Investigation of logarithmic signatures for feature extraction and application to marine engine fault diagnosis

Chaitanya Patil[1], Gerasimos Theotokatos[1], Yue Wu[2], and Terry Lyons[3,4]

[1]Maritime Safety Research Centre, Department of Naval Architecture, Ocean and Marine Engineering, University of Strathclyde, Glasgow G4 0LZ, UK

[2]Department of Mathematics and Statistics, University of Strathclyde, Glasgow G1 1XH, UK

[3] Mathematical Institute, University of Oxford, Oxford, OX2 6GG, UK

[4]The Alan Turing Institute, London, NW1 2DB, UK

September 3, 2024

Abstract

Although the advancements in marine engines diagnosis technologies and systems, estimating faults combinations at the entire operating envelope is challenging. This study aims at first investigating the path logarithmic signatures (logS) method for information extraction and dimensions reduction from the in-cylinder pressure signals, and secondly, proposing the most effective data-driven hybrid approach employing logS as input to artificial neural networks (ANN) regression to estimate the severity of critical faults in marine engines. A large four-stroke marine diesel engine is considered and in-cylinder pressures are generated using a validated physics-based digital twin by simulating scenarios with four faults combinations of varying severity in the entire operating envelope. A parametric study is performed to quantify the logS number impact on the ANN regression models accuracy and training time. Four data pre-processing approaches, which consider elementary or high variance logS, without or with the use of Principal Components Analysis (PCA), are also comparatively assessed. The results demonstrate that the approach involving the use of eight elementary logS, Principal Components Analysis (PCA), standardisation and an ANN regression model comprising two hidden layers with ten neurons each is the most

effective, as it exhibits the lowest values on both the root mean square and the standard error 95% confidence interval. This is the first study on logS application for marine engines faults severity estimation, and as such it impacts the development of future data-driven diagnostics methods.

1 Introduction

1.1 Background

Maritime industry is impacted by the industry 4.0 concepts and tools. Emerging technologies, such as smart and autonomous systems, advanced materials, big data analysis, robotics, sensors, and communications, as well as artificial intelligence (AI) are significant drivers for the future development of the shipping industry [1]. The development of advanced fault diagnosis systems for marine engines has shifted from traditional methods with the use of experiential knowledge to the use of technologies such as acoustic emission testing, thermographic inspection, and machine learning (ML) algorithms for the acquired data analysis.

1.2 Literature review

Lv et al. [2] conducted an extensive review on fault diagnosis studies for marine machinery highlighting the ML algorithms advances. Guo et al. [3] proposed a universal fault diagnosis framework for marine machinery based on domain adaptation. Zhang et al. [4] developed a binary adversarial autoencoder deep neural network model for marine anomaly detection in system-level. Kougiatsos et al. [5] demonstrated by simulations a distributed cyber-physical framework for marine engines sensor fault diagnosis. Coraddu et al. [6] reported the health monitoring of marine engines by using weekly supervised algorithms. Unsupervised fault detection using the clustering techniques and estimation of Remaining Useful Life (RUL) for rotating mechanical systems was reported by Calabres et al [7]. Xi et al. [8] developed a classifier to avoid human errors using independent component analysis (ICA) of vibration data obtained from several engine locations. Tsaganos et al. [9] proposed an ensemble model (multi-technique) to classify common faults of two-stroke marine engines. Extreme learning ensemble was developed by Kowalski et al. [10] to classify 14 faults in four stroke diesel engines. Cai et al. [11] developed a fault localisation algorithm targeting the engine subsystems separately using an association rule-based method for mining a database with historical faults. Wang et al. [12] proposed a fault diagnosis framework consisting of supervised and unsupervised algorithms for classifying limited number of diesel

engines faults at specific operating points. The diagnosis of a marine engine fault based on a comprehensive similarity method was studied by Chai et al. [13], considering faults related to high pressure fuel pump wear, carbon formation in the fuel injector, piston ring and exhaust valve leakages. Perera et al. [14] demonstrated a localised model with data points clustering using Gaussian mixture submodels associated with an expectation maximisation algorithm for the detection of marine engine faults. Graph Learning Neural Networks were used for semi supervised multiple graph learning for faults classification by Jiang et al. [15]. A robust feature extraction-based image classification neural network was used to detect the damages on road by Roy et al. [16], which however requires image data processing on a larger scale. Velasco-Gallego and Lazakis [17] developed a methodology integrating the first-order Markov chain with both ResNet50V2 and Convolutional Neural Networks (CNN) to identify anomalies in diesel generator sets by implementing time series imaging. Meng et al. [18] conducted reliability analusis by employing physics-informed neural networks. Recently, Song et al. [19, 20] proposed advanced methods for fault tolerant dynamic control methods demonstrating their application for the case studies of a quadrotor unmanned aerial vehicle. However, they are not relevant to marine engine mechanical components diagnosis applicable for ship machinery due to their high frequency of occurrence from long duration operation. Nevertheless, they could be considered for addressing the topic of marine engines fault tolerant control or sensor fault diagnosis in future studies.

Above discussed methods require extensive datasets as input with several engine performance parameters (dimensions). The excessive amount of information required to develop data-driven models is known as the curse of dimensionality [21], which results in deteriorating the machine learning algorithms performance. The large number of dimensions increases both computational effort and the training datasets required to develop a meaningful data-driven model [22], rendering the selection of input with the minimum dimensions quintessential. Hybrid approaches combining signal processing with machine learning algorithms can address the dimensionality issues [2]. For example, Li et al. [23] developed a predictive model to detect oil leaks in marine diesel engines using the Wigner bispectrum of instantaneous angular speed along with fault classification based on ANN, reporting coefficient of determination (R^2) values around 94%. Kuai et al. [24] proposed an engine valve wear detection model by employing the wavelet packet decomposition of the vibration signal and ANN classification. Wang et al. [25] employed a classification model for marine engine fault detection including valve wear, injection timing delay, turbocharger and air cooler fouling, by using efficient channel attention of convolution neural networks, resulting in R^2 values around 97%. Jian et al. [26] developed a classification model to detect mechanical faults in marine blowers combining time-averaging decomposition (TAD)

with confusion score feature selection (CSFS), reporting R^2 values around 94%. However, the previous studies only considered single or limited engine operating points instead of the entire operating envelope, whereas simultaneous multiple faults occurrence was not considered in the pertinent literature [27, 28].

In-cylinder pressure has been used for diagnostics, as it conveys crucial information characterising thermodynamic phenomena that occur within the engine cylinders, reflecting the faults and/or degradation of the engine components [29]. Hence, in-cylinder pressure signals are expected to be effective input parameters for data-driven fault estimation models. However, the input dimensions of such models depend on the pressure signal sampling frequency, which typically lies in the range of 0.5–1^{o}CA. Such sampling rates, along with the consideration of the engine operating envelope, the expected faults and their combinations, may lead to non-practical data-driven models dimensions for real applications. Therefore, the reduction of the in-cylinder pressure signals dimensions is required, so that input parameters of manageable dimensions are employed in such models to diagnose faults and estimate the fault severity in marine engines.

The rough path is a generalisation of smooth paths and can represent evolving features in real phenomena modelled by using controlled differential equations. According to Lyons et al. [30], the rough path method is effective in reducing the input features in multi-dimensional datasets to path signatures. The path signatures are features representing the order of events and the evolving pattern(s) of the considered system [31, 32, 33]. Signatures have been successfully used in several applications, such as handwriting recognition [34], and healthcare early-warning prediction [35, 36]. The logarithm of these signatures, called logarithmic signatures (logS), provides a more concise representation of the data, which makes it preferable for cases where multidimensional features are required [37]. Logarithmic signatures are more effective to transform wide stream of data into negligible class [38] compared to methods such as Fourier transform, Wavelet transform, Cepstral analysis [39], and Hilbert transform [40], whereas they are well compatible with modern deep learning algorithms [38]. The logS can convey different levels of information, as revealed from their applications in medical, finance and art industries. Based on the preceding discussion, it is considered that the logS method can be used to extract information and reduce dimensions of the in-cylinder pressure signals, providing advantages in retaining crucial information characterising faults (and their combinations), over alternative methods. However, a rigorous investigation to identify the most effective approach and logS levels for marine engine fault diagnosis is needed.

The preceding discussion leads to the identification of the following research challenges pertinent to the faults severity estimation for marine engines: (a) the effective and practical fault

severity estimation should consider faults combinations and the the whole operating envelope; (b) effective data-driven models should exhibit high accuracy employing the lowest number of input parameters; (c) models for marine engine faults severity estimation especially employing in-cylinder pressure signals are not available; and (d) datasets for benchmarking purposes are not available.

1.3 Aim and novelty

This study aims at investigating the use of logarithmic signatures (logS) for marine engine fault diagnosis. The selected logS from in-cylinder pressure are derived and used to develop a hybrid data-driven approach that combines logS and artificial neural networks (ANN) regression to estimate the severity for four critical and most frequent component faults (and their combinations) throughout the operating envelope of the engine.

To the best of authors' knowledge, this is the first study to propose the application of logarithmic signatures for the estimation of fault severity in marine engines using only in-cylinder pressure signals. In addition, this study addresses the research challenges (a) to (c) identified in the preceding section, as it considers four most frequent faults with varying severity and their combinations in the whole engine operating envelope, whilst providing pathways and recommendations to develop data-driven models for estimating the faults severity in marine engines. Moreover, this study overcomes the literature limitations, where mostly single faults classification models are reported.

This study provides the following contributions:

1. Address the need for effective and practical fault severity estimation in marine engines considering combinations of simultaneous faults and the entire engine operating envelope.

2. Develop a novel data-driven hybrid approach that employs as input the in-cylinder pressure signal to estimate the severity of critical faults in marine engines.

3. Generate datasets by using a physics-based digital twin, which can be used to benchmark fault diagnosis algorithms.

1.4 Study structure

The remainder of this study is structured as follows. Section 2 describes the employed methodology, methods, and tools, as well as the four approaches for extracting the corresponding logS.

Section 3 provides the results and their discussion. Section 4 discusses the study implications, whereas section 5 summarises the main findings.

2 Methodology

The methodology employed in this study is illustrated in the flowchart (top box) of Figure 1. It consists of following main four phases: (1) Data Generation, (2) Data pre-processing approaches; (3) Regression models training; (4) Regression models testing. The application of the developed data-driven model for the prediction of the faults severity of marine engines using the recommended approach (as discussed in the results section) is illustrated in the flowchart at the bottom in Figure 1.

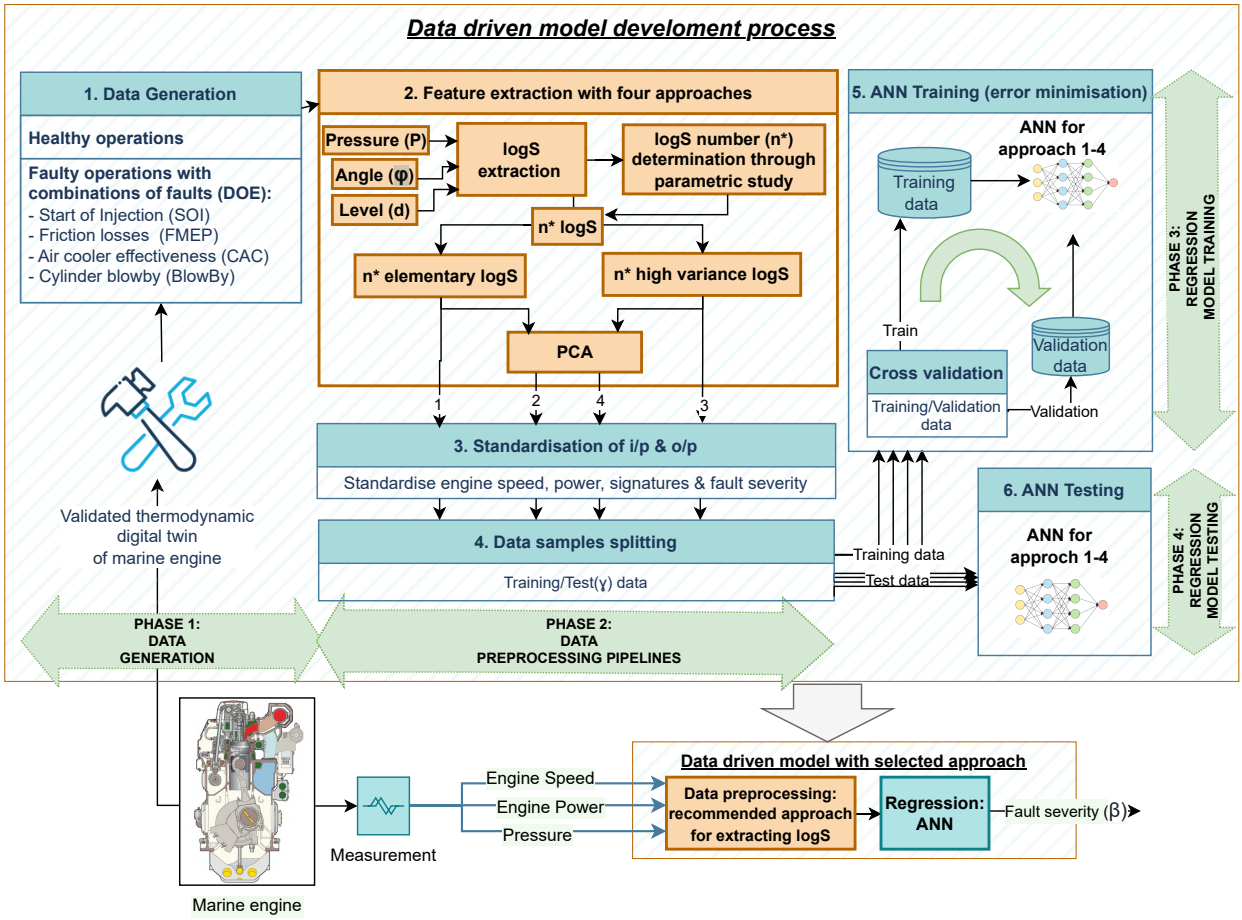


Figure 1: Model development methodology flowchart (top box) along with application of the final model at the bottom.

Phase 1–Data generation: This phase deals with the generation of the in-cylinder pressure diagrams for the engine operating envelope considering several faults and their combinations. A validated thermodynamic digital twin of the investigated marine engine is employed. Simula-

tion runs are performed to derive datasets corresponding to both healthy and faulty operating conditions. Several operating scenarios with combinations of the four most frequent faults of varying severity are considered. A design of experiment (DOE) method is used to set faults with different severity indices as explained in Subsection 2.1.

Phase 2–Data pre-processing: This phase focuses on deriving the datasets required to set up, train, and test the data-driven regression models (in Phase 3). The following four approaches are proposed:

- Approach 1: Elementary n* number of logS
- Approach 2: Elementary n* number of logS transformed by principal component analysis (PCA)
- Approach 3: Selective n* number of logS with high variance
- Approach 4: Selective n* number of logS with high variance and transformed by PCA

Initially, the in-cylinder pressure signals are fed to the logS algorithm (also providing the logS level as input) to derive the logS (termed elementary). Different logS levels are considered as explained in section 2.2.1), to perform a parametric study and quantify the impact of the logS level (and number) on the regression models accuracy. Variance analysis is carried out to identify logS with high variance on the considered faults. Principal component analysis is also employed to derive the principal components of logS (elementary and high variance). All the parameters are standardised and datasets consisting of the input and output parameters required for the regressions models (Phase 3) corresponding to approaches 1–4 are formed. Finally, the standardised datasets are separated into training and test datasets (considering the test to train ratio, γ), which in turn are used to train and validate the regression models (Phase 3).

Phase 3–Regression model training: This phase involves the set-up and training of regression models based on artificial neural networks (ANN). The training of the regression models (ANN) is based on the prediction error minimisation, whereas the cross validation is performed on the validation datasets, which are iteratively separated from the training data to avoid the ANNs over-fitting. The performance of these ANNs is evaluated using metrics including Root Mean Square Error (RMSE) and standard errror (SE). The details of the ANN structure and training parameters are provided in Subsection 2.3.

Phase 4–Regression model testing: Lastly, the trained ANNs corresponding to each approach are evaluated by employing the testing datasets, which were separated and are not

used at all in the training phase. The findings of the four phases are evaluated and recommendations for the development of data-driven fault severity estimation models for marine engines are provided, according to the bottom part Figure 1.

2.1 Data generation for in-cylinder pressure

This section describes the generation of the datasets required for the training and validation of the proposed data-driven fault diagnosis model. Experimental measurements from faulty engine conditions are not available, due to the followed maintenance practices. However, synthetic datasets that contain performance parameters from both healthy and faulty conditions are generated by using a physics-based digital twin (DT) based on thermodynamic zero-dimensional modelling, which was validated against shop trials and limited operational parameters.

The engine used in this study is the four-stroke Wärtsilä 9L46C marine engine, which is a nine cylinder turbocharged medium speed engine. This engine is installed in a RoPax ferry, the propulsion plant of which consists of two identical engines, each driving a controllable pitch propeller and a power take-off generator through a gearbox. The main particulars of this engine are listed in Table 1 [41].

Table 1: Investigated marine engine specifications.

Parameter	Value
Maximum Continuous Rating point	9,450 kW @ 500 r/min
No. of Cylinders	9
Cylinder Bore	460 mm
Clutch-in Speed	300 RPM
Turbocharger	ABB TPL 77-A30

The DT for this engine is developed and calibrated according to the framework provided by Tsitsilonis et al. [42]. Subsequently, it is employed to generate in-cylinder pressure datasets considering both healthy and faulty conditions; the latter are associated with several engine components faults and degradation. This study considers four faults pertaining to fuel injection, friction, blowby leakage, and charge air cooler that occur the most frequently, and for which experimental data collection remains challenging [43, 44, 45]. The modelling of these faults considering their severity level β_{fault} is described as follows.

1. **Fuel injection issues**: The advance and retard of the start of injection (SOI) occurs due to wear and tear of the fuel injection system (pumps and injectors), resulting in deterioration

of the engine performance. This study employs SOI advance and retard to represent the fuel injection fault. This fault is captured by changing the Start of Injection (SOI) for each cylinder (which is input parameter of the employed DT), according to the following equation:

$$\theta_{SOI,\text{faulty}} = \theta_{SOI,\text{healthy}}(1 \pm \beta_{\text{SOI}}) \tag{1}$$

2. **Increased friction losses**: Increased engine friction losses due to lubrication system deficiencies, or excessive bearings wear and tear result in reducing the engine brake power. This fault is injected to the DT by adjusting the Friction Mean Effective Pressure (FMEP) according to the following equation:

$$fmep_{\text{faulty}} = fmep_{\text{healthy}}(1 + \beta_{\text{fmep}}) \tag{2}$$

3. **Blowby**: Leakages of exhaust gases from the cylinder due to wear and tear in cylinder liner and piston rings cause the engine performance deterioration. This fault is modelled by using the blowby area and the nozzle flow equation to calculate the leakage gas mass flow rate [46]. The Blowby area ($A_{BB,\text{faulty}}$) in this study and expressed by:

$$A_{BB,\text{faulty}} = \pi \frac{[D_{\text{cyl}}^2 - (D_{\text{cyl}} - 2t)^2]}{4} \beta_{\text{Blowby}} \tag{3}$$

Where t is gap thickness between the piston rings and the cylinder wall.

4. **Charge Air Cooler (CAC) fouling**: This fault is attributed to the deposits and fouling in both sides (air, cooling water) of the CAC, and results in reducing the heat transfer rate (from the air to the cooling water). It is modelled by reducing the charge air cooler effectiveness according to the following equation:

$$\eta_{CAC,\text{faulty}} = \eta_{CAC,\text{healthy}}(1 - \beta_{\text{CAC}}) \tag{4}$$

To capture the combinations of these faults, a full factorial design method reported in [47] is employed. The fault severity level β_{fault} varies from 0 to 0.6 (-0.6 to 0.6 for the SOI fault) with increments of 0.15. The healthy conditions correspond to $\beta_{\text{fault}} = 0$, whereas the most severe fault corresponds to $\beta_{\text{fault}} = \pm 0.6$. In total, 1,125 ($5^3 \times 9$) combinations were generated with the full factorial design method for each operating point (steady state conditions) of the investigated marine engine. To retain the engine power output with varying the fault/degradation severity and combinations, the engine injected fuel amount was adjusted by using a PID controller. To facilitate the fault/degradation estimation over the whole engine operating envelope, 50 operating points corresponding to steady state conditions were considered. Cumulatively, 56,250

$(1,125 \times 50)$ simulation runs were performed with the fault combinations. In this respect, the heath status of the investigated marine engine throughout their operating envelope can be represented. It should be noted that this study considers similar faults/degradation for all the cylinders of a multi-cylinder engine. The cases of different faults/degradation for different cylinders is out of this study scope.

2.2 Data pre-processing using logS

This section describes the logS derivation and the parametric study considering levels of logS (corresponding to different logS numbers) to determine number of logS required for fault diagnosis task. Subsequently, the four approaches are proposed (based on the rationale behind the logS selection) including data standardisation and splitting, to derive the datasets required for the regression model training and testing (for Phases 3 and 4).

2.2.1 Log signatures (logS) from pressure path

For an integer J, consider a $\mathbb{R}^J$-valued time-dependent, piece wise-differentiable path X_t of finite length over $t \in [s,t]$. Each coordinate of X is a real-valued path denoted by $X^i, i \in \{1,\ldots,J\}$. The *signature* of a path $X : [a,b] \to \mathbb{R}^J$, denoted by $S(X)_{s,t}$, is the infinite collection of all iterated integrals of X:

$$S(X)_{s,t} := (1, S(X)^1_{s,t}, \ldots, S(X)^J_{s,t}, S(X)^{1,1}_{s,t}, S(X)^{1,2}_{s,t}, \ldots), \tag{5}$$

where, the first term is 1 by convention, whereas the superscripts of the terms after the first term run along the set of all multi-index $\{(i_1,\ldots,i_k)|k \geq 1, 1 \leq i_1,\ldots,i_k \leq J\}$ with the coordinate iterated integral being

$$S(X)^{i_1,\ldots,i_k}_{s,t} := \int_{s<t_k<t} \cdots \int_{s<t_1<t_2} \mathrm{d}X^{i_1}_{t_1} \ldots \mathrm{d}X^{i_k}_{t_k}. \tag{6}$$

The finite collection of all terms $S(X)^{i_1,\ldots,i_k}_{s,t}$ with the multi-index of fixed length d is termed as the d^{th} level of the signature. Each level has its geometric meaning; for instance, the first level of the signature of a path captures the incremental effect of the path whilst the second level captures its area effect. The signature can be truncated up to the d^{th} level for practical use, denoted by $\lfloor S(X)_{s,t} \rfloor_d$. The length of the truncated signature up to the d^{th} level is $J(J^d-1)/(J-1)$.

The logarithmic signature of a path is defined as the logarithm of the signature of the path X: $\log(S(X))$. The advantage of logarithmic signatures compared to signatures is that they further reduce the dimensions of the input whilst retaining most of the signature properties.

The resulting dimension is $\sum_{k=1}^{d} \frac{1}{k} \sum_{i|k} \mu\left(\frac{k}{i}\right) J^i$, where $i \mid k$ represents the value k can be divided by i and $\mu\left(\frac{k}{i}\right)$ is the Mobius function, i.e., the sum of the primitive $\frac{k}{i}$th roots of unity.

It must be noted that the log signatures do not have universality, which means other paths can have same log-signatures. Therefore, the signatures needs to be combined with non-linear models for learning tasks [48, 49].

Path signatures are used for various applications in stochastic analysis, including the integration of rough paths, the construction of stochastic differential equations, and the computation of rough path spaces. In real applications, truncated signatures and logarithmic signatures can be derived from datasets by employing several available Python packages (e.g., esig [50], iisignature [51], signatory [52]). Importantly, logS are used primarily in ML to represent sequential data in a way that is invariant to time-warping and re-parametrisation [33], which is the case for the application herein. More information on the logS calculation is provided in the pertinent literature.

This study calculates the logS for in-cylinder pressure signal, which is $\mathbb{R}^J$-valued angle-dependent (pressure values are recorded as function of the crank angle (ϕ) for the range $0 - 720^{o}$CA for four stroke engines), piece wise-differentiable path p_ϕ of finite length over ascending values of $\phi \in [0, 720]$ for each cylinder. Figure 2 provides the in-cylinder pressure diagrams for one of the nine engine cylinders at a specific engine operating point.

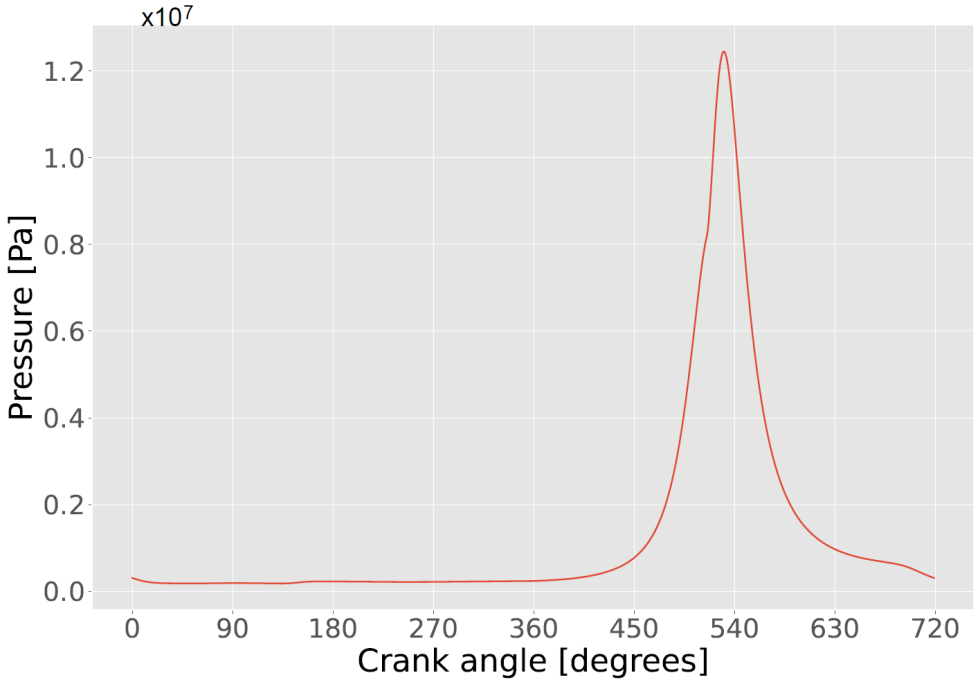


Figure 2: Typical in-cylinder pressure signal (path) for one cylinder for an engine operating point with healthy conditions.

The crank angle (ϕ) is used instead of the time for the logS calculation. The crank angle

values are normalised in the range [0,1]. The signature of the pressure path p_ϕ, denoted by $S(p)_{0,1}$, is the infinite collection of all iterated integrals of p, according to the following equation:

$$S(p)_{0,1} := (1, S(p)^1_{0,1}, S(p)^{1,1}_{0,1}, S(p)^{1,2}_{0,1}, \ldots) \tag{7}$$

The python package (*stream2logsig* from *esig.tosig* [50]) is employed to calculate the logS using as input the in-cylinder pressure (p),the normalised crank angle (ϕ), and the signature level (d), according to the following equation:

$$S(p)_{0,1} = esig.tosig.stream2logsig(p, \phi, d) \tag{8}$$

where p denotes the in-cylinder pressure values for a cylinder corresponding to normalised ascending crank angle ϕ. All the pressure profiles for the $56,250$ operating points (Section 2.1) are used to derive their corresponding logS.

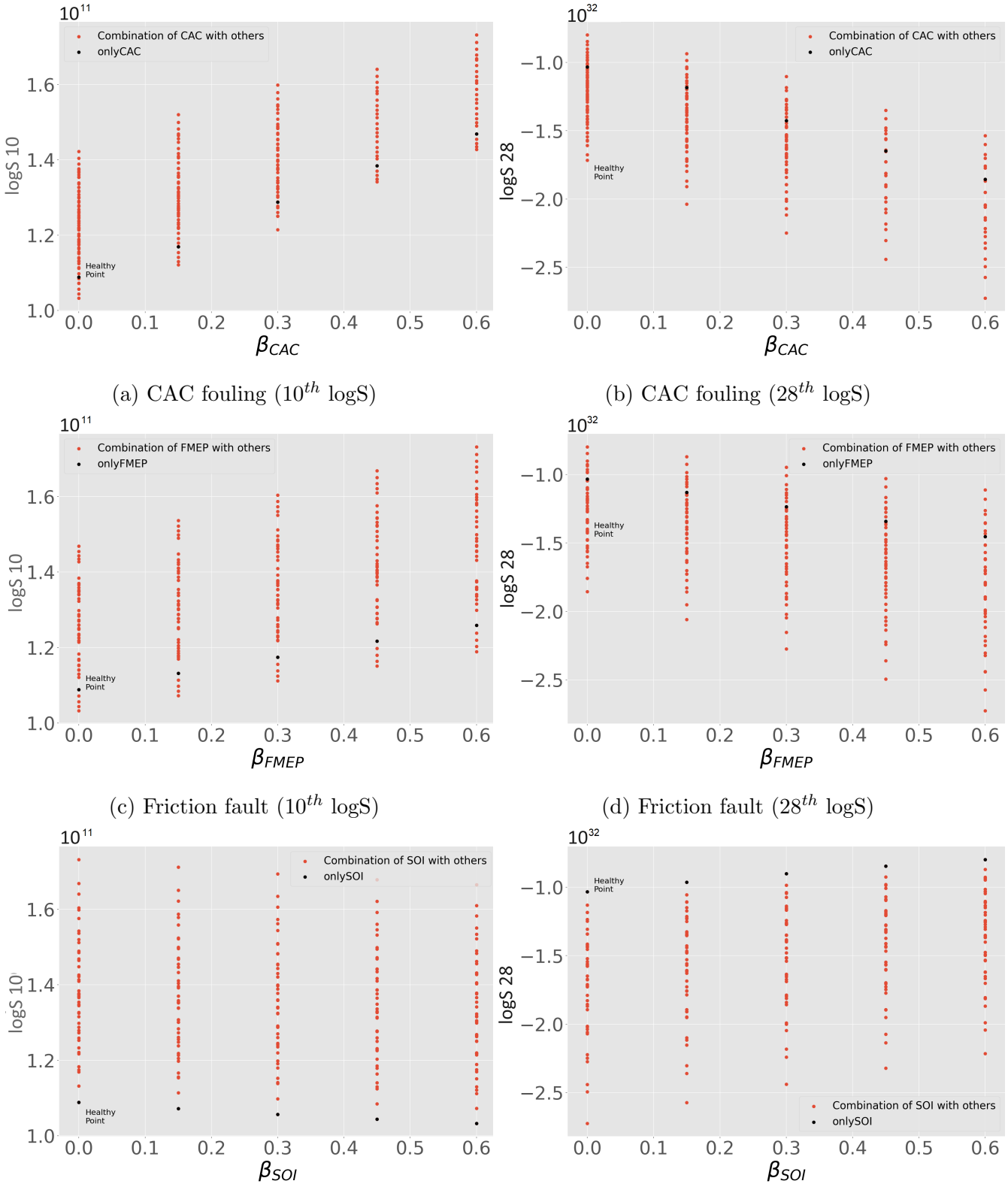
2.2.2 Parametric study

Initially, a parametric study is performed on level of logS (derived from in-cylinder pressure signals) to determine minimum level and their corresponding logS number (n*) which can be used as input for estimation of fault severity. The selection is performed by comparison of the accuracy of the fault estimation of regression models (ANN) that accept extracted logS as input. The logS levels from 2 to 6 are used for the parametric study. The results of the parametric study are presented in the results section.

2.2.3 Variance analysis and standardisation

This subsection describes the activities to identify which signatures can convey information pertinent to the fault severity estimation.

Figure 3 shows the variation in the randomly selected 10^{th} and 28^{th} logS due to the considered faults. Each plot focuses on a specific single fault (black markers) as well as its combination with other faults (red markers). These results showcase that each combination of faults exhibits different impact on the logS. No specific relationship is observed between the fault severity and the logS value. For example, the 10^{th} signature increases with the fault severity (β_{fault}), whereas the 28^{th} increases with the reciprocal severity of the faults. Moreover, it is inferred from the vertical axes scales that the variation of logS values for a specific fault changes dramatically from the 10^{th} to the 28^{th} signature. This change increases further from the earliest to the later signatures. This implies that the information contained in the logS increases with the level (number of signatures).



(a) CAC fouling (10^{th} logS)

(b) CAC fouling (28^{th} logS)

(c) Friction fault (10^{th} logS)

(d) Friction fault (28^{th} logS)

(e) SOI advance (10^{th} logS)

(f) SOI advance (28^{th} logS)

Figure 3: Impact of faults severity on 10^{th} (left) and 28^{th} (right) logS for the cases of single faults (black symbols) and faults combinations (orange symbols) at 462.5 r/min and 100% load.

Higher variance of the input parameters (logS herein) has higher potential to be used as input for data driven models [53]. The logS variance analysis for each fault is performed to

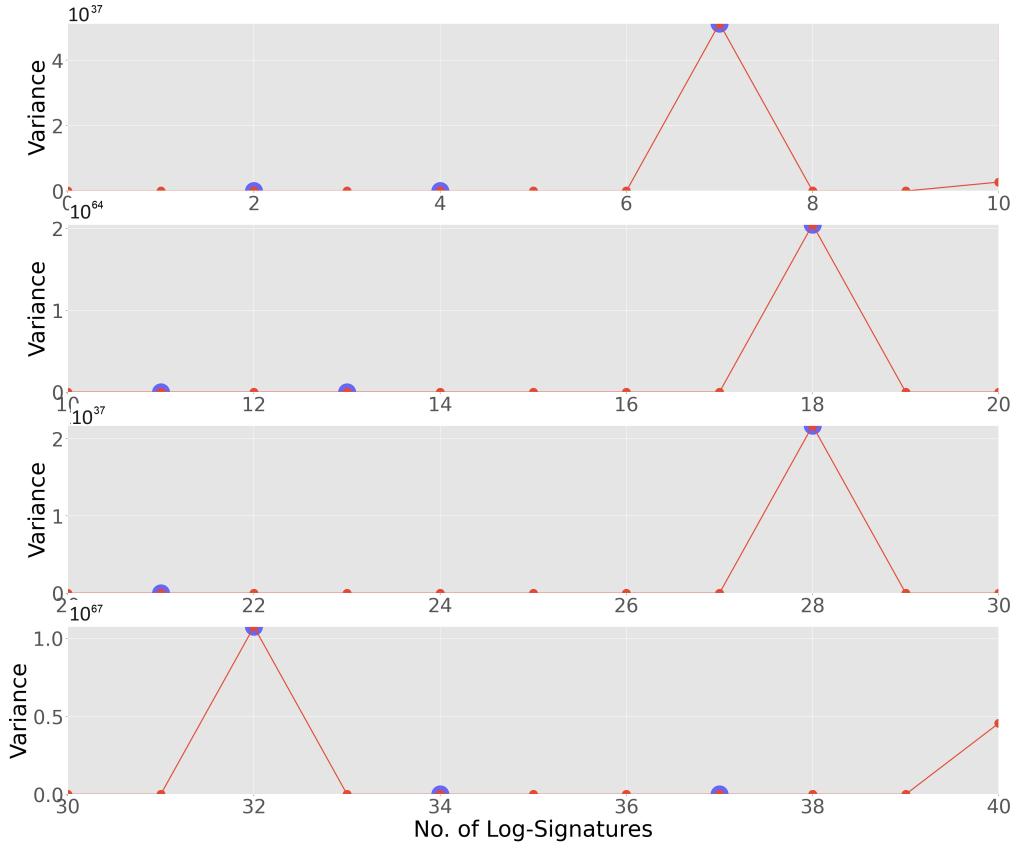


Figure 4: Variance of logS values (red symbols) for the charge air cooler fouling fault (and combinations with other faults) for the logS No. 1 to 40 (plot is presented in four sections). Blue points represents the local maxima of logS.

identify whether using specific logS (with high variance) can lead to ANN higher accuracy of fault severity estimation. For example, Figure 4 presents the variance of logS for the charge air cooler fouling. Due to the vertical-axes scale limitations, the variance peaks of several logS remain unnoticed. The logS from 1 to 40 are presented in the four plots of Figure 4, each using different y-axis scale to identify the variance peaks. The local maxima (peaks) of the variance are represented by the blue points. Among the first 40 logS, the numbers 2, 4, 7, 11, 13, 18, 21, 28, 32, 37, and 40 exhibit high variance. Hence, the third approach considers the use of the high variance logS as input for the regression models.

Principal Component Analysis (PCA) [54] is a method to identify underlying patterns and transform multi-dimensional datasets into sets of uncorrelated variables, called principal components [55], which exhibit high variance.

Approaches 2 and 4 employ PCA to transform the elementary and high variance logS, respectively, in a new set of input parameters (for the regression models). The four approaches listed in Section 2 are comparatively assessed as presented in the results section.

The datasets include the following parameters for each approach: elementary logS (approach

1); principal components of elementary logS (approach 2); high variance logS (approach 3); and principal components of high variance logS (approach 4). In addition to these parameters, for all approaches, the datasets include the following parameters: engine speed, engine load, β_{SOI}, β_{fmep}, β_{Blowby}, and β_{CAC}.

These datasets are standardised to zero mean and unit variance, using the StandardScaler function [56]. In this respect, considerable differences in magnitudes of and output features are avoided.

2.2.4 Data splitting

The pre-processed datasets are divided into two sets: training and testing. The training set is used to train the regression (ANN) models for each approach, whilst the testing set is reserved (from the training phase) to test those models. The ratio of test to train datasets is determined by employing the following equation:

$$\gamma = \frac{F_{\text{Test}}}{F_{\text{Total}}} \tag{9}$$

where F_{Test} denotes the test datasets, and F_{Total} denotes the complete datasets. This study employs $\gamma = 0.15$ to separate the test data from the complete data-set.

2.3 Regression model

Regression models are mathematical representations of the relationships between the input features and the continuous target variables. As several regression techniques including linear, polynomial, support vector, decision trees, and artificial neural networks are available, the selection of the most appropriate technique is crucial. Comparison of different regression techniques on the same datasets were reported in previous authors' studies [57][58]. Therein, regression based on artificial neural networks (ANN) was found to be the most accurate and robust technique. Hence, ANNs are selected to develop regression models.

ANN consists of an input layer (accepting input features) and an output layer (estimating output parameters) with hidden layers between them. The selection of input and output parameters depends on the features and target variables, respectively.

Several ANNs were set up corresponding to the considered approaches. The ANNs output parameters (target variables) include the severity (β) for each fault. The input parameters (features) include the engine speed, and power, as well as the log-signatures (elementary or high variance) or their principal components, depending on the approach (1 to 4). These input and

Table 2: Parameters of the feed forward Artificial Neural Network (ANN) used for the fault severity estimation.

Parameter	**Value**
Optimiser	Adam
Batch size	30
Hidden layers	2
Nodes per hidden layer	10
Activation function (hidden layer)	Exponential Linear Unit (ELU)
Activation function (output layer)	Linear
Learning rate	0.001

output parameters are standardised as described in the previous section. However, a typical selection technique does not exist to determine the hidden layers and their neurons numbers. Therefore, these parameters are determined by testing several combinations of the hidden layers number (1, 2, and 3) and neurons number per layer (5, 10, and 15) for the ANN with 10 input parameters targeting to minimise the faults severity estimation errors [58], according to the guidelines reported in Xu and Chen [59]. As the input parameters number for the different ANN does not vary substantially between the different approaches, the selected (optimal) ANN structure was employed for the four approaches.

Figure 5 presents the selected optimal structure of the developed ANNs for estimating the faults severity. The same structure is used to compare the different approaches described in the pre-processing phase. The ANNs optimised parameters are listed in Table 2. The Exponential Linear Unit (ELU) function is used as the activation function for the hidden layers due to its fast convergence [60]. As the fault severity is a continuous variable (Section 2.1), the linear function is used for the output layer.

The training datasets (that were separated in Phase 2) are further classified into training and validation datasets. The training datasets are used with a batch size of 30 samples to calculate (fit) the weights and biases of the ANNs. The Adam back-propagation algorithm reported in [61, 62] was employed to minimise the prediction error (using the testing data). To avoid over-fitting [63], cross validation is performed through several iterations by further separating the training data in training and validation datasets (85% training and 15% validation) in each iteration. The error monitored during these iterations is presented in the results section. The

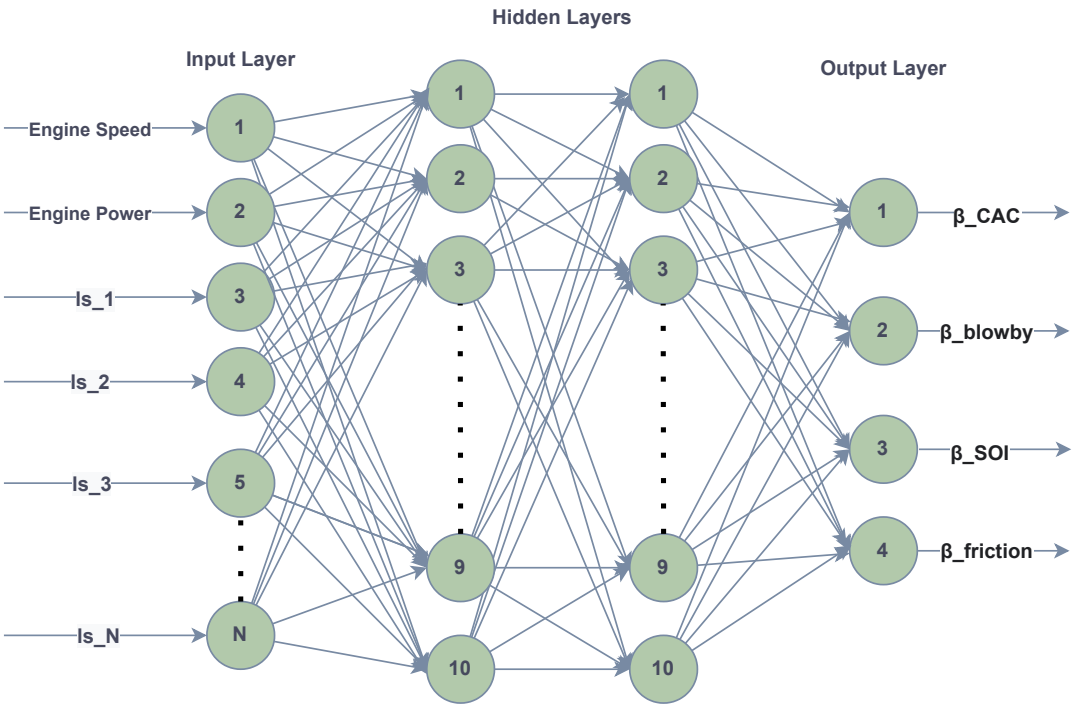


Figure 5: Structure of the Artificial Neural Network (Multi Layer Perceptron) used for the estimation of the four considered faults and their severity

loss function representing the prediction error is expressed by the following equation:

$$e_{\text{ANN}} = \underset{\theta_{cyl}}{argmin} \left\{ \sqrt{\sum_{i=1}^{D} (\beta_{\text{fault},cyl,i} - \beta_{\text{fault},cyl,i}(C_i, \theta))^2} \right\} \tag{10}$$

where C_i denotes the ANN input (logS, engine speed and power) for the i^{th} data-set, θ represents the ANN weights, D is the total training datasets number, $\beta_{\text{fault,cyl},i}$ denotes the training data-set fault severity, whereas $\beta_{\text{fault,cyl}}(C_i, \theta)$ denotes the predicted fault severity.

The final adjusted values of weights and biases at each node, which provide the minimum error when employing the validation datasets, are determined at the end of the training process considering 200 epochs (total number of times the ANN uses the data-set). The error during the cross-validation iterations over 200 epochs for approaches 1–4 are presented and discussed in the results section.

The Standard Error (SE) is calculated by subtracting the predicted values (estimated by the regression model) from the actual values, according to the following equation:

$$\text{SE} = y_i - \hat{y}_i \tag{11}$$

where y_i is the actual value of β_{fault}, $\hat{y}_i$ is the predicted value of β_{fault}

The Adam optimiser minimises the Mean Square Error (MSE) to identify the ANN weights

and biases. The MSE is calculated by the following equation:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{12}$$

The Root Mean Square Error (RMSE) is calculated according to the following equation to compare the ANNs overall performance:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

The coefficient of determination (R^2) is also used to compare the accuracy of ANNS for each approach to estimate faults on validation data. The R^2 is calculated according to the following equation:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{14}$$

3 Results and discussion

The RMSE variation for the ANNs employed in the parametric study are presented in Figure 6; the training phase results appear on the right, whereas the validation phase results are shown on the left. The levels of the log signatures from 2 to 6 resulted into extraction of 5 to 25 logS, respectively. When the logS obtained from level 2 and 3 are considered as input (to the respective ANNs), the RMSE becomes saturated after 50 and 100 epochs respectively, which is attributed to the limited information provided by these logS. The information supplied through deeper logS (with level more than 4) reduces the RMSE, showcasing potential for its further reduction when in greater epochs number. To avoid over-fitting the training data, the RMSE of the validation process is monitored, which also exhibits a similar behaviour. However, the RMSE calculated during the validation phase becomes unstable as the logS level increases

Table 3: Characteristics of the ANNs corresponding to the 1st approach for different logS levels.

Level (d)	**No of signatures (n)**	**No of Inputs (N)**	**Hyper-Parameters**	**Training time (s)**	**RMSE on validation data**
2	3	5	214	399.3	0.17
3	5	7	234	427.9	0.11
4	**8**	**10**	**264**	**443.8**	**0.095**
5	14	16	324	453.3	0.079
6	23	25	414	614.7	0.071

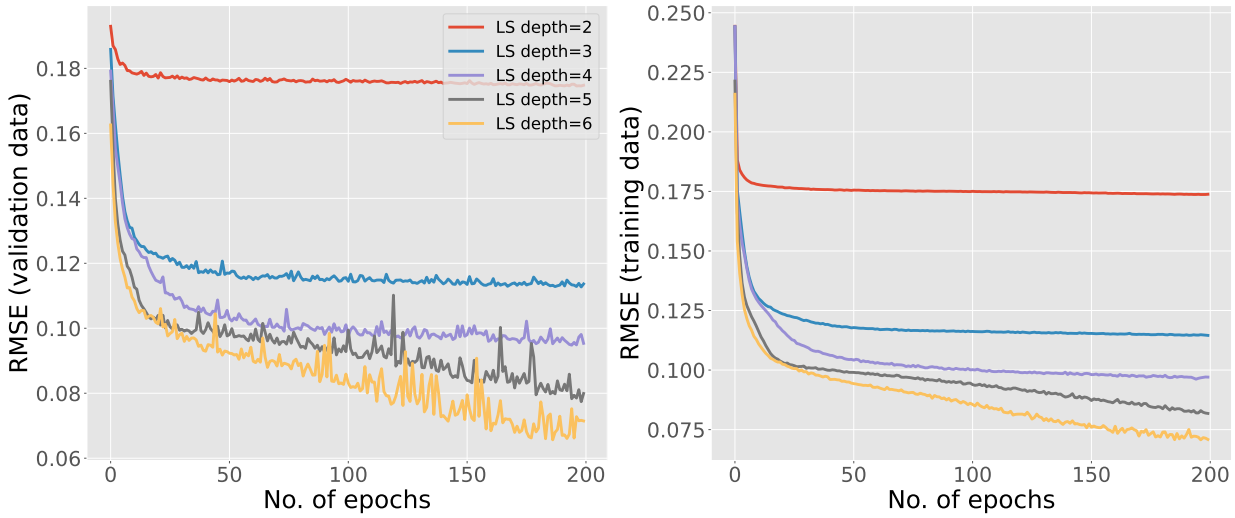


Figure 6: RMSE variation of ANNs corresponding to the 1st approach for the investigated marine engine faults estimation during the ANN validation phase (left), and training phase (right) over 200 epochs with varying logS level.

Table 3 presents the characteristics (input parameters, number of hyper-parameters) of the developed ANNs considering different logS levels, as well as the required training time and the calculated RMSE by using the validation datasets. It is evident that greater logS level increases the number of the input parameters (logS, engine speed and power) and the training time, whereas the RMSE reduces. A 5% increase of the training time with around 15% RMSE reduction is found between ANNs with logS levels up to 5. However, logS levels beyond 5 considerably prolongs the ANN training duration, exhibiting though 5% RMSE reduction.

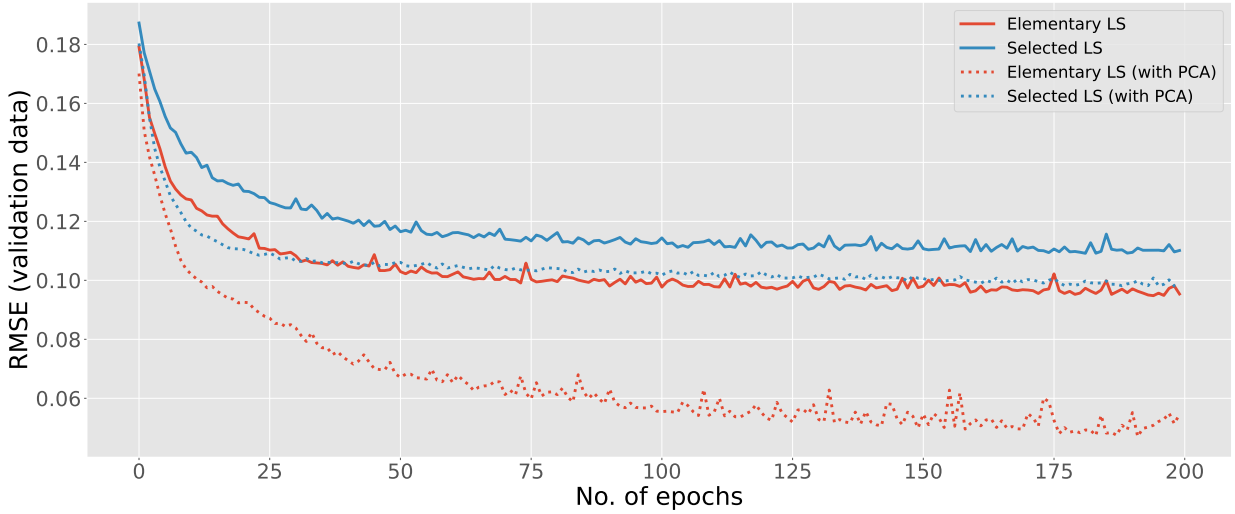


Figure 7: RMSE variation of the ANNs corresponding to four approaches on validation data-set over 200 epochs.

Based on these findings, level 4 corresponding to 8 elementary logS (n*=8) is selected to

further compare the ANNs performance considering the four approaches listed in Phase 2. Figure 7 illustrates the RMSE for the validation datasets of the ANNs for these approaches. These ANNs employed the same number of input parameters (8 logS, engine speed and power). The RMSE of the ANN using (as input) the elemental logS is 15% lower compared to the RMSE of the ANN using the selected high variance logS (following the variance analysis). This showcases that the selected logS with high variance do not contribute to the fault diagnosis compared to similar number of elementary logS. This implies that the elementary logS collectively provide the required information pertinent to the considered four faults. Figure 7 also provides the RMSE for the approaches 2 and 4 with the use of the 8 logS principal components. It is inferred that the use of PCA improves the ANN accuracy (lower RMSE) for both approaches (elementary and selected logS). Overall, these results indicate that approach 2 (elementary logS with PCA) is more effective exhibiting the lowest RMSE.

Table 4 summarises the performance of ANNs for the four approaches. Approach 2 (elementary logS with PCA) exhibits around 80% reduction in the training time associated with 41% RMSE reduction compared with first approach (without PCA). Approach 4 (selected logS with PCA) exhibits only around 11% RMSE reduction with almost same training time compared to approach 3 (selected logS). This implies that approach 2 (elementary logS with PCA) is the more effective in terms of capturing the datasets information and improve the regression model accuracy for estimating the faults severity.

Table 4: Characteristics of regression models (ANNs) pertaining to the four approaches listed in Phase 2.

No	**Approach Description**	**No of Inputs (N)**	**Trainable Parameters**	**Training time (s)**	**RMSE on validation data**	**R^2 on validation data**
1	Elementary logS	10	264	443.8	0.095	0.78
2	Elementary logS with PCA	10	264	96.39	0.055	0.92
3	Selected logS	10	264	94.90	0.11	0.70
4	Selected logS with PCA	10	264	93.70	0.098	0.77

The developed ANNs corresponding to the four approaches are further tested on 15% of the testing datasets (which were separated in Phase 2). Figure 8 presents the overall RMSE (bars) calculated by employing eq. (13) considering the predicted faults severity (with combinations of the four faults) at particular engine cylinders, whereas the whiskers represent the 95% confidence interval of the standard error calculated by eq. (11) for each approach. Similarly to the findings discussed in the preceding paragraphs (corresponding to the ANNs training performance), approach 2 (elementary logS with PCA) is found to be the most effective, as it exhibits the lowest

overall RMSE (0.059 compared to 0.095–0.115 of the other approaches). Moreover, approach 2 exhibits the lowest 95% confidence interval fo the standard error.

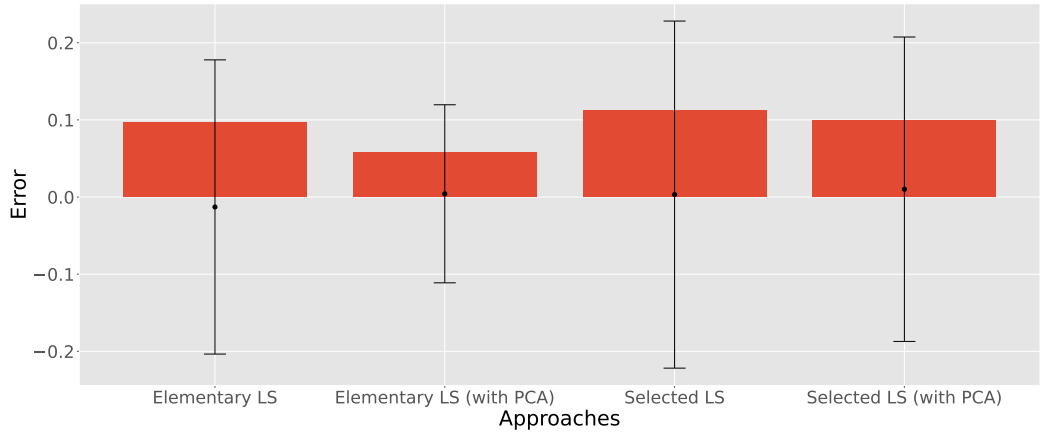


Figure 8: Estimated error metrics for the ANNs corresponding to approaches 1–4 for considering the four faults for testing datasets. Bars denote RMSE, whereas whiskers represent the 95% confidence interval of the standard error.

Figure 9 presents the predicted error metrics on each fault severity (β_{fault}) for the ANNs corresponding to approaches 1–4. The bars represent the RMSE, whereas the whiskers represent the 95% confidence interval of the standard error. These results also confirm that the approach 2 (elementary logS with PCA) exhibits the lowest RMSE in all the considered faults. However, it must be noted that RMSE for the faults of friction, blowby and start of injection exhibits the lowest values (in the range 0.02–0.03 for Approach 2), whereas the RMSE for the charge air cooler fault is around 0.1. This is attributed to the fact that the first three faults directly impact the engine in-cylinder pressure signals. The developed ANNs predictions exhibit lower accuracy compared to the other faults, as the charge air cooler fouling indirectly impacts the in-cylinder pressure.

The standard error (SE) represents the difference between the target and predicted values of the faults severity. The SE confidence intervals trade-offs (as function of the approach) follow the respective RMSE trade-offs. Approach 2 (elementary logS with PCA) exhibits the smallest SE 95% confidence interval. The SE 95% confidence interval average values are offset (both positive and negative offsets were estimated). It is reaffirmed that using PCA improves the ANNs accuracy (reduction of both RMSE and SE). Hence, this study recommends the PCA use along with logS.

Based on the previous findings, this study recommends the use of an integrated approach to develop data driven models for marine engines fault diagnosis as presented in Figure 1. This approach fuses the logarithmic signatures as the feature reduction method, the Principal

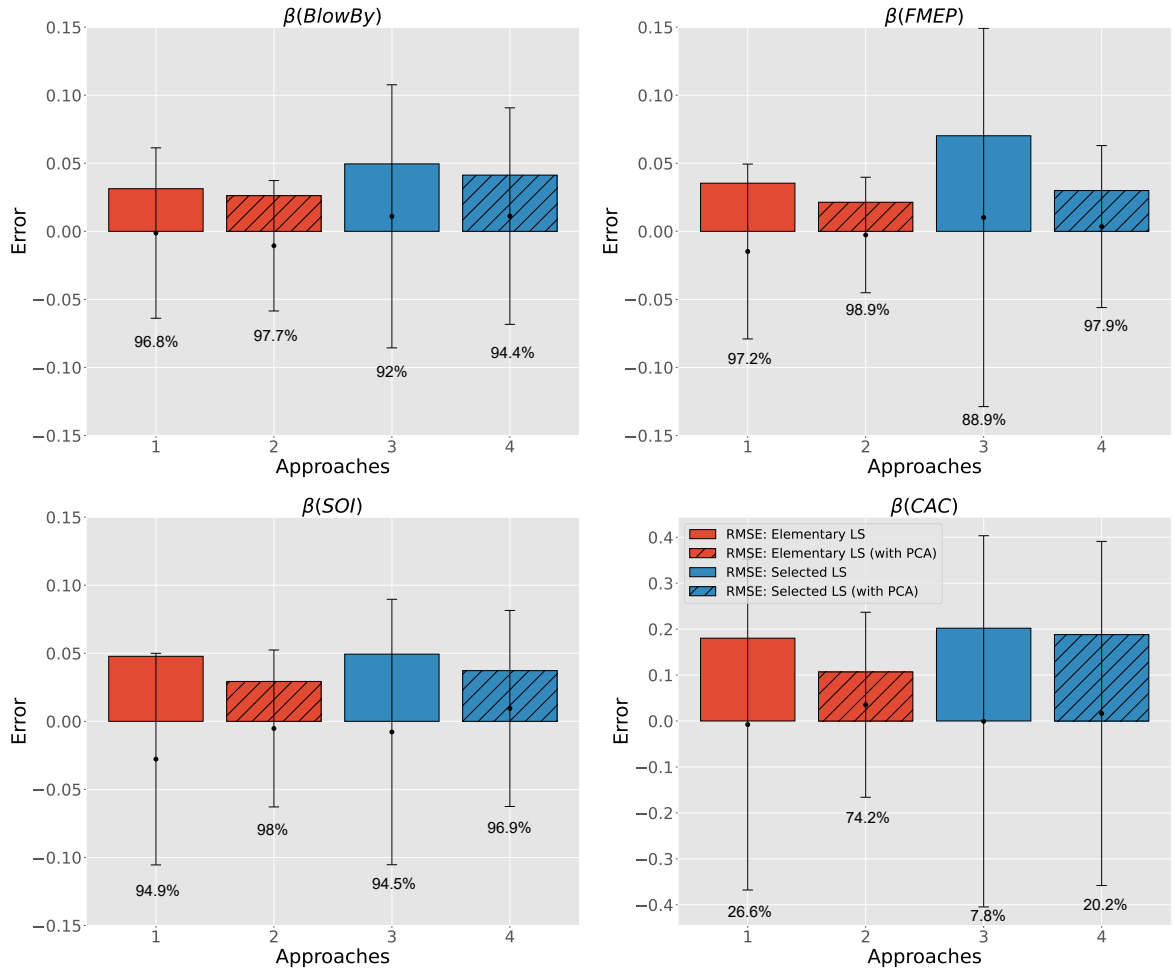


Figure 9: Estimated error metrics for ANNs corresponding to approaches 1–4 for each fault by using the testing datasets. The bars denote RMSE, whereas the whiskers represent the 95% confidence interval of the standard error. Accuracy for each approach in percentage (R^2) below every bar.

Components Analysis as a method to transform the logS to variable with high variance, the input/output parameters standardisation, and the ANN as the regression method to estimate the faults severity.

The proposed data-driven model consists of two blocks; the first deals with the extraction of information from the in-cylinder pressure signals (using logS, PCA, and standardisation), whereas the second consists a typical ANN, which is employed as the regression model to estimate the faults severity. For marine engines similar to the one investigated herein, the ANN structure is expected to follow the optimal found in this study (2 hidden layers, 10 neurons per layer). However, future studies can investigate other structures or modern regression methods to further minimise the prediction errors. This model can be developed by using labeled data sets generated by a digital twin or synthetic datasets (measured and simulated data). Following model development, its practical application includes the use of measured in-cylinder pressure,

engine speed, and load as input parameters, to estimate engine faults and their severity, as proposed in the bottom part of the flow chart shown in Figure 1. It must be noted though that pre-processing of the measured engine parameters are needed to remove offsets and outliers prior to their use (as input) in fault estimation data-driven models.

4 Study implications

The study demonstrates that the extraction of log signatures effectively captures information from the pressure signals in the cylinder to support engine fault diagnosis. Hence, it is expected that elementary logS processed with PCA is an effective method in addressing the need for reducing large datasets, whilst retaining the conveyed information of the original datasets. Simple regression techniques (such as typical ANNs) facilitate the use of the extracted logS to estimate specific fault severity, thus providing a computationally effective approach for ship machinery diagnosis applications.

The selected hybrid approach (Approach 2) demonstrated herein for marine engines fault diagnosis, which employed techniques including logS, PCA, standardisation, and ANNs, exhibited R^2 values exceeding 97% for estimating each fault (alone or in presence of other faults) in the entire operating envelope) shown in Figure 9. The pertinent literature (Section 1.2) reported that state-of-the-art (SOTA) algorithms exhibited R^2 in the range $94-97\%$ for estimating single faults at specific operating points. Hence, it is inferred that the selected approach outperforms SOTA algorithms. However, future studies are recommended to provide specific comparative assessments between the proposed approach and SOTA algorithms.

The selected hybrid approach will allow the development of intelligent health monitoring and management systems for the next-generation ships machinery, which are expected to use a high level of autonomy and perform partially crewed operations even without crews. These approaches and methods can be extended to other internal combustion engine types including propulsion and electricity generation, providing that in-cylinder pressure signals are available, thus leading to the development of effective tools for these engines' health management.

The limitations of this study include verification of the fault severity estimation model in measured and / or published data sets, consideration of other components faults, and consideration of different fault severity levels in each engine cylinder. These limitations need to be addressed in future studies. Despite the preceding limitations, this study impact is considerable, as it provides: (a) insights on the use of path logarithmic signatures to marine engines fault diagnostics applications, and (b) recommendations to develop data-driven faults severity

estimation models.

5 Conclusions

This study investigated the applicability of logarithmic signatures (logS) for the fault severity estimation of marine engines using only in-cylinder pressure signals and the engine operating point (speed and load). The in-cylinder pressure signals were generated using the digital twin of the investigated marine engine. Four pre-processing approaches to generate labelled datasets with the regression models input and output parameters were comparatively assessed to quantify their impact on the regression models accuracy. These approaches included the use of elementary or high variance logS without/with the use of Principal Component Analysis (PCA), and additionally employing features (parameters) standardisation. Regression models based on typical ANN were developed for each approach.

The main findings of this study are summarised as follows.

1. The engine fault severity estimation accuracy increases with the log signatures level. The regression ANN with log signatures level up to five as input exhibited 5% increase of the required training time and around 15% root mean square error reduction.

2. The trade-off between accuracy and training time for the regression ANN was determined. The log signatures with level four (eight log signatures) were found to provide the highest accuracy.

3. The second pre-processing approach, using eight elementary log signatures as input along with principle component analysis is found to be the most effective.

4. The overall fault severity estimation error (Root Mean Square Error) for the second approach was found around 0.05, whereas the 95% confidence interval of the standard error was estimated ± 0.1 for testing data-set.

5. The severity of faults originating from combustion chamber, including blowby, friction, and start of injection change were estimated with 0.025 root mean square error and ± 0.05 for the 95% confidence interval of the standard error for the testing data-set.

This study considered the same faults severity for all nine engine cylinders, whereas only a limited number of measured in-cylinder pressure signals corresponding to healthy conditions and specific operating points were employed. This is the first study reported in the relevant literature

that addresses the estimation of the fault severity; therefore, comparison of the results with reference datasets is not performed as such datasets are not available. Moreover, the comparison of the logS with traditional information extraction methods such as Fourier transform, Hilbert transform, wavelet, and Cepstral analysis is proposed for future studies to quantify the impact on the fault severity prediction. This study affirmed that the use of in-cylinder pressure signals supports the diagnosis and estimation of severity of faults and their combinations for marine engines. However, to address faults on other engine components (including charge air cooler and turbocharger), additional datasets are required as input. Future studies could investigate combinations of faulty cylinders, the use of additional datasets for the severity estimation of marine engines faults (and their combinations), alternative information extraction methods, and bench-marking the developed methods against reference or experimentally obtained datasets.

Acknowledgements

This study was carried out in the framework of the i-HEATS project, which is funded by the Innovate UK Smart Grants under the grant agreement No 99958. The authors affiliated with the MSRC also greatly acknowledge the funding from DNV AS and RCCL for the MSRC establishment and operation. The opinions expressed herein are those of the authors and should not be construed to reflect the views of Innovate UK, DNV AS and RCCL.

Terry Lyons was funded in part by the EPSRC [grant number EP/S026347/1], in part by The Alan Turing Institute under the EPSRC grant EP/N510129/1, the Data Centric Engineering Programme (under the Lloyd's Register Foundation grant G0095), the Defence and Security Programme (funded by the UK Government) and the Office for National Statistics & The Alan Turing Institute (strategic partnership) and in part by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA).

References

[1] M Meier, G Sudwoj, P Theodossopoulos, E Tzanos, and I Karakas. A real time comprehensive analysis of the main engine and ship data for creating value to ship operators. In *Proceedings of the 29th CIMAC World Congress on Combustion Engine Technology, Vancouver, BC, Canada*, pages 10–14, 2019.

[2] Yaqiong Lv, Xueting Yang, Yifan Li, Jialun Liu, and Shijie Li. Fault detection and diagnosis of marine diesel engines: A systematic review. *Ocean Engineering*, 294:116798, 2024. ISSN 0029-8018. doi: https://doi.org/10.1016/j.oceaneng.2024.116798. URL https://www.sciencedirect.com/science/article/pii/S0029801824001355.

[3] Yu Guo, Jundong Zhang, Bin Sun, and Yongkang Wang. A universal fault diagnosis framework for marine machinery based on domain adaptation. *Ocean Engineering*, 302: 117729, 2024. ISSN 0029-8018. doi: https://doi.org/10.1016/j.oceaneng.2024.117729. URL https://www.sciencedirect.com/science/article/pii/S0029801824010667.

[4] Peng Zhang, Chaozhe Li, Huanyun Xu, Yongjiu Zou, Kai Wang, Yuewen Zhang, and Peiting Sun. Bi-aae: A binary adversarial autoencoder deep neural network model for anomaly detection in system-levels marine diesel engines. *Ocean Engineering*, 302:117700, 2024.

[5] Nikos Kougiatsos and Vasso Reppa. A distributed cyber-physical framework for sensor fault diagnosis of marine internal combustion engines. *IEEE Transactions on Control Systems Technology*, 2024.

[6] Andrea Coraddu, Luca Oneto, Davide Ilardi, Sokratis Stoumpos, and Gerasimos Theotokatos. Marine dual fuel engines monitoring in the wild through weakly supervised data analytics. *Engineering Applications of Artificial Intelligence*, 100:104179, 2021. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2021.104179.

[7] Francesca Calabrese, Alberto Regattieri, Lucia Botti, Cristina Mora, and Francesco Gabriele Galizia. Unsupervised fault detection and prediction of remaining useful life for online prognostic health management of mechanical systems. *Applied Sciences*, 10(12):4120, 2020.

[8] Wenkui Xi, Zhixiong Li, Zhe Tian, and Zhihe Duan. A feature extraction and visualization method for fault detection of marine diesel engines. *Measurement: Journal of the International Measurement Confederation*, 116:429–437, 2 2018. ISSN 02632241. doi: 10.1016/j.measurement.2017.11.035.

[9] G Tsaganos, N Nikitakos, D Dalaklis, AI Ölcer, and D Papachristos. Machine learning algorithms in shipping: improving engine fault detection and diagnosis via ensemble methods. *WMU Journal of Maritime Affairs*, 19(1):51–72, 2020.

[10] Jerzy Kowalski, Bartosz Krawczyk, and Michał Woźniak. Fault diagnosis of marine 4-stroke diesel engines using a one-vs-one extreme learning ensemble. *Engineering Applications of Artificial Intelligence*, 57:134–141, 1 2017. ISSN 09521976. doi: 10.1016/j.engappai.2016.10.015.

[11] Chengtao Cai, Xiangyu Weng, and Chuanbin Zhang. A novel approach for marine diesel

engine fault diagnosis. *Cluster Computing*, 20:1691–1702, 6 2017. ISSN 15737543. doi: 10.1007/s10586-017-0748-0.

[12] Sibo Wang, Jin Wang, and Ruihan Wang. A novel scheme for intelligent fault diagnosis of marine diesel engine using the multi-information fusion technology. In *IOP Conference Series: Materials Science and Engineering*, volume 782. Institute of Physics Publishing, 4 2020. doi: 10.1088/1757-899X/782/3/032022.

[13] Yanyou Chai, Xiuyan Peng, Liufeng Xu, and Jiuyu Shi. Research on fault diagnosis of marine diesel engine based on integrated similarity. *CCIS*, 227:678–685, 2011.

[14] Lokukaluge P. Perera. Marine engine centered localized models for sensor fault detection under ship performance monitoring. In *IFAC-PapersOnLine*, volume 49, pages 91–96. Elsevier B.V., 2016. doi: 10.1016/j.ifacol.2016.11.016.

[15] Bo Jiang, Si Chen, Beibei Wang, and Bin Luo. Mglnn: Semi-supervised learning via multiple graph cooperative learning neural networks. *Neural Networks*, 153:204–214, 2022.

[16] Arunabha M. Roy and Jayabrata Bhaduri. Densesph-yolov5: An automated damage detection model based on densenet and swin-transformer prediction head-enabled yolov5 with attention mechanism. *Advanced Engineering Informatics*, 56:102007, 2023. ISSN 1474-0346. doi: https://doi.org/10.1016/j.aei.2023.102007. URL https://www.sciencedirect.com/science/article/pii/S1474034623001350.

[17] Christian Velasco-Gallego and Iraklis Lazakis. Development of a time series imaging approach for fault classification of marine systems. *Ocean Engineering*, 263:112297, 2022.

[18] Zeng Meng, Qiaochu Qian, Mengqiang Xu, Bo Yu, Ali Rıza Yıldız, and Seyedali Mirjalili. Pinn-form: A new physics-informed neural network for reliability analysis with partial differential equation. *Computer Methods in Applied Mechanics and Engineering*, 414:116172, 2023.

[19] Xiaona Song, Peng Sun, Shuai Song, and Vladimir Stojanovic. Finite-time adaptive neural resilient dsc for fractional-order nonlinear large-scale systems against sensor-actuator faults. *Nonlinear dynamics*, 111(13):12181–12196, 2023.

[20] Xiaona Song, Chenglin Wu, Shuai Song, Vladimir Stojanovic, and Inés Tejado. Fuzzy wavelet neural adaptive finite-time self-triggered fault-tolerant control for a quadrotor unmanned aerial vehicle with scheduled performance. *Engineering Applications of Artificial Intelligence*, 131:107832, 2024.

[21] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[22] Naomi Altman and Martin Krzywinski. The curse (s) of dimensionality. *Nat Methods*, 15 (6):399–400, 2018.

[23] Zhixiong Li, Xinping Yan, Chengqing Yuan, and Zhongxiao Peng. Intelligent fault diagnosis method for marine diesel engines using instantaneous angular speed. *Journal of Mechanical Science and Technology*, 26:2413–2423, 2012.

[24] Zhenyi Kuai and Guoyong Huang. Fault diagnosis of diesel engine valve clearance based on wavelet packet decomposition and neural networks. *Electronics*, 12(2):353, 2023.

[25] Jihui Wang, Hui Cao, Zhichao Cui, Zeren Ai, and Kuo Jiang. Intelligent fault diagnosis of marine diesel engines based on efficient channel attention-improved convolutional neural networks. *Processes*, 11(12), 2023. ISSN 2227-9717. doi: 10.3390/pr11123360. URL https://www.mdpi.com/2227-9717/11/12/3360.

[26] Jiawei Jiang, Yihuai Hu, Yanzhen Chen, and Guohua Yan. Feature extraction and intelligent fault diagnosis of marine machinery. *Journal of Vibration Engineering & Technologies*, 12 (1):201–211, 2024.

[27] Muhittin Orhan and Metin Celik. A literature review and future research agenda on fault detection and diagnosis studies in marine machinery systems. *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment*, 0 (0):14750902221149291, 0. doi: 10.1177/14750902221149291. URL https://doi.org/10.1177/14750902221149291.

[28] Denys P. Viana, Dionísio H. C. de Sá Só Martins, Amaro A. de Lima, Fabrício Silva, Milena F. Pinto, Ricardo H. R. Gutiérrez, Ulisses A. Monteiro, Luiz A. Vaz, Thiago Prego, Fabio A. A. Andrade, Luís Tarrataca, and Diego B. Haddad. Diesel engine fault prediction using artificial intelligence regression methods. *Machines*, 11(5), 2023. ISSN 2075-1702. doi: 10.3390/machines11050530. URL https://www.mdpi.com/2075-1702/11/5/530.

[29] Habil István Lakatos and Peter Dely. In-cylinder pressure indication of internal combustion engines for diagnostic purposes. In *2015 IEEE 13th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 37–40, 2015. doi: 10.1109/SAMI.2015.7061902.

[30] Terry J Lyons. Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2):215–310, 1998.

[31] Terry Lyons. Rough paths, signatures and the modelling of functions on streams. *arXiv preprint arXiv:1405.4537*, 2014.

[32] Terry Lyons and Andrew D McLeod. Signature methods in machine learning. *arXiv preprint arXiv:2206.14674*, 2022.

[33] Ilya Chevyrev and Andrey Kormilitzin. A primer on the signature method in machine learning. *arXiv preprint arXiv:1603.03788*, 2016.

[34] Zecheng Xie, Zenghui Sun, Lianwen Jin, Hao Ni, and Terry Lyons. Learning spatial-semantic context with fully convolutional recurrent network for online handwritten chinese text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1903–1917, 2017.

[35] James Morrill, Andrey Kormilitzin, Alejo Nevado-Holgado, Sumanth Swaminathan, Sam Howison, and Terry Lyons. The signature-based model for early detection of sepsis from electronic health records in the intensive care unit. In *2019 Computing in Cardiology (CinC)*, pages Page–1. IEEE, 2019.

[36] James H Morrill, Andrey Kormilitzin, Alejo J Nevado-Holgado, Sumanth Swaminathan, Samuel D Howison, and Terry J Lyons. Utilization of the signature method to identify the early onset of sepsis from multivariate physiological time series in critical care monitoring. *Critical Care Medicine*, 48(10):e976–e981, 2020.

[37] Shujian Liao. *Log signatures in machine learning*. PhD thesis, UCL (University College London), 2022.

[38] Patric Bonnier, Patrick Kidger, Imanol Perez Arribas, Cristopher Salvi, and Terry Lyons. Deep signatures. 2019.

[39] Wangying Chen, Mingyue Yu, and Peng Wu. Rotor–stator rub-impact fault and position identification of aero-engine based on vmd-mf-cepstrum-knn. *Tribology Transactions*, 66(1): 23–34, 2023.

[40] Simon Rit, Rolf Clackdoyle, Jan Hoskovec, and Jean Michel Létang. List-mode proton ct reconstruction using their most likely paths via the finite hilbert transform of the derivative of the backprojection. *Fully3D, vol. Submitted*, pages 324–327, 2015.

[41] Wartsila. Project guide for marine applications: Wartsila 46c, 2001. Accessed: 2022-10-06.

[42] Konstantinos-Marios Tsitsilonis and Gerasimos Theotokatos. A novel method for in-cylinder pressure prediction using the engine instantaneous crankshaft torque. *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment*, 236(1):131–149, 2022.

[43] VT Lamaris and DT Hountalas. A general purpose diagnostic technique for marine diesel engines–application on the main propulsion and auxiliary diesel units of a marine vessel. *Energy conversion and management*, 51(4):740–753, 2010.

[44] Dimitrios T Hountalas. Prediction of marine diesel engine performance under fault conditions. *Applied thermal engineering*, 20(18):1753–1783, 2000.

[45] Jose Antonio Pagán Rubio, Francisco Vera-García, Jose Hernandez Grau, Jose Muñoz Cámara, and Daniel Albaladejo Hernandez. Marine diesel engine failure simulator based on thermodynamic model. *Applied Thermal Engineering*, 144:982–995, 2018.

[46] Marcellin Perceau, Philippe Guibert, Adrian Clenci, Victor Iorga-Simăn, Mihai Niculae, and Stéphane Guilain. Investigation of the aerodynamic performance of the miller cycle from transparent engine experiments and cfd simulations. *Machines*, 10(6):467, 2022.

[47] Marco Cavazzuti. Design of experiments. In *Optimization methods*, pages 13–42. Springer, 2013.

[48] Terry Lyons. Rough paths, signatures and the modelling of functions on streams, 2014. URL https://arxiv.org/abs/1405.4537.

[49] Shujian Liao, Terry Lyons, Weixin Yang, and Hao Ni. Learning stochastic differential equations using rnn with log signature features, 2019. URL https://arxiv.org/abs/1908.08286.

[50] David Maxwell and Terry Lyons. Esig, 2020. URL https://github.com/maxwelld90/esig_documentation.

[51] Jeremy Reizenstein and Benjamin Graham. The iisignature library: efficient calculation of iterated-integral signatures and log signatures, 2018. URL https://arxiv.org/abs/1802.08252.

[52] Patrick Kidger and Terry Lyons. Signatory: differentiable computations of the signature and logsignature transforms, on both cpu and gpu, 2020. URL https://arxiv.org/abs/2001.00706.

[53] Xiaofei He, Ming Ji, Chiyuan Zhang, and Hujun Bao. A variance minimization criterion to feature selection using laplacian regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:2013–2025, 2011.

[54] Sara Sharifzadeh, Ali Ghodsi, Line H Clemmensen, and Bjarne K Ersbøll. Sparse supervised principal component analysis (sspca) for dimension reduction and variable selection. *Engineering Applications of Artificial Intelligence*, 65:168–177, 2017.

[55] Karol Warne, Girijesh Prasad, Sina Rezvani, and L Maguire. Statistical and computational intelligence techniques for inferential model development: a comparative evaluation and a novel proposition for fusion. *Engineering applications of artificial intelligence*, 17(8):871–885, 2004.

[56] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[57] Chaitanya Patil and Gerasimos Theotokatos. Comparative analysis of data-driven models for marine engine in-cylinder pressure prediction. *Machines*, 11(10), 2023. ISSN 2075-1702. doi: 10.3390/machines11100926. URL https://www.mdpi.com/2075-1702/11/10/926.

[58] Chaitanya Patil Gerasios Theotokatos. Engine fault diagnosis techniques using machine learning techniques (regression). Technical report, Naval Architecture, Ocean and Marine Engineering department, University of Strathclyde, Glasgow, UK, 2022.

[59] Shuxiang Xu and Ling Chen. A novel approach for determining the optimal number of hidden layer neurons for fnn's and its application in data mining. 2008.

[60] Gang Wang, Georgios B. Giannakis, and Jie Chen. Learning relu networks on linearly separable data: Algorithm, optimality, and generalization. *IEEE Transactions on Signal Processing*, 67(9):2357–2370, 2019. doi: 10.1109/TSP.2019.2904921.

[61] E.M. Johansson, F.U. Dowla, and D.M. Goodman. Backpropagation learning for multilayer feed-forward neural networks using the conjugate gradient method.

https://doi.org/10.1142/S0129065791000261, 02:291–301, 11 2011. ISSN 0129-0657. doi: 10.1142/S0129065791000261.

[62] Zijun Zhang. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, pages 1–2, 2018. doi: 10.1109/IWQoS.2018.8624183.

[63] Marianne Defernez and E. Katherine Kemsley. Avoiding overfitting in the analysis of high-dimensional data with artificial neural networks (anns). *Analyst*, 124:1675–1681, 1999. doi: 10.1039/A905556H. URL http://dx.doi.org/10.1039/A905556H.