

Vision and Outlook on High-Tech Equipment Diagnostics

TNO 2024 R12778 – 20 december 2024

Vision and Outlook on High-Tech Equipment Diagnostics

Auteurs	Leonardo Barbini, Thomas Nägele, Alvaro Piedrafita
Rubricering rapport	TNO Public
Titel	TNO Public
Rapporttekst	TNO Public
Aantal pagina's	24 (excl. voor- en achterblad)
Aantal bijlagen	0
Projectnaam	KIP Vision on HT Equipment Diagnostics 2024
Projectnummer	060.62772

Alle rechten voorbehouden

Niets uit deze uitgave mag worden verveelvoudigd en/of openbaar gemaakt door middel van druk, fotokopie, microfilm of op welke andere wijze dan ook zonder voorafgaande schriftelijke toestemming van TNO.

© 2025 TNO

Inhoudsopgave

1	Introduction.....	4
2	The high-tech equipment diagnostic challenges and needs	6
2.1	Diagnostics, a domain of many facets	6
2.2	Industry needs	9
3	State of the art	13
4	State of practice.....	15
5	TNO-ESI diagnostic proposition.....	17
5.1	Diagnostic vision	17
5.2	Past activities.....	19
6	Conclusions	21
7	References.....	22

1 Introduction

TNO-ESI has established five Product Market Combination (PMC) teams to maximize the impact of embedding world-leading methodologies in the high-tech equipment industry. These PMC teams are tasked with identifying socio-technical trends pertinent to the high-tech sector. From these trends, PMCs will extrapolate both current and anticipated needs facing the industry. Ultimately, they will formulate a strategic roadmap to address these needs through collaborative efforts with ESI. The five PMCs are:

- PMC team Engineering for System Qualities
- PMC team Engineering for Lifecycle Excellence
- PMC team Engineering for System Evolution/Diversity
- PMC team Engineering for Systems of Systems Integration
- PMC team Systems Architecting/Systems Engineering.

This document is part of the work conducted by the PMC Engineering for Lifecycle Excellence (LCE), which refers to the comprehensive management and optimization of a system throughout its entire lifecycle. From the initial concept and design, via its usage and integration with other systems, through to decommissioning and disposal. Within the work of the PMC on LCE four key areas of interest have been identified as relevant for the high-tech sector: diagnostics, testing, security and sustainability. This document covers the TNO-ESI vision related to diagnostics. For the other key areas and the overarching PMC LCE, a vision will be provided in separate documents.

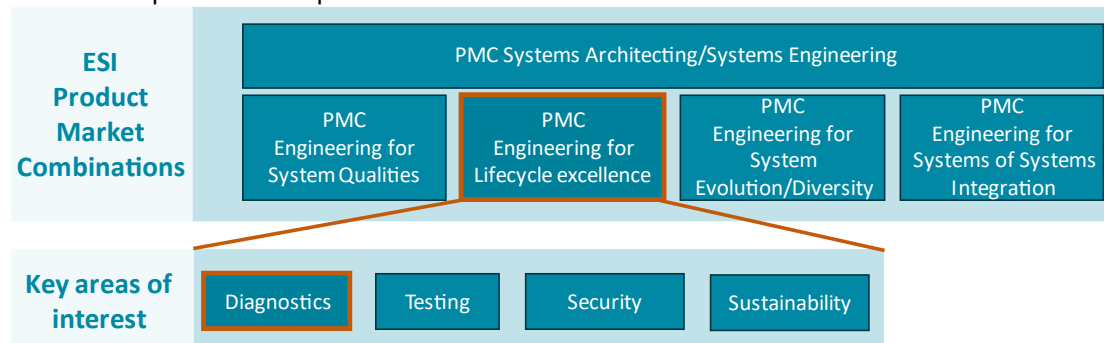


Figure 1-1 Diagnostics as key area of interest within TNO-ESI Product Market Combinations (PMCs).

The overarching PMC Systems Architecting/Systems Engineering has identified, with particular emphasis on the Dutch sector, eight key socio-technical trends for the high-tech industry. The identified trends are:

1. Increasing dependence on complex systems
2. Growing system diversity
3. Growing role of Artificial Intelligence (AI)
4. Continuous innovation and updating of long-life systems
5. Climbing the value chain
6. Innovation in ecosystems
7. Growing scarcity of engineering experts
8. Growing emphasis on Sustainability.

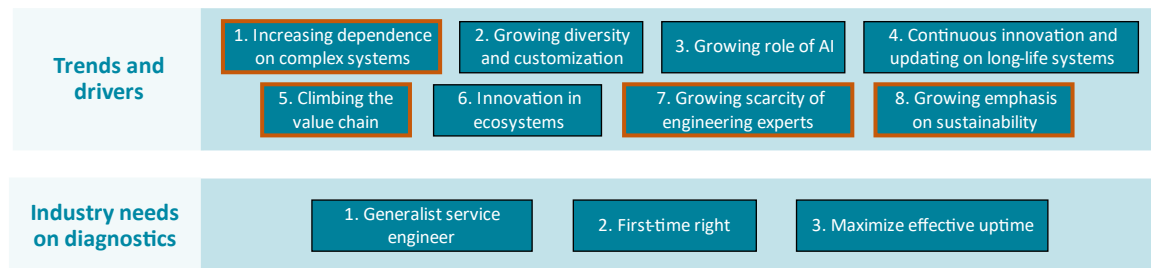


Figure 1-2 Key trends and drivers with corresponding industry needs in diagnostics.

For their detailed description, we refer the reader to [1]. Although all eight trends to some extent have an impact on all PMCs, some trends are most impactful in specific PMCs. For the diagnostics area within the PMC engineering for LCE, we identify the four key trends 1, 5, 7 and 8.

Diagnostic specific trends and corresponding industry needs

The rationale for the choice of these trends, together with the related industry needs is given below.

Increasing dependence of society and economy on complex high-tech systems, Trend 1, directly implies that their Overall System Effectiveness (OSE) must be maximized. Here with OSE we mean the system's ability to meet operational requirements, maintain reliability, and achieve desired outcomes while minimizing costs and downtime throughout its lifecycle. However, the increasing complexity of these systems together with the dynamic environment in which they operate makes maximization of OSE more and more challenging. Many factors contribute to the overall effectiveness of high-tech systems, among which diagnostics. Here, diagnostics means a set of approaches to promptly facilitate the resolution of any issue in the operation of these system. Ultimately, such approaches shall also prevent such issues from occurring, i.e. losses of overall effectiveness, by defining proactive actions. We refer to this need as to **maximize effective uptime**.

Increasing scarcity of engineering experts, Trend 7, is limiting the high-tech industry's traditional escalation-based approach to solving complex diagnostic problems. This combines with a tendency to have unique systems system (related to Trend 2), i.e. a wide range of different system configuration that need specific service. Not only does it take too much time to equip service engineers with the required knowledge, but even worse, it is problematic to find enough engineers to match the service needs of the installed base of systems. Given this trend there is a need for a **generalist service engineer**: a service engineer who does not need to be an expert in a specific domain to solve complex escalations in this domain.

Climbing the value chain, Trend 5, manifests as a shift from selling high-tech systems, such as lithography scanners, printers, and other complex equipment, to selling services, e.g. wafers per hour or prints per day. This shift in the business proposition has a direct effect on diagnostics. It introduces a strong incentive for manufacturers to build know-how to diagnose and repair systems as quickly and efficiently as possible, without wasting time, resources, or energy. This trend, coupled with the **growing emphasis on sustainability**, Trend 8, impacting many companies, results in a need for **first-time-right diagnostics**.

2 The high-tech equipment diagnostic challenges and needs

The three needs identified in Section 1 are based on observations and feedback from the TNO-ESI high-tech ecosystem. In this section, each of these needs will be presented in more detail together with a description of the challenges specific to the high-tech equipment diagnostic domain.

2.1 Diagnostics, a domain of many facets

The high-tech diagnostics domain is one of many facets. The complexity of cyber-physical systems manifests in a wide variety of failure behaviours, ranging from software (SW), hardware (HW) and problems between SW and HW. On top of this, these systems are usually deployed in demanding, complex and heavily varying environments, ranging from complex factories to interactions with human operators using the system in very different ways.

Furthermore, diagnostics inherently requires a systems-thinking approach as issues arise from the interplay of various components. For instance, the image quality of an industrial printer results from the interplay of many factors, including the life-of-ink², life-of-paper, life-of-droplets and the operation of the printer's components involved in their integration. Analogously, in lithography, the product quality is the results of the life-of-light, the life-of-wafer, and complex interactions between many other processes, some in control of OEM supplier and some out of control.

The choice of abstraction level for diagnostic reasoning directly impacts the scope of potential diagnoses. A higher level of abstraction focusing on system-wide behaviours, e.g. on-product key performance indicators, may be insufficient to diagnose component-level issues. Conversely, a lower level of abstraction, considering component details, will miss system-level anomalies arising from the interactions between components. Finding the optimal trade-off between different abstraction levels is a common systems thinking problem, particularly relevant for diagnostics.

All these factors together make the high-tech diagnostics domain a challenging one. This section addresses some of the most relevant complicating factors in the domain.

²With life-of-X we refer to all the factors that contribute to the handling and processing of X, throughout its journey in an industrial system

Hardware, software and system: the nature of the system

Modern high-tech systems are great examples of complex cyber-physical systems (CPS). These systems integrate computational (cyber) and physical components to achieve specific goals. The cyber component, a sophisticated software stack, governs the physical component to fulfil these goals. System goals are typically defined as performance metrics, such as accuracy or throughput, quantifying desired system behaviour. The realization of a CPS involves the synergistic interaction of hardware and software components to meet these predefined goals. Recently, the increasing integration of Fault Detection, Identification, and Recovery (FDIR) techniques into systems has made diagnostics more complex. This is due to the added layers of adaptive behaviour and potential failure masking introduced by these techniques [2].

From a diagnostic point of view, an issue can occur at any place in the CPS. That holds for both the cyber and physical components. At the same time, failures often lead to the system failing to achieve its goal(s), as described above. A description of the issue would therefore often be at goal-level – ‘the system does not do X sufficiently well’ – while the real problem is at a much lower level, either in the hardware or the software domains, as these make the system.

On top of that, most of what the system reports to the service engineer is created by software. Due to some hardware problem, the software may detect that it is unable to properly control the system, after which it reports an error and may just stop its execution. On the other hand, software may also fail in delivering the correct messages and may obfuscate failure behaviour because of that, complicating the troubleshooting a service engineer has to do. Furthermore, software may also have problems, requiring proper testing before its release. This latter topic connects to the *testing*, a key area of interest of PMC2 as shown in Figure 1-1. A separate document will describe the challenges and needs on the future of testing for high-tech systems.

Overall, the hardware-software interplay of high-tech systems can be both a gift and a curse, depending on where the real problem is and how well the entire system can cope and report based on such a failure.

Context, system and system-of-systems: the nature of the system’s environment

High-tech systems often represent a complex integration of various subsystems, some of which may be designed and manufactured by the same entity responsible for the overall system, while others may originate from suppliers. This heterogeneous composition makes it challenging to ensure sufficient transfer of diagnostic information throughout the supply chain. This results in limited diagnostic knowledge, at the high-tech system level, for the subsystems from suppliers. Figure 2-1 (bottom left) shows the high-tech system of interest in light blue. It contains other systems, respectively in light and dark blue for the system designed and manufactured in-house and from a supplier.

Furthermore, the deployment of these high-tech systems within a factory necessitates their interaction with numerous other (high-tech) systems. Often, diagnostic knowledge of these other systems is limited. This is not primarily a necessity for diagnosis problems within each system, but becomes relevant when a customer has a problem on the product itself. As a result, this hinders the

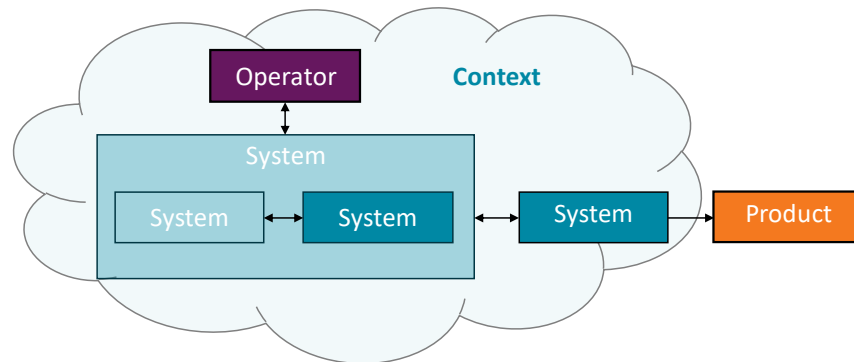


Figure 2-1 An high-tech system in its environment. What is the cause of a problem on a product?

availability of a comprehensive diagnostics knowledge at factory level. Figure 2-1 shows this challenge by the interaction of the high-tech systems in light and dark blue.

Moreover, these high-tech systems operate within physical environments, which influence their intended functionality; and the interactions with human operators, who configure and control these systems, can further introduce complexities and potential sources of unintended behaviour. Figure 2-1 shows the context in which the system is deployed, together with the system's operator, with the grey and purple cloud and box, respectively.

Reactive, performance and predictive: the nature of the issues to diagnose

Diagnostics in the high-tech systems domain is a complex task with multiple dimensions. One crucial dimension is the type of diagnostics to perform. While all types of diagnostics attempt to maximize the system's effective uptime, they target different diagnostic challenges. For example, reactive diagnostics involves responding to a problem after it has already occurred, while predictive diagnostics aims to anticipate potential issues before they manifest. By monitoring system behaviour and identifying potential anomalies, it is possible to take corrective actions proactively. This can involve updating software, replacing hardware components, or adjusting system configurations or coefficient to avoid/compensate failures and maintain optimal performance. The biggest benefit of being able to predict is that maintenance can be scheduled, which also allows the production line to anticipate for the interruption.

Another dimension to consider is the severity of the problem, which ranges from hard-down situations to performance issues. Hard-down situations, where a system completely fails to function, are often deemed easier to diagnose due to the clear absence of output. However, for high-tech systems, this task is still cumbersome due to the system's complexity. Systems that operate sub-optimally are extremely hard to diagnose. Fault isolation, or root cause analysis, is a key aspect of diagnostics. This involves tracing a problem back to its origin to actively fix it and restore functionality. In the case of system down situations, fault isolation can help identify faulty components or software errors. For performance degradation, understanding the root causes can lead to adaptations that restore system functionality. **These dimensions of the diagnostic domain are independent and combine in various ways.** For instance, a system could experience a hard-down due to a software bug, which would require a reactive diagnosis. On the other hand, the same system might exhibit performance degradation due to hardware aging, which could be predicted through monitoring and proactive maintenance. In both cases, the severity and root cause are independent factors that can be addressed through different diagnostic approaches.

Therefore, for a comprehensive diagnostic approach, one may need to address all possible combinations of severity, root cause, and diagnostic timing (reactive or predictive).

2.2 Industry needs

Generalist service engineer

Currently, customer service departments employ many service engineers with very specific areas of expertise. This may be limited to a certain subsystem or a specific domain. Due to the scarcity of experts and sometimes a lack of in-depth understanding of the system's behavior, especially for newly introduced systems, relying solely on this group of experts becomes infeasible. There is the need for a new class of service engineers who, while not deeply specialized in any particular domain, can efficiently troubleshoot a broad range of complex system issues, i.e. generalists. This will reduce reliance on a limited number of highly specialized engineering experts, as well as enlarge the pool of potential candidates to hire.

In order to maintain troubleshooting efficiency, generalist service engineers need to be equipped with **timely access to relevant information, tailored for the specific issue being addressed**. We need to avoid an information overload towards the service engineer, to assure that he or she can focus on understanding what is going on in the system. By better scoping information presented to the service engineer, less experience is required from the service engineer to skim through bulky information to narrow it to the relevant parts. We envision that a digital assistant will become indispensable for service engineers. By providing rapid access to system design details, historical data, and analytics, this assistant will enhance the efficiency of root cause analysis. Furthermore, computer-aided troubleshooting powered by this digital assistant will enable engineers to identify and resolve issues more effectively, even with limited expertise. In cases where a better understanding of the system has been gained, this feedback should again be provided to the digital assistant to improve future diagnostics.

Figure 2-2 shows the current approach for diagnostic knowledge build-up in many organizations. A subset of the design knowledge is transferred from Research and Development (R&D) to the customer Support (CS) department, depicted as a blue full circle, e.g. by conducting Failure Mode Effect Analysis (FMEA) and writing Service Manuals (SM). Of this subset of knowledge, only a small portion is actually used by CS, depicted as the intersection between the full blue circle and the full purple circle. Over time, by the cumulative learning process driven by solving real diagnostic cases (incidents), CS builds up troubleshooting knowledge. This is depicted with a full purple circle. Clearly, this process is inefficient, requires CS expertise, and does not allow for a generalist service engineer.

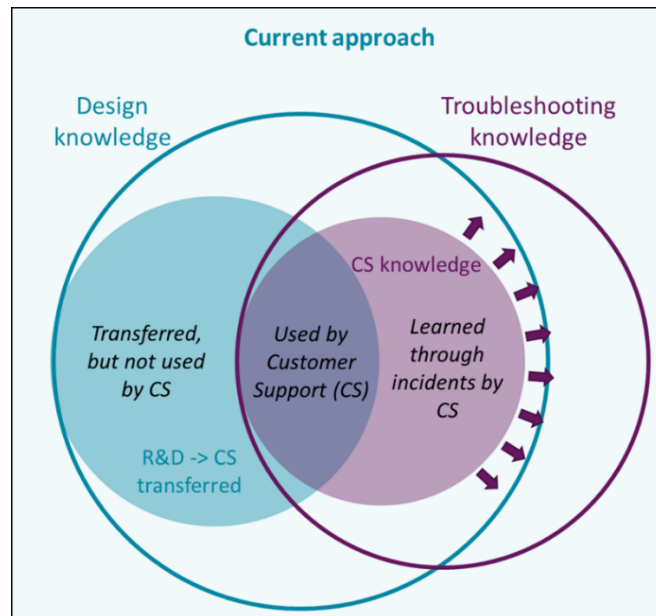


Figure 2-2 The current approach of building up troubleshooting knowledge starts with transferred information from the R&D department, as shown on the left. A part of all the available design knowledge is transferred to the customer support (CS) department by means of manuals or other types of documents. Over the lifetime of the system, only part of this transferred information is used. CS also learns how to troubleshoot efficiently through incidents that are solved in the field, slowly increasing the size of the filled purple circle on the right.

First-time right

The need for first-time right has multiple interpretations, but they all strive for reducing resources and time spent on resolving a system malfunction. First-time right is about being able to solve the malfunction on the first service visit, and about avoiding the unnecessary replacement of components that are not broken. Industry should stop following diagnostic trial-and-error approaches, in which parts are swapped without need.

The main challenge to achieve first-time right is to **have meaningful and sufficient data gathering from the system**: one should only collect system data that is meaningful for the type of diagnostics that should be performed. This could include both automated data gathering, such as logging, and dynamic optimal testing procedures by a service engineer, where dynamic means tailored to the specific diagnostic issue under investigation. This will enable the service engineer to efficiently narrow down the search space for the issue and to make well-thought decisions on how to proceed. However, different types of diagnostic (strategies) may require different data to be collected: predictive diagnostics or prognosis depends mostly on automatically acquired data, while reactive diagnostics is less strict.

The necessity and approaches for achieving first-time right vary between failures that occur rarely and those that occur frequently. Failures typically follow a Pareto like distribution, where 20% of the failure types account for 80% of the occurrences [3]. Figure 3-3 illustrates a schematic representation of this distribution, in purple and blue the types of failures occurring frequently and rarely, respectively.

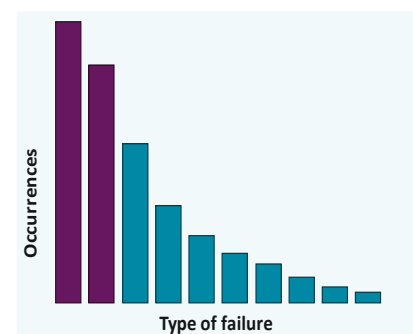


Figure 2-3 Pareto distribution of failure types, sorted by number of occurrences

A significant portion of diagnostic knowledge is built around frequent occurrences, which can be effectively addressed using statistical methods, such as pattern matching and generally data-driven approaches. To achieve first-time right for these types of failures, organizations should learn from field feedback quickly, by employing methods to utilize learnings from feedback in a structured way.

Conversely, little diagnostic knowledge is built around the more rare issues, which cause long downtimes even if there is a structured method of learning from feedback in place. To achieve first-time right for these types of failures, digital reasoning systems are needed, which gather and reason with the most relevant system information on the system given the current data.

Maximize effective uptime

The need to maximize effective uptime can be addressed by a wide range of different approaches. Maximizing effective uptime is about minimizing downtime and underperformance. There are two angles to this goal. First, the time required to resolve an issue, i.e. issue resolution time, could be reduced, so that the underperformance or non-performing situation lasts as shortly as possible. Second, the frequency of having an underperforming or non-performing system could be minimized.

Issue resolution time is influenced by various factors, including travelling time, diagnostic time, repair time, logistics, and reporting time. This document primarily focuses on diagnostic time, which involves identifying the necessary corrective action, however it's essential to acknowledge that reducing the overall frequency of interruptions requires a multifaceted approach. Historically, this has been addressed by enhancing system reliability through improved robustness or via redundancy. In the context of this document, we identify the need to mitigate the negative impact of unexpected system failures by implementing predictive maintenance and proactive measures.

Some of the main challenges for reducing the issue resolution time have already been addressed in the other two industry needs: access to the relevant information and having meaningful and sufficient data gathering in place. The challenge to avoid unexpected system interruptions is to **properly assess the current and future** system's effectiveness **and suggest meaningful countermeasures**, based on the collected relevant system data interpreted in a useful way.

Predictive diagnostic tasks focus on predicting future system failures and improving maintenance efficiency rather than diagnosing issues that have already occurred. This temporal dimension of the prediction task requires inferring and tracking the system's performance over time. To address these unique aspects, predictive diagnostics often needs continuous variables to represent varying degrees of the system health, ranging from fully operational to completely broken. This contrasts with traditional diagnostic methods that typically rely on Boolean or multi-valued variables. Such a shift from discrete to continuous variables significantly increases the complexity of the problem. In addition, the temporal aspect of prediction adds another dimension to it.

Figure 2-4 shows a schematic representation of the complexity build-up through different types of diagnostic approaches, all contributing to increased effective uptime of a system. Rather elementary is the situation in which the system just stops and does not work anymore and needs to be diagnosed. While the diagnosis itself may be challenging, detecting a problem is usually not in such a situation. Second is the situation in which the system's performance is inadequate. In contrast to a system down, it is much harder to both recognize as well as to diagnose such a situation, due to many contributors, which all



Figure 2-4 Different types of diagnostic capabilities, ordered from less complex to more complex from left to right.

combined result in the issue manifestation. Finally, prognosis not only includes diagnosing the present state, but it also comprises prognosing the future state of the system based on the observations.

3 State of the art

There are three academic fields that attempt to address the previously introduced industry needs and challenges: Prognostics and Health Management (PHM), Reliability, Availability, Maintainability and Safety (RAMS), and Fault Detection and Isolation (FDI), each one with a somewhat different focus. Diagnostics research sits in the intersection of all three. It is helpful for a discussion of the state of the art of the field to analyse it from two different angles: what challenges is each field working on, and how are they being approached.

The *What* in PHM, FDI and RAMS

All three fields aim at minimizing risk and Total Cost of Ownership (TCO), each one focusing on different approaches to achieve that.

In the RAMS field, the focus is on characterizing, already at design time, the reliability, availability, maintainability and safety of a system, and determining ways to change said design to improve its performance in these key aspects. The diagnostic challenge in this field is to assess the observability of failures in a given design. However, the focus in RAMS goes beyond fault identification and into assessing fault incidence rates, modelling fault impacts on system-level reliability, and development of mitigation and maintenance strategies, such as introduction of redundancy or modular design.

In the FDI field, the focus is on the identification of abnormal behaviour (anomaly detection) and failure modes in operational conditions. FDI is specifically concerned with the impact of failures on controlled systems, forecasting their future behaviour, and designing robust control systems that can maintain performance despite the presence of faults. FDI is a subfield of control engineering and is heavily influenced by control theory and signal processing.

Prognostics and Health Management (PHM) is an interdisciplinary field focused on predicting the future health and performance of systems to optimize maintenance and ensure reliability. The emphasis in PHM rests on prognostic forecasting, for which the concept of Remaining Useful Life (RUL) is commonly used, although anomaly detection and, to a smaller degree, fault identification are also studied.

In all three fields, most research remains concentrated at the component or subsystem level [4], [5] (e.g., ball bearings, rotors, gears, pumps, sensors). Further, as said above, with different focus for each field: design for diagnostics in RAMS, reactive diagnostics in FDI, and prognosis in PHM.

At the system level, research in the PHM and FDI domains is geared towards predicting remaining useful life and anomaly detection, leveraging deep learning's strength in handling many-to-one, non-linear regression problems. System-level assessments in the RAMS field are more common and considerable effort is spent on assessing overall reliability and availability of systems using knowledge of the same concepts related to their components.

The *How* in PHM, FDI and RAMS

Given the current trends in AI research, most efforts are nowadays directed towards applying deep learning to PHM, FDI and RAMS problems.

Most AI-based approaches for diagnostics use convolutional neural networks, Long Short-Term Memory neural networks (LSTMs) and/or autoencoders [6] combined with historical data of machines, such as sensor readings or time-to-failure [7] [8] [9]. Unsupervised

learning is used for classification and clustering in fault identification, while supervised learning is more common for RUL estimation and anomaly detection.

These deep learning approaches have shown improved performance over traditional techniques based on signal processing and feature extraction [10]. Moreover, little domain expertise is required to make these models. Nevertheless, deep learning approaches still face several limitations common to industrial AI applications [7] [9]:

- **data rigidity:** AI-based solutions require complete and consistently shaped data for training and usage. The necessity of existing training data limits their application at product introduction and for rare events.
- **poor extrapolation:** AI solutions trained in specific configurations may not perform well under changing conditions, which is particularly challenging in high-tech domains with significant variations in equipment configuration and usage.
- **domain knowledge agnosticism:** incorporating domain knowledge into deep learning methods is often difficult, if not impossible.
- **explainability:** the complexity of deep learning models can make them difficult to interpret and explain.

Hybrid data- and knowledge-driven approaches are just beginning to appear. Where detailed models exist in the form of digital twins, the models have been used to simulate failures and generate synthetic data on which Machine Learning (ML) models can be trained to do fault and anomaly detection through clustering and classification [9], or prediction [4] [7].

Alternatively, purely data-driven approaches have been augmented with physics or system knowledge to achieve higher accuracy [11] [12] [13] [14] [15]. There is work using physics informed neural network architectures (PINNs) and/or graph neural networks (GNNs) [16], which can capture physics and structure much better than standard deep learning architectures. These tend to outperform the purely data-driven or knowledge-driven approaches at the cost of added complexity.

A smaller portion of research is devoted to model-based techniques. Model-based diagnostics requires a higher degree of domain expertise than data-driven or hybrid approaches. The flipside is that model-based techniques do not have any requirements on the availability of field data, and can be applied at design time, when field data is non-existent, in rare cases where not enough data exist to train data-driven models, or when field data is incomplete or of varying size.

At component level, model-based research is most prevalent in industries where very detailed models exist (energy [17], aerospace [18]) and/or high diagnosability on product introduction is a necessity (nuclear [19], aerospace). With detailed models at hand, fault identification, as well as state estimation and prognostics is to some degree possible.

At system level, most research focuses on root-cause analysis using graphical models, although methodologies for system-level prognostics based on component-level prognostic models are beginning to emerge [20], [21], [22], [23].

4 State of practice

The reality of the state of practice of diagnostics in industry is complex and fragmented. Companies across all industrial domains use different ad-hoc combinations of expert, data and knowledge-driven diagnostic approaches with different scopes and goals. A useful way of making sense of the state of practice in diagnostics is to view it from two orthogonal directions. One point of view considers the scope of a diagnostic solution, with component-level diagnostics on one side, system-level diagnostics on the other, and subsystem-level diagnostics somewhere in the middle. Another way of looking at it is through the approach taken, which comes roughly in three flavours: data-driven, knowledge-driven, and expert driven.

At component level, data-driven diagnosis and prognosis are primarily used for the most critical components or the most common failures. These approaches rely on standard machine learning and artificial intelligence tools, which can sometimes be developed in-house, although much off-the-shelf tooling also exists [24].

Model-based diagnosis and prognosis at component level is typically found only in industries with very high reliability and safety requirements, such as aerospace, defence, or nuclear sectors. For components that are not critical, model-based prognostic methods are generally not applied. Diagnostics, if conducted at all, are usually based on expert-driven documents, like Failure Modes and Effects Analysis (FMEA), Failure Modes, Effects, and Criticality Analysis (FMECA), and static diagnostic procedures. Alternatively, diagnostics may be left to the maintenance technicians.

At system level, the most common approach to diagnostics involves the creation of expert-driven FMEA documents for diagnosability analysis and static diagnostic procedures to guide technicians performing diagnostics in the field. When these methods fail, companies often escalate issues to experts. Data-driven models are used at the subsystem level for the most common failures, provided there is sufficient data and the failure signatures are uniquely identifiable. Integration of component and subsystem diagnostics into dashboards is also practiced. Model-based diagnostics at the system level, similar to the component level, are typically carried out only by industries that are required to do so, such as aerospace, defence, and nuclear power sectors.

Tooling view

In the course of the diagnostics projects undertaken by TNO-ESI in the last decade, the diagnostics expertise team has reviewed a number of commercial tools for model-based diagnostics, see Table 4.1. The criteria for a tool to be considered were that it must offer a scalable modelling approach, be already successfully applied in an industrial context, and offer, to some degree, both design and operational diagnostics support.

Furthermore, the selected tools all implement a model-based diagnostic approach. Several tools are nowadays being introduced in the diagnostic landscape that use a more data-driven approach, compiling a library of data-driven models at component level. As it is not yet clear how to use these component-level data-driven models for system-level diagnostics, which is TNO-ESI's focus, these tools have not yet been investigated. It is suggested that such tools are investigated in follow-up projects, in which data-driven and knowledge-driven approaches may strengthen each other.

The tools that were identified as most promising were MADe and ISDD Tool Suite, see Table 4.1. A full evaluation of MADe in the high-tech industry context was conducted as part of a

project carried at TNO-ESI and the conclusion was that the MADe tool does not yet offer enough diagnostic support to be adopted in that industry. A preliminary investigation of ISDD Tool Suite suggests that this tool might be a better fit for the high-tech manufacturing industry, especially when applied to design for diagnostics and model-based system engineering models are available. It is suggested that this tool is investigated in future projects.

Overall, our conclusion is that, for these tools, modelling complexity is high to very high, but using the tools can sometimes be worth the effort since it can replace manual work that is also very costly and error-prone, e.g. FMEA and creation of diagnostic procedures. Furthermore, it can save time in diagnosing a system by avoiding inefficient trial-and-error maintenance or expensive escalations. Additionally, tool maturity is higher for reliability analysis than for diagnostics, and adoption in industries not used to work model-based is low. Therefore, we conclude that there is not a fit-for-all off-the-shelf diagnostic tool and further research is needed to bring system-level diagnostics to the high-tech industry

Name	Source
MADe	www.phmtechnology.com [25]
TEAMS	www.teamqsi.com [26]
Kairos Workbench	www.kairostechno.com [27]
RODON	www.combitech.com [28]
ISDD Tool Suite	www.dsiintl.com [29]

Table 4.1: Overview of off-the-shelf diagnostic tools, non-exhaustive.

5 TNO-ESI diagnostic proposition

Based on the identified socio-technical trends and the challenges and needs specific to the diagnostic domain of high-tech systems, this section will outline ESI's comprehensive proposition. It will present a future vision, exploring potential advancements and innovative approaches to address the emerging diagnostic needs identified above.

Furthermore, it will briefly present ESI's most recent activities on diagnostics, examining their focus and contributions to the field. By combining the historical insights with a forward-thinking perspective, this section aims to provide a roadmap for the future of high-tech system diagnostics.

5.1 Diagnostic vision

ESI's future diagnostic framework is depicted by Figure 5-1. At the core of the framework sits the *diagnostic model*. The precise definition of such a mathematical model is complex and varies across specific diagnostic approaches or implementations. For the purpose of this document, we consider it as a computational entity that encapsulates all diagnostic-relevant system information and serves three key diagnostic use cases: design for diagnostics, reactive diagnostics, and prognosis. Each of these three will be briefly explained in the following paragraphs. Our vision is that we will continue our efforts in developing model-based methodologies for diagnostics, in which such a diagnostic model will have a central role.

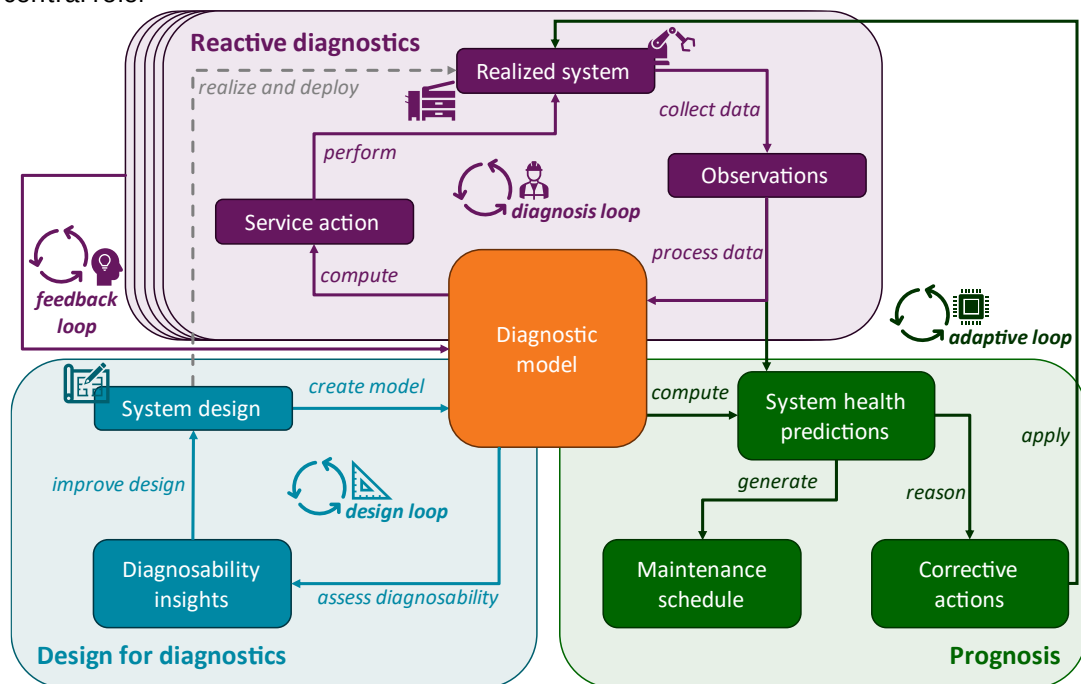


Figure 5-1 Envisioned TNO-ESI diagnostic framework.

Design for diagnostics

In the initial phase of a system's lifecycle, while the system is being designed, one has to consider how well the system could be diagnosed once it is deployed in the field. To do this, an initial diagnostic model should be created to assess the system's **diagnosability** and to identify observability gaps in the system under design. Based on the model's findings, the designers of the system could make an informed decision about improving the observability or not. This is related to the academic field of RAMS, as described in Section 3.

Early stages of the system design phase should also build the foundation of the diagnostic model. All the detail that comes in later stages of the design will add granularity to the model, enriching its resolution and accuracy. The creation of this diagnostic model should not be a separate activity. The creation of the diagnostic model should fit well in existing model-based design processes, tapping into those other bodies of knowledge that are being used in industry [30].

While design for diagnostics initially checks for the system's readiness to be diagnosed (fit-for-diagnosis) based on the observables and tests present in the system, the notion of diagnosis may change in the future. Moving towards more predictive strategies may induce different constraints to be of effective use. Design for diagnostics will change over time, depending on the planned diagnostic needs from the system.

Finally, design for diagnostics will not only help design engineers in ensuring their system to be diagnosable, but it will also help business to determine a fitting diagnostic strategy for their system.

Reactive diagnostics

Once the system is realized and deployed in the field, it will eventually fail. As described in Section 2.1, the observations of the failure may range from a complete system stop to unacceptable performance degradation. Either way, based on the observation of a failure, one needs to identify its cause to be able to resolve the issue. Fault diagnosis is usually an iterative process of conducting tests to acquire new information from the system to refine some intermediate diagnosis, until a final diagnosis is reached.

The diagnostic model will be used as a digital assistant to guide a service engineer to come to a diagnosis. For large and complex systems, where humans struggle to properly oversee how things influence one another, such an assistant can reason beyond what the human could. Also, the assistant will provide unbiased suggestions, possibly suggesting hypotheses the engineer would not have considered.

Data being logged by the system needs to be interpreted automatically by the assistant, as much as possible, to unburden the service engineer. Also, information about the system's state, mode of operation and anything else relevant for diagnosing the system should be considered and properly weighed by the assistant.

Even though such a digital assistant brings a lot of value in complex escalations in which a lot of knowledge needs to be combined to come to a root cause, it may also be used for the simpler cases. At the start of an escalation that requires root cause identification, one does not know the complexity of the case. As such, this reactive diagnostic assistant should be applicable for any diagnostic case.

Prognosis

While the system is in operation, a lot of data is being generated by it. This data will be used to assess the system's health during operation. Besides assessing its current health, predictions could be made of its future health. Even though this is already being done at small scale for individual components, see Section 3, doing this at system scale will bring

more value. By considering the big picture, better estimates of the system's effectiveness could be made, as redundancy or specific use of components or modules in the system could be incorporated.

Being able to predict the system's effectiveness based on prognosis of health assessments allows for efficient maintenance scheduling. Maintenance could be scheduled based on the prediction that the system may no longer perform its function in the near future. By predicting the system's condition at various levels of granularity, one could decide to also inspect parts that are mildly degraded during a planned service visit. This results in fewer more effective service visits compared to when one acts just on the fully degraded parts. This informed way of scheduling maintenance will lead to a higher overall effectiveness of the system.

While the maintenance schedule still needs a human in the loop for ensuring system effectiveness, it would be better if the system can self-adapt to its continuously changing situation. Based on predictions, the system may maximize its effectiveness by changing configuration by itself. For redundant components or modules, this could be condition-based load balancing, but it may also automatically change movement speeds or throughput to temporarily reduce wear to some parts. Such self-adapting actions may extend the lifetime of the system sufficiently long while it awaits maintenance actions.

Field feedback

The diagnostic model that results after the system design has been completed is based solely on design information. After deployment of the system in the field, observations made on the installed base of systems may be used to improve the model. Also, learnings from service visits or in-field diagnostics – preferably assisted by the model – will be included in the model itself. This stream of feedback improves the quality and accuracy of the model. The diagnostic model therefore provides a structured way of capturing these learnings. Combining this design-driven initial model with the data-driven feedback creates a joint design- and data-driven model. This model may also reveal discrepancies over time between the intended way the system works and the way the system actually works once it is deployed. This could provide valuable feedback for the next iteration of product design.

5.2 Past activities

In recent years, TNO-ESI has collaborated with high-tech industrial partners on several diagnostic projects. While in all the projects we have explored approaches for diagnostics of high-tech systems, their specific scopes and contributions have differed. Some example of these are given below:

- model-based diagnostics methodologies focusing on both physical (structural) and functional descriptions of a system [31]. The main difference being that a structure-based approach has a higher diagnostic accuracy at the cost of modelling effort, compared to the more abstract functional approach.
- digital assistants for diagnosing hardware failures that led to hard-down situations, making use of probabilistic graphical models for reasoning. To achieve this, we introduced a method to iteratively suggest the next best service action, i.e. the reactive diagnostics loop in Figure 5-1 [32] [33] [34].
- methods to compute the diagnosability of a system at the design stage, i.e. the design for diagnostics loop in Figure 5-1. This involved evaluating the system's ability to be diagnosed during operation, rather than focusing on immediate diagnostic actions [35] [36].

- Methods and tools for predicting performance (timing) behaviour and diagnosing related issues [37] [38].

Overall, TNO-ESI diagnostic activities have been on diagnosing both HW and SW failures, mostly separately. Further research is needed to address the interplay of SW and HW, particularly at system level. Despite the different scopes, the main approach adopted in the different projects is overlapping. All the projects implement a model-based approach with diagnostic computational models created or generated using knowledge on the system design.

Figure 5-2 shows some of the elements that have already been (partially) addressed in past TNO-ESI activities. Research in three different modelling paradigms – structure, function, and performance – has been done, each leading to a methodology supporting one or more use cases. To support these use cases, algorithms have been developed to use the information captured in the diagnostic model by the paradigm. The types of algorithms that have been developed are shown by the overlapping blue circles on the purple paradigms. To enable the algorithms developed, one or more formalisms were used to structure information or to execute computations. These are shown as orange circles. Overlapping circles indicate tight relations between the formalism and the algorithms or modelling paradigm.

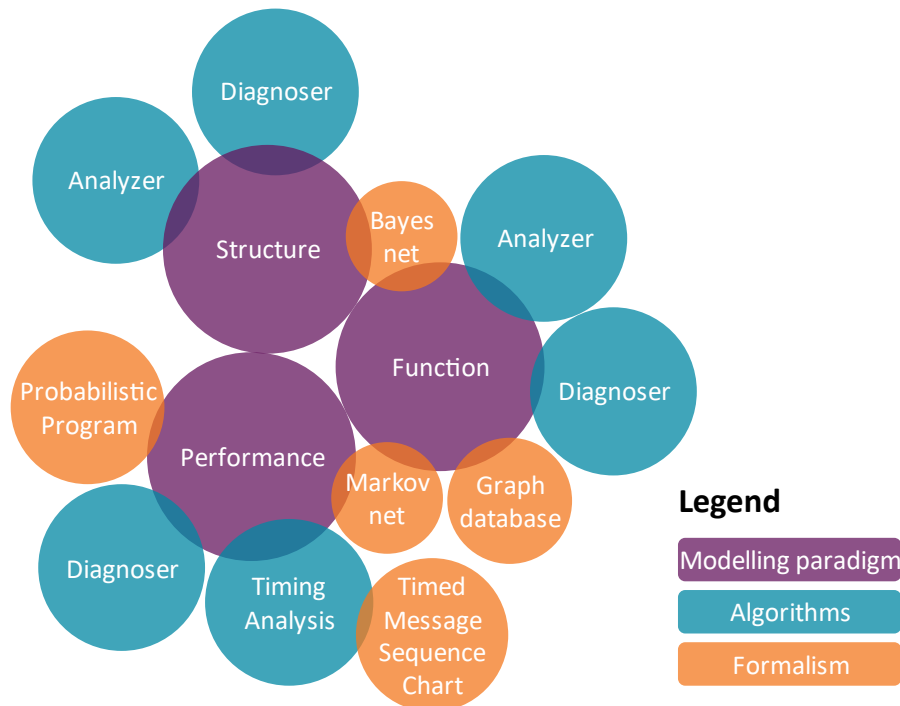


Figure 5-2 Schematic overview of the diagnostic work already (partly) addressed by TNO-ESI in the past years. Overlapping bubbles indicate a tight relation between them, e.g., both the performance and function modelling paradigms use Markov nets as a formalism for reasoning. Bubble sizes are only for visualization purpose.

6 Conclusions

This document outlined the TNO-ESI vision on diagnostics of high-tech industrial system, by addressing the three main needs identified in diagnostics:

- **generalist service engineer:** make diagnostics accessible to a wider range of professionals, i.e. less reliance on experts in a world where these are scarce. TNO-ESI supports through the development of methods that provide decision support in the field for **reactive diagnostics**. Digital assistants will be developed that combine knowledge of the system with real-time data to help service engineers in identifying and resolving system failures.
- **first-time right:** decrease the wasting of time, materials and energy in a world of increased system dependability and where sustainability requirements will be enforced. This will be supported by model-based methods that include system knowledge and field learnings for assisting service engineers. **Design for diagnostics** methods will be developed to ensure availability of the right data to reason with.
- **maximize effective up-time:** shift maintenance strategies from reactive to proactive, in a world in which there is a transition from selling systems to selling services. **Prognosis** methods will be developed to accurately predict the current and future system's effectiveness. Based on these assessments, maintenance plans can be made or mitigating actions can be taken.

For more details on these needs and how they relate to socio-technical trends of the high-tech sector, see Section 2. To learn more about the key diagnostic areas TNO-ESI will contribute to, see Section 5.1.

Overall, we foresee that the high-tech sector will have to invest in model-based diagnostic methodologies, where the diagnostic model is not static, but is continually improved via a feedback loop from the field providing new knowledge. This approach leverages data-based and AI approaches, and will result in enhanced system reliability, reduced downtime, and optimized maintenance efforts.

Acknowledgements

The research is carried out as part of the Kennisinvesteringsproject (KIP) Feasibility study of probabilistic programming applications in industrial engineering, under the responsibility of TNO-ESI. The research is supported by the Netherlands Organisation for Applied Scientific Research TNO.

7 References

- [1] A. J. Hendriks and S. Acur, "Vision and Outlook for Systems Architecting and Systems Engineering in the High-Tech Equipment Industry," TNO Report, 2024.
- [2] K. Classens, J. v. d. Wijdeven, M. Heemels and T. Oomen, "From Fault Diagnosis and Predictive Maintenance to Control Reconfiguration," *DSPE MIKRONIEK*, vol. 63, no. 5, pp. 5-12, 2023.
- [3] W. Tiddens, "Setting sail towards predictive maintenance," UTwente, PhD thesis, 2018.
- [4] S. Hanqi and J. Lee, "Machine Learning Approaches for Diagnostics and Prognostics of Industrial Systems Using Open Source Data from PHM Data Challenges: A Review," *International Journal of Prognostics and Health Management*, vol. 15, 2024.
- [5] W. Cheng, H. Ahmad, L. Gao, J. Xing, N. Zelin, X. Chen, Z. Xu and R. Zhang, "Diagnostics and Prognostics in Power Plants: A systematic review," *Reliability Engineering and System Safety*, 2024.
- [6] A. Trilla, N. Mijatovic and X. Vilasis-Cardona, "Unsupervised Probabilistic Anomaly Detection over Nominal Subsystem Events through a Hierarchical Variational Autoencoder," *International Journal of Prognostics and Health Management*, vol. 14, 2023.
- [7] C. Huang, S. Bu, H. H. Lee, C. H. Chan, S. W. Kong and W. K. Yung, "Prognostics and health management for predictive maintenance: A review," *Journal of Manufacturing Systems*, vol. 75, 2024.
- [8] "Prisma 2016 project cooperation agreement (doc no 2015-10196)".
- [9] M. Fernandes, J. M. Corchado and G. Marreiros, "Machine learning techniques applied to mechanical fault diagnosis and fault prognosis in the context of real industrial manufacturing," *Applied Intelligence*, 2022.
- [10] S. Bi, C. Wang, B. Wu, S. Hu, W. Huang, W. Ni, Y. Gong and X. Wang, "A comprehensive survey on applications of AI technologies to failure analysis of industrial systems," *Engineering Failure Analysis*, 2023.
- [11] J. Bofill, M. Abisado, J. Villaverde and G. Avelino Sampedro, "Exploring Digital Twin-Based Fault Monitoring: Challenges and Opportunities," *Sensors*, vol. 23, 2023.
- [12] A. Ifeanyi and J. Coble, "Online and offline fault detection and diagnostics in an nuclear power plant condenser," in *Proceedings of the Annual Conference of the PHM Society 2024*, 2024.
- [13] Y. Liu and X. Jiang, "Towards Predictive Maintenance of a Heavy-Duty Gas Turbine A New Hybrid Intelligent Methodology for Performance Simulation," in *Annual Conference of the PHM Society*, 2022.
- [14] E. Pesola, K. Kocio, M. Prather and A. Ildefonso, "A Hybrid Model-Based and Data-Driven Framework for Automated Spacecraft Fault Detection," in *Annual Conference of the PHM Society*, 2023.
- [15] B. Xiao, J. Zhong, X. Bao, L. Chen, J. Bao and Z. Yu, "Digital twin-driven prognostics and health management for industrial assets," *Scientific Reports*, vol. 14, 2024.

- [16] K. F. Niresi, H. Bissig, H. Baumann and O. Fink, "Physics-Enhanced Graph Neural Networks For Soft Sensing in Industrial Internet of Things," in *32nd European Signal Processing Conference*, 2024.
- [17] A. e. al., "Modeling and Simulation of Thermal Effects on Electrical Behavior," in *ANNUAL CONFERENCE OF THE PROGNOSTICS AND HEALTH MANAGEMENT SOCIETY 2024*, 2024.
- [18] C. Teubert, K. J. Griffith, M. Corbetta, C. Kulkarni, P. Banerjee and M. Daigle, *ProgPy Python Prognostics Packages*, 2024.
- [19] A. Ifeanyi, M. Zanotelli and J. Coble, "Data-Driven Remaining Useful Life Estimation: Inference-Based Versus Direct Prediction," in *ANNUAL CONFERENCE OF THE PROGNOSTICS AND HEALTH MANAGEMENT SOCIETY*, 2024.
- [20] B. Pidaparthi, R. Jacobs, S. Ghosh, S. K. Ravi, A. W. Amer, L. Luan, R. M. Krishnan, F. Zhang, V. Perez and L. Wang, "Proactive Aircraft Engine Removal Planning with Dynamic Bayesian Networks," in *Annual Conference of the Prognostics and Health Management Society 2024*, 2024.
- [21] K. Bouhadra and F. Forest, "Knowledge-based and Expert Systems in Prognostics and Health Management: A Survey," *International Journal of Prognostics and Health Management*, vol. 15, 2024.
- [22] F. Tamssaouet, K. T. P. Nguyen, K. Medjaher and M. Orchard, "A Fresh New Look on System-level Prognostic: Handling Multi-component Interactions, Mission Profile Impacts, and Uncertainty Quantification," *International Journal of Prognostics and Health Management*, 2021.
- [23] D. Mandelli, C. Wang, K. A. Manjunatha, V. Agarwal and L. Lin, "Rethinking Reliability in Terms of Margins," in *ANNUAL CONFERENCE OF THE PROGNOSTICS AND HEALTH MANAGEMENT SOCIETY 2023*, 2023.
- [24] "GE SmartSignal," GE Vernova, [Online]. Available: <https://www.gevernova.com/software/products/asset-performance-management/equipment-downtime-predictive-analytics>. [Accessed 3 12 2024].
- [25] A. Hess, J. Stecki and S. Rudov-Clark, *The Maintenance Aware Design environment: Development of an Aerospace PHM Software Tool*, 2008.
- [26] S. Deb, K. Pattipati, V. Raghavan, M. Shakeri and R. Shrestha, "Multi-Signal Flow Graphs: A Novel Approach for System Testability Analysis and Fault Diagnosis.," in *Autotestcon*, 1994.
- [27] M. Lind, "An introduction to multilevel flow modeling," *Nuclear Safety and Simulation*, vol. 2, 2011.
- [28] S. Adén and K. Stjernström, "Guidelines for modeling hydraulic components and model based diagnostics of hydraulic applications".
- [29] C. DePaul, *Model-Based Safety Analysis and Assessment*, 2019.
- [30] E. van Gerwen, L. Barbini and T. Nägele, "Integrating System Failure Diagnostics Into Model-based System Engineering," *INCOSE INSIGHT*, vol. 25, no. 4, pp. 51-57, December 2022.
- [31] T. Nägele, L. Barbini, G.-J. van den Braak, A. Piedrafita Postigo and M. Lipplaa, "Guided diagnosis of functional failures in cyberphysical systems," TNO Report, 2025.
- [32] E. van Gerwen, "Guided root cause analysis of machine failures - Status 2023," 2024.
- [33] E. van Gerwen, L. Barbini and M. Borth, "Differential Diagnosis with Active Testing," in *4th Asia Pacific Conference of the Prognostics and Health Management*, Tokyo, Japan, 2023.

- [34] L. Barbini, C. Bratosin and E. van Gerwen, “Model based diagnosis in complex industrial systems: a methodology,” in *PHM Society European Conference*, 2020.
- [35] L. Barbini, C. Bratosin and T. Nägele, “Embedding Diagnosability of Complex Industrial Systems Into the Design Process Using a Model-Based Methodology,” in *PHM Society European Conference*, 2021.
- [36] P. v. Hertum and T. Nägele, “4.5.1. Automated design-driven diagnostics for lithographymachines at ASML,” in *Demystifying Artificial Intelligence: Symbolic, Data-Driven, Statistical and Ethical AI*, Berlin, Boston, De Gruyter, 2024, pp. 127-136.
- [37] B. v. d. Sanden, Y. Li, J. v. d. Aker, B. Akesson, T. Bijlsma, M. Hendriks, K. Triantafyllidis, J. Voeten and T. Basten, “Model-driven system-performance engineering for cyber-physical systems,” in *Proceedings of the 2021 International Conference on Embedded Software*, New York, NY, USA, 2021.
- [38] K. Triantafyllidis, S. Niknam and Y. Blankenstein, “PPS: From Methodology to Application in Industry : A method for analyzing performance on complex Cyber Physical Systems,” TNO, 2023.

ICT, Strategy & Policy

High Tech Campus 25
5656 AE Eindhoven
www.tno.nl