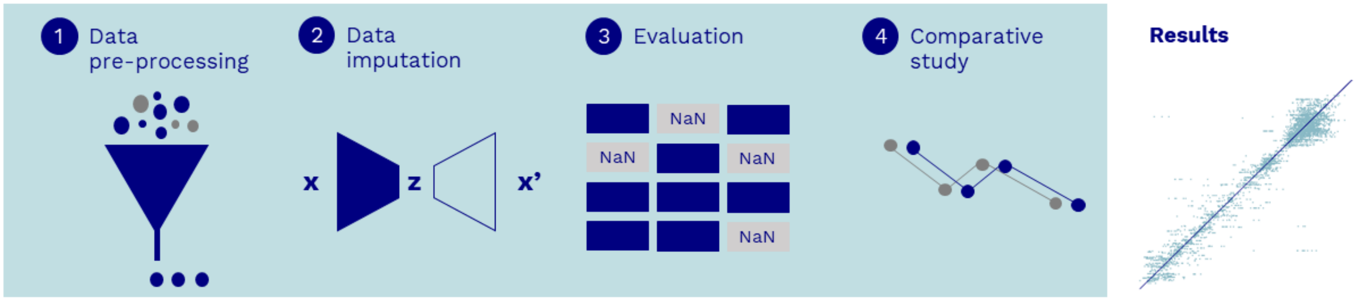
Analysis of variational autoencoders for imputing missing values from sensor data of marine systems

Christian Velasco-Gallego[1], **Iraklis Lazakis**[1]
1. Department of Naval Architecture, Ocean and Marine Engineering, University of Strathclyde, 100 Montrose Street, G4 0LZ, Glasgow, United Kingdom



Of all the causes of accidents to ships, 14% pertains to damage due to ship equipment. Accordingly, the maritime industry is currently considering state-of-the-art maintenance and inspection processes, an example of which is Condition-Based Maintenance (CBM). This is a strategy that hinges on the condition monitoring of assets. Condition Monitoring (CM) has proven to increase efficiency, reliability, profitability, and performance of vessel. To enable this maintenance strategy, sensors need to be installed along the most critical ship components and around the environment where these assets are operating through the application of Internet of Ships (IoS). IoS has demonstrated to be effective for collecting data in real time as well as performing diagnosis and prognosis to assess the current and future health of machinery to assist instant decision-making. The employment of IoS presents several challenges, an example of which is the imputation of missing values. Data imputation is a compelling pre-processing step, the aim of this is to estimate identified missing values to avoid under-utilisation of data. This data preparation step has gained popularity over the last few years due to its importance when dealing with Industrial Internet of Things (IIoT) sensor data. Although some articles presented new methodologies to impute missing values from sensor data of marine machinery based on machine learning methodologies, deep learning models have not yet been considered. For this reason, variational autoencoders for imputing missing values from sensor data of marine systems are analysed in this paper. To assess the performance of variational autoencoders as imputation methods, a comparative study is performed with widely implemented imputation techniques. Mean imputation, Forward Fill and Backward Fill, and k-Nearest Neighbors are considered. To that end, a case study on marine machinery system parameters obtained from sensors installed on a diesel generator of a tanker ship is performed. Results demonstrate the applicability of variational autoencoders when dealing with missing values of marine machinery systems sensor data, achieving a coefficient of determination of 0.99 when imputing missing values of the diesel generator power parameter.

KEY WORDS: data imputation, deep learning, neural networks, variational autoencoders, marine machinery systems, smart maintenance

INTRODUCTION

The utilisation of data provides greater opportunities regarding predictive maintenance in order to anticipate forthcoming failures in marine machinery. As such, costs can be diminished by averting random preventive maintenance and crisis-related reactive maintenance. Due to the prosperity of Condition-Based Maintenance (CBM) within the maritime industry, the volume of data accessible to implement instant data-driven decision-making strategies for enhancing operations and maintenance activities is growing exponentially. Accordingly, analysis can empower predictive maintenance by implementing Maintenance Analytics (MA). As outlined by Karim et al. (2016), and Jasiulewicz-Kaczmarek and Gola (2019), MA is constituted by four interconnected time-line phases (maintenance descriptive, maintenance diagnostic analytics, maintenance predictive analytics, and maintenance prescriptive analytics), the aim of which is the promotion of maintenance actions by improving the understanding of data and information.

Maintenance descriptive analytics summarises the data collected from various maintenance sources to provide measures and visualisations. For instance, based on fault data, measures in relation to the number of failures per components and/or per period can be obtained, as well as their respective graphical representations. If the outcome of this stage is combined with reliability data, the next maturity phase is achieved. Maintenance Diagnostic Analytics is constituted by fault detection (detection of faults and malfunctions), fault isolation (root cause analysis implementation), and fault identification (description of the fault type and its nature). Once the current health of marine machinery is determined, maintenance predictive analytics can be employed to approximate future outcomes, such as either predicting the likelihood of marine machinery operating without a failure up to certain times or estimating the Remaining Useful Life (RUL) before a failure may occur. This is achieved by considering current marine machinery conditions concurrently with past operation profiles. Finally, the higher level of maturity is maintenance prescriptive analytics, which transforms the outcomes obtained in the preceding phases into actions to optimise, among other aspects, spare parts inventory lists, capital invested in condition monitoring equipment, contractors' employment, degradation performance minimisation, Operation and Maintenance (O&M) activities, and generated operational income.

However, as Internet of Ships is in its infancy, there is a lack of data quality due to unreliable outcomes caused by certain anomalies and missing values that are originating from device failure, network collapse, and human error (Balakrishnan and Sangaiah, 2018; Izonin et al., 2019; Noor et al., 2014). Accordingly, the adequate implementation of data pre-processing steps, such as data synchronization and data imputation, is essential to guarantee reliable data-driven models. Various studies have been performed in relation to data imputation methodologies in industries in which the utilisation of sensors is highly expanded. Liu et al. (2020) proposed a univariate data imputation method to recover large gaps of missing values from Industrial IoT manufacturing sensor data. Hadeed et al. (2020) implemented an evaluation process to assess both univariate and multivariate imputation techniques. Predictive Mean Matching, Random Markov, and Kalman Filter were some of the approaches assessed by the latter. The proposed methodology was applied in the environmental sector. Although the suggested studies presented promising imputation results in their respective sector, further application and elaboration on the respective results need to be performed to determine if such frameworks are robust to the challenges to be addressed in the maritime sector. An example of these is the consideration of different states (e.g., operational, and non-operational conditions) and abrupt adjustments that refer to small changes applied due to the contractual agreements between the charterer and the ship owner in relation to the vessel speed and the fuel oil consumption per day (Velasco-Gallego and Lazakis, 2020).

Therefore, although data imputation is a compelling pre-processing step that has gained popularity recently, there is a lack of formalisation and analysis thus far within the maritime industry (Velasco-Gallego and Lazakis, 2020; Cheliotis et al., 2019). This indicates that the deployment of such methodologies within the maritime domain is also yet to be adequately formalised, thus needing further research in this field.

To that end, the present paper addresses the above issue by introducing a methodology that analyses variational autoencoders for imputing missing values from sensor data of marine systems. Deep learning has demonstrated their capability to automatically adapt to different typologies and complex datasets, and the provision of accurate predictions. Furthermore, feature engineering and data labelling, which are concepts especially complex to address in the maritime industry, are not required. Thus, such methodologies are assessed to determine if they also perform accurately when dealing with missing values, and subsequently lead to a bias reduction in further steps of the data-driven models. Moreover, future work guidelines are also indicated to deal with some of the challenges that deep learning methodologies present, such as the lack of transparency and the computational resources that are required.

The rest of the paper is structured as follows. Section 2 presents the current literature on data imputation methods within the maritime industry. Section 3 describes the proposed methodology. Section 4 reflects on the results obtained after implementing the proposed methodology through a case study. Finally, in Section 5 the conclusions of this paper are presented.

LITERATURE REVIEW

A total of three articles have been identified that present data imputation methodologies to deal with missing data collected from marine machinery. Cheliotis et al. (2019) developed a data imputation framework by combining both *k*-Nearest Neighbors (*k*-NN) and Multiple Imputation by Chained Equations (MICE) techniques. To highlight the accurate performance of the hybrid imputation method a case study was presented. Specifically, data collected from a total of 8 sensors coupled to the turbocharger and to the main engine of a chemical tanker were analysed. The proposed hybrid methodology was compared to *k*-NN and MICE methods to demonstrate the enhanced results of the proposed methodology.

Velasco-Gallego and Lazakis (2020) implemented a comparative study to examine the real-time imputation performance of a total of 20 machine learning and time series forecasting algorithms. Examples of these are the mean imputation, *k*-NN, Neural Networks (NNs) with 1, 2, and 3 hidden layers, and Autoregressive Integrated Moving Average (ARIMA). A case study was also implemented on a total of 7 machinery parameters, such as the main engine rotational speed, the lubrication oil inlet pressure, and the jacket water cooling system inlet pressure, obtained from sensors installed on a cargo vessel to assess their performance, suggesting that ARIMA outperformed the remaining imputation models in terms of accuracy and computational cost.

Velasco-Gallego and Lazakis (2021) developed a Data Assessment Imputation Framework (DAIF) to assess the accuracy of any imputation model. Specifically, the Kernel Ridge Regression and the GA-ARIMA models were evaluated by the proposed methodology as multivariate and univariate imputation techniques, respectively. This was done to demonstrate the applicability of the suggested framework in the case of marine machinery systems. Additionally, the Exponentially Weighted Moving Average (EWMA) model was also assessed as a denoising method, and a real-time imputation tool based on an open-source stack was also put forward. A case study based on the analysis of time-series data collected from a main propulsion engine of a cargo vessel was presented. Results demonstrated the importance of: 1) applying denoising when time series data contain high noise and the model applied is sensitive to it, 2) preventing failures that lead to the collection of either incorrect or missing values, 3) the influence of an effective data dashboard on the prevention of sensor failure, 4) the necessity to implement data assessment imputation frameworks, as there is not a unique model that outperforms the remaining imputation techniques for all possible characteristics and contexts described in the maritime industry.

With regards to data pre-processing approaches, Dalheim and Steen (2020) developed a data preparation toolbox for analysis of time series data. The methodology was constituted by feature selection, synchronization, outlier detection, validation, and extraction. Karagiannidis and Themelis (2021) performed the following pre-processing steps: "imputation" algorithm, outliers' identification, data smoothing, data quality control, and feature engineering. The imputation algorithm assumes that past values do not change drastically in the future when an adequate short time window is considered, the performance of which may decrease if the data contain large gaps of missing values. Perera and Mo (2016) implemented Gaussian Mixture Models (GMMs) with an Expectation Maximization (EM) algorithm to identify marine engine operating regions. Additionally, Principal Component Analysis (PCA) was utilised for structure understanding of each GMM in relation to ship and navigation performance. Gkerekos and Lazakis (2020) performed engine transients' rejection, recording anomalies rejection, weather forecast imputation, and feature engineering. Multivariate Imputation by Chained Equations (MICE) algorithm was applied as an imputation approach. Although all the above papers considered pre-processing steps, none of them analysed comprehensively the data imputation phase, despite the fact it has been perceived that datasets of marine machinery systems usually contain from 4.4% to 26% missing values (Cheliotis et al., 2019).

In addition to the above, several studies have been conducted in relation to deep learning methodologies within the maritime industry for the implementation of fault diagnosis and Remaining Useful Life (RUL) prediction (Senemmar and Zhang, 2021; Han et al., 2021; Wang et al., 2020; Ellefsen et al., 2019). However, to the best of the authors' knowledge, there is no evidence that such methodologies have been performed to implement data imputation, except for the analysis of multilayer perceptron, as indicated in preceding paragraphs. Deep learning methodologies have demonstrated their capability to automatically adapt to different typologies and complex datasets, and the provision of accurate predictions. Furthermore, feature engineering and data labelling, which are concepts especially complex to address in the maritime industry due to the identification of stationary parts of in-service measurement data (Dalheim and Steen, 2020), for instance, are not required. Despite of the previous studies undeniable results, various challenges are yet to be addressed, such as the need for large volumes of data, the lack of flexibility, the risks of obtaining an overfitted model, and the consideration of such methodologies as "black-box" models due to their lack of transparency. Therefore, it is expected that further research will encompass strategies that address such challenges and ensure the further implementation of deep learning methodologies within the maintenance analytics context. In an effort to further promote and enhance the application of smart maintenance methodologies within the maritime industry, the present study suggests a novel framework for the analysis of variational autoencoders for regression modified by adding Long Short-Term Memory (LSTM) layers in both the encoder and decoder to consider the characteristics of time series data.

METHODOLOGY

The proposed methodology is graphically represented in (Fig. 1). The first phase refers to data pre-processing, which is essential to be implemented prior to model training due to the characteristics of the data. For instance, data may contain idle states that need to be excluded from the analysis. Subsequently, the LSTM-VAE-based regressor analysed in this study is introduced. To assess the imputation performance of such an approach, several contexts and metrics are considered. Finally, to evaluate if the analysed methodology can enhance other imputation techniques widely implemented within the maritime industry, a comparative study is introduced.

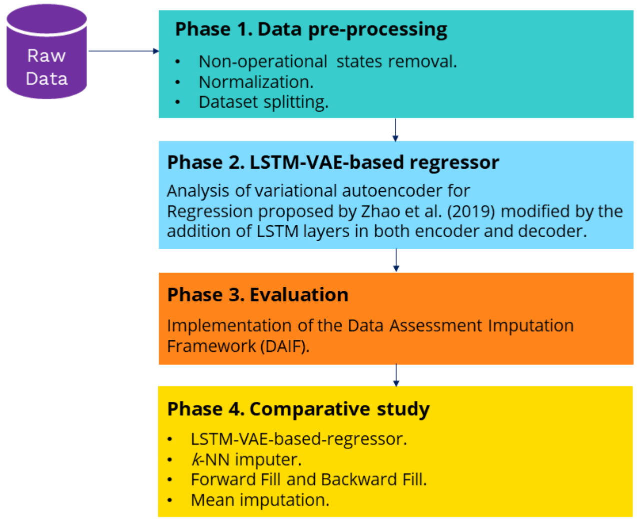

Raw Data
Phase 1. Data pre-processing
• Non-operational states removal.
• Normalization.
• Dataset splitting.
Phase 2. LSTM-VAE-based regressor
Analysis of variational autoencoder for Regression proposed by Zhao et al. (2019) modified by the addition of LSTM layers in both encoder and decoder.
Phase 3. Evaluation
Implementation of the Data Assessment Imputation Framework (DAIF).
Phase 4. Comparative study
• LSTM-VAE-based-regressor.
• k-NN imputer.
• Forward Fill and Backward Fill.
• Mean imputation.

Fig.1. Graphical representation of the proposed methodology.

Data pre-processing

Prior to the implementation of this section, data understanding needs to be performed to determine the steps to apply within this section. The usual pre-processing steps that need to be addressed when dealing with sensor data of marine systems are data synchronization, data resampling, non-operational states detection, data understanding, data denoising, outliers' detection, data imputation, data transformation, feature engineering, non-stationarity assessment, and multicollinearity identification.

All parameters may have been either recorded at different timestamps or present delays due to communication issues. Therefore, the parameters need to be synchronised prior to the implementation of the data imputation model. To that end, data is resampled to ensure that all features' timestamps are aligned. Additionally, downsampling may be required to reduce the noise that time series data contain when dealing with high-frequency data. Interpolation is applied to address this matter.

Subsequently, non-operational states need to be adequately identified and discarded. Original Equipment Manufacturers (OEMs) of the systems being analysed are usually consulted to address this matter. However, the expert knowledge obtained from the original engine/equipment manufacturers' manuals are complemented with data-driven models to evaluate if the accuracy of identifying such states increases. Accordingly, Gaussian Mixture Models (GMMs) with Expectation-Maximization (EM) algorithm implementation for fitting the model is applied. This probabilistic model considers that data are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. EM is utilised for fitting the models and Bayesian information Criterion to evaluate the possible number of clusters. This step is performed by the application of the scikit-learn Python library (Pedregosa et al., 2011).

Prior to the implementation of the data imputation model, data transformation is also implemented. Normalization is applied so that the parameters lie between 0 and 1 values. To adequately assess the performance of the data imputation approach, the remaining steps identified within this section, such as data denoising and feature engineering, are not implemented within this methodology, as they could interfere with the imputation performance. However, the implementation of such approaches is highly recommended, as they have been proven to enhance the imputation performance of the analysed models (Velasco-Gallego and Lazakis, 2021).

In relation to exploratory data analysis, correlation analysis is performed by the estimation of both the Pearson's correlation coefficient and Spearman's rank correlation coefficient to identify linear and non-linear relationships between features. To finalise the pre-processing step, the data are split into training (80% of the entire dataset), validation (20% of the training dataset), and test (20% of the entire dataset) sets to avoid model overfitting.

LSTM-VAE-based regressor analysis

The deep learning model implemented as a data imputation technique in this study is the variational autoencoder for regression. As the name indicates, this is a type of autoencoder that learns the parameters of a probability distribution, which enables the model to be generative. The model is constituted by an autoencoder, the aim of which is to learn both how to reduce the input dimensions and compress the inputs into an encoded representation. This compressed state, a.k.a latent space representation, presents the lowest possible dimensions of the inputs. Subsequently, the decoder is utilised to learn how to reconstruct the data contained in the latent space representation to reproduce the inputs as analogously as possible. To achieve the variational autoencoder neural network, the training process for such autoencoder is regularised by encoding the input as a distribution over the latent space (Fig. 2).

The methodology proposed by Zhao et al. (2019) is modified to enable the model to impute time-series data. Such methodology is constituted by an encoder with 2 intermediate layers of dimension (128, 32) with *tanh* as activation function. The resulting output is independently connected to two layers, the dimension of which is 8, to determine both the mean and the standard deviation of the latent representation. The regressor, which shared the intermediate layers of the encoder, is utilised to determine the mean and standard deviation for the predicted feature. Finally, the model is also constituted by the decoder. By utilising the latent representation as the input, the reconstruction is accomplished. Therefore, the architecture proposed considers a feedforward artificial neural network. Specifically, a multilayer perceptron, which does not deal with sequentiality, and thus does not consider the characteristics of time series data. Accordingly, the variational autoencoder based regression model is adapted to learn temporal dynamic behaviour by the implementation of Long Short-Term Memory Network (LSTM), which is a type of recurrent neural network introduced by Hochreiter and Schmidhuber (1997). Fig. 3 presents a diagram of the VAE-based regression model highlighting the modification of both the encoder and decoder by the addition of LSTM layers. Specifically, the encoder is formed by 2 layers (128, 64) and *tanh* activation function. Analogously, the decoder is constituted by 2 layers (64, 128) and *tanh* activation function. The ratio of validation set has been set to 0.20. Adam optimizer has been applied to compile such a model. Subsequently, the model has been trained, setting the number of epochs to 100 and the batch size to 32. The prior hyperparameters have been defined based on prior experience and heuristic evaluation. This step is performed by the implementation of the Python libraries Tensorflow and Keras.

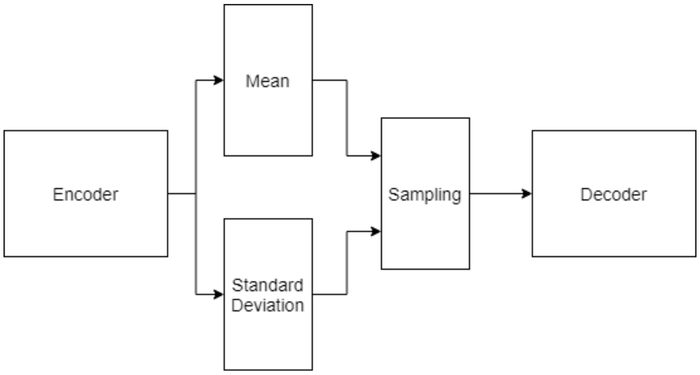


Fig. 2. Variational autoencoder

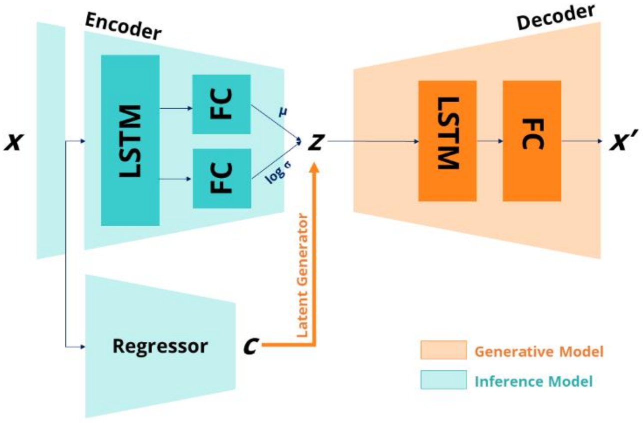


Fig. 3. Diagram of the VAE-based regression model modified to include LSTM layers in both encoder and decoder.

Evaluation

To analyse if the LSTM-VAE-based regression model can be implemented as an imputation model, different contexts need to be analysed. Generally, a total of three different missingness of data mechanisms are considered. The first mechanism, Missing Completely at Random (MCAR), involves those situations in which the missingness is independent of the data. An example of which is a random failure produced in the fuel flowmeter (Cheliotis et al. 2019). The second mechanism is the Missing at Random (RAM). In this case the missingness is dependent on another feature (e.g., if a component of a main engine fails, the operating condition of dependent components may be altered). The third mechanism is identified as Missing Not at Random (MNAR), which refers to those scenarios in which the missingness is related to the feature itself. The evaluation of the imputation performance of this model has been implemented by the Data Assessment Imputation Framework (DAIF) presented by Velasco-Gallego and Lazakis (2021). Specifically, the first mechanism, MCAR, is employed ~~presented~~ in this study, as it has been identified as the most common mechanism when dealing with data collected from marine systems. The procedure of which is described hereunder.

- All the sequences of the target value obtained from the original dataset are considered as the input of ~~the~~ DAIF. The remaining features are considered as explanatory variables.
- *n* samples with different missing ratios ($r_1, r_2, \ldots, r_m$) are generated. Each sequence contains values missing completely at random.
- The missing values are initially either masked or imputed to fit the VAE-regression model.
- The missing values are imputed by implementing the VAE-regression model.

Additionally, six metrics are estimated to determine the imputation performance of the imputation approach (Root Mean Square Error (RMSE), Mean Squared Error (MSE), Median Absolute Error (MedAE), Mean Absolute Error (MAE), Max. Error, and coefficient of determination (R^2)).

Comparative study

To compare the imputation performance of the LSTM-VAE-regression model with data imputation methodologies widely implemented by both industry and academia a total of three models are utilised to perform a comparative study.

The first method considered is the mean imputation, which is a simple forecasting method widely used to impute missing values by estimating the mean of the sample. It is usually implemented as it is easy to interpret, easy to apply, and the execution time is low. The second method is employed by Makridis et al. (2020). This method consists of applying Forward Fill and, subsequently, Backward Fill algorithms. Finally, to also consider a multivariate imputation approach, *k*-NN is also utilised, which was analysed by Cheliotis et al (2019), and Velasco-Gallego and Lazakis (2020).

RESULTS

Having explored the methodology being analysed as a data imputation technique, a case study is introduced to assess its imputation performance. Specifically, a total of 14 parameters (see Table 1) collected from a diesel generator of a tanker ship are considered.

Table 1. Parameters of the diesel generator considered for the case study.

Id	Parameter
P1	Power
P2	Exhaust gas outlet temperature of cylinder 6
P3	Exhaust gas outlet temperature of cylinder 5
P4	Exhaust gas outlet temperature of cylinder 4
P5	Exhaust gas outlet temperature of cylinder 3
P6	Exhaust gas outlet temperature of cylinder 2
P7	Exhaust gas outlet temperature of cylinder 1
P8	Winding temperature T phase
P9	Winding temperature S phase
P10	Winding temperature R phase
P11	Turbocharger exhaust gas outlet temperature
P12	Cooling air temperature
P13	Lube oil inlet temperature
P14	Cylinder exhaust gas outlet temperature (average)

In order to apply the methodology described in the previous section of this paper, the target variable is defined, and the main outcomes obtained after performing the pre-processing phase with regards to non-operational states identification are presented. Furthermore, the descriptive statistics, the histograms, and the correlation analysis for each parameter are also shown for a better understanding of the introduced case study. Subsequently, the missing ratios analysed are presented. Then, the results obtained in the comparative study phase are introduced and comprehensively discussed. Finally, elements for future work are presented to conclude this section.

The parameter P1 is considered as the target variable. The remaining features are considered as the explanatory variables. These data have been collected in a 1-minute frequency and include a total of 66207 instances. Fig. 4 represents graphically the time series data of such a parameter. As observed, there are several idle states that need to be excluded from the analysis. Furthermore, adjustments introduced due to either contractual agreements between the charterer and the shipowner or weather conditions can be also perceived. To identify those idle states, GMMs with EM is applied, the minimum of mixture components analysed being 1 and the maximum 10. Four different types of covariance are also assessed (full, tied, diagonal, and spherical). As indicated in Fig. 5, a total of 2 components and the spherical covariance type have been selected as the parameters to train the model. Once the model has been fitted, the idle states have been removed, thus obtaining a total of 33745 instances. Therefore, more than 49% of the dataset refers to idle states.

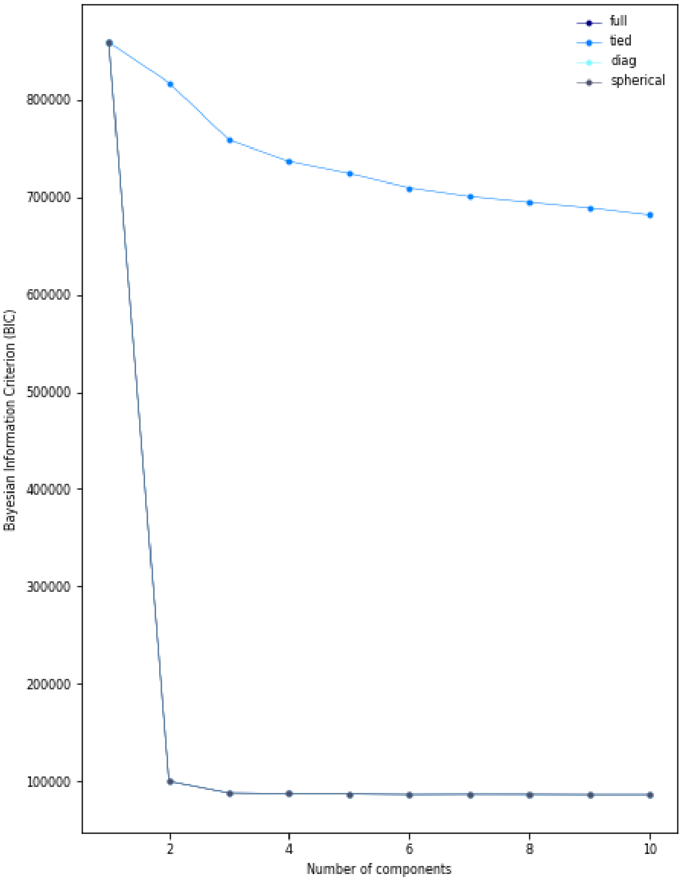


Fig. 5. Parameters' selection of the GMMs.

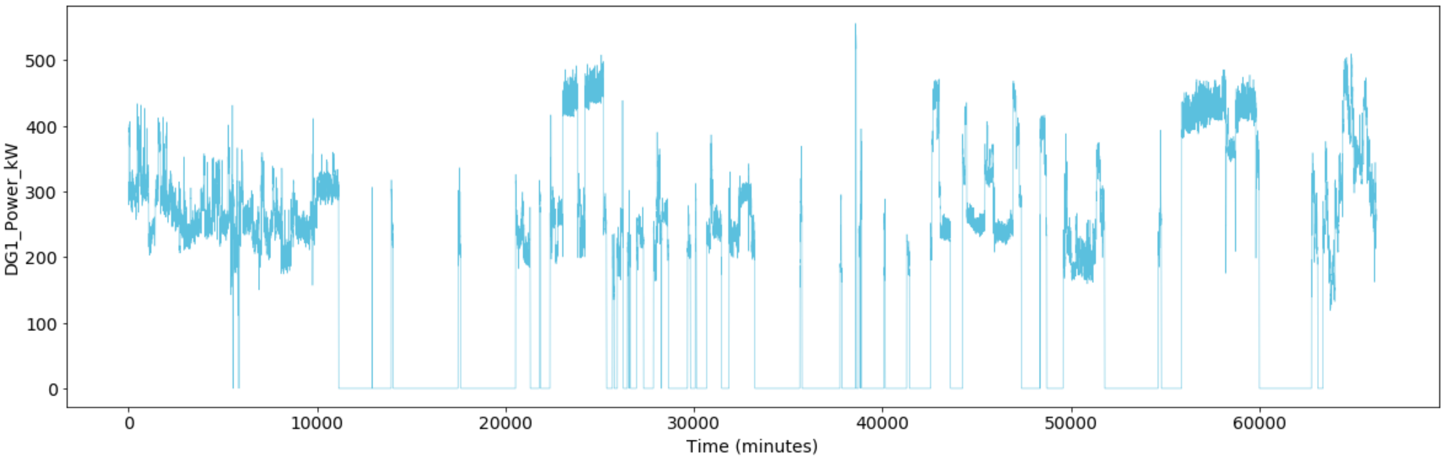
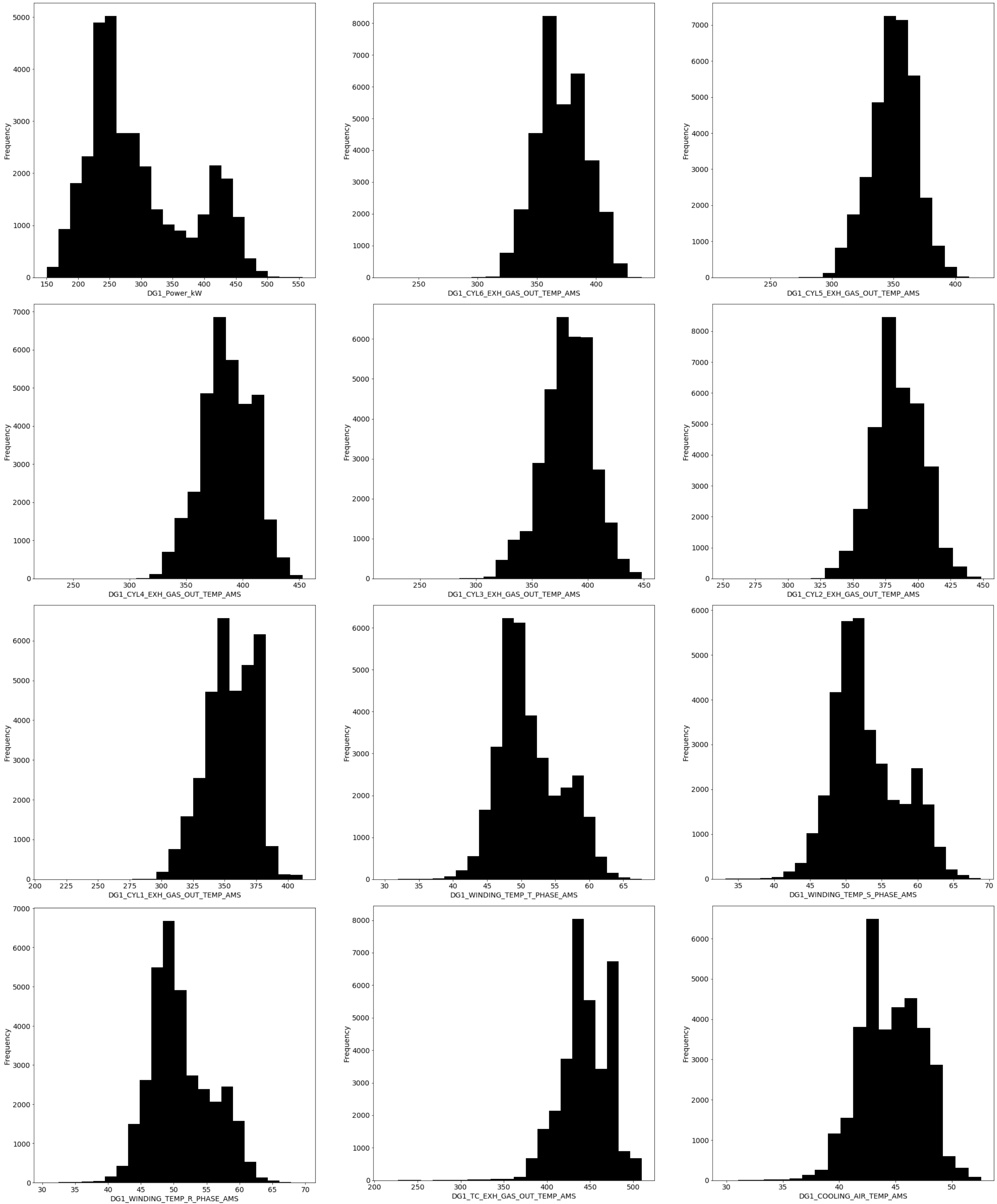


Fig. 4. Time series plot of the diesel generator power.

Table 2. Descriptive statistics of the monitored features.

	P1	P2	P3	P4	P5	P6	P7
Count	33745	33745	33745	33745	33745	33745	33745
Mean	296.93	370.61	349.95	386.52	382.16	384.43	354.64
Std.	81.09	20.95	18.31	22.79	22.43	18.91	19.12
Min.	150.19	222.7	213.5	226.5	219.9	251.9	209.4
25%	236.97	355.6	338.9	371.6	368.2	372.1	341.6
50%	270.94	369.3	350.8	386	383.1	383.1	354.9
75%	356.14	385.8	362.2	404.1	397.5	399.5	371.6
Max.	555.93	438.7	421.1	453.1	448.1	448.5	411.7

	P8	P9	P10	P11	P12	P13	P14
Count	33745	33745	33745	33745	33745	33745	33745
Mean	51.30	52.89	51.06	444.82	44.59	62.00	371.39
Std.	4.612	4.79	4.67	28.64	2.72	0.84	19.36
Min.	30.2	33.2	30.7	213.7	29.9	45.6	223.9
25%	48.1	49.5	47.8	429.1	42.6	61.7	359.66
50%	50.3	51.9	50	443.5	44.5	62.1	371.06
75%	54.3	55.9	54.2	469.1	46.7	62.5	388.16
Max.	67.7	68.8	69.6	510.1	52.6	65.2	434.55


Frequency
DG1_Power_kW
150 200 250 300 350 400 450 500 550
0 1000 2000 3000 4000 5000
DG1_CYL6_EXH_GAS_OUT_TEMP_AMS
250 300 350 400
0 1000 2000 3000 4000 5000 6000 7000 8000
DG1_CYL5_EXH_GAS_OUT_TEMP_AMS
250 300 350 400
0 1000 2000 3000 4000 5000 6000 7000
DG1_CYL4_EXH_GAS_OUT_TEMP_AMS
250 300 350 400 450
0 1000 2000 3000 4000 5000 6000 7000
DG1_CYL3_EXH_GAS_OUT_TEMP_AMS
250 300 350 400 450
0 1000 2000 3000 4000 5000 6000
DG1_CYL2_EXH_GAS_OUT_TEMP_AMS
250 275 300 325 350 375 400 425 450
0 1000 2000 3000 4000 5000 6000 7000 8000
DG1_CYL1_EXH_GAS_OUT_TEMP_AMS
200 225 250 275 300 325 350 375 400
0 1000 2000 3000 4000 5000 6000
DG1_WINDING_TEMP_T_PHASE_AMS
30 35 40 45 50 55 60 65
0 1000 2000 3000 4000 5000 6000
DG1_WINDING_TEMP_S_PHASE_AMS
35 40 45 50 55 60 65 70
0 1000 2000 3000 4000 5000 6000
DG1_WINDING_TEMP_R_PHASE_AMS
30 35 40 45 50 55 60 65 70
0 1000 2000 3000 4000 5000 6000 7000
DG1_TC_EXH_GAS_OUT_TEMP_AMS
200 250 300 350 400 450 500
0 1000 2000 3000 4000 5000 6000 7000 8000
DG1_COOLING_AIR_TEMP_AMS
30 35 40 45 50
0 1000 2000 3000 4000 5000 6000

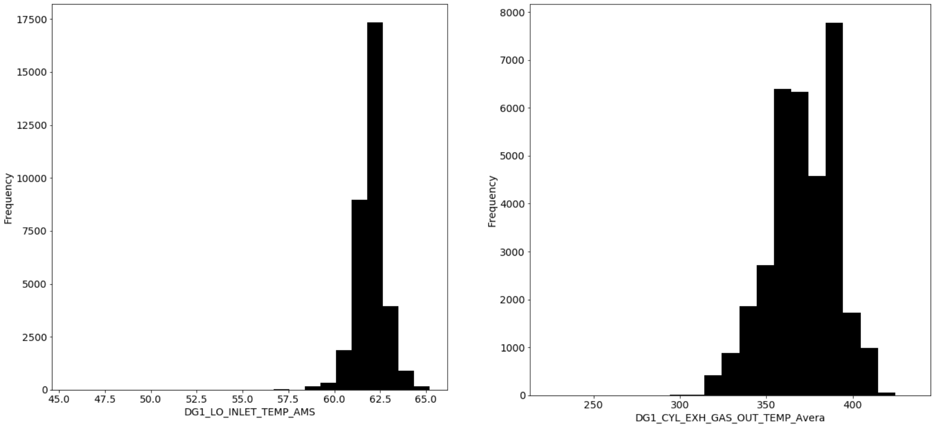


Fig. 6. Histograms of the monitored features (P1-P14).

Subsequently, exploratory data analysis is performed. The descriptive statistics are presented in Table 2, the histograms are represented in Fig. 6., and the Pearson's correlation coefficients are described in Table 3. As indicated, all parameters present a strong correlation with the P1, except for the parameter P13

To assess the imputation performance of the fitted model, a total of five ratios of missing values have been analysed (0.05, 0.15, 0.25, 0.5, 0.8). Therefore, three contexts of missing values are evaluated (small, medium, and large). Results for the four models being analysed (LSTM-VAE-based regressor, mean imputation, application of Forward Fill and, subsequently, Backward Fill algorithms, and *k*-NN) are presented in Tables 4-7. The numbers of neighbours selected for the *k*-NN imputer is estimated by the square root of the number of instances of the sample, thus avoiding overfitting.

Table 3. Pearson's correlation coefficient (absolute values).

Parameter	coefficient
P2	0.88
P3	0.82
P4	0.87
P5	0.80
P6	0.84
P7	0.81
P8	0.86
P9	0.87
P10	0.86
P11	0.88
P12	0.77
P13	0.51
P14	0.89

Results demonstrated that LSTM-VAE-based regressor outperforms the remaining analysed models. However, it can be perceived that, when data contain small numbers of missing values, the difference is not significant. For instance, if a missing ratio of 0.05 is considered, it can be observed that the RMSE of the LSTM-VAE-based regressor is 8.91 kW (Table 4), while the value of this metric is 9.1 kW if the Forward Fill and Backward Fill (FF-BF) algorithms are applied. By contrast, when a large rate is considered (missing ratio of 0.8) the RMSE of the LSTM-VAE-based regressor and FF-BF are 12.20 kW and 22.00 kW, respectively. This indicates that the LSTM-VAE-based regressor is more robust than the FF-BF model, which only presents an analogous performance when the dataset contains small numbers of missing values, as such an approach imputes the missing values based on the prior and subsequent observations. Thus, further analysis with regards to the implication of noise and abnormal values in the observations considered by FF-BF in each instance needs to be made. In relation to the performance of the *k*-NN imputer, it can be perceived that the imputations are more robust than when FF-BB is applied for medium and large gaps of missing values. However, several limitations need to be highlighted with regards to such an approach. Examples of which are the performance degradation when the sample contains large gaps of missing values and when both dimensions and number of instances are high. Additionally, the number of neighbours need to be optimally selected. In this study it has been selected by estimating the square root of the number of instances of the sample to avoid overfitting, although more sophisticated approaches may need to be applied to enhance its imputation performance. The worst results are achieved when the mean imputation is applied, accomplishing the maximum RMSE perceived when the missing ratio is 0.8 (108.07 kW). The variability of the data and the different operational states identified within the dataset may be the cause of such results. Furthermore, when the mean imputation is applied it can be perceived that the parameter distribution is distorted, which may lead to a disruption of the relationship between features. As results did not vary significantly when applying this methodology, this has not been included in the analysis described in Fig. 7. However, an example to observe the distortion of the parameter is expressed in Fig. 8. The study of performing mean imputation in each operational state can be implemented instead of estimating the mean for the entire dataset, thus determining if the bias and the limitations of such an approach still occur.

If results are analysed generally, it can be observed that the imputation performance is reduced when the missing ratio increases; as less observed data can be utilised to train the different models. Therefore, this indicates that the prevention of errors that yield either corrupted or missing values is of paramount importance when building a data-driven model. In addition, should an imputation method need to be applied due to the impossibility of preventing these errors, a comprehensive analysis needs to be performed to avoid adding bias estimates that may lead to an inaccurate model that could ultimately be utilised for decision-making strategies.

Table 4. Imputation results of the LSTM-VAE-based regressor.

Missing ratio	RMSE (kW)	MSE (kW^2)	Max. Error (kW)	MedAE (kW)	MAE (kW)	R^2
0.05	8.91	79.47	41.48	5.33	6.69	0.99
0.15	9.80	96.12	196.66	4.72	6.48	0.98
0.25	10.10	102.13	144.23	4.98	6.86	0.99
0.5	10.74	115.39	153.57	5.06	7.23	0.98
0.8	12.20	148.88	196.06	7.29	8.87	0.98

Table 5. Imputation results of the *k*-NN imputer.

Missing ratio	RMSE (kW)	MSE (kW^2)	Max. Error (kW)	MedAE (kW)	MAE (kW)	R^2
0.05	14.68	215.51	69.41	7.97	10.92	0.97
0.15	15.26	232.89	70.68	8.51	11.44	0.97
0.25	15.89	252.56	151.11	8.64	11.6	0.96
0.5	16.22	262.96	149.24	8.88	11.97	0.96
0.8	18.31	335.37	150.91	10.87	13.72	0.95

Table 6. Imputation results of the application of Forward Fill and, subsequently, Backward Fill algorithms.

Missing ratio	RMSE (kW)	MSE (kW^2)	Max. Error (kW)	MedAE (kW)	MAE (kW)	R^2
0.05	9.1	82.9	42.61	4.68	6.53	0.99
0.15	10.89	118.69	107.46	4.99	7.28	0.98
0.25	12.2	148.72	107.46	5.27	7.91	0.98
0.5	16.89	285.18	221.99	6.24	9.58	0.96
0.8	22.00	484.29	262.13	8.52	13.59	0.94

Table 7. Imputation results of the mean imputation technique.

Missing ratio	RMSE (kW)	MSE (kW^2)	Max. Error (kW)	MedAE (kW)	MAE (kW)	R^2
0.05	107.87	11635.41	180.1	113.53	100.16	0
0.15	107.68	11595.06	181.68	113.63	99.95	0
0.25	107.89	11641.99	181.68	113.44	100.39	0
0.5	108.47	11764.97	189.86	113.75	101.14	0
0.8	108.07	11679.65	189.42	112.89	100.66	0

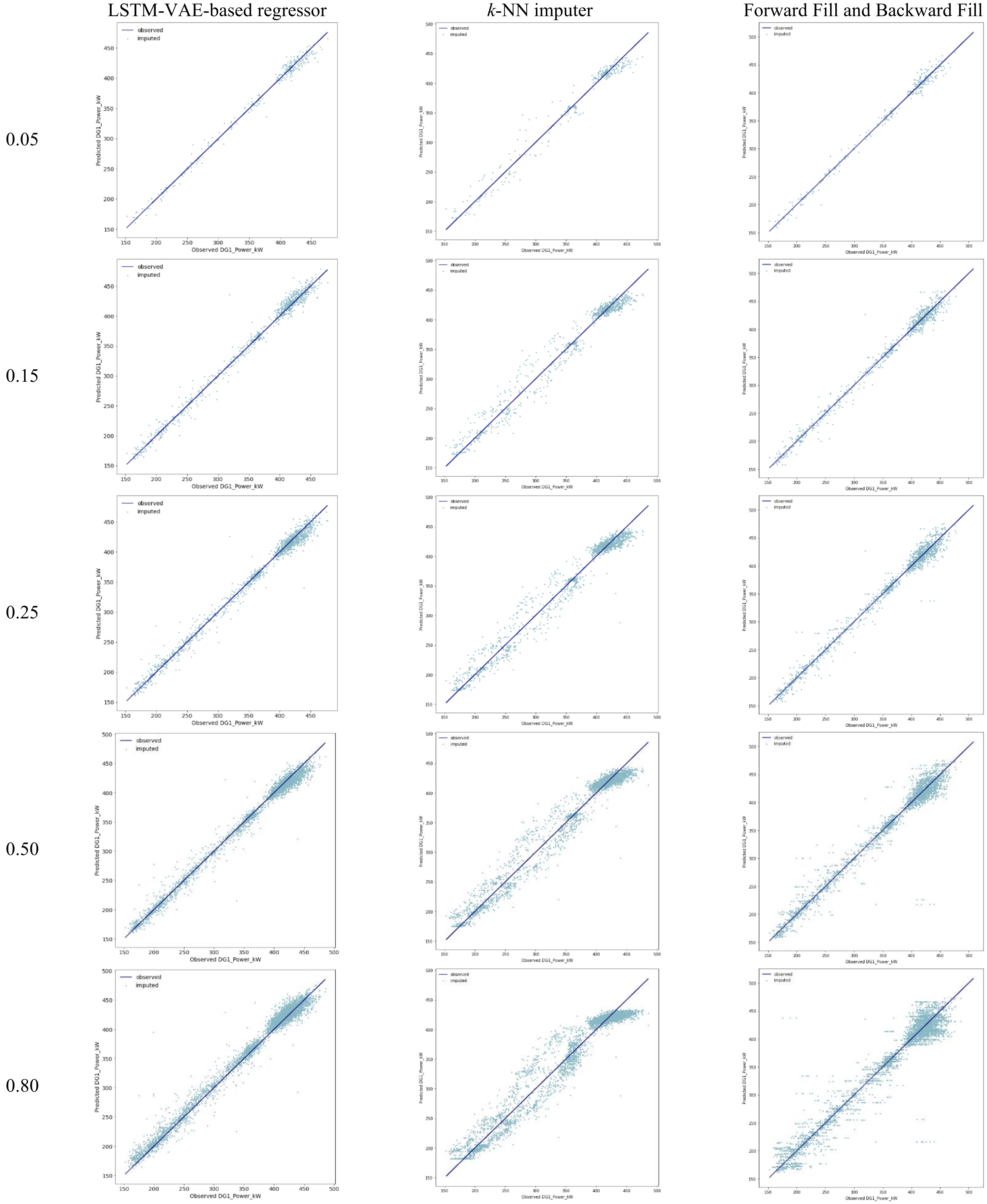


Fig. 7. Comparison between observed values and imputed values of the diesel generator power.

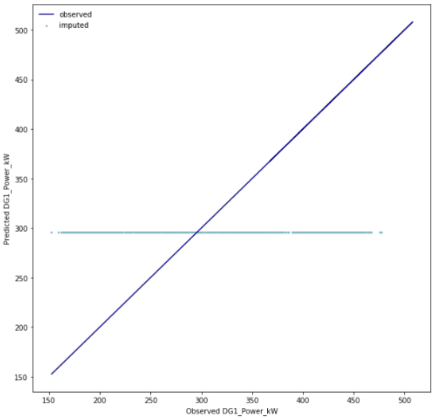


Fig. 8. Comparison between observed values and imputed values of the diesel generator power when the mean imputation is applied, and the missing ratio is 0.15.

Although the LSTM-VAE-based regression model has demonstrated an accurate imputation performance, it presents several limitations that need to be addressed. The first major limitation of this approach is the dependency on the volume of data utilised for training purposes. In addition, its computational cost is higher than other methodologies that presented analogous imputation results in specific contexts. Thus, the application of other methodologies may need to be utilised as imputation techniques instead when there are insufficient resources to deal with such models and when a low execution time is also required. Other limitations are the lack of transparency, lack of flexibility, and risk of obtaining overfitting models. These matters may generate a lack of trust towards these models within the industry if they are not adequately addressed. Therefore, further research needs to be implemented. Some of the aspects that are in the research agenda in relation to deep learning and data imputation analyses are listed hereunder.

- Perform a comprehensive analysis of other deep learning models to assess their imputation performance.
- Analyse the implication of corrupted data in the imputation performance of the analysed models.
- In this study only non-operational states were identified and discarded from the analysis. However, there are other states that could not be determined with the current approach that may negatively impact the imputation performance. Accordingly, further research needs to be performed to identify in a more accurate manner these transitions so that the implication of such states in the imputation performance can be assessed.
- Study of implementing deep learning models for real-time data imputation.
- The analysis performed in this study presented as a target variable a feature that was highly correlated with the explanatory variables. Thus, there is a need of studying and introduce modifications to the current model to determine possible enhancements in the architecture of the neural network and obtain a more robust model while considering the data imputation performance and the resources needed to adequately implement it. This also includes the analysis of optimisation techniques to adequately select the architecture of the deep neural network and determine the hyperparameters of the applied models, as this study selected the hyperparameters based on different tested architectures.
- Apply other data pre-processing steps that have not been implemented to evaluate if the imputation performance increases. Examples of these are feature selection, feature extraction, and time-series denoising. Moreover, this analysis has demonstrated the presence of collinearity between independent variables. To deal with such a collinearity, a proper analysis in the data pre-processing step needs to be performed by implementing a more comprehensive correlation analysis and by estimating relevant metrics, such as Variance Inflation Factor (VIF). Then, based on whether collinearity is identified or not, it is possible that some features may need to be removed from the analysis or techniques such as Principal Component Analysis (PCA) and Partial Least Squares (PLS) may need to be further analysed and implemented accordingly.
- Introduce explainable intelligence models to deal with the current scepticism towards “black-box” models within the maritime industry.
- Perform more validations with other available datasets and parameters.

CONCLUSIONS

Data accessibility in the maritime industry is already a fact, which has led to an exponential increase in data-driven applications to assist decision-making strategies. However, there are several challenges that need to be addressed. An example of which is the existence of either corrupted or missing values due to diverse typology errors, including device failure and inadequate data manipulation. The inappropriate treatment of such values may lead obtaining of unreliable models.

Accordingly, data imputation needs to be implemented in these conditions. This pre-processing step is not yet formalised, however, within the maritime industry. An indicator of this is the application of imputation techniques that lack robustness as shown in various studies. Consequently, further research in relation to this matter is required within the maritime industry, as it has been noted that marine machinery sensor datasets usually contain from 4.4% to 26% missing values.

This deals with missing values from marine systems data by analysing the possible application of deep learning methodologies as data imputation techniques. Specifically, variational autoencoders for regression are evaluated. To that end, a data imputation assessment framework is utilised to simulate various missing contexts, thus enabling the analysis of their imputation performance. To complement the analysis, a comparative study is considered by implementing other data imputation techniques currently applied within this industrial sector. Mean imputation, Forward Fill and Backward Fill, and *k*-NN imputers were considered. A case study on a diesel generator of a tanker ship was introduced to evaluate the performance of the proposed data imputation methodology. Specifically, the power parameter was considered as a target variable. Results demonstrated that the imputation performance is enhanced when variational autoencoders for regression are implemented. However, it presents analogous results to Forward Fill and Backward Fill when leading with small missing ratios. However, the latter may not be appropriate if the dataset contains either other missing scenarios (e.g., large gaps or large ratios of missing values) or unexpected behaviours, which indicates its lack of robustness. The worse results were achieved when mean imputation was applied, as it both distorts the parameter distribution and disrupts the relationship between features when considering parameters of analogous characteristics from the one presented in this study.

Consequently, variational autoencoders for regression are considered the most adequate models to perform data imputation tasks. Their suitability notwithstanding, the limitations cannot be avoided, such as their lack of transparency and the need of large volumes of data. Meanwhile, future work guidelines, including the utilisation of explainable artificial intelligence models and the development of a comprehensive comparative study of deep learning methodologies for data imputation, have been indicated accordingly to advance with the formalisation of data imputation within the maritime industry.

REFERENCES

Balakrishnan S M, Sangaiah A K. Chapter 6 - aspect oriented modeling of missing data imputation for Internet of Things (IoT) based healthcare infrastructure. Intelligent Data-Centric Systems (2018): 135-145, doi: https://doi.org/10.1016/B978-0-12-813314-9.00006-2.

Cheliotis M, Gkerekos C, Lazakis I, Theotokatos G. A novel data condition and performance hybrid imputation method for energy efficient operations of marine systems. Ocean Engineering 188 (2019): 1-14, doi: https://doi.org/10.1016/j.oceaneng.2019.106220.

Dalheim Ø Ø, Steen S. Preparation of in-service measurement data for ship operation and performance analysis. Ocean Engineering 212 (2020): 1-17, doi: https://doi.org/10.1016/j.oceaneng.2020.108261.

Ellefsen A. L., Cheng X., Holmeset F. T., Æsøy V., Zhang, H, Ushakov S., 2019. Automatic Fault Detection for Marine Diesel Engine Degradation in Autonomous Ferry Crossing Operation. IEEE International Conference on Mechatronics and Automation (ICMA) 2019, pp. 2195-2200, doi: https://doi.org/10.1109/ICMA.2019.8816600.

Gkerekos C, Lazakis I. A novel, data-driven heuristic framework for vessel weather routing. Ocean Engineering 197 (2020): 1-10, doi: https://doi.org/10.1016/j.oceaneng.2019.106887.

Hadeed S J, O'Rourke M K, Burgess J L, Harris R B, Canales R A. Imputation methods for addressing missing data in short-term monitoring of air pollutants. Science of the Total Environment 730 (2020): 1-7, doi: https://doi.org/10.1016/j.scitotenv.2020.139140.

Han P., Ellefsen A. L., Li G., Æsøy V, Zhang H., 2021. Fault Prognostics Using LSTM Networks: Application to Marine Diesel Engine. IEEE Sensors Journal, pp. 1-8, doi: https://doi.org/10.1109/JSEN.2021.3119151.

Hochreiter S., Schmidhuber, J., 1997. Long short-term memory 9(8): pp. 1735-1780, doi: https://doi.org/10.1162/neco.1997.9.8.1735.

Izonin I, Kryvinska N, Tkachenko R, Zub K An approach towards missing data recovery within IoT smart system. Procedia Computer Science 155 (2019): 11-18, doi: https://doi.org/10.1016/j.procs.2019.08.006.

Jasiulewicz-Kaczmarek M, Gola A. Maintenance 4.0 Technologies for Sustainable Manufacturing - an Overview. IFAC-PapersOnLine 52:10 (2019): 91-96, doi: https://doi.org/10.1016/j.ifacol.2019.10.005.

Karagiannidis P, Themelis T. Data-driven modelling of ship propulsion and the effect of data pre-processing on the prediction of ship fuel consumption and speed loss. Ocean Engineering 222 (2021): 1-15, doi: https://doi.org/10.1016/j.oceaneng.2021.108616.

Karim R, Westerberg J, Galar D, Kumar U. Maintenance Analytics - The New Know in Maintenance. IFAC-PapersOnLine 49:28 (2016): 214-219, doi: https://doi.org/10.1016/j.ifacol.2016.11.037.

Liu Y, Dillon T, Yu W, Rahayu W, Mosafa F. Missing value imputation for Industrial IoT sensor data with large gaps. IEEE Internet of Things Journal 7:8 (2020): 1-13, doi: https://doi.org/10.1109/JIOT.2020.2970467.

Noor N M, Abdullah M M A B, Yahaya A S, Ramli N A. Comparison of linear interpolation method and mean method to replace the missing values in environmental data set. Materials Science Forum 803 (2014): 278-281, doi: https://doi.org/10.4028/www.scientific.net/MSF.803.278.

Pedregosa F et al. Scikit-learn: Machine Learning in Python. JMLR 12 (2011): 2825-2830.

Perera L. P. and Mo B. Data analysis on marine engine operating regions in relation to ship navigation. Ocean Engineering 128 (2016): 163-172, doi: https://doi.org/10.1016/j.oceaneng.2016.10.029.

Senemmar S., Zhang J., 2021. Deep Learning-based Fault Detection, Classification, and Locating in Shipboard and Power Systems. 2021 IEEE Electric Ship Technologies Symposium, pp. 1-6, doi: http://doi.org/10.1109/ESTS49166.2021.9512342.

Velasco-Gallego C, Lazakis I. Real-time data-driven missing data imputation for short-term sensor data of marine systems. A comparative study. Ocean Engineering 218 (2020): 1-23, doi: https://doi.org/10.1016/j.oceaneng.2020.108261.

Velasco-Gallego C, Lazakis I. Data imputation of missing values from marine machinery systems sensor data. Evaluation, visualisation, and sensor failure detection. Proceedings of the RINA Maritime Innovation and Emerging Technologies (2021): 15-23.

Wang S., Wang J., Wang R., 2020. A novel scheme for intelligent fault diagnosis of marine diesel engine using the multi-information fusion technology. IOP Conference Series Materials Science and Engineering 782, pp. 1-12, doi: https://doi.org/10.1088/1757-899X/782/3/032022.

Zhao Q., Adeli E., Honnrorat N., Leng T., Pohl K. M. Variational AutoEncoder For Regression: Application to Brain Aging Analysis. Application to Brain Aging Analysis. Med Image Comput Comput Assist Interv 11765 (2019): 823-831, doi: https://doi.org/10.1007/978-3-030-32245-8_91.