4th Asia Pacific Conference of the Prognostics and Health Management,
Tokyo, Japan, September 11 – 14, 2023 R03-02

# Unsupervised Anomaly Detection in Marine Diesel Engines using Transformer Neural Networks and Residual Analysis

Qin Liang[1, 2], Knut Erik Knutsen[1], Erik Vanem [1], Vilmar Æsøy [2], Houxiang Zhang [2]

[1] *Group Research and Development - DNV, Høvik, Norway*
*Qin.Liang@dnv.com*
*Knut.Erik.Knutsen@dnv.com*
*Erik.Vanem@dnv.com*

[2] *Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology, Ålesund, Norway*
*vilmar.aesoy@ntnu.no*
*hozh@ntnu.no*

## Abstract

This paper presents a novel unsupervised approach for detecting anomalies in marine diesel engines using a Transformer Neural Network based autoencoder (TAE) and residual analysis with Sequential Probability Ratio Test (SPRT) and Sum of Squares of Normalized Residuals (SSNR). The proposed method can capture temporal dependencies in normal time-series data without the need for labeled failure data. To assess the effectiveness of the proposed approach, a dataset of faulty data is generated under the same operational profile as the normal training data. The model is trained using normal data, and the faulty data is reconstructed using the trained model. SPRT and SSNR are then used to analyze the residuals from the observed and reconstructed faulty data, with significant deviations exceeding a predefined threshold being identified as anomalous behavior. The experimental results demonstrate that the proposed approach can accurately and efficiently detect anomalies in marine diesel engines. Therefore, this approach can be considered as a promising solution for early anomaly detection, leading to timely maintenance and repair, and preventing costly downtime.

## 1. Introduction

The maritime industry plays a critical role in global trade and transportation, over 80% of the volume of international trade in goods is carried by sea (Stalk, 2021). Ships and equipment onboard are the backbone of the operation of maritime industry. Onboard ship equipment maintenance is a critical aspect of ensuring safe and efficient vessel operations. The maintenance of the marine equipment, however, poses a significant challenge due to the remote and harsh environment of the sea. To ensure that equipment performs optimally, ship owners and operators need to have an effective maintenance strategy in place that balances safety, cost, and operational efficiency. In addition to the challenges of maintaining equipment, the development of autonomous ships adds another layer of complexity. Autonomous ships require higher safety and reliability standards for equipment, making the need for an effective maintenance strategy even more critical.

In this context, it is essential to continue to explore innovative and effective maintenance approaches to ensure the safe and reliable operation of marine equipment. The development of advanced data analytics, the Internet of Things (IoT) technologies, and machine learning algorithms holds great promise for improving onboard equipment maintenance and achieving optimal performance (Knutsen et al., 2022). Over the recent years, much has been implemented on these topics through two main approaches: data-driven approaches and model-based approaches (Bernardo & Reichard, 2017). Especially data-driven approaches applying Deep Learning (DL) techniques has become a popular direction with successful implementation in different domains. Neural networks are a type of machine learning algorithm that are modeled after the structure and function of the human brain. Deep learning, is a subset of machine learning, and neural networks make up the backbone of deep learning algorithms (Kriegeskorte & Golan, 2019). In fact, it is the number of node layers, or depth, of neural networks that distinguishes a single neural network from a deep learning algorithm, which normally have more than three.

Neural networks and deep learning models can be categorized based on their architecture and working mechanisms. There are various types of neural networks, such as Convolutional

Neural Networks (CNN), Recurrent Neural Networks (RNN), and Autoencoder (AE). On the other hand, new architectures are actively being developed by researchers. Over the last couple of years, there has been a surge in the development of large generative language models like ChatGPT. The key to their success lies in the Transformer architecture, which serves as the foundational pillar for these models.

This paper proposes and tests a new method for detecting anomalies in marine diesel engines using a Transformer Neural Network. The approach is unsupervised, meaning that the model is trained on normal operational data and tested on data from faulty operations. Once reconstruction is complete, human domain knowledge, SPRT and SSNR are used to evaluate the model's performance.

The remainder of the paper is organized as follows: Section II discusses the latest trends in data-driven equipment anomaly detection. Section III presents the methodology used in this study. Section IV covers the model and data used in the study, including the collection process and model training. Results and analysis are presented in Section V. Finally, Section VI concludes the paper and proposes future work.

## 2. RELATED STUDIES

In recent years, data-driven prognostics has gained significant attention as a research area. (Vanem & Brandsæter, 2021) presents a comprehensive implementation of cluster-based anomaly detection for marine engine systems. This study highlights the potential of employing statistical techniques for effective anomaly detection. Concurrently, numerous researchers are also exploring the application of DL methods in this domain, i.e., (Han, Li, Skulstad, Skjong, & Zhang, 2020), (Ellefsen, Bjørlykhaug, Æsøy, Ushakov, & Zhang, 2019) and (Hu, Cheng, Wu, Zhu, & Shao, 2021). An autoencoder (AE) is a type of artificial neural network that is used for unsupervised learning of efficient data representations. Because of the mechanism, AEs can be used for anomaly detection, where they are trained on healthy data and then used to detect any deviations from the normal behavior of the machinery or system. The key advantage is that AEs adopt an unsupervised learning architecture, they do not require large amounts of data to be labeled. (Han, Ellefsen, Li, Holmeset, & Zhang, 2021) proposed an LSTM-based variational autoencoder (LSTM-VAE) for fault detection in maritime components. (Listou Ellefsen et al., 2020) proposed a fault-type independent spectral anomaly detection algorithm for marine diesel engine degradation based on variational autoencoder (VAE). (Hemmer, Klausen, Khang, Robbersmyr, & Waag, 2020) introduced an unsupervised learning approach for detecting defects in large, slow-rotating axial bearings by developing a Health Indicator (HI). The proposed method utilizes variational inference and involves the use of a VAE and a conditional variational autoencoder (CVAE).

RNN are a type of neural network that are designed to handle sequential data. Unlike feed-forward neural networks like Multilayer perceptron (MLP), which process input data in a fixed order and don't have any memory (Liang, Tvete, & Brinks, 2019) and (Liang, Tvete, & Brinks, 2020), RNN maintain an internal state that allows them to process sequences of varying lengths and capture the temporal dependencies between successive inputs. (Hu et al., 2021) introduced a new deep bidirectional recurrent neural networks (DBRNNs) ensemble method for Remaining Useful Life (RUL) prediction of aircraft engines. CNN has shown promising results in detecting faults based on acoustic signals, vibration data, and thermal images. For example, (Massoudi, Verma, & Jain, 2021) used CNN to classify engine sounds based on the type and severity of the fault.

After conducting a literature survey, it appears that the use of Transformer Neural Network for prognostics or anomaly detection purposes is not widely explored. To address this gap, (Zhang, Song, & Li, 2022) proposed a new deep method for RUL prediction called Dual-Aspect Self-Attention based on Transformer (DAST) to improve the overall efficiency of predictive maintenance tasks. The results demonstrated that DAST outperforms BiLSTM and CNN methods in terms of RMSE (Root Mean Squared Error) and score values for most engines. It is important to note that DAST is a supervised learning approach that requires labeled data for training. However, in reality, obtaining sufficient fault or RUL data can be challenging, which may limit the performance of the model. In another transformer related study, (Tuli, Casale, & Jennings, 2022) introduced TranAD, a deep transformer network for efficient and accurate anomaly detection and diagnosis in multivariate timeseries data. TranAD outperforms state-of-the-art baseline methods in both detection and diagnosis performance while offering data and time-efficient training. The paper uses seven publicly available datasets in their experiments. The authors acknowledge some concerns about the lack of quality benchmark datasets for time series anomaly detection.

## 3. METHODOLOGY

### 3.1. Proposed TAE

Transformer is a deep learning model introduced by (Vaswani et al., 2017). in 2017, and it has gained significant popularity in the field of natural language processing. Transformer Networks were introduced as an alternative to RNNs for sequence modeling tasks. Unlike RNNs, Transformer Neural Networks (TNN) have a parallelizable architecture, making them faster for certain tasks. They also require fewer training iterations and are less prone to the vanishing gradient problem than RNNs. Although RNNs have been widely used for sequence modeling, their limitations have led to the development of TNN. TNN have shown superior performance in several natural language processing tasks, including language translation, compared to RNNs. They are faster, require less training data, and have a better ability to handle long sequence dependencies (Karita et al., 2019). RNNs use recurrent connections to process sequential data, while TNNs rely on self-attention mechanisms to capture dependencies between all elements in a sequence, without using any connections between the elements themselves. One key challenge in applying self-attention to sequential data is that the order of the elements in the sequence is lost when computing the attention weights. This is because the attention mechanism computes the attention weights based on the similarity between the query vector and the keys associated with each element in the sequence, regardless of their position. This makes it challenging for the model to differentiate between elements at different positions in the sequence. To address this issue, the TNN introduces positional encoding. Positional encoding is a technique that adds a fixed positional vector to the input embeddings, providing the model with information about the position of each element in the sequence. The purpose of this is to provide the model with positional information, allowing it to distinguish between different elements in the sequence.

Instead of using TNN for prediction, in this study TNN was used in an autoencoder manner to reconstruct the data. The proposed architecture can be seen in 1. The details of the proposed architecture are illustrated as follows.

1. Feed-forward Neural Network: An FNN is a fundamental type of artificial neural network that can be used as a building block for constructing more complex models such as MLPs. It is characterized by its fully connected structure, where each unit in one layer is directly connected to all units in the subsequent layer via weight connections.
2. Positional encoding is a way of incorporating position information into the input embeddings by adding a fixed vector to each embedding, which varies based on its position in the sequence. There are various ways of positional encoding methods to choose. In this paper, the method from (Vaswani et al., 2017) is used.
3. Residual connection and layer normalization layer (Add & Norm): This layer is added after each sublayer in TNN encoder. The function of residual connections is to ease the challenge of training deep neural networks. Meanwhile, layer normalization can quicken the training progress and promote faster convergence of the model by normalizing the activation value of each layer.
4. Multihead self-attention layer: The encoder employs multihead self-attention to extract the significance of various sensors along the sensor dimension, enabling it to autonomously learn to prioritize characteristics with higher weights. As a consequence, there is no need for human intervention during the training process, resulting in an automated and efficient feature selection process.
5. Layer normalization and residual connections: Each sublayer in the TNN, including multi-head self-attention and position-wise feed-forward networks, is surrounded by residual connections and followed by layer normalization. Residual connections allow the output of each sublayer to be added to its input, which helps to mitigate the vanishing gradient problem by allowing gradients to flow directly to earlier layers. Layer normalization, on the other hand, is a technique that is used to normalize the activations of each layer. This helps to stabilize the learning process by reducing the internal covariate shift.

   In this study, the proposed TAE used two identical layers, with each layer consisting of four attention heads. Rather than using the TNN architecture again for decoding, an MLP architecture is utilized to reconstruct the original input data. It is important to note that in this study, the Transformer model is used in an AE manner. This means that the decoder does not receive any shifted features as inputs, but rather only the learned representations from the encoder. This approach reduces the complexity of the model and makes the reconstruction process more efficient.

### 3.2. Sequential Probability Ratio Test

SPRT is a statistical method (Wald, 1992) used to make decisions about a hypothesis based on a sequential analysis of data. The method involves taking samples of data sequentially and updating the probability of a hypothesis after each sample is taken. (Vanem & Storvik, 2017) , (Brandsæter, Vanem, & Glad, 2019) and (Brandsæter, Manno, Vanem, & Glad, 2016) have already explored the application of SPRT on maritime equipment and proved its capability of detecting anomalies.

The trained model from introduced previously provides a reconstruction $x_t^{'}$ of the observed signal values $x_t$ at each time step $t$. The residuals, i.e. the difference between the reconstructed and the observed value $r_t = x_t^{'} - x_t$ are analyzed sequentially by the SPRT to determine if the signal indicates a normal or anomalous state of the system. To employ SPRT for analyzing residual data, it is necessary to define two competing hypotheses: a null hypothesis $H_0$ and an alternative hypothesis $H_1$. Typically, the null hypothesis asserts that residuals are normally distributed with mean 0 and some standard deviation $\sigma$ which present is the system is in normal state. While the alternative hypothesis $H_1$ which assumes that the residuals are normally distributed with specific mean $\mu$ and/or standard deviation $\sigma'$ different from the null hypothesis if the system is in the anomalous state. The SPRT is performed for each feature independently.

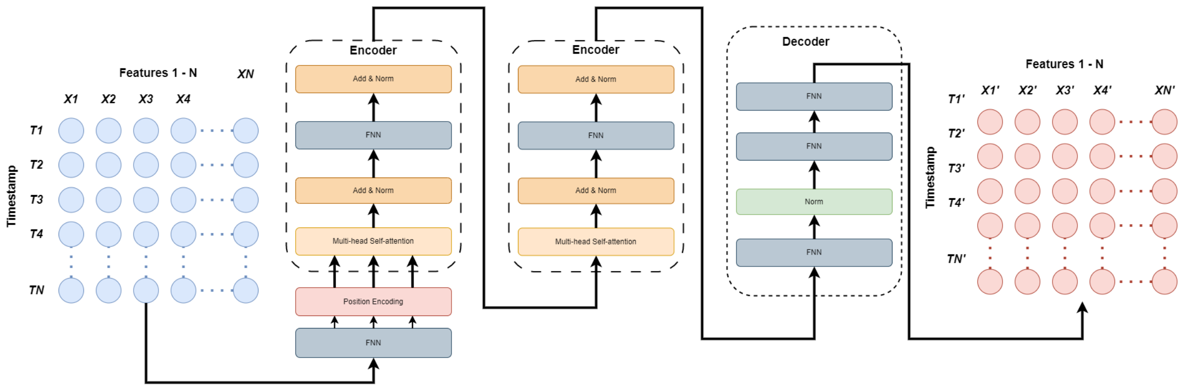



Figure 1. Proposed architecture

$$
\begin{aligned}
H_0 &: x \sim N(0, \sigma) \\
H_1 &: x \sim N(\mu, \sigma')
\end{aligned} \tag{1}
$$

The SPRT can be calculated in the following steps. It is assumed both follow normal distribution, the normal distribution probability function is:

$$
f(r) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(r-\mu)^2}{2\sigma^2}\right) \tag{2}
$$

where $r$ is the residuals of reconstructed and observed signal value. Then, the likelihood ratio can be calculated based on $L_1$ and $L_0$ when $H_0$ has a mean of 0. If $L_1$ is greater than $L_0$, it indicates that the distribution aligns more closely with $H_1$ than with $H_0$, and vice versa. The log likelihood ratio can be calculated as:

$$
log\frac{L_1}{L_0} = log(\prod_{i=1}^{n} \exp\left[(r_i - \frac{\mu}{2})\frac{\mu}{\sigma^2}\right]) \tag{3}
$$

$$
= \sum_{i=1}^{n} (r_i - \frac{\mu}{2})\frac{\mu}{\sigma^2} \tag{4}
$$

In this case, $r_i$ represents the residuals at each time step $i$. Once the hypotheses are defined and the log likelihood is calculated, the SPRT index can be sequentially calculated and updated. To achieve this, two threshold values, $A$ and $B$, must be specified. The calculated SPRT index at each time step is then compared with these lower and upper decision boundaries. At each time step, three possible outcomes can occur:

- If the value falls below the lower limit ($A$), it indicates the acceptance of the normal state ($H_0$). Consequently, the test statistic is reset.
- If the value exceeds the upper limit ($B$), it suggests the acceptance of the anomalous state ($H_1$). Accordingly, the test statistic is reset.
- When the value lies between the defined threshold values, it signifies an insufficiency of available information to reach a conclusive decision.

The thresholds $A$ and $B$ can be calculated based on the following equations:

$$
\begin{aligned}
A &= \log\left(\frac{\beta}{1-\alpha}\right) \\
B &= \log\left(\frac{1-\beta}{\alpha}\right)
\end{aligned} \tag{5}
$$

where $\alpha$ is the probability of Type I error (false alarm), which represents the probability of rejecting the true $H_0$. $\beta$ is the probability of Type II error (missed alarm), which represents the probability not rejecting $H_0$ when it is false.

### 3.3. Sum of Squares of Normalized Residuals

The chi-square distribution, also written as $\chi^2$ distribution, is a continuous probability distribution widely used in statistical inference and hypothesis testing. It is particularly relevant in scenarios where the sum of squared independent, identically distributed random variables is being analyzed. The chi-square distribution is a special case of the gamma distribution and is often used in goodness-of-fit tests, independence tests for contingency tables, and the estimation of confidence intervals.

The chi-square distribution is characterized by its degrees of freedom, which determine the shape of the distribution. The degrees of freedom are typically related to the number of independent observations or constraints in a given problem. Specifically, the sum of the squares of $k$ independent standard normal distribution variables follows a chi-square distribution with $k$ degrees of freedom. This concept forms the basis of the SSNR. In this study, the assumption is made that the SSNR follows a chi-square distribution with $k$ degrees of freedom equal to the number of features, which is 17.

To assess the significance of the SSNR, a comparison is made with the corresponding chi-squared distribution. This enables the determination of the probability of observing an SSNR value as large or larger than a defined threshold. The threshold for hypothesis testing can be derived using the inverse cumulative distribution function (CDF), which allows the mapping of probabilities back to values from the distribution. Three confidence levels are selected in this study: 99.99%, 99.7%, and 95% for evaluation. The confidence level represents the threshold probability. For instance, a confidence level of 99.7% implies a 0.3% chance of making a false alarm. By considering the given confidence level and degrees of freedom (equal to the number of features), threshold values of 47.56, 37.37, and 27.58 are obtained using the inverse CDF. The selection of a 99.7% confidence level is guided by the three-sigma rule (Pukelsheim, 1994), which serves as a widely recognized benchmark. However, in practical applications, the confidence level can be adjusted to meet specific requirements. Different applications may exhibit varying degrees of sensitivity to inaccuracies in reconstructed signals. A higher confidence level of 99.99% is also selected based on the condition of this study.

By applying the threshold to the SSNR, the reconstruction error can be effectively monitored. The SSNR is defined as follows:

$$SSNR = \sum_{i=1}^{d_i} \left( \frac{r_i - \mu_o}{\sigma_o} \right)^2 \tag{6}$$

where $d_i$ is the number of features, $r_i$ is the residuals of reconstructed faulty data, $\mu_o$ is the mean of residuals of reconstructed normal data, $\sigma_o$ is the standard deviation of residuals of reconstructed normal. Further details regarding the test results will be discussed in the following sections.

## 4. EXPERIMENTAL STUDY

### 4.1. Data collection and processing

The Department of Ocean Operations and Civil Engineering at the Norwegian University of Science and Technology in Ålesund has established a hybrid power lab for the purpose of data collection. The laboratory consists of a compact marine diesel engine integrated with a generator, a marine battery system, a marine DC switchboard equipped with essential power converters, and a comprehensive marine automation system that supervises the entire operational process.

The data collection process involves running the engine on an operating profile that emulates an actual ferry crossing on the west coast of Norway. The ferry departs from shore at a safe and constant speed, then accelerates until it reaches a suitable speed. The speed is maintained at a constant level before safely decreasing and finally braking just before docking. The entire ferry crossing process takes 20 minutes, and the complete engine operating profile is depicted in Figure 2. Both the normal operation data and the faulty degradation data are collected while running the engine operating profile. The only difference between the two data sets is that a fault is introduced at an unknown time step in the faulty degradation data. Therefore, the primary objective is to predict the fault time step on time.

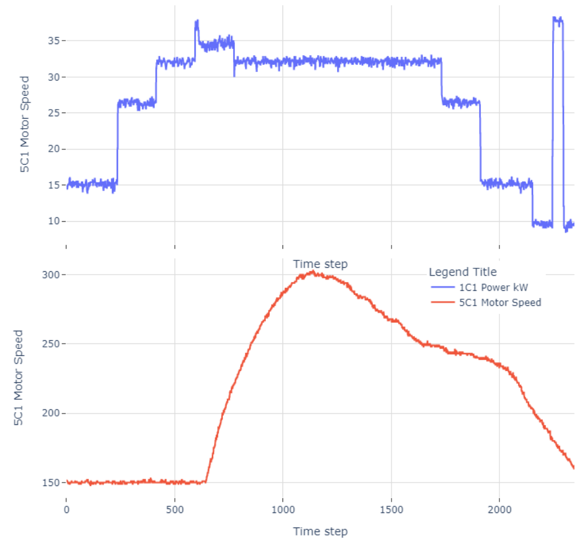



Figure 2. Engine operation profile

The engine is equipped with two water cooling systems - a primary and a secondary system, where the latter cools the former. The primary cooling system is regulated by an internal bimetal thermostatic valve, which commences opening at a temperature of 78°C and reaches full opening at 90°C. On the other hand, the secondary cooling system relies on a frequency operated fan that circulates air through a heat exchanger. A malfunction of the fan is intentionally introduced to create a fault that subsequently leads to a decline in cooling efficiency within the secondary cooling system. To prevent potential issues, the system is equipped with an alarm that activates when the cooling water temperature exceeds 85°C. A total of 2336 time steps were recorded over a 1168-second period, with a frequency of 2Hz.

The process of feature selection is a critical step that can significantly impact the performance of a study. In this study, the feature selection process was initiated by employing the domain knowledge and expertise of the engine operator to select 21 input features. To further refine the feature selection, a correlation matrix was utilized to identify and remove highly correlated features. The pairs of highly correlated features were identified by comparing each column of the upper triangular matrix against a threshold value of 0.95. The highly correlated features were dropped from the dataset to reduce redundancy and potentially enhance the performance. After this step, 17 features are kept. In the present study, the objective is to investigate the potential of the Transformer algorithm for time series reconstruction. As such, no further advanced feature selection methods, such as dimensionality reduction techniques are adopted.

The data used for training has undergone zero-mean and unit variance normalization. This technique involves scaling the features to have a zero mean and unit variance, which ensures that all features are on a comparable scale and prevents the dominance of features with large variances during the learning process. Moreover, such normalization can improve the stability and performance of machine learning models during training. It is worth noting that the normalization statistics derived from the normal operation data are also utilized for the faulty degradation data.

The reconstruction of the data is a time series problem. When dealing with time series data, it is often useful to divide it into smaller subsequences or sequences that have a fixed length. This can be done using a sliding window approach where a window of fixed length is moved across the time series data at a fixed stride. At each window position, the subsequence of data within the window is extracted and added to a list of subsequences.

### 4.2. Model training

In the proposed architecture for time series reconstruction using the TNN, several key hyper-parameters must be defined to ensure its effective implementation. The number of TNN layers refers to the number of encoding and decoding layers in the architecture. Increasing the number of layers can improve the model's ability to capture complex temporal patterns but may also increase the risk of overfitting. The number of heads is the number of parallel attention mechanisms that are applied in each encoder and decoder layer. Increasing the number of heads can enhance the model's ability to attend to multiple parts of the input sequence simultaneously. The time sequence length is the length of the input time series sequence that is fed into the model. This parameter determines how much historical data the model can use to make its predictions. The dimension of the feedforward network (FNN) is the size of the hidden layer in the FNN component of the TNN. Increasing the dimension of the FNN can allow the model to learn more complex relationships between features but may also increase the risk of overfitting. The selected hyper-parameters are summarized in Table 1. In addition to that, early stopping technique is involved in the training process. Early stopping is a regularization technique used in machine learning to prevent overfitting of the training data.

Table 1. Selected hyper-parameters

| **Hyper-parameter** | **Value** |
|---|---|
| Learning rate | 0.001 |
| Batch size | 32 |
| Time sequence length | 10 |
| Number of Transformer layers | 2 |
| Number of heads | 4 |
| Number of epochs | 100 |
| Number of FNN layers | 3 |

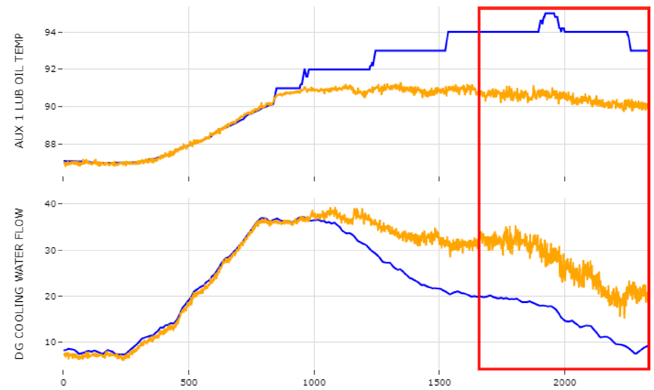



Figure 3. Observed vs Reconstructed data

## 5. RESULT ANALYSIS

As previously mentioned, the system triggers an alarm if the cooling water temperature exceeds 85°C. In this particular case, a cooling fault occurred due to a malfunction in the fan at the beginning of the test, resulting in a reduction in cooling efficiency in the secondary cooling system. The evaluation of the trained model is carried out by analyzing both the observed and reconstructed data. Figure 3 displays several features of the observed (blue) and reconstructed (orange) data, with the red square indicating the time-step (1658) at which the system raised an alarm, which persisted until the end of the test.

In the paper, the evaluation of the model focuses on its ability to detect anomalies during the anomalies period and prior to the activation of the actual alarm. This prognostic information can provide valuable insights to operators in real-world scenarios, enabling them to take proactive measures.

### 5.1. Evaluation on SPRT

As introduced previously, the SPRT is computed using the log likelihood and two thresholds. The SPRT index is reset whenever it surpasses both thresholds. In this study, the normal data set is referred to as the training data, while the faulty data is referred to as the test data. In this study, the standardization of residuals from the reconstructed test data plays a crucial role. The residuals of the reconstructed test data are standardized using the mean and standard deviation derived from the first 500 time steps of the same data. It is important

Table 2. Anomalous time step comparison - SPRT

| Feature | Negative mean change | Positive mean change |
|---|---|---|
| Feature 1 | 708 | N/A |
| Feature 2 | 1043 | 1786 |
| Feature 3 | 910 | 678 |
| Feature 4 | 846 | 522 |
| Feature 5 | 1025 | 1162 |
| Feature 6 | N/A | 714 |
| Feature 7 | 658 | 1339 |
| Feature 8 | 1400 | 603 |
| Feature 9 | 605 | 2239 |
| Feature 10 | 630 | 972 |
| Feature 11 | 1314 | 686 |
| Feature 12 | 1503 | 1009 |
| Feature 13 | 611 | 1016 |
| Feature 14 | 1143 | 600 |
| Feature 15 | 681 | 759 |
| Feature 16 | 955 | 2276 |
| Feature 17 | 1204 | 990 |
| Average | 1116 | 1084 |

to note that during these initial 500 time steps, the system is in normal operation mode, with all signals falling within their respective normal working ranges. This standardization approach ensures a stable performance of the SPRT.

In this study, two alternative hypotheses are examined: deviations in the positive and negative directions of the mean. The three-sigma rule is applied in this step, with the alternative means of -4 and 4 being utilized in the tests. The results of the negative test are presented in Figure 4 as an example.

For the anomalies detected in the first 500 time steps are taken as faulty warning. It is worth noting that the SPRT indices are calculated for all features, with some features indicating errors earlier than others. The average of the identified errors across all features serves as the final measure for evaluating the performance of the models. The detailed anomalous time steps identified by both tests are presented in Table 2. The average number of detected anomalous time steps from both tests is 1116 and 1084, respectively. These averages are compared with the time step at which the system alarms are activated, demonstrating that the detections are timely and accurate.

### 5.2. Evaluation on Sum of Squares of Normalized Residuals

In addition to assessing the models' performance using SPRT, the performance is also evaluated using SSNR, as previously introduced. To implement SSNR, the test data residuals undergo standardization, adopting the same approach used in SPRT. The mean and standard deviation of the test data residuals from the first 500 time steps are utilized for standardization. A time step is considered a potential warning if its SSNR value exceeds the average. The SSNR performance is illustrated in Figure 5, where it is evident that the SSNR can detect anomalies well before the system alarm is activated. Figure

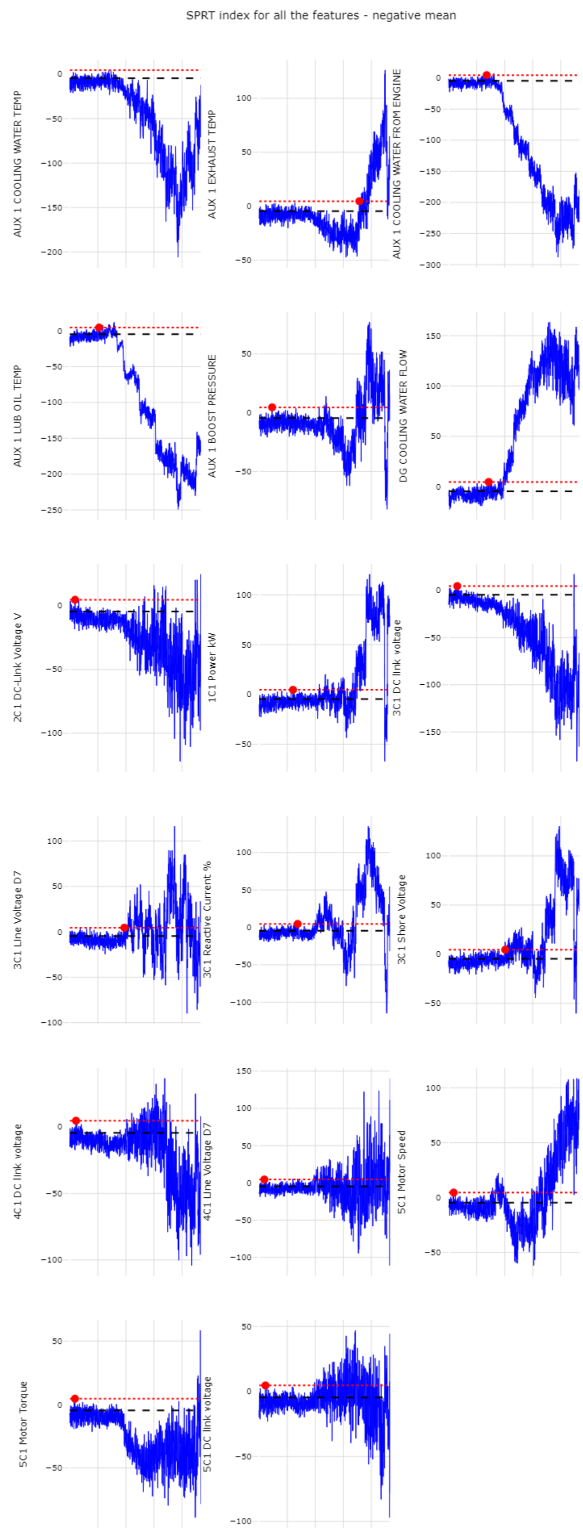



Figure 4. SPRT on test data - negative mean change

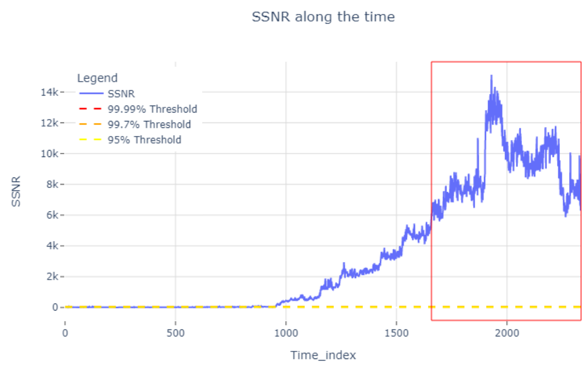



Figure 5. SSNR with full scale

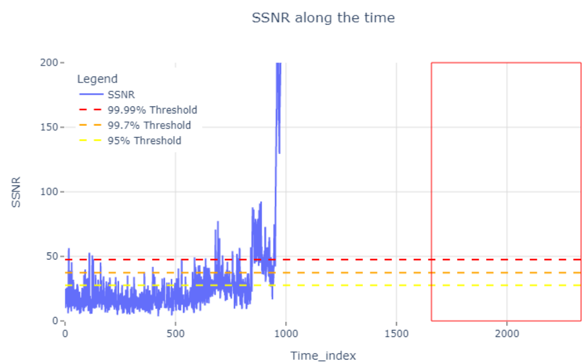



Figure 6. SSNR with selected scale

6 provides a closer examination of the SSNR performance. Similarly, anomalies detected within the first 500 time steps are considered as faulty warnings. The SSNR frequently indicates faults between time steps 850 and 950. Comparing with the system alarm time step, the SSNR demonstrates efficient and timely anomaly detection.

The results obtained from the SSNR analysis highlight the early anomaly detection capability of the TAE model. It consistently identifies anomalies between time steps 850 and 950, triggering timely warnings. In comparison to the system alarm at time step 1658, the SSNR exhibits efficient and timely detection of anomalies. These findings suggest that in real-world scenarios, operators would have ample time to implement preventive measures and prevent the system alarm from activating."

## 6. CONCLUSION AND FUTURE WORK

This paper proposes a novel approach to detect anomalies in marine diesel engines. The approach consists of two parts: a Transformer-based autoencoder and residuals evaluation methods based on SPRT and SSNR. The data used in this study consists of normal and faulty data collected under the same operation profile. The normal data is used to train the model, while the faulty data is used to test it. In this paper, a detailed explanation of the proposed architecture is presented, along with an outline of the optimal combination of hyperparameters. The evaluation of the observed and reconstructed data shows that the proposed TAE demonstrates stable performance and can detect anomalies in a timely manner.

In this study, it was found that the potential of TNN is not fully utilized. Firstly, the amount of training data is limited to only 20 minutes of operation. Although this is sufficient to demonstrate the potential of the TNN model, it is likely that the model's performance could be further improved with larger training datasets. One of the major advantages of TNN over RNN is its ability to be trained in parallel, which could be exploited with larger datasets. In addition, longer time sequences could also be explored to see the attention mechanism against the memory of LSTM. Secondly, the data used in this study only contained one operation profile, which may not be representative of the full range of working conditions encountered in marine diesel engines. Thirdly, the study did not explore other feature selection and dimensionality reduction techniques, which could further enhance the performance of the TAE. Additional techniques such as principal component analysis or independent component analysis could be used to reduce the dimensionality of the data, while preserving the most important features. Lastly, the SPRT and SSNR methods used in this study could be further explored to improve its effectiveness. Alternative statistical methods or modifications to the existing methods could be investigated to improve its accuracy and reliability in detecting anomalies in marine diesel engines.

## ACKNOWLEDGMENT

This work has partly been carried out within the RealTOPs project, supported by DNV and the Research Council of Norway. The authors report there are no competing interests to declare.

## NOMENCLATURE

| | |
|---|---|
| $TAE$ | Transformer neural network base autoencoder |
| $SPRT$ | Sequential Probability Ratio Test |
| $SSNR$ | Sum of Squares of Normalized Residuals |
| $CNN$ | Convolutional Neural Networks |
| $AE$ | Autoencoder |
| $RNN$ | Recurrent Neural Networks |
| $TNN$ | Transformer Neural Networks |
| $LSTM$ | Long Short-Term Memory |

## REFERENCES

Bernardo, J. T., & Reichard, K. M. (2017). Trends in research techniques of prognostics for gas turbines and diesel engines. In *Annual conference of the phm society* (Vol. 9).

Brandsæter, A., Manno, G., Vanem, E., & Glad, I. K. (2016). An application of sensor-based anomaly detection in the maritime industry. In *2016 ieee international conference on prognostics and health management (icphm)* (pp. 1–8).

Brandsæter, A., Vanem, E., & Glad, I. K. (2019). Efficient online anomaly detection for ship systems in operation. *Expert Systems with Applications*, *121*, 418–437.

Ellefsen, A. L., Bjørlykhaug, E., Æsøy, V., Ushakov, S., & Zhang, H. (2019). Remaining useful life predictions for turbofan engine degradation using semi-supervised deep architecture. *Reliability Engineering & System Safety*, *183*, 240–251.

Han, P., Ellefsen, A. L., Li, G., Holmeset, F. T., & Zhang, H. (2021). Fault detection with lstm-based variational autoencoder for maritime components. *IEEE Sensors Journal*, *21*(19), 21903-21912. doi: 10.1109/JSEN.2021.3105226

Han, P., Li, G., Skulstad, R., Skjong, S., & Zhang, H. (2020). A deep learning approach to detect and isolate thruster failures for dynamically positioned vessels using motion data. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–11.

Hemmer, M., Klausen, A., Khang, H. V., Robbersmyr, K. G., & Waag, T. I. (2020). Health indicator for low-speed axial bearings using variational autoencoders. *IEEE Access*, *8*, 35842-35852. doi: 10.1109/ACCESS.2020.2974942

Hu, K., Cheng, Y., Wu, J., Zhu, H., & Shao, X. (2021). Deep bidirectional recurrent neural networks ensemble for remaining useful life prediction of aircraft engine. *IEEE Transactions on Cybernetics*.

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., ... others (2019). A comparative study on transformer vs rnn in speech applications. In *2019 ieee automatic speech recognition and understanding workshop (asru)* (pp. 449–456).

Knutsen, K. E., Liang, Q., Karandikar, N., Ibrahim, I. H. B., Tong, X. G. T., & Tam, J. J. H. (2022). Containerized immutable maritime data sharing utilizing distributed ledger technologies. In *Journal of physics: Conference series* (Vol. 2311, p. 012006).

Kriegeskorte, N., & Golan, T. (2019). Neural network models and deep learning. *Current Biology*, *29*(7), R231–R236.

Liang, Q., Tvete, H., & Brinks, H. (2020). Prediction of vessel propulsion power from machine learning models based on synchronized ais-, ship performance measurements and ecmwf weather data. In *Iop conference series: Materials science and engineering* (Vol. 929, p. 012012).

Liang, Q., Tvete, H. A., & Brinks, H. W. (2019). Prediction of vessel propulsion power using machine learning on ais data, ship performance measurements and weather data. In *Journal of physics: Conference series* (Vol. 1357, p. 012038).

Listou Ellefsen, A., Han, P., Cheng, X., Holmeset, F. T., Æsøy, V., & Zhang, H. (2020). Online fault detection in autonomous ferries: Using fault-type independent spectral anomaly detection. *IEEE Transactions on Instrumentation and Measurement*, *69*(10), 8216-8225. doi: 10.1109/TIM.2020.2994012

Massoudi, M., Verma, S., & Jain, R. (2021). Urban sound classification using cnn. In *2021 6th international conference on inventive computation technologies (icict)* (pp. 583–589).

Pukelsheim, F. (1994). The three sigma rule. *The American Statistician*, *48*(2), 88–91.

Stalk, P. (2021). Review of maritime transport. *PUNITED NATIONS CONFERENCE ON TRADE AND DEVELOPMENT*, 1-177.

Tuli, S., Casale, G., & Jennings, N. R. (2022). Tranad: Deep transformer networks for anomaly detection in multivariate time series data. *arXiv preprint arXiv:2201.07284*.

Vanem, E., & Brandsæter, A. (2021). Unsupervised anomaly detection based on clustering methods and sensor data on a marine diesel engine. *Journal of Marine Engineering & Technology*, *20*(4), 217–234.

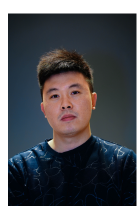
Vanem, E., & Storvik, G. O. (2017). Anomaly detection using dynamical linear models and sequential testing on a marine engine system. In *Annual conference of the phm society* (Vol. 9).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wald, A. (1992). *Sequential tests of statistical hypotheses*. Springer.

Zhang, Z., Song, W., & Li, Q. (2022). Dual-aspect self-attention based on transformer for remaining useful life prediction. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1-11. doi: 10.1109/TIM.2022.3160561

## BIOGRAPHIES



**Qin Liang** works as a Senior Researcher in Group Research and Development - Maritime programe DNV. In additon, he is currently pursuing the Ph.D. degree with the Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology (NTNU). His current research interests include ship performance, equipment condition monitoring, machine learning and deep learning.

**Knut Erik Knutsen** is a Principal Researcher at DNV – Group Research and Development – Maritime Transport and team lead for the Data Driven Services group. He holds a PhD in Semiconductor Physics (2013) and a MSc in Materials, Energy and Nanotechnology (2009) from the University of Oslo. He also has experience as an Avionics Technician from the Royal Norwegian Airforce (1999-2004) where he was performing maintenance and calibration of aircraft equipment. Currently his research focuses on Data Driven Services and in particular data integrity solutions on the edge to increase the level of trust in data driven applications, and further enable novel digital solutions to support DNVs purpose of safeguarding life, property and the environment.

**Erik Vanem** received the Cand. Scient. degree (Master of Science equivalent) in physics and the Ph.D. degree in statistics from the University of Oslo in 1996 and 2012, respectively. He worked three years at the Research Department of Telenor, three years at PGS Reservoir, one year at the Oslo University College, and spent some time at the Norwegian Defence Research Establishment. He has been working at legacy DNV R&I since 2003 on a number of research projects related to maritime safety and risk assessment. Since 2016, he has also been an Associate Professor with the University of Oslo in a 20% position. He is currently working as a Principal Researcher with the Maritime Transport Group, DNV Group Research and Development, Høvik, Norway. As a Researcher, he has authored and coauthored a number of papers in international journals and international conference proceedings and authored a recent monograph.

**Vilmar Æsøy** graduated from the Norwegian University of Science and Technology (NTNU), in 1989, and continued his research on natural gas fueled marine engines at NTNU/MARINTEK, until 1997. He received the Ph.D. degree for his research on natural gas ignition and combustion through experimental investigations and numerical simulations, in 1996. From 1989 to 1997, he was engaged in several large R&D projects developing gas fueled engines and fuel injection systems for the diesel engine manufacturers, Wärtsilä, and Bergen Diesel, Roll-Royce. From 1998 to 2002, he was a R&D Manager for Rolls-Royce Marine Deck Machinery. Since 2002, he has been employed in teaching with the Ålesund University College, where he is also developing and teaching courses in marine product and systems design on bachelor's and master's level. In 2010, he received the green ship machinery professorship. His current research interest includes energy and environmental technology, with focus on combustion engines and the need for more environmental friendly and energy efficient systems.

**Houxiang Zhang** (M'04-SM'12) received the Ph.D. degree in mechanical and electronic engineering, in 2003, and the Habilitation degree in informatics from the University of Hamburg, in 2011. Since 2004, he has been a Postdoctoral Fellow and a Senior Researcher with the Department of Informatics, Faculty of Mathematics, Informatics and Natural Sciences, Institute of Technical Aspects of Multimodal Systems, University of Hamburg, Germany. He was with the Aalesund University College, in 2016. In 2011, he joined the Norwegian University of Science and Technology, Norway, where he is currently a Professor on mechatronics. His current research interests include biological robots and modular robotics, especially on biological locomotion control, and virtual prototyping in demanding marine operation. He has applied for and coordinated more than 20 projects supported by the Norwegian Research Council, German Research Council, and industry. In these areas, he has published over 160 journal and conference papers as author or co-author. He received four best paper awards and four finalist awards for Best Conference Paper at the International Conference on Robotics and Automation.