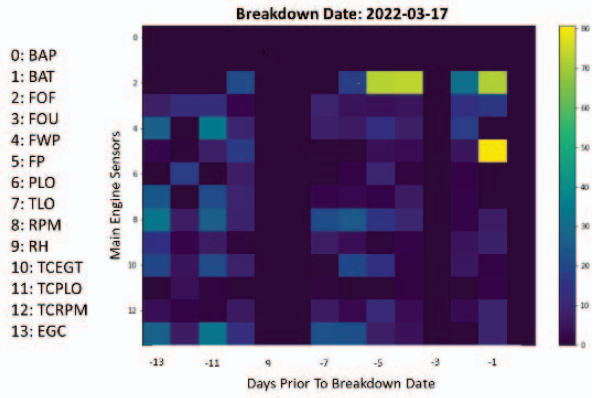

Fig. 8. Example of Heatmap Showing Sensors Attributed to Breakdown

We perform the same analysis for 27 breakdowns of V1 vessel and 12 breakdowns of V2 vessel and find out the top 5 sensors that commonly occur in our analysis and diagnosis. The top common sensors for breakdowns of V1 are FOF, EGC, RPM, FWP, and PLO while the top common sensors for breakdowns of V2 include TCPLO, FP, RPM, FOF, and PLO.

V. Conclusions

In this paper, we develop an autoencoder based anomaly detection method towards predictive maintenance of main engines in marine vessels. In addition, we also propose a method to evaluate the detection results objectively. Our experimental results show that the single autoencoder approach achieves comparable performance with the approach using multiple autoencoders for detection of anomalies prior to marine main engines' breakdowns. The developed autoencoders achieve 76.6%, 98.1%, and 85.3% of precision, recall, and F1 scores on average when testing on historical time-series data collected over a year of 2 vessels' operations. Specifically, the models achieve 82% to 96% of F1 score for all engines except the SB engine of V2 vessel. Additionally, we develop a method to identify sensors that are highly likely contributing to the breakdowns of marine main engines. The method enables users to visualize the trends of faulty sensors prior to breakdowns. Further verification is required to evaluate the correctness and usefulness of our method to aid components diagnosis of main engines in practice. Our immediate future work is to try improving the detection accuracy especially the one of SB engines by exploring diverse network architecture or different types of autoencoders such as LTSM autoencoders.

Acknowledgement

This research is supported by the National Research Foundation, Singapore, and the Maritime and Port Authority of Singapore / Singapore Maritime Institute under the Maritime Transformation Programme (Maritime Artificial Intelligence (AI) Research Programme – Grant number SMI-2022-MTP-06).

References

[1] C. Kandemir, M. Celik, "A human reliability assessment of marine auxiliary machinery maintenance operations under ship pms and maintenance 4.0 concepts," *Cognition, Technology & Work*, vol. 22, no. 3, pp. 473–487, 2020.

[2] T.M. Allen, "Us navy analysis of submarine maintenance data and the development of age and reliability profiles," *Department of the Navy SUBMEPP*, 2001.

[3] A.A. Jaber and R. Bicker, "The state of the art in research into the condition monitoring of industrial machinery," *Int. J. of Current Engineering and Technology*, vol. 4, no. 3, pp. 1986–2001, 2014.

[4] A. Brandsæter, G. Manno, E. Vanem, and I.K. Glad, "An application of sensor-based anomaly detection in the maritime industry," in *2016 IEEE international conference on prognostics and health management (ICPHM)*. IEEE, 2016, pp. 1–8.

[5] E. Vanem and G.O. Storvik, "Anomaly detection using dynamical linear models and sequential testing on a marine engine system," in *Annual Conference of the PHM Society*, 2017, vol. 9.

[6] E. Vanem and A. Brandsæter, "Cluster-based anomaly detection in condition monitoring of a marine engine system," in *2018 Prognostics and System Health Management Conference (PHM-Chongqing)*. IEEE, 2018, pp. 20–31.

[7] D. Kim, S. Lee, and J. Lee, "An ensemble-based approach to anomaly detection in marine engine sensor streams for efficient condition monitoring and analysis," *Sensors*, vol. 20, no. 24, pp. 7285, 2020.

[8] C. Qu, Z. Zhou, Z. Liu, and S. Jia, "Predictive anomaly detection for marine diesel engine based on echo state network and autoencoder," *Energy Reports*, vol. 8, pp. 998–1003, 2022.

[9] C. Fu, et al., "Comparative study on health monitoring of a marine engine using multivariate physics-based models and unsupervised data-driven models," *Machines*, vol. 11, no. 5, pp. 557, 2023.

[10] D. Kim, G. Antariksa, M. Handayani, S. Lee, and J. Lee, "Explainable anomaly detection framework for maritime main engine sensor data," *Sensors*, vol. 21, no. 15, pp. 5200, 2021.

[11] A. Brandsæter, E. Vanem, I.K. Glad, "Efficient on-line anomaly detection for ship systems in operation," *Expert Systems with Applications,* 2019, 121, 418–437.

[12] M. Cheliotis, I., Lazakis, G. Theotokatos, "Machine learning and data-driven fault detection for ship systems operations," *Ocean Enginering*, 2020, vol. 216, 107968

[13] G.R., Garcia, G. Michau, M. Ducoffe, J.S. Gupta JS, O. Fink, "Temporal signals to images: Monitoring the condition of industrial assets with deep learning image processing algorithms," *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*. 2022;236(4):617-627. doi:10.1177/1748006X21994446

[14] E. Vanem and A. Brandsæter, "Unsupervised anomaly detection based on clustering methods and sensor data on a marine diesel engine," *Journal of Marine Engineering & Technology*, 20.4 (2021): 217-234.