



Metrics and Methods to Assess Building Fault Detection and Diagnosis Tools

Stephen Frank,¹ Guanjing Lin,² Xin Jin,¹ Rupam Singla,³
Amanda Farthing,⁴ Liang Zhang,¹ and Jessica Granderson²

1 National Renewable Energy Laboratory

2 Lawrence Berkeley National Laboratory

3 TRC Energy Services

4 University of Michigan

**NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC**

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Technical Report
NREL/TP-5500-72801
March 2019



Metrics and Methods to Assess Building Fault Detection and Diagnosis Tools

Stephen Frank,¹ Guanjing Lin,² Xin Jin,¹ Rupam Singla,³
Amanda Farthing,⁴ Liang Zhang,¹
and Jessica Granderson²

1 National Renewable Energy Laboratory

2 Lawrence Berkeley National Laboratory

3 TRC Energy Services

4 University of Michigan

Suggested Citation

Frank, Stephen, Guanjing Lin, Xin Jin, Rupam Singla, Amanda Farthing, Liang Zhang, and Jessica Granderson. 2019. *Metrics and Methods to Assess Building Fault Detection and Diagnosis Tools*. Golden, CO: National Renewable Energy Laboratory. NREL/TP-5500-72801. <https://www.nrel.gov/docs/fy19osti/72801.pdf>.

**NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC**

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Technical Report
NREL/TP-5500-72801
March 2019

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Building Technologies Office. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government.

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

U.S. Department of Energy (DOE) reports produced after 1991 and a growing number of pre-1991 documents are available free via www.OSTI.gov.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

NREL prints on paper that contains recycled content.

Acknowledgments

This work was authored by the National Renewable Energy Laboratory (NREL), operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308, and by Lawrence Berkeley National Laboratory, operated for the DOE under Contract No. DE-AC02-05CH11231. Funding was provided by the DOE Assistant Secretary for Energy Efficiency and Renewable Energy Building Technologies Office Emerging Technologies Program. The views expressed in the report do not necessarily represent the views of the DOE or the U.S. Government.

The authors thank Marina Sofos and Amy Jiron of the DOE Building Technologies Office for their support of this work. In addition, we thank the members of the DOE automated fault detection and diagnosis project technical advisory group for their reviews and feedback, and Kim Trenbath of NREL for her assistance with report preparation.

Table of Contents

1	Introduction	1
2	Methodology	2
2.1	Problem Statement	2
2.2	General Performance Evaluation Framework	2
2.2.1	Input Scenarios	2
2.2.2	Input Samples	3
2.2.3	Ground Truth	3
2.2.4	Algorithm Execution	3
2.2.5	Algorithm Outputs	3
2.2.6	Evaluation and Results	3
3	Definition of a Fault	4
3.1	Condition-Based	4
3.2	Behavior-Based	4
3.3	Outcome-Based	5
4	Definition of an Input Sample	6
4.1	Single Instant of Time	6
4.2	Regular Slice of Time	6
4.3	Other Definitions for Input Samples	6
5	Performance Metrics	8
5.1	Classification of Algorithm Outcomes	8
5.2	Static Performance Metrics	10
5.2.1	Detection Metrics	11
5.2.2	Diagnosis Metrics	11
5.3	Unified Metrics	15
6	Discussion	17
6.1	Summary of Industry Expert Opinion	17
6.1.1	Impact of Evaluation Design Choices on Evaluation Outcomes	17
6.1.2	Considerations for Data Set Generation	18
6.1.3	Considerations for Algorithm Comparison	19
7	Conclusion	20
7.1	Best Practices	20
7.2	Recommended Future Work	21

List of Figures

Figure 1. FDD performance evaluation framework 3
Figure 2. Various ways to define an input sample 7
Figure 3. Classification of fault detection and diagnosis outcomes during algorithm evaluation. 9
Figure 4. An illustrative confusion matrix with three fault types 10
Figure 5. Example ROC curve 11

List of Tables

Table 1. A summary of commonly used detection metrics 12
Table 2. A summary of commonly used diagnosis metrics 13
Table 2. – continued from previous page 14

1 Introduction

Faults and operational inefficiencies are pervasive in today's commercial buildings (Roth et al. 2005; Katipamula 2015; Yu, Yuill, and Behfar 2017). Fault detection and diagnosis (FDD) tools use building operational data to identify the presence of faults and isolate their root causes. Widespread adoption of such tools and correction of the faults they identify would deliver an estimated 5%–15% energy savings across the commercial buildings sector (Brambley et al. 2005; Roth et al. 2005). In the United States, this opportunity represents 260–790 TWh (0.9–2.7 Quadrillion BTU) of primary energy, or approximately a 2% reduction in national primary energy consumption (EIA 2012, 2018).

Fault detection is a process of detecting faulty behavior and fault diagnosis is a process of isolating the cause(s) of the fault after it has been detected. Fault detection and diagnosis are sometimes performed separately but are often combined in a single step. In the last three decades, the development of automated fault detection and diagnosis (AFDD) methods for building heating, ventilation, and air conditioning (HVAC) and control systems has been an area of active research. Two International Energy Agency Annex Reports (Hyvärinen and Satu 1996; Dexter and Pakanen 2001) and literature reviews by Katipamula and Brambley (2005a, 2005b), Katipamula (2015), and Kim and Katipamula (2018) are the major review publications in the HVAC FDD area.

Kim and Katipamula (2018) indicate that since 2004, more than 100 FDD research studies associated with building systems have been published. A great diversity of techniques are used for FDD, including physical models (Bonvini et al. 2014; Muller, Rehault, and Rist 2013), black box (Jacob et al. 2010; Wang, Zhou, and Xiao 2010), grey box (Sun et al. 2014; Zogg, Shafai, and Geering 2006), and rule-based approaches (Bruton et al. 2014; House, Vaezi-Nejad, and Whitcomb 2001). Commercial AFDD software products represent one of the fastest growing and most competitive market segments in technologies for building analytics. There are dozens of AFDD products for buildings now available in the United States, and new products continue to enter the market (Granderson et al. 2017; DOE 2018). However, considerable debate continues and uncertainties remain regarding the accuracy and effectiveness of both research-grade FDD algorithms and commercial AFDD products—a state of affairs that has hindered the broad adoption of AFDD tools.

Far more effort has gone into developing FDD algorithms than into assessing their performance. Indeed, there is no generally accepted standard for evaluating FDD algorithms. There is an urgent need to develop a broadly applicable evaluation procedure for existing and next-generation FDD tools. Such a procedure would provide a trusted, standard method for validation and comparison of FDD tools at all stages of development, from early-stage research to mature commercial products. Given the wide variety of FDD use cases and competing techniques, establishing a standard evaluation methodology is a daunting challenge. Significant progress has been made in establishing FDD test procedures and metrics within both the buildings sector (Reddy 2007b; Yuill and Braun 2013) and other industries (Kurtoglu, Mengshoel, and Poll 2008; SAE 2008). Nevertheless, existing approaches to evaluation differ significantly and much ambiguity remains.

Therefore, this report describes a general, systematic framework for evaluating the performance of FDD algorithms that leverages and unifies prior work in FDD evaluation. We outline the process required to evaluate an FDD algorithm and examine three critical questions that must be answered to apply this evaluation process:

1. What defines a fault?
2. What defines an evaluation input sample?
3. What metrics should be used to evaluate algorithm performance?

In the sections of the paper that follow we present the research methodology and findings related to fault definition, input samples, and evaluation metrics. We discuss these findings in light of key considerations for FDD algorithm performance testing, and conclude with recommendations and suggested areas of future work.

2 Methodology

The objective of the research was to develop a general and practical performance evaluation framework for FDD algorithms by synthesis of the prior research with industry domain expertise. To inform the framework, we reviewed more than 40 articles, book chapters, and technical reports related to FDD evaluation in five industries: buildings, aerospace, power systems, manufacturing, and process control. In addition, we solicited input from six FDD experts in the buildings industry. Our intended audience is the buildings industry, however, the principles outlined are broadly applicable and inform FDD evaluation methodologies for other industries.

2.1 Problem Statement

The purpose of an FDD algorithm is to determine whether building systems and equipment are operating improperly (fault detection) and, in the case of abnormal or improper operation, to isolate the root cause (fault diagnosis). The purpose of FDD performance evaluation is to quantify how well an FDD algorithm performs these two tasks. Achieving a credible outcome from FDD performance evaluation requires adherence to a clear and well-designed evaluation procedure. The purpose of the general evaluation framework presented in this report is to provide a rigorous foundation upon which such FDD evaluation procedures may be constructed. The framework is therefore descriptive rather than prescriptive: we outline the process required to evaluate an FDD algorithm and we document the choices faced by an FDD evaluator.

2.2 General Performance Evaluation Framework

Yuill and Braun (2013) describe a general FDD evaluation approach that has been successfully applied in the buildings domain. With this procedure as a starting point, Figure 1 presents a general FDD performance evaluation framework consisting of six components or steps:

1. A set of **input scenarios**, which define the driving conditions, fault types, and fault intensities (fault severity with respect to measurable quantities).
2. A set of **input samples** drawn from the input scenarios, each of which is a test data set for which the performance evaluation will produce a single outcome.
3. **Ground truth** information associated with each input sample.
4. Execution of the **FDD algorithm** that is being evaluated. The FDD algorithm receives input samples and produces fault detection and fault diagnosis outputs.
5. FDD algorithm **fault detection and fault diagnosis outputs**.
6. The FDD performance **evaluation** produces **results**, which are summarized using a set of performance metrics. The metrics are generated by comparing the FDD algorithm output and the ground truth information for each sample, then aggregating.

Components 1, 2, 4, and 5 are original to the evaluation procedure presented by Yuill and Braun (2013), while components 3 and 6 are novel.

2.2.1 Input Scenarios

Each input scenario defines a test case consisting of one or more input samples. Input scenarios may specify (Reddy 2007b; Yuill and Braun 2013):

- Building types and characteristics (age, size, use patterns, etc.)
- Equipment types
- Faults types, intensities, and prevalence
- Environmental conditions
- Data available to the FDD algorithm (*e.g.*, from sensors, meters, or a control system)

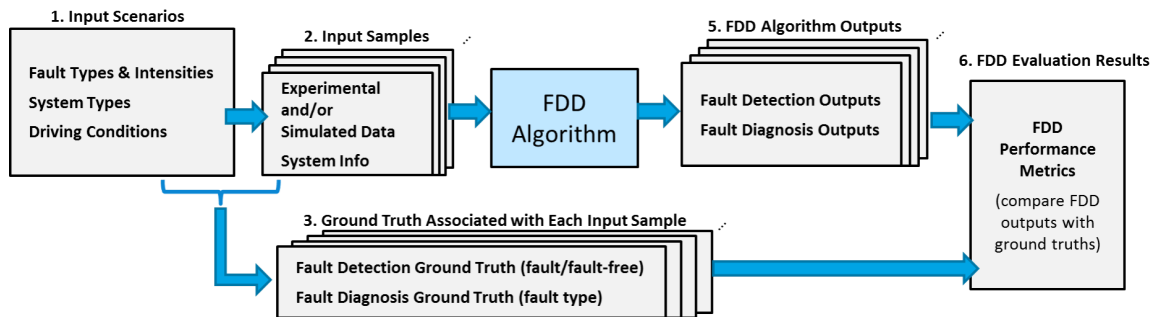


Figure 1. FDD performance evaluation framework
(expanded and generalized from Yuill and Braun [2013, Figure 1])

- Cost data (if applicable for calculating performance metrics).

2.2.2 Input Samples

Input samples are drawn from the input scenarios make up the AFDD evaluation data set. Each input sample is a collection of data for which the AFDD performance evaluation should produce a single result (Section 5.1). Input samples may include system information (metadata) and time series trend data from building sensors and control systems.

2.2.3 Ground Truth

In order to evaluate whether the output of an AFDD algorithm is correct for a given input sample, it is first necessary to establish the state of the system represented by that sample: faulted or unfaulted, and if faulted, which fault cause or causes are present. In this step, each input sample is assigned a ground truth state.

2.2.4 Algorithm Execution

In this step, the FDD algorithm is first initialized and then executed for each input sample. Initialization may include input of metadata specific to the selected input scenario(s), supervised learning using a training data set (input samples labeled with ground truth), or tuning (parameter adjustment) to adjust the algorithm's sensitivity.

2.2.5 Algorithm Outputs

For each input sample, the FDD algorithm is expected to produce a detection output that indicates whether a fault is present and a diagnosis output that presents further information about the precise nature or root cause of the fault.

2.2.6 Evaluation and Results

Evaluation results are generated by comparing the FDD algorithm's output for each sample with the ground truth data, then aggregating the results into one or more FDD performance metrics.

3 Definition of a Fault

The presence of a fault may be—and has been—defined in many ways. The existing literature and commercial FDD tools use three general methods or categories of fault definition: condition-based, behavior-based, or outcome-based.

As an introductory example, consider an air handling unit (AHU) with its cooling coil valve stuck open that is experiencing a call for heating. The unit's faulted state may be defined by the unit's condition (the chilled water valve is stuck open), behavior (the unit is simultaneously heating and cooling), or outcome (the unit's chilled water consumption is greater than expected). If, however, the same unit was cooling instead of heating, it would still be considered faulted under the condition-based definition (the valve is still stuck), but not under the behavior-based definition (it is no longer simultaneously heating and cooling). The unit's state under the outcome-based definition would be determined by the amount of chilled water flow through the stuck valve compared to an expected level of chilled water consumption.

Although rarely identified explicitly, these three categories of fault definition are used consistently in disparate fields, including aerospace, industrial process control, power systems, and buildings. With respect to building HVAC systems, Wen and Regnier (2014) distinguish between the condition-based and behavior-based categories while Yuill and Braun (2013, 2014) describe the outcome-based category. Here, we extend these prior works by formally defining and comparing the three categories.

3.1 Condition-Based

The condition-based definition of a fault is the presence of an improper or undesired physical *condition* in a system or piece of equipment. Examples of condition-based fault definitions include stuck valves, fouled coils, and broken actuators. In the case of control systems, the definition may be extended to encompass an error in the underlying control code. Although the faulty condition may (and typically will) cause improper or undesired system or equipment operation, the presence or absence of such operation does not define the presence or absence of the fault. Rather, the system is faulted so long as the faulty condition is present, regardless of whether its behavior is presently exhibiting symptoms of the fault.

Many existing articles on FDD evaluation use exclusively condition-based ground truth. Examples can be found in the aerospace (Kurtoglu, Mengshoel, and Poll 2008), defense (DePold, Siegel, and Hull 2004), power systems (Cusidó et al. 2008), water treatment (Corominas et al. 2011), and buildings industries (Gouw and Faramarzi 2014; Ferretti et al. 2015; Mulumba et al. 2015). Among articles that use different categories of fault definition for different faults, condition-based definitions are also common, for example, Morgan et al. (2010).

3.2 Behavior-Based

The behavior-based definition of a fault is the presence of improper or undesired *behavior* during the operation of a system or piece of equipment. Examples of behavior-based fault definitions include simultaneous heating and cooling and short cycling. Typically, the faulty behavior is caused by some underlying faulty condition; Wen and Regnier (2014) observe that many faults can be described in terms of either symptoms (behavior) or sources (underlying conditions). However, the key difference between the condition-based and behavior-based fault definitions is the treatment of the case when a fault condition is physically present but the system or equipment is not symptomatic: a condition-based definition still considers the system faulted, but a behavior-based definition does not.

Faulty behavior is typically defined with respect to rules—logical statements that dictate expected behavior. Alternatively, faulty behavior may be defined using observability criteria, for instance, the results of a hypothesis test that the observed sensor readings differ statistically from normal operation. Analysis of fault observability (detectability) is widely used in chemical and industrial process monitoring (Yue and Qin 2001; Joe Qin 2003).

A few articles describe mixes of faults, of which some have a behavior-based ground-truth definition: diesel engine overheating (Morgan et al. 2010), reduced condenser and evaporator water flow rates for chillers (Reddy 2007a), and failure to maintain AHU temperature and pressure set points (Wen and Regnier 2014). Regardless of the ground truth definition, use of equipment behavior as the primary fault detection criteria is common in FDD algorithms, particularly rule-based algorithms that leverage indirect sensor readings (Reddy 2007b; Yuill and Braun 2013; Ferretti et al. 2015; Zhao et al. 2017).

3.3 Outcome-Based

The outcome-based definition of a fault is a state in which a quantifiable *outcome* or performance metric for a system or piece of equipment deviates from a correct or reference outcome, termed the expected outcome. Examples of outcome-based fault definitions include increased hot or chilled water consumption (compared to an expected value), reduced coefficient of performance (compared to an expected or rated value), and zone temperature outside of comfort bounds. Although there is significant overlap between behavior-based and outcome-based fault definitions, the key feature of an outcome-based definition is the presence of an expected, or baseline, outcome against which the system or equipment performance is compared.

Use of an outcome-based fault definition is common in manufacturing and industrial process control, in which the key criterion is whether the output of the production process conforms to expected metrics or tolerances (MacGregor and Kourti 1995; Taguchi, Chowdhury, and Wu 2005). In the buildings industry, Yuill and Braun (2013, 2014) have proposed that ground truth samples for unitary equipment faults be classified as faulted or unfaulted according to their fault impact ratio (FIR), which is the ratio between the measured and baseline value of some metric of interest,

$$\text{FIR} = \frac{\text{Value}_{\text{faulted}} - \text{Value}_{\text{unfaulted}}}{\text{Value}_{\text{unfaulted}}}. \quad (1)$$

Aside from the process control industry, only a few articles surveyed used an outcome-based detection method within the FDD algorithm. Frank et al. (2016) use deviation of building energy consumption outside of normal bounds as the fault detection criteria. This approach is similar to energy monitoring tools that flag abnormal energy consumption in monthly utility bills, for example, Reichmuth and Turner (2010).

4 Definition of an Input Sample

AFDD performance evaluation requires a data library consisting of a large set of input samples, which the AFDD algorithm will process to produce raw outcomes for evaluation. There are several ways to define an input sample (Figure 2). The existing academic literature uses two common methods: a single instant of time and a regular slice of time.

4.1 Single Instant of Time

An input sample defined as a single instant of time (Figure 2a) consists of a single set of simultaneous measurements of the selected system variables, representing a snapshot of system parameters under a certain condition. This type of input sample has been used in diverse contexts, including for aerospace applications (SAE 2008), diesel engines (Morgan et al. 2010), wastewater treatment (Corominas et al. 2011), chillers (Reddy et al. 2006), and air conditioning equipment (Yuill and Braun 2013; Gouw and Faramarzi 2014).

4.2 Regular Slice of Time

An input sample defined as a regular slice of time (Figure 2b) contains multiple measurements of the selected system variables recorded within a fixed time window (for example, one day or one week). In the academic literature, time slices are typically on a repeating cycle (for example, every hour on the hour) and measurements within the time slice are recorded at a regular interval (for example, each minute). Use of this type of input sample is also common in the academic literature (Cusidó et al. 2008; Kurtoglu, Mengshoel, and Poll 2008; Jiang, Yan, and Zhao 2013; Mulumba et al. 2015; Ferretti et al. 2015; Zhao et al. 2017). In some evaluation approaches (for example, Zhao et al. (2017)), the fault is imposed for the full duration of the time slice. In other cases (for example, Ferretti et al. (2015)), the fault is imposed for only a portion of the time slice but the entire sample is nevertheless considered to represent a fault.

4.3 Other Definitions for Input Samples

Other, less common definitions for input samples include rolling time horizons, event-based windows, and hybrid windows that combine nonconsecutive measurements or combine concepts from the single instant in time and regular slice of time definitions. The rolling time horizon definition for an input sample (Figure 2c) is similar to a regular slice of time (Figure 2b), but the time window shifts through time at a fixed interval of less than the window width (for example, 60-minute windows centered on each minute of the day). Event-based input samples define a sample as a set of measurements taken within a window of time immediately before, during, and/or after a triggering event. An event may be a large change in a monitored variable (Figure 2d) or an external action, such as takeoff of an aircraft (DePold, Siegel, and Hull 2004; Simon et al. 2008) or insertion of a fault condition (Kurtoglu, Mengshoel, and Poll 2008). Use of rolling time horizon-based or event-based input samples for evaluation is uncommon in the academic literature, and the few available literature examples of event-based samples are all outside of the buildings domain. However, some commercial AFDD algorithms use these types to determine AFDD outputs.

The three papers mentioned above also illustrate hybrid definitions of an input sample. To evaluate FDD algorithms for aircraft engines, DePold, Siegel, and Hull (2004) and Simon et al. (2008) use a hybrid sample consisting of two sets of nonconsecutive steady-state measurements recorded during two separate events: takeoff and cruise. Kurtoglu, Mengshoel, and Poll (2008) combine event-based and single instant in time definitions for input samples. The evaluation samples consist of variable-length time series data collected after a fault is inserted in an electrical power system (an event). The authors compute temporal performance metrics with respect to single instances of time within this time series but use the AFDD algorithm outputs for the final instant of time within the event window to compute static metrics.

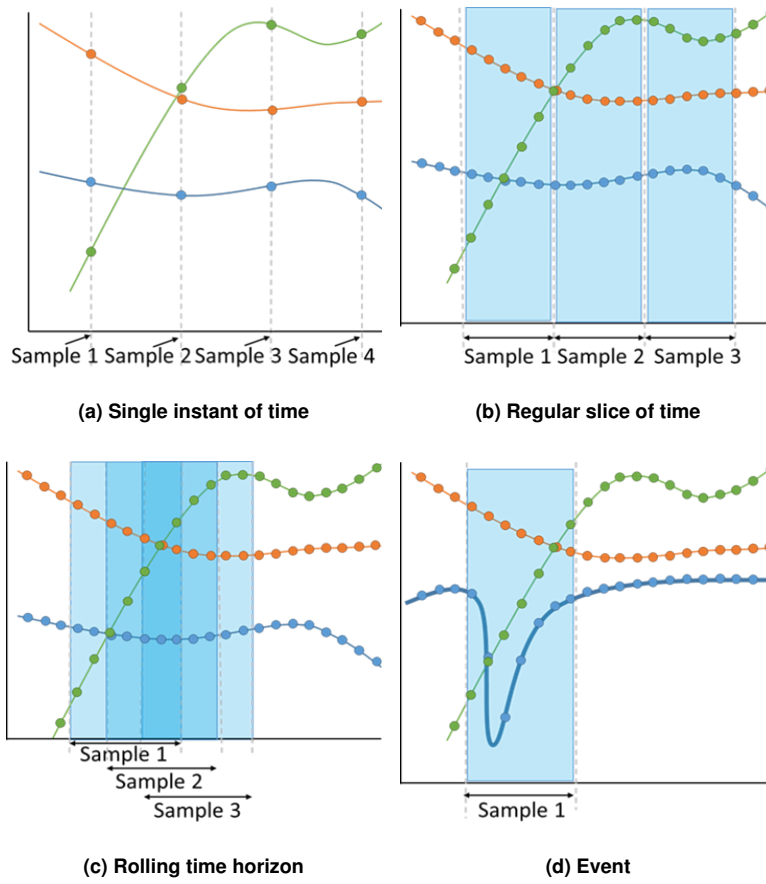


Figure 2. Various ways to define an input sample for FDD algorithm evaluation

5 Performance Metrics

FDD performance metrics are abundant in the literature (Reddy 2007b; Kurtoglu, Mengshoel, and Poll 2008; Yuill and Braun 2013), and most of them are quantitative measures. Existing AFDD performance metrics may be divided into two categories: temporal and static (Kurtoglu, Mengshoel, and Poll 2008). Temporal metrics quantify an FDD algorithm's evolving response to a time-varying fault signal, while static metrics quantify an FDD algorithm's performance with respect to a collection of samples independent of their ordering in time.

Temporal performance metrics require consideration of FDD algorithm output as a time series. Kurtoglu, Mengshoel, and Poll (2008) define five temporal FDD performance metrics: time to estimate (time to acquire a response), time to detect, time to isolate (diagnose), detection stability factor, and isolation (diagnosis) stability factor. Temporal metrics such as these are of interest in applications for which rapid response is a key requirement. For example, time from fault incipience to first detection is a common metric used in aerospace applications (Vachtsevanos et al. 2006; SAE 2008), for which rapid fault detection is critical to ensure continued safe operation. Temporal metrics also evaluate response time for fault monitoring of continuous chemical or industrial processes, such as waste water treatment (Corominas et al. 2011). Temporal metrics are less relevant to FDD in buildings, for which FDD outputs are reviewed periodically rather than continuously and fault detection is not typically time critical.

Static performance metrics describe the time-independent performance of an FDD algorithm. Most static performance metrics are computed using the same basic set of possible algorithm outcomes. This section describes these basic outcomes, reviews the use of a confusion matrix to summarize outcomes, and presents a set of standard mathematical formulas for commonly used static performance metrics.

5.1 Classification of Algorithm Outcomes

Conceptually, an FDD algorithm labels a sample as faulty or fault-free (detection), and if faulty, the possible cause(s) of the fault (diagnosis). The algorithm may also fail to provide an output for either the detection stage or the diagnosis stage. Combining these possibilities for algorithm output with possible ground truth states yields five possible outcomes for fault detection and three for fault diagnosis (Figure 3).

The possible detection outcomes are false positive (FP), false negative (FN), true positive (TP), true negative (TN), and no detection (ND).

False positive refers to the case in which the ground truth indicates a fault-free state but the algorithm reports the presence of a fault. Also known as a false alarm or Type I error,

False negative refers to the case in which the ground truth indicates a fault exists but the algorithm reports a fault-free state. Also known as missed detection or Type II error.

True positive refers to the case in which the ground truth indicates a fault exists and the algorithm correctly reports the presence of the fault.

True negative refers to the case in which the ground truth indicates a fault-free state and the algorithm correctly reports a fault-free state.

No detection refers to the case in which the algorithm cannot be applied (for example, due to insufficient data) or the algorithm gives no response because of excessive uncertainty. No detection outcomes may be further subdivided into no detection positive (ND_P) (ground truth is faulty) and no detection negative (ND_N) (ground truth is fault-free).

Although the ND outcome is not widely used in the FDD literature, it is important to consider: an FDD algorithm's detection performance may be overrated if the ND_N cases are confused or combined with TN cases.

Given TP detection outcome, the possible fault diagnosis outcomes are correct diagnosis (CD), misdiagnosis (MD), and no diagnosis (ND_G , to distinguish from no detection).

Correct diagnosis refers to the case in which the predicted fault type (cause) reported by the algorithm matches the true fault type.

Misdiagnosis refers to the case in which the predicted fault type does not match the true fault type.

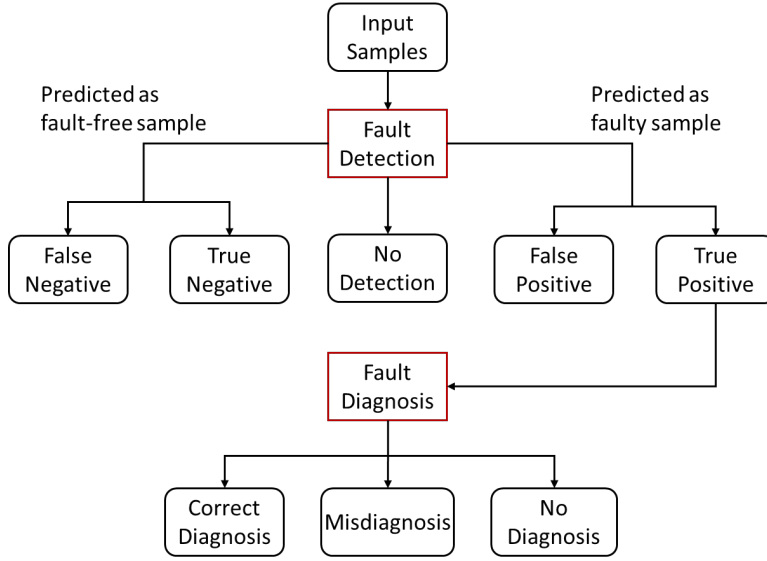


Figure 3. Classification of fault detection and diagnosis outcomes during algorithm evaluation.
(Adapted from Reddy [2007b, Figure 1])

No diagnosis refers to a case in which the algorithm does not or cannot provide a predicted fault type, for example, because of excessive uncertainty.

Deterministic diagnosis algorithms provide unique diagnoses, whereas probabilistic algorithms may provide multiple possible diagnoses. Possible ways to treat multiple fault predictions during FDD evaluation include:

- Treat non-unique diagnoses as ND_G (most strict)
- Treat non-unique diagnoses as CD if the prediction with the highest probability matches the ground truth
- Treat non-unique diagnoses as CD if at least one of the predictions above the decision threshold matches the ground truth (least strict).

In cases with multiple fault types indicated in the ground truth, an evaluation may require an algorithm to correctly predict all types (most strict), any type (least strict), or some intermediate number or fraction of types.

The confusion matrix (Figure 4) provides an intuitive way to relate the FDD algorithm output (prediction conditions) to ground truth (true conditions). To populate the confusion matrix, the basic outcomes of Figure 3 are further partitioned as follows:

- FN_i represents false negatives associated with ground truth fault type i
- ND_P and ND_N represent no detection outcomes associated with faulty ground truth and fault-free ground truth, respectively
- $ND_{P,i}$ represents no detection positive outcomes associated with ground truth fault type i
- CD_i represents correct diagnoses associated with ground truth fault type i
- $MD_{i,j}$ represents misdiagnoses where ground truth is fault type i is incorrectly diagnosed as fault type j
- $ND_{G,i}$ represents no diagnosis outcomes associated with ground truth fault type i .

Absent subscripts for fault type, these symbols represent the summation of outcomes across all fault types: $FN = \sum_{i=1}^N FN_i$, $MD = \sum_{i=1}^N \sum_{j=1, j \neq i}^N MD_{i,j}$, etc., in which N is the total number of fault types in the evaluation.

Each individual cell in the matrix represents the number of evaluation samples that resulted in each distinct outcome, while the sum of all cells represents the total number of samples used in the evaluation. In some cases, the matrix is normalized such that each cell contains the fraction of samples associated with each outcome and the sum of all cells

			Protocol Output (Prediction Condition)					
			Predicted Condition Positive (PCP)				Prediction Condition Negative (PCN)	No Detection
			Fault 1	Fault 2	Fault 3	No Diagnosis		
Ground Truth (True Condition)	True Condition Positive (TCP)	Fault 1	Correct diagnosis (CD ₁)	Misdiagnosis (MD _{1,2})	Misdiagnosis (MD _{1,3})	No diagnosis (ND _{G,1})	False negative (FN ₁)	No detection positive (ND _{P,1})
		Fault 2	Misdiagnosis (MD _{2,1})	Correct diagnosis (CD ₂)	Misdiagnosis (MD _{2,3})	No diagnosis (ND _{G,2})	False negative (FN ₂)	No detection positive (ND _{P,2})
		Fault 3	Misdiagnosis (MD _{3,1})	Misdiagnosis (MD _{3,2})	Correct diagnosis (CD ₃)	No diagnosis (ND _{G,3})	False negative (FN ₃)	No detection positive (ND _{P,3})
	True Condition Negative (TCN)	False positive (FP)				True negative (TN)	No detection negative (ND _N)	

Figure 4. An illustrative confusion matrix with three fault types

equals one. The bold box in Figure 4 includes all possible outcomes from the fault diagnosis stage, the summation of which is equal to the TP outcome total in the detection stage. Mathematically,

$$TP = \sum_{i=1}^N \left(CD_i + ND_{G,i} + \sum_{j=1, j \neq i}^N MD_{i,j} \right). \quad (2)$$

To simplify the equations for performance metrics presented in subsequent sections, the following terms are defined:

True Condition Positive the total number of positive (faulty) samples, $TCP = TP + FN + ND_P$

True Condition Negative the total number of negative (fault-free) samples, $TCN = FP + TN + ND_N$

Predicted Condition Positive the total number of samples predicted as positive (faulty), $PCP = TP + FP$

Predicted Condition Negative the total number of samples predicted as negative (fault-free), $PCN = FN + TN$

Population for fault i number of samples associated with ground truth fault type i ,

$$Pop_i = CD_i + \sum_{j=1, i \neq j}^N MD_{i,j} + FN_i + ND_{P,i} + ND_{G,i}$$

Total Population the total number of samples in the evaluation,

$$\begin{aligned} TotPop &= TCP + TCN \\ &= PCP + PCN + ND_P + ND_N \\ &= FP + TN + ND_N + \sum_{i=1}^N Pop_i. \end{aligned}$$

5.2 Static Performance Metrics

A wide variety of static performance metrics for FDD algorithms can be found in the literature. Some are specific to FDD, while others are borrowed from the more general field of classification and clustering algorithms. Tables 1

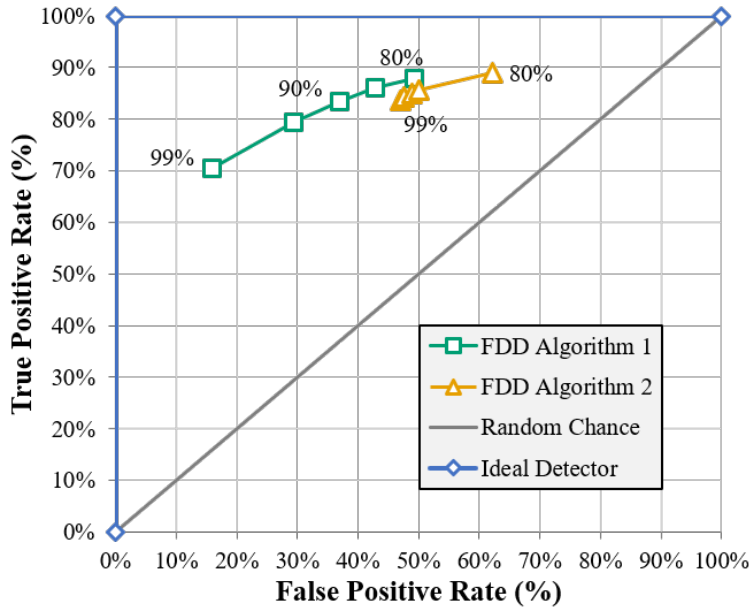


Figure 5. Example ROC curve
(adapted from Frank et al. [2016, Figure 3])

and 2 present standard formulas for commonly used metrics that leverage the notation for FDD algorithm outcomes introduced in 2.2.5. For each metric, the tables also provide commonly used synonyms; a brief definition; the range of possible values, with the best and worst scores indicated; representative citations; and usage comments.

5.2.1 Detection Metrics

Table 1 summarizes commonly used detection metrics. The following metrics can also be found in the literature, but they are used less frequently: false discovery rate (FDR), false omission rate (FOR), positive likelihood ratio (LR^+), negative likelihood ratio (LR^-), diagnostic odds ratio, and Gini coefficient (G_1).

Several common detection metrics can be visualized in a single graphical plot using the receiver operating characteristic (ROC) curve (Bradley 1997). An ROC curve is a graphical plot that plots the TPR versus the FPR of a detector or binary classifier as the discrimination threshold is varied. Another common detection metric is Area Under the Curve (AUC), which refers to the area obtained by integrating the ROC curve from $FPR = 0$ to $FPR = 1$. AUC has a minimum score of 0 and a maximum score of 1.

Figure 5 shows an example of a ROC curve for building fault detection (adapted from Frank et al. (2016)). The green and yellow curves represent the performance of two FDD algorithms as their decision thresholds are varied. The diagonal gray line represents the curve that would result from random classification of samples, while the blue curve represents an ideal detector ($AUC = 1$). In this example, FDD Algorithm 1 has better detection performance than FDD Algorithm 2 because its curve is closer to the upper left corner.

5.2.2 Diagnosis Metrics

Table 2 summarizes commonly used diagnosis metrics. Some diagnosis metrics represent multiclass versions of related detection metrics. For example, CDR_{Tot} is a multiclass version of the detection metric ACC, which only supports binary-class scenarios. Similarly, the diagnosis metric Kappa Coefficient follows the same structure as its detection metric counterpart, and the only differences are that ACC_{detect} and ACC_{random} in the detection metric are replaced by CDR_{Tot} and CDR_{random} , respectively.

Precision, Recall, and F_1 score are traditionally used for binary classification and applied as detection metrics. Sokolova and Lapalme (2009) extended these definitions into the multi-class context. The multi-class definitions of

Table 1. A summary of commonly used detection metrics

Metric	Synonyms	Definition	Equation	Range [Best, Worst]	Representative Citation	Comments
False Positive Rate (FPR)	False alarm rate, probability of false alarm, fall-out	Proportion of negatives that yield positive outcomes	$FPR = \frac{FP}{TCP}$	[0,1]	Detroja, Gudi, and Patwardhan (2007) and Ferretti et al. (2015)	$FPR = 1 - TNR$ if there are no ND_N outcomes
False Negative Rate (FNR)	Missed detection rate, miss rate	Proportion of positives that yield negative outcomes	$FNR = \frac{FN}{TCN}$	[0,1]	Detroja, Gudi, and Patwardhan (2007) and Ferretti et al. (2015)	$FNR = 1 - TPR$ if there are no ND_P outcomes
True Positive Rate (TPR)	Fault detection rate, sensitivity, recall probability of detection, hit rate	Proportion of positives that are correctly identified	$TPR = \frac{TP}{TCP}$	[1,0]	DePold, Siegel, and Hull (2004) and Banjanovic-Mehmedovic et al. (2017)	
True Negative Rate (TNR)	Specificity (SPC)	Proportion of negatives that are correctly identified	$TNR = \frac{TN}{TCN}$	[1,0]	DePold, Siegel, and Hull (2004) and Banjanovic-Mehmedovic et al. (2017)	
Negative Predictive Value (NPV)	Negative predictive rate	Proportion of negative results	$NPV = \frac{TN}{FP+TN}$	[1,0]	Banjanovic-Mehmedovic et al. (2017)	
Positive Predictive Value (PPV)	Precision	Proportion of positive results	$PPV = \frac{TP}{TP+FP}$	[1,0]	Banjanovic-Mehmedovic et al. (2017)	
Accuracy (ACC)		Proportion of correct predictions	$ACC = \frac{TP+TN}{TotPop}$	[1,0]	Vachtsevanos et al. (2006) and Banjanovic-Mehmedovic et al. (2017)	
F ₁ Score	F-score, F-measure	Harmonic mean of PPV and TPR	$F_1 = \frac{2}{\frac{1}{TPR} + \frac{1}{PPV}} = \frac{2 \times TPR \times PPV}{TPR + PPV}$	[1,0]	Derbali et al. (2017) and Banjanovic-Mehmedovic et al. (2017)	F ₁ score does not take TN into consideration. MCC and Cohen's kappa do consider TN.
Area Under Curve (AUC)	Area under the ROC curve, A',c-static	The area enclosed by a ROC curve and horizontal axis	Can be calculated by using maximum likelihood estimation or trapezoidal integration	[1,0]	DePold, Siegel, and Hull (2004) and Vachtsevanos et al. (2006)	Indicating the performance of a binary classifier. AUC of random guess is 0.5.
Kappa Coefficient	Cohen's Kappa, Kappa statistics	Compares detection accuracy to the accuracy of random chance	$\kappa = \frac{ACC_{detect} - ACC_{random}}{1 - ACC_{random}}$ where $ACC_{random} = \frac{(TN+FP) \times PCN + (FN+TP) \times PCP}{TotPop^2}$	[1,-1]	Simon et al. (2008)	Kappa provides a chance corrected coefficient of agreement. $Kappa \geq 0.75$ is considered good.
Matthews Correlation Coefficient (MCC)	Phi coefficient	A measure of the quality of binary classifications	$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	[1,-1]	Sa et al. (2017)	Generally regarded as a balanced measure even for classification problems with very different sizes

Table 2. A summary of commonly used diagnosis metrics

Metric	Synonyms	Definition	Equation	Range: [Best, Worst]	Representative Citation	Comments
Correct Diagnosis Rate, Single Fault (CDR_i)	Isolation classification rate, correct diagnosis fraction, percent correctly classified	Proportion of samples for a single fault that are correctly identified	$CDR_i = \frac{CD_i}{CD_i + \sum_{j=1, j \neq i}^N MD_{i,j} + ND_{G,i}}$	[1,0]	Simon et al. (2008)	Metric for correct diagnosis of fault i
Correct Diagnosis Rate, Total (CDR_{Tot})	Isolation classification rate, correct diagnosis fraction, diagnosis accuracy	Proportion of samples for all faults that are correctly identified	$CDR_{Tot} = \frac{\sum_{i=1}^N CD_i}{TP}$	[1,0]	Lu et al. (2016)	Metric for diagnosis of all faults
Misdiagnosis Rate (MDR)	Isolation misclassification rate	Proportion of all faults that are incorrectly or unable to be identified	$MDR = \frac{\sum_{i=1}^N (\sum_{j=1, j \neq i}^N MD_{i,j} + ND_{G,i})}{TP}$	[0,1]	Reddy (2007b)	Equal to $1 - CDR_{Tot}$
Kappa Coefficient	Cohen's Kappa, Kappa statistics	A measure for comparing the total CDR to the accuracy of random chance	$\kappa = \frac{CDR_{Tot} - CDR_{random}}{1 - CDR_{random}}$ where $CDR_{random} = \frac{\sum_{i=1}^N CD_i (CD_i + \sum_{j=1, j \neq i}^N MD_{i,j})}{TP^2}$	[1,-1]	Simon et al. (2008)	Similar definition to the detection case
Unable to Diagnose Fraction (UDF)		Proportion of no response from a diagnosis algorithm	$UDF = \frac{\sum_{i=1}^N ND_{G,i}}{TP}$	[0,1]	Reddy (2007b)	Observed fault patterns do not correspond to any rule within diagnosis rule
Misdiagnosis Cost (MDC)	Misclassification Cost	Misdiagnosis rate weighted by a cost matrix	$MDC = \frac{\sum_{i=1}^N \sum_{j=1}^N C_{i,j} CM_{i,j}}{TP}$ where the cost matrix is $C_{i,j} \in [0, 1]$, and the confusion matrix is $CM_{i,j} = \begin{cases} CD_i & \text{if } i = j \\ MD_{i,j} & \text{if } i \neq j \end{cases}$	[0,1]	Sarkar, Jin, and Ray (2011), Jin et al. (2011), and Vachtsevanos et al. (2006)	Can assign unequal costs to different classes such that misdiagnosis of more important classes incurs higher cost

Table 2. – continued from previous page

Metric	Synonyms	Definition	Equation	Range: [Best, Worst]	Representative Citation	Comments
Micro-averaging Precision (Precision _μ)	Micro-averaging Positive Predictive Value	Agreement of the ground truth fault types with those of a diagnosis algorithm if calculated from sums of per-fault-type decisions	$\text{Precision}_\mu = \frac{\sum_{i=1}^N \text{TP}_{G,i}}{\sum_{i=1}^N (\text{TP}_{G,i} + \text{FP}_{G,i})}$	[1,0]	Sokolova and Lapalme (2009)	Ratio between correct diagnosis of fault type <i>i</i> and total samples classified as fault type <i>i</i> . Diagnosis results of the more prevalent fault types have larger impact than the less prevalent ones.
Micro-averaging Recall (Recall _μ)	Micro-averaging True Positive Rate, Sensitivity	Effectiveness of a diagnosis algorithm to predict fault types if calculated from per-fault-type decisions	$\text{Recall}_\mu = \frac{\sum_{i=1}^N \text{TP}_{G,i}}{\sum_{i=1}^N (\text{TP}_{G,i} + \text{FN}_{G,i})}$	[1,0]	Sokolova and Lapalme (2009)	Ratio between correct diagnosis of fault type <i>i</i> and the total number of samples with ground truth fault type <i>i</i> . No diagnosis outcomes ND _{G,i} are not considered.
Micro-averaging F ₁ score (F ₁ score _μ)	Micro-averaging F-score, F-measure	Harmonic mean of Precision _μ and Recall _μ	$\text{F}_1 \text{score}_\mu = \frac{2 \times \text{Precision}_\mu \times \text{Recall}_\mu}{\text{Precision}_\mu + \text{Recall}_\mu}$	[1,0]	Sokolova and Lapalme (2009)	The equation is similar to the detection version but the variables are different
Macro-averaging Precision (Precision _M)	Macro-averaging Positive Predictive Value	An average per-fault-type agreement of the ground truth fault types with those of a diagnosis algorithm	$\text{Precision}_M = \frac{1}{N} \sum_{i=1}^N \frac{\text{TP}_{G,i}}{\text{TP}_{G,i} + \text{FP}_{G,i}}$	[1,0]	Sokolova and Lapalme (2009)	The average of the precision for all fault types. Unlike Precision _μ , Precision _M treats all fault types equally.
Macro-averaging Recall (Recall _M)	Macro-averaging True Positive Rate, Sensitivity	An average per-fault-type effectiveness of a diagnosis algorithm to predict fault types	$\text{Recall}_M = \frac{1}{N} \sum_{i=1}^N \frac{\text{TP}_{G,i}}{\text{TP}_{G,i} + \text{FN}_{G,i}}$	[1,0]	Sokolova and Lapalme (2009)	The average of the recall for all fault types. Unlike Recall _μ , Recall _M treats all fault types equally.
Macro-averaging F ₁ score (F ₁ score _M)	Macro-averaging F-score, F-measure	Harmonic mean of Precision _M and Recall _M	$\text{F}_1 \text{score}_M = \frac{2 \times \text{Precision}_M \times \text{Recall}_M}{\text{Precision}_M + \text{Recall}_M}$	[1,0]	Sokolova and Lapalme (2009)	Similar definition as the detection and micro-averaging cases but different variables

TP, TN, FP, and FN (marked with subscript G) are different from the binary-class version that was originally defined for detection purpose only:

- $TP_{G,i}$ represents the correct diagnoses associated with ground truth fault type i and is the same as CD_i
- $FP_{G,i}$ represents the incorrect diagnoses with predicted fault type i and is equal to $\sum_{j=1, i \neq j}^N MD_{j,i}$
- $FN_{G,i}$ represents the incorrect diagnoses with ground truth fault type i and is equal to $\sum_{j=1, i \neq j}^N MD_{i,j}$
- $TN_{G,i}$ represents the samples with ground truth fault type other than i predicted as fault type other than i (not necessarily correctly diagnosed) and is equal to $\sum_{i=1, i \neq j}^N (CD_i + MD_{i,j} + MD_{j,i})$

With these new definitions, the overall diagnosis performance can be assessed using the multi-class version of Precision, Recall, and F_1 score in two ways: one way is to directly calculate the average of the same measures for all fault types (macro-averaging marked with a subscript M), and the other way calculate the individual cumulative measure first and then calculate the performance metrics (micro-averaging marked with a subscript μ). Macro-averaging treats all fault types equally while micro-averaging favors more prevalent fault types. The macro-averaging and micro-averaging version of multi-class precision, recall, and F_1 score are shown at the end of Table 2, following the definitions in Sokolova and Lapalme (2009).

5.3 Unified Metrics

Several metrics unify detection and diagnosis results into a single index, or score, that represents the overall performance of the AFDD algorithm. Multiclass Accuracy (MAcc) describes the fraction of correctly classified samples for both detection and diagnosis,

$$MAcc = \frac{TN + \sum_{i=1}^N CD_i}{TotPop}. \quad (3)$$

We propose a similar metric, Combined Detection and Diagnosis Rate (CDDR), which describes the overall algorithm performance for condition positive (faulty) samples only. For a single fault i ,

$$CDDR_i = \frac{CD_i}{Pop_i}, \quad (4)$$

in which the denominator is the number of condition positive cases whose true condition is fault i . CDDR may also be defined for all faults,

$$CDDR_{Tot} = \frac{\sum_{i=1}^N CD_i}{\sum_{i=1}^N Pop_i} \quad (5)$$

$CDDR_{Tot}$ is essentially the same as Combined Detection and Diagnosis Recall. The minimum score is 0 and the maximum score is 1 for both MAcc and CDDR. Unlike MAcc, CDDR does not consider false positives. Combined Detection and Diagnosis Precision (CDDP), which is similar to CDDR, considers false positives but not false negatives:

$$CDDP_{Tot} = \frac{\sum_{i=1}^N CD_i}{TP + FP} \quad (6)$$

Combined Detection and Diagnosis F_1 Score (CDD F_1 Score) is proposed to combine $CDDR_{Tot}$ and $CDDP_{Tot}$ and reflect both false negatives and false positives in a single number:

$$CDD F_1 \text{ Score} = \frac{2 \times CDDR_{Tot} \times CDDP_{Tot}}{CDDR_{Tot} + CDDP_{Tot}} \quad (7)$$

Several researchers have combined the basic outcomes outlined in Section 2.2.5 with information about fault prevalence and cost to construct comprehensive performance metrics for FDD algorithms:

- Reddy (2007b) proposes an evaluation score in the form of an optimization problem: the goal of a fault detection algorithm is to minimize the combined cost of FP and FN outputs, taking into account both service costs and energy costs. The approach is applicable to both algorithm tuning and to the comparison of different algorithms. Using this approach, the author also proposes generalized and normalized detection and overall FDD scores under the condition that all AFDD algorithms to be compared have been tuned to yield the same FPR.
- Similarly, Yuill and Braun (2017) proposed a figure of merit for FDD algorithms that compares the net value of FDD (benefits minus costs) compared to a baseline case that does not employ FDD. The figure of merit is computed by comparing cost differences for multiple scenarios representing equipment type, fault type, fault intensity, ambient conditions, etc., then performing a summation weighted with respect to the probability of each scenario. Yuill and Braun (2016) provide guidelines for calculating the net value of FDD in the context of unitary equipment.
- Corominas et al. (2011) propose a time-based fault detection evaluation index for waste water treatment systems that penalizes FPs at a flat rate but penalizes FNs using an exponential decay function that starts at zero and converges to a maximum penalty value over time. The evaluation index is normalized with respect to the penalty that would be awarded for the worst possible performance. The authors observe that the penalties for FP and FN outcomes should be assigned based on the relative costs of those two outcomes to system operators.

Unified FDD performance metrics such as these are attractive because they summarize the all costs and benefits associated with using an FDD algorithm in a single score. However, the difficulty with any metric based on cost/benefit analysis is that fault prevalence and cost data are variable and highly uncertain, as acknowledged by both Reddy (2007b) and Yuill and Braun (2016). Moreover, because different FDD users have different cost structures, unified metrics must be recomputed for each individual FDD user in order to be meaningful. If an FDD evaluator retains the raw algorithm outcomes, this is possible so long as user-specific cost data are known.

6 Discussion

In order to ground the review presented in this report in the actual practice of FDD algorithm developers, vendors, implementers, and end users, we interviewed six domain experts with deep knowledge of the building analytics industry: three in the commercial sector and three in the academic sector. We provided each expert with a condensed version of the background information covered in Sections 2–5 of this report together with a list of focus questions corresponding to each of the three key topics: fault definition, input sample definition, and performance metrics for AFDD evaluation. Each expert participated in a brief individual phone interview during which they provided answers to and comments on the focus questions. Finally, we compiled and correlated the various responses. This section presents the result of these interviews, followed by a discussion of the impact of evaluation procedure choices on evaluation outcomes and on data set generation.

6.1 Summary of Industry Expert Opinion

All six industry experts agreed that both commercially available and research AFDD algorithms can be found that leverage all three fault definitions for fault detection. Experts were split on the question of what fault definition to use in a ground truth data set intended for FDD algorithm evaluation. All experts interviewed were extremely hesitant to select a single approach, citing the need for more context. Nearly all experts noted that condition-based definitions are more widely used and more appropriate for fault diagnosis, even when the detection algorithm is behavior-based or outcome-based. Experts noted that behavior-based and outcome-based fault definitions have little diagnostic power. However, experts disagreed as to whether algorithms should be penalized for differences in the fault definitions used for detection and diagnosis.

Within a given FDD algorithm, an input sample may be preprocessed into one or several analysis elements required by the algorithm. Most experts stated that they are familiar with at least one algorithm that uses each of the four ways to define an analysis: a single instant of time, a regular slice of time, a rolling time horizon, and an event. Experts noted that algorithms typically produce one output for each analysis element. When multiple analysis elements are used, these outputs may require aggregation to yield a single outcome for the input sample. All experts agreed that some form of notification delay setting commonly exists in FDD algorithms, especially in commercially available AFDD tools. The delay setting may be based on fault duration or number of fault appearances counted from intermediate AFDD results. Most experts recommended using a “regular slice of time” (time window) of one day or longer for evaluation samples, as this length is well-aligned with the design and typical use of commercially available AFDD products for buildings. The exception was for handheld diagnostic devices, for which “single instant of time” is a better choice for evaluation samples.

The six experts interviewed expressed mixed opinions regarding the selection of performance metrics. Two experts preferred combined metrics, reasoning that detection is not very useful without diagnosis. However, three other experts preferred separate metrics. They indicate that key evaluation metrics—such as FPR, FNR, and CDR—should be used collectively. The sixth expert did not offer a preference.

6.1.1 Impact of Evaluation Design Choices on Evaluation Outcomes

The evaluation design choices made for fault and input sample definitions have direct effects on FDD evaluation outcomes. In general, use of a condition-based fault definition results in the largest number of samples being classified as faulted in the ground truth data, while use of an outcome-based definition results in the smallest number of faulted samples. Therefore, all else being equal (including the samples in the evaluation data set), using condition-based ground truth will result in fewer false alarms and more missed detections, while outcome-based ground truth will result in more false alarms and fewer missed detections. Because systems and equipment may exhibit some fault symptoms (adverse behaviors) without significantly altering performance outcomes, using behavior-based ground truth is likely to yield evaluation results that fall somewhere between the results for the other two definitions. These trade-offs are apparent in the literature (Yuill and Braun 2013; Zhao et al. 2017).

One key way that the definition of an input sample affects evaluation outcomes is by defining the number of cases counted in the evaluation, which is important for ratio-based metrics. For example, if the evaluator uses a single instant of time sample definition for evaluating algorithm A and a regular slice of time (one-hour) sample definition for evaluating algorithm B, then the false alarm rates of the two algorithms cannot be fairly compared side-by-side

as the referencing point differs due to the inconsistent input sample definition. In short, for fair comparison, the definition of input sample should be consistent across all the FDD algorithm candidates involved in an evaluation. Furthermore, as confirmed by industry experts, algorithms differ in reporting timescale. As a result, regardless of the input sample definition selected, there will be instances in which FDD algorithms generate outputs at a different timescale from the input sample. The FDD evaluator should clearly document how this mismatch is handled. Zhao et al. (2017) provide an example of good practice for such documentation.

6.1.2 Considerations for Data Set Generation

To generate a data set for FDD evaluation, ground truth must be assigned to each input sample. Because fault impact varies, the evaluator must establish severity thresholds that distinguish between faulted and unfaulted samples. These thresholds should be consistent with the ground truth fault definition method that the evaluator has elected to use.

- **Condition-based ground truth:** Yuill and Braun (2013) propose the term *fault intensity* (FI), which is defined for each fault in terms of measurable numeric quantities related to the physical condition of the system or its control parameters. FI may be binary (*e.g.*, power failure) or continuous (*e.g.*, refrigerant 15% undercharged). For each fault, the evaluator should document the range of FI values that are considered sufficiently severe to include as faults in the data set.
- **Behavior-based ground truth:** the evaluator should define and document either a set of rules for expected behavior, violation of which establishes a fault, or a statistical significance test for fault observability that establishes when a fault is symptomatic. In the former case, the rules are similar to rules used in rule-based AFDD algorithms: they typically take the form of if/then statements describing expected system actions and may include tunable numeric thresholds.
- **Outcome-based ground truth:** the evaluator should first define the performance metrics (outcomes) of interest. For each outcome, the evaluator must establish and document both a baseline (expected) value (possibly different for each input sample) and the FIR that defines a fault. The requirement for a baseline complicates generation of ground truth. Yuill and Braun (2013) discuss the relative merits of various methods for obtaining the baseline.

Evaluation data may be supplied from simulation, laboratory experiments, or field measurements from a real building. Each approach has advantages and disadvantages. The closer the evaluation procedure can adhere to the realism of a field study, the greater the credibility, but the more difficult it is to obtain and sufficiently screen the data. It is important to recognize that all data sets make implicit assumptions about fault prevalence, and these assumptions affect computed performance metrics.

The input sample definition should also be considered when selecting a data set generation approach, because input sample definition constrains the available approaches for generating data and determines the efforts required to process the raw data.

- **Single instant of time type of input sample:** It is a snapshot of system operation conditions. Thus, it is usually desirable that the measurements be taken when the system is at a steady state. The steady-state requirement means that the laboratory or model should have the capability to control the operation conditions at a desired value throughout the data generation period. Steady-state operating conditions are hard to find in field data.
- **Regular slice of time type of input sample:** Longer time durations require more laboratory time, which may not be feasible for experiments due to resource constraints. In this case, simulation or building field data may be better data sources.
- **Other types of input sample** (for example, rolling window horizon and event): If a more esoteric type of input sample is selected, considerable computing or programming efforts may be required to convert the raw data to the needed structure.

6.1.3 Considerations for Algorithm Comparison

An AFDD evaluator also faces a choice between using a fixed, independent fault definition for all AFDD algorithms evaluated or tailoring the fault definition to align with the AFDD algorithm's methodology. If the same fault definition is used for evaluating multiple algorithms, then the evaluation metrics may be compared directly. However, the choice of fault definition may be perceived to disadvantage certain algorithms. Conversely, use of tailored fault definitions and associated ground truth for different AFDD algorithms allows evaluation of each algorithm with respect to its own design philosophy and limitations but complicates comparisons among algorithms.

If AFDD algorithms (particularly commercial ones) are placed in competition (whether actual or perceived), then algorithm developers may question the evaluation methodology. Such questions may include:

1. **If a fault is not observable by an algorithm, is it fair to penalize the algorithm for failing to detect it?** This situation may arise when using condition-based ground truth to evaluate a behavior-based algorithm that cannot access the underlying system state.
2. **If a fault has insignificant impact, should an algorithm really report it?** This question may arise if an algorithm employs an outcome-based detection mechanism or philosophy but the ground truth is condition-based or behavior-based.
3. **If an algorithm detects a valid fault, is it treated as a false alarm just because it does not currently have a significant impact?** This question may arise in the reverse situation from the previous question: an algorithm employs a condition-based or behavior-based detection mechanism but the ground truth is outcome-based.
4. **Should the fault definition conform to end user expectations?** This question raises the thorny issue of what exactly the end user expects, which is likely to vary depending on the end user.

Inconsistency in philosophy among AFDD algorithm developers or vendors may lead to disputes about the accuracy or fairness of the evaluation methodology, but transparency throughout the evaluation process can mitigate such concerns. Therefore, whatever approach is selected, the fault types, fault definitions, fault thresholds, and methodology for determining ground truth should be clearly documented and made available with the evaluation results.

7 Conclusion

This report proposes a general FDD performance evaluation framework and documents the design decisions required to implement the framework. The key decisions required are the definition of a fault, the definition of an input sample for evaluation, and the set of metrics to be used in the evaluation. A fault can be defined by the condition or state of a physical system, by a system's undesired or improper behavior, or by deviation of a quantitative outcome from an expected value or range. The choice of fault definition determines the ground truth classification of evaluation input samples and, by extension, affects the values of the metrics computed from FDD outcomes associated with those samples.

In the existing literature, input samples for FDD evaluation are usually defined as a single instant in time (a set of simultaneous measurements) or a regular, repeating slice of time. Commercial FDD tools may also use rolling time horizons or event-based windows. The definition of an input sample has implications for evaluation data set generation, mapping FDD outputs to performance evaluation results, and comparison of FDD algorithms.

A thorough understanding of FDD algorithm performance often requires examination of multiple metrics. The most common of these are false positive rate (false alarm rate), false negative rate (missed detection rate), and correct diagnosis rate, but many others have also been used. The most technically advanced are unified metrics: metrics that combine detection and diagnosis results into a single score, often by leveraging cost/benefit analysis. Unified metrics rely on accurate knowledge of fault prevalence, fault impact, and cost data for both energy and maintenance. Because these data are not readily available, unified metrics have, to date, been difficult to apply in practice.

7.1 Best Practices

The proposed FDD performance evaluation framework accommodates many options for specific evaluation parameters. This report provides examples of these options and design decisions from the FDD literature for buildings and other industries. Regardless of the specific options chosen, it is critical to clearly disclose and fully document all aspects of the performance evaluation for it to be credible and replicable. Documentation should address the fault, sample, and metric definitions employed; the scenarios used; and all relevant assumptions about fault prevalence, cost, etc. Additionally, "apples-to-apples" comparison of the performance of AFDD algorithms requires (i) that the algorithms be tested using consistent fault, input sample, and performance metric definitions; and (ii) that they be tested using the same evaluation data set (the same scenarios, input samples, and ground truth). If different data sets must be used (for instance, if evaluators are working independently with access to diverse data sets), then efforts should be made to align the samples statistically (*e.g.*, for similar fault prevalence and severity). These efforts should be clearly documented.

Although there is no single choice of evaluation parameters that will universally be perceived as ideal, the findings from this work indicate some consensus for design of FDD evaluation procedures. Condition-based fault definitions are commonly used in the literature for both algorithm development and as ground truth in FDD performance evaluation. Subject matter experts also noted that condition-based ground truth is the most widely employed and best aligned with diagnosis. In contrast, behavior-based approaches are relatively less frequently used for ground truth in the literature, while outcome-based approaches can present challenges for experimentally generated data sets and data sets drawn from field studies. Taken together, these findings suggest that a condition-based approach to ground truth definition represents the most practical near-term choice.

For input sample definition, regular daily time slices are well-suited for evaluating typical FDD algorithms because many such tools provide results that building operators review daily or weekly. For handheld diagnostic tools, which are often used to perform "spot checks", the best input sample definition is a single point in time. In the case of metrics, false positive rate, false negative rate, and correct diagnosis rate are the most common and therefore lend themselves to ease of interpretation across a broad audience.

7.2 Recommended Future Work

Further research can support the evolution of the proposed general AFDD performance evaluation framework into a set of standard, trusted evaluation procedures. To this end, we recommend further investigation into user and stakeholder expectations for AFDD algorithm performance and comparative analysis, development of publicly available fault performance evaluation data sets that facilitate independent comparison of FDD algorithms, and implementation of case studies that compare the effect of evaluation design choices on evaluation outcomes. Together, these will enhance the industry's understanding of the trade-offs inherent in FDD performance evaluation and the desired form and content of outcomes. To support effective implementation of unified performance metrics, high priority longer-term efforts include research to estimate fault prevalence, impact, and cost; and the quantification of the non-energy costs and benefits of acting on FDD algorithm outputs, whether accurate or inaccurate.

References

- Aerospace, SAE. 2008. *Health and Usage Monitoring Metrics: Monitoring the Monitor*.
- Banjanovic-Mehmedovic, Lejla, Amel Hajdarevic, Mehmed Kantardzic, Fahrudin Mehmedovic, and Izet Dzananovic. 2017. "Neural Network-Based Data-Driven Modelling of Anomaly Detection in Thermal Power Plant." *Automatika* 58 (1): 69–79. doi:10.1080/00051144.2017.1343328. eprint: <https://www.tandfonline.com/doi/pdf/10.1080/00051144.2017.1343328>. <https://www.tandfonline.com/doi/abs/10.1080/00051144.2017.1343328>.
- Bonvini, Marco, Michael D. Sohn, Jessica Granderson, Michael Wetter, and Mary Ann Piette. 2014. "Robust On-Line Fault Detection Diagnosis for HVAC Components Based on Nonlinear State Estimation Techniques." *Applied Energy* 124 (): 156–166. ISSN: 0306-2619. doi:10.1016/j.apenergy.2014.03.009.
- Bradley, Andrew P. 1997. "The use of the area under the ROC curve in the evaluation of machine learning algorithms." *Pattern Recognition* 30 (7): 1145–1159. ISSN: 0031-3203. doi:[https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2). <http://www.sciencedirect.com/science/article/pii/S0031320396001422>.
- Brambley, Michael R., Philip Haves, Sean C. McDonald, Paul Torcellini, David G. Hansen, David Holmberg, and Kurt W. Roth. 2005. *Advanced Sensors and Controls for Building Applications: Market Assessment and Potential R&D Pathways*. Technical Report PNNL-15149. Richland, WA: Pacific Northwest National Laboratory.
- Bruton, Ken, Paul Raftery, Peter O'Donovan, Niall Aughney, Marcus M. Keane, and D. T. J. O'Sullivan. 2014. "Development and Alpha Testing of a Cloud Based Automated Fault Detection and Diagnosis Tool for Air Handling Units." *Automation in Construction* 39 (): 70–83. ISSN: 0926-5805. doi:10.1016/j.autcon.2013.12.006.
- Corominas, Lluís, Kris Villez, Daniel Aguado, Leiv Rieger, Christian Rosén, and Peter A. Vanrolleghem. 2011. "Performance Evaluation of Fault Detection Methods for Wastewater Treatment Processes." *Biotechnology and Bioengineering* 108, no. 2 (): 333–344. ISSN: 1097-0290. doi:10.1002/bit.22953.
- Cusidó, J., L. Romeral, J. A. Ortega, J. A. Rosero, and A. García Espinosa. 2008. "Fault Detection in Induction Machines Using Power Spectral Density in Wavelet Decomposition." *IEEE Transactions on Industrial Electronics* 55, no. 2 (): 633–643. ISSN: 0278-0046. doi:10.1109/TIE.2007.911960.
- DePold, Hans, Jason Siegel, and Jon Hull. 2004. "Metrics for Evaluating the Accuracy of Diagnostic Fault Detection Systems." *ASME Turbo Expo: Power for Land, Sea, and Air, Volume 2: Turbo Expo 2004*: 835–841. doi:10.1115/GT2004-54144.
- Derbali, M., S. M. Buhari, G. Tsaramirsis, M. Stojmenovic, H. Jerbi, M. N. Abdelkrim, and M. H. Al-Beiruty. 2017. "Water Desalination Fault Detection Using Machine Learning Approaches: A Comparative Study." *IEEE Access* 5:23266–23275. doi:10.1109/ACCESS.2017.2716978.
- Detroja, K.P., R.D. Gudi, and S.C. Patwardhan. 2007. "Plant-wide detection and diagnosis using correspondence analysis." *Control Engineering Practice* 15 (12): 1468–1483. ISSN: 0967-0661. doi:<https://doi.org/10.1016/j.conengprac.2007.02.007>. <http://www.sciencedirect.com/science/article/pii/S0967066107000391>.
- Dexter, Arthur, and Jouko Pakanen, eds. 2001. *Demonstrating Automated Fault Detection and Diagnosis Methods in Real Buildings*. Finland: Technical Research Centre of Finland.
- Ferretti, Natascha Milesi, Michael A. Galler, Steven T. Bushby, and Daniel Choinière. 2015. "Evaluating the Performance of Diagnostic Agent for Building Operation (DABO) and HVAC-Cx Tools Using the Virtual Cybernetic Building Testbed." *Science and Technology for the Built Environment* 21, no. 8 (): 1154–1164. ISSN: 2374-4731. doi:10.1080/23744731.2015.1077670.
- Frank, Stephen; Michael; Heaney, Xin; Jin, Joseph; Robertson, Howard; Cheung, Ryan; Elmore, and Gregor Henze. 2016. "Hybrid Model-Based and Data-Driven Fault Detection and Diagnostics for Commercial Buildings." In *2016 ACEEE Summer Study on Energy Efficiency in Buildings*. Pacific Grove, CA: ACEEE.
- Gouw, Sean, and Ramin Faramarzi. 2014. "Is This My Fault? A Laboratory Investigation of FDD on a Residential HVAC Split System." In *2014 ACEEE Summer Study on Energy Efficiency in Buildings*, 1:84–95. Pacific Grove, CA: ACEEE.
- Granderson, Jessica, Rupam Singla, Ebony Mayhorn, Paul Ehrlich, Draguna Vrabie, and Stephen Frank. 2017. *Characterization and Survey of Automated Fault Detection and Diagnostic Tools*. Technical Report LBNL-2001075. Berkeley, CA: Lawrence Berkeley National Laboratory.
- House, J. M., H. Vaezi-Nejad, and J. M. Whitcomb. 2001. "An Expert Rule Set for Fault Detection in Air-Handling Units." *ASHRAE Transactions* 107 ().
- Hyvärinen, Juhani, and Kärki Satu, eds. 1996. *Building Optimization and Fault Diagnosis Source Book*. OCLC: 246254321. Finland: Technical Research Centre of Finland. ISBN: 978-952-5004-10-6.

- Jacob, Dirk, Sebastian Dietz, Susanne Komhard, Christian Neumann, and Sebastian Herkel. 2010. "Black-Box Models for Fault Detection and Performance Monitoring of Buildings." *Journal of Building Performance Simulation* 3, no. 1 (): 53–62. ISSN: 1940-1493. doi:10.1080/19401490903414454.
- Jiang, Qingchao, Xuefeng Yan, and Weixiang Zhao. 2013. "Fault Detection and Diagnosis in Chemical Processes Using Sensitive Principal Component Analysis." *Industrial & Engineering Chemistry Research* 52, no. 4 (): 1635–1644. ISSN: 0888-5885. doi:10.1021/ie3017016.
- Jin, X., Y. Guo, S. Sarkar, A. Ray, and R. M. Edwards. 2011. "Anomaly Detection in Nuclear Power Plants via Symbolic Dynamic Filtering." *IEEE Transactions on Nuclear Science* 58 (1): 277–288. ISSN: 0018-9499. doi:10.1109/TNS.2010.2088138.
- Joe Qin, S. 2003. "Statistical Process Monitoring: Basics and Beyond." *Journal of Chemometrics* 17, numbers 8–9 (): 480–502. ISSN: 0886-9383, 1099-128X. doi:10.1002/cem.800.
- Katipamula, Srinivas. 2015. *Improving Commercial Building Operations Thru Building Re-Tuning: Meta-Analysis*. Richland, WA.
- Katipamula, Srinivas, and Michael R. Brambley. 2005a. "Methods for Fault Detection, Diagnostics, and Prognostics for Building Systems—A Review, Part I." *HVAC&R Research* 11, no. 1 (): 3–25. ISSN: 1078-9669. doi:10.1080/10789669.2005.10391123.
- . 2005b. "Methods for Fault Detection, Diagnostics, and Prognostics for Building Systems—A Review, Part II." *HVAC&R Research* 11, no. 2 (): 169–187. ISSN: 1078-9669. doi:10.1080/10789669.2005.10391133.
- Kim, Woohyun, and Srinivas Katipamula. 2018. "A Review of Fault Detection and Diagnostics Methods for Building Systems." *Science and Technology for the Built Environment* 24, no. 1 (): 3–21. ISSN: 2374-4731. doi:10.1080/23744731.2017.1318008.
- Kurtoglu, T., O. J. Mengshoel, and S. Poll. 2008. "A Framework for Systematic Benchmarking of Monitoring and Diagnostic Systems." In *2008 International Conference on Prognostics and Health Management*, 1–13. doi:10.1109/PHM.2008.4711454.
- Lu, Chen, Yang Wang, Minvydas Ragulskis, and Yujie Cheng. 2016. "Fault Diagnosis for Rotating Machinery: A Method based on Image Processing." *PLOS ONE* 11, no. 10 (): 1–22. doi:10.1371/journal.pone.0164111. <https://doi.org/10.1371/journal.pone.0164111>.
- MacGregor, J. F., and T. Kourti. 1995. "Statistical Process Control of Multivariate Processes." *Control Engineering Practice* 3, no. 3 (): 403–414. ISSN: 0967-0661. doi:10.1016/0967-0661(95)00014-L.
- Morgan, I., H. Liu, B. Tormos, and A. Sala. 2010. "Detection and Diagnosis of Incipient Faults in Heavy-Duty Diesel Engines." *IEEE Transactions on Industrial Electronics* 57, no. 10 (): 3522–3532. ISSN: 0278-0046. doi:10.1109/TIE.2009.2038337.
- Muller, T., N. Rehaalt, and T. Rist. 2013. "A Qualitative Modeling Approach for Fault Detection and Diagnosis on HVAC Systems." In *Proceedings of the 13th International Conference for Enhanced Building Operations*. Montreal, Canada.
- Mulumba, Timothy, Afshin Afshari, Ke Yan, Wen Shen, and Leslie K. Norford. 2015. "Robust Model-Based Fault Diagnosis for Air Handling Units." *Energy and Buildings* 86 (): 698–707. ISSN: 0378-7788. doi:10.1016/j.enbuild.2014.10.069.
- Reddy, T. Agami. 2007a. "Application of a Generic Evaluation Methodology to Assess Four Different Chiller FDD Methods (RP-1275)." *HVAC&R Research* 13, no. 5 (): 711–729. ISSN: 1078-9669. doi:10.1080/10789669.2007.10390982.
- . 2007b. "Formulation of a Generic Methodology for Assessing FDD Methods and Its Specific Adoption to Large Chillers." *ASHRAE Transactions* 113 (): 334–342.
- Reddy, T. Agami, James Braun, Satyam Bendapudi, Ashish Singhal, and John Seem. 2006. *Evaluation and Assessment of Fault Detection and Diagnostic Methods for Centrifugal Chillers - Phase II*. Technical Report 1275-RP. Philadelphia, PA: ASHRAE.
- Reichmuth, Howard, and Cathy Turner. 2010. "A Tool for Efficient First Views of Commercial Building Energy Performance." In *2010 ACEEE Summer Study on Energy Efficiency in Buildings*, 3:325–338. Pacific Grove, CA: ACEEE.
- Roth, Kurt W., Detlef Westphalen, Michael Y. Feng, Patricia Llana, and Louis Quartararo. 2005. *Energy Impact of Commercial Buildings Controls and Performance Diagnostics: Market Characterization, Energy Impact of Building Faults and Energy Savings Potential*. Technical Report D0180. Cambridge, MA: TIAX LLC.

- Sa, Jaewon, Younchang Choi, Yongwha Chung, Jonguk Lee, and Daihee Park. 2017. "Aging Detection of Electrical Point Machines Based on Support Vector Data Description." *Symmetry* 9 (12). ISSN: 2073-8994. doi:10.3390/sym9120290. <http://www.mdpi.com/2073-8994/9/12/290>.
- Sarkar, S., X. Jin, and A. Ray. 2011. "Data-Driven Fault Detection in Aircraft Engines With Noisy Sensor Measurements." *J. Eng. Gas Turbines Power* 133 (8): 081602–1–081602–10. doi:10.1115/1.4002877.
- Simon, Donald L., Jeff Bird, Craig Davison, Al Volponi, and R. Eugene Iverson. 2008. "Benchmarking Gas Path Diagnostic Methods: A Public Approach." In *AASME Turbo Expo 2008: Power for Land, Sea, and Air*, 325–336. doi:10.1115/GT2008-51360.
- Sokolova, Marina, and Guy Lapalme. 2009. "A Systematic Analysis of Performance Measures for Classification Tasks." *Inf. Process. Manage.* (Tarrytown, NY, USA) 45, no. 4 (): 427–437. ISSN: 0306-4573. doi:10.1016/j.ipm.2009.03.002. <http://dx.doi.org/10.1016/j.ipm.2009.03.002>.
- Sun, B., P. B. Luh, Q. S. Jia, Z. O'Neill, and F. Song. 2014. "Building Energy Doctors: An SPC and Kalman Filter-Based Method for System-Level Fault Detection in HVAC Systems." *IEEE Transactions on Automation Science and Engineering* 11, no. 1 (): 215–229. ISSN: 1545-5955. doi:10.1109/TASE.2012.2226155.
- Taguchi, Genichi, Subir Chowdhury, and Yun Wu. 2005. *Taguchi's Quality Engineering Handbook*. OCLC: 728091434. Hoboken, NJ: Wiley [u.a.] ISBN: 978-0-471-41334-9.
- U.S. Department of Energy (DOE). 2018. *Find a Product or Service*. <https://smart-energy-analytics.org/product-service>.
- U.S. Energy Information Administration (EIA). 2012. *Commercial Building Energy Consumption Survey (CBECS)*. www.eia.gov/consumption/commercial/.
- . 2018. *Annual Energy Outlook 2018*. <https://www.eia.gov/aeo>.
- Vachtsevanos, George, Frank Lewis, Michael Roemer, Andrew Hess, and Biqing Wu. 2006. "Fault Diagnosis and Prognosis Performance Metrics." In *Intelligent Fault Diagnosis and Prognosis for Engineering Systems*, 355–399. John Wiley & Sons, Inc. ISBN: 978-0-470-11784-2. doi:10.1002/9780470117842.ch7.
- Wang, Shengwei, Qiang Zhou, and Fu Xiao. 2010. "A System-Level Fault Detection and Diagnosis Strategy for HVAC Systems Involving Sensor Faults." *Energy and Buildings* 42, no. 4 (): 477–490. ISSN: 0378-7788. doi:10.1016/j.enbuild.2009.10.017.
- Wen, Jin, and Adam Regnier. 2014. "AHU AFDD." In *Automated Diagnostics and Analytics for Buildings*, 1 edition, ed. by Barney L. Capehart and Michael R. Brambley, 467–489. Lilburn, GA: Fairmont Press. ISBN: 978-1-4987-0611-7.
- Yu, Yuebin, David Yuill, and Alireza Behfar. 2017. *Fault Detection and Diagnostics (FDD) Methods for Supermarkets - Phase I*. Technical Report 1615-RP. Omaha, NE: ASHRAE.
- Yue, H. Henry, and S. Joe Qin. 2001. "Reconstruction-Based Fault Identification Using a Combined Index." *Industrial & Engineering Chemistry Research* 40, no. 20 (): 4403–4414. ISSN: 0888-5885. doi:10.1021/ie000141+.
- Yuill, David, and James Braun. 2013. "Evaluating the Performance of Fault Detection and Diagnostics Protocols Applied to Air-Cooled Unitary Air-Conditioning Equipment." *HVAC&R Research* 19, no. 7 (): 882–891. ISSN: 1078-9669. doi:10.1080/10789669.2013.808135.
- . 2014. "Methodology for Evaluating FDD Protocols Applied to Unitary Systems." In *Automated Diagnostics and Analytics for Buildings*, 1 edition, ed. by Barney L. Capehart and Michael R. Brambley, 491–517. Lilburn, GA: Fairmont Press. ISBN: 978-1-4987-0611-7.
- Yuill, David P., and James E. Braun. 2016. "Effect of the Distribution of Faults and Operating Conditions on AFDD Performance Evaluations." *Applied Thermal Engineering* 106 (): 1329–1336. ISSN: 1359-4311. doi:10.1016/j.applthermaleng.2016.06.149.
- . 2017. "A Figure of Merit for Overall Performance and Value of AFDD Tools." *International Journal of Refrigeration* 74 (): 651–661. ISSN: 0140-7007. doi:10.1016/j.ijrefrig.2016.11.015.
- Zhao, Yang, Jin Wen, Fu Xiao, Xuebin Yang, and Shengwei Wang. 2017. "Diagnostic Bayesian Networks for Diagnosing Air Handling Units Faults – Part I: Faults in Dampers, Fans, Filters and Sensors." *Applied Thermal Engineering* 111 (): 1272–1286. ISSN: 1359-4311. doi:10.1016/j.applthermaleng.2015.09.121.
- Zogg, D., E. Shafai, and H. P. Geering. 2006. "Fault Diagnosis for Heat Pumps with Parameter Identification and Clustering." *Control Engineering Practice* 14, no. 12 (): 1435–1444. ISSN: 0967-0661. doi:10.1016/j.conengprac.2005.11.002.