DELPHI – Diagnosis by ExpLoiting PHysical Insights in neural network models

Public report

Project within Elektronik, mjukvara och kommunikation

Author Daniel Jung, Linköping University
Håkan Warnquist, Scania CV

Date 2025-04-30

FFI
Fordonsstrategisk
Forskning och
Innovation

Content

FFI in short

FFI, Strategic Vehicle Research and Innovation, is a joint program between the state and the automotive industry running since 2009. FFI promotes and finances research and innovation to sustainable road transport.

For more information: www.ffisweden.se

1. Summary

DELPHI was a 3-year project led by Scania CV AB together with Linköping University as an academic partner starting in January 2022. The total budget of the project was 9,511,000 SEK, of which 4,755,000 SEK was financed through Vinnova/FFI.

The goal of the project was to develop data-driven models for fault diagnosis of dynamic systems and isolation of unknown faults by using physical insights about the system. Machine learning and data-driven fault diagnosis of technical systems is complicated by the fact that it is difficult to collect representative training data from different types of faults because these rarely occur and it can be difficult to know exactly how these manifests in the system. A solution to limited training data is therefore to construct data-driven models based on structural model knowledge and causal information to make the models less data-hungry and to be able to draw conclusions about faults that the models have not been trained on.

The results of this project have contributed to the sub-program area Intelligent and Reliable Systems. The contributions include techniques and algorithms for constructing data-driven models where it is possible to pinpoint where in the system a fault has occurred even if the fault has not been observed before. The techniques developed can be used to streamline the development of diagnostic systems by being able to use both models and data in a systematic way. The technique was evaluated using case studies from several different subsystems of a truck in the different applications of workshop diagnosis and remote diagnosis.

The project was divided into five work packages with the following content and results:

1a. Method development for the design of grey-box RNN from structural models and causal information

In this work package, methods have been developed to generate neural ODE models from structural models. Experiments with both simulation models and data from vehicles show that data-driven residuals can be used to isolate faults even when fault data is missing. As part of training neural ODE models, we have shown the connection between the properties of the trained model and the stability regions of numerical solvers. This was used in a novel method to initialize network models that significantly improved the convergence rate of the model. The methods in this work package have been validated on both simulation models and data from different case studies.

1b. Decision-making based on data-driven models trained on limited data

In this work package, methods to detect out-of-distribution data were proposed to reduce false alarms so that faults can be isolated using consistency-based diagnosis methods. A hybrid diagnosis system architecture is developed for combining model-based diagnosis and machine learning methods. A fault diagnosis benchmark has been published to inspire more researchers.

2a. Case studies for workshop diagnosis

The purpose of this work package was to evaluate the method in the scenario of workshop diagnosis where test cycles and sensors not normally available on the road can be used. A test environment was created that collects data from internal and external sensor on the vehicle and streams them to the cloud where it can be processed by machine learning algorithms. Case studies were carried out on an emission control system, a fuel injection system, and the air system of a combustion engine. This work package also studied the industrializability of the method.

2b. Case studies for remote diagnosis

The purpose of this work package was to evaluate the method in the scenario of remote diagnosis where data can be processed on board as well as being sent to the cloud for centralized processing. A test environment was created for federated learning where multiple decentralized edge devices can collaboratively be trained using local data without needing to share the data itself. In addition to case studies, this work package also studied the industrializability of a federated learning system.

Overall, the project work has been successful and all deliverables in the work packages have been met. In total the project has resulted in 19 academic publications of which one was awarded best paper. The method has been shown to work and be useful for improving the diagnostic capabilities in the automotive industry.

2. Sammanfattning på svenska

DELPHI var ett treårigt projekt lett av Scania CV AB tillsammans med Linköpings universitet som akademisk partner med start i januari 2022. Den totala budgeten för projektet var 9 511 000 SEK, varav 4 755 000 SEK finansierades genom Vinnova/FFI.

Målet med projektet var att utveckla datadrivna modeller för felsökning av dynamiska system och isolering av okända fel genom att använda fysikaliska insikter om systemet. Maskininlärning och datadriven felsökning av tekniska system kompliceras av att det är svårt att samla in representativa träningsdata från olika typer av fel eftersom dessa sällan uppstår och det kan vara svårt att veta exakt hur dessa manifesteras i systemet. En lösning på begränsade träningsdata är därför att konstruera datadrivna modeller baserade på strukturell modellkunskap och kausal information för att göra modellerna mindre datahungriga och kunna dra slutsatser om fel som modellerna inte har tränats på.

Resultaten från detta projekt har bidragit till delprogramområdet Intelligenta och Tillförlitliga System. Bidragen inkluderar tekniker och algoritmer för att konstruera datadrivna modeller där det är möjligt att identifiera var i systemet ett fel har inträffat även om felet inte har observerats tidigare. De utvecklade teknikerna kan användas för att effektivisera utvecklingen av diagnossystem genom att kunna använda både modeller och data på ett systematiskt sätt. Tekniken utvärderades med hjälp av fallstudier från flera olika delsystem av en lastbil i olika tillämpningar av verkstadsdiagnos och fjärrdiagnos.

Projektet delades in i fem arbetspaket med följande innehåll och resultat:

1a. Metodutveckling för design av greybox-RNN från strukturella modeller och kausal information

I detta arbetspaket har metoder utvecklats för att generera neurala ODE-modeller från strukturella modeller. Experiment med både simuleringsmodeller och data från fordon visar att datadrivna residualer kan användas för att isolera fel även när feldata saknas. Som en del av träningen av neurala ODE-modeller har vi visat sambandet mellan egenskaperna hos den tränade modellen och stabilitetsregionerna för numeriska lösare. Detta användes i en ny metod för att initiera nätverksmodeller som avsevärt förbättrade modellens konvergenshastighet. Metoderna i detta arbetsområde har validerats på både simuleringsmodeller och data från olika fallstudier.

1b. Beslutsfattande baserat på datadrivna modeller tränade på begränsade data

I detta arbetsområde föreslogs metoder för att upptäcka data som avviker från träningsdata för att minska falsklarm så att fel kan isoleras med hjälp av konsistensbaserade diagnosmetoder. En hybrid diagnossystemarkitektur utvecklades för att kombinera modellbaserad diagnos och maskininlärningsmetoder. Ett felsökningsbenchmark har publicerats för att inspirera fler forskare.

2a. Fallstudier för verkstadsdiagnos

Syftet med detta arbetspaket var att utvärdera metoden i scenariot för verkstadsdiagnos där testcykler och sensorer som normalt inte är tillgängliga på vägen kan användas. En testmiljö skapades som samlar in data från interna och externa sensorer på fordonet och strömmar dem till molnet där de kan bearbetas av maskininlärningsalgoritmer. Fallstudier genomfördes på ett emissionskontrollsystem, ett bränsleinsprutningssystem och luftsystemet för en förbränningsmotor. Detta arbetsområde studerade också metodens industrialiserbarhet.

2b. Fallstudier för fjärrdiagnos

Syftet med detta arbetspaket var att utvärdera metoden i scenariot för fjärrdiagnos där data kan bearbetas ombord samt skickas till molnet för centraliserad bearbetning. En testmiljö skapades för federerad inlärning där flera decentraliserade edge-enheter kan samarbeta genom att använda lokala data utan att behöva dela själva datan. Förutom fallstudier studerade detta arbetspaket också industrialiserbarheten för ett federerat inlärningssystem.

Sammantaget har projektarbetet varit framgångsrikt och alla leveranser i arbetspaketen har uppfyllts. Totalt har projektet resulterat i 18 akademiska publikationer varav en har fått pris för bästa artikel. Metoden har visat sig fungera och vara användbar för att förbättra diagnostikförmågan inom fordonsindustrin.

3. Background

Fault diagnosis involves detecting when a technical system deviates from expected behavior due to degradation or failing components. In the automotive industry, diagnosis systems have long been essential for detecting faults in components that may affect vehicle emissions. Another motivation is that unplanned roadside breakdowns are costly, especially in heavy transportation. A diagnosis system should detect faults in complex systems at an early stage, provide timely warnings that allow for planned maintenance, and assist mechanics in quickly identifying the faulty component. If the diagnosis system fails to detect faults in time or generates false alarms, trust in the vehicle decreases, and customer satisfaction is negatively impacted due to unnecessary workshop visits and increased service costs.

One way to diagnose components using time series data is model-based diagnosis, see e.g. [20]. In this case, a physically based model of all components in the system is first created. By comparing various actual measurement values with values produced through simulations of different sub-models, referred to as residual generators, it is possible to identify which of these sub-models deviate from the nominal behavior and thus should contain a fault. Given a good physical model, this technique can be very successful in diagnosing the type of system found on a truck. The problem, however, is that it is difficult and costly to produce models with sufficient accuracy in practice.

When training data is available from various operating conditions, a promising approach is the use of data-driven fault diagnosis methods and machine learning. In this case, models are trained using data from the system with which they can classify new data with the correct diagnosis. This typically requires collecting training data that comes from the faults that should be diagnosed, which is a problem since each individual fault rarely occurs, and perhaps only after the system has been used for a long time. It is possible to train a model with data from fault-free cases and in this way identify deviations that indicate faults. The difficulty then is to get good enough fault isolation for the diagnosis to be useful off-board.

In model-based diagnosis, residual generators are modeling nominal system behavior to detect abnormal behavior caused by faults. Fault isolation is possible by designing multiple residual generators where each residual is only sensitive to a subset of faults. Diagnosis candidates are computed based on the residual patterns using some fault isolation logic. If it is possible to design data-driven models for residual generation that are only sensitive to a subset of faults, fault isolation can be done using the same fault isolation logic without the need of representative data from the different faults. This observation has been a motivation for this project.

Anomaly classifiers can be trained using fault-free data to detect abnormal system behavior. However, data-driven models are not able to reason about its cause when data from faults are not available. Therefore, it is necessary to bridge theory from model-based diagnosis with data-driven fault diagnosis to understand how to design anomaly classifiers that can be used to reason about abnormal behavior. Instead of talking about

model-based diagnosis and data-driven diagnosis there is a need for a common fault diagnosis framework where different methods can be systematically combined in a hybrid diagnosis system [21].

Previous research in, e.g., [22] has shown that including information about model structure and causal relationships between signals in data-driven models based on physical insights can be a solution. It was shown that basic physical understanding about the system is sufficient to derive machine learning model structures for fault isolation without the need of representative faulty data to train the models, see e.g. [23]. Deriving model structures from physical insights can be used to automate the residual design process instead of using trial-and-error which is a common design principle in machine learning. The physical modeling then becomes simple enough to be performed by engineers on a large scale. With the help of insights from this limited physical model, a data-driven model trained on mostly nominal data can still achieve fault isolation of unknown faults.

In 2020, Scania supported research in this area that was conducted at Linköping University. Based on literature surveys that were made it was concluded that this project is in the absolute forefront in fault diagnosis research. A proof-of-concept was implemented and evaluated with promising results [24].

4. Purpose, research questions and method

The purpose of this FFI project has been to improve powertrain fault diagnosis by understanding how to combine machine learning with physical insights for detection of abnormal system behavior and reasoning about its cause. More specifically, this project has focused on neural networks designed using physical insights for fault isolation when training data is scarce and how to combine machine learning and model-based diagnosis in a consistency-based diagnosis framework. Another important purpose is to develop these techniques to make them more applicable in automotive industry. This requires systematic methods and decision tools to generate models and evaluate the model performance.

The developed neural network-based residuals improve the fault isolation process, e.g. at the workshop, by efficient testing procedures or during operation by analyzing time-series data. Deriving neural networks from physical insights simplifies the diagnosis system design process by modeling complex relations between signals from training data. If each neural network models a specific subsystem, data-driven residuals can be used in a consistency-based diagnosis framework to compute diagnoses. This requires a theoretical framework to systematically combine model-based diagnosis methods and machine learning in a diagnosis system. The conclusions drawn from different diagnosis methods should be combined to avoid rejecting the true diagnosis.

The main research questions of the project are the following:

- How can different neural network structures be realized for a dynamic system based on structural and causal information? This also includes to investigate how information about dynamics should be integrated into the network.
- How can a suitable network structure be systematically selected for fault isolation? How can available information of analytical dependencies between components in the system be integrated into the neural network model to reduce the risk of overtraining and to improve the interpretability of the models?
- How can data-driven residuals be used in a consistency-based diagnosis framework when training data is not representative of all operating conditions? This is also important to combine information from model-based residuals and data-driven residuals when computing diagnoses.
- How can this method of fault diagnosis be applied to two different automotive use cases? The first use case is workshop diagnosis when tests can be run in controlled environments with additional external sensors that normally do not exist on the vehicle. The second use case is remote diagnosis when data can be processed both onboard and offboard? These use cases are important to investigate how the method can be industrialized, e.g. what infrastructure is needed and how the work process should be?

The method in this project is to develop theory and algorithms for generating neural network model structures and model training. A framework will also be developed on how to use data-driven residuals for fault isolation in a consistency-based framework. To validate the developed methods, data from both simulations and real case studies will be used. The different case studies will be used to identify various types of problems and properties of data, especially when there is a lack of representative data from relevant faults and operating conditions. The data collection should represent realistic operating conditions of the system but also industrially relevant faults.

To address the research questions, the project has been separated into four different work packages. The first two work packages focus on the theoretical research questions where the second two target the question regarding industrialization. The first work package develops methods for designing neural networks from physical insights and addresses research question 1 and 2. The second work package focus on how to do fault isolation using data-driven residuals using a consistency-based diagnosis framework and addresses research questions 2 and 3. The third and fourth work packages focus on the implementation and evaluation of the methods developed in the first two work packages addressing research question 4. This includes setting up relevant case studies and investigate the industrialization aspects.

5. Objective

In this project, the goal was to develop techniques for systematically designing data-driven predictive models for a given system, which can be used not only to detect when a fault has occurred but also to pinpoint where it has occurred, even if the fault has not been observed before. The intention was to explore neural networks and how their structure can be designed to model physical insights about the system as well as causal information, including details about dynamic states and relationships between different actuators and measurement signals. Physical insights and causal information are assumed to be represented in the form of a structural model that describes how the system is constructed, how different components are connected, and how they interact based on physical principles, even when the exact analytical model relationships are unknown.

The project aims to address the following unresolved challenges with this technology:

- We must be able to handle the fact that the vehicles from which we collect data come in many different variants. The data is not rich enough to be limited to only vehicles that are built the same.
- Another challenge is that the technology performs poorly in operating conditions it has not been sufficiently trained on. Therefore, a method for handling model uncertainties and missing data needs to be developed.
- The project also aims to investigate whether models can be improved through smart data collection strategies to cover operating points where the model is weak. The technology is designed to be trained on nominal data, which is the most common type of data. For more efficient data utilization, we also want to be able to handle data with known faults when available.

Industrialization of machine learning solutions in the automotive industry requires good understanding of the system and knowledge transfer between machine learning developers and application engineers. Addressing issues with limited training data and model interpretability is necessary for improving trustability of machine learning models in automotive applications. To improve the competitiveness of Swedish automotive industry in machine learning motivates research projects in this topic to strengthen the competence in applied machine learning in the industry.

6. Results and deliverables

The program area within Electronics, Software, and Communication where this project contributes the most is the subarea called Intelligent and Reliable Systems, specifically the cornerstone of Machine Learning.

The work in this project was organized in five work packages:

1a Method development for design of grey-box recurrent neural networks using structural models and causal information

1b Decision-making based on data-driven models trained on non-representative training data

2a Case studies for workshop diagnosis

2b Case-studies for remote diagnosis

3 Project management

The main part of the research in this project has been done as part of a Ph.D. research project. The Ph.D. student, Arman Mohammadi, started in January 2022 and planned future project delivery is the dissertation is planned in January 2026. From Linköping University, the project has also involved three senior researchers: Daniel Jung, Mattias Krysander, and Erik Frisk. The work related to training of neural ODEs has been done in collaboration with Ph.D. student Theodor Westny. Eight master's thesis projects, two student summer projects, and one student project, have also been organized within the project.

Scania was responsible for providing data and resources for the case studies. This was done by both Scania engineers and master thesis students. Data collection for the case study on the engine located in the Vehicular Systems lab at Linköping University was carried out by Linköping employees.

6.1 Results from work package 1a: Model development for design of grey-box RNN from structural models and causal information

The main research question of this work package is to develop methods for systematic design of neural networks of dynamic systems for residual generation using structural and causal information. This includes how to integrate information about dynamics in the model structure and how the selected model structure affects its generalizability. Since many model candidates can be derived from a structural model it is relevant to identify suitable neural network model structures and understand how physical insights should be integrated in the model structure. The developed techniques are applied and evaluated using case studies from Work packages 2a and 2b.

In the project application, the originally planned deliverables for this work package were the following:

- Develop methods to identify and generate data-driven residuals from a structural model.
- Methods should be implemented in a toolbox and evaluated on the case studies-
- Results are presented in scientific papers and master's thesis reports.
- The results will also be presented in the PhD thesis of Arman Mohammadi.

Methods have been developed for the automated design of neural ODE models for modeling dynamic systems. A systematic design of neural ODEs from a structural model has been proposed. From the structural model, different computational graphs are derived based on structurally redundant equation sets. Each computational graph describes the relationship between a set of input and output signals. From the structural model, each computational graph models a set of components which gives an interpretation of what subsystem is modeled by each computational graph, see [1,2,3,5].

To improve the generalizability of the models, the number of signals used as inputs to the models has been reduced. Since the developed neural networks model different subsystems, we have shown how training data can be used more efficiently by including data from fault scenarios in fault-free data to train residuals that are insensitive to these faults [1].

This project has shown that deriving neural ODE model structures based on physical insights can be used to design residuals with desirable fault isolation properties without the need of training data from faults. The project has demonstrated that these neural ODE models are useful for isolating unknown faults [1]. Methods have been developed for the automatic generation of models in Python code using structural models. The model code is standardized to enable systematic training and evaluation in PyTorch. The proposed methods have been validated on different case studies: an aftertreatment system [1,2], a fuel injection system [5], and an internal combustion engine [9].

To model dynamic systems, neural networks need to integrate state variables. In general, both the neural network and the numerical integration process are trained simultaneously. In neural ODE models, established numerical solvers, such as Euler forward or Runge-Kutta, are used instead of learning this from data. The project has evaluated how the choice of solver affects the trained model. In [4], we have demonstrated that the stability region of a chosen solver, i.e., how fast dynamics can be simulated without the model diverging, limits the type of dynamics that the neural network can learn. A model trained with a higher-order solver can learn faster dynamics compared to a lower-order solver. Even though the computational time per epoch is higher when using a higher order method, the convergence rate is also faster. In this project, it has been shown experimentally that when linearizing a neural network around the operating conditions in training data, the eigenvalues of the trained model become bounded by the stability region of the selected solver [6].

It is also shown that a model trained using a higher order method can be simulated using a lower order method but not the other way around, see [4]. The reason is that the learned dynamics of the trained model is bounded by the stability region of the solver used during training. This means, for example, that a model that is trained using a higher order method has eigenvalues outside of the stability region of the lower-order method. We showed that the stability properties when simulating dynamic models also put a constraint when learning models from data [6]. These results are important when training prediction models, e.g. for residual generation, because the time-series data often has a given sampling time, and the selected solver can then be used to define what dynamics to learn by the model.

Based on the results from numerical solver selection, new initialization methods have been proposed for more efficient training of neural networks, see [6]. The initialization of model parameters in the neural network accounts for the stability region of the solver used to simulate the neural network model, preventing instability in the simulation and ensuring faster convergence compared to standard methods suggested in frameworks like PyTorch. Experiments using data from multiple applications validate that the model converges much faster using the proposed initialization method compared to standard techniques that are recommended in e.g. PyTorch. The developed techniques have been used in this project but have also been made publicly available in git-repos:

- https://github.com/westny/neural-stability, and
- https://github.com/westny/neural-residual.

To investigate different methodologies and case studies, various student projects have been organized. For example, in the master's thesis project [12], different data-driven models, e.g. SINDy and Random Forests, derived from physical insights for fault detection were evaluated. A student project developed a prototype of a remote diagnosis system for the aftertreatment case study utilizing data-driven models [19].

We have also explored the use of ensemble models to handle different types of model uncertainties and avoid misclassifications, both ensembles of neural ODEs and ensembles of probabilistic neural networks. The probabilistic neural network predicts a target distribution instead of only the target signal. Based on the ensemble predictions, experimental results show that it is possible to distinguish between aleatoric and epistemic uncertainties, i.e. uncertainties caused by lack of information in the selected input signals and the wrong model structure and uncertainties caused by out-of-distribution data. The ensemble models avoid false alarms caused by out-of-distribution data and calibrate an adaptive threshold based on the estimated model uncertainties.

6.2 Results from work package 1b: Decision making based on data-driven models that are trained on limited data

In this work package, methods have been developed for fault isolation of abnormal system behavior based on the principles of consistency-based diagnosis using the residual generators in work package 1a. The purpose has been to develop theory and algorithms to combine the results from model-based and data-driven residuals to compute diagnoses. The proposed methods handle model inaccuracies of data-driven models caused by limited training data and unknown faults. The developed methods are evaluated using the residuals developed in work package 1a.

In the project application, the originally planned deliverables for this work package were the following:

- Methods to compute diagnoses using data-driven residuals and consistency-based diagnosis taking into consideration non-representative training data

- Evaluations of the developed methods using simulated data and data from case studies are presented in a technical report or scientific papers

A methodology has been developed for using data-driven residuals in a consistency-based diagnosis framework [1,10]. This requires structured residuals, i.e. residuals that are sensitive to some faults but insensitive to others. The structural sensitivity to faults can be derived from the structural model and the computational graph used to derive the neural network model structure. However, false alarms must be avoided as they will falsely reject the true diagnosis. This is complicated by that when a fault happens in the system that the residual should be insensitive to, the fault could change the operating conditions of the system that are not represented in training data. A fault can introduce system behavior that cannot be represented by a fault-free system only.

One contribution is the detection of anomalies in the inputs to a data-driven model, enabling conclusions to be drawn about whether a detected anomaly is caused by a fault [10]. When model inputs in test data deviates from training data it is assumed that a data-driven model does not generalize well, and it is not known what is causing the anomaly. In this project, the test validity region is formulated using a one-class support vector machine trained on model inputs using training data.

In [10], a decision logic is formulated that works in a consistency-based framework, where detected abnormal behavior that is outside of the model's validity region is only considered to be 'out of range'. This information is not used to reject fault hypotheses but can be used to identify when something has happened that requires special attention by an engineer. Another relevant application of the test validity region is to identify new datasets that can be relevant to augment training data to cover more operating conditions to improve generalizability. A training process is formulated to implement a data-driven residual with a test validity region that takes into consideration available training data and adaptive thresholds [1].

Which signals that are used as input and target signals have a significant impact on a residual's test validity region. Results show that multiple residuals based on the same signals are needed to maximize the validity region for the set of residuals [10].

Hybrid diagnosis systems combine model-based and data-driven methods to leverage their respective strengths and mitigate individual weaknesses in fault diagnosis. In [11], we have proposed a unified framework for analyzing and designing hybrid diagnosis systems, focusing on the principles underlying the computation of diagnoses from observations. The framework emphasizes the importance of assumptions about fault modes and their manifestations in the system. The proposed architecture supports both fault decoupling and classification techniques, allowing for the flexible integration of model-based residuals and data-driven classifiers.

The novelty of the proposed architecture is that it is shown to be a generalization of classical model-based diagnosis system design and a data-driven classification. Having this diagnosis assumption-based perspective simplifies the design of hybrid diagnosis systems since it gives a general principle of how to utilize different diagnosis methods, such as classical model-based and data-driven methods, to compute diagnoses. The

proposed framework emphasizes that the key factor in categorizing fault diagnosis methods is not whether they are model-based or data-driven, but rather their ability to decouple faults which is crucial for rejecting diagnoses when fault training data is limited.

We have also shown the connection between properties in model-based diagnosis, such as analytical redundancy, and data properties, such as the intrinsic dimension of data, see [12]. These results are important to bridge the theory and methods developed in the model-based diagnosis community to data-driven fault diagnosis to deal with incomplete training data and unknown faults.

In [11], we have demonstrated that rather than distinguishing between model-based and data-driven diagnosis, it is more important to focus on the ability of different models to decouple faults. Since training data is limited, it is not possible to reject fault hypotheses if the test data deviates from the training data. To isolate unknown faults, it is crucial that residuals are designed so that the model input does not deviate from the training data when a fault occurs. This has been demonstrated through both simulation studies and experimental data from case studies. If faults cannot be decoupled, residuals can be used as inputs to other data-driven models to generate a prioritized list of the most probable diagnoses.

To inspire other researchers to work on fault diagnosis with limited training data and structural models, the LiU-ICE fault diagnosis benchmark has been published which is based on an internal combustion engine test bench at Linköping university [13]. Data from the test bench has been published as part of the benchmark and was used in a scientific competition that we organized in 2024 at the 12th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS) in Ferrara, Italy. The competition had six participating teams from all over the world.

6.3 Results from work package 2a – Case studies for workshop diagnosis

The purpose of this work package was to study how the methods developed in work packages 1a and 1b can be applied for workshop diagnosis. It also had the purpose of evaluating how the methods developed in the project could be industrialized.

In the project application, the originally planned deliverables for this work package were the following:

- A test environment for collecting data from external and internal sensors on a vehicle in the workshop.
- The modelling and evaluation of three different subsystems.
- Master theses with published reports

Although a modern vehicle has many sensors, it is typically designed to not have any more sensors than needed for the control of the vehicle. This poses a challenge for the method proposed in this project as there will be too little analytical redundancy to be able

to isolate the faults with sufficient granularity to be useful for the workshop technician. To counter this, it is possible to add additional external sensors to the system that are used only when running tests in the workshop.

In this work package, a test system was built that enabled a technician to easily add more sensors to vehicle that could be used when running tests in the workshop. The system is able to simultaneously collect data from these external sensors and the internal sensors the vehicle and send it up to cloud servers where the data could be analyzed and visualized. The measurement system was designed to be modular and affordable so that several different measurement configurations and sensor types could be supported. By having the system integrated to the cloud ensures that data from all tests is always stored so that more diagnostic models can be trained in the future.

The first use case was on the Selective Catalytic Reduction (SCR) system that is responsible of reducing the amount of nitrogen oxides (NOx) in the emissions by injecting the appropriate amount of urea into the exhaust system. This system was selected because of it is particularly difficult to troubleshoot because it contains few sensors to give it sufficient redundancy to do fault isolation. Additional pressure sensors were added in the dosing system and data was collected from trucks in different workshop scenarios where different types of common faults were implemented. Data from this case study was used in the publications [1, 2, 4]. It was shown that the proposed method works for this scenario. With the additional sensors it was possible to isolate faults that could not be isolated before. A known challenge with this method is that when diagnostics is done using data from unknown faults for which the model is not trained on, it can sometimes misclassify this data because the fault causes the system to enter operating conditions that are never entered under normal operating conditions. With the test system built for this work package, it was possible to control the system into operating points that the system normally does not enter and thereby improve the quality of the trained models so that they better can isolate unknown faults.

The second use case was performed on the fuel injection system responsible for delivering fuel at the correct pressure to the injectors. Additional pressure sensors were added to the low-pressure part of the system and data was collected from a vehicle running on the test track with implemented blockage faults with varying degrees. Data from this case study was used in the publication [5].

The third use case was performed on the air intake and exhaust system of the turbo-charged engine in the LiU laboratory. Data was collected in from several different fault scenarios. This data was used in the publications [6,9,13] and made publicly available in LiU-ICE fault diagnosis benchmark [13]. By being used in the diagnostic competition arranged in work package 1b, the method developed in this project could be evaluated against other state of the art diagnostic methods of other research teams.

To evaluate industrial applicability of the methods, the first use case was revisited at the end of the project and new models were created using Scania engineers and tested using the methods and the toolbox provided by LiU in work packages 1a and 1b. This confirmed that the knowledge transfer from university to industry was successful and that the methods were mature enough to be industrialized.

6.4 Results from work package 2b – Case studies for remote diagnosis

The purpose of this work package was to develop a test environment that could be used for applying the method to the remote diagnosis scenario. In this scenario diagnosis occurs while the vehicle is in operation, and it has a connection to cloud servers via the mobile network.

In the project application, the originally planned deliverables for this work package were the following:

- A test environment for collecting data from internal sensors on a vehicle on the road capable of collecting buffered data from interesting events and being able to locally execute diagnostic algorithms.
- The modelling and evaluation of three different subsystems.
- Master theses with published reports

In this work package a test environment was created for federated learning. Federated learning is a machine learning approach where multiple decentralized edge devices can collaboratively be trained using local data without needing to share the data itself. This is particularly useful for remote diagnosis where high frequency time series data is plentily available on-board, but the mobile connection is too limited for it all to be sent up to the cloud. The test environment was created such that it is possible to cross-compile it for multiple ECU platforms making it flexible as a research platform. The client runs onboard the vehicle and communicates with a federated learning framework in the cloud that makes it possible to send down machine learning models and orchestrate them. A hardware rig was built consisting of 6 edge devices running the client and being able to process pre-recorded signal data.

In a case study, on the same systems as for the workshop diagnosis, it was difficult to achieve the same level of fault isolation in the remote scenario due to not having enough analytical redundancy. On the other hand, in the remote diagnosis scenario data is available at higher frequency making it possible to train residual detecting deviations in behaviors with high dynamics. With the federated learning setup, data is also more readily available making it possible to train models for individual vehicles rather having one model for a larger population. In the case studies, data from several systems were studied, the SCR system, the propulsion battery, and the fuel injection system. With regard to the ability to industrialize the project results, it was also studied how a federated learning framework can be deployed in an automated way and how to adapt models so that they are best trained on an edge device.

6.5 Contribution to EMK's Program Area

The program area within Electronics, Software, and Communication where this project contributed the most is the subarea called Intelligent and Reliable Systems, specifically the cornerstone of Machine Learning.

The new technology developed in this project contributes to advancing data-driven product and service development for more efficient diagnosis of complex vehicle systems. The project results also improve the development of diagnostic functionality by systematically utilizing available data from different driving scenarios and mathematical models for data-driven modeling of complex dynamic systems for fault diagnosis. The project has shown how data-driven models can be used to reason about abnormal system behavior which is also important in safety critical applications to understand under which operating conditions a model can be trusted.

The developed methods also contribute to more efficient data collection by identifying data scenarios that can enrich the existing datasets. Overall, the automated design of neural networks and efficient training speed up industrialization of data-driven models in automotive applications by supporting engineers in the diagnosis system design process.

Today's workshop diagnostics rely heavily on manually developed troubleshooting information. By combining physical models with machine learning algorithms trained on vehicle data, much of this process can be automated, enabling individualized functionality for each unique vehicle with its specific characteristics.

A correct fault diagnosis at an early stage is essential for predicting maintenance needs in various components and subsystems. Through the use case of remote diagnostics, the project introduces new possibilities for monitoring vehicle health status by leveraging time-series data. For remote diagnostics, the project has explored complex vehicle functions and system-of-systems approaches by adapting algorithms for distributed machine learning.

6.6 Contribution to FFI's Overall Goals

The project has successfully contributed to FFI’s overall goal of increasing Sweden’s research and innovation capacity by evaluating the latest academic research on real-world problems relevant to the automotive industry. While machine learning has shown promising results in many computer applications, automotive applications are mechatronic systems that should be predictable and reliable even though training data is scarce. The project has successfully bridged model-based diagnosis theory with machine learning and showed its applicability using real industrial case studies. This also helped strengthen the competence of Swedish industry and academia in machine learning to maintain global competitiveness.

Research and innovation capacity in Sweden has also increased through the project’s education of a PhD student and a workshop with Scania engineers. The project has promoted collaboration between industry and academia, as both are partners in the project has been working towards a common research question. Academia gained access to data and support for conducting case studies from the industry, while the industry benefits from the latest academic research. The access to realistic data from relevant case studies have helped to identify questions to be investigated that are academically relevant but also industrially relevant.

7. Dissemination and publications

7.1 Dissemination

How are the project results planned to be used and disseminated?	Mark with X	Comment
Increase knowledge in the field	X	The results during the project have been spread within the project group through workshops and externally through a large number of scientific publications.
Be passed on to other advanced technological development projects	X	The methods researched are ready for further evaluation in internal predevelopment projects at the industry part.
Be passed on to product development projects	X	The infrastructure developed is ready for direct application in further development projects at the industry part.
Introduced on the market		
Used in investigations / regulatory / licensing / political decisions		

7.2 Publications

Within DELPHI, one journal article, eight conference papers, two technical reports, and eight master's thesis projects have been published. Additionally, paper [9] *"A Study on Redundancy and Intrinsic Dimension for Data-Driven Fault Diagnosis"* was awarded Best Conference Paper at *the 35th International Conference on Principles of Diagnosis and Resilient Systems*, 2024.

Journal papers

[1] **Consistency-based diagnosis using data-driven residuals and limited training data**, A Mohammadi, D Jung, M Krysander, *Control Engineering Practice, 2025*

Conference papers

[2] **Fault Diagnosis of Exhaust Gas Treatment System Combining Physical Insights and Neural Networks**, D Jung, B Kleman, H Lindgren, H Warnquist, *IFAC Advances in Automotive Control*, 2022

[3] **Analysis of grey-box neural network-based residuals for consistency-based fault diagnosis**, A Mohammadi, M Krysander, D Jung, *IFAC Safe Process*, 2022

[4] **Analysis of Numerical Integration in RNN-Based Residuals for Fault Diagnosis of Dynamic Systems**, A Mohammadi, T Westny, D Jung, M Krysander, *IFAC World Congress*, 2023

[5] **Fuel injection fault diagnosis using structural analysis and data-driven residuals**, N Allansson, A Mohammadi, D Jung, *IFAC Safe Process*, 2024

[6] **Stability-Informed Initialization of Neural Ordinary Differential Equations**, T Westny, A Mohammadi, D Jung, E Frisk, *ICML*, 2024

[7] **Fault diagnosis using data-driven residuals for anomaly classification with incomplete training data**, D Jung, M Krysander, A Mohammadi, *IFAC World Congress*, 2023

[8] **Assumption-based design of hybrid diagnosis systems: analyzing model-based and data-driven principles**, D Jung, M Krysander, *PHM*, 2024

[9] **A study on redundancy and intrinsic dimension for data-driven fault diagnosis,** D Jung, D Axelsson, *DX*, 2024

Master's thesis reports

[10] **Evaluation of model-based fault diagnosis combining physical insights and neural networks applied to an exhaust gas treatment system case study**, B Kleman och H Lindgren, 2021

[11] **Data-Driven Diagnosis For Fuel Injectors Of Diesel Engines In Heavy-Duty Trucks**, F Eriksson och E Björkkvist, 2024

[12] **Fault Detection of Internal Combustion Engine: Exploring Dynamic Relations with SINDy and AR Models for Engine Sensor Fault Detection**, M Sadeghi Naeini, 2024

[13] **Battery Degradation and Health Monitoring in Lithium-Ion Batteries: An Evaluation of Parameterization and Sensor Fusion Strategies**, S Saber, 2024

[14] **Machine Learning-Based Prediction of Diagnostic Trouble Codes in Electric Vehicle Batteries: A Multi-Temporal Analysis**, A O Abubaker, 2025

[15] **Contrastive Representation Learning for Engine Fault Diagnosis - Learning Time-Series Feature Representations to Improve Open-Set Fault Classification**, M Collard, 2025

[16] **Training Neural Networks on Embedded Devices**, P T Shaji, 2023

[17] **Online and Federated Learning for the heavy vehicle industry**, S Vijayvergiya, 2025

Technical reports

[18] **The LiU-ICE Benchmark -- An Industrial Fault Diagnosis Case Study**, D Jung, E Frisk, M Krysander, arXiv:2408.13269, 2024

[19] **AI-based remote diagnosis for heavy vehicles – Technical Report**, E Björkkvist, I Ederlöv, M Karlsson, A Levin, M Nibell, E Wigström, A Öberg, Student project report, https://tsrt10.gitlab-pages.liu.se/2023/scania/, 2023

8. Conclusions and future research

Data-driven residuals have been shown to be advantageous with respect to other anomaly classifiers because they can reason about faults without faulty data. This project has shown that the data-driven residuals, where the network structure is derived from a structural model, can be used to isolate unknown faults. The main task of feature selection is to identify a subset of input signals with sufficient information to predict the target signals. However, depending on e.g. excitation in data there is a risk that some relevant signals are missed which will result in model inaccuracies. A structural model can help identify different sensor combinations based on physical insights which is useful for model explainability and interpretability. However, since many data-driven models, such as neural networks, do not generalize well, the quality of training data is an important factor that affects the false alarm rate. One way to reduce the need for training data is to find model candidates that use few signals to improve generalizability which is supported by using the structural model.

A test validity measure is proposed to detect when test data deviates from training data to avoid false alarms caused by out-of-distribution data. Experiments show that the false alarm rate can be significantly reduced during nominal operation and decoupled fault scenarios. The test validity region is important when applying data driven residual generation, but the developed techniques are also relevant in a model-based framework to handle model inaccuracies and generalization issues.

Another important result is the evaluation of numerical solvers and the effect on the trained model. The proposed neural ODE model parameter initialization method shows that there is lots of potential in improving training of neural networks and model

interpretability by combining classical modeling and simulation with machine learning methods.

Ensemble-based models show promising results to model prediction uncertainties and detect out-of-distribution data. Thus, they are capable of handling both aleatoric and epistemic uncertainties. However, the higher computational cost of ensemble models makes them more suitable for remote applications than onboard if onboard computational capacity is limited.

This research project has bridged model-based diagnosis with data-driven fault diagnosis. Even though data-driven diagnosis is often treated as a standard classification problem, it is shown that theory from model-based diagnosis, e.g. redundancy, can be used in a data-driven context. When representative training data is limited, there is an increased risk of misclassifications caused by out-of-distribution data. This is even more relevant for structured residuals since they should not false alarm when a fault in another part of the system changes the operating conditions. Thus, theory and methods for data-driven fault diagnosis are needed.

As future work it would be relevant to investigate how the proposed data-driven residuals can be used on a vehicle fleet level where different vehicles have slightly different behavior. This means that residual models should adapt to each vehicle to maximize detection performance. The models should update or retrain as new training data becomes available but also share information between vehicles to improve residual model generalizability. The success in combining methods from physical modeling and simulation, e.g. when initializing neural network parameters, has shown great potential in adopting techniques from classical model simulation to neural ODEs. Another interesting research direction is to continue the work on combining model-based diagnosis theory and data-driven fault diagnosis to develop new machine learning methods that can reason about abnormal system behavior when a structural model is not available. This could also be important for using machine learning in, e.g., safety-critical systems to determine when under which operating conditions a model is reliable.

9. Participating parties and contact persons

Scania

Contact person: Håkan Warnquist

hakan.warnquist@scania.se

Linköping University

Contact person: Daniel Jung

daniel.jung@liu.se

10. References

[20] **Introduction to diagnosis and fault-tolerant control**. Blanke, M., Kinnaert, M., Lunze, J., Staroswiecki, M., Blanke, M., Kinnaert, M., and Staroswiecki, M. Springer Berlin Heidelberg. 2016.

[21] **Analysis of fault isolation assumptions when comparing model-based design approaches of diagnosis systems.** Jung, D., Khorasgani, H., Frisk, E., Krysander, M., & Biswas, G. (2015). *IFAC-PapersOnLine*, *48*(21), 1289-1296.

[22] **Isolation and Localization of Unknown Faults Using Neural Network-Based Residuals.** Jung, D. *Annual Conference of the PHM Society*. 2019.

[23] **State space neural networks and model-decomposition methods for fault diagnosis of complex industrial systems.** Pulido, B., Zamarreño, J. M., Merino, A., & Bregon, A. (2019). *Engineering Applications of Artificial Intelligence*, *79*, 67-86.

[24] **Evaluation of model-based fault diagnosis combining physical insights and neural networks applied to an exhaust gas treatment system case study.** B Kleman, H Lindgren, Master's thesis, LiTH-ISY-EX--21/5415—SE*, 2021*