Data-Driven Open Set Fault Classification and Fault Size Estimation Using Quantitative Fault Diagnosis Analysis

Andreas Lundgren and Daniel Jung

Abstract—Data-driven fault classification is complicated by imbalanced training data and unknown fault classes. Fault diagnosis of dynamic systems is done by detecting changes in time-series data, for example residuals, caused by faults or system degradation. Different fault classes can result in similar residual outputs, especially for small faults which can be difficult to distinguish from nominal system operation. Analyzing how easy it is to distinguish data from different fault classes is crucial during the design process of a diagnosis system to evaluate if classification performance requirements can be met. Here, a data-driven model of different fault classes is used based on the Kullback-Leibler divergence. This is used to develop a framework for quantitative fault diagnosis performance analysis and open set fault classification. A data-driven fault classification algorithm is proposed which can handle unknown faults and also estimate the fault size using training data from known fault scenarios. To illustrate the usefulness of the proposed methods, data have been collected from an engine test bench to illustrate the design process of a data-driven diagnosis system, including quantitative fault diagnosis analysis and evaluation of the developed open set fault classification algorithm.

Index Terms—Open set classification, Fault diagnosis, Fault estimation, Kullback-Leibler divergence, Engine fault diagnosis.

I. INTRODUCTION

Fault diagnosis of industrial systems is about detecting faults in the system by comparing model predictions of nominal system behavior and sensor data mounted on the monitored system [1]. Early detection of faults, and identifying their root cause, are important to improve system reliability and be able to select suitable counter measures at an early stage. Connected systems can make use of remote diagnosis solutions which have access to more computational capabilities and data analysis compared to what is available in an on-board diagnosis system [2]. However, this also put some restrictions, for example, on how much data that can be transmitted for fault diagnosis purposes. Two common approaches used for fault diagnosis are model-based and data-driven fault diagnosis.

Model-based fault diagnosis relies on a mathematical model of the technical system describing the nominal system behavior. Residuals are computed by comparing model predictions and sensor data to detect inconsistencies caused by faults [3]. Data-driven models use training data from different operating conditions and faulty scenarios to capture the relationship between a set of input and output data. The output data could be a feature or sensor value to be predicted, which is referred to as regression, or the class label that input data belongs to, referred to as classification [4].

Fault detection and isolation are complicated by model inaccuracies and measurement noise [5]. Developing mathematical models of technical systems is a time-consuming process which requires expert knowledge about the system to be monitored. This has motivated the use of machine learning and data-driven fault diagnosis methods to learn system behavior from collected operating data.

Designing data-driven diagnosis systems to model data from different fault classes is crucial to achieve satisfactory fault diagnosis performance. However, collecting representative data from various fault scenarios that can occur in the system is a complicated task, resulting in limited training data and unknown fault classes [6]. Therefore, conventional multi-class classification algorithms are not suitable for fault diagnosis applications since these assume that training data are representative of all data classes. Another aspect is when different faults have similar impact on the system behavior, for example when trying to classify small faults at an early stage, which will result in overlapping fault classes causing classification ambiguities. In these cases, it is not desirable to only identify the most likely fault hypothesis but to find all plausible fault hypotheses that can explain the observed data.

Quantitative fault detection and isolation analysis gives useful information during the diagnosis system design process regarding how easy it is to detect and isolate different fault classes [5]. Applying quantitative analysis early during the system development phase can be used to evaluate, for example, if fault diagnosis performance requirements can be met, which data features to use to classify different faults, and where to put additional sensors to best improve fault diagnosis performance [7].

Here, a quantitative analysis for data-driven fault classification is proposed based on the framework developed in [5]. The proposed method uses the Kullback-Leibler divergence to analyze fault detection and isolation performance for a given set of features and can be applied early in the diagnosis system design process before a classification algorithm has been selected. A second contribution is a data-driven open set fault classification algorithm based on the same framework. Time-series data are classified using the Kullback-Leibler divergence where time intervals of the signals are represented by estimated probability density functions (pdf). In addition,
a data-driven fault size estimation algorithm is proposed that can be used to estimate the fault size using training data from different fault magnitudes. The developed algorithms are evaluated using real data from an internal combustion engine test bench [1].

The outline of this paper is as follows. First, the problem formulation is presented in Section II. Related research is summarized in Section III and some background to fault diagnosis and quantitative fault detection and isolation analysis are given in Section IV. Then, the proposed framework for data-driven quantitative analysis is presented in Section V, the proposed open set fault classification algorithm in Section VI, and the fault size estimation algorithm in Section VII. The internal combustion engine case study is described in Section VIII and the results of the experiments are presented in Section IX. Finally, some conclusions are summarized in Section X.

II. PROBLEM FORMULATION

One objective in this work is to develop a data-driven method for quantitative fault detection and isolation analysis. The proposed method should be able to quantify how easy it is to detect and isolate the different classes of faults represented in training data. Let \( r_i = (r_{1, t}, r_{2, t}, ..., r_{n, t}) \) denote a sample at time \( t \) from \( n \) signals or features, for example residuals. Each sample \( r_i \) in training data belongs to one of \( m \) known fault classes \( \{f_1, f_2, ..., f_m\} \). It is assumed that training data are collected from different fault realizations with known fault magnitudes where the fault size is represented by a variable \( \theta_{i,t} \in \mathbb{R} \) for fault class \( f_i \).

Based on the proposed quantitative analysis framework, the second objective is to develop an open set classification algorithm for fault diagnosis applications. Since faults are rare events, it is expected that training data are not representative of all fault realizations. This means that there will be unknown fault scenarios caused by new realizations of known fault classes or the occurrence of previously unknown fault classes. The proposed classification algorithm should be able to detect and identify all known classes that can explain the observations but also detect when the observations are likely to belong to an unknown fault scenario. Instead of analyzing raw time-series data, different time intervals of the signals are modeled using pdfs. Representing time-series data using pdfs is useful to reduce the amount of data to be logged or transmitted when used as a remote diagnosis service while preserving relevant information about the system health. In addition, a data-driven method is developed to estimate the magnitude of a detected known fault based on training data to track the system health when no mathematical model is available to estimate the fault.

As a case study, an internal combustion engine test bench is considered in this work. Data from both nominal and different faulty system operation have been collected. To evaluate the proposed methods, a set of neural network-based residual generators described in [8], is used to compute features that will be used for data-driven classification. The proposed algorithms are evaluated using data from a set of different fault scenarios, including sensor faults and leakages.

III. RELATED RESEARCH

Quantitative fault detection and isolation analysis have been considered in, for example, [5], [9]. In [5], the Kullback-Leibler divergence is used to analyze time-discrete linear descriptor models. In [9], diagnosis performance is measured based on the distance between different kernel subspaces. One application of quantitative analysis is sensor selection, see for example [7], [10]. With respect to these previous works, a data-driven approach is proposed here for analysis of non-linear systems.

Several recent papers consider hybrid diagnosis system designs combining, e.g. residual generators and machine learning. In [11], both sensor data and residual data are used as input to a tree augmented naive Bayes fault classifier. In [12], feature selection using neural networks is applied before training the fault classifiers. In [13], model-based residuals and sensor data are used as inputs to a Bayesian network to perform fault classification and in [14] model data features are extracted and fed into a neural network classifier.

Open set classification has been considered in computer vision applications to deal with unknown classes not covered by the training data [15]. Different algorithms have been proposed to solve the open set classification problem, for example Weibull-calibrated support vector machines [16] and extreme value machines [17]. Different approaches have been proposed for open set fault classification to handle both known and unknown fault scenarios, for example, one-class support vector machines [11], conditional Gaussian network [18], and hidden Markov models [19]. With respect to previous works, an open set fault classification algorithm is proposed classifying batch data represented by pdfs, instead of classifying individual samples, using the Kullback-Leibler divergence.

Some work on data-driven fault severity estimation has been done, see for example [20], [21]. In [20], faults are assumed to appear as pulses in the time-domain data which is inherently tied to the bearing case. In [21], Paris’ formula estimating crack growth in bearings, is used to interpolate between distributions from known fault sizes. With respect to previous work, a data-driven fault size estimation algorithm is proposed based on pdfs representing feature data from different types of fault realizations.

IV. BACKGROUND

To formulate the data-driven fault diagnosis problem, the framework presented in [5] is used to model data from different fault classes and quantify fault classification performance. A summary of the relevant results and definitions is presented here.

A. Modeling Fault Classes

To capture the impact of model uncertainties and measurement noise, the feature vector \( \bar{r} \) is modeled as a random variable where its pdf is denoted \( p = p(\bar{r}) \). The pdf \( p(\bar{r}) \) varies depending on different system operating conditions, such as, operating point and fault realization. Let \( \Omega_i = \Omega_i(\bar{r}) \) denote the set of probability density functions of data \( \bar{r} \) that can be explained by fault \( f_i \), where \( p(\bar{r}) \in \Omega_i(\bar{r}) \) is used to denote...
one pdf \( p(\bar{r}) \) in the set. Then, the following definition is used to model each fault class.

**Definition 1 (Fault mode):** A fault mode is defined by the set \( \Omega_i(\bar{r}) \) of all observations \( \bar{r} \) with corresponding pdfs \( p(\bar{r}) \in \Omega_i(\bar{r}) \) that can be explained by fault class \( f_i \).

To simplify notation, \( p \) and \( \Omega_i \) are used where the dependence on \( \bar{r} \) is left out. Different fault classes \( f_i \) are represented by different sets \( \Omega_i \) where \( \Omega_{NF} \) is used to denote the fault-free case (No Fault).

The ability to detect and isolate faults is based on if there are observations that can be explained by one fault class but not another. Thus, it is possible to analyze fault detection and isolation performance by comparing the different sets \( \Omega_i \). An implication from this modeling approach is that it is not possible to isolate fault-free data from faults. This is also consistent with model-based fault isolation algorithms such as \[24\].

### B. Quantitative Fault Diagnosis Analysis

The similarities between the different fault modes, modeled by the sets \( \Omega_i \), can be used to analyze fault diagnosis performance for a given system. However, only analyzing qualitative performance, such as fault isolability in Definition 2, does not give sufficient information of how easy it is to detect and isolate different faults. One way to quantify fault diagnosis performance is to use the Kullback-Leibler (KL) divergence to measure the similarity between two pdfs.

The KL divergence is a similarity measure between pdfs defined as \[25\]:

\[
K(p||q) = \int p \log \left( \frac{p}{q} \right) dp = E_p \left[ \log \frac{p}{q} \right] \tag{1}
\]

From a fault diagnosis perspective, \(1\) can be interpreted as the expected value of a log-likelihood ratio test determining if \( \bar{r} \) is drawn from \( p \) or \( q \) when \( p \) is the true density function. If \( p \) and \( q \) are two pdf’s representing two different fault realizations, the larger the value of \( K(p||q) \) the easier it is to distinguish \( p \) from \( q \) when \( p \) is true. However, if a fault can have different realizations, a measure of how easy it is to distinguish a fault \( f_i \) with realization \( p \) from another fault \( f_j \), which could have any realization \( q \in \Omega_j \), is defined by the smallest value of \( K(p||q) \) for all \( q \in \Omega_j \). This measure is proposed in \[5\], called distinguishability, which is defined as:

\[
D_{i,j}(p) = \min_{q \in \Omega_j} K(p||q) \tag{2}
\]

If \( p \) belongs to fault \( f_i \), i.e., \( p \in \Omega_i \), the distinguishability measure \( D_{i,j}(p) \) quantifies how easy it is to isolate \( f_i \) from \( f_j \) given that \( \bar{r} \) has a pdf \( p \) where a large value corresponds to a realization that is easier to distinguish. Fault detection performance is denoted \( D_{i,NF}(p) \). The distinguishability measure \( D_{i,j}(p) \) is non-negative and zero if and only if \( p \in \Omega_j \), i.e., when it is not possible to isolate from fault class \( f_j \).

If \( p \sim \mathcal{N}(\mu_p, \Sigma_p) \) and \( q \sim \mathcal{N}(\mu_q, \Sigma_q) \) are two multivariate normal distributions with dimension \( k \) and known mean vectors, \( \mu_p, \mu_q \in \mathbb{R}^k \), and covariance matrices, \( \Sigma_p, \Sigma_q \in \mathbb{R}^{k \times k} \), \( K(p||q) \) can be computed analytically as \[26\]:

\[
K(p||q) = \frac{1}{2} \left[ \text{Tr} \left( \Sigma_q^{-1} \Sigma_p \right) + \left( \mu_q - \mu_p \right)^\top \Sigma_q^{-1} \left( \mu_q - \mu_p \right) - k + \log \left( \frac{\det \Sigma_q}{\det \Sigma_p} \right) \right]. \tag{3}
\]

### V. Approximated Distinguishability Measure Using Training Data

In many applications, the sets \( \Omega_i \) modeling each fault mode are, at least partially, unknown. This means that the distinguishability measure \(2\) cannot be used. Instead, an approximated distinguishability measure is proposed based on training data.

Training data from different fault realizations only represent a subset of all possible realizations of each fault class. Assuming samples in training data are correctly labelled, the estimated pdfs belonging to fault class \( f_i \) in training data can be used to make a lower approximation of the true fault mode \( \Omega_i \) denoted \( \hat{\Omega}_i \subseteq \Omega_i \). Then, an approximation of \(2\) can be computed as

\[
D_{i,j}(p) = \min_{q \in \hat{\Omega}_j} K(p||q) \tag{4}
\]

i.e., distinguishability is computed based on the set of already observed realizations of each fault \( f_i \), \( \hat{\Omega}_i \).

The approximate distinguishability measure \(4\) is illustrated in Figure 1 where it is evaluated for a set of pdfs \( p_1, p_2, \) and \( p_3 \) to a fault mode \( \Omega_j \). The dashed lines show each pdf \( q \) that minimizes \(4\) for each \( p_i \). The lower plot shows the computed Kullback-Leibler divergence from \( p_2 \) to all \( q \in \hat{\Omega}_j \) which, generally, increases as \( q \) are located further away from \( p_2 \).

Since \( \hat{\Omega}_j \) is a subset of \( \Omega_j \), the relation between \(2\) and estimated distinguishability in \(4\) is given by the inequality

\[
0 \leq D_{i,j}(p) \leq \hat{D}_{i,j}(p) \tag{5}
\]

The approximation \(4\) gives an upper bound of how easy it is to reject \( f_j \) for \( f_i \) given a pdf \( p \).

The inequality \(5\) is intuitive since when the set \( \hat{\Omega}_j \) is not representative of \( \Omega_j \), there is a high risk that a classifier, based on available training data, would falsely reject fault class \( f_i \) even though \( p \in \Omega_i \). By incrementally updating each fault mode \( \hat{\Omega}_i \) with new training data from new realizations of fault \( f_i \), the upper bound will become tighter. This will reduce the risk over time of falsely rejecting the true fault class. Another property of the distinguishability measure is that if \( \Omega_{NF} \subseteq \Omega_j \), then \(5\)

\[
D_{i,NF}(p) \geq \hat{D}_{i,j}(p) \tag{6}
\]

This result can be interpreted as that it is easier to detect a fault \( f_i \) than to isolate it from another fault \( f_j \).
A. Ranking of Fault Hypotheses

Assume that there are multiple fault modes and the objective is to identify the most likely fault mode given an observation with pdf $p$. A statistical test can be formulated using the generalized log-likelihood ratio test [27]

$$\lambda_{LR} = \log \frac{\max_{q_i \in \Omega_i} q_i(r)}{\max_{q_0 \in \Omega_0} q_0(r)}$$

Then the expected value of $\lambda_{LR}$ given that $r \sim p$

$$E_p[\lambda_{LR}] = \int p(x) \log \frac{\max_{q_i \in \Omega_i} q_i(x)}{\max_{q_0 \in \Omega_0} q_0(x)} dx$$

which can be reformulated as

$$E_p[\lambda_{LR}] = \int p(x) \log \frac{\max_{q_i \in \Omega_i} q_i(x)}{\max_{q_0 \in \Omega_0} q_0(x)} dx - \int p(x) \log \frac{\max_{q_0 \in \Omega_0} q_0(x)}{\max_{q_i \in \Omega_i} q_i(x)} dx - \min_{q_0 \in \Omega_0} \int p(x) \log \frac{p(x)}{q_0(x)} dx - \min_{q_i \in \Omega_i} \int p(x) \log \frac{p(x)}{q_i(x)} dx = D_0(p) - D_1(p)$$

The generalized log-likelihood ratio test $E_p[\lambda_{LR}] > 0$ when fault class $f_1$ is more likely, and $\lambda_{LR} < 0$ when fault class $f_0$ is more likely. If $p \in \Omega_0$ then $E_p[\lambda_{LR}] = 0$. By computing $D_j(p)$ for all fault classes $f_j$, selecting the class with the smallest value of $D_j(p)$ corresponds to selecting the most likely fault class based on pair-wise comparison of all faults using the generalized log-likelihood ratio test. If $D_j(p)$ is large, it means that $p$ is not likely to be explained by fault $f_j$. If $D_j(p)$ is large for all fault classes $f_j$, this means that no known fault class is likely to explain observation $p$, thus indicating the occurrence of an unknown fault class. These results give a systematic approach to compute fault hypotheses, including the unknown fault case, by evaluating and comparing the evaluated distinguishability measure $D_j(p)$ for each known fault class $f_j$ where $f_j$ is a diagnosis candidate if $D_j(p)$ is sufficiently small and an unknown fault is identified if $D_j(p)$ is large for all $f_j$.

B. Tuning of The One-Class Classifier Thresholds Using Within-Class Distinguishability

Selecting the threshold $J_i$ for each fault class models how different a pdf $p$ can be from all elements in $\Omega_i$ while still being explained by fault class $f_i$. A small threshold $J_i$ increases the risk of falsely rejecting the true fault class while a large threshold $J_i$ means that fault class $f_i$ is more likely to be a fault hypothesis increasing fault classification ambiguity.

One approach to select $J_i$ is to make sure that the majority of the samples $p \in \Omega_i$ should be explained by fault class $f_i$ if that pdf $p$ is removed from $\Omega_i$. Let

$$D_{i,j}(p) = \min_{q \in \Omega_i \setminus p} K(p||q)$$

Fig. 1: The upper plot show an illustration of the approximate distinguishability measure [4] from a set of pdfs $p_1$, $p_2$, and $p_3$ to a fault mode $\hat{\Omega}_j$. The dashed lines show each pdf $q$ that minimizes [4] for each $p_i$. The computed Kullback-Leibler divergence in the lower plot given $p_2$.
which is here referred to as within-class distinguishability. Analyzing the distribution of \( D_{1,\alpha}(p) \) for all \( p \in \Omega_i \) can be used to select a threshold. Note that (10) does not state anything about the relation between the sets \( \Omega_i \) and \( \Omega_i \) but rather give information about how scattered training data are in that fault mode.

The threshold \( J_i \) is tuned based on the distribution of the within-class distinguishability (10) for all \( p \in \Omega_i \). The distribution will have non-negative support and is here approximated using a kernel density estimation method [4] as illustrated in Figure 2. Let \( \Phi_i(x) \) denote the cumulative density estimation (cdf) of the estimated distribution and let \( \alpha \) denote a desired false alarm rate. Then, the threshold \( J_i \) is selected such that \( \Phi_i(J_i) = 1 - \alpha \). The lower plot in Figure 2 illustrate selection of a threshold corresponding to \( \alpha = 5\% \).

![Kernel density estimation of within-class distinguishability for the fault-free class](image)

C. Fault Classification Algorithm Summary

A data-driven open-set fault classifier can then be formulated as follows: The diagnosis output \( D \) of the algorithm is computed as follows: If \( D_{NF}(p) < J_{NF} \) then \( D = \{\text{NF}\} \). Otherwise, if \( D_{NF}(p) \geq J_{NF} \) then \( D = \{f_i : D_i(p) \geq J_i\} \). Otherwise if \( D = \emptyset \), then \( D = \{f_x\} \) where \( f_x \) denoted the unknown fault case.

VII. DATA-DRIVEN FAULT SIZE ESTIMATION

The method presented in Section VI provides a means to classify new data, but it does not give any information about the severity of these faults. If each training distribution \( q_i \) has a known fault size \( \theta_i \), this information can be utilized to estimate the severity of new faults by comparing how similar the data is to the training data. One approach that has been suggested in [28] is to model faults into qualitative classes, such as {normal, slight, large}. Another way, which is a method that is largely unexplored, is to find a quantitative severity estimation \( \hat{\theta} \). This study suggests a method for estimating \( \hat{\theta} \) as a convex optimization problem by using the KL divergence as a dissimilarity measure. The method is based on the following fundamental assumption. New data, collected from a fault mode \( f_i \) of severity \( \theta_i \), with distribution \( p \) should be "close" (in a KL divergence sense) to training data \( q_i \in \Omega_i \) with severity \( \theta_i \), if \( \theta_i \approx \theta_i \).

The fault size of a pdf \( p \) is estimated by finding a representation of \( p \) using a set of training distributions. Assuming that \( p \in \Omega_i \), the severity \( \theta \) of the corresponding fault class \( f_i \) is estimated as a linear combination of training distributions \( \{q_1, q_2, \ldots, q_N\} = \Omega_i \), such that the estimated distribution \( \hat{p} \) minimizes \( K(p||\hat{p}) \) where \( \hat{p} = \sum_{i=1}^{N} \lambda_i q_i \). This can be expressed as

\[
\lambda_1^*, \ldots, \lambda_N^* = \arg \min_{\lambda_1, \ldots, \lambda_N} K(p||\lambda_1 q_1 + \ldots + \lambda_N q_N) \\
\text{s.t.} \sum_{i=1}^{N} \lambda_i = 1 \\
\lambda_i \geq 0, \ \forall i
\]

The fault size estimate \( \hat{\theta} \) is then computed as a weighted sum of the fault sizes \( \theta_1 \ldots \theta_N \) corresponding to \( q_1 \ldots q_N \) as

\[
\hat{\theta} = \sum_{i=1}^{N} \lambda_i^* \theta_i. \tag{12}
\]

The estimate \( \hat{\theta} \) in (11) is obtained by using all training distributions of \( \Omega_i \) in the optimization. This is an ineffective strategy since it increases the computational cost by adding numerous distributions \( q_i \) likely to correspond to \( \lambda_j = 0 \). If data has been collected from a variety of severities and conditions, it is unlikely that new data would have a distribution that is equally similar to all available training data. Using this line of reasoning, only a small subset of all realizations are reasonably of interest for estimating \( \hat{\theta} \).

Let \( \{q_1, q_2, q_3, \ldots\} \) denote an ordered set of all elements \( q_i \in \Omega_i \) such that \( K(p||q_1) \leq K(p||q_2) \leq K(p||q_3) \leq \ldots \). Then \( \Omega_i^Q \) is defined as the first \( k \) elements in the ordered set. Thus, the number of optimization parameters can be reduced by selecting a subset \( \Omega_i^Q \subseteq \Omega_i \). The parameter \( k \leq N \) is the cardinality of \( \Omega_i^Q \) and can be considered a design parameter. If \( q_1, q_2, \ldots, q_k \) are multivariate normal distributions then \( \hat{\theta} = \sum_{i=1}^{k} \lambda_i q_i \) is a Gaussian Mixture Model [4]. The Kullback-Leibler divergence \( K(p||\hat{p}) \) has no analytical expression [29]. There are several different methods that can be used to approximate \( K(p||\hat{p}) \) numerically. For a comparison of different approximation methods see [29]. In this study, Monte Carlo sampling is used as an approximation method. The KL divergence \( K(p||\hat{p}) \) is estimated as

\[
K_{MC}(p||q) = \frac{1}{n} \sum_{i=1}^{n} \log \left( \frac{p(x_i)}{q(x_i)} \right) \tag{13}
\]

by generating a large number \( n \) of samples \( \{x_i\}_{i=1}^{n} \) from \( p \) to approximate the integration in (1). By the law of large numbers \( \lim_{n \to \infty} K_{MC}(p||q) = K(p||q) \). A closer examination of the actual upper and lower bounds of this approximation is found in [30].
Using (13) and (11) gives the updated fault size estimation algorithm:

$$\lambda_1, ..., \lambda_N = \arg \min_{\lambda_1, ..., \lambda_N} \frac{1}{n} \sum_{i=1}^{n} \log \left( \lambda_1 q_1(x_i) + ... + \lambda_k q_k(x_i) \right)$$

s.t. \( \sum_{i=1}^{N} \lambda_i = 1 \)

\( \lambda_i \geq 0, \ \forall i \)

where \( \{q_1, q_2, ..., q_n\} = \hat{\Omega}_i^Q \).

### VIII. CASE STUDY

The diagnostic framework is tested by using experimental data collected from an engine test bench. The engine is a commercial, turbo charged, four cylinder, internal combustion engine from Volvo Cars, and the test bench in question is shown in Figure 3. The sensor and actuator setup is the standard configuration for the engine. Figure 4 shows a schematic view of the engine along with the monitored signals where \( y \) denote sensor measurements and \( u \) denote actuator signals.

![Figure 3: The engine test bench which was used for data collection. The engine is a commercial four cylinder combustion engine with standard sensor and actuator configuration.](image)

![Figure 4: A schematic of the model of the air flow through the model. Available output signals are sensors \( y \) and actuators \( u \). The figure is used with permission from [31].](image)

The sensor faults were introduced by altering the sensor output gain in the engine control system. Since the errors are injected this way, the faulty signal output is actually used in the engine control scheme which gives a more realistic fault realization compared to if the error is simulated in the data using post processing. Each sensor fault is injected by multiplying the measured variable \( x_i \) by a factor \( \theta \) such that the resulting output is given as: \( y_i = (1 + \theta) \cdot x_i \) where \( \theta = 0 \) corresponds to the nominal case.

**TABLE I: Fault classes considered in the case study.** All sensor fault are induced as multiplicative faults with severities ranging from \(-20\% \) to \(+20\%\).

| Fault Class | Description |
|-------------|-------------|
| \( NF \)   | Fault-free class |
| \( f_{y\text{pim}} \) | Fault in intake manifold pressure sensor |
| \( f_{y\text{pic}} \) | Fault in intercooler pressure sensor |
| \( f_{y\text{mas}} \) | Fault in air-mass flow sensor |
| \( f_{\text{iml}} \) | Leakage in the intake manifold |

**TABLE II: Faults classes and known magnitudes represented in training data.**

| Fault Class | Fault magnitudes |
|-------------|-----------------|
| \( NF \)   | -20\%, -15\%, -10\%, -5\%, 5\%, 10\%, 15\% |
| \( f_{y\text{pim}} \) | -20\%, -15\%, -10\%, -5\%, 5\%, 10\%, 15\% |
| \( f_{y\text{pic}} \) | -20\%, -15\%, -10\%, -5\%, 5\%, 10\%, 15\%, 20\% |

The data were generated using the class 3 Worldwide harmonized Light-duty vehicles Test Cycles (WLTC), which...
is part of the World harmonized Light-duty vehicles Test Procedure (WLTP). The cycle is explained in detail in [32], and the velocity profile of the cycle is shown in Figure 5. The cycle is used since it covers a variety of operating conditions. The benefit of collecting data from different operating points is to account for any variance in the model output error due to varying speed and load. One example of why it is reasonable to assume that the model error is operating point dependent is the pressure measurements. Consider an air leakage in the intake manifold. The air mass flow through the hole would depend on the difference in pressure between the manifold and its surroundings, which at high load operating conditions would be greater than at lower loads, and thus have a greater effect on the system.

![Fig. 5: Speed profile for the Worldwide Harmonised Light Vehicles Test Procedure (WLTP) cycle.](image)

**B. Residual Generation**

The proposed method can be applied to any set of features to be used for fault diagnosis. In dynamic systems operated in various transient operating conditions, as the engine, it is complicated to use raw sensor data as features since these signals can vary significantly over time complicating fault detection of small faults. Here, a set of four residual generators \( \vec{r} = (r_1, r_2, r_3, r_4) \) was generated based on a set of Recurrent Neural Networks to filter out the system dynamics. The design process of the residual generators is described in [8]. The residual generators are generated based on a structural model representation [33] of an existing non-linear dynamic engine model, similar to the one described in [34], which models the flow of air through the engine. These residual outputs are then normalized to have zero mean in the fault-free case.

The prediction performance of two of the four residual generators are shown in Figure 6 and Figure 7, respectively. The residual generators filter out most of the system dynamics and have a small relative prediction error. To show the impact of different faults on the residual output, three of the four residuals are plotted against each other for different fault classes in Figure 8. The different faults are projected into different directions in the residual space. However, some fault classes are partially overlapping, e.g. a fault in the sensor measuring pressure after the intake manifold, \( f_{\text{pim}} \), and a leakage in the intake manifold, \( f_{\text{int}} \). It is expected that it is more difficult to distinguish between these two faults since they occur in the same part of the engine. Also, note that residual data from small faults are overlapping with the fault-free case.

![Fig. 6: The upper plot compares data from sensor \( y_{\text{pim}} \) and model prediction from a recurrent neural network. The lower plot show the resulting residual \( r_3 \).](image)

![Fig. 7: The upper plot compares data from sensor \( y_{\text{pim}} \) and model prediction from a recurrent neural network. The lower plot show the resulting residual \( r_4 \).](image)

![Fig. 8: Residual data from three of four residual generators are plotted against each other. The different colors correspond to data from different fault classes.](image)

**IX. Evaluation**

The proposed methods for quantitative fault diagnosis analysis and open-set fault classification are evaluated using data from the engine case study. First, different time segments of the multi-variate residual data are represented by estimated
multivariate normal distributions to model the different fault modes. The distinguishability measure (4) is used to evaluate fault detection and isolation performance. Finally, the proposed data-driven fault classification algorithm is evaluated, including classification of unknown faults and fault size estimation.

A. Data Processing

Data from the four residual generators are here segmented into consecutive intervals of 100 samples. For each interval, the mean and covariance matrix of a four-dimensional multivariate normal distribution is estimated. An illustration in the one-dimensional case is shown in Figure 9. The figure shows residual data and the estimated normal distribution for each interval of data.

Fig. 9: An illustration of the one-dimensional residual where a normal distribution is estimated for every 100 sample interval.

The set of estimated pdfs is then randomly split into training and validation data where 80% are used for training. Training data from each fault class are used to model each fault mode $\Omega_i$ for the different fault classes represented in training data, see Table II including pdfs from different realizations and magnitudes of each fault. Note that pdfs estimated from the fault-free case are included in all fault modes to model small faults.

B. Evaluating Fault Diagnosis Performance Using Distinguishability

The first step of the diagnosis system design process is to evaluate available data from different faults to quantify how easy it is to distinguish data from each fault class. Distinguishability is evaluated for all pdfs $p_i \in \Omega_i$ with respect to all other fault modes and the distribution of distinguishability values for different fault magnitudes is plotted in Figure 10. The subplot at position $(i,j)$ shows distinguishability of fault $f_i$ from fault $f_j$. The marks on each vertical line represent the 10%, 25%, 50%, 75%, and 90% quantiles of distinguishability values. The results show that detection and isolation performance, in general, improves with increasing fault magnitudes. It also shows that fault $f_{ypic}$ should be the easiest of the three sensor faults to distinguish while $f_{ywaf}$ is most difficult since the distinguishability measure is significantly smaller. Another observation is that the results indicate that distinguishability is not a symmetric measure, i.e. it might not be as easy to distinguish mode $f_i$ from $f_j$ as the other way around. For example, it is easier to distinguish $f_{ypic}$ from $f_{ypim}$ than vice versa since the distinguishability measure is larger. Another observation is that it is easier to distinguish each fault mode from the fault-free mode, the left most column in Figure 10 than the other fault modes, according to (6).

C. Fault Classification

The open set fault classification algorithm described in Section VI is implemented where a one-class classifier is modeled using the distinguishability measure for each known fault class, as described in Section VI. A threshold is selected based on the distribution of the within-class distinguishability (10) measure for each fault mode using kernel density estimation as illustrated in Figure 2. Each threshold for each fault mode is calibrated to have a 5% outlier rate. For comparison, the figure shows the kernel density estimation based on both training and validation data.

1) Classification of known fault classes: The set of distinguishability based one-class classifiers are first evaluated using validation data from the known fault classes. Ideally, the probability to reject a fault class should be small for all fault magnitudes of data from the true class and large for other fault classes. For comparison, two sets of one-class support vector machines (1SVM) are trained. The 1SVM classifiers are implemented using the function fitcsvm in Matlab and their kernel parameters are fit to training data using a subsampling heuristic. The first set uses the mean of the pdfs as inputs and the second set uses the raw residual data as inputs. For both sets, the decision function of each 1SVM, modeling each fault class, is calibrated to have an outlier rate of 5%. The probabilities of rejecting each fault class given data from different fault realizations using the three different open set fault classifiers are shown in Figure 11.
Classification performance is consistent with the analysis results in Figure 10, showing that the easiest fault to classify is \( f_{ypic} \), since the probabilities to reject the other fault modes is higher than the other faults. The most difficult fault to correctly isolate is \( f_{ywaf} \). When comparing the results between the three algorithms, the 1SVM classifiers based on raw residual data have the overall worst performance. However, note that the other methods use features computed from batch data. The most significant difference between the distinguishability-based classifiers and the 1SVM classifiers using residual mean is when classifying \( f_{ywaf} \). The probabilities of rejecting the other fault modes \( f_{ypic} \) and \( f_{ywaf} \) when \( f_{ypim} \) is higher for the distinguishability based classifier. The results show that the distinguishability based classifier has an overall better detection and isolation performance compared to the 1SVM classifiers.

2) Classification of unknown fault class: To evaluate detection of unknown fault classes, data from a leakage is evaluated. The distribution of the distinguishability measure \( D_i(p) \) is shown for each of the known fault modes and the corresponding thresholds for each known fault mode in Figure 12. It is shown that around 40% of the distribution of \( D_i(p) \) is exceeding the thresholds of each fault mode, which is significantly larger than the calibrated 5% outlier rate, indicating that none of the known faults can explain the leakage data. For reference, the corresponding results when evaluating on data from sensor fault \( f_{ypim} \) are shown in Figure 13. Most of the distinguishability values for the new data are only located within the threshold for fault mode \( f_{ypim} \), correctly identifying the true fault class.

### D. Fault Size Estimation

The fault estimation algorithm (11) described in Section VII is applied to validation data. The numerical Kullback-Leibler divergence (13) used in the fault size estimation of each new pdf \( p \) is estimated using 1000 samples based on the 10 pdfs in training data with the smallest \( D_i(p) \) values. The prediction error results for each fault class are shown in Figure 14. To evaluate the proposed data-driven fault size estimation algorithm (11), a histogram of the average fault size is estimated based on the 10 pdfs with the smallest \( D_i(p) \) values. The algorithm (11) has a smaller prediction error for all evaluated fault modes.

The prediction error correlates with the analysis of distinguishability between different fault modes in Figure 10. Fault \( f_{ypic} \) have a higher distinguishability to detect which indicates that the impact of that fault is significant making it easier to distinguish realizations of different magnitudes. Fault \( f_{ywaf} \) have a significantly lower distinguishability which does not change much between different fault magnitudes. This could indicate that different magnitudes of that fault have similar impact on the residual outputs thus resulting in more ambiguities when estimating the fault size. One solution to improve estimation accuracy is to averaging the estimated fault size over consecutive segments of the data.

### X. Conclusions

A data-driven framework is developed for both quantitative fault detection and isolation analysis of non-linear systems and open set fault classification. Since data from different
fault classes are overlapping, a set of one-class classifiers are designed using the Kullback-Leibler divergence as a similarity measure when evaluating if new data can be explained by that class or not. Training data are used to model the different fault modes using a set of distinguishability-based one-class classifiers. Experiments using real data from an internal combustion engine test bench show that the proposed methods are able to quantify fault detection and isolation performance and also classify both known and unknown faults, including estimation of the fault size. The proposed algorithms show promising results when compared to other data-driven methods. Interesting applications of the proposed open set fault classification algorithm are remote diagnosis and data-driven condition monitoring when it is relevant to keep the transmission rate of diagnosis data low during system operation.

REFERENCES

[1] D. Jung, K. Ng, E. Frisk, and M. Krysander, “Combining model-based diagnosis and data-driven anomaly classifiers for fault isolation,” Control Engineering Practice, vol. 80, pp. 146–156, 2018.

[2] S. You, M. Krage, and L. Jalic, “Overview of remote diagnosis and maintenance for automotive systems,” in SAE 2005 World Congress & Exhibition. SAE International, apr 2005.

[3] R. Isermann, “Model-based fault-detection and diagnosis—status and applications,” Annual Reviews in control, vol. 29, no. 1, pp. 71–85, 2005.

[4] T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media, 2009.

[5] D. Eriksson, E. Frisk, and M. Krysander, “A method for quantitative fault diagnosability analysis of stochastic linear descriptor models,” Automatica, vol. 49, no. 6, pp. 1591–1600, 2013.

[6] C. Sankavaram, A. Kodali, K. Pattipati, and S. Singh, “Incremental classifiers for data-driven fault diagnosis applied to automotive systems,” IEEE access, vol. 3, pp. 407–419, 2015.

[7] D. Jung, Y. Dong, E. Frisk, M. Krysander, and G. Biswas, “Sensor selection for fault diagnosis in uncertain systems,” International Journal of Control, vol. 93, no. 3, pp. 629–639, 2020.

[8] D. Jung, “Residual generation using physically-based grey-box recurrent neural networks for engine fault diagnosis,” arXiv preprint arXiv:2008.04644, 2020.

[9] L. Li and S. Ding, “Gap metric techniques and their application to fault detection performance analysis and fault isolation schemes,” Automatica, vol. 118, p. 109029, 2020.

[10] D. Jiang and W. Li, “Multi-objective optimal placement of sensors based on quantitative evaluation of fault diagnosability,” IEEE Access, vol. 7, pp. 117850–117854, 2019.

[11] H. Khorasgani and G. Biswas, “A methodology for monitoring smart buildings with incomplete models,” Applied Soft Computing, vol. 71, pp. 396–406, 2018.

[12] W. Zhang, G. Biswas, Q. Zhao, H. Zhao, and W. Feng, “Knowledge distilling based model compression and feature learning in fault diagnosis,” Applied Soft Computing, p. 105958, 2019.

[13] K. Tidritri, T. Tiplica, N. Chatti, and S. Verron, “A generic framework for decision fusion in fault detection and diagnosis,” Engineering Applications of Artificial Intelligence, vol. 71, pp. 73–86, 2018.

[14] I. Matei, M. Zhenirovskyy, J. de Kleer, and A. Feldman, “Classification-based diagnosis using synthetic data from uncertain models,” in PHM Society Conference, vol. 10, no. 1, 2018.

[15] W. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. Boult, “Toward open set recognition,” IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 7, pp. 1757–1772, 2013.

[16] W. Scheirer, L. Jain, and T. Boult, “Probability models for open set recognition,” IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 11, pp. 2317–2324, 2014.

[17] E. Rudd, L. Jain, W. Scheirer, and T. Boult, “The extreme value machine,” IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 3, pp. 762–768, 2018.

[18] M. Atoui, A. Cohen, S. Verron, and A. Kobi, “A single bayesian network classifier for monitoring with unknown classes,” Engineering Applications of Artificial Intelligence, vol. 85, pp. 681–690, 2019.

[19] Y. Yan, P. Luh, and K. Pattipati, “Fault diagnosis of components and sensors in hvac air handling systems with new types of faults,” IEEE Access, vol. 6, pp. 21682–21696, 2018.

[20] N. Sawalhi, R. Randall, and H. Endo, “The enhancement of fault detection and diagnosis in rolling element bearings using minimum entropy deconvolution combined with spectral kurtosis,” Mechanical Systems and Signal Processing, vol. 21, no. 6, pp. 2616–2633, 2007.

[21] X. Guo, L. Chen, and C. Shen, “Hierarchical adaptive deep convolutional neural network and its application to bearing fault diagnosis,” Measurement: Journal of the International Measurement Confederation, vol. 93, pp. 490–502, 2016.

[22] P. Paris and F. Erdogan, “A critical analysis of crack propagation laws,” Journal of Fluids Engineering, Transactions of the ASME, vol. 85, no. 4, pp. 528–533, dec 1963.

[23] D. Jung, H. Khorasgani, E. Frisk, M. Krysander, and G. Biswas, “Analysis of fault isolation assumptions when comparing model-based design approaches of diagnosis systems,” IFAC-PapersOnLine, vol. 48, no. 21, pp. 1289–1296, 2015.

[24] J. De Kleer and B. Williams, “Diagnosing multiple faults,” Artificial intelligence, vol. 32, no. 1, pp. 97–130, 1987.

[25] S. Kullback and R. Leibler, “On information and sufficiency,” The annals of mathematical statistics, vol. 22, no. 1, pp. 79–86, 1951.

[26] C. Bishop, Pattern recognition and machine learning. springer, 2006.

[27] M. Basseville, I. Nikiforov et al., Detection of abrupt changes: theory and application. prentice Hall Englewood Cliffs, 1993, vol. 104.

[28] J. Grezmak, P. Wang, C. Sun, and R. Gao, “Explainable convolutional neural network for gearbox fault diagnosis,” in Procedia CIRP, vol. 80, 2019, pp. 476–481.

[29] J. Hershey and P. Olsen, “Approximating the Kullback Leibler divergence between Gaussian mixture models,” in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 4, 2007.

[30] J. Durrieu, J. Thiran, and F. Kelly, “Lower and upper bounds for approximation of the Kullback-Leibler divergence between Gaussian mixture models,” in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2012, pp. 4833–4836.

[31] L. Eriksson, S. Frei, C. Onder, and L. Guzzella, “Control and optimization of turbocharged Spark ignited engines,” in IFAC Proceedings Volumes (IFAC-PapersOnline), vol. 15, no. 1, 2002, pp. 283–288.

[32] M. Tutuianu, A. Marotta, H. Steven, E. Ericsson, T. Hanii, N. Ichikawa, and H. Ishii, “Development of a World-wide Worldwide harmonized Light duty driving Test Cycle,” Technical Report, vol. 03, no. January, pp. 7–10, 2014.

[33] E. Frisk, M. Krysander, and D. Jung, “A toolbox for analysis and design of model based diagnosis systems for large scale models,” in IFAC World Congress, Toulouse, France, 2017.

[34] L. Eriksson, “Modeling and control of turbocharged SI and DI engines,” in Oil and Gas Science and Technology, vol. 62, no. 4, 2007, pp. 523–538.

[35] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, “Support vector method for novelty detection,” in Advances in neural information processing systems, 2000, pp. 582–588.

[36] “Matlab 2018b statistics and machine learning toolbox,” 2018, the MathWorks, Natick, MA, USA.