Assumption-based Design of Hybrid Diagnosis Systems: Analyzing Model-based and Data-driven Principles

Daniel Jung[1] and Mattias Krysander[2]

[1,2] *Department of Electrical Engineering, Linköping University, Linköping, SE-581 83, Sweden*
daniel.jung@liu.se
mattias.krysander@liu.se

ABSTRACT

Hybrid diagnosis systems combine model-based and data-driven methods to leverage their respective strengths and mitigate individual weaknesses in fault diagnosis. This paper proposes a unified framework for analyzing and designing hybrid diagnosis systems, focusing on the principles underlying the computation of diagnoses from observations. The framework emphasizes the importance of assumptions about fault modes and their manifestations in the system. The proposed architecture supports both fault decoupling and classification techniques, allowing for the flexible integration of model-based residuals and data-driven classifiers. Comparative analysis highlights how classical model-based and pure data-driven systems are special cases within the proposed hybrid framework. The proposed framework emphasizes that the key factor in categorizing fault diagnosis methods is not whether they are model-based or data-driven, but rather their ability to decouple faults which is crucial for rejecting diagnoses when fault training data is limited. Future research directions are suggested to further enhance hybrid fault diagnosis systems.

1. INTRODUCTION

A diagnosis system can be described as a function that uses observations from the monitored system to compute diagnoses. A diagnosis is a statement about the system's health that is consistent with the observations. The output from a diagnosis system is updated over time, as new observations are collected, and used as input to other functionalities, e.g., fault-tolerant control and fault mitigation (Amin & Hasan, 2019), computer-assisted troubleshooting (Pernestål, Nyberg, & Warnquist, 2012), and prognostics (Zio, 2022). Thus, the diagnosis system must draw reliable conclusions about the system's health, at every time instance, even when there are classification ambiguities, to be able to take appropriate countermeasures. Drawing the wrong conclusions about detected abnormal behavior can be both hazardous and expensive. This can be summarized in the following design principles when developing a diagnosis system:

1. To avoid drawing the wrong conclusion about the system's health, the diagnosis system must not falsely reject the true diagnosis candidate.
2. The diagnosis system should be as precise as possible rejecting diagnosis candidates that are not consistent with system operation.
3. Faults should be detected and isolated at an early stage for the system to act accordingly.

Designing a diagnosis system that fulfills these objectives is a nontrivial problem and requires a good understanding of the behavior of the system to be monitored and the characteristics of the faults to be diagnosed.

1.1. Fault Diagnosis Methods

Because of its industrial and scientific relevance, the fault diagnosis problem has been approached in many communities. Two common approaches are referred to as model-based diagnosis and data-driven diagnosis (Jung, Ng, Frisk, & Krysander, 2018). In model-based diagnosis, a mathematical model of the system derived from physical insights is used to detect inconsistencies between model predictions and observations, mainly by designing residual generators. Fault isolation is then performed by matching residual patterns with different fault hypotheses derived from model analysis (Travé-Massuyès, 2014). Data-driven fault diagnosis refers to methods that use historical data from different fault scenarios to learn the relation between observations and class labels (diagnoses). A common data-driven approach is to formulate a classification problem where fault diagnosis is a matter of assigning which class label best explains the observations based on previous (training) data (Qin, 2012).

Still, the main principle of both model-based and data-driven methods is to compare measurements with expected, or predicted, system behavior to detect inconsistencies. Since both approaches utilize some type of model of the system behavior to describe the relation between different signals, the main difference is based on what type of information the model is derived. However, there is no clear distinction between model-based diagnosis and data-driven diagnosis when comparing how observations are mapped to different diagnoses.

1.2. Modeling Assumptions and Fault Diagnosis

There are several complicating factors of fault diagnosis, such as measurement noise and model inaccuracies but also limited information about potential fault scenarios and fault manifestations (Sankavaram, Kodali, Pattipati, & Singh, 2015). Developing sufficiently accurate models for fault diagnosis can be a time-consuming process that requires expert system knowledge. Another factor is that faults are rare events which results in imbalanced training data and lack of representative data of relevant fault scenarios. These issues need to be addressed in the diagnosis system design.

Ideally, a diagnosis system should compute all diagnoses that can explain the observations. However, defining the set of observations that can be explained by each fault mode is in many applications not feasible. Instead, different diagnosis system design choices are made to approximate this relationship. Examples are the 'single-fault' assumption that maximally one fault is present in the system at the same time, and the 'exoneration' assumption that the risk of missed detections is negligible (Travé-Massuyès, 2014).

Many design choices are based on assumptions about fault modes, i.e., how different faults manifest in the system (Jung, Khorasgani, Frisk, Krysander, & Biswas, 2015). Note that these design choices do not have to be intentionally made by the developer but have a significant impact on the diagnosis output. For example, if formulating fault diagnosis as a classification problem, the diagnosis output will depend on whether a closed-set classifier or an open-set classifier is used, i.e., are all fault modes assumed to be known by the diagnosis system or should it identify scenarios with potentially unknown faults (Scheirer, Jain, & Boult, 2014). Identifying and applying valid assumptions for a given application can reduce diagnosis complexity and at the same time improve the fault diagnosis accuracy (Travé-Massuyès, 2014). On the other hand, a diagnosis system that is designed based on the wrong assumptions could result in falsely rejecting the true diagnosis.

The performance demand of the diagnosis system requires careful identification of a suitable diagnosis solution for a given application. This is a nontrivial task because it depends on many factors, such as behavioral characteristics of different faults, performance requirements, access to training data and models, etc. Also, when validating a proposed diagnosis system design, performance is evaluated based on a limited set of fault realizations, which can be misleading if test data is not representative of all possible fault realizations (Frisk, Jarmolowitz, Jung, & Krysander, 2022).

Understanding how fault diagnosis methods are based on different assumptions about fault modes and the relationship between observations and diagnoses, is necessary to compare different diagnosis system solutions. It is also possible to avoid fault diagnosis solutions that are based on non-valid assumptions for a given application. This is also important when combining methods from different approaches in a hybrid diagnosis system design. If different models draw different conclusions about the system's health, a hybrid diagnosis system should make use of all information and avoid conflicts, i.e. reject diagnoses when possible or rank diagnoses based on which is most likely when there are diagnosis ambiguities, see e.g. (Jung et al., 2018).

1.3. Problem Formulation

The scope of this work is the design of hybrid diagnosis systems. Different hybrid diagnosis systems, tailored for specific applications, have been proposed in the literature. However, designing a hybrid diagnosis system that takes advantage of different diagnosis methods, requires a unified view of how to compute diagnoses.

The first objective is to develop a framework to analyze and compare diagnosis system designs. The purpose is not on how to evaluate the performance of a given design (e.g. false alarm rate or classification accuracy) but on understanding the fundamental principles of how different diagnosis system solutions compute diagnoses, i.e. how it reasons about fault hypotheses (Jung et al., 2015). The focus is on how the set of diagnoses computed by the diagnosis system is dependent on the assumptions made about how different faults manifest in the system.

The second objective is, based on the developed framework, to present a hybrid diagnosis system architecture that can be used to combine different diagnosis methods in accordance to the stated design principles. The purpose of the architecture is a foundation to combine methods based on their properties which should be a generalization of classical model-based and data-driven architectures. Based on the results from this study relevant directions for future research in hybrid fault diagnosis are also proposed.

2. RELATED RESEARCH

Several survey papers have been published that describe fault diagnosis state-of-the-art with a focus on a specific approach, e.g., model-based diagnosis and signal-based fault diagnosis (Gao, Cecati, & Ding, 2015), data-driven fault diagnosis (Abid, Khan, & Iqbal, 2021; Z. Xu & Saleh, 2021), physics-informed machine learning for fault diagnosis (Y. Xu, Kohtz, Boakye, Gardoni, & Wang, 2023), condition monitoring

(Stetco et al., 2019), or a given application, e.g., batteries (Xiong, Sun, Yu, & Sun, 2020), traction systems in high-speed trains (Chen, Jiang, Ding, & Huang, 2020), HVAC systems (Mirnaghi & Haghighat, 2020), and wind-turbines (Liu & Zhang, 2020; Stetco et al., 2019). The potential of evolutionary algorithms and neural networks for fault diagnosis with respect to model-based methods are discussed in (Witczak, 2006). The survey paper (Theissler, Pérez-Velázquez, Kettelgerdes, & Elger, 2021) focused on machine learning for predictive maintenance and identified challenges, e.g. limited access to labeled data, and lack of public datasets. The authors state that ML will not replace model-based methods but help in hybrid designs. A survey on prognostics and health management (PHM) is presented in (Zio, 2022) that discuss the future needs in fault diagnosis research, e.g. deployment of diagnosis systems in industrial applications. The importance of training data for data-driven fault diagnosis and how this affects the selection of method is discussed in (Dai & Gao, 2013). These mentioned surveys give a good overview of different fault diagnosis methods. However, there is limited focus on the underlying differences between different fault diagnosis methods related to their design objectives.

To address the limitations of classical model-based and data-driven fault diagnosis methods, several papers have proposed hybrid diagnosis system solutions. In early work, see e.g. (Becraft, Lee, & Newell, 1991; Senjen, De Beler, Leckie, & Rowles, 1993), hybrid diagnosis systems have been proposed combining neural networks (NN) as input to expert systems. Another proposed solution is to use physically-based methods to generate features that are fed to a data-driven classifier, see e.g. (Luo, Namburu, Pattipati, Qiao, & Chigusa, 2009; Yan, Ji, & Shen, 2017; Destro, Facco, Munoz, Bezzo, & Barolo, 2020). The proposed methods are evaluated using different case studies, but there is little motivation for why the selected diagnosis system design is chosen based on given performance requirements.

The authors in (Mylaraswamy & Venkatasubramanian, 1997) propose a hybrid fault diagnosis framework combining e.g. digraphs, observers, trend analysis, expert rules, and NN, based on the conclusion that no diagnosis approach is good for everything. To deal with conflicts in diagnosis outputs, a voting scheme is proposed based on the confidence of each method. Different decision-making strategies for collaboration between heterogeneous fault diagnosis methods are evaluated in (Ghosh, Ng, & Srinivasan, 2011). The results show that utility-based methods (e.g. majority-voting) work if all diagnosis methods are equally good but evidence-based methods (e.g. Bayesian and Dempster-Shafer) work better if there is a bigger diversity in performances of different diagnosis methods. Issues with limited training data from faults are discussed but not considered explicitly in the study. In (Yan et al., 2017), an extended Kalman filter (EKF) is used to generate more stationary features, compared to raw sensor data, that are fed into a recursive one-class SVM (1SVM) to detect faults in a chiller (HVAC). EKF is also used in (Destro et al., 2020) to generate features (estimated states) combined with actuator and sensor signals as input to a PCA classifier to diagnose a process system. The authors state that the proposed method has difficulties with out-of-distribution data.

Several papers propose model-based residuals to generate features as input to data-driven classifiers, see e.g. (Jung & Sundström, 2017; Purbowaskito, Lan, & Fuh, 2024; Lundgren & Jung, 2022). In (Yu, Shields, & Daley, 1996), the output from a model-based parity space is fed to a set of shallow NN used as a one vs rest classifier for each fault. In (Svärd, Nyberg, Frisk, & Krysander, 2013), model-based residuals are evaluated using a Kullback-Leibler divergence-based change detection method. Several authors use signed digraphs to identify subsets of signals to train a set of PLS-based residuals to monitor different sub-models, see e.g. (Lee, Han, & Yoon, 2004; Lee, Tosukhowong, Lee, & Han, 2006; Ahn et al., 2008). In (Garcia-Alvarez, Bregon, Pulido, & Alonso-Gonzalez, 2023), residuals are fed to a PCA model. Fault isolation is done by identifying which residuals made the PCA detect an anomaly and then performing Consistency-Based Diagnosis (De Kleer & Williams, 1987) based on the residuals. Experiments showed that the hybrid approach outperforms each approach individually but also when compared with a black-box NN approach.

The authors in (Tidriri, Tiplica, Chatti, & Verron, 2018) propose a Bayesian Network-based information fusion of model-based structured residuals and discriminant analysis. The authors in (Atoui, Cohen, Verron, & Kobi, 2019) propose a Bayesian Network (BN) for fault diagnosis that combines discrete and continuous variables. It handles unknown faults by detecting when an observation deviates too much from data. The use of structural residuals and BN for fault isolation using a fault signature matrix is proposed in (Atoui & Cohen, 2021) to address the problem with limited training data. A BN is also proposed in (Wang et al., 2023) to fuse residuals, knowledge-based, and data-driven features to perform fault diagnosis. The authors in (Khorasgani, Farahat, Ristovski, Gupta, & Biswas, 2018) propose a unifying framework for model-based and data-driven methods. The framework discusses the use of both real data and simulation data for feature extraction using model-based methods, domain knowledge, and data-driven techniques. The features are then fed to supervised and unsupervised methods.

Surveys of hybrid fault diagnosis methods can be found in e.g. (Wilhelm, Reimann, Gauchel, & Mitschang, 2021; Tidriri, Chatti, Verron, & Tiplica, 2016; Goupil, Chanthery, Travé-Massuyès, & Delautier, 2022). The authors in (Wilhelm et al., 2021) discuss different ways of hybrid designs focusing on series and parallel architectures. Combining different methods in series is claimed to utilize the pros and cons of each approach. Parallel solutions are claimed to be more robust since there is less risk that an error early in a serial solution would propagate and reduce the performance of others. There is no proposal of a given framework, but the authors present different approaches. Several papers, e.g. (Tidriri et al., 2016) and (Frisk et al., 2022), compare model-based diagnosis and data-driven diagnosis and discuss the advantages of hybrid diagnosis system designs combining both models and data.

A comparison of six different diagnosis system designs submitted to a Wind turbine FDI competition is presented in (Odgaard & Stoustrup, 2012). The authors in (Feiyi & Jinsong, 2015) reviewed multiple proposed diagnosis system designs developed for the NASA Advanced Diagnostics and Prognostics Test-bed (ADAPT). In (Jung et al., 2015), it was shown that the design objectives of different model-based diagnosis system designs have implications on the assumptions made about observations that can be explained by the different fault modes. A review of public process monitoring benchmarks and proposed solutions are presented in (Melo, Câmara, Clavijo, & Pinto, 2022).

With respect to previous works, the contribution here is to understand the connection between diagnosis systems and modeling assumptions about fault behavior. For a diagnosis system developer, it is important to understand the impact of different design choices when implementing a diagnosis system for a given application.

3. FAULT DIAGNOSIS DEFINITIONS

Before analyzing the relation between assumptions about faults and diagnoses, a set of general definitions for fault diagnosis is presented describing the relationship between observations and faults. Then, a set of common assumptions about faults are defined using the general definitions.

3.1. Fault Modes and Observation Sets

A diagnosis system is used to monitor a system to detect and diagnose abnormal behavior. In principle, a diagnosis system can be interpreted as a function $D(z)$ which outputs a set of diagnosis candidates (fault hypotheses) $D \subseteq \mathcal{F}$ given a set of observations z. The observations consist of sensor and actuator signals but could also be input from a technician or operator. If many diagnosis candidates can explain a given observation, each diagnosis $d \in D$ can also have a value that represents a ranking of how likely it is with respect to the other diagnoses which will be discussed later.

Let $\{f_1, f_2, ...\}$ denote a set of faults that could be present in the system. Since several faults could be present at the same time, the term *fault mode* F_i is used to describe the system state and is defined as a set of faults that is present in the system. A fault mode could be a single fault $F_i = \{f_1\}$ but also multiple faults $F_i = \{f_1, f_2, \ldots\}$. The fault-free mode when no faults are present is denoted NF (No Fault). Let $\mathcal{F} = \{\text{NF}, F_1, F_2, \ldots\}$ denote the set of all fault modes that the system can be in.

Table 1. The list of possible observations and corresponding diagnoses in the switch-lamp example.

Observations		
Switch (S)	**Lamp (L)**	**Diagnoses**
Open	Off	NF, f_1, f_3, $\{f_1, f_3\}$, $\{f_2, f_3\}$
Open	On	f_2
Closed	Off	f_1, f_2, f_3, $\{f_1$,$f_3\}$, $\{f_2$,$f_3\}$
Closed	On	NF, f_2

Let $\mathcal{O}_{F_i}(z)$ denote the set of observations z from the system that can be explained by fault mode F_i. The set of observation z of the fault-free mode is denoted $\mathcal{O}_{\text{NF}}(z)$. Note that the $\mathcal{O}_{F_i}(z)$ describes the true set of observations that could be measured from any realization of F_i and is a system property. Based on the observation sets, it is possible to define fault detectability and isolability.

Definition 1 (Fault isolability) *A fault mode F_i is isolable from another fault mode F_j if $\mathcal{O}_{F_i} \setminus \mathcal{O}_{F_j} \neq \emptyset$, i.e. there exist an observation z that can be explained by F_i but not F_j. If F_i is isolable from the fault-free mode NF, i.e. $\mathcal{O}_{F_i} \setminus \mathcal{O}_{NF} \neq \emptyset$, then F_i is said to be* detectable.

Fault isolability defines the conditions when F_i is isolable from F_j. However, if F_i is isolable from F_j it does not guarantee that F_i is isolable for all $z \in \mathcal{O}_{F_i}$ and from all fault modes $F_j \in \mathcal{F}$. Note that fault isolability is not a symmetric property, i.e., if F_i is isolable from F_j, it does not mean that F_j is isolable from F_i.

To illustrate the relationship between different observations and diagnoses, a small example is considered.

Example 1 *Consider a system consisting of a lamp that is controlled by a switch. The possible observations are that the lamp can be on or off and the switch can be open or closed. The potential fault modes are that the switch can be ok, stuck open (f_1), or stuck closed (f_2) and that the lamp can be ok or broken (f_3). The mode when both the switch and lamp are ok is referred to as NF (No Fault). The set of observations and corresponding diagnoses is listed in Table 1. Brackets are used to define fault modes where multiple faults are present. Table 2 shows the observation sets for fault modes f_1 and f_2. If the switch (S) is closed and the lamp (L) is on both f_1 and f_2 are diagnosis candidates. If the switch is open and the lamb is off f_1 is a diagnosis but f_2 is not and f_2 is isolable from f_1 if the lamp is on.* □

Even for this small example, several diagnoses can be explained by each observation. In general, e.g. if there are continuous signals, it is not feasible to list all observations.

Table 2. The list of possible observations for the diagnoses $\{f_1\}$ and $\{f_2\}$ in the switch-lamp example.

Diagnoses	Observation set
f_1	{S Open, L off}, {S Closed, L off}
f_2	{S Open, L on}, {S Closed, L off}, {S Closed, L on}

3.2. Assumptions when Modeling Fault Modes

The diagnosis system models the observation set, i.e., which observations z that result in F_i being a diagnosis candidate. Consider a diagnosis system $D(z)$ and a mode $F_i \in \mathcal{F}$. The set of observations where F_i is a diagnosis candidate is denoted by $\hat{\mathcal{O}}_{F_i} = \{z | F_i \in D(z)\}$. If the models of the observation sets are inaccurate, there is a risk that the true diagnosis is either falsely rejected, i.e. if there are observations z such that $z \notin \hat{\mathcal{O}}_{F_i}$ and $z \in \mathcal{O}_{F_i}$, or that the number of diagnoses is large because false diagnoses are not properly rejected.

Design principle 1, i.e., no mode should be falsely rejected, implies that

$$\mathcal{O}_{F_i} \subseteq \hat{\mathcal{O}}_{F_i} \tag{1}$$

for all modes $F_i \in \mathcal{F}$. A diagnosis system with this property is said to be *complete* since all diagnoses are diagnosis candidates.

Design principle 2, i.e., all diagnosis candidates should be a diagnosis implies that

$$\hat{\mathcal{O}}_{F_i} \subseteq \mathcal{O}_{F_i} \tag{2}$$

for all modes $F_i \in \mathcal{F}$. A consequence of (1) and (2) is that these sets should be equal for all modes, i.e., the objective when designing a diagnosis system is to model $\mathcal{O}_{F_i}$ as accurately as possible for all fault modes F_i. In noisy environments, the observation sets can be replaced with probability distributions transferring property (1) to achieve a low false alarm probability and (1) to achieve a low missed detection probability. Typically, there is a trade-off between a low false alarm probability and a low missed detection rate and in consistency-based diagnosis low false alarm probability is prioritized, i.e. similar to (1), not to falsely reject diagnoses.

3.2.1. Approximating the Set of Fault Modes

One assumption is that all fault modes $\mathcal{F}$ that the system can be in are modeled by the diagnosis system, called the closed-world assumption. If $\hat{\mathcal{F}}$ is the set of considered modes in a diagnosis system the assumption can be formalized as:

Assumption 1 (Closed-world) *The modeled set of fault modes is equal to the set of all fault modes, i.e., $\hat{\mathcal{F}} = \mathcal{F}$.*

Under the closed-world assumption, each observation is mapped to at least one known fault mode, i.e., there are no other possible fault modes than the modes in $\hat{\mathcal{F}}$. This means that the complete diagnosis system always outputs at least one fault mode, the true one, from $\hat{\mathcal{F}}$ as a diagnosis candidate.

Different methods are used to systematically identify potential faults that could occur in a system, e.g. FMEA (Spreafico, Russo, & Rizzi, 2017) and FTA (Ruijters & Stoelinga, 2015). Still, it can be difficult to identify all potential faults or it is not possible to properly model all fault modes resulting in observations z such that $z \notin \hat{\mathcal{O}}_{F_i}$ for all F_i, i.e., $D(z) = \emptyset$.

Multi-class classifiers, e.g. Random Forest (Breiman, 2001), and diagnosis algorithms, see e.g. (De Kleer & Williams, 1987), that return diagnoses based on $\hat{\mathcal{F}}$ and do not consider the unknown fault scenario are based on the closed-world assumption. Methods that do not rely on the closed-world assumption are, e.g., open-set classifiers can return that data can come from an unknown class, e.g. (Scheirer et al., 2014) or fault isolation logics that can return an unknown fault as a diagnosis, see e.g. (Jung et al., 2018).

Since faults are rare events, the set of diagnoses can be reduced by assuming that at most one fault is present in the system, i.e. no multiple faults. This means that the computed diagnosis candidates only consist of single faults which is referred to as the single-fault assumption:

Assumption 2 (Single-fault) *The modeled set of fault modes $\hat{\mathcal{F}}$ only includes modes representing single faults.*

The single-fault assumption is common in both model-based and data-driven diagnosis. In model-based diagnosis, the single-fault assumption is often used to reduce the number of diagnosis candidates while in data-driven diagnosis, e.g. supervised learning, it is often implicit when training classifiers using data from single-faults only (Atoui & Cohen, 2021).

3.2.2. Approximating the observation sets

Other approximations are directly related to how the diagnosis system approximates the observation sets of different fault modes. Some types of faults are expected to manifest in a specific number of ways.

Assumption 3 (Limited ways of fault manifestation) *Each fault has a limited number of ways it could manifest or evolve in the system.*

An example is faults that occur abruptly in the system. This assumption is used in diagnosis systems utilizing signal transient information to identify the fault. In (Mosterman & Biswas, 1999), transient information patterns are identified and compared to different fault manifestations to isolate the fault. This assumption is also found in data-driven methods when it is assumed that training data is representative of fault modes to be classified, e.g. fault magnitudes and operating conditions.

There are some assumptions here that are related to fault manifestation that restrict the possible observations that can be explained by each fault mode.

Assumption 4 (Strong discriminability) *Two modes F_i and F_j are strongly discriminable from each other if*

$$\mathcal{O}_{F_i} \cap \mathcal{O}_{F_j} = \emptyset \tag{3}$$

If the observation sets $\mathcal{O}_{F_i}$ and $\mathcal{O}_{F_j}$ are strongly discriminable, then each observation z can only be explained by maximally one of the two fault classes. The term strong discriminability is used in (Travé-Massuyès, 2014) but is also applied in, e.g., multi-class classifiers returning the most likely class for a given observation. If the diagnosis system is designed such that strong discriminability holds for all pairs of fault modes in $\hat{\mathcal{F}}$, then each observation z is mapped to maximally one diagnosis candidate. This is the case in many data-driven classifiers which outputs the most likely class for a given observation. A special case of strong discriminability that is sometimes used in model-based diagnosis is the assumption that a fault in the system will trigger all fault detectors that are sensitive to that fault, see e.g. (Travé-Massuyès, 2014).

Assumption 5 (Exoneration) *All fault modes are strongly discriminable from the fault-free mode, i.e.*

$$\mathcal{O}_{F_i} \cap \mathcal{O}_{NF} = \emptyset \quad \forall F_i \tag{4}$$

This means that a fault is always detectable, i.e. the missed detection rate is negligible. The exoneration assumption is valid in systems where the impact of faults is significant compared to model uncertainties and noise. However, exoneration can result in falsely rejecting the true diagnosis if the fault is small or if the fault detectors have varying detection performance, see e.g. (Sanchez, Escobet, Puig, & Odgaard, 2015).

4. FEATURE GENERATION

In general, it is too difficult to design a diagnosis system that directly maps the observations to diagnoses, i.e. modeling the sets $\mathcal{O}_{F_i}$ directly using z. A common approach is to use some type of signal processing or filtering to generate features to simplify the diagnosis process. Since many diagnosis systems contain a feature generation step it is necessary to discuss how assumptions about the modeled feature outputs relate to assumptions about the observation sets. Some main reasons for feature generation are:

- System dynamics complicates fault detection using only raw observations z.
- The complex relation between faults and observations z complicates fault isolation.
- Some faults are affecting the dynamic behavior of the system or introducing noise.
- The number of signals is large making it dimensionally difficult to model the observation sets. Strong correlations between signals could be used to reduce the feature space compared to the original observation space without losing information.

Therefore, a feature generation step is performed to map the observations z to a feature space r where it is easier to compute diagnoses. Feature generation can be performed in several steps, e.g. first generating residuals and then processing the residual output to extract features (Frisk et al., 2022).

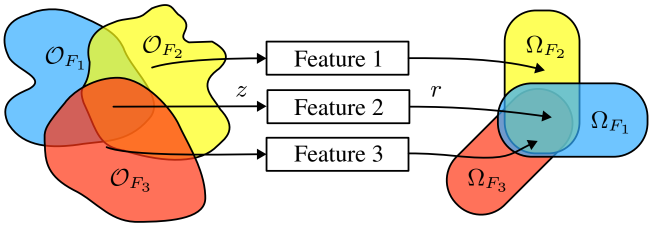

Figure 1. Mapping observations to feature space.

Let $\Omega_{F_i}(r)$ denote the set of feature output combinations $r = r(z)$ that can be explained by fault mode F_i

$$\Omega_{F_i}(r) = \{r(z) | z \in \mathcal{O}_{F_i}\} \tag{5}$$

The feature set maps observations to feature space $r : \mathcal{O}_{F_i} \to \Omega_{F_i}$ and $\hat{\Omega}_{F_i}$ denotes the modeled set. Note that the set of features might not map all information that is available in the observation space and thus lose fault diagnosis properties.

Diagnosability properties, such as detectability and isolability, can also be defined in feature space Ω. The same assumptions can be made for the feature space, such as strong discriminability and exoneration. However, assumptions made about Ω_{F_i} have implications on $\mathcal{O}_{F_i}$. For example, if Ω_{F_i} and Ω_{F_j} are strongly discriminable, it implies that $\mathcal{O}_{F_i}$ and $\mathcal{O}_{F_j}$ are also strongly discriminable (Jung et al., 2015):

$$\Omega_{F_i} \cap \Omega_{F_j} \neq \emptyset \to \mathcal{O}_{F_i} \cap \mathcal{O}_{F_j} \neq \emptyset \tag{6}$$

This means, e.g., that designing a diagnosis system solution that only returns the most likely fault mode as a diagnosis assumes that there is no overlap in possible observations from different fault modes. Note that a design goal of feature generation is to transfer the detection and isolation performance in the observation space to the feature space. However, defining Ω_{F_i} given r for each fault mode F_i is still necessary, e.g. by estimating it from data from that fault class.

There are other design objectives when generating features that can be used to model Ω_{F_i} without the need for data from F_i. One feature property that is commonly used in model-based diagnosis is fault decoupling, i.e., designing features that are insensitive to certain faults.

Definition 2 (Fault decoupling) *For a given subset of fea-*

tures $r_s \subseteq r$, a mode F_j is said to be decoupled *if*

$$\Omega_{F_j}(r_s) \subseteq \Omega_{NF}(r_s) \tag{7}$$

i.e., a decoupled fault mode is not detectable with r_s.

Fault decoupling can be achieved, e.g., by designing residuals that monitor different subsystems, see e.g. (Venkatasubramanian, Rengaswamy, Yin, & Kavuri, 2003). Another approach to decouple sensor faults is to generate a feature without using that signal, or modeling the fault to be decoupled and estimate it using an observer, to make the residual insensitive to the fault (Commault, Dion, Sename, & Motyeian, 2002).

4.1. Residual Generators

One reason for discussing residuals specifically as a feature is that they are central in model-based diagnosis to achieve fault decoupling for fault isolation, referred to as structured residuals. A residual generator is designed as a function of observations that is (asymptotically) zero in the nominal case. A simple example of a residual generator is $r = y - \hat{y}(u)$ where the model prediction $\hat{y}$ is computed based on some actuator signal u. An advantage of residuals is that they are approximately zero in the nominal case since the system dynamics are filtered out which simplifies anomaly detection.

The ability to design residual generators requires redundancy. When a mathematical model is available, analytical redundancy is evaluated as a model property and refers to the ability to derive a mathematical expression describing the relationship between different observations (and their derivatives). From a data-driven perspective, redundancy relates to the intrinsic dimension of data (Camastra & Staiano, 2016), i.e., that observations from the fault-free system are located on a low-dimensional manifold in observation space, see e.g. (Mohammadi, Krysander, & Jung, 2022).

Designing features, such as residuals, where faults are decoupled is central for consistency-based reasoning and fault isolation when training data from faults is not available. The key is that fault hypotheses containing decoupled faults can be rejected if a residual is detecting a fault. However, this requires accurate models to avoid false alarms. Because of model inaccuracies a residual can be nonzero if there is a fault but also if the model inaccuracies are too significant. One example is data-driven models which do not generalize well for out-of-distribution data. This relates to the concept of aleatoric and epistemic uncertainty (Abdar et al., 2021). Aleatoric uncertainty refers to process and measurement noise that cannot be captured by the model, while epistemic uncertainty refers to the generalizability of the model. The validity of the feature model for a given observation z is an important aspect when reasoning about the cause of detected anomalies, see (Jung, Krysander, & Mohammadi, 2023).

5. CONSTRUCTION OF DIAGNOSIS SYSTEMS

The strong connection between assumptions made in different fault diagnosis methods and diagnoses, adds complexity when designing hybrid diagnosis systems. A general problem when combining methods is how to handle potential ambiguities or contradictions between the methods when computing diagnoses, especially if the two methods are based on different assumptions about faults. Two general hybridization principles are mentioned in, e.g., (Wilhelm et al., 2021): serial and parallel design but also combinations of both.

5.1. Series and Parallel Hybrid Diagnosis

Hybrid diagnosis systems where the output from one method is used as input in another are referred to as *series hybrid*. One example is when a set of model-based residuals is computed and then fed as input to a data-driven classifier to generate diagnoses (Atoui & Cohen, 2021), or when the residual output is processed to extract a new set of features (Frisk et al., 2022). Series hybrid diagnosis systems follow the principle discussed in the previous section where one method is used for feature generation to map the observation set of a feature set that is easier to process, e.g. by a data-driven classifier. Model-based diagnosis systems often consist of a series of methods where the output from one method is fed to the next method, e.g. a set of residual generators where the outputs are fed to a change detection or anomaly detection step, followed by a fault isolation logic, see Figure 2.

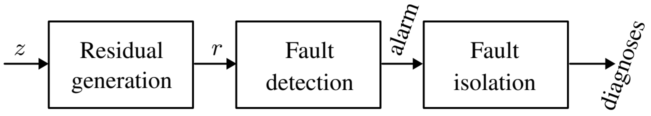

Figure 2. Fault detection and isolation.

Diagnosis systems where different methods are evaluated in parallel are referred to as *parallel hybrid*. A decision fusion method is needed to combine the results from each method to compute the diagnoses. The decision fusion method can compute the global diagnosis based on the assumptions made on the observation sets, or the feature sets. Conflicting statements from different methods, e.g., if $r_1(z) \in \Omega_{F_i}(r_1)$ but $r_2(z) \notin \Omega_{F_i}(r_2)$ then F_i is a diagnosis given r_1 but not given r_2. An example of a fusion method is the use of Bayesian Networks to compute the most probable set of diagnoses given the individual outputs, see e.g. (Tidriri et al., 2018; Atoui et al., 2019; Wang et al., 2023). Then, the mapping from the observations to the diagnoses is determined by how the fusion strategy is calibrated. Another option is to model the sets $\Omega_{F_i}(r)$ in such a way that conflicts are omitted.

5.2. Computation of Diagnoses

Computing diagnosis candidates taking into consideration model uncertainties and non-representative training data is nontrivial if a constraint is to avoid falsely rejecting the true diagnosis, i.e. the diagnosis system should be complete. Two principles considered here are rejection and ranking of diagnosis candidates. Rejection is the ability to narrow down the set of candidates by removing diagnoses that cannot explain the observations, and ranking is a way to order the diagnoses in a priority list based on which diagnosis is more likely, without rejecting any diagnoses. Ranking diagnosis candidates is not as strong as rejecting diagnoses but gives useful information when prioritizing which sets of actions or counter-measures to take. Some methods can be used to reject diagnosis candidates while other approaches cannot due to limited information.

Rejection of Diagnosis Candidates As previously discussed, computing reliable diagnoses requires knowledge of how the diagnosis system approximates $\hat{\mathcal{O}}_{F_i}$ for each fault mode. If $\hat{\mathcal{O}}_{F_i} \supseteq \mathcal{O}_{F_i}$, then it is possible to reliably reject F_i when $z \notin \hat{\mathcal{O}}_{F_i}$ since it is guaranteed that $z \notin \mathcal{O}_{F_i}$. However, achieving this is not trivial especially since this would require training data that is representative of all realizations of each fault to model the observation sets of all fault modes. However, if there are features, e.g. structured residual, where F_i is decoupled, then if $r \notin \Omega_{\text{NF}}$ also means that $r \notin \Omega_{F_i}$. In that case, it is sufficient to use nominal data to determine when to reject diagnoses. However, this would then require some other information to generate the feature where the fault is decoupled, e.g. a mathematical model of the system.

Ranking of Diagnosis Candidates When it is not possible to reliably reject diagnoses, e.g. when training data is not representative of the different fault classes, a second approach is to rank them using some quantitative measure. This means that all fault modes are plausible, but a ranking is used to prioritize how likely each diagnosis candidate is. Ranking can be done in different ways. One example is to rank each diagnosis independently of other diagnoses. This can be done when $\hat{\mathcal{O}}_{F_i} \subseteq \mathcal{O}_{F_i}$. This can be achieved, e.g., by using a one-class classifier to model the data support using data from that fault mode (Jung et al., 2018). Another advantage of using one-class classifiers to model each fault mode is that the training of the models will not suffer from imbalanced training data. Also, as new training data becomes available, it is sufficient to update the models of the fault modes that are represented in the new data, see e.g. (Jung et al., 2018).

A binary or multi-class classifier does not model each fault mode individually but relative to the other modes. Since such classifiers always return the most likely diagnosis, even if training data from two fault modes overlap or if test data is significantly different from training data, there is no guarantee that the modeled observation sets are over-estimations or under-estimations of the true observation set. Thus, the output from such a model only gives a relative ranking between the two diagnoses.

6. A PROPOSED HYBRID DIAGNOSIS SYSTEM ARCHITECTURE

Because of the strong connection between assumptions and design choices, a baseline diagnosis system design can assist in the development process and avoid misclassifications due to conflicts between methods used in the diagnosis system. Here, a generic diagnosis system architecture is proposed, inspired by (Jung et al., 2018), to systematically combine different types of fault diagnosis methods, where the risk of falsely rejecting the true diagnosis is low.

The proposed architecture, see Figure 3, contains both series and parallel hybrid components. In the design, there is a feature evaluation step. The generated features, e.g. residuals, are fed to a fault detection step. When a fault is detected, the isolation step is activated which consists of a diagnosis rejection part and a fault ranking part.

6.1. Feature Evaluation and Fault Detection

The fault detection step uses a set of anomaly classifiers, here referred to as fault detectors, where each anomaly classifier can be based on either single or multiple features. The purpose of the fault detection step is to detect abnormal system behavior, i.e. to reject the fault-free mode. If all features that are fed to a fault detector, are insensitive to a subset of faults, then that fault detector can be used to reject diagnoses. This means that all features that are used by a fault detector must be insensitive to a fault for the fault detector to be insensitive to the same fault (Jung et al., 2018). The set of features is designed to maximize fault detection performance. However, each fault detector is calibrated to fulfill requirements on fault alarm rate. Note that the detection step is separated from the isolation step to handle different performance requirements regarding detection and isolation performance. Since it is often crucial to avoid false alarms, it can be difficult to detect small faults using a sample-by-sample detection approach. Time-series analysis and statistical change detection methods such as the CUSUM test can be used to improve the detection of (small) faults over time, see e.g. (Gustafsson, 2007).

6.2. The Fault Isolation Process

If a fault is detected then the isolation step is activated. The fault isolation process consists of a fault isolation logic, that is used to reject diagnosis candidates and a data-driven ranking of the diagnosis candidates. The fault isolation logic requires fault detectors that are insensitive to faults and the fault ranking step requires training data from fault modes.

6.2.1. Diagnosis Rejection using Fault Isolation Logic

If there are fault detectors that are insensitive to some fault modes, and a fault has been detected with one of these fault detectors, this information is then fed to a consistency-based fault isolation logic (CBD), see (De Kleer & Williams, 1987). The fault isolation logic rejects diagnoses based on the fault sensitivity of the features that detect faults. Diagnoses are rejected that cannot be explained by the decoupled fault modes. An advantage is that CBD will not reject the true diagnosis if there are no false alarms, i.e. there is no exoneration or strong discriminability assumption (Jung et al., 2018). It can identify both single-fault and multiple-fault diagnoses without the need for training data from faults since it is sufficient to model the feature's nominal distribution. However, CBD is conservative since the number of diagnosis candidates can be large. Note that if there are no fault detectors where faults are decoupled, no diagnoses are rejected. If there are features without fault decoupling properties, these can still be fed to the CBD. However, they will not reject any diagnoses if they detect abnormal behavior (except the diagnosis that the system is fault-free). Thus, there is no need to treat the features differently in the diagnosis system design.

6.2.2. Diagnosis Ranking using One-Class Classifiers

Based on the computed diagnoses, each candidate is ranked using the feature outputs in the hybrid diagnosis architecture. The ranking of the diagnoses is done using data-driven one-class classifiers trained on available data from each fault mode to model its data support. The fault-free mode is not modeled here since this step is not activated before abnormal behavior is activated. Note that multi-class classifiers are not used since they are based on both the closed-world assumption and the strong discriminability assumption. This means that the classifiers cannot identify when there is an unknown fault scenario but also if there are multiple classifiers used to rank the diagnoses, different classifiers likely identify different faults leading to diagnosis conflicts.

Modeling each fault mode individually using anomaly classifiers, e.g. 1SVM, makes it possible to identify likely unknown fault scenarios and abnormal behavior that have not been observed before (Scheirer et al., 2014). It also allows for identifying multiple diagnoses that can explain the observations if training data from different fault modes are overlapping. Another advantage is that imbalanced data is not an issue during training and that it is possible to incrementally learn the data support of each fault mode individually as new data are available without having to retrain all classifiers. Note that, in general, if no training data from a fault mode is available, it is not possible to rank that diagnosis. If there are features where faults are decoupled it is possible to design classifiers to rank multiple-fault modes, using only training data from single faults (Jung et al., 2018).

The output of the diagnosis system is a set of diagnosis candidates and a ranking of each candidate. Multiple-fault diagnosis is handled by the CBD fault isolation logic. However, ranking multiple-fault diagnoses, in general, requires training data from multiple-fault scenarios to model the multiple-fault mode. In (Jung et al., 2018), it is shown that if there are features that are insensitive to faults, training data from single-faults is sufficient by ranking each fault individually in the multiple-fault diagnosis when the other faults are decoupled.

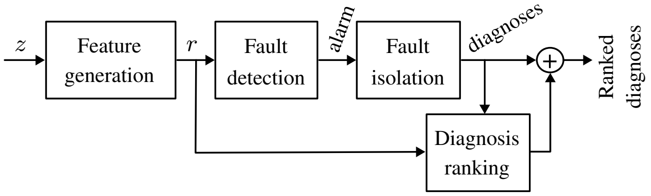

Figure 3. Hybrid diagnosis system design.

6.3. Comparison to Conventional Diagnosis Systems

The proposed architecture has been evaluated in (Jung et al., 2018) using a set of model-based residuals to evaluate data from an internal combustion engine. However, the selected architecture has not been motivated from the assumption-based perspective. Here, it is also motivated that a wider range of feature generation techniques can be used as long as their fault decoupling properties are taken into consideration.

A central part of the proposed architecture is the distinction between generated features where fault modes are decoupled and those where fault decoupling is not achieved because fault decoupling allows for rejecting diagnoses. Otherwise, the diagnosis principle is to rank the remaining diagnoses. Thus, the proposed architecture shows how to combine, e.g., model-based methods and data-driven methods when training data is limited. An interesting property of the proposed architecture is that classical model-based diagnosis systems or data-driven fault diagnosis designs become special cases depending on what type of information is available in the design process. If the ranking module is not available (e.g. if no training data from faults is available) and there are features that are insensitive to some faults, e.g. structured residuals, the diagnosis system design becomes a typical model-based diagnosis system, see e.g. (Gao et al., 2015), where diagnosis candidates are derived from isolation logics. On the other hand, if no faults are decoupled in any of the features, the fault isolation logic will not reject any diagnoses. Then, the diagnosis system becomes a typical data-driven fault classifier where the set of features is fed to one or more classifiers to identify the most likely fault class (Dai & Gao, 2013).

7. DISCUSSION AND CONCLUSIONS

There is a continuous scientific development of fault diagnosis methods. Still, design principles and guidelines are needed to support the implementation of diagnosis system solutions in real applications. Even though it is convenient to treat the fault diagnosis problem as a generic classification problem, different complicating factors require careful consideration when selecting an appropriate diagnosis system design. A general guideline to design a diagnosis system is proposed which can both reject and rank diagnoses based on the properties of the generated features. It is not claimed that the proposed architecture is always the optimal choice. However, having this diagnosis assumption-based perspective simplifies the design of hybrid diagnosis systems since it gives a general principle of how to utilize different diagnosis methods, such as classical model-based and data-driven methods, to compute diagnoses. The proposed framework shows that when categorizing fault diagnosis methods it is not model-based vs data-driven that is important. Instead, an important aspect is how observation sets for different fault models are modeled where fault decoupling is important to reject diagnoses when training data from faults is limited.

As seen in the literature, residuals are popular as features for fault diagnosis, especially the ability to filter out system dynamics and isolate faults. Residuals are also natural when utilizing physical insights together with machine learning in hybrid diagnosis systems. There is a need for methods to design data-driven residual generators that can decouple faults when a system model is not available. More investigations are needed around the connection between model properties, such as analytical redundancy, and data properties, such as the intrinsic dimension of data, to bridge the theory and methods developed in the model-based diagnosis community to data-driven fault diagnosis to deal with incomplete training data and unknown faults.

REFERENCES

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., … Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, *76*, 243–297. doi: 10.1016/j.inffus.2021.05.008

Abid, A., Khan, M., & Iqbal, J. (2021). A review on fault detection and diagnosis techniques: basics and beyond. *Artificial Intelligence Review*, *54*, 3639–3664. doi: 10.1007/s10462-020-09934-2

Ahn, S., Lee, C., Jung, Y., Han, C., Yoon, E., & Lee, G. (2008). Fault diagnosis of the multi-stage flash desalination process based on signed digraph and dynamic partial least square. *Desalination*, *228*(1-3), 68–83. doi: 10.1016/j.desal.2007.08.008

Amin, A., & Hasan, K. (2019). A review of fault tolerant control systems: advancements and applications. *Measurement*, *143*, 58–68. doi: 10.1016/j.measurement.2019.04.083

Atoui, M., & Cohen, A. (2021). Coupling data-driven and model-based methods to improve fault diagnosis. *Computers in Industry*, *128*, 103401. doi: 10.1016/j.compind.2021.103401

Atoui, M., Cohen, A., Verron, S., & Kobi, A. (2019). A single bayesian network classifier for monitoring with unknown classes. *Engineering Applications of Artificial Intelligence*, *85*, 681–690. doi: 10.1016/j.engappai.2019.07.016

Becraft, W. R., Lee, P. L., & Newell, R. B. (1991). Integration of neural networks and expert systems for process fault diagnosis. In *Proceedings of the 12th international joint conference on artificial intelligence-volume 2* (pp. 832–837).

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32. doi: 10.1023/A:1010933404324

Camastra, F., & Staiano, A. (2016). Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, *328*, 26–41. doi: 10.1016/j.ins.2015.08.029

Chen, H., Jiang, B., Ding, S., & Huang, B. (2020). Data-driven fault diagnosis for traction systems in high-speed trains: A survey, challenges, and perspectives. *IEEE Transactions on Intelligent Transportation Systems*, *23*(3), 1700–1716. doi: 10.1109/TITS.2020.3029946

Commault, C., Dion, J., Sename, O., & Motyeian, R. (2002). Observer-based fault detection and isolation for structured systems. *IEEE Transactions on Automatic Control*, *47*(12), 2074–2079.

Dai, X., & Gao, Z. (2013). From model, signal to knowledge: A data-driven perspective of fault detection and diagnosis. *IEEE Transactions on Industrial Informatics*, *9*(4), 2226–2238. doi: 10.1109/TII.2013.2243743

De Kleer, J., & Williams, B. (1987). Diagnosing multiple faults. *Artificial intelligence*, *32*(1), 97–130. doi: 10.1016/0004-3702(87)90063-4

Destro, F., Facco, P., Munoz, S., Bezzo, F., & Barolo, M. (2020). A hybrid framework for process monitoring: Enhancing data-driven methodologies with state and parameter estimation. *Journal of Process Control*, *92*, 333–351. doi: 10.1016/j.jprocont.2020.06.002

Feiyi, R., & Jinsong, Y. (2015). Fault diagnosis methods for advanced diagnostics and prognostics testbed (adapt): A review. In *2015 12th ieee international conference on electronic measurement & instruments (icemi)* (Vol. 1, pp. 175–180). doi: 10.1109/ICEMI.2015.7494248

Frisk, E., Jarmolowitz, F., Jung, D., & Krysander, M. (2022). Fault diagnosis using data, models, or both–an electrical motor use-case. *IFAC-PapersOnLine*, *55*(6), 533–

538. doi: 10.1016/j.ifacol.2022.07.183

Gao, Z., Cecati, C., & Ding, S. (2015). A survey of fault diagnosis and fault-tolerant techniques—part i: Fault diagnosis with model-based and signal-based approaches. *IEEE transactions on industrial electronics*, *62*(6), 3757–3767. doi: 10.1109/TIE.2015.2417501

Garcia-Alvarez, D., Bregon, A., Pulido, B., & Alonso-Gonzalez, C. (2023). Integrating pca and structural model decomposition to improve fault monitoring and diagnosis with varying operation points. *Engineering Applications of Artificial Intelligence*, *122*, 106145. doi: 10.1016/j.engappai.2023.106145

Ghosh, K., Ng, Y., & Srinivasan, R. (2011). Evaluation of decision fusion strategies for effective collaboration among heterogeneous fault diagnostic methods. *Computers & chemical engineering*, *35*(2), 342–355. doi: 10.1016/j.compchemeng.2010.05.004

Goupil, L., Chanthery, E., Travé-Massuyès, L., & Delautier, S. (2022). A survey on diagnosis methods combining dynamic systems structural analysis and machine learning. In *33rd international workshop on principle of diagnosis–dx 2022*.

Gustafsson, F. (2007). Statistical signal processing approaches to fault detection. *Annual Reviews in Control*, *31*(1), 41–54. doi: 10.1016/j.arcontrol.2007.02.004

Jung, D., Khorasgani, H., Frisk, E., Krysander, M., & Biswas, G. (2015). Analysis of fault isolation assumptions when comparing model-based design approaches of diagnosis systems. *IFAC-PapersOnLine*, *48*(21), 1289–1296. doi: 10.1016/j.ifacol.2015.09.703

Jung, D., Krysander, M., & Mohammadi, A. (2023). Fault diagnosis using data-driven residuals for anomaly classification with incomplete training data. *IFAC-PapersOnLine*, *56*(2), 2903–2908.

Jung, D., Ng, K., Frisk, E., & Krysander, M. (2018). Combining model-based diagnosis and data-driven anomaly classifiers for fault isolation. *Control Engineering Practice*, *80*, 146–156. doi: 10.1016/j.conengprac.2018.08.013

Jung, D., & Sundström, C. (2017). A combined data-driven and model-based residual selection algorithm for fault detection and isolation. *IEEE Transactions on Control Systems Technology*, *27*(2), 616–630. doi: 10.1109/TCST.2017.2773514

Khorasgani, H., Farahat, A., Ristovski, K., Gupta, C., & Biswas, G. (2018). A framework for unifying model-based and data-driven fault diagnosis. In *Annual conference of the phm society* (Vol. 10).

Lee, G., Han, C., & Yoon, E. (2004). Multiple-fault diagnosis of the tennessee eastman process based on system decomposition and dynamic pls. *Industrial & engineering chemistry research*, *43*(25), 8037–8048. doi: 10.1021/ie049624u

Lee, G., Tosukhowong, T., Lee, J., & Han, C. (2006). Fault diagnosis using the hybrid method of signed digraph and partial least squares with time delay: The pulp mill process. *Industrial & engineering chemistry research*, *45*(26), 9061–9074. doi: 10.1021/ie060793j

Liu, Z., & Zhang, L. (2020). A review of failure modes, condition monitoring and fault diagnosis methods for large-scale wind turbine bearings. *Measurement*, *149*, 107002. doi: 10.1016/j.measurement.2019.107002

Lundgren, A., & Jung, D. (2022). Data-driven fault diagnosis analysis and open-set classification of time-series data. *Control Engineering Practice*, *121*, 105006. doi: 10.1016/j.conengprac.2021.105006

Luo, J., Namburu, M., Pattipati, K., Qiao, L., & Chigusa, S. (2009). Integrated model-based and data-driven diagnosis of automotive antilock braking systems. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *40*(2), 321–336. doi: 10.1109/TSMCA.2009.2034481

Melo, A., Câmara, M., Clavijo, N., & Pinto, J. (2022). Open benchmarks for assessment of process monitoring and fault diagnosis techniques: a review and critical analysis. *Computers & Chemical Engineering*, 107964. doi: 10.1016/j.compchemeng.2022.107964

Mirnaghi, M., & Haghighat, F. (2020). Fault detection and diagnosis of large-scale hvac systems in buildings using data-driven methods: A comprehensive review. *Energy and Buildings*, *229*, 110492. doi: 10.1016/j.enbuild.2020.110492

Mohammadi, A., Krysander, M., & Jung, D. (2022). Analysis of grey-box neural network-based residuals for consistency-based fault diagnosis. *IFAC-PapersOnLine*, *55*(6), 1–6. doi: 10.1016/j.ifacol.2022.07.097

Mosterman, P., & Biswas, G. (1999). Diagnosis of continuous valued systems in transient operating regions. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *29*(6), 554–565. doi: 10.1109/3468.798059

Mylaraswamy, D., & Venkatasubramanian, V. (1997). A hybrid framework for large scale process fault diagnosis. *Computers & chemical engineering*, *21*, S935–S940. doi: 10.1016/S0098-1354(97)87622-3

Odgaard, P., & Stoustrup, J. (2012). Results of a wind turbine fdi competition. *IFAC Proceedings Volumes*, *45*(20), 102–107. doi: 10.3182/20120829-3-MX-2028.00015

Pernestål, A., Nyberg, M., & Warnquist, H. (2012). Modeling and inference for troubleshooting with interventions applied to a heavy truck auxiliary braking system. *Engineering applications of artificial intelligence*, *25*(4), 705–719. doi: 10.1016/j.engappai.2011.02.018

Purbowaskito, W., Lan, C., & Fuh, K. (2024). The potentiality of integrating model-based residuals and machine-learning classifiers: An induction motor fault diagnosis case. *IEEE Transactions on Industrial Informatics*, *20*(2), 2822-2832. doi: 10.1109/TII.2023.3299111

Qin, S. J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *Annual reviews in control*, *36*(2), 220–234. doi: 10.1016/j.arcontrol.2012.09.004

Ruijters, E., & Stoelinga, M. (2015). Fault tree analysis: A survey of the state-of-the-art in modeling, analysis and tools. *Computer science review*, *15*, 29–62. doi: 10.1016/j.cosrev.2015.03.001

Sanchez, H., Escobet, T., Puig, V., & Odgaard, P. (2015). Fault diagnosis of an advanced wind turbine benchmark using interval-based arrs and observers. *IEEE Transactions on Industrial Electronics*, *62*(6), 3783–3793. doi: 10.1109/TIE.2015.2399401

Sankavaram, C., Kodali, A., Pattipati, K., & Singh, S. (2015). Incremental classifiers for data-driven fault diagnosis applied to automotive systems. *IEEE access*, *3*, 407–419. doi: 10.1109/ACCESS.2015.2422833

Scheirer, W., Jain, L., & Boult, T. (2014). Probability models for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, *36*(11), 2317–2324. doi: 10.1109/TPAMI.2014.2321392

Senjen, R., De Beler, M., Leckie, C., & Rowles, C. (1993). Hybrid expert systems for monitoring and fault diagnosis. In *Proceedings of 9th ieee conference on artificial intelligence for applications* (pp. 235–241). doi: 10.1109/CAIA.1993.366605

Spreafico, C., Russo, D., & Rizzi, C. (2017). A state-of-the-art review of fmea/fmeca including patents. *Computer Science Review*, *25*, 19–28. doi: 10.1016/j.cosrev.2017.05.002

Stetco, A., Dinmohammadi, F., Zhao, X., Robu, V., Flynn, D., Barnes, M., … Nenadic, G. (2019). Machine learning methods for wind turbine condition monitoring: A review. *Renewable energy*, *133*, 620–635. doi: 10.1016/j.renene.2018.10.047

Svärd, C., Nyberg, M., Frisk, E., & Krysander, M. (2013). Automotive engine fdi by application of an automated model-based and data-driven design methodology. *Control Engineering Practice*, *21*(4), 455–472. doi: 10.1016/j.conengprac.2012.12.006

Theissler, A., Pérez-Velázquez, J., Kettelgerdes, M., & Elger, G. (2021). Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. *Reliability engineering & system safety*, *215*, 107864. doi: 10.1016/j.ress.2021.107864

Tidriri, K., Chatti, N., Verron, S., & Tiplica, T. (2016). Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges. *Annual Reviews in Control*, *42*, 63–81. doi: 10.1016/j.arcontrol.2016.09.008

Tidriri, K., Tiplica, T., Chatti, N., & Verron, S. (2018). A generic framework for decision fusion in fault detection and diagnosis. *Engineering Applications of Artificial Intelligence*, *71*, 73–86. doi: 10.1016/j.engappai.2018.02.014

Travé-Massuyès, L. (2014). Bridging control and artificial intelligence theories for diagnosis: A survey. *Engineering Applications of Artificial Intelligence*, *27*, 1–16. doi: 10.1016/j.engappai.2013.09.018

Venkatasubramanian, V., Rengaswamy, R., Yin, K., & Kavuri, S. (2003). A review of process fault detection and diagnosis: Part i: Quantitative model-based methods. *Computers & chemical engineering*, *27*(3), 293–311. doi: 10.1016/S0098-1354(02)00160-6

Wang, Z., Liang, B., Guo, J., Wang, L., Tan, Y., & Li, X. (2023). Fault diagnosis based on residual–knowledge–data jointly driven method for chillers. *Engineering Applications of Artificial Intelligence*, *125*, 106768. doi: 10.1016/j.engappai.2023.106768

Wilhelm, Y., Reimann, P., Gauchel, W., & Mitschang, B. (2021). Overview on hybrid approaches to fault detection and diagnosis: Combining data-driven, physics-based and knowledge-based models. *Procedia Cirp*, *99*, 278–283. doi: 10.1016/j.procir.2021.03.041

Witczak, M. (2006). Advances in model-based fault diagnosis with evolutionary algorithms and neural networks. *International Journal of Applied Mathematics and Computer Science*, *16*(1), 85–99.

Xiong, R., Sun, W., Yu, Q., & Sun, F. (2020). Research progress, challenges and prospects of fault diagnosis on battery system of electric vehicles. *Applied Energy*, *279*, 115855. doi: 10.1016/j.apenergy.2020.115855

Xu, Y., Kohtz, S., Boakye, J., Gardoni, P., & Wang, P. (2023). Physics-informed machine learning for reliability and systems safety applications: State of the art and challenges. *Reliability Engineering & System Safety*, *230*, 108900. doi: 10.1016/j.ress.2022.108900

Xu, Z., & Saleh, J. (2021). Machine learning for reliability engineering and safety applications: Review of current status and future opportunities. *Reliability Engineering & System Safety*, *211*, 107530. doi: 10.1016/j.ress.2021.107530

Yan, K., Ji, Z., & Shen, W. (2017). Online fault detection methods for chillers combining extended kalman filter and recursive one-class svm. *Neurocomputing*, *228*, 205–212. doi: 10.1016/j.neucom.2016.09.076

Yu, D., Shields, D., & Daley, S. (1996). A hybrid fault diagnosis approach using neural networks. *Neural computing & applications*, *4*, 21–26. doi: 10.1007/BF01413866

Zio, E. (2022). Prognostics and health management (phm): Where are we and where do we (need to) go in theory and practice. *Reliability Engineering & System Safety*, *218*, 108119. doi: 10.1016/j.ress.2021.108119

BIOGRAPHIES

Daniel Jung was born in Linköping, Sweden in 1984. He received a Ph.D. degree in 2015 from Linköping University, Sweden. In 2017 he was a Research Associate at the Center for Automotive Research at The Ohio State University, Columbus, OH, USA. Since 2022, he has been an Associate Professor at Linköping University. Since 2022, Daniel is affiliated with the Swedish research excellence center ELLIIT. His current research interests include model-based and data-driven fault diagnosis and electrification of transportation.

Mattias Krysander was born in Linköping, Sweden in 1977. He received a Ph.D. degree in 2006 from Linköping University, Sweden. Since 2012, he has been an Associate Professor at Linköping University. His current research interests include model-based and data-driven fault diagnosis and prognosis and battery systems.