Uncertainty-Aware Fault Diagnosis of Unknown Faults Using Ensemble-Based NODE Residuals

Daniel Jung
Linköping University, Sweden

Theodor Westny
Linköping University, Sweden

Abstract

This extended abstract presents an uncertainty-aware approach to data-driven fault diagnosis using ensembles of neural ordinary differential equations (NODE). The proposed method leverages physical insights in the design of NODE models for residual generation and employs bootstrap ensembles to quantify prediction uncertainty and adapt detection thresholds. The adaptive thresholds enable robust detection and isolation of faults in scenarios with limited training data and reduces the risk of false alarms due to out-of-distribution data. The methodology is validated on real-world data from the LiU-ICE benchmark, providing insights into how training data quality and residual design impact uncertainty-aware fault diagnosis.

2012 ACM Subject Classification Computing methodologies

Keywords and phrases Data-driven diagnosis, model-based diagnosis, structural methods, neural ordinary differential equations, ensemble models

Digital Object Identifier 10.4230/OASIcs.DX.2026.19

Category Extended Abstract

Funding *Daniel Jung*: D. Jung is partially funded by the Swedish excellence center ELLIIT.
Theodor Westny: T. Westny was with the Division of Vehicular Systems at Linköping University.

1 Introduction

This is an extended abstract based on the journal paper:

> Jung, D., & Westny, T. (2026). Uncertainty-aware fault diagnosis of unknown faults using ensemble-based NODE residuals. Mechanical Systems and Signal Processing, 242, 113599.

Machine learning (ML) and data-driven modeling for fault diagnosis use training data to model the relationship between different signals. The purpose of the models is to detect inconsistencies between the model predictions and observations (fault detection) and compute diagnoses (fault isolation). Still, data-driven fault diagnosis is complicated by the fact that most faults are rare events, thus, collecting relevant training data is challenging, expensive, and sometimes infeasible. Because of the lack of representative training data in data-driven fault diagnosis, one challenge is to distinguish whether a detected anomaly is caused by a faulty component or model inaccuracies caused by out-of-distribution (OOD) data.

An ensemble-based NODE approach is proposed to address the challenges of OOD data in data-driven fault diagnosis. The solution applies data-driven residuals for fault detection as well as adaptive thresholds that capture aleatoric and epistemic uncertainties. The adaptive threshold is used to reduce the risk of false alarms when the model accuracy is low. While neural networks possess the capacity to learn complex relationships between signals, they are

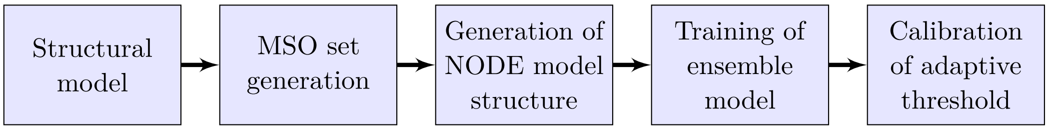

Figure 1 A summary of the NODE-based residual generator design process.

inherently susceptible to overfitting to the training domain, which can limit their ability to generalize under distribution shifts or operating conditions that are insufficiently represented in the training data.

The use of data-driven residuals based on Neural Ordinary Differential Equations (NODE) models derived from physical insights has been successfully proposed before, e.g., in [7]. The data-driven residuals can be applied in a consistency-based diagnosis framework to reason about faults even though training data from faults is scarce. However, fault isolation is complicated by false alarms caused by OOD data which motivated the need to capture the model validity or prediction accuracy of the data-driven residuals in relation to training data.

The focus of this work has been to avoid false alarms by using ensemble models instead of single NODE models to estimate model inaccuracies. The motivation is the advantage of ensemble models to simplify the calibration of an adaptive threshold since the ensemble model approximates prediction uncertainties even for complex models. A summary of the design process is illustrated in Figure 1. The proposed method is evaluated using the LiU-ICE benchmark [5] as a case study and several important conclusions have been identified for systematic design of data-driven residuals and efficient use of training data.

2 Design and Evaluation of NODE-Based Residuals

In model-based diagnosis, residual generators are constructed based on mathematical models derived based on physical insights about the system to be monitored. A structural representation of the mathematical model can be used to perform diagnosability analysis and to find redundant equation sets for residual generation using the Dulmage-Mendelsohn's decomposition [6]. Based on the structural model of the system and available measurements, a number of Minimally Structurally Over-determined (MSO) candidate sets are derived using the Fault Diagnosis Toolbox [3]. From the redundant equation sets, e.g., MSO sets, computational graphs describing the relationship between different signals describe how to compute the residual generators. Based on these computational graphs, the NODE network model structures are derived. Here, a subset of the MSO sets is selected based on three criteria:

- **i)** It must be possible to construct a residual generator in state-space form.
- **ii)** The target signal must have sufficient excitation for the NODE model to learn the system dynamics.
- **iii)** When there are multiple MSO sets that have the same target signal, select the MSO set with the largest set of input signals.

The first two criteria are needed for efficient implementation and training of the NODE models. The first criterion is fulfilled if the MSO set can be realized with integral causality [2]. The second criterion is determined by selecting NODE model candidates where the target signal span a wide range of values with respect to the operating range. For the LiU-ICE model, residual candidates are selected where pressure or air mass flow signals are used as target signals, since they vary significantly during a driving cycle, but not, e.g., temperatures, since their vary slowly during a driving cycle. The third criterion is added to maximize prediction accuracy by providing a maximum number of informative input signals. Note

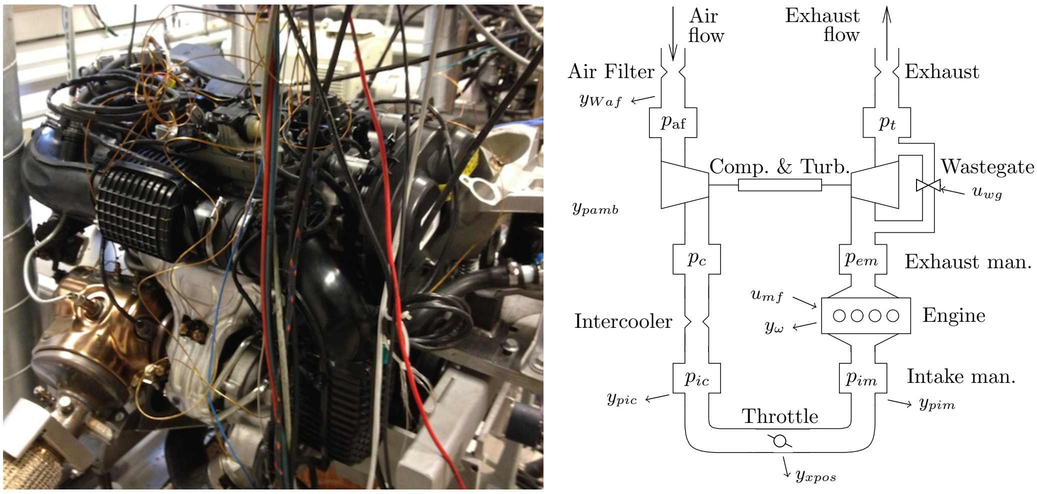

Figure 2 The LiU-ICE benchmark lab setup showing the engine test bench in the left picture and a schematic of the air path of the internal-combustion engine in the right picture.

that there is a design trade-off between maximizing detection performance and isolation performance. Since the focus in this work is on the use of ensemble models for reducing the risk of false alarms caused by OOD data, the MSO sets are selected to maximize detection performance. However, the ensemble-based methods in this work can be applied to systematic residual selection for fault detection and isolation, see, e.g., [4].

3 Data-Driven Residual Design for the LiU-ICE Benchmark

The LiU-ICE benchmark is used as a case study, which considers the air path through an internal-combustion engine, see Figure 2. A model structure is available derived from component models described in, e.g., [1]. The set of available measurements reflects a standard configuration in a production vehicle, comprising eight sensor signals: intercooler pressure y_{pic}, intake manifold pressure y_{pim}, ambient pressure y_{pamb}, intercooler temperature y_{Tic}, ambient temperature y_{Tamb}, air mass flow after air filter y_{Waf}, engine speed y_ω, throttle position y_{xpos}, as well as two actuator signals: wastegate position u_{wg}, and injected fuel mass into the cylinders u_{mf}.

The selected NODE model structures are derived from three selected MSO sets derived based on the structural model provided in the LiU-ICE benchmark. Each MSO set defines a specific residual generator structure with its corresponding input signals and output target. The general NODE model structure is in the form:

$$\begin{aligned} r_i &= y_i - h_i(x_i) \\ \dot{x}_i &= g_i(x_i, u_i) \end{aligned} \tag{1}$$

where r_i is the residual, y_i is the target signal, x_i is an internal state vector and u_i are the model inputs based on the selected MSO set. For example residual r_1 is derived based on MSO 591_85 as:

$$\begin{aligned} r_1 &= y_{pic} - h_1(x_1) \\ \dot{x}_1 &= g_1(x_1, u_1) \\ u_1 &= [y_{Tamb}, y_{Waf}, y_{xpos}, y_\omega, y_{pamb}, y_{pim}, u_{wg}, u_{mf}] \end{aligned} \tag{2}$$

In addition, two more residuals, r_2 and r_3, are implemented with target signals y_{pim} and y_{Waf}, respectively.

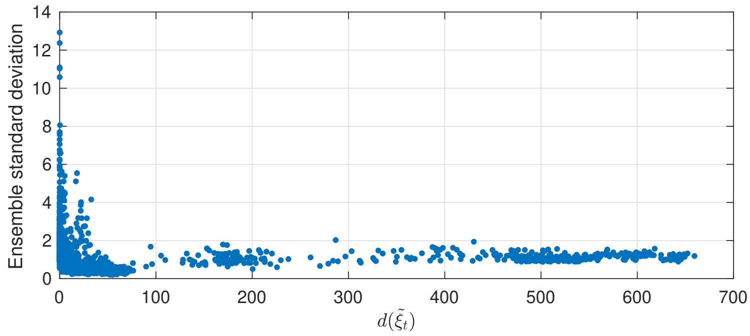

Figure 3 Example of the standard deviation of the ensemble prediction model as a function of $d(\tilde{\xi}_t)$. The standard deviation is almost constant for larger $d(\tilde{\xi}_t)$ while it varies significantly when $d(\tilde{\xi}_t)$ is close to zero.

The models were implemented in PyTorch [8] and trained on fault-free data collected from different driving cycles to minimize the mean squared error (MSE) between the model prediction and the target signal. The layer parameters in the NODE models were initialized using stability-informed initialization [9]. The Runge-Kutta (RK4) ODE solver was set to a fixed time step of 0.05 seconds, which is consistent with the sampling rate of the input signals.

4 Evaluation

The principle used to identify when anomalies are likely caused by faults and not OOD data, is when the residual output exceeds a level that can be explained by model inaccuracy alone. Since the model parameters are optimized by minimizing the MSE cost function, training will prioritize fitting to operating conditions that are well represented in training data. Thus, the idea is that model accuracy of the ensemble output is evaluated as a function of density of training data, here denoted $d(\xi_t)$ where $\tilde{\xi}_t = (\tilde{u}_t, \tilde{\hat{x}}_t)$ represents the model inputs and internal states, for different operating conditions. The generalizability of the model is assumed to degrade as soon as the model inputs ξ_t become OOD. To quantify OOD, the local density of the sample $\tilde{\xi}_t$, is computed by counting the number of training samples within a predefined neighborhood.

The ensemble standard deviation $\hat{\sigma}$ can be used as a measure of prediction uncertainty. Figure 3 shows the results of one trained residual that $\hat{\sigma}$ is almost constant when $d(\tilde{\xi}_t)$ is sufficiently large while it varies more as $d(\tilde{\xi}_t)$ goes to zero. Similar results are observed for other residuals and datasets, indicating that it is not sufficient to analyze the range of data when training a model but also the density in different operating conditions. Experiments also show that fault detection performance drops significantly when $d(\tilde{\xi}_t)$ is small and faults become easier to detect in operating conditions where $d(\tilde{\xi}_t)$ is large.

The results show that it can be easier to detect a small fault than a large fault due to increasing model inaccuracies due to OOD data when a larger fault occurs. The left plot in Figure 4 shows the residual output during a fault in sensor y_{pic} with a fault magnitude of -15% and the adaptive threshold computed based on the ensemble output variance. The fault occurs at 120 s, but the adaptive threshold exceeds the residual output resulting in missed detections. The right plot in Figure 4 shows the corresponding residual output of a fault in the same sensor, but the magnitude is +5%. Even though the threshold is varying during the scenario, the residual exceeds the threshold. Thus, the detection accuracy is higher for the small fault size.

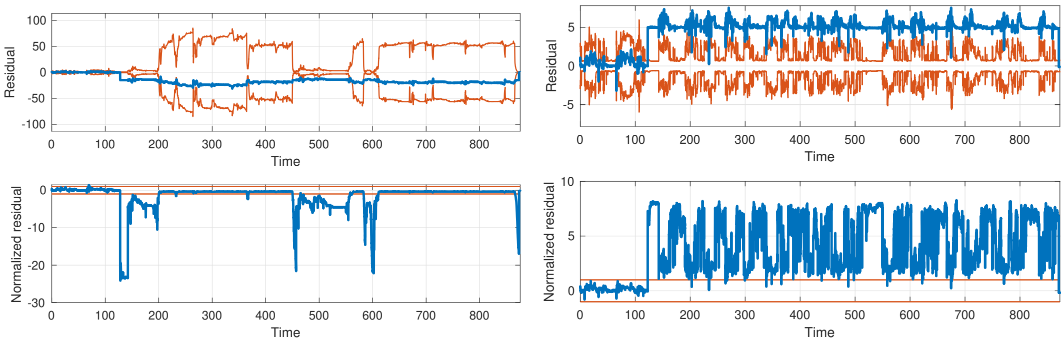

Figure 4 Evaluation of r_1 on the fault scenario $f_{ypic} = -15\%$ (left plot) and the fault scenario $f_{ypic} = +5\%$ (right plot) during the ftp75 city2 cycle. The OOD input data for the larger fault result in a higher threshold which increases the missed detection rate with respect to the smaller fault.

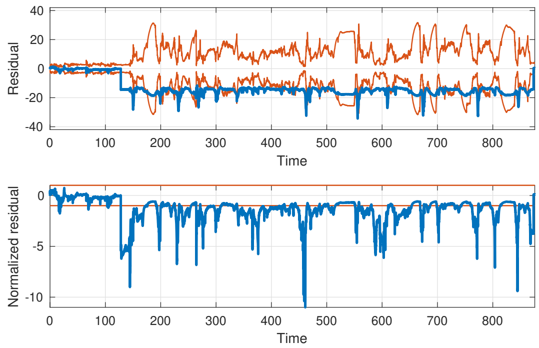

Figure 5 Evaluation of the ensemble model r_1 on the same fault scenario as in the left plot of Figure 4 but with an augmented training dataset. The adaptive threshold varies less, resulting in better detection performance.

To illustrate the impact of training data, the same NODE model for r_1 is trained with an augmented training dataset with a higher density of data from various operating conditions from additional drive cycles. With respect to the residual output in the left plot in Figure 4, Figure 5 shows the residual output from the same fault scenario, but the residual model trained on the augmented dataset. It is visible that the threshold does not increase as much during the fault scenario, because the augmented training data covers a larger variety of operating conditions. This leads to that the residual exceeds the threshold more consistently.

The results from the experiments have been validated on other residual models derived from the other MSO sets.

5 Conclusions

This study highlights the importance of handling OOD data in data-driven fault diagnosis to improve the ability to reason about previously unseen faults. Experiments show that ensemble methods can effectively quantify prediction uncertainty and thereby reduce false alarms due to OOD inputs. This advantage, however, is accompanied by a higher risk of missed detections for large faults. The results underscore the need for systematic experimental design when constructing compact yet informative training datasets. Moreover, model selection, especially the choice of input and target signals, is crucial, as OOD inputs can severely impair detection performance. At the same time, variability in ensemble predictions offers a valuable complement to residuals for identifying abnormal behavior.

References

1 L. Eriksson and L. Nielsen. *Modeling and control of engines and drivelines*. John Wiley & Sons, 2014.

2 E. Frisk, A. Bregon, J. Åslund, M. Krysander, B. Pulido, and G. Biswas. Diagnosability analysis considering causal interpretations for differential constraints. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 42(5):1216–1229, 2012. doi:10.1109/TSMCA.2012.2189877.

3 E. Frisk, M. Krysander, and D. Jung. A toolbox for analysis and design of model based diagnosis systems for large scale models. *IFAC-PapersOnLine*, 50(1):3287–3293, 2017.

4 D. Jung and E. Frisk. Residual selection for fault detection and isolation using convex optimization. *Automatica*, 97:143–149, 2018. doi:10.1016/J.AUTOMATICA.2018.08.006.

5 D. Jung, E. Frisk, and M. Krysander. The liu-ice benchmark–an industrial fault diagnosis case study. *arXiv preprint*, 2024. arXiv:2408.13269.

6 M. Krysander, J. Åslund, and M. Nyberg. An efficient algorithm for finding minimal overconstrained subsystems for model-based diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 38(1):197–206, 2007. doi:10.1109/TSMCA.2007.909555.

7 A. Mohammadi, M. Krysander, and D. Jung. Consistency-based diagnosis using data-driven residuals and limited training data. *Control Engineering Practice*, 159:106283, 2025.

8 Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems (NeurIPS)*, 2019.

9 Theodor Westny, Arman Mohammadi, Daniel Jung, and Erik Frisk. Stability-informed initialization of neural ordinary differential equations. In *International Conference on Machine Learning (ICML)*, 2024.