Diss. ETH No. 29153

FAULT DIAGNOSTICS UNDER LABEL AND DATA SCARCITY

A dissertation submitted to attain the degree of

Doctor of Sciences of ETH Zurich
(Dr. sc. ETH Zurich)

presented by

Katharina Rombach

MSc EST, ETH Zurich

born on 16.12.1986

citizen of Germany

accepted on the recommendation of

Prof. Dr.	Olga Fink,	[examiner]
Prof. Dr.	Gabriel Michau,	[co-examiner]
Prof. Dr.	Giovanni Sansavini,	[co-examiner]
Prof. Dr.	Piero Baraldi,	[co-examiner]

2023

Abstract

Deep learning based on representative and large-scale training datasets has lead to impressive performance gains in various fields. With an increasing availability of condition monitoring data from industrial assets, deep learning holds much potential to be applied to fault detection and diagnostics. Unfortunately, in the context of safety-critical systems, the available training datasets are typically neither representative nor large-scale. The reasons for that are two-fold: First, faults occur very rarely in safety critical systems, leading to scarcity of faulty samples. Moreover, even if anomalies or faults occur, it is difficult to determine the exact point in time of their initiation, resulting in scarce and imprecise labeling. Second, there is a large variety of operating conditions affecting the condition monitoring data. It is impossible to collect a representative dataset within a limited observation time period that covers all the relevant conditions. The resulting data and label scarcity can limit the performance of deep learning models for condition assessment. While some solutions have been proposed for data and label scarcity, they often make implicit or explicit assumptions that are often unrealistic in real-world scenarios. The dissertation addresses these real-world challenges and limitations and proposes four main contributions.

Firstly, while robustness of fault diagnostics models to changing operational environment has been approached before, the previously proposed approaches did not consider that in addition to changes in the operational environment also a new health condition might emerge at deployment time. We propose a methodology based on contrastive feature learning that enables to achieve both objectives simultaneously: (1) robustness towards changes in the operational environment and (2) sensitivity to novel faults.

Secondly, to enable the transfer of fault diagnostics models between different operational environments, common restrictions are that the same fault classes must have occurred under both operational environments and that the change in the operational environments is small. These restrictions severely limit the application of existing methods to domain adaptation tasks in the context of real industrial applications. To lift these restrictions, we propose a data generative approach that is particularly beneficial for domain adaptation with extreme label space discrepancies and thus, is suitable for realistic settings of fault diagnostics under changes in the operational environment.

Thirdly, to tackle the challenge of label noise, previously proposed methods typically require prior knowledge about the label noise. This is often not available in reality and thus, existing methods are not suited to be applied in a realistic operational context. We eliminate this limitation and develop a method that solely relies on rough estimates of the label noise level.

Lastly, in the first stage of monitoring an asset where presumably only healthy data is available, different fault detection methods have been proposed. Existing anomaly detection algorithms typically disregard the fact that anomalies in the data are not only caused by faults but also by changes in the operational environment. If faults are detected each time the operational environment changes, the anomaly detection will raise many false alarms and thus, is not deployable in real operations. To counteract that, we adapt contrastive feature learning to be applicable to the anomaly detection setup and demonstrate the superiority of the method on two datasets recorded under real in-service conditions.

We demonstrate that our proposed methods can significantly improve the fault detection and diagnostics performance under real-world constraints, mitigate previously existing limitations and extends the applicability of deep learning to realistic settings in condition monitoring. Thus, the proposed framework extends the applicability of deep learning models

for real-world industrial applications and is a key component in enabling safe operations in all phases of the life cycle of an asset.

Zusammenfassung

Tiefe neuronale Netze zeigen bei verschiedenen Aufgaben eine sehr gute Leistungsfähigkeit, wenn sie auf repräsentativen und grossen Datensätzen trainiert werden. Mit der steigenden Verfügbarkeit von Zustandsüberwachungsdaten bergen sie grosses Potential zur Erkennung und Diagnose von Defekten in Industrieanlagen. Allerdings sind die verfügbaren Zustandsüberwachungsdaten von sicherheitskritschen Industrieanlagen oft weder repräsentativ noch noch haben die eine ausreichende Menge. Folgenden Gründe kommen dafür in Betracht: Zum einen treten Defekte in sicherheitskritischen Anlagen nur selten auf. Das führt dazu, dass Defektdaten oft nur sehr spärlich in den gesammelten Daten repräsentiert sind. Selbst wenn ein Defekt auftritt, ist es schwierig festzustellen, wann genau der Defekt initiiert wurde. Das wiederum führt zu spärlichen und unzuverlässigen Labels in den Daten. Zum anderen werden die Zustandsüberwachungsdaten von vielen Betriebsfaktoren beeinflusst. Es ist nicht möglich, einen repräsentativen Datensatz innerhalb eines beschränkten Zeitraums zu sammeln, der alle relevanten Betriebsfaktoren abdeckt. Die daraus folgende Knappheit der Daten und Labels verschlechtert die Leistungsfähigkeit der neuronalen Netze, die für Zustandsüberwachung eingesetzt werden. Einige Methoden wurden bereits vorgeschlagen für Daten- und Label Knappheit. Allerdings gehen diese Methoden häufig von impliziten oder expliziten Annahmen aus, die in realen industriellen Anwendungen nicht erfüllt sein können. In dieser Dissertation schlagen wir vier wissenschaftliche Beiträge vor, um den Einsatz von neuronalen Netzen in realen industriellen Anwendungen zu ermöglichen.

Erstens werden in vorherigen Publikationen zwar die Robustheit von neuronalen Netzen in Bezug auf wechselnde Betriebsbedingungen angegangen, allerdings nehmen diese Publikationen nicht in Betracht, dass nicht nur die Betriebsbedingungen sich ändern können, sondern auch der Gesundheitszustand der Industrieanlage. Wir schlagen die Methode des 'contrastive learning' vor um beide Ziele zu erreichen: (1) Robustheit in Bezug auf wechselnde Betriebsbedingungen und (2) Sensitivität in Bezug auf neue Defekte.

Zweitens um ein Defekt-Diagnose-Model zwischen verschiedenen Betriebsbedingungen transferieren zu können, müssen in der Regel zwei Anforderungen erfüllt sein: dieselben Defekte müssen unter beiden Betriebsbedingungen vorgekommen sein und die Änderung der Betriebsbedingungen muss klein sein. Diese Anforderungen limitieren die Anwendung von sogenannten Domänenanpassungsmethoden für realistische industrielle Anwendungen. Um diese Anforderungen aufzuheben, schlagen wir eine generative Methode vor, die insbesondere geeignet ist, wenn verschiedene Defekttypen unter verschiedenen Betriebsbedingungen beobachtet wurden und eignet sich daher besonders für realistische industrielle Anwendungen der Domänenanpassung von Defekt-Diagnose-Modellen.

Drittens bisherige Methoden, die die Herausforderung von unzuverlässigen Labels angehen, benötigen in der Regel vorherige Kenntnis über die Art oder Menge der Unverlässlichkeit der Labels. Dies ist allerdings nicht verfügbar in realistischen industriellen Anwendungen. Wir eliminieren diese Anforderung und entwickeln eine Methode, die nur eine grobe Einschätzung der Unverlässlichkeit der Labels benötigt.

Als letztes, frühere Studien haben Methoden zur Defekterkennung vorgeschlagen für die erste Phase der Zustandsüberwachung, in welcher vermutlich nur gesunde Zustände aufgetreten sind. Allerdings nehmen diese Anomalie-Detektions-Methoden nicht in Betracht, dass Anomalien in den Daten nicht nur von Defekten herrühren, sondern auch von veränderten Betriebsbedingungen. Wenn ein Modell aufgrund veränderter Betriebsbedingungen Defekte detektiert, kann es nicht eingesetzt werden unter wechselnden Betriebsbedingungen, da es

sonst zu vielen Betriebsunterbrechungen aufgrund von falschen Alarmen kommen würde. Um dem entgegenzuwirken, adaptieren wir die Methode des 'contrastive learning' so, dass es auch ohne echte Defektdaten anwendbar wird. Wir zeigen die Überlegenheit von unserer Methode an zwei Datensätzen die unter echten Betriebsbedingunghen aufgezeichnet wurden.

In dieser Forschung zeigen wir, dass unsere Methoden die Leistungsfähigkeit der Defekterkennung und -diagnose in Industrieanlagen signifikant verbessern, dass vorherige Limitationen aufgehoben wurden und dass die Anwendung von neuronalen Netzen unter realistischen Bedingungen erweitert wurde. Das entwickelte Framework erweitert somit die Anwendbarkeit von neuronalen Netzen für industrielle Anwendungen und ermöglicht einen sicheren Betrieb in verschiedenen Phasen des Lebenszyklus von Industrieanlagen.

Contents

Acknowledgments

I would like to express my deepest gratitude to my supervisor Prof. Olga Fink for letting me be participate in her research and for the guidance, inspiration, and ideas in the last few years. It was truly inspiring to work with you, learn from you and your on-point feedback. Thanks for your patience (especially in the second year) and encouragement to develop as a researcher and as a person.

Equally so, I could not have undertaken this journey without my second supervisor Dr. Gabriel Michau. I truly valued your feedback, guidance and patience. Thank you so much for you time, for generously providing me with knowledge and expertise, if work or life related. It was such a pleasure and inspiration to learn from you.

I am extremely grateful to my partners at SBB for providing me with valuable insights. Dear Dr. Wilfried Bürzle, Dr. Stefan Koller, Dominik Imfeld, Silvan Rohrbach and Arturo Vivas, thanks so much for the time and support.

I would also like to express my gratitude to Prof. Giovanni Sansavini, Prof. Piero Baraldi and Prof. Ioannis Anastasopoulos. It is an honor to have you in my PhD committee.

And to my colleagues and friends in the IMS/IMOS team, for the discussions, support, and joy in the last few years. Thanks so much for making me feel so welcome and part of the team although I was mainly joining remotely.

Lastly, words cannot express my gratitude to my family and friends so I will keep it short. Dear mom, dad and Franzi. Thanks a lot for the unconditional support and for equipping me with the tools to enjoy life and what I do, even in tough times. I obviously would have achieved nothing without you. Dear Lotti, thanks a lot for the emotional support. Dear Marc, thanks for bearing with me even if I might complain a lot and thanks supporting me with english. Finally, I would like to thank Christoph for everything. Thanks for your support, love and patience. Thanks for being able to accept me in any state of mind.

1 Introduction

1.1 Motivation

Critical services such as transport, energy or communication are essential pillars for the proper functioning of our modern and globalized world. Traditionally, industrial and infrastructure assets have been monitored and assessed by human experts who visually inspect the assets (Jing et al., 2021; Yu et al., 2019). Physics based models have also been proposed for monitoring industrial assets (Luo et al., 2003; Cubillo et al., 2016). However, the implementation and performance of physics based models is often obfuscated by noisy environments and system complexities (Khan and Yairi, 2018). With the increasing complexity of industrial systems and the increasing requirements on system availability and performance, a reliable and generalizable solution for real-time condition monitoring is needed to enable safe, efficient and effective operations (Vrignat et al., 2022).

Major pillars for condition monitoring are the early and accurate detection of an incipient fault (fault detection) and the identification of its specific fault type and severity level (fault diagnostics) (Fink et al., 2020). In recent years, there has been much effort conducted to develop and implement algorithms to monitor the health condition of complex industrial assets - we refer to these algorithms as condition assessment solutions.

With the development of low cost sensors and the increasing capacity to store and process data, more and more condition monitoring systems have been installed, providing real-time information on the system's condition (Jardine et al., 2006). Data-driven algorithms based on the collected real-time condition monitoring data have shown to be a very promising direction for monitoring the condition of industrial assets (Zhang et al., 2019; Chien and Chen, 2020). Their potential has been demonstrated in a multitude of studies which have shown that superior empirical results have been achieved in various application fields such as bearing or wind turbine fault diagnostics compared to other methods (Zhang et al., 2020; Hoang and Kang, 2019).

Especially deep learning, a subcategory of machine learning based on deep neural networks, has led to impressive performance gains in recent years, especially in applications such as computer vision or natural language processing (Ramesh et al., 2021; Goyal et al., 2021). While traditional machine learning approaches required manual and time-consuming feature engineering, the main difference in deep learning is that representative features can be learnt automatically.

Motivated by the promising achievements of deep learning in many fields of research, there has been a high interest in recent years to transfer the full potential of deep learning from other applications to Prognostics and Health Management (PHM) tasks (Fink et al., 2020). Contrary to other research studies conducted on deep learning, PHM tasks mostly rely on time series data which has not been in the center of deep learning research which mainly focused on image datasets. Different types of deep learning methods have been proposed for fault detection, diagnostics and prognostics - also based on time series data (Serradilla et al., 2022). However, certain challenges arise when working on deep learning for PHM tasks such as the lack of or scarce representation of fault data or the multitude of operational or environmental factors that can strongly influence the condition monitoring data.

1.2 Challenges

The success of deep learning has largely relied on the availability of representative and labeled large-scale datasets for the training of the neural network models. In particular, the training

datasets need to be representative of the conditions and settings that the model might endure when it is deployed (test dataset). One of the main challenges of PHM applications has been a lack of such representative, labeled and sufficiently large datasets. This data and label scarcity hinders to transfer the full potential of deep learning to PHM tasks. Real condition monitoring datasets of in-service systems, for example, might lack data samples of faulty conditions or might contain a lot of variability in the quality of labels. There are three main factors responsible for data and label scarcity:

Rareness of faults. As safety-critical faults cannot be tolerated in operation, industrial assets are operated in safe regimes (Michau and Fink, 2021). Thus, faults in safety-critical systems occur very rarely (Fink et al., 2020; Biggio and Kastanis, 2020). While this is desired from the operational perspective, from the data perspective, it means that faulty conditions are not represented well in the available condition monitoring datasets used for the training of machine learning algorithms. However, we are particularly interested in detecting, diagnosing and predicting those faulty system conditions. Traditional supervised learning algorithms would require a respresentative fault dataset in order to learn the specificities of the distinct fault patterns and classify them precisely. If, however, no fault or only few faults occurred at training time, such a representative dataset is not available.

When no fault data is available in the training dataset, anomaly detection models have been deployed to detect any anomalous patterns that deviate from the healthy training dataset (Zhao et al., 2018; McKinnon et al., 2020). However, not every novel pattern necessarily corresponds to a fault in the system. Instead, other factors can also cause anomalous patterns in the data such as new operating or environmental conditions (details are elaborated below) (Michau and Fink, 2021). Previously proposed anomaly detection models that raise an alarm if any anomalous pattern is observed could result in a high false alarm rate (Michau and Fink, 2021). For adequate applicability in real applications, it is important that anomaly detection models are only sensitive to changes in the health conditions resp. faults and not to other changes in the data caused by changing operating or environmental conditions.

Contrary to the case of missing faulty samples in the training dataset, in other cases some faults have been observed during the data acquisition time and hence, the respective fault data is represented in the training dataset. In this case, deep classification models can be trained to distinguish between known fault types and/or severity levels if the respective information on the fault types and severity levels is available. While superior performances have been achieved by employing deep classification models for fault diagnostics (Li et al., 2021a; Zhang et al., 2020), new health conditions i.e. new fault types and severity levels might emerge at any point in time. Thus, a mature solution to condition assessment should not only entail the correct distinction of known health conditions but also the detection of novel health conditions. Achieving both of these objectives jointly is particularly challenging when dealing with condition monitoring data which is subjected to a variety of factors that cause variations in the data such as operational or environmental conditions (see below).

Non-informative factors of variation in the data. Not only evolving health conditions, but also other factors in an operational environment can cause variations in the condition monitoring data. Typical factors are operating or environmental conditions such as operational speed or load, the ambient temperature or sensor location (Shi et al., 2022). Since all of the above mentioned factors are unrelated to the health condition of the asset, we will refer to them as non-informative factors of variation in this thesis. These non-informative factors can cause a discrete or continuous shift in the underlying data distribution (Michau and Fink, 2021). In deep learning terminology, these shifts are referred to as domain shifts (Zhou et al., 2022c). The performance of data-driven models can drop significantly under

domain shifts (Wang et al., 2019). For example, an anomaly detection model could raise a false alarm if confronted with anomalies in the data due to a domain shift. This will lead to a high false positive rate and disqualifies a model from being deployed for PHM tasks as it will hinder efficient operations and would create additional cost. If the anomaly detection model is adjusted to be less sensitive, it might also fail to detect early signs of an defect. This will lead to a higher false negative rate, which also disqualifies a model from being deployed as a missed fault can pose a safety risk. Equally so, the performance of a classification model is negatively impacted under domain shifts since data-driven models are prone to provide wrong predictions if the underlying data distribution has shifted. Thus, for assets that are monitored with data-driven fault detection and diagnostics models, the non-informative factors of variation can pose a severe safety and reliability risk if an algorithm's training dataset was recorded under different non-informative factors of variation compared to those it encounters when deployed (test dataset).

Some of the factors causing variations might be known and controllable in the sense that (a) they can be set to each desired value to record a representative dataset and (b) each change of the operational parameters can easily be detected or can be measured. For example, the operational speed or load typically can be set or are measured and thus, are controllable. Domain shifts caused by controllable factors can be detected and appropriate mitigation strategies can be pursued to prevent a performance drop of the deployed condition assessment solutions. For example, unsupervised domain adaptation approaches have been proposed to adapt a model from a labeled source domain to an unlabeled target domain (Farahani et al., 2021). Domain adaptation has also been successfully applied to PHM tasks where domains typically represent different operational or environmental conditions under which the data is recorded (Li et al., 2020a). Most of the proposed methods require that the same classes are represented in the source and the target domain (Rombach et al., 2023). This requirement, however, is not realistic in the context of a complex industrial environment. Since faults are rare in these environments, it is not realistic to assume that the same fault types and severity levels occur during the data acquisition time in the source and target domain. There has been an increasing interest in recent years to develop domain adaptation methods that meet the requirements of a complex industrial environment, where different fault classes are represented in the source and the target domain (Zhou et al., 2022a; Zhang et al., 2021b). Different types of this label space discrepancies are distinguished. For example, if only a subset of classes in the source domain is represented in the target domain, researchers refer to this setting as the *Partial* setting (Li et al., 2020b). Or, the setting where both domains have classes that are not represented in the other domain is referred to as the *Open-Partial* setting. Different approaches have been developed which are suited for one specific setting of the above mentioned label space discrepancy settings. For example, a method for *Partial* domain adaptation is proposed for bearing fault diagnostics (Zhou et al., 2022a). For the *OpenSet* domain adaptation setting on similar data, another method has been proposed (Zhang et al., 2021b). These methods are often not universally applicable in multiple settings of label space discrepancies. Moreover, in practice, extreme cases of label space discrepancy, where only the healthy class is shared between two domains, are common as faults occur rarely (Rombach et al., 2023). These extreme cases have only been tackled by researchers insufficiently so far. Furthermore, most publications evaluated the proposed method on rather small domain gaps such as slight changes in very few operating parameters and might, therefore, not perform well on larger domain gaps.

Contrary to the known and controllable factors, where distinct domain shifts in the data can be identified *a priori*, there are also other factors of variation that might be unknown or not measurable or controllable. For example, the temperature might impact the recorded measurement data. While the ambient temperature often is measured or tracked, the temperature at the sensor location might substantially differ from the ambient temperature. If there

is no temperature sensor in place at the sensor location, there is no possibility of measuring thistemperature. Moreover, in an open environment, there might exist a multitude of other factors that influence the condition monitoring data that might be unknown to the operator. In these cases, when the factors are not known or cannot be measured, domain shifts cannot be identified *a priori*. This poses a safety critical risk as an undetected domain shift results in an undetected performance drop of the deployed condition assessment solution. Domain generalization has been proposed to tackle this challenge of unforeseen and unknown domain shifts that might emerge after the model has been trained, mainly in the field of computer vision (Zhou et al., 2022c; Wang et al., 2022). Very recently, domain generalization methods have also been proposed and applied for fault diagnostics in industrial assets (Zhang et al., 2021a; Zhao and Shen, 2022a; Liao et al., 2020). These methods were developed to be robust towards unforeseen or unknown domain shifts such that reliable classification of known health conditions is possible even if the underlying data distribution shifts. In addition to reliable classification of known health conditions under domain shifts, it is equally important to be able to detect the emergence of novel faults in PHM tasks. In other words, in the context of PHM it is critical that not only robustness towards any novel data variation is achieved but in addition to the robustness towards domain shifts, sensitivity to novel data variations caused by a change in the health condition is required. Distinguishing novel variations in the data that are caused by a domain shift from those that are caused by a new health condition is a very challenging task in absence of the corresponding data of either the future domain shift or the future health condition. However, it is a critical requirement for PHM to consider both types of data scarcity: (1) scarce domain representation and (2) scarce fault representations. A mature solution for condition assessment needs to be able to cope with both simultaneously.

Label Noise. Ground truth information on the exact health condition of assets is often difficult to obtain. For example, assets that operate in an open environment such as railway wheels cannot be monitored constantly while in operation. Instead, the ground truth information about the assets' health condition is often assigned by domain experts in hindside at a workshop visit, where the exact point in time of a change in the health condition cannot be reconstructed anymore, or is extracted based on pre-defined rules or assumptions such as fixed timeperiods after maintenance in which an asset is considered to be healthy. Since such rules are not adapted to the specific health conditions or the operational environment of a component, they can quickly result in noisy information on the true health condition of an industrial asset. For the application of deep learning, it means that the available training dataset might be impacted by label noise. Label noise can have a critical effect on the learning process (Algan and Ulusoy, 2021). It encourages the model to memorize individual samples rather than to generalize over certain class-specific attributes, even if the data is drawn from the same underlying data distribution (Arpit et al., 2017). Since label noise is ubiquitous in real world datasets, some methods have been proposed that aim to prevent the memorization of mislabeled samples. These typically rely on prior information on either the characteristics of label noise or a clean validation dataset without label noise (Ren et al., 2018; Vahdat, 2017). Since this prior information is typically not available in real PHM tasks, existing methods are not applicable in a realistic scenario. For PHM applications in realistic scenarios, however, not much research has been conducted to tackle label noise. This is, however, required in order to develop methods that are applicable in scenarios where ground truth information is not only time consuming to obtain but also is often even impossible to obtain.

These three challenges, (1) the rareness of faults, (2) the non-informative factors of variation and (3) the label noise result in the fact that high quality representative labeled datasets

are scarce. Many of the problems encountered in PHM, once formulated as deep learning problems result in 'extreme' setups of known deep learning problems. We define 'extreme' setups as those where the data scarcity does not only relate to a lacking representation of different domains but also to an extremely scarce representation of faults. One example of an 'extremely' scarce representation of faults is often encountered in domain adaptation, where only the healthy class is represented in both, the target and source domain, but fault types were only observed in either one of the domains. I.e. the fault data is scarcely represented in the domains and this results in an extreme label space discrepancy for the tasks of domain adaptation. An example of lacking domain representation is that the training dataset might only be recorded under certain environmental conditions (e.g. summer) or operating conditions. It is however critical that the a model developed under one condition performs also well if exposed to new conditions that were not represented in the training dataset (e.g. winter). Additionally, the data scarcity is often accompanied with scarce labeling. All three challenges combined presents quite a substantial challenge for deep learning methods; especially since reliable condition assessment solutions need to be robust towards unknown domain shifts but, simultaneously, sensitive to novel, previously unobserved fault types. This is quite difficult to achieve as both objectives cannot be imposed directly in absence of data from future domain shifts and future health conditions. Existing methods are not quite satisfying in overcoming these challenges under real conditions. In this thesis, we address these problems and propose new methods to better handle these 'extreme' cases of data and label scarcity to cover the whole life-time of an asset.

1.3 Research Gaps and Overriding Research Questions

In this thesis, we propose a framework for condition assessment, that is robust towards novel domain shifts, is robust to label noise and is sensitive towards novel health conditions. To achieve this, we address several research gaps with respect to data scarcity and label quality.

Research Gap 1 A mature fault diagnostics solution needs to ensure safe operation even under unforeseen changes of operating or environmental conditions while also enabling the detection of novel fault types. These two objectives have not yet been tackled simultaneously.

Research Question 1 *How can we train a fault diagnostics model that is both, able to perform well on known and unknown domains as well as able to detect novel fault types?*

Research Gap 2 It is crucial to be able to develop robust and reliable fault diagnostics solutions also for systems that have not yet experienced any faults but whose operating profiles are dissimilar to those where faults are known. In other words, it is crucial to adapt or transfer fault diagnostics models from one domain to another under real conditions. Contrary to previously proposed domain adaptation approaches, a solution that is applicable under real conditions needs to be able to deal with extreme label space discrepancies, deal with large domain gaps and deal with different types of label space discrepancies (universally applicable). For example, if different fault types are observed in the source and target domain during the data acquisition time and only the healthy class is shared between the domains, resulting in a extreme label space discrepancy. Furthermore, the operating regime of the asset might have changed significantly between the source and target domain, resulting in large domain gaps. These scenarios are common in condition monitoring datasets and need to be addressed. Moreover, the developed method needs to be applicable under realistic conditions, where no representative data from the missing health conditions is available to tune the methodology.

Research Question 2 *How can we enable universal domain adaptation for fault diagnostics models with extreme label space discrepancies and large domain gaps?*

Research Gap 3 Since label noise is ubiquitous in condition monitoring datasets, it is crucial to have a mechanism that enables robust training of deep neural networks despite the presence of label noise. Existing approaches that aim to stabilize the training of data-driven models in the presence of label noise either require a clean validation dataset that can be used as a reference or knowledge about the characteristics of the label noise (type and extent). However, in real applications of fault diagnostics these requirements cannot be fulfilled. Typically, there is no ground truth information available such as knowledge about the type and extent of the label noise or a clean validation dataset. In the context of PHM, the challenge of label noise is particularly pronounced since it can pose a safety-critical risk in PHM if no mitigation strategy is taken. In this dissertation, we aim to develop a methodology that can deal with label noise without any exact knowledge about the label noise or a clean dataset on which hyperparameter (HP) tuning can be performed.

Research Question 3 *How can effective fault diagnostics be enabled in the presence of label noise if no preliminary knowledge about the amount of label noise and no clean validation dataset is available?*

Research Gap 4 Lastly, we address the challenge of distinguishing variations in the data that are caused by a change in the health condition from those that are caused by non-informative factors (domain shifts) in cases where no or only little fault data is available. Contrary to previous works for fault detection, we aim to move from an anomaly detection model that is sensitive to any kind of anomaly in the data to a fault detection model that is only sensitive to faults.

Research Question 4 *How to concurrently achieve invariance to non-informative factors and sensitivity to fault types for fault diagnostics but also for fault detection, where only healthy data and no fault data is available?*

In this dissertation, we will investigate various sizes of domain shifts originating from changing operating conditions only. However, there is no limitation to applying the developed methods to domain shifts caused by other factors of variation.

2 Background

In this section, we present a background on general topics that are closely related to this dissertation. We focus the literature review mainly on publications that are relevant in the field of PHM. A detailed introduction on the topics that are specific to the papers, are introduced in the respective chapters. Moreover, to avoid redundancy, we provide the literature review on certain topics like research on label noise only in the respective chapters.

Condition Monitoring and Assessment. According to Williams et al. (1994), condition monitoring is defined as the continuous or periodic measurement and interpretation of data to indicate the condition of an item to determine the need for maintenance. The term is used in the context of data acquisition and also in the context of interpreting the acquired data (Widodo and Yang, 2007). Therefore, the term condition monitoring does not allow to distinguish between the the process of acquiring data and the interpretation of the data with e.g. data-driven models. The term condition assessment is a less widely used in literature. However, it explicitly refers to a tool resp. model for interpreting and assessing the condition of an asset e.g. in the context of technical performance of the building to long-term maintenance expectations (Yacob et al., 2022). In this thesis we will, on the one hand, use the term 'condition assessment solutions' when referring to concrete tools, methods or models that allow us to assess the health condition of the industrial asset based on the condition monitoring data. On the other hand, we will use the term *condition monitoring* when referring to the entire process of acquiring data and drawing conclusions from it.

Transfer Learning. The major underlying assumption in machine and deep learning is that the training and future data must have the same data distribution and that the same learning task needs to be fulfilled (Pan and Yang, 2009). In other words, neither the learning task nor the data distribution should change between the time period in which the model is trained (source domain) and a new time period in which the model is deployed (target domain). However, in many real-world applications and especially in the context of PHM, this assumption may not hold. The underlying data distribution might have shifted between the source and the target domain or the learning task might have changed. For both of these scenarios, the generic concept of transfer learning has been a widely approached topic (Zhuang et al., 2020; Pan and Yang, 2009; Tan et al., 2018). It aims to improve the learning of a predictive function $f = h_T(.)$ in the target domain for a target task using source data and the source task (Pan and Yang, 2009). As a generic concept, different categories of transfer learning are distinguished by Pan and Yang (2009). First, *inductive transfer learning* aims to address the setting where the source and target task differ, regardless if there is also a shift between the source and the target data. Second, *transduction transfer learning* aims to address the setting where there is a domain shift between the source and target domain, but the learning task is identical in both domains. In the context of fault diagnostics, transfer learning has raised a lot of attention in recent years (Li et al., 2020a; Li et al., 2022). The goal is typically to either enable the transfer between different working conditions, the transfer between different machine components or the transfer from simulation to real-world data (Yao et al., 2022). In this thesis, we will therefore focus on the *transduction transfer learning* setting, where the shift in the underlying data distribution is caused by a change in the operational environment (see below). *Transduction transfer learning* tasks are usually tackled with domain adaptation techniques.

Domain Shifts. If the assumption that the training and the test data are drawn from the same data distribution is violated, the performance of deep models can drop significantly (Zhou et al., 2022c). Unfortunately, changes in the underlying data distribution, so called domain shifts, are common in real-world applications and especially in condition monitoring applications where a domain shift can be caused by any change in the operational environment (Siahpour et al., 2020; Ding et al., 2023) (see below for examples). A formal **definition** of a domain is provided by Wang et al. (2022) as:

Definition 1 (Domain)**.** Let X denote a nonempty input space and Y an output space. A domain is composed of data that is sampled from a distribution. We denote it as $S = \{(x_i, y_i)\}_{i=1}^{n} \sim P_{XY}$, where $x \in X \subset \mathbb{R}^d, y \in Y \subset \mathbb{R}$ denotes the label, and P_{XY} denotes the joint distribution of the input sample and output label. X and Y denote the corresponding random variables.

Building on that definition of a domain, a domain shift is defined as:

Definition 2 (Domain Shift)**.** Let $S = \{(x_i, y_i)\}_{i=1}^{n} \sim P_{XY}^{S}$ be data sampled from a source domain and $T = \{(x_i, y_i)\}_{i=1}^{n} \sim P_{XY}^{T}$ data sampled from a target domain. There is a domain shift between S and T, if the joint distributions between the domains are different: $P_{XY}^{S} \neq P_{XY}^{T}$.

The operational environment under which fault condition assessment solutions are deployed can vary substantially (Ragab et al., 2022). For example, the operating conditions can change, the environmental conditions might vary, the configuration of the asset might be updated or new fleets are taken into operation. These changes cause domain shifts that can significantly decrease the performance of deployed data-driven models for condition assessment (Wang et al., 2019). Therefore, domain shifts pose a reliability risk for the industrial operations if the models are not appropriately adapted.

Domain Adaptation for Classification. Domain adaptation aims to adapt or transfer a model between two distinct domains: from one labeled source domain to one specific unlabeled target domain that is subject to a domain shift. In the context of PHM applications, the source domain, for example, could differ from the target domain by the operating conditions under which the data is recorded. A formal definition of domain adaptation is given as:

Definition 3 (Domain Adaptation)**.** Let $S = \{(x_i, y_i)\}_{i=1}^{n} \sim P_{XY}^{S}$ be data from a labeled source domain, $T = \{(x_i)\}_{i=1}^{m} \sim P_{X}^{T}$ be data from an unlabeled target domain and $D_{test} = \{(x_i)\}_{i=1}^{o} \sim P_{X}^{T}$ a test dataset from the same target domain, whereby the underlying data distribution differs between the source and target domain ($P_{XY}^{S} \neq P_{XY}^{T}$ as defined in Definition 2). The variable x is the input data to the model, and y describes the classes $\{0, ..., C\text{-}1\}$ to predict, i.e. health conditions in the context of fault diagnostics. The goal of domain adaptation is to learn a predictive function h : $X \longrightarrow Y$ from the labeled source domain S and the unlabeled target domain T to achieve a minimum prediction error on a dataset sampled from the target domain D_{test}:

$$\min_{h} \mathbf{E}_{(x,y)\in D_{test}}[l(h(x), y)], \tag{2.1}$$

where E is the expectation and l$(\cdot, \cdot)$ is the loss function

Different approaches have been proposed in computer vision including discrepancy-based, adversarial-based and reconstruction-based methods (Wang and Deng, 2018). Due to the relevance of domain shifts in condition monitoring data, methods have recently been applied and adapted for fault diagnostics (Li et al., 2020a). However, negative transfer i.e. leveraging

source domain data undesirably reduces the learning performance in the target domain, has been a long-standing and challenging problem in domain adaptation in general (Zhang et al., 2022). Common causes can be a large domain divergence, poor source or target data quality as well as an inappropriate choice of domain adaptation algorithm (Zhang et al., 2022). Another reason for negative transfer in domain adaptation can be if the label spaces of the source and target domain are not congruent (Cao et al., 2018, 2019). Therefore, most of the previously proposed domain adaptation methods require that the same classes are represented in the source and target domain. This scenario is referred to as the *ClosedSet* domain adaptation setting. Unfortunately, this scenario does not meet the reality of real world datasets (Rombach et al., 2023). Instead, discrepancies in the label space i.e. that different classes are represented in the source and target domain are very common in real applications. In the literature, different scenarios of label space discrepancies are distinguished: In the *Partial* domain adaptation scenario, for example, the target domain covers only a subset of the classes in the source domain. *Vica versa*, in the *OpenSet* domain adaptation scenario, the source domain covers only a subset of classes compared to the target domain. In the *Open-Partial* domain adaptation scenario, both the target and source domain have private classes, i.e. classes that are not represented in the other domain. Due to the relevance of any kind of label space discrepancy to real world applications, there has been an increasing effort to develop methods that are applicable in different scenarios where the label spaces are not congruent (Zhang et al., 2021b,c). Although these developments present a milestone in transferring models to unlabeled domains, domain adaptation approaches are transductive methods and as such, require data from an target dataset at the model development time. That means, they cannot deal with unforeseen domain shifts at deployment time.

Domain Generalization. deals with a challenging setting where one or several different but related domain(s) are given, and the goal is to learn a model that can generalize to an unseen test domain (Wang et al., 2022; Zhou et al., 2022b). Contrary to domain adaptation, domain generalization does not require access to some test resp. target data at development time. For example, in the context of PHM applications, the goal is to train a model on data from different operating conditions (multiple source domains) such that the model can generalize well to data that is recorded under novel, previously unobserved operating conditions at test time. A formal **definition** is provided by Wang et al. (2022) as:

Definition 4 (Domain Generalization)**.** Given M training (source) domains $S_{train} = \{S_i | i = 1, \ldots, M\}$ where $S_i = \{(x_{i,j}, y_{i,j})\}_{j=1}^{n_i}$ denotes the i-th domain. The joint distributions between each pair of domains are different: $P_{XY}^i \neq P_{XY}^j, 1 \leq i \neq j \leq M$. The goal of domain generalization is to learn a robust and generalizable predictive function h : $X \longrightarrow Y$ from the M training domains to achieve a minimum prediction error on an unseen test domain S_{test} (i.e., S_{test} cannot be accessed in training and $P_{XY}^{test} \neq P_{XY}^i$ for $i \in \{1, \ldots, M\}$:

$$\min_h \mathbf{E}_{(x,y)\in S_{test}}[l(h(x), y)], \tag{2.2}$$

where E is the expectation and l(·, ·) is the loss function

Thus, the goal of domain generalization is to train a model that performs well on previously unobserved data that experiences a domain shift compared to the training datasets. Domain generalization methods have been developed mainly in the field of computer vision to address the generalization issue, where data manipulation techniques (Shankar et al., 2018), representation learning techniques (Li et al., 2018a) as well as specific meta-learning strategies (Balaji et al., 2018) have been proposed. Overviews on the different domain generalization techniques for computer vision are in the surveys by Wang et al. (2022) as well as by Zhou et al. (2022c).

Solving the domain generalization issue is not only crucial in computer vision applications but also very critical in PHM applications since changes in the operating conditions, for example, can occur at any point in time causing a domain shift. While domain adaptation for PHM has been addressed in a lot of research studies, domain generalization has only recently started to be addressed for fault diagnostics tasks (Zheng et al., 2019). Mainly, domain invariant feature representations have been trained by employing adversarial techniques, metric learning techniques and even combining it with data augmentation techniques (Li et al., 2020c; Liao et al., 2020; Nejjar et al., 2022). For example, adversarial feature alignment combined with pseudolabeling was proposed to enable domain generalization for fault diagnostics on rotary machines given one labeled and one unlabeled source domain (Liao et al., 2020). To account for the specific challenges in condition monitoring datasets, (Li et al., 2020c) combined three different techniques to train a generalizable fault diagnostics model for bearing datasets based on multiple labeled source domains: a data augmentation technique was proposed combined with adversarial alignment and metric learning. The above mentioned methods explicitly decrease the distinct domain gap between defined the source domains to learn domain invariant features. However, the identification of distinct domains is only possible if the factors causing the domain shift are known or/and can be controlled. For assets that operate in an open environment, some of these factors might not be known, might not be measurable or controllable. If the identification of distinct domains is not possible, methods based on adversarial alignment, for example, can not be applied in a straightforward manner but would require methodological adaptions. Therefore, to satisfy more realistic requirements of real operations, it is preferable to develop methods that do not require the identification of distinct domains.

Contrastive Learning. Contrastive learning aims to learn a feature representation that groups semantically similar data close to each other while pushing semantically dissimilar ones far apart (Schroff et al., 2015). In the PHM context, contrastive learning transfers very well when assuming that data samples recorded under similar health conditions should be considered as semantically similar and those samples recorded under dissimilar health conditions as semantically dissimilar. To achieve this, a contrastive loss function is employed to train an encoder model (see Equation 2.3), whereby x_a is the anchor sample, x_p the positive sample (that shares the semantic meaning with the anchor), and x_n is the negative sample with a different semantic meaning (Schroff et al., 2015), $f(.)$ is the encoded sample, $||.||$ is a distance metric, and ϵ a margin parameter.

$$\begin{aligned} L(x_a, x_p, x_n) = \sum_{x_a \in X} max(0, ||f(x_a) - f(x_p)|| \\ - ||f(x_a) - f(x_n)|| + \epsilon) \end{aligned} \tag{2.3}$$

Contrastive learning has shown to enable generalization for downstream tasks in a theoretical investigation (Saunshi et al., 2019) and thus, lends itself to be applied in the operational context of PHM applications, where generalization towards unforeseen domain shifts is required. Regardless of the exact downstream task for PHM (e.g. fault diagnostics), the deployed condition assessment solution needs to be robust towards a large variety of non-informative factors causing variations in the condition monitoring data. By selecting samples in the positive pair which share the same health condition but are recorded under different operational environments, we can directly impose the invariance towards non-informative factors and, thus, train the encoder model to filter out content in the data that is not relevant. If the encoder model is able to generalize well, it will also show robustness towards novel, unforeseen domain shifts at test time. Simultaneously, the condition assessment solution needs to be very sensitive to changes in the health condition. By maximizing the distance in the

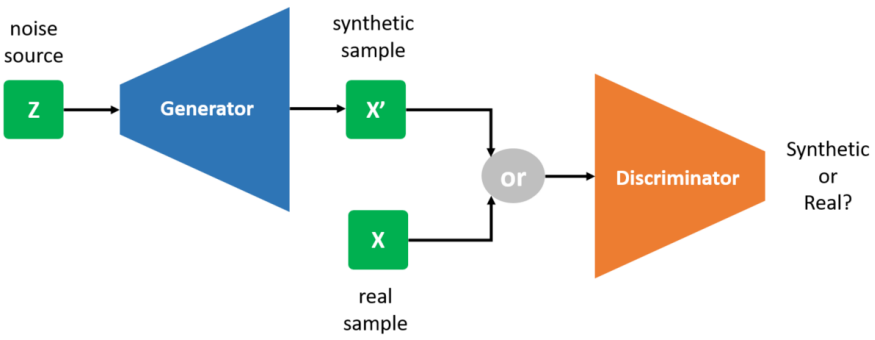


Figure 2.1: Basic structure of a GAN.

feature space between two samples with different health conditions (negative pair), we can directly impose the sensitivity towards changes in the health condition.

The potential of contrastive learning is particularly pronounced in the context of condition monitoring datasets as it does not require to have distinct domains available to reach a domain-invariant embedding, contrary to adversarial approaches that typically rely on distinct source and target domains. Contrastive learning has already been applied in the context of domain generalization. For example, Han et al. (2021) combined contrastive learning with adversarial training to achieve domain-invariant features for domain generalization for fault diagnostics of a planetary gearbox. However, the additional adversarial alignment step requires again the distinct differentiation between the different source domains and thus, does not take full advantage of applying contrastive learning. After the publication of our work, Zhao and Shen (2022b) proposed a method based on contrastive learning to enable the detection of novel faults in a target domain while enabling generalization between domains.

Furthermore, while contrastive learning has been applied to obtain generalization for fault diagnostics models, it has not yet been attempted to apply it for fault detection. However, the same type of generalization towards variability in the healthy class caused by domain shifts is desirable for fault detection models. Michau and Fink (2021) worked on this task that enables the transfer of anomaly detection models between a source and a target domain via adversarial alignment. The generalization of anomaly detection models to unforeseen domains, however, was not investigated much yet. While the contrastive learning has shown to be useful for fault diagnostics, the transfer is not straightforward to an anomaly detection setting where only healthy conditions are available and faulty conditions are missing. In absence of any fault data, it is not possible to explicitly train a feature space that positions dissimilar health conditions far apart in the feature space. Developing an adaptation of contrastive learning in absence of fault data is subject of this dissertation.

Generative Models - Generative Adversarial Networks A generative adversarial network (GAN) is a model that can generate synthetic data and is trained in an adversarial manner (Goodfellow et al., 2020). In its basic form, it can be depicted as shown in Figure 2.1.

The generator model (blue model in Figure 2.1) is trained to map a noise sample to a synthetic data sample (adversarial samples) that can "fool" the discriminator. The discriminator is trained to distinguish real data samples from synthetic ones (Creswell et al., 2018). Over the years, different architectures and extensions of GANs have been proposed such as the conditional GAN (cGAN) (Mirza and Osindero, 2014) or the Wasserstein GAN with gradient penalty (Gulrajani et al., 2017). The latter has especially raised a lot of attention as it has led to a more stable optimization process compared to other implementations. The discriminator of the Wasserstein GAN approximates the Wasserstein distance of the real data samples to the generated ones.

In the context of PHM applications, GANs have been used to tackle the challenge of data scarcity resp. class imbalance in training datasets. Underrepresented class data has been synthetically generated to get a better representation of the classes (Luo et al., 2021; Zareapoor et al., 2021). Despite the great achievements in providing richer class representations, GANs have mainly been used to generate data that is similar to the training data i.e. drawn from the same data distribution. So far, GANs have not been able to generate data that is drawn from a distinct distribution or domain that was not observed before. Thus, these approaches are not suited to generate faults that are recorded e.g. under a novel operating condition.

Data generative models have also been used to enable domain adaptation (Hong et al., 2018; Bousmalis et al., 2017; Menke et al., 2022) as well as domain generalization (Zhou et al., 2020a,b) - mainly in the field of computer vision. For domain adaptation, the generative models are, for example, trained to transfer source images to target images. The generated target data is then labeled by inheriting the labels from the source domain (Bousmalis et al., 2017). For domain generalization, there have been very recent approaches aiming to generate out-of-distribution data to improve the generalizability of the models (Zhou et al., 2020a,b). These approaches, however, are not suited to generate data that is specific to a distinct previously unobserved domain but rather data that is different from the observed data or interpolates between existing domains. Thus, the approaches are not suited to generate faults in a specific domain, where they have not been observed before.

3 Proposed Framework and Contribution

3.1 Framework

Over the life cycle of a monitoring system for industrial assets, the data availability resp. the data scarcity changes and with it the requirement and expectation on the solution for condition assessment. In this thesis, we propose a framework including four different modules. Each of the modules addresses an open research question for different phases of the life cycle of a monitoring system and is developed to overcome current limitations of the specific data scarcity or label quality setting in the respective phase (see Section 1.3). By developing methods that can perform well in more extreme and realistic cases of data scarcity or label quality, each module helps reaching a higher maturity level of a condition assessment solution with less information (data or labels) available. In other words, the proposed framework enables to progressively increase the maturity level of condition assessment solution within a shorter amount of time compared to existing approaches.

The framework is shown in Figure 3.1. On the left, a railway system is used as an example to represent the different life cycles of condition monitoring. Railway systems exemplify the complexity of industrial assets well as they have long life times and operate in an open environment i.e. are exposed to changing environmental and operating conditions. Moreover, their fleets show a high diversity of configurations. The railway infrastructure system is equipped with wayside monitoring devices (green icons in Figure 3.1). Thus, the monitoring devices are exposed to a variety of non-informative factors such as different fleets, different environmental conditions and different operating conditions causing shifts in the underlying data distribution (domain shifts). On the right in Figure 3.1, different phases of development and deployment are shown.

In the initial phase of condition monitoring, the available condition monitoring data is typically unlabeled but is assumed to originate from healthy conditions (illustrated as the grey boxes in the lower row in Figure 3.1). Anomaly detection models can be developed, that are able to detect any anomalous pattern in the data. In this phase, ***Research Question 4*** arises since some novel or anomalous variations might not originate from a change in the health condition but might originate from a change in the non-informative factors that have not yet been observed such as novel operating conditions. For reliable operations, it is important to distinguish anomalies caused by a change in the non-informative factors from anomalies caused by a change in the health condition to prevent a high false alarm rate or missed faults. This challenge is addressed in the *'No Fault Label'* module.

In a latter phase of condition monitoring, some faults might have occurred (second lowest row in Figure 3.1). If they were successfully detected and the corresponding data was labeled e.g. by domain experts, a classification model can be implemented for fault diagnostics, enabling also the distinction between different fault types and potentially also severity levels. This therefore provides a more mature solution to condition assessment compared to a fault detection model. Because condition monitoring data is often affected by label noise, ***Research Question 3*** needs to be addressed on how to enable robust training despite the presence of uncharacterized label noise. This is addressed in the *Label Noise* module.

If a robust fault diagnostics model for one operating condition or one unit of a fleet can be developed, it may be required to transfer the existing model to different fleets or to new operating conditions. From a data perspective, this means that a model needs to be transferred to another domain (second upper row in Figure 3.1). For condition monitoring datasets, often specific challenges apply when adapting models to new domains. One challenge arises when

only the healthy class is shared between the domains. In this case, there exist many possible solutions to adapt the model resp. to align the two data distributions from the source and target domain. Moreover, it is not possible to evaluate the fitness of the individual adaptation solution in absence of fault classes that are represented in both domains. This challenge is particularly pronounced when another challenge applies: If the domain gap is large and a big adaptation step needs to be performed. From these challenges, ***Research Question 2*** arises. To enable the transfer of diagnostics models also under these extreme settings of label space discrepancies, we propose the *'Extreme Domain Adaptation'* module.

Lastly, it is desirable to reach a mature level of condition assessment solutions that does not require intervention each time a new domain shift occurs that is caused e.g. by a change in the operating conditions. Instead, the trained model should generalize well to unforeseen domain shifts (upper row in Figure 3.1). Simultaneously, it is utterly important to remain sensitive to novel health conditions to enable safe operations. To satisfy both of these requirements, ***Research Question 1*** needs to be addressed. We propose the *'Domain Generalization'* module to achieve that.

3.2 Modules

The proposed framework comprises four different modules with their corresponding methods. In the following, the modules are briefly introduced. Some more detailed information on the modules can be found in Section 3.3 and ultimately, the proposed methods is presented in Chapter 4-Chapter 7.

1. Module 1: *Domain Generalization module* aims to enable not only reliable diagnostics in known source domains (S_i as defined in Definition 4) but also in novel domains (S_{test} as defined in Definition 4) that have not been observed before. The proposed solution integrates not only (1) the reliable distinction of known health conditions but also (2) the detection of novel faults. Contrastive learning is proposed to achieve both tasks simultaneously. The module provides a mature solution for reliable condition assessment as it does not require model adjustments each time a domain shift occurred but is still sensitive to novel health conditions. The module is illustrated in Figure 3.2a.

2. Module 2: *Extreme Domain Adaptation module* is able to transfer models to new domains given extreme label space discrepancy i.e. if only the healthy class is shared between the domains and the domain gaps are potentially large. We propose to address the challenge raised by the label space discrepancies for domain adaptation by enabling the generation of domain- and class-specific data from fault conditions that have not been observed before in the target domain (T as defined in Definition 3). The generated fault data can compensate for unseen domain-specific fault classes and, thereby, transform the given *Partial* or *Open-Partial* DA setting into a *ClosedSet* DA setting. The module is illustrated in Figure 3.2b.

3. Module 3: *Label Noise module* addresses the real world challenge that often condition monitoring datasets are affected by label noise. Typically, there is no additional information available about the characteristics of the noise available and no clean and correctly labeled validation dataset. In this module, we propose to enable robust fault diagnostics in the presence of label noise without requiring any concrete knowledge of the label noise. Instead, our proposed method relies on a rough assumption regarding the level of label noise solely. We propose a two-step method that first identifies outliers based on the samples' consistency with the hypothesis update and second, modifies the training dataset based on the identified outlier samples. The module is illustrated in Figure 3.2c.

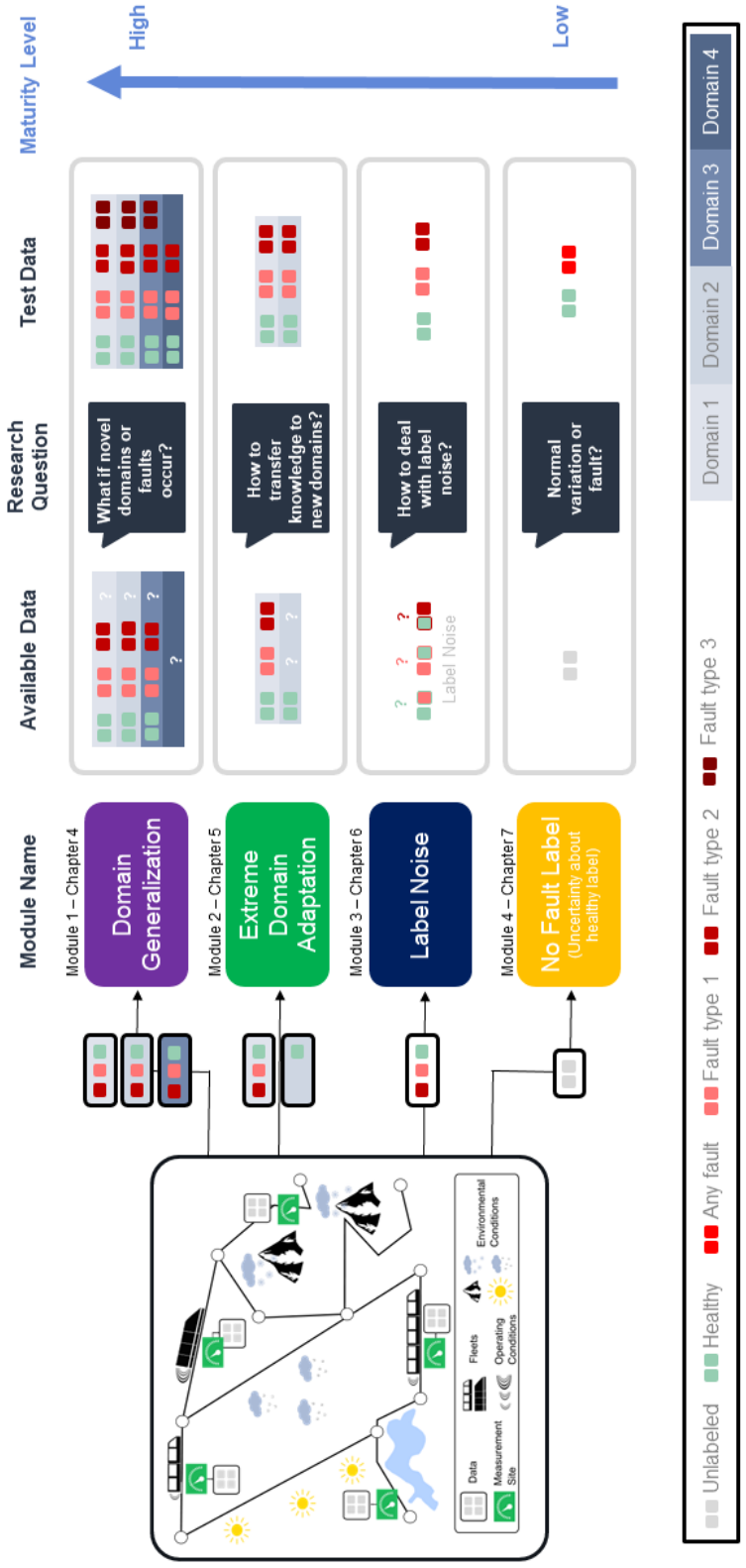
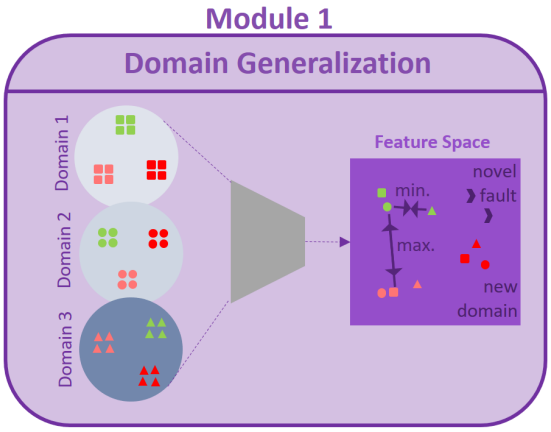
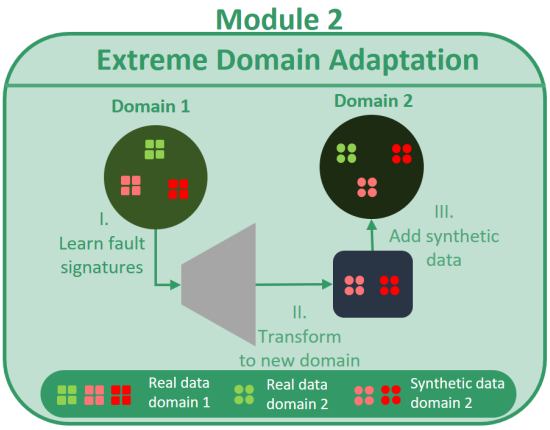


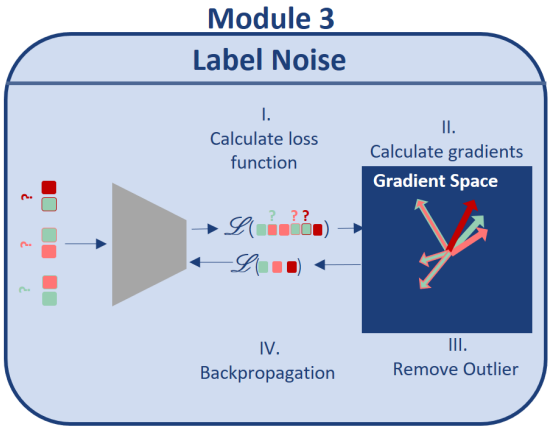
Figure 3.1: Overview of the proposed framework which deals with various real-world limitations. The modules of the framework are flexible applicable in different life cycle phases and with different levels of data and label scarcity. The framework is applicable in different types of assets that face the same challenges as addressed in the modules.



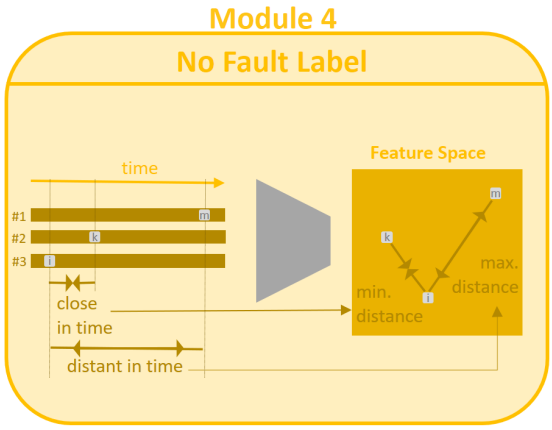
(a) Module 1 - Chapter 4: Contrastive Learning for Fault Detection and Diagnostics in the Context of Changing Operating Conditions and Novel Fault Types



(b) Module 2 - Chapter 5: Controlled Generation of Unseen Faults for *Partial* and *Open-Partial* Domain Adaptation



(c) Module 3 - Chapter 6: Improving generalization of deep fault detection models in the presence of mislabeled data



(d) Module 4 - Chapter 7: Contrastive Feature Learning for Fault Detection and Diagnostics in Railway Applications

Figure 3.2: Overview of the proposed domain adaptation framework which deals with various real-world limitations.

4. Module 4: *No Fault Label module* aims to achieve robustness towards normal variations in the healthy data that is caused by a change in the operational environment such as changing operational conditions. Simultaneously, the module aims to train a model that is particularly sensitive to possible future faults although these have not been observed yet. Unsupervised contrastive learning has been proposed to achieve that. The proposed module can be applied when the exact health condition is not known but presumably only healthy data has been observed and, therefore, provides a stable solution in the beginning of the condition monitoring phase. The module is illustrated in Figure 3.2d.

The different modules enable to progressively reach a higher maturity level of the condition assessment solution within a shorter period of time while enabling more reliable operations compared to state of the art methods. The following advantages can be achieved under this framework:

- Chapter 4: The ability to simultaneously provide a fault diagnostics solution that is robust under unforeseen domain shifts, robust to label noise while being sensitive to detecting new health conditions.

- Chapter 5: The ability to transfer fault knowledge between domains and enable domain adaptation under extreme label space discrepancies.
- Chapter 6: The ability to achieve robust fault diagnostics under unknown label noise types and levels.
- Chapter 7: The ability to achieve robustness towards normal variations within the healthy class for fault detection models and sharpening the sensitivity to potential future faults.

3.3 Contributions

This cumulative thesis incorporates four published articles in the fields of deep learning, as well as its applications in fault diagnosis. The key papers are included in Chapters 4 - 7. A summary of the methodology proposed in those articles and specific contributions is described in subsections 3.3.1 - 3.3.4 and additional contributions are described in subsection 3.3.5. A list with all publications can be found in Section 3.6.

3.3.1 Module 1: Contrastive Learning for Fault Detection and Diagnostics in the Context of Changing Operating Conditions and Novel Fault Types

To enable safe operations, not only the robustness to domain shifts is important but also the ability of the solution to be able to detect novel health conditions. These are somewhat competing objectives where (1) robustness to novel data variations due to domain shifts is required but (2) also sensitivity to novel variations in the data due to a change in the health condition. We address this challenge of achieving the two above mentioned objectives simultaneously by proposing a contrastive learning algorithm. A feature representation is learned such that data samples that were recorded under the same health conditions are positioned close to each other in the feature space. This corresponds to the positive pair x_a and x_p as defined in Equation 2.3 and is illustrated in Figure 3.2a by the green icons. *Vica versa*, samples recorded under different health conditions correspond to the negative pair x_a and x_n as defined in Equation 2.3 and are positioned far apart in the feature space as illustrated in Figure 3.2a.

Specific Contributions

- The proposed method does not require an explicit distinction of different source domains nor do the domains need to be discrete.
- Considering the operational environment of an industrial asset, the proposed method does not only enable generalization to new domains but also enables to detect novel faults.
- The proposed method outperforms the comparison methods on both tasks (1) classifying known health conditions under domain shifts and (2) detecting novel faults on a bearing dataset.

3.3.2 Module 2: Controlled Generation of Unseen Faults for Partial and Open-Partial Domain Adaptation

To enable domain adaptation under realistic scenarios in PHM, the respective method also needs to be applicable in extreme settings i.e. with extreme label space discrepancies between the source domain (S) and the target domain (T) as well as with large domain gaps i.e. when the underlying data distributions of the source domain and target domain differ considerably. The proposed method enables the generation of domain- and class-specific data samples from fault conditions that have not been observed before in the target domain. The generated fault data can compensate for unseen domain-specific fault classes and, thereby, transform the given *Partial* or *Open-Partial* DA setting into a *ClosedSet* DA as illustrated in Figure 3.2b, where the target domain is enhanced with synthetic data. Thus, it is particularly suited for domain adaptation under extreme label space discrepancies and large domain gaps.

Specific Contributions

- The proposed method enables the controlled generation of data that has not been observed before by adapting data to the specificities of a desired domain as well as to a desired fault class. Thus, unsupervised domain mapping is enabled.

- The proposed method is applicable even if no validation dataset of unobserved data is available and thus, contrary to other approaches proposed in literature, satisfies real-world requirements.
- The proposed method is applicable in different settings of label space discrepancies.
- The proposed method is evaluated on two bearing datasets with different domain gap sizes and outperforms comparison methods significantly in cases of large domain gaps.

3.3.3 Module 3: Improving generalization of deep fault detection models in the presence of mislabeled data

The availability of either ground truth information on the type or amount of label noise or a clean validation dataset is not realistic in condition monitoring datasets. Previous approaches that aim to enable robust optimization of deep learning methods in presence of label noise typically require prior knowledge on label noise or a clean validation datasets and thus, are not applicable to many PHM applications. We aim to lift the above mentioned requirements by developing a method that can perform well only based on a rough estimate of the label noise level. Instead of investigating the model's output of the classification model compared to the different ground truth classes, we propose to investigate the gradient space to prevent potential independent of the class characteristics as illustrated in Figure 3.2c.

Specific Contributions

- The proposed methodology requires solely a rough estimate on the label noise and does neither require exact prior knowledge on the type or amount of label noise nor a clean validation dataset. Therefore, it satisfies the requirements of real applications more.
- The benefits of the proposed methodology are evaluated on both, a computer vision and a condition monitoring dataset. Both, the computer vision and fault diagnosis experiments demonstrate the effectiveness of the proposed method, even under severe label noise levels.

3.3.4 Module 4: Contrastive Feature Learning for Fault Detection and Diagnostics in Railway Applications

It is crucial to prevent anomaly detection models to be too sensitive to normal variations in the healthy class and consequently, raise false alarms. Equally so, anomaly detection need to be very sensitive to novel variations in the data that are caused by a fault. Achieving both objectives of (1) invariance to normal variations and (2) sensitivity to possible faults is a quite challenging task in the anomaly detection setup i.e. in absence of any fault data as the sensitivity to faults can not be imposed directly. We approach this challenge by inducing an encoding of the data in the feature space that aims to achieve invariance to normal variations and a high sensitivity to changes in the health condition. We propose to apply contrastive learning, whereby we enabled the application of contrastive learning in an anomaly detection setup by using the 'time passed since the last maintenance action' as a surrogate to describe the health condition (see Figure 3.2d).

Specific Contributions

- We evaluate the proposed methodology on real condition monitoring datasets contrary to a dataset recorded under laboratory conditions, that only partially reflect the difficulties of real condition monitoring data recorded under real in-service conditions.
- The experiments demonstrate the superiority of contrastive learning in both, a supervised and an anomaly detection setup.

3.3.5 Additional Contributions

In addition to the specific contributions elaborated above, we also address more general limitations that currently hinder the application of data-driven models for PHM. In the following we will elaborate on the contributions that this thesis provides to overcome these general limitations.

Hyperparameter Tuning without Access to a Representative Validation Dataset. Many methods based on deep learning models require the tuning of a multitude of hyperparameters. Tuning these optimally influences the performance of the final model significantly. Due to the importance of finding optimal hyperparameters, many methods rely on available additional information. For example, many methods that tackle the challenge of label noise rely on previous knowledge about the type and amount of label noise or a clean validation dataset. Moreover, domain adaptation methods that were proposed in cases with discrepancies in the label spaces require the availability of a labeled and representative validation dataset of the target domain including the missing fault classes. This additional information is not available for tasks of real industrial applications. For example, if a fault has not been observed before in a specific target domain, a representative validation dataset including the missing fault classes does not exist in a realistic scenario. This limits the application of many existing methods to real applications where neither ground truth information nor a representative validation datasets are typically available. In this dissertation, we aim to develop methods that can be applied in absence of this additional information.

Extreme Data Scarcity. Many methods developed in literature addressed data scarcity challenges before. In PHM tasks, however, we are often faced with extreme cases of data scarcity, where not only domain data is scarcely represented but also the fault data representation is extremely scarce. For example, if we want to transfer fault diagnostics models to a new domain, often not the same fault types occurred in the source and target domain. Often, the representation of potential fault types and severity levels is scarce in either one or both domains. This can pose a critical challenge for existing domain adaptation methods, especially in the extreme case that only the healthy class is shared between the domains and the domain gap is large (fault scarcity). Otherwise, existing anomaly detection methods take into account that fault data is not only scarce but often not available at all in safety critical systems. However, it has hardly been addressed that, in addition to the missing fault data, also the normal variations within the healthy class is only scarcely represented (domain scarcity). An anomaly detection model might raise false alarms if there is a unforeseen domain shift at deployment time, which is not desired from a operational point of view. *Vica versa*, domain generalization methods in the context of fault diagnostics do consider that the available training dataset might only scarcely represent the conditions under which the asset is operated in the future and thus, unforeseen or unknown domain shifts might occur at deployment time. However, contrary to anomaly detection models, domain generalization methods do not take into account that also the health conditions might be scarcely represented as well and thus, novel health condition might emerge at deployment time (domain and fault scarcity). In this dissertation, we push the boundaries of deep learning under data scarcity and aim to alleviate the limitations of existing approaches.

3.4 Aim, Scope and Thesis Outline

The aim of this research is *to develop reliable fault detection and diagnostics models which can efficiently deal with various types of label and data scarcity that are specific to different phases of the life cycle of an asset or the life cycle of the installed measurement system.* This thesis is dedicated to filling existing research gaps elaborated in Section 1.3 that hinder the

reliable application of deep learning methods. The remainder of the dissertation is organized as follows:

Chapter 4 addresses the research question: *How can we train a fault diagnostics model that is both, able to perform well on known and unknown domains as well as able to detect novel fault types?*. We propose contrastive learning to achieve the two objectives: (1) robustness towards domain shifts and (2) sensitivity to novel faults. The proposed method is evaluated on a benchmark bearing dataset. This chapter enables condition monitoring with data-driven models over multiple known and unknown domains and enables to distinguish known and unknown faults. The method in this chapter can be applied if the monitoring of a complex system has advanced and presents a mature solution to condition assessment as it does not require to retrain a fault diagnostics model each time a domain shift occurs but it still enables safe operations since novel faults can be detected.

Chapter 5 addresses the research question: *How can we enable universal domain adaptation for fault diagnostics models with extreme label space discrepancies and large domain gaps?* This chapter further extends the framework to the scenario where a fault diagnostics model needs to be adapted to be applicable in new domains - for new fleets or significantly different operating conditions. To address more realistic scenarios in condition monitoring compared to previously proposed domain adaptation methods, the data generative method developed in 4 is particularly suited for large domain gaps and is applicable in extreme label space discrepancies i.e. if only the healthy class is shared between two domain. The superiority of the developed method is demonstrated on two benchmark bearing datasets with different sizes of domain gaps.

Chapter 6 addresses the research question: *How can effective fault diagnostics be enabled in the presence of label noise if no preliminary knowledge about the amount of label noise and no clean validation dataset is available?* It can be applied in a condition monitoring phase, where faults have occurred, were detected and labels are available. However, the quality of the labels is not consistent and it is suspected that label noise is affecting the dataset. A method is proposed in Section 6 that does not rely on concrete knowledge about the label noise but relies solely on a rough estimation of the level of label noise. Therefore, the proposed method addresses a more realistic setup compared to other approaches proposed in literature. The proposed method considers the gradient space before updating the neural network and thus, prevents overfitting to mislabeled samples from the beginning. Experiments on an image classification task and a condition monitoring dataset demonstrate that the proposed method results in robust classification models also, under large levels of label noise.

Chapter 7 addresses the research question: *How to concurrently achieve invariance to non-informative factors and sensitivity to fault types for fault diagnostics but also for fault detection, where only healthy data and no fault data is available?* To address this question, contrastive feature learning is adapted in Chapter 7 to be suited not only for fault diagnostics but also for the anomaly detection setup where only the healthy class is available at training time. This enables to transfer the benefits of contrastive learning of learning generalizable features from the fault diagnostics setup to the fault detection setup. In the anomaly detection setup, invariance to operational conditions and sensitivity to degradation processes is imposed. It is evaluated to which extent sensitivity to degradation can be transferred to sensitivity to faults. Experiments are conducted on two real condition monitoring datasets within the railway system. The dataset cover two different settings: fault diagnostics and fault detection.

Chapter 8 discusses the key findings in the individual works.

Chapter 9 completes this thesis with conclusions and an outlook for future research possibilities.

The four developed approaches in Chapter 4-7 mitigate the current limitations in applying deep learning to PHM problems that have been elaborated earlier. They have been developed

to suit the different phases when introducing a new monitoring solution for an industrial asset: from early deployment when no faults and little operating conditions have yet been observed to a mature fault detection and diagnostics model that is robust towards changes in the non-informative factors (domain shifts) and simultaneously sensitive to novel fault types. The four developed approaches can be employed consecutively and compose a framework as described in Section 3.1 that enables reliable operation in the different phases of monitoring an asset. The developed methods will ultimately allow to reach a mature solution for condition assessment while enabling safe and efficient operations on the way of reaching a high level of maturity (see Figure 3.1).

3.5 Relevance to Science and Economy

In this thesis, we provide a holistic framework that can be applied to different real-world industrial assets for robust and reliable fault detection and diagnostics. For each of the proposed modules, we integrate real world constraints. In order to incorporate artificial intelligence into real-world problems methods are required that are able to learn from little data and with as little supervision as possible. From a scientific perspective, we need to develop methods that perform well in data scarcity settings that are common the the context of real industrial applications and thus, differ from traditional deep learning applications where representative and large-scale datasets are available. For example, by developing a method that enables the generation of previously unobserved data in a controlled and directed manner in Chapter 6, domain adaptation can be applied even if no faults were observed in the target domain yet. Developing methods that can overcome the specific challenges that arise under different data scarcity settings is one of the main goals of this thesis.

The methodological advances have great implications for economy. By developing methods that can perform reliably with less amount of information or data available, the condition assessment can be advanced quicker compared with other state-of-the-art methods as less data acquisition time is required to gather a more representative dataset. Furthermore, increasing the model's robustness towards e.g. distributional shifts in the data makes the use of deep learning methods much more attractive to the industry as they need less adaptation or human supervision. Each of the modules can be applied individually and flexible for problems where data scarcities apply, as it is the case in many real-world applications. Moreover, the modules applied progressively within one framework provide solutions for practitioners that just started to monitor an industrial asset and have not gained much experience on the system or the data yet. The framework will guide the practitioner to increase the maturity level of the condition monitoring solution within a shorter period of time compared to other methods. By providing reliable models for fault detection and diagnostics on a component level, the modules resp. the framework do not only enable safe operations but can significantly support efficient maintenance planning.

3.6 Publications

The following works are published or submitted during my doctoral study:

- Chapter 4: Rombach, Katharina, Gabriel Michau, and Olga Fink (2021). "Contrastive Learning for Fault Detection and Diagnostics in the Context of Changing Operating Conditions and Novel Fault Types". In: *Sensors* 21.10, p. 3550. DOI: https://doi.org/10.3390/s21103550
- Chapter 5: Rombach, Katharina, Gabriel Michau, and Olga Fink (2023). "Controlled generation of unseen faults for Partial and Open-Partial domain adaptation". In: *Reliability Engineering & System Safety* 230, p. 108857. DOI: https://doi.org/10.1016/j.ress.2022.108857
- Chapter 6: Rombach, Katharina, Gabriel Michau, and Olga Fink (2020). "Improving generalization of deep fault detection models in the presence of mislabeled data". In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, pp. 3103–3110. DOI: https://doi.org/10.1109/SMC42975.2020.9283002
- Chapter 7: Rombach, Katharina and Michau, Gabriel and Ratnasabapathy, Kajan and Ancu, Lucian-Stefan and Bürzle, Wilfried and Koller, Stefan and Fink, Olga. "Contrastive Feature Learning for Fault Detection and Diagnostics in Railway Applications." *under review in scientific journal.* 2022.

Besides, there are an additional works published during the doctoral study:

- Rombach, Katharina, Gabriel Michau, Kajan Ratnasabapathy, Lucian-Stefan Ancu, Wilfried Bürzle, Stefan Koller, and Olga Fink (2022). "Contrastive feature learning for railway infrastructure fault diagnostic". In: *32nd European Safety and Reliability Conference (ESREL 2022)*

4 Contrastive Learning for Fault Detection and Diagnostics in the Context of Changing Operating Conditions and Novel Fault Types

This chapter corresponds to the published article:[1]

Rombach, Katharina, Gabriel Michau, and Olga Fink (2021). "Contrastive Learning for Fault Detection and Diagnostics in the Context of Changing Operating Conditions and Novel Fault Types". In: *Sensors* 21.10, p. 3550. DOI: https://doi.org/10.3390/s21103550.

Abstract: Reliable fault detection and diagnostics are crucial in order to ensure efficient operations in industrial assets. Data-driven solutions have shown great potential in various fields but pose many challenges in Prognostics and Health Management (PHM) applications: Changing external in-service factors and operating conditions cause variations in the condition monitoring (CM) data resulting in false alarms. Furthermore, novel types of faults can also cause variations in condition monitoring data. Since faults occur rarely in complex safety critical systems, a training dataset typically does not cover all possible fault types. To enable the detection of novel fault types, the models need to be sensitive to novel variations. Simultaneously, to decrease the false alarm rate, invariance to variations in CM data caused by changing operating conditions is required. We propose contrastive learning for the task of fault detection and diagnostics in the context of changing operating conditions and novel fault types. In particular, we evaluate how a feature representation trained by the triplet loss is suited to fault detection and diagnostics under the aforementioned conditions. We showcase that classification and clustering based on the learned feature representations are 1) invariant to changing operating conditions while also being 2) suited to the detection of novel fault types. Our evaluation is conducted on the bearing benchmark dataset provided by the Case Western Reserve University (CWRU).

4.1 Introduction

Modern industrial processes are increasingly subject to oversight by condition monitoring (CM) devices. The recorded data opens up the possibility of data-driven maintenance models (Fink, 2020). Purely data-driven solutions are especially interesting with regard to complex assets for which model-based approaches are limited or don't exist. Recent successes in deep learning have demonstrated the potential of data-driven solutions (Devlin et al., 2019; Chen et al., 2020). However, for the task of fault detection and diagnostics, particular challenges arise when applying deep learning to CM data from an industrial asset.

Complex industrial assets are often subject to a variety of operating conditions as well as external (e.g. environmental) factors that strongly influence the acquired data. Changing ambient temperature, for example, might affect the roughness of the asset, which could then be sensed by accelerometer measurements resulting in changes of the signals. The ambient

[1] Please note, this is the author's version of the manuscript published in *Sensors*. Changes resulting from the publishing process, namely editing, corrections, final formatting for printed or online publication, and other modifications resulting from quality control procedures may have been subsequently added. The final publication is available at https://doi.org/10.3390/s21103550.

temperature is therefore a factor that causes variations in the data but cannot be controlled. That means that a complete training dataset that is recorded in summer will deviate from the data experienced in the winter season. Predicting or foreseeing all of these influential factors is not always possible as some factors of variations are simply not known or cannot be controlled. Even if all future operating conditions are completely controllable and known (e.g. defined in the specifications of a working environment), the multitude of possible combinations makes it often infeasible to collect a dataset with a sufficient representation of all possible combinations of operating conditions within the specifications. Hence, a training dataset might only represent a subset of all possible conditions. Ultimately, often in real applications, it is not realistic to assume that a training dataset contains all possible future conditions that the asset will experience (Michau and Fink, 2021). In this paper, we distinguish between conditions or factors that are represented in the training dataset and those that are not. The later ones are referred to as **novel** operating conditions. Yet, the performance of data-driven models often relies on the fact that the data collected during inference time is similar to the training dataset (independent and identically distributed (IID)) (Tan et al., 2018). I.e. the training dataset needs to be representative of all ambient factors and operating conditions the asset will encounter in the future. If a model is subjected to new variations in the data caused by, e.g. unexpected ranges of ambient temperature, it might perform poorly in identifying the exact system condition of the asset (Fink et al., 2020). This can result in false alarms. To prevent this, a fault diagnostic model needs to be invariant to all variations in the data that correspond solely to varying operational or environmental factors rather than to a change in the asset's condition.

On the other hand, while faults arise very rarely in operating industrial assets, there is a multitude of different fault types with various severities that can possibly occur (Michau and Fink, 2021). It is not realistic to assume that the training dataset contains all possible fault types at all possible intensities. However, robust fault diagnostics entails the task of identifying fault types in general. This includes those faults that are unknown at training time and, therefore, are not represented in the training dataset. Similarly to the terminology used for operating conditions that are not reflected in a training dataset, we refer to these faults as **novel** fault types. A safety issue can arise if a model is not capable of detecting novel fault types or is underestimating a fault's severity. Therefore, to ensure safe operation, a robust fault diagnostic model needs to be sensitive to novel variations in the data that correspond to novel fault types.

Ultimately, the goal is to train a fault diagnostics model that is both invariant to the variability in the CM data caused by novel operating conditions or external factors and, simultaneously, sensitive to the changes corresponding to novel fault types that were not considered or known when the model was developed. In this work, we show that features trained with contrastive learning are able to achieve both of the aforementioned objectives. This is the first work that applies contrastive learning to PHM applications in order to tackle both of the above objectives: 1) invariance of the models to novel operating conditions and 2) sensitivity of the models with respect to variations caused by novel fault types.

4.2 Related Work

Contrastive learning is a discriminative approach that aims to group semantically similar samples close to each other in the feature space while pushing semantically dissimilar samples far apart from each other (Jaiswal et al., 2021; Hermans et al., 2017). To achieve this, a contrastive loss is formulated based on a similarity metric quantifying how close different features are (Hadsell et al., 2006). In contrast to other frequently used losses - such as cross-entropy loss or mean squared error loss, whose objective is to directly predict a label or values - contrasive learning aims to train a semantically meaningful feature representation of the data. This has recently shown great promise, mainly in the context of computer vision,

achieving or exceeding state-of-the-art results in both a supervised (Hadsell et al., 2006; Gomez et al., 2018; Hermans et al., 2017) and unsupervised setting (Dosovitskiy et al., 2014; Oord et al., 2018; Bachman et al., 2019). Franceschi et al. applied contrastive learning also successfully to timeseries data (Franceschi et al., 2019).

If the contrastive loss function is based on triplets of training data samples, it is referred to as **triplet loss**. The idea of using data triplets (instead of data pairs) for contrastive learning was first introduced in 2009 for nearest-neighbor classification (Weinberger and Saul, 2009). For each sample (the "anchor" x_a), the distance to both a positive sample (x_p) and a negative one (x_n) is calculated in order to formulate the loss function. Different techniques have been proposed to select these positive and negative samples. For supervised tasks, for example, the hard triplet loss (Hermans et al., 2017) chooses the sample with the same label that is farthest away from the anchor (x_a) as the positive sample. Whereas, the nearest sample with a different label is selected as the negative sample. By contrast, the soft margin loss function (Schroff et al., 2015) randomly selects a negative sample and regards all samples with the same labels within the batch as the positives. Regardless of the exact implementation, the objective is to group data with the same label and increase the distance to other classes of data in the feature space, i.e. to give the feature clusters a semantic meaning.

Feature extraction or learning has been identified as one of the most important elements in PHM applications (Nguyen et al., 2018). Manually engineered features (feature extraction) as well as learned features (feature learning) have been proposed for the purpose of fault detection and diagnostics (Krummenacher et al., 2017; Chao et al., 2021). The resulting feature space is then classified (Patel and Giri, 2016; Krummenacher et al., 2017) or clustered (Yoon et al., 2017; Chao et al., 2021) in order to detect and classify faults and their severity, but also to detect novel fault types (Chao et al., 2021). Robust feature learning is the objective of many publications of fault diagnosis (Abid et al., 2019; Shen et al., 2018). These works typically focus on robustness with respect to noisy environments. That means they assume to have representative (but noisy) samples of all classes. On the contrary, this paper focuses on robustness with respect to a shift of the underlying data distribution e.g. caused by changing operating conditions. Contrastive learning has been applied in domain adaptation settings for PHM applications (see below) (Wang and Liu, 2020) but not yet for robust feature learning in the context of unknown changing conditions and novel fault detection. However, the idea of learning low-dimensional representations of high-dimensional data that correspond solely to their semantic meaning is very promising. It offers the potential to filter out variations of the data that are caused by changing conditions and do not contain information regarding the asset's condition.

Transfer learning in general relaxes the hypothesis that the training data must be IID with the test data (Tan et al., 2018). By transferring knowledge that is learned in source tasks to a related target task, it aims to alleviate the issue of insufficient training data (Torrey and Shavlik, 2010; Tan et al., 2018). This has attracted a lot attention in machinery fault diagnostics, where, for example, changing operating conditions or external factors cause a shift in the CM data that is not reflected in the training dataset (Li et al., 2020a). Means of domain adaption - a branch of transfer learning - have been widely used to address the challenge of adapting a model to new conditions (Wang et al., 2020a, 2019; Zhang et al., 2018; Lu et al., 2016). Noteworthy is the approach of Wang and Liu where contrastive learning is used for domain adaptation. However, these approaches require both a) a clear identification of the target domain and b) representative data for all classes from this target domain. Pioneering work by Wang et al. (Wang et al., 2020a) has enabled the application of domain adaptation even if certain class data (e.g. certain faults) is missing in the target domain. Nevertheless, it still requires to identify and foresee the target domain, which is not always possible (e.g. if these new conditions are caused by external factors that are neither known nor controllable). Further, representative data of all classes is required in the source

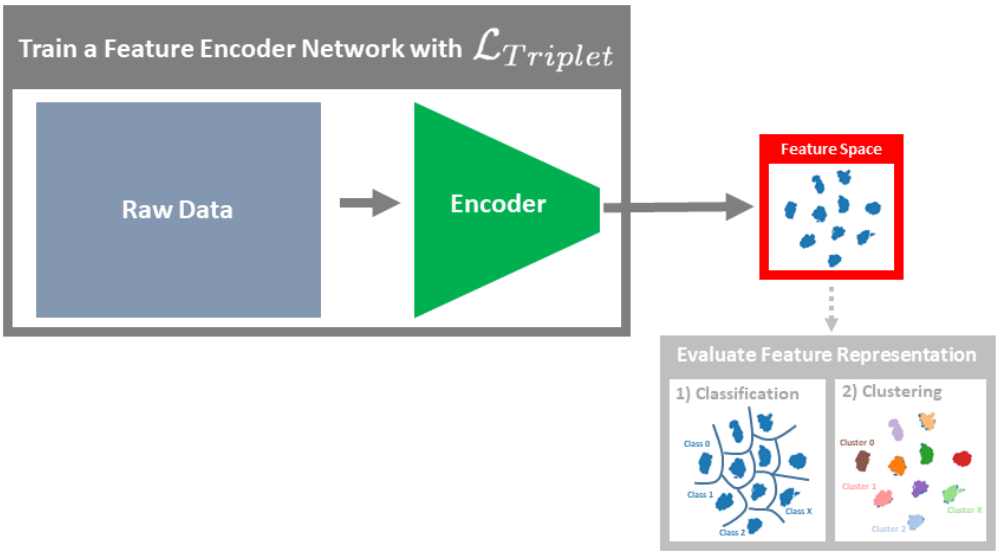
domain. This is not given if the novel emerging fault types are those that have not been anticipated before.

4.3 Methodology

Contrastive learning is evaluated in the context of the PHM application of detecting, classifying, and determining the type and severity of bearing faults. Specifically, we evaluate whether fault detection and diagnostics based on the learned feature representation is, on the one hand, invariant to variations in the CM data caused by novel operating conditions and, on the other hand, sensitive to variations caused by novel fault types. To achieve that, the retrieved features are both, classified and clustered. The feature representations are learned via the semi-hard implementation of the triplet loss $\mathcal{L}_{Triplet}$ (Schroff et al., 2015), where the negative loss is calculated based on one negative sample that is randomly sampled within a batch. The positive loss is computed based on the average distance of all positive samples within the batch to the anchor sample. The distance metric used for all case studies is the L2 Norm. The feature learning models are then applied to test datasets that contain novel operating conditions in Case Study 1 and in Case Study 2 the models are exposed to novel fault types.

To evaluate the suitability of the learned feature representation for detecting and classifying known fault types (but also for detecting novel fault types), the learned features are classified and clustered. A support vector machine (SVM) is trained for classification. The classification performance showcases whether the models are affected by a change in the operating conditions. For the identification of novel fault types, the feature space is clustered with two different clustering algorithms: *Ordering points to identify the clustering structure* (OPTICS) (Ankerst et al., 1999) as well as k-means (Lloyd, 1982), for which the silhouette score (Rousseeuw, 1987) is used to determine the number of clusters.

A scheme of the methodology can be seen in Figure 4.1.



(a) (b)

Figure 4.1: Methodology schemes of (**a**) training a feature representation with the triplet loss and (**b**) evaluating the learned feature representation (classification and clustering) with respect to the objectives of achieving invariance to novel operating conditions and sensitivity to novel faults.

4.4 Case Studies

4.4.1 Dataset

All case studies are conducted on a bearing dataset provided by the Case Western Reserve University Bearing Data Center (CWRU dataset) (Smith and Randall, 2015). The publicly available dataset is often used as a benchmark dataset in the field of PHM in general. It has been used for different tasks within the field of fault detection and diagnostics. Recently published methods include stacked denoising autoencoder (Lu et al., 2017) or recurrent neural networks (Shenfield and Howarth, 2020) (a comprehensive overview is given by Neupane and Seok (Neupane and Seok, 2020)). The dataset is especially suited to demonstrate solutions related to diagnosing faults under different operating conditions (different loads in this case) and transferring models between these different conditions (domain adaptation) (Wang et al., 2020a, 2019; Zhang et al., 2018; Lu et al., 2016).

However, we would like to emphasize that the setup that we are dealing with in this research, has not yet been tackled by other researchers: the algorithms we are seeking to develop are on the one hand supposed to be sensitive to novel types of faults, however on the other hand they are supposed to be robust to novel operating conditions. Unfortunately, there are no other case studies that could be used to compare our proposed approach to directly. In fact, we reformulate the problem setup to make it applicable to the problem of novel fault type detection. Therefore, previous results obtained on this dataset are also not directly comparable.

The accelerometer measurements are recorded under four different loads 0, 1, 2, 3, which correspond to different operating conditions in our case studies. Ten different health conditions of the bearing are represented in the dataset (see Table 4.1): Healthy condition (N), three different fault types (inner race faults [IR], outer race faults [OR], and ball faults [B]), and three different fault severities for each of the fault types (7, 14, 21). The sample dataset was collected from the CWRU dataset with sampling frequency of 48 kHz.

Class	0	1	2	3	4	5	6	7	8	9
Severity [mils]	-	7	7	7	14	14	14	21	21	21
Type	N	B	IR	OR	B	IR	OR	B	IR	OR

Table 4.1: Classes in the CWRU dataset.

Preprocessing: The original signals are divided into sequences of 512 points with no overlap between the sequences. Each sequence is scaled by the mean and standard deviation of the healthy data. This results in a dataset containing one-dimensional timeseries of length 512, each labeled by the label of the original signal.

The proposed algorithm and most of the baseline methods (see Section 4.4.3) uses raw signals as input data. However, we also compared the performance to that of algorithms based on feature engineering and used the frequently applied Fast Fourier Transform (FFT) for extracting features in the frequency domain (Heideman et al., 1984).

The FFT features are calculated based on the previously extracted timeseries dataset whereby the absolute value of the FFT coefficients is considered as the FFT features. Due to the symmetry of the resulting features, only the first half is considered, resulting in a 256-dimensional feature space.

4.4.2 Case Study Setup

Two case studies are conducted to evaluate the suitability of contrastive learning with respect to the objectives of achieving 1) invariance of the models similar but novel operating conditions (interpolation - see Experiment 1) as well as 2) sensitivity to novel fault types (extrapolation - Experiment 2). In the following, these objectives and their corresponding

setups are elaborated.

Case Study 1: Invariance to Novel Operating Conditions

This case study tests the invariance of the trained models to novel changes in the operating conditions. As defined in Section 4.1, novel operating conditions are those that are not represented in the training dataset. In the CWRU dataset, the different loads are considered as different operating conditions (see Section 4.4.1). The models are trained under a subset of operating conditions and evaluated on two test datasets: Data recorded under the same operating conditions as the training dataset ($\mathbb{T}$) and a second test dataset containing data recorded under the operating condition that was not part of the training dataset ($\mathbb{T}_p$). For example, if no data under the load 1 is available at training, the training dataset is defined as $\mathbb{D}_{train} = \mathbb{D}_{train}/\mathbb{D}_{load=1}$ (19,129 samples) and the two test datasets are defined as 1) $\mathbb{T} = \mathbb{D}_{test}/\mathbb{D}_{load=1}$ and 2) $\mathbb{T}_p = \mathbb{D}_{load=1}$. This case study setup corresponds to the scenario where a model experiences novel operating conditions or factors influencing the measurements during inference time that were not known at training time. The goal here is not to extrapolate to novel operating conditions but rather to train a feature representation that is not impacted by a shift in operating conditions. Therefore, the case study includes two data selections, whereby the two intermediate loads are being withhold for training. (This setup deviates from the typical experimental setup in the field of domain adaptation since we do not assume any knowledge about the missing conditions or target domain during training time.)

Case Study 2: Sensitivity to Novel Fault Types

To test the ability of the model to distinguish known fault types and severities from novel ones, a model is trained on a subset of fault types and evaluated on two test datasets: One containing the same subset of fault types as the training dataset ($\mathbb{T}$) and the second test dataset including the novel fault types that were not in the training dataset ($\mathbb{T}_p$). The CWRU dataset used in this research (see Section 4.4.1) allows for multiple data selection choices to evaluate the objective at hand. Two different exemplary data selections are chosen to evaluate the objective at hand. First, the fault B is withheld from the training dataset and second, the IR fault with all fault severities. For example, the first data selection results in a the training dataset $\mathbb{D}_{train} = \mathbb{D}_{train}/\mathbb{D}_{fault=1,4,7}$ (18,195 samples) and the test datasets $\mathbb{T} = \mathbb{D}_{test}/\mathbb{D}_{fault=1,4,7}$ (4,549 samples) and $\mathbb{T}_p = \mathbb{D}_{fault=1,4,7}$ (4,998 samples).

Evaluation: To evaluate the learned features with respect to the objective of achieving invariance to changing operating conditions, a classification model is trained based on the known classes at training time (see Section 4.4.4). To evaluate the objective of achieving sensitivity to novel fault types, the feature space of the test dataset containing the novel fault types is clustered. To evaluate the clustering performance, we closely follow the work of Chao et al. (Chao et al., 2021) by reporting the following metrics: **R**: the number of detected clusters; ***AMI***: the adjusted mutual information, measuring how closely the clustering algorithm replicates the true classes (Vinh et al., 2010); ***h***: the homogeneity, which indicates whether clusters contain only data points which are members of a single class; ***c***: the completeness, which measures whether members of a given class are elements of the same cluster (Rosenberg and Hirschberg, 2007). Furthermore, a two-dimensional t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008) is used for visualization of the feature representation with a fixed perplexity value of 100.

4.4.3 Baseline Methods

Contrastive learning results in models that provide an informative feature representation of the data. To evaluate the performance of the contrastive learning framework, we defined several baseline models with the focus on encoding features in the latent space with different types of learning setups, ranging from supervised learning to autoencoding architectures. Different loss functions are used to optimize the encoder network. First, an autoencoder is trained with the objective to reconstruct the input signal with the mean squared error loss.

The bottleneck layer activations provide the feature representation. Second, a classification model that is directly trained to predict the labels with cross-entropy loss. The latent space activations provide the feature representation. To provide a clear comparison for the evaluation of the different loss functions with respect to the different objectives, the same encoder model architecture is used for all the encoding models (the concrete choice is explained in Section 4.4.4). Third, experiments are also conducted on features extracted from the raw input signals: Fast Fourier Transform (FFT) coefficients. The forth evaluation model is an autoencoder architecture that is optimized with respect to the goal to ideally reconstruct the FFT coefficient and not the raw data.

4.4.4 Models

Encoder

A small latent feature space dimensionality is chosen arbitrarily with the purpose of creating a bottleneck that needs to select the most informative content and, thus, may help to remove some factors of variability. Therefore, the dimensionality of the latent feature space was set to 16. The feature encoders share the same architecture - with one exception (see below). The architecture was chosen such that good performance could be achieved on all training objectives given a feature space of 16 dimensions. The encoder network consists of four 1D-convolution layers (64, 32, 16, 8 kernels) with a kernel size of 12, activated with Leaky ReLu (alpha=0.5), followed by a MaxPooling (with strides of 2), and a Dropout layer (with a dropout rate of 0.1). The output of the convolution layers is flattened before passing it to a fully connected layer with 16 dimensions, again, activated by Leaky ReLu (alpha=0.5). The triplet encoder has an additional L2 normalization layer. The classifier is followed by a fully connected layer with number of classes in the training dataset and softmax activation. The autoencoder (AE) model is followed by a decoder model (reverse architecture of the encoder). To enable convergence, all models are trained with the Adam optimizer for 100 epochs and a batch size of 64.

The process to encode the FFT features is elaborated in Section 4.4.1. While the fixed, small feature space size allows for comparison of the different feature spaces, training an autoencoder successfully (minimizing the reconstruction error of the input signal) required an adaption of the model architecture. Additionally, it is beneficial to train it on the FFT features and not on the raw signals (as often done in literature (Neupane and Seok, 2020)). Therefore, a second autoencoder model to reconstuct the FFT features is trained to enable a fair comparison (see Section 4.4.3) with the following encoder architecture. It consists of four 1D-convolution layers (64, 32, 16, 8 kernels) with a kernel size of 12 and a stride of 2, activated with Leaky ReLu (alpha=0.5). The output of the convolution layers is flattened before passing it to a fully connected layer with 64 dimensions, again, activated with Leaky ReLu (alpha=0.5).

Classification

To evaluate the performance of the learned or extracted features, a supervised architecture was chosen that uses the learned or extracted features as input. It is important to highlight that supervised evaluations are not feasible for all the case studies. For the supervised evaluation case studies, an SVM with a Radial Basis Function kernel is trained based on the learned or extracted feature representations. For the supervised classifier, the outputs of the classifier are used directly without training an additional SVM on the learned features as in the case of the other two models. In Section 4.4.5, the specific hyperparameters are shown.

Clustering

Since particularly the discovery of novel fault types requires unsupervised evaluation of the feature space, clustering approaches were applied to the learned or extracted features. Two

different clustering methods are used for comparison purposes: a partitioning clustering approach and a density-based clustering approach.

The features of the classifier encoder and the AE are scaled by the mean value before applying the clustering.

OPTICS: the density-based algorithm *Ordering points to identify the clustering structure* uses a distance metric to group points that are close to each other. Compared to density-based spatial clustering of applications with noise (DBSCAN) (Ester et al., 1996), OPTICS allows for clusters of varying density. The utilized implementation deviates from the original OPTICS algorithm by first performing k-nearest-neighborhood searches on all points. This is then used to calculate core distances in order to identify core sizes. For details, please refer to (Buitinck et al., 2013). One benefit of using OPTICS is that it has the ability to detect "noisy samples" as outliers. These are samples that are not contained in any cluster as they are not density-reachable as defined in (Ankerst et al., 1999). This property is particularly useful for detecting novel fault types.

K-means + silhouette score: K-means is a clustering algorithm which assigns each sample to the cluster with the nearest mean (MacQueen, 1967). In our research, the number of clusters is determined by the silhouette score (Rousseeuw, 1987). It measures how similar an object is to its own cluster (cohesion) as compared to other clusters (separation) based on the euclidean distance.

4.4.5 Hyperparameter Tuning

The hyperparameters of the supervised classification algorithm SVM are tuned on a validation dataset split from the training dataset (see first columns in Table 4.2). Although the unsupervised clustering algorithms do not rely on the availability of labels, it is beneficial to tune certain hyperparameters. To do this, we again exploit the availability of the labeled training dataset: The minimum number of clusters considered for *Kmeans + Silhouette* was set to the number of classes in the training dataset (ten for case study 1 and seven for case study 2). The maximum number of clusters was set to a fixed value of 20. When applying *OPTICS*, the explicit clustering method can be chosen, as well as the minimal number of samples per class and the maximum distance between two samples for one to be considered as being in the neighborhood of the other ϵ. These parameters were tuned to achieve high performance on a fraction of the training dataset corresponding to the size of the dataset $\mathbb{T}$. Whenever possible, the smallest fixed value of ϵ was chosen such that an AMI of 98% was achieved in the fraction of the training dataset. Otherwise, the value was set to infinity. Each setting is shown in Table 4.2.

	Classification - SVM		Clustering - Exp. 1			Clustering - Exp. 2		
	C	γ	method	#	ϵ	method	#	ϵ
AE/AE$_{FFT}$/FFT	5.99	0.001	xi	10	∞	xi	10	∞
CLE	-	-	xi	10	∞	xi	10	∞
TE	1.67	0.046	DBSCAN	10	0.2	DBSCAN	10	0.08

Table 4.2: Classification and Clustering Hyperparameters Based on the Feature Spaces of the FFT, the Autoencoder based on the FFT (AE$_{FFT}$), the Autoencoder (AE), the Classifier Encoder (CLE), and Triplet Encoder (TE) Models

4.5 Results

4.5.1 Case Study 1: Invariance to Novel Operating Conditions

For visualization purposes, the 2-D t-SNE of the feature spaces of the models that share the same encoder architecture (AE, the classifier encoder and the triplet encoder) are shown with the true labels (y_{true}) on $\mathbb{T} \cup \mathbb{T}_p$ in Figure 4.2. Exemplary, the figures of sample selection 1 ($\mathbb{D}_{train} = \mathbb{D}_{train}/\mathbb{D}_{load=1}$) are displayed. Visually, the triplet encoder features appear to

cluster the different classes best (see Figure 4.2). All clusters are well separated, cohesive, and contain only one class of data. The silhouette score per class confirms the visual impression (see Table 4.3), as the feature space of the triplet encoder shows the highest silhouette score calculated on the true labels. However, a slight deviation is visible within the classes from $\mathbb{T}$ to $\mathbb{T}_p$.

FFT	AE_{FFT}	AE	CLE	TE
0.04	0.10	-0.18	0.38	**0.81**

Table 4.3: Silhoutte Score of the Class Clusters in the feature representation based on the FFT features (FFT), the autoencoder with FFT features (AE_{FFT}), the autoencoder (AE), classifier encoder (CLE) and triplet encoder (TE) on $\mathbb{T} \cup \mathbb{T}_p$

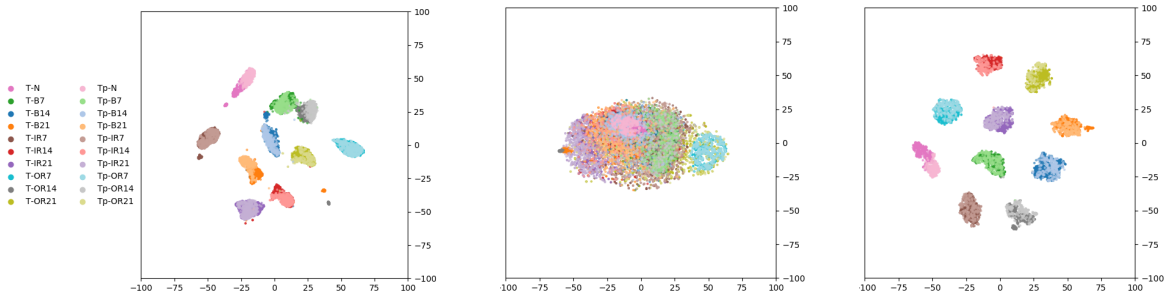
Hence, the classification performance based on the triplet encoder features is not impacted by the change in operating conditions of the different test datasets (accuracy of 100% on $\mathbb{T}$ and $\mathbb{T}_p$ on both of the sample selections - see classification results in Table 4.4). Similarly, the classification performance based on classifier encoder features is hardly impacted by the change of operating conditions - only a negligible performance drop of 1% is observed from $\mathbb{T}$ to $\mathbb{T}_p$ for both sample selections - see Table 4.4. On the contrary, all other models show a more significant accuracy drop from test dataset $\mathbb{T}$ to $\mathbb{T}_p$ on both sample selections (more pronounced in sample selection 1). This showcases the issue of changing operating condition for the fault diagnostic task and disqualifies these methods to be used in these scenarios of changing operating conditions.

Clustering methods are used for the second objective of detecting novel fault types (see Exp. 2). However, the clustering needs to perform well on $\mathbb{T} \cup \mathbb{T}_p$, even if no novel fault types - but rather only a shift in the operating conditions - is present. If this were not be the case, it would not be possible to distinguish between variations in the data due to changes in the operating conditions and the presence of novel fault types. In Table 4.4, the clustering performance on the respective feature representations is shown. It is apparent that only the clustering of the triplet encoder feature space is not impacted by the change in operating conditions. Hardly any performance change is observed between clustering based on $\mathbb{T}$ and clustering based on $\mathbb{T} \cup \mathbb{T}_p$ on the sample selection 1 ($\mathbb{D}_{train} = \mathbb{D}_{train}/\mathbb{D}_{load=1}$) - see clustering results in Table 4.4. For data selection 2 ($\mathbb{D}_{train} = \mathbb{D}_{train}/\mathbb{D}_{load=2}$), a slight change is observed when using OPTICS (AMI changed by 3%). However, this is still the highest AMI compared to the other methods. Clustering based on the other features perform considerably worse. For example on sample selection 1 ($\mathbb{D}_{train} = \mathbb{D}_{train}/\mathbb{D}_{load=1}$), OPTICS underestimated the number of classes present in the feature spaces of the classifier encoder (R=6) and the AE (R=3). Hence, data from different classes are assigned to the same cluster, resulting in a lower h score compared to the c score. Clustering with k-means performs slightly better for all methods. This is due to the fact that the minimum number of clusters was set to 7 (see Section 4.4.5). Hence, the number of clusters is closer to the number of true classes in the data resulting in a better performance compared to the density-based clustering method OPTICS. We again see a higher c score compared to the h score for all evaluation models. This means that multiple classes are assigned to one cluster, whereas other classes are split into multiple clusters.

4.5.2 Case Study 2: Missing Faults

In Figure 4.3 the feature space is depicted for $\mathbb{T} \cup \mathbb{T}_p$ for sample selection 1 ($\mathbb{D}_{train} = \mathbb{D}_{train}/\mathbb{D}_{fault=1,4,7}$) and the models sharing the same encoder architecture (AE, classifier encoder, triplet encoder). The true labels as well as the predicted cluster class of the two methods (OPTICS and k-means) is displayed.

In the first row of the figure, the true labels of the different feature spaces are shown. The light grey (fault B7), light orange (fault B14), and light purple (fault B21) correspond to



(a) CLE (b) AE (c) TE

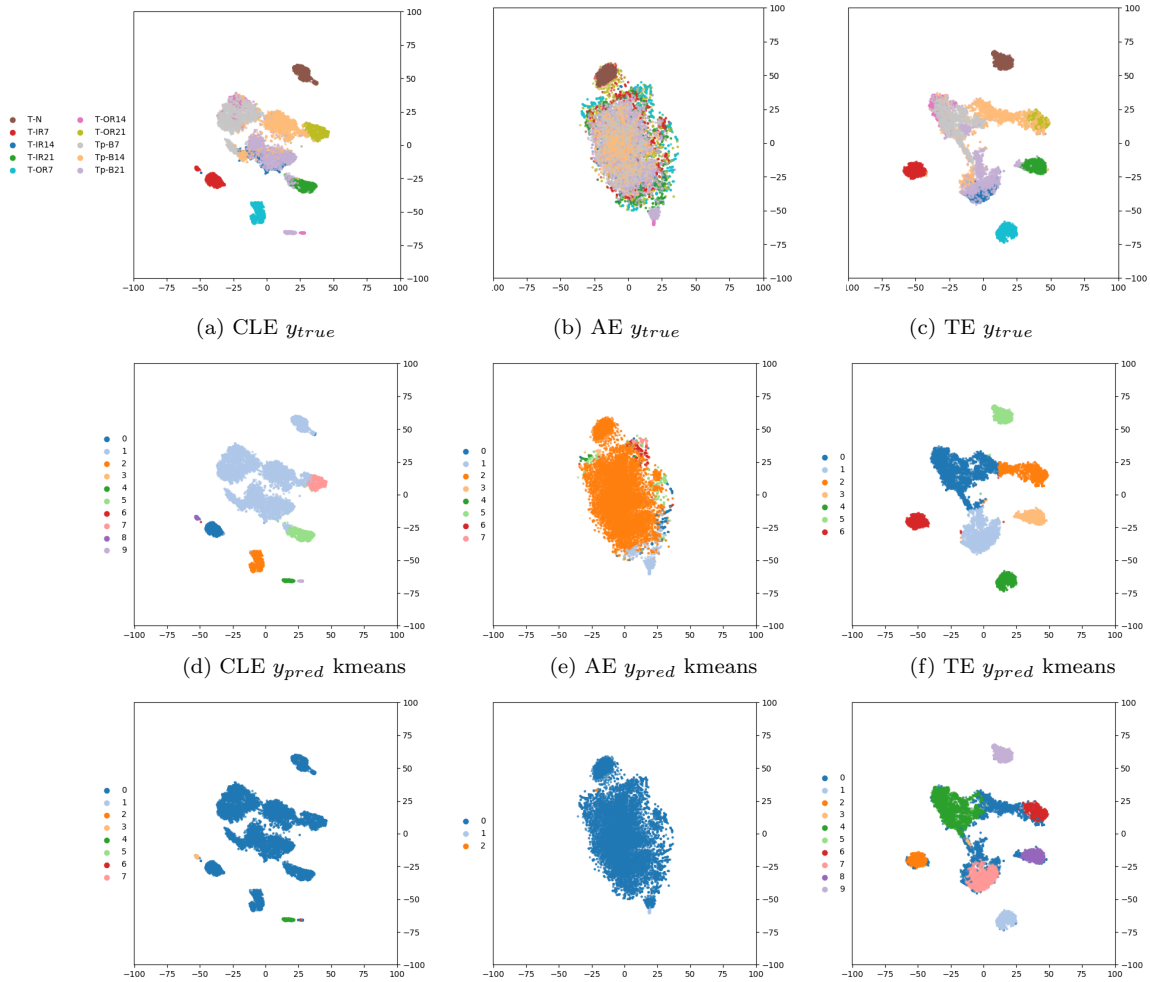
Figure 4.2: Case Study 1: t-SNE Plot of Feature Space on $\mathbb{T} \cup \mathbb{T}_p$ of the Classifier Encoder (CLE), the Autoencoder (AE) and the Triplet Encoder (TE).

	Classification		Clustering - OPTICS								Clustering - k-means							
	$\mathbb{T}$	$\mathbb{T}_p$	$\mathbb{T}$				$\mathbb{T} \cup \mathbb{T}_p$				$\mathbb{T}$				$\mathbb{T} \cup \mathbb{T}_p$			
	acc	acc	R	AMI	h	c	R	AMI	h	c	R	AMI	h	c	R	AMI	h	c
Sample Selection 1: $\mathbb{D}_{train} = \mathbb{D}_{train}/\mathbb{D}_{load=1}$; $\mathbb{T} = \mathbb{D}_{test}/\mathbb{D}_{load=1}$ and $\mathbb{T}_p = \mathbb{D}_{load=1}$																		
FFT	97%	91%	5	27%	16%	84%	5	26%	15%	89%	11	47%	41%	56%	10	53%	47%	60%
AE_{FFT}	97%	91%	3	26%	17%	88%	6	26%	16%	87%	11	46%	40%	56%	10	46%	38%	58%
AE	67%	60%	3	1%	1%	36%	4	1%	1%	32%	11	29%	22%	46%	14	29%	22%	46%
CLE	**100%**	99%	6	23%	14%	67%	6	23%	13%	80%	11	70%	62%	81%	11	70%	61%	82%
TE	**100%**	**100%**	11	**96%**	**98%**	**95%**	11	**97%**	**98%**	**95%**	10	**100%**	**100%**	**100%**	10	**99%**	**99%**	**99%**
Sample Selection 2: $\mathbb{D}_{train} = \mathbb{D}_{train}/\mathbb{D}_{load=2}$; $\mathbb{T} = \mathbb{D}_{test}/\mathbb{D}_{load=2}$ and $\mathbb{T}_p = \mathbb{D}_{load=2}$																		
FFT	97%	95%	6	26%	15%	87%	5	25%	15%	94%	11	47%	41%	56%	11	47%	39%	58%
AE_{FFT}	97%	94%	7	6%	4%	42%	6	25%	15%	93%	10	46%	40%	56%	10	45%	38%	57%
AE	65%	59%	2	4%	2%	5%	3	2%	1%	4%	20	29%	23%	43%	20	30%	23%	45%
CLE	99%	98%	8	28%	17%	72%	7	26%	16%	80%	13	70%	63%	80%	11	70%	60%	85%
TE	**100%**	**100%**	11	**96%**	**97%**	**94%**	10	**93%**	**92%**	**95%**	10	**99%**	**99%**	**99%**	10	**99%**	**99%**	**99%**

Table 4.4: Case Study 1: Classification and Clustering Results on Various Operating Conditions Based on Feature Spaces of the FFT, the Autoencoder based on the FFT (AE_{FFT}), the Autoencoder (AE), the Classifier Encoder (CLE), and Triplet Encoder (TE) Models.

the novel fault types. These are not well isolated in any of t-SNE visualizations, as can be seen in the first row Figure 4.3. Therefore, none of the clustering algorithms can identify the novel fault types as distinct clusters. Yet the original clusters in $\mathbb{T}$ in the triplet encoder feature space are still being found in $\mathbb{T} \cup \mathbb{T}_p$ using both clustering methods (third column in Figure 4.3). While k-means simply assigns the novel fault types to the already existing clusters, OPTICS identifies some data of the novel fault types as outliers (labeled with 0): For example on data selection 1 ($\mathbb{D}_{train} = \mathbb{D}_{train}/\mathbb{D}_{fault=1,4,7}$) considering the triplet encoder feature space, a total of 2,036 noisy samples or outliers are detected, of which 166 are B7 faults (11% of all B7 faults), 1,140 are B14 faults (76% of all B14 faults), and 623 are B21 faults (43% of all B21 faults). Ultimately, 94% of the outliers are from the novel fault type.

The evaluation metrics are shown in Table 4.5. Only the class clusters of the triplet encoder are similarly compact such that a fixed value of ϵ could be set (see Section 4.4.5). Therefore, outliers could be identified as samples that are not density reachable. For the other baseline methods, this was not possible (see Section 4.4.5). As the novel faults are not well isolated in either of the resulting feature spaces, OPTICS performs well only in identifying novel faults as outliers on the triplet encoder features, where the cluster densities are compact. It is apparent that the feature space of the different AE as well as the FFT features does not provide a feature representation that is able to group the different fault types. This is also true for the feature space of the classifier. Many classes are grouped in the same cluster whereas other classes are split into multiple clusters, resulting in a higher c score compared to the h score for both clustering methods.



(a) CLE y_{true} (b) AE y_{true} (c) TE y_{true}

(d) CLE y_{pred} kmeans (e) AE y_{pred} kmeans (f) TE y_{pred} kmeans

(g) CLE y_{pred} OPTICS (h) AE y_{pred} OPTICS (i) TE y_{pred} OPTICS

Figure 4.3: Case Study 2: t-SNE plot of feature space of the classifier encoder (first column), AE (second column), and triplet encoder (last column) model on $\mathbb{T} \cup \mathbb{T}_p$ with the true labels (first row), the predicted labels with k-means (second row), and the predicted labels with OPTICS (last row)

	Classification		Clustering - OPTICS								Clustering - k-means							
	$\mathbb{T}$	$\mathbb{T}_p$	$\mathbb{T}$				$\mathbb{T} \cup \mathbb{T}_p$				$\mathbb{T}$				$\mathbb{T} \cup \mathbb{T}_p$			
	acc	acc	R	AMI	h	c	R	AMI	h	c	R	AMI	h	c	R	AMI	h	c
Sample Selection 1: $\mathbb{D}_{train} = \mathbb{D}_{train}/\mathbb{D}_{fault=1,4,7}$; $\mathbb{T} = \mathbb{D}_{test}/\mathbb{D}_{fault=1,4,7}$ and $\mathbb{T}_p = \mathbb{D}_{fault=1,4,7}$																		
FFT	100%	0%	4	37%	24%	86%	5	23%	13%	87%	9	57%	53%	61%	7	41%	31%	59%
AE_{FFT}	100%	0%	4	37%	24%	86%	4	23%	13%	89%	8	51%	47%	57%	7	39%	30%	56%
AE	84%	0%	3	4%	2%	4%	3	3%	1%	43%	13	21%	17%	29%	8	16%	11%	35%
CLE	100%	0%	8	7%	4%	34%	8	7%	4%	41%	9	75 %	68 %	83%	5	54%	41%	80%
TE	100%	0%	8	96%	98%	94%	10	73%	69%	77%	7	100%	100%	100%	7	72%	64%	82%
Sample Selection 2: $\mathbb{D}_{train} = \mathbb{D}_{train}/\mathbb{D}_{fault=2,5,8}$; $\mathbb{T} = \mathbb{D}_{test}/\mathbb{D}_{fault=2,5,8}$ and $\mathbb{T}_p = \mathbb{D}_{fault=2,5,8}$																		
FFT	97%	0%	3	36%	23%	88%	6	23%	13%	85%	7	35%	28%	50%	7	50%	41%	65%
AE_{FFT}	97%	0%	5	37%	24%	84%	5	23%	14%	87%	7	42%	36%	51%	11	53%	48%	59%
AE	73%	0%	2	1%	1%	3%	2	1%	0%	4%	18	26%	23%	36%	7	28%	21%	43%
CLE	**100%**	0%	9	36%	25%	68%	9	24%	14%	73%	7	66%	58%	77%	7	61%	50%	78%
TE	**100%**	0%	8	**97%**	**99%**	**96%**	7	**76%**	**65%**	**93%**	7	**99%**	**99%**	**99%**	7	**79%**	**71%**	**90%**

Table 4.5: Case Study 2: Classification and Clustering Results with Novel Faults Based on Feature Spaces of the FFT, the Autoencoder based on the FFT (AE_{FFT}), the Autoencoder (AE), the Classifier Encoder (CLE), and Triplet Encoder (TE) Models.

4.6 Discussion

The goal of this research is to learn a feature representation that allows for robust classification under changing operating conditions as well as identification of novel faults. None of goals is a classification task per se. However, the classification results on test dataset ($\mathbb{T}$) allow for comparison with results of other State-of-the-Art (SOTA) publications on the used benchmark dataset. Accuracies above 99% have been achieved by various SOTA methods (see Section 4.4.1). Despite the rather simple model architectures evaluated in this paper (compared to other SOTA models - see Section 4.4.1), the classification results on the test dataset $\mathbb{T}$ of up to 100% showcase the validity of the proposed methods including the chosen baseline methods.

Over all the case studies, the performance based on the AE with the 16-dimensional feature space is very low. However, we consider a low-dimensional feature space more suited to filtering out uninformative variations from the input data, which is one of the objectives of this work. Therefore, we consider this a fair comparison. The lack of robustness of these autoencoding methods to new operating conditions becomes particularly apparent in the high classification performance drop in case study 1 from $\mathbb{T}$ to $\mathbb{T}_p$ (see Table 4.4). This is not surprising as the AE is trained to fully reconstruct the input signal. Hence, the objective is to pass all information regarding the measurements through the bottleneck layer, including information related to various operating conditions. Therefore, variations in the operating conditions appear in the feature space as well, making this approach not suitable if the objective is to achieve invariance or robustness to operating conditions. Similarly, the FFT features contain all information of the signal including variations caused by operating conditions. Therefore, the classification performance based in these features are equally affected by the change in operating conditions.

The labels are directly considered when training the classifier encoder and triplet encoder, enabling the models to focus on the semantic meaning. This results in a better classification performance. Remarkably, the classification performance based on the classifier encoder and triplet encoder features is hardly affected if the operating conditions change at inference time. The clustering performance on the features of these two models varies significantly, both on $\mathbb{T}$ and $\mathbb{T} \cup \mathbb{T}_p$. As the features of a certain class are represented in a more compact way by the triplet encoder, the space is more suited for clustering. Yet, a shift can be visually observed between the data of $\mathbb{T}$ and $\mathbb{T}_p$ within the respective clusters. This means that the model is not invariant to the shift in operating conditions. However, the different classes in $\mathbb{T} \cup \mathbb{T}_p$ are still cohesive and separable. Therefore, neither the classification nor the clustering performance is negatively impacted by the novel operating conditions. Both clustering methods perform well on $\mathbb{T} \cup \mathbb{T}_p$, with k-means even delivering results comparable to the classification performance.

All feature encodings are sensitive to variations in the data corresponding to novel faults. However, they do not provide a representation that allows the clustering algorithms to isolate them in the feature space. Therefore, none of the clustering methods identifies clusters including most of a novel fault class. Yet the compactness of the learned feature representations per class of triplet encoder enables to set a fixed value of ϵ in OPTICS, i.e. a fixed maximal value for two samples to be considered neighbors in a cluster. This enables us to detect novel faults at least as outliers (if not as distinct clusters). A detected outlier could raise an alarm to the operator and initiate a further evaluation. For example, in the sample selection 1 of case study 2, 94% of the outliers actually correspond to novel faults, relatively few false alarms will be raised. However, many novel faults will not be detected but simply registered as another fault class. In this case, fault detection will still be ensured.

Limitations: The performance of the OPTICS clustering algorithm depends strongly on the data at hand: If the dataset contains mainly novel faults ($|\mathbb{T}_p| >> |\mathbb{T}|$), these will primarily determine the clusters and will not be detected as outliers anymore. Therefore, it is important to keep the dataset $\mathbb{T}$ with known conditions as a reference for the clustering algorithm. Continuously, a novel dataset with unknown conditions $\mathbb{T}_p$ can be added. Our case studies have been conducted under an approximate balance between the two datasets ($|\mathbb{T}| \approx |\mathbb{T}_p|$); this ratio can be tuned according to the safety criticality of the system.

4.7 Conclusion

In this research, contrastive learning has been evaluated in the context of PHM applications. Specifically two typical scenarios in PHM were investigated: A trained model is faced with new operating conditions and new faults at inference time. We were able to show that a feature representation trained with a contrastive learning paradigm is well suited to the clustering of classes under different and partially novel operating conditions. This enables clustering that is invariant to fluctuation in the data corresponding to similar but novel operating conditions, as seen before. Simultaneously, the compactness of the retrieved feature representations enables density-based clustering that is sensitive to novel faults. Ultimately, contrastive learning seems to be a promising paradigm for PHM applications. To further establish contrastive learning in PHM applications, we propose to dedicate future work to the question of how contrastive learning can be applied in a semi-supervised or unsupervised setting.

5 Controlled Generation of Unseen Faults for *Partial* and *Open-Partial* Domain Adaptation

This chapter corresponds to the published article: [1]

Rombach, Katharina, Gabriel Michau, and Olga Fink (2023). "Controlled generation of unseen faults for Partial and Open-Partial domain adaptation". In: *Reliability Engineering & System Safety* 230, p. 108857. DOI: https://doi.org/10.1016/j.ress.2022.108857.

Abstract: New operating conditions can result in a significant performance drop of fault diagnostics models due to the domain shift between the training and the testing data distributions. While several domain adaptation approaches have been proposed to overcome such domain shifts, their application is limited if the fault classes represented in the two domains are not the same. To enable a better transferability of the trained models between two different domains, particularly in setups where only the healthy data class is shared between the two domains, we propose a new framework for *Partial* and *Open-Partial* domain adaptation based on generating distinct fault signatures with a Wasserstein GAN. The main contribution of the proposed framework is the controlled synthetic fault data generation with two main distinct characteristics. Firstly, the proposed methodology enables to generate unobserved fault types in the target domain by having only access to the healthy samples in the target domain and faulty samples in the source domain. Secondly, the fault generation can be controlled to precisely generate distinct fault types and fault severity levels. The proposed method is especially suited in extreme domain adaption settings that are particularly relevant in the context of complex and safety-critical systems, where only one class is shared between the two domains. We evaluate the proposed framework on *Partial* as well as *Open-Partial* domain adaptation tasks on two bearing fault diagnostics case studies. Our experiments conducted in different label space settings showcase the versatility of the proposed framework. The proposed methodology provided superior results compared to other methods given large domain gaps.

5.1 Introduction

A reliable operation of complex (safety-critical) assets can be achieved by monitoring the condition of the assets in real time, detecting the faults in an early stage and distinguishing between the different fault types to enable an informed schedule of the recovery maintenance or fault mitigation actions (Abid et al., 2021). Data-driven models based on real-time condition monitoring (CM) data have shown a great potential for fault detection and diagnostics (Zhao et al., 2020; Guan et al., 2021). However, CM data is often affected by distributional shifts (referred to as domain shifts), that can significantly decrease the performance of data-driven models (Miao et al., 2022; Zhou et al., 2022d). For example, changing operating conditions can cause such a distributional shift (Michau and Fink, 2021; Rombach et al.,

[1] Please note, this is the author's version of the manuscript published in *Reliability Engineering and System Safety*. Changes resulting from the publishing process, namely editing, corrections, final formatting for printed or online publication, and other modifications resulting from quality control procedures may be subsequently added. The final publication is available at https://doi.org/10.1016/j.ress.2022.108857.

2021). Similarly, CM data of two units of a fleet can differ quite significantly due to differences in their configurations and operating regimes (Michau and Fink, 2019; Li et al., 2021c). To enable the transfer of a data-driven model to new operating conditions or new units in a fleet, domain adaptation (DA) methods have been successfully applied in fault diagnostics (Deng et al., 2022; Lee et al., 2022). Most of the proposed approaches, however, require that the same fault classes are represented in the source and the target domain. This DA setting, where the source and target domain datatsets cover the same classes, is referred to as *Closed-Set* DA - see Figure 5.1. However, in real-world datasets, the classes represented in the two domains are not always congruent. Due to the rareness of faults in complex industrial assets, for example, observing each possible fault in all assets of a fleet and/or under all possible operating conditions may not be practically feasible, particularly for safety-critical systems (Michau and Fink, 2021). Practical fault diagnostics solutions, typically, need to be taken into operation within a short period of time, not allowing to wait until all possible fault types have occurred. This results in cases where not all fault classes have been observed in all units or under all operating conditions, leading to label space discrepancies in CM datasets.

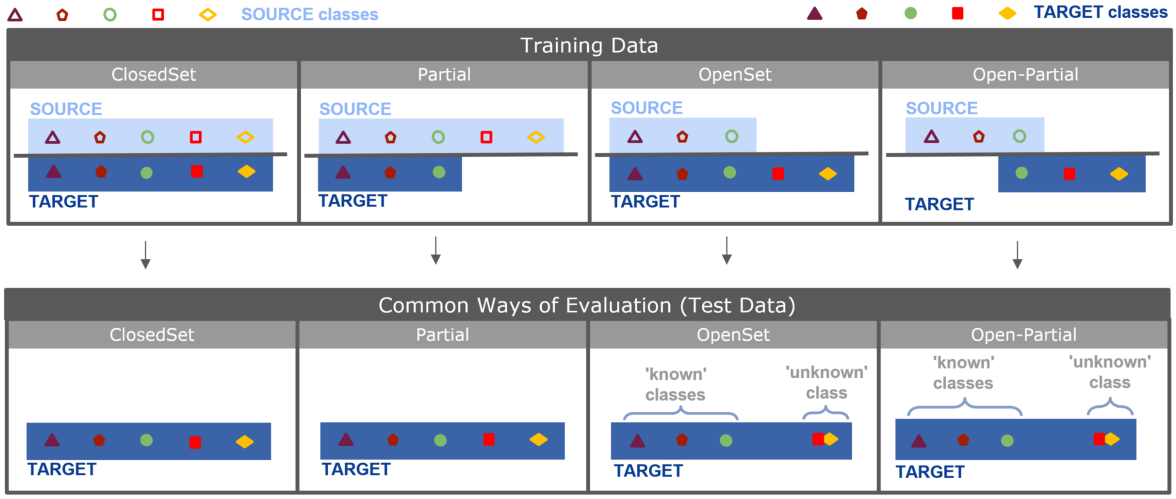
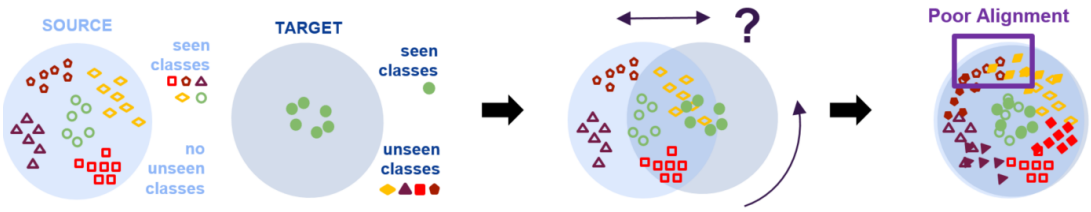


Figure 5.1: Four DA configurations according to label space discrepancies: (a) *ClosedSet*; (b) *Partial*; (c) *OpenSet*; (d) both *Open-Partial* (Boris et al., 2021).

In the literature, different DA settings have been distinguished by their type of discrepancy in the label space (Boris et al., 2021). This is illustrated in Figure 5.1. In the *Partial* DA setting, the target domain covers only a subset of the source classes (source domain has private classes), whereas in the *OpenSet* DA setting, the source domain covers only a subset of the target domain classes (target domain has private classes). The *Open-Partial* DA setup is a combination of both previous settings where both domains have private classes that are not represented in the other domain. Most existing DA methods are designed for only one of the above mentioned DA settings (mainly *ClosedSet* DA) and are often not transferable to other DA scenarios (see Section 5.2). This poses a challenge for successful DA in real applications, where different types of discrepancies in the label space can occur. Since safety-critical systems are reliable by design, faults occur very rarely. In some cases, only healthy data (one class) is available in different domains, not allowing to perform fault diagnostics at all. Instead, Michau and Fink (2021), proposed to train a one-class classification model that is transferable between domains. If faults did occur in one of the domains, fault diagnostic is possible. For the DA task, however, we are presented with an extreme case of label space discrepancy in prognostics and health management (PHM) applications, where often only one class, the healthy one, is shared between the two domains (Wang et al., 2020a). For example, if a system starts operating under a new operating condition, only data of

the assets's current condition will be available. For safety-critical systems, this is usually the healthy condition, meaning that only the healthy class is shared between datasets from various operating conditions (an extreme case of *Partial* DA). As illustrated in Figure 5.2, such an extreme case of label space discrepancy between two domains can pose a significant challenge for *Partial* DA methods based on feature alignment. With only one class shared between the two domains, there exist many possible alignment solutions (see Figure 5.2b) and their performance can only be evaluated after the model is employed and the real target faults have been observed (see Figure 5.2c). Extreme discrepancies in the label space of training datasets can also arise if two units of a fleet are experiencing different fault types during the data collection (and model development) period. Then, in the available training dataset, the only common health class experienced so far by both units may be the healthy class. However, both of the units can be affected during their life times (during the deployment of the developed models) by the same failure modes. Therefore, the fault diagnostics algorithms should be able to diagnose all possible fault types and not only those that have been experienced by the specific unit at the training time. The results of previous studies show that, generally, the less classes are shared between the domains, the harder the DA task becomes (Wang et al., 2020a; Zhang et al., 2021b). For example, compared to the *ClosedSet* setting, the classification performance on a bearing dataset dropped by 20% when only three out of ten classes were shared between the two domains (Zhang et al., 2021b). Despite the relevance to PHM applications, there is hardly any work tackling the extreme cases of discrepancies in the label spaces (with only one shared class between two domains) for fault diagnostics in different DA settings. These extreme scenarios are in the focus of the research in this paper.



(a) Real source and target datasets in the extreme *Partial* DA setting.

(b) Alignment problem.

(c) A-posteriori evaluation.

Figure 5.2: Illustration of the source and target alignment challenge when only one class is shared between the domains on the example of the *Partial* DA setting: The source and the target datasets are shown in Figure 5.2a whereby only one class (green class) is represented in the target domain. The alignment step based on one class only is shown in Figure 5.2b, whereby the challenge of finding the optimal alignment is indicated. The quality of chosen alignment method can only be tested during the a-posteriori evaluation, when the target classes have been observed (see Figure 5.2c).

We propose to address the challenge raised by the label space discrepancies for DA by enabling the generation of domain- and class-specific data from fault conditions that have not been observed before in the target domain. The generated fault data can compensate for unseen domain-specific fault classes and, thereby, transform the given *Partial* or *Open-Partial* DA setting into a *ClosedSet* DA setting. The generation of previously unobserved target fault data is based on observed faults in the source domain i.e. we propose to perform unsupervised domain mapping. This is particularly challenging since it is unknown how an unobserved fault in the target domain should look like. The unsupervised target fault generation needs to fulfill two requirements. Firstly, the generated data should be adapted to the specificities of the desired domain and secondly, the faults should be specific to a desired class in the label space. We address for the first time such an unsupervised but controlled generation of fault data based only on the healthy data in the target domain and faulty data in the source domain. The proposed work is based on the hypothesis that the Fourier spectrum from faulty data can be disentangled in data signatures that represent (1)

solely fault class characteristics and (2) domain-specific characteristics within the data. The validity of this hypothesis is evaluated implicitly by conducting different DA experiments. The main contribution of this research is a novel framework *FaultSignatureGAN* based on a Wasserstein GAN (Arjovsky et al., 2017) that enables to generate domain-independent fault class signatures that are transferable to any new domain, given only healthy data of that domain. This is, on the one hand, a particularly challenging task since no samples of faulty data in the target domain are available and, on the other hand, a particularly relevant case for real safety-critical applications where a representative dataset of fault data is typically not available. *FaultSignatureGAN* enables a controlled way to generate physically plausible faults of previously unobserved distinct classes in the target domain and thereby, enables to complement label spaces with different types of class discrepancies for DA tasks. Since the proposed framework relies solely on the availability of source faults and healthy target data, its benefits are particularly pronounced for targeting the extreme case of DA where only one class (the healthy class) is shared between the two domains. However, it is applicable to any number of shared and missing fault classes in the two domains. The proposed framework *FaultSignatureGAN* is not limited to only one type of label space discrepancy since it is applicable in *Partial* as well as *Open-Partial* setups.

The remainder of the paper is organized as follows. First relevant related work is summarized in Section 5.2, the proposed framework is explained in Section 5.3. The case studies are introduced in Section 5.4 and the exact setup of the conducted experiments is stated in Section 5.5. The results of the conducted experiments for *Partial* DA are shown in Section 5.6.1 and for *Open-Partial* settings in Section 5.6.2. The findings are discussed in Section 5.7 and conclusions are drawn in Section 5.8.

5.2 Related Work

Domain Adaptation has been intensively studied in recent years in the context of PHM applications (Li et al., 2022). Most of the proposed approaches, however, have been exclusively developed for the *ClosedSet* DA setting where the source and target domain cover the same classes - see left column in Figure 5.1. As exemplified in Section 5.1, the assumption of a *ClosedSet* setting is not realistic in many practical applications. Hence, *ClosedSet* DA methods do not meet the requirements of industrial applications. There has been an increasing interest to develop methods that address more realistic DA scenarios with label space discrepancies. Approaches for *Partial* (Liang et al., 2020), *OpenSet* (Li et al., 2021b) as well as *Open-Partial* (Saito and Saenko, 2021) DA have been mainly developed in the field of computer vision. Recently, several research studies have developed the ideas further to adapt them to the challenges of real CM data. In the context of fault diagnostics, adversarial approaches have been proposed for *OpenSet* DA with different degrees of label space discrepancies i.e. with a different number of shared classes between the two domains. For example, in (Zhao and Shen, 2022c), an auxiliary domain discriminator was introduced to attribute less weight to private target samples and fault diagnostics experiments were conducted on three bearing datasets with two shared classes between the two domains. In another study on *OpenSet* DA for fault diagnostics, Zhang et al. (2021b) used an instance-level weighted mechanism to identify private target classes and tested the proposed method i.a. with three (out of ten) shared classes between the domains on two rotating machinery datasets. The results demonstrated that, generally, the less classes are shared between the domains, the harder the DA task becomes. Another method proposed a source class-wise and target instance-wise weighting mechanism combined with an additional outlier identifier for *OpenSet* fault diagnostics on two rotating machine datasets. The proposed method has even been applied on multiple label space discrepancy settings (Zhang et al., 2021c). Despite its relevance to fault diagnostics in safety-critical complex technological system, none of the above mentioned studies has tackled the extreme case of the *OpenSet* or *Open-Partial* DA

setting, where only one class (the healthy class) is shared between the domains. Another major limitations of the previously proposed approaches on *OpenSet* DA is that they only aim to classify known classes (i.e. source classes) and do not enable to distinguish private target samples into different classes. In safety-critical applications, however, it is important to distinguish between different health conditions within the private target samples to plan appropriate maintenance actions.

Methods targeting the *Partial* DA setting have also been developed for PHM applications. For example, a class-weighted adversarial DA method was proposed that uses the domain discriminator's output to detect private classes (Li et al., 2020d). The output of two classifiers has also been employed to estimate the target distribution and train domain-invariant representations (Jiao et al., 2019). Also, randomly selected source data is used to augment the target domain to align the conditional distributions combined with a class-wise adaptation (Zhao et al., 2022). Some research studies have even dealt with the extreme case of *Partial* DA where only the healthy state is shared between the two domains (Li and Zhang, 2020; Li et al., 2018b; Wang et al., 2020a). For example, Li and Zhang (2020) proposed a conditional data alignment step (using the maximum mean discrepancy) that is only applied to the healthy data from the source and target domain to prevent misalignment due to the label space discrepancy. In addition to the conditional alignment, the authors proposed prediction consistency schemes using multiple classifier models for fault diagnostics in *Partial* DA settings. Wang et al. (2020a) proposed a unilateral alignment approach (*Unilateral*) for *Partial* DA with extreme label space discrepancy. The proposed method made use of the inter-class relationships of the source domain and aligned the target features to the pre-trained source domain features. Although the results of previous studies using different feature source and target alignment techniques in extreme *Partial* DA settings are promising, the methods have mainly been tested on CM datasets with small domain gaps (indicated by the high Baseline classification performance). The employed methods may fail under large domain shifts, where the inter-class relationships might have changed significantly.

One fault data generation approach was investigated in the extreme case of *Partial* DA combined with an additional alignment step (Li et al., 2018b). However, the proposed target data generation method required extrapolation abilities of the generative model. Given the limited extrapolation abilities of deep models, it is not to be expected that the generated data resembles realistic target faults - especially given large domain gaps. Instead of generating target data as performed in (Li et al., 2018b), Zhao et al. (2022) adapted the idea of Liang et al. (2020) and proposed to augment the target data with source data to compensate the missing class data and performed adversarial feature alignment on the augmented and class-weighted datasets combined with a class-center-alignment loss. While the source data augmentation stabilized the alignment process, the proposed method may fail in settings where the inter-class relationships might have changed significantly. Further, the above mentioned approaches tackling different settings of label space discrepancy in DA have usually been developed for one specific DA setting, either*Partial* or *OpenSet*, and are typically not applicable in other settings. Furthermore, large domain gaps have not been tackled so far in the extreme case of label space discrepancy where only the healthy class is shared between the domains. Another limiting factor in applying the above mentioned DA methods based on feature alignment to new safety-critical assets is finding an optimal hyperparameter setting. With only one class being shared between the domains, there exist multiple possible alignment solutions and their quality can only be evaluated a posteriori, posing a safety risk in industrial assets. Therefore, previous works used, e.g., data and labels from target faults for one validation domain shift to tune the hyperparameters (Wang et al., 2020a). This solution to find the optimal hyperparameter settings is, however, not possible in real applications where data from unobserved target faults is not available.

In this work, we aim to develop a framework that performs well in the extreme case of

DA under different label space discrepancy settings with a particular focus on the *Partial* and *Open-Partial* setting. We aim to develop a framework that enables DA also in the cases where the domain gaps are large. Further, we aim to achieve this without relying on target validation data to tune our methodology, as this is one of the limiting factors in existing DA methods to new safety-critical assets.

Domain generalization addresses the challenge of fault diagnostics under unforeseen domain shifts (contrary to one explicit shift between two domains). Different techniques have been proposed in the context of fault diagnostics (Zhou et al., 2022d; Zhao and Shen, 2022b). However, these methods generally require access to multiple source domains, mainly with shared labeled spaces . This is often not given in industrial applications and therefore, domain generalization approaches are not applicable to the challenges addressed in this research.

Controlled Synthetic Data Generation has raised a lot of attention in recent years (Gui et al., 2021). In the context of DA for computer vision tasks, for example, conditional generative models have been employed for domain mapping i.e. to translate a source input image to an image that closely resembles the target distribution (Wilson and Cook, 2020). However, these approaches require a *ClosedSet* DA setting since the target domain typically inherits the labels from the source domain. In the context of PHM applications, generative models have mainly been applied to balance imbalanced datasets, whereby e.g. conditional GANs have been used to control the generation process to generate desired distinct classes (Luo et al., 2021). However, those approaches are solely suited to generate data from classes that have been observed before and not to generate previously unobserved classes in a specific domain. The latter is the focus of our research.

Contrary to using generative models, Wang et al. (2021) proposed to use expert knowledge about different fault type patterns to generate synthetic fault data without access to any real fault data. This enabled to address fault diagnostics if only little real target fault data is available by performing DA in a subsequent step to close the synthetic-to-real domain gap. The expert knowlege enables to transfer different fault types to different types of bearings. However, this approach requires a substantial domain knowledge. Furthermore, patterns of different fault types (as in (Wang et al., 2021)) are typically easier to distinguish compared to different severity levels of the same fault type, as addressed in this research. To distinguish between different fault severities as well as types, synthetic data representing also different fault severities is required.

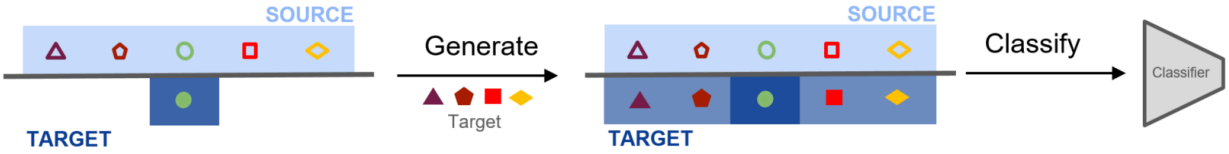
The concept of **disentenglement**, which is based on the hypothesis that real-world data is generated by a few independent explanatory factors of variation (Locatello et al., 2019), has also been studied in the context of controlled data generation. However, although disentangled representations should be general and are expected to be generalizable to new domains, recent studies found that disentanglement does not guarantee combinatorial generalization (understand and produce novel combinations of familiar elements) (Schott et al., 2021). To mitigate the lack of generalizability of disentangled representations, it is possible to constrain the disentanglement using a-priori knowledge on the data structure. For example, Yang and Soatto (2020) assumed that, in image datasets, the domain-specific information is solely represented in the low frequency range whereas the semantic information is reflected only in the high frequency range. This assumption allowed the authors to generate unseen target data simply by swapping the domain-specific low-frequency block of the source and target images and perform DA with the synthetically generated data. This block-wise distinction into a domain-specific and a semantic-specific frequency ranges can be considered as a disentanglement in the Fourier space. Unfortunately, such a block-wise distinction of fixed frequency ranges representing either solely the domain-specific or semantic-specific components is not possible for CM data from mechanical systems with complex dynamic behaviour. Instead, we expect a fault as well as a domain shift to affect the entire frequency spectrum. Based on the intuition that the OCs are independent of defects, we can assume that faults create

disturbances on top of existing signals. We, therefore, assume that the Fourier spectrum can be expressed as the sum of domain-specific components and fault-specific components. This assumption, that both information content (domain-specific and semantic-specific) does not only affect constrained frequency ranges but rather impact the entire spectrum, generalizes the work of (Yang and Soatto, 2020) to data from other application domains, in particular to CM data from complex industrial systems. Further, the assumptions enables the generation of unseen data while neither relying on combinatorial generalization of disentangled features nor relying on extrapolation abilities of the generative model.

In this work, we aim to develop a framework that enables a controlled generation of novel distinct fault classes in a target domain where the fault condition has not been observed before. Thereby, the **contribution of our proposed framework** is the generation of unseen domain-specific fault data, that enables DA with extreme label space discrepancies, also under large domain gaps. Contrary to other generative approaches, we do not only control the class being generated but also the specific domain of the data. Further, the data generation is unsupervised since the respective target fault has not been observed before in the specific target domain i.e. we enable the controlled generation of out-of-distribution data. Although the developed framework enables the generation of previously unseen data, it does not rely on extrapolation abilities of the generative model but instead, it relies on a disentanglement assumption. This assumption enables to transfer the fault information between different domains and, ultimately, to generate physically plausible data of unseen fault types and fault severities. The proposed framework, therefore, enables to generate data that can substitute for missing domain-specific class data for DA problems with label space discrepancies. Our methodology is especially suited for DA in the extreme case of label space discrepancy, where only one class is shared between the domains and thereby, addresses an important requirement of reliable fault diagnostics in complex industrial (safety-critical) assets. However, it is also applicable to DA setups with any number of missing classes. Furthermore, contrary to other DA approaches, the proposed methodology is universally applicable to both *Partial*, *Open-Partial* DA setups.

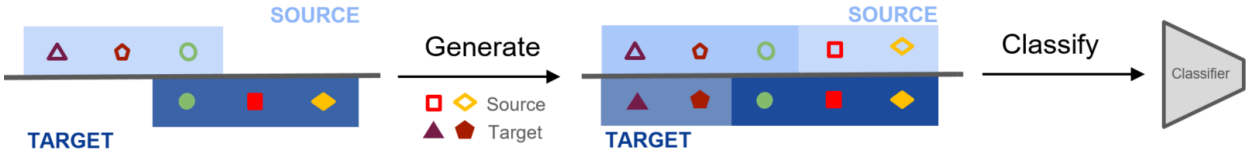
5.3 Methodology

We propose a framework, referred to as *FaultSignatureGAN*, that enables to generate distinct domain-independent fault signatures based on the hypothesis that the faulty signal can be represented as the sum of domain-specific components and fault-specific components. These fault signatures can be transferred to new target domains such that the transferred data is representative of distinct fault classes in a target domain where they have not been observed before. The generated data is then used in a subsequent step to substitute for missing class data in different DA settings with label space discrepancies: *Partial* (see Figure 5.3) and *Open-Partial* DA (see Figure 5.4). Finally, a classification model is trained on the augmented datasets.



(a) Real data. (b) Generate data. (c) Real and generated data. (d) Classifier.

Figure 5.3: **FaultSignatureGAN** in the *Partial* DA settings: the original data setting is depicted in Figure 5.3a; the missing target classes are generated in Figure 5.3b; the target dataset is augmented with synthetically generated data in Figure 5.3c and a classifier is trained on the augmented dataset in Figure 5.3d.



(a) Real data. (b) Generate data. (c) Real and generated data. (d) Classifier.

Figure 5.4: **FaultSignatureGAN** in the *Open-Partial* DA settings: the original data setting is depicted in Figure 5.4a; the missing source and target classes are generated in Figure 5.4b; the source and target dataset is augmented with synthetically generated data in Figure 5.4c and a classifier is trained on the augmented dataset in Figure 5.4d.

5.3.1 Training the generative model

FaultSignatureGAN comprises three parts (A-C) as illustrated in Figure 5.5: (A) The first part ensures that generated fault signatures are easily transferable to a specific domain; (B) the second part ensures that the transformed fault signatures represent plausible domain data; and (C) the last part ensures that the transformed fault signatures are representative of the desired fault classes. Part **(A)** of the framework is tackled by a generative network that generates domain-independent fault signatures from distinct classes in the Fourier domain. These fault signatures are then transferred to a specific domain by adding them to randomly sampled data from the domain's healthy class. The ability of the generated data to represent true domain fault data is imposed in part **(B)** by an adversarial discriminator. The semantic plausibility of the generated data to represent a desired fault class (as sampled from the sampling module) is tackled by a cooperative classifier in part **(C)**. The different parts of the framework are detailed below.

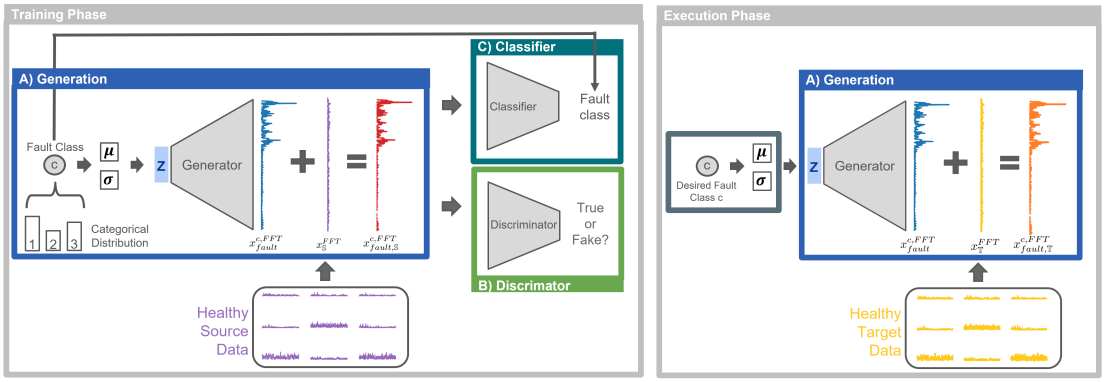


Figure 5.5: *FaultSignatureGAN*: **Training Phase:** Training the A) generative model to generate domain independent fault characteristics while imposing B) plausibility with the discriminator in the source domain and C) semantic consistency with the classifier. **Execution Phase:** The generation of unseen target data.

The underlying hypothesis: The proposed approach is based on the hypothesis that the Fourier spectrum of fault data can be expressed as the sum of 1) domain-specific components (the spectrum of a signal from normal operation) and 2) of fault-specific components representing the specific fault characteristics. Further, we assume that the latter (spectrum of a signal representing the specific faulty condition) corresponds to a general domain-independent fault signature that is adjusted to new domains simply by linear scaling. In other words, this hypothesis allows us to express Fourier coefficients (Cooley and Tukey, 1965) of the fault data of a certain class c from a specific domain $\mathbb{X}$ ($x^{c,FFT}_{fault,\mathbb{X}}$) as a sum of (1) domain-specific characteristics that are represented by the healthy class data of that domain $x^{FFT}_{\mathbb{S}}$ and (2) the fault class specific characteristics that are domain-independent $x^{c,FFT}_{fault}$ and scaled by a factor w- see Equation 5.1.

$$x_{fault,\mathbb{X}}^{c,FFT} = x_{\mathbb{X}}^{FFT} + w * x_{fault}^{c,FFT} \tag{5.1}$$

The linear scaling with w is performed to account for the fact that the fault signature is affected by operational changes and, therefore, we alleviate the strong assumption that the fault-specific variations of real fault data are independent from operating conditions. Between a source domain $\mathbb{S}$ and target domain $\mathbb{T}$, the weight factor w is defined as in Equation 5.2.

$$w = E(\mathbb{T}_{h,\mathbb{S}}/\mathbb{T}_{h,\mathbb{T}}) \tag{5.2}$$

(Part A) The generative model G_θ: The final goal is to generate faults in a target domain that have not been observed before. However, from the two components of the faulty signal in the target domain as given in Equation 5.1, we only have access to the healthy data representing the domain-specific variations ($x_{\mathbb{T}}^{FFT}$) in the target domain $\mathbb{T}$. Therefore, to generate unseen faults in the target domain $x_{fault,\mathbb{T}}^{c,FFT}$, we need to design a framework that enables the generation of the domain-independent characteristics of a fault class $x_{fault}^{c,FFT}$. We train the proposed architecture on the data from the source domain, where we do not only have access to the healthy data $x_{\mathbb{S}}^{FFT}$ but also to true fault data $x_{fault,\mathbb{S}}^{c,FFT}$. In the source domain $\mathbb{S}$, the scaling factor equals to 1. Due to the variability in the healthy class, simply subtracting the individual healthy samples from faulty ones (the reverse operation) is not sufficient to retrieve a domain independent fault signature. Therefore, we propose a generative model. The generative model is trained such that its output (blue signal in Figure 5.5) can be transformed in a real source fault by adding it to a healthy source sample (according to Equation 5.1). Thus, the generated signal can be transformed to real domain faults with any of the samples from the healthy data distribution. In this study, we train one generative model to generate all severity levels of one fault type. This process is depicted in Figure 5.5. To ensure plausibility of the generated signals in the specific domain, the generator is trained to fool **a discriminator** D_w (see below). To ensure semantic or class consistency, we condition the generative model on the desired fault class by simply sampling the distinct desired fault class from a categorical distribution. Each of the discrete values drawn from the categorical distribution corresponds to a specific fault class. The probability of each category is defined based on the class distribution in the training dataset $T_{f,\mathbb{S}}$, from which the fault signatures should be learned from. In other words, the probability of category i is defined by Equation 5.3. The value sampled from the uniform distribution is then passed to two vectors (μ and σ in Figure 5.5), that parameterize a Gaussian distribution (mean and deviation), from which we sample using the reparametrization trick (Kingma and Welling, 2013). The generative model is updated based on the consistency of the desired class with the **classifier's** prediction (see below).

$$p_i = \frac{|\{(x_j, y_j)|((x_j, y_j) \in T_{f,\mathbb{S}})\&(y_j = i)\}|}{|T_{f,\mathbb{S}}|} \tag{5.3}$$

To enable a better distinction, we will refer to the signal representing the domain independent fault characteristics (depicted in blue in Figure 5.5) as the *generated fault signature*, and to the signal representing the domain-specific fault data (depicted in red or orange in Figure 5.5) as the *generated data sample* throughout the paper. Further, in the following, we will consider the data always in the Fourier domain without emphasizing it specifically.

(Part B) The discriminator D_w: We need to ensure that the generated data represents plausible domain data. Our final goal is to generate unseen data from a target domain. However, this target fault data has not been observed so far. Hence, we cannot ensure

plausibility of the generated data in the target domain directly while training the generative model. Instead, we train the generator to generate plausible fault samples in the source domain. Therefore, the discriminator is trained to discriminate between real fault data of the source domain and the generated synthetic source data. We implement a Wasserstein GAN (Arjovsky et al., 2017) that is optimized with gradient penalty (Gulrajani et al., 2017) since its training has proven to be more stable compared to other GAN implementations, mitigating mode collapse. The adversarial loss function is defined by Equation 5.4.

$$L_D = \mathbb{E}_{\tilde{x}^c, x_{h,\mathbb{S}}}[D_w(\tilde{x}^c + x_{h,\mathbb{S}})] - \mathbb{E}_{x^c_{f,\mathbb{S}}}[D_w(x^c_{f,\mathbb{S}})] + \lambda_{GP}\mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(||\nabla_{\hat{x}} D_w(\hat{x})||_2 - 1)^2], \tag{5.4}$$

where D_w is the discriminator model, $\tilde{x}$ is a generated fault signature, $x_{h,\mathbb{S}}$ is a healthy source sample, $x_{f,\mathbb{S}}$ is a faulty source sample and $\hat{x}$ is drawn from $\mathbb{P}_{\hat{x}}$, a newly defined data distribution used to impose the gradient penalty. For more details on the calculation of the gradient penalty, the interested reader is referred to (Gulrajani et al., 2017).

(Part C) The classifier C_γ: A classifier is added to the framework to ensure semantic consistency of the generated data to a desired class. The classifier is optimized with the semi-hard triplet loss (Schroff et al., 2015) on real source data. In Equation 5.5, the corresponding loss function is shown, where C_γ is the classifier network, α is a fixed margin, x_a is the anchor sample, x_p the positive sample and x_n the negative one.

$$L_C = max(||C_\gamma(x_a) - C_\gamma(x_p)||^2 - ||C_\gamma(x_a) - C_\gamma(x_n)||^2 + \alpha, 0) \tag{5.5}$$

For updating the generative model G_θ, the semi-hard triplet loss is calculated using only synthetic data $(\tilde{x} + x_h)$ as anchors and real fault data $x_{f,\mathbb{S}}$ as positive resp. negative samples. A pseudo algorithm of the **Training Phase** is shown in Algorithm 1.

5.3.2 The generation of unseen data in the execution phase

After training the generative model G_θ in the **Training Phase**, the generation of target faults in the **Execution Phase** is straight forward: First, we sample the input of the generative model from a categorical distribution, which determines the desired fault classes to generate. The number of generated samples per class can be chosen freely. This input is then passed to the generative model to generate the respective fault class signatures. The fault signature is then transferred into the target domain (instead of the source domain) by (1) linearly scaling the fault signature (with w) and (2) adding it to the healthy data of the target domain (yellow data in **Execution Phase** of Figure 5.5). The scaling of the fault signature is defined as the ratio between the mean signal of the healthy source data and the mean of the healthy target data per frequency component (as defined in Equation 5.2). Hence, the unseen target data is generated as defined in Equation 5.6.

$$x^{c,FFT}_{fault,\mathbb{T}} = x^{FFT}_{\mathbb{T}} + w * x^{c,FFT}_{fault} \tag{5.6}$$

5.3.3 Alternative approaches used for comparison

In this work, we address two DA settings with label space discrepancies: *Partial* DA and *Open-Partial* DA, with a particular focus on the extreme case where only one class is shared between two domains. While for *Partial* DA, some approaches have been proposed, only few are suitable for this extreme scenario. These few approaches are used for comparison for the *Partial* DA experiments. (1) First, we report the *Baseline* results, where we train a classifier on real source domain data only. It shows the minimal achievable performance if no adaptation is performed. (2) The adversarial feature alignment approach *Unilateral* DA (Wang et al., 2020a) is chosen as a comparison method as it has been evaluated in a *Partial* DA setting before (as elaborated in Section 5.2). It aims to achieve the same goals but uses a

Algorithm 1 Training Phase of *FaultSignatureGAN*

Require: $T_{\mathbb{S}}$ (Source Dataset); $\lambda_{GP}, \lambda_D, \lambda_E, \alpha$ (Loss Function Parameter); n_{critic}, es (Early Stopping Criteria), m (Batch Size)
Ensure: G_θ
▷ **Prepare Dataset**
$T_{h,\mathbb{S}} = \{(x, y) \in T_{\mathbb{S}} \mid \text{y is healthy }\}$; $T_{f,\mathbb{S}} = \{(x, y) \in T_{\mathbb{S}} \mid \text{y is a fault class }\}$
$Cat(T_{f,\mathbb{S}})$ Categorical Distribution of the classes in $T_{f,\mathbb{S}}$
while es == False **do**
  for t = 1,..,n_{critic} **do**
  ▷ **Sample data batches**
    $\{z^{(i)}\}_{i=0}^{m} \sim Cat(T_{f,\mathbb{S}})$
    $\{(x_{f,\mathbb{S}}, y_{f,\mathbb{S}})\}_{i=0}^{m} \sim \mathbb{T}_{f,\mathbb{S}}$; $\{x_{h,\mathbb{S}}\}_{i=0}^{m} \sim \mathbb{T}_{h,\mathbb{S}}$, $\{(x_{\mathbb{S}}, y_{\mathbb{S}})\}_{i=0}^{m} \sim \mathbb{T}_{\mathbb{S}}$
    $\epsilon \sim U[0, 1]$
    ▷ **Generate data**
    $\tilde{x} \leftarrow G_\theta(z)$
    $\tilde{x}_f \leftarrow \tilde{x} + x_{h,\mathbb{S}}$
    $\hat{x} \leftarrow \epsilon x_{f,\mathbb{S}} + (1 - \epsilon)\tilde{x}_f$
    ▷ **Update discriminator D**
    $L_D^i \leftarrow D_w(\tilde{x_f}) - D_w(x_{f,\mathbb{S}}) + \lambda_{GP}((||\nabla_{\hat{x}} D_w(\hat{x})||_2 - 1)^2)$
    $w \leftarrow Adam(\nabla_w 1/m \sum_{i=1}^{m} L_D^i, w)$
  end for
  ▷ **Update classifier C**
  From $\{x_{h,\mathbb{S}}\}_{i=0}^{m}$ form triplets (Schroff et al., 2015) according to label $x_{\mathbb{S},a}$, $x_{\mathbb{S},p}$ and $x_{\mathbb{S},n}$
  $L_C^i \leftarrow max(||C_\gamma(x_{\mathbb{S},a}) - C_\gamma(x_{\mathbb{S},p})||^2 - ||C_\gamma(x_{\mathbb{S},a}) - C_\gamma(x_{\mathbb{S},n})||^2 + \alpha, 0)$
  $\gamma \leftarrow Adam(\nabla_\gamma 1/m \sum_{i=1}^{m} L_C^i, \gamma)$
  ▷ **Update generator G**
  $L_D^i \leftarrow -D_w(G_\theta(z))$
  $L_C^i \leftarrow max(||C(\tilde{x}_f) - C(x_{\mathbb{S},p})||^2 - ||C(\tilde{x}_f) - C(x_{\mathbb{S},n})||^2 + \alpha, 0)$
  $L_G^i = \lambda_D * L_D^{(i)} + \lambda_C * L_C^{(i)}$
  $\theta \leftarrow Adam(\nabla_\theta 1/m \sum_{i=1}^{m} L_G^i, \theta)$
end while
return G_θ

different strategy (feature alignment vs. data generation in our proposed framework). While originally proposed as a completely unsupervised DA method, the authors also conducted experiments on the extreme scenario (where only the healthy class is shared between the two domains). For these experiments, the healthy data label from the target domain was used for alignment (Wang et al., 2020a). We compare our method to both implementations and denote the completely unsupervised implementation as *Unilateral* and the one using the target's healthy label for alignment as *Unilateral**. (3) The adversarial approach *BA3US* balances each batch of target data with randomly sampled source data. It, therefore, presents an interesting comparison method to the proposed *FaultSignatureGAN*, where we balance the target domain with generated data that has been mapped to the target domain in an unsupervised manner. (4) Last, the data generation approach *GenAlign* is used for comparison (Li et al., 2018b) (see Section 5.2), where target data is generated by passing novel input to the generative model. This approach is used to challenge the hypothesis that generative models are limited in their extrapolation abilities and therefore, the novel target data generation should not rely on extrapolation abilities of the model (as we do in our work).

For the *Open-Partial* domain experiments, however, there is no other suitable comparison method that is applicable to the same extreme case scenario as we consider here where only the healthy class is shared between the two domains. Therefore, only the *Baseline* is used for comparison in these experiments.

5.4 Case Studies

The proposed approach is tested on two bearing datasets that have been commonly applied for DA tasks in fault diagnostics in different settings. Our proposed framework is evaluated on both datasets in *Partial* and *Open-Partial* DA experiments. Both datasets are adjusted to the problem formulation to the respective DA setup.

5.4.1 CWRU

The CWRU dataset is a publicly available benchmark bearing dataset (bearing type SKF 6205) provided by the Case Western Reserve University Bearing Data Center (CWRU dataset) (Smith and Randall, 2015). The data was collected on a test rig in laboratory conditions. It contains data recorded under four different loads (referred to as domain 0,1,2 and 3). The different load settings resulting in different rotational speeds are shown in Table 5.1. Data under healthy and nine different faulty conditions is available: Three fault types - Ball, Inner Race and Outer Race - with three severity levels each. An overview of the fault types and severity levels is shown in Table 5.2. The CWRU dataset has been extensively used to demonstrate *ClosedSet* DA methods under different operating conditions as well as for *Partial* DA setups (Wang et al., 2020a; Li and Zhang, 2020).

5.4.2 Paderborn

The Paderborn dataset is a publicly available bearing dataset (bearing type SKF 6203) provided by the Chair of Design and Drive Technology from Paderborn University (Lessmeier et al., 2016). It incorporates both artificially induced bearing faults and realistic damages caused by accelerated lifetime tests (Zhang et al., 2020) under different operating conditions (rotational speed, load torque and radial force) (Chen et al., 2018). In this study, we only consider real fault data and not the artificially induced one. The represented health conditions in the dataset are healthy conditions, Inner Race faults (three severity levels) as well as Outer Race faults (two severity levels). The different operating conditions are shown in Table 5.1 and the different classes in Table 5.2. The data was also collected on a test rig under laboratory conditions and was also previously used in different DA studies (Pandhare et al., 2019; Chen et al., 2018; Wang et al., 2020a). Previous publications mainly focused on 3-class classification of the different fault types (Wang et al., 2020a) and suggested that the domain gaps in the Paderborn dataset are larger compared to the CWRU dataset. Further, previous publications typically neglected the domain 1, since the domain gap to the other domains is considerably large compared to the other domain gaps. In this research, we focus on the type and severity classification (6-class classification) and also aim to bridge large domain gaps. Contrary to previous works, we, therefore, included domain 1 in our DA evaluation.

Moreover, we use less data compared to previous publications, such as e.g. (Wang et al., 2020a), for our evaluation (only using the datasets K002-5; KA04, KA15-16 and KI16,18,21 whereas KA22, KA30, KI04 and KI14 have not been used in this study). This enables us to evaluate if we can extract transferable fault characteristics from only limited fault data.

	CWRU	Paderborn			
Domain	Rotational Speed [rpm]	Rotational Speed [rpm]	Load Torque [Nm]	Radial Force [N]	Setting Name
0	1797	1500	0.7	1000	N15_M07_F10
1	1772	900	0.7	1000	N09_M07_F10
2	1750	1500	0.1	1000	N15_M01_F10
3	1730	1500	0.7	400	N15_M07_F04

Table 5.1: Operating conditions under which the two case studies (CWRU and Paderborn) are recorded. Each setting corresponds to one domain.

		Healthy	Outer Race (OR)			Inner Race (IR)			Ball (B)		
CWRU	Size	-	7	14	21	7	14	21	7	14	21
	Class	0	1	2	3	4	5	6	7	8	9
Paderborn	Extent of Damage	-	1	2	-	1	2	3	-	-	-
	Class	0	1	2	-	3	4	6	-	-	-

Table 5.2: Health conditions represented in the the case studies (CWRU and Paderborn).

5.5 Experimental Setup

To test if *FaultSignatureGAN* is capable of generating unseen domain faults, *Partial* and *Open-Partial* DA experiments are conducted, whereby the different domains correspond to the different operating conditions in the case studies. The experimental setups are shown in Table 5.3 (*Partial*) and Table 5.4 (*Open-Partial*) on an exemplary domain shift from some domain X to some target domain Y ($X \longrightarrow Y$).

Partial

Dataset	Domain Shift	Source Domain	Source Classes	Target Domain	Target Classes during Training
CWRU	X → Y	X	0,1,2,3,4,5,6,7,8,9	Y	0
Paderborn	X → Y	X	0,1,2,3,4,5	Y	0

Table 5.3: Experimental Setup for *Partial* DA on an exemplary domain shift $X \longrightarrow Y$.

Open-Partial

Task	Domain Shift	Source Domain	Source Classes	Target Domain	Target Classes during Training
Source (IR) ⇌ Target (OR)	X → Y	X	0,3,4,5	Y	0,1,2
Source (OR) ⇌ Target (IR)	X → Y	X	0,1,2	Y	0,3,4,5

Table 5.4: Experimental Setup for *Open-Partial* DA on an exemplary domain shift $X \longrightarrow Y$ on the Paderborn dataset.

The experiments are conducted as follows: First, a generative model is trained on data from one domain (as described in Section 5.3 and depicted in **Training Phase** in Figure 5.5). In this work, we train one generative model to generate all severity levels of one fault type. Second, the label space of the target domain is completed by generating synthetic target fault data as depicted in **Execution Phase** in Figure 5.5 based on healthy target data. The number of generated data samples per class is chosen to match the mean number of samples per class in the source domain. In the third step, a new training dataset is composed of the generated and real data from both domains and used to train a classification model. The performance of the classifier is then evaluated on a test dataset composed of all unseen faults and 30% of the class data, from which the conditions have been observed before.

Hyperparameter Tuning: Data-driven solutions based on neural networks come with many hyperparameters to tune including those of the network architecture (layer type, activation, kernel size, initialization etc.). These choices strongly influence the performance of the final model including its generalizability to new data. There is no commonly accepted procedure for optimizing the hyperparameters for an unknown target domain (Bousmalis et al., 2017). Some works rely, therefore, on a target validation dataset (Bousmalis et al., 2017) or validation tasks (Wang et al., 2020a). In many practical applications, especially in the context of safety-critical systems, where no target fault data is available, this is not possible. Hence, in this work, we do not make the assumption of having target data available for hyperparameter tuning, since it is a strong limitation of applying existing DA methods to real PHM applications.

For training the **generative model** (**Training Phase** in Figure 5.5), only criteria related to the source dataset are used: In addition to optimizing the loss functions (see Section 5.3) on the source dataset, a stopping criterion is implemented. The training is stopped if an auxiliary classifier trained on the synthetic source data returns an accuracy of at least 98%, evaluated on the real source data. Since this callback function is computationally expensive, it is only executed after each 50 epochs of training.

Further, the hyperparameters of the final **classification model** need to be tuned as well. In absence of real target fault data, we used synthetically generated data as a validation dataset. To showcase and evaluate the impact that hyperparameter settings have on the ability of the model to generalize to an unseen domain, we trained three different model architectures: Model (1) equals the one used in previous publications (Li et al., 2018b; Wang et al., 2020a) , Model (2) equals Model (1) but has the ReLu activation function and Model (3) equals Model (1) except that the kernel size is set to 12 (compared to 3 used in (Li et al., 2018b)). Exemplary, we only evaluate the domain shifts from source domain 0 on the CWRU dataset for hyperparameter tuning. The final accuracies on a source validation dataset, a synthetic fault dataset as well as on the true target test dataset of the three models are shown in Figure 5.6. The performance on the target dataset varies considerably depending on the architecture used. For example, on domain shift $0 \rightarrow 2$, the final performance on the target dataset varies by 10% depending on the model used. This evaluation shows clearly that even small changes in hyperparameters can impact the generalization ability strongly i.e. have a big effect on the performance in the target domain. The source validation dataset does not provide a good indication which model to choose since the performance on the source dataset always results in 100%. The accuracy of the synthetically generated dataset does not correlate strongly with the target accuracy in all instances. However, it gives a clear indication to choose Model 3 in all instances. This is also the best choice for the highest accuracy on the target dataset in the first two domain shifts. Only for domain shift $0 \rightarrow 3$ this is not an ideal choice. Although not ideal, we want to emphasize that the synthetic validation dataset provides information on which model to choose compared to the source validation dataset. On average, that information results in the best final model choice. Therefore, we conduct our experiments on Model (3) for the CWRU dataset. We train the classification model for 2000 epochs (since the source as well as the synthetic validation datasets suggest that no considerable change happens after 2000 epochs). To enable a better comparison, we use only one model architecture for all domain shifts per case study.

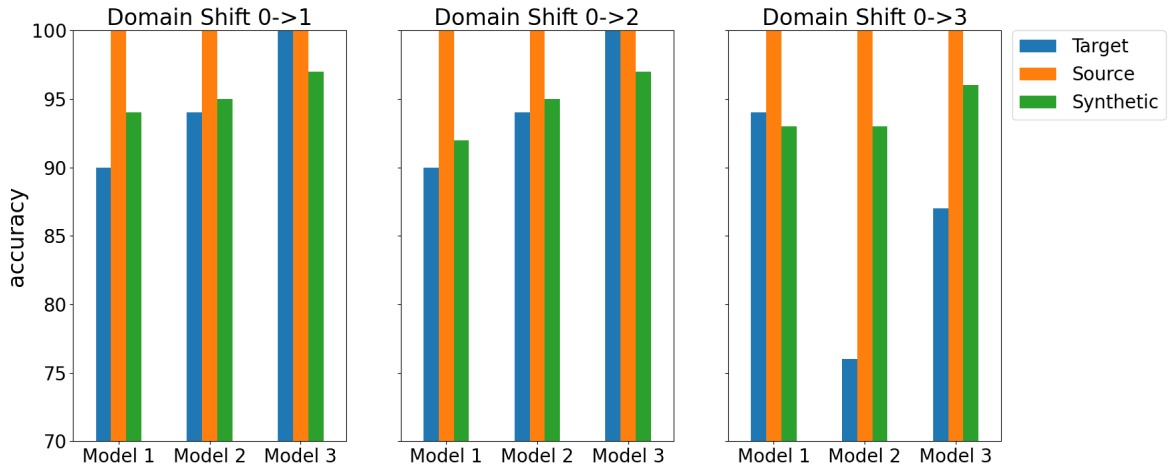


Figure 5.6: Visualisation of the effect of hyperparameter tuning on the generalizability of different model architectures on different domain shifts: Model 1 as in (Wang et al., 2020a), Model 2 as in (Wang et al., 2020a) but with the ReLu activation function, Model 3 as in (Wang et al., 2020a) with the kernel size of 12 (compared to 3 used in (Wang et al., 2020a)). Three different datasets are used for evaluation: 1)source validation dataset (orange); 2) dataset with synthetic faults (green), and 3) the real target dataset (blue).

Apart from using a synthetic validation dataset, we propose to use the following strategy for certain hyperparameters: (1) Applying a heavy regularization (since it leads to better generalization); (2) running the optimization for multiple epochs - more than indicated by the validation dataset. The latter choice is motivated by the findings of learning theory that

hypothesize that there are two phases of deep learning: a fitting and a compression phase. It is indicated that the latter is responsible for the excellent generalization performance of deep networks (Shwartz-Ziv and Tishby, 2017). Even though this hypothesis has been challenged recently (Saxe et al., 2019), we still decided to set the number of optimization steps high.

The final model architectures being used and hyperparameter settings are elaborated in Section 5.9.

Label Availability: Similarly to previous studies conducted on the extreme case of label discrepancy (Wang et al., 2020a; Li and Zhang, 2020), we assume to know the label of the healthy data in the target domain for the *Partial* and *Open-Partial* DA experiments. Since healthy data is ubiquitous, this is considered to be a realistic assumption. In addition, for the *Open-Partial* setup, we assume that if at training data acquisition time fault classes have been observed, we also have the labels for the fault classes, both for source and target domains. This is a particularly realistic setup for fleets of complex systems with different ages for each of the units (each of the domains). Some units will have experienced one subset of fault types and other units will have experienced another subset of fault types. However, at testing time, we would like to be able to diagnose all of the fault types for all of the units.

Data Pre-Processing: To enable a fair comparison, the datasets are pre-processed in the same way as in previous publications (Wang et al., 2020a; Li et al., 2018b). The CWRU datasets are first truncated (at 12000 timesteps) and divided into 200 sequences of 1024 points. After applying the Fast Fourier Transform (Cooley and Tukey, 1965), only the first 512 coefficients are used (excluding the first one).

The same process is applied to the Paderborn dataset. However, the data is not truncated and the 1024 long samples are sliced with a stride of 4096.

5.6 Evaluation and Results

Partial (see Section 5.6.1) as well as *Open-Partial* DA experiments are conducted (see Section 5.6.2) to test the ability of the generated data to bridge domain gaps in different DA settings. Last, to evaluate the physical plausibility of the generated data qualitatively, we visualise the generated data (see Section 5.6.3).

5.6.1 Partial DA

First, we conduct experiments in the extreme case of *Partial* DA where only the healthy class is shared between the domains. The experimental setup is exemplified in Figure 5.7 for the Paderborn dataset, where only the healthy data of the target domain is available. Therefore, we first generate the missing data in the target domain (darker blocks in Figure 5.7) and evaluate on a target test dataset.

(a) Real datasets. (b) Real and generated training datasets. (c) Real test dataset.

Figure 5.7: Example of an extreme experimental *Partial* DA Settings on the Paderborn case study. In the Figure 5.7a The real datasets are shown where only the healthy class is shared between the source and the target dataset and the source dataset has five private classes. In Figure 5.7b The training dataset is shown where the missing fault classes in the target domain are synthetically generated and in Figure 5.7c the test dataset consisting of real target data is shown.

CWRU: We compare the results of the proposed methodology with the methods outlined in Section 5.3.3. If available, the exact results from previous publications are shown. If not, we re-implement the methods while using exactly the same setup as in the original publications.

Only for *Ba3US*, a new architecture needed to be tuned as elaborated in Section 5.9 since this method has not yet been applied to any of the presented case studies. Since hyperparameter tuning without fault data did not lead to satisfying results, we followed the protocol of Wang et al. (2020a) and used the domain shift experiment $0 \longrightarrow 3$ as a validation task. For the source-only experiments, we report on the one hand the previously reported results based on the originally proposed classifier architecture (Wang et al., 2020a) with a kernel size of 3 (referred to as *Baseline*). On the other hand, we report the results of the source-only experiments conducted with the classifier architecture optimized based on the synthetic data as reported in Section 5.5. This architecture has a kernel size of 12 (referred to as $Baseline^{Syn}$). The balanced accuracy of all experiments is shown in Table 5.5. The overall performance of all approaches is high - even for the two source-only *Baselines*, suggesting that the domain gaps are small in this dataset. This also leaves only limited room for improvement. The baseline model with optimized hyperparameters ($Baseline^{Syn}$), however, outperforms the existing *Baseline* by 1.71%, showcasing that the generated data is beneficial for hyperparameter tuning. Moreover, adding the data generated by *FaultSignatureGAN* to the training dataset, results in an additional improvement of 0.96% - resulting in a total improvement of 2.67% compared to the previously reported baseline method. This shows that the generated data is beneficial in order to bridge domain gaps.

The two methods based on data generation used for comparison, ($PixelDA^{+}$ and *GenAlign*), where the generative model is conditioned on novel input data (the real signal in $PixelDA^{+}$ and the target features in *GenAlign*) perform worse than *FaultSignatureGAN*. These results suggest that it is not beneficial to condition the generative model on unseen input and rely on extrapolation abilities of the generative model as in *GenAlign* and *PixelDA*. Therefore, these approaches are not used as comparison methods in the following experiments on the Paderborn dataset. From the two unsupervised adversarial alignment approaches (*Unilateral* and *BA3US*), the *Unilateral* approach performed consistently better on all domain shifts. *Unilateral**, where the label of the healthy target data is used for alignment, results in the highest performance compared to all other approaches. On average, *FaultSignatureGAN* performs within the same range of *Unilateral**. In the following experiments on the Paderborn dataset, only *Unilateral** is used as a comparison method.

Domain Shift	Baseline[1]	$Baseline^{Syn}$	Unilateral[2]	Unilateral*[1]	BA3US	Pixel DA^{+}	Gen Align (Li et al., 2018b)[1]	*Fault Signature GAN*
$0 \rightarrow 1$	93.49±1.75	99.49±0.06	97.04±0.86	98.08±0.16	91.07±3.98	92.45	97.81	**99.87±0.07**
$0 \rightarrow 2$	93.65±0.96	99.96±0.02	96.38±2.34	**99.56±0.18**	91.12±1.28	91.37	96.02	99.36±0.36
$0 \rightarrow 3$	91.02±0.02	90.27±0.69	94.14±0.56	**98.22±0.65**	96.33±1.71	86.01	94.24	94.50±1.10
$1 \rightarrow 0$	97.93±0.93	96.79±0.45	97.48±0.45	98.08±0.32	96.98±1.02	**99.59**	97.27	97.62±0.19
$1 \rightarrow 2$	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	99.95±0.04	99.38	96.32	99.95±0.00
$1 \rightarrow 3$	98.26±1.63	**99.46±0.19**	98.40±0.91	99.20±0.19	99.23±0.66	93.81	94.59	99.35±0.11
$2 \rightarrow 0$	91.63±1.63	96.15±0.15	90.13±3.66	96.43±0.43	93.7±3.59	94.98	95.44	**96.50±0.16**
$2 \rightarrow 1$	97.09±0.09	97.78±0.09	97.84±0.26	97.48±0.40	95.58±1.93	**98.97**	96.55	97.06±0.09
$2 \rightarrow 3$	**99.78±0.17**	99.63±0.12	99.71±0.10	98.97±0.21	99.75±0.21	96.00	96.13	99.63±0.09
$3 \rightarrow 0$	87.96±0.18	88.58±0.19	86.50±4.56	94.85±2.16	86.75±0.21	**96.65**	92.82	92.81±0.92
$3 \rightarrow 1$	89.42±0.42	92.68±0.64	93.22±0.97	**96.18±0.50**	85.53±3.19	94.07	93.04	95.41±0.21
$3 \rightarrow 2$	99.65±0.17	99.68±0.44	99.82±0.04	99.78±0.09	98.68±2.23	98.72	95.49	**99.99±0.01**
Mean	94.99	96.70	95.88	**98.07**	94.56	95.17	95.49	97.66

[1] Results as reported in the original publication.
[2] Models used as in original publication for reproducing results.

Table 5.5: Extreme *Partial* DA results on the CWRU dataset (10-class classification) under all domain shifts.

Paderborn: In this case study, we only use the best performing comparison DA method *Unilateral**. We neglect those that were not performing well on the CWRU case study. Moreover, contrary to other publications on the Paderborn case study that focus only on fault type classification (3-class classification), we focus on the task of fault type and severity classification in this work (6-class classification). Since the size of the domain gap differs considerably from *domain 1* to the other three domains (as indicated by the *Baseline* results), we report the results for all DA tasks related to the domain 1 separately. Please note that in

previous DA studies, the results on these DA tasks were never reported (Wang et al., 2020a). Therefore, we report the results separately. In the upper part of Table 5.6, only the results on *domains 0,2* and *3* are reported (smaller domain gaps) and in the lower part all results on DA shifts related to *domain 1* are reported (large domain gap).

The DA approaches (*Unilateral** and *FaultSignatureGAN*) outperform the *Baseline* on all domain shift experiments - see Table 5.6. While the performance gain is comparable between the two approaches on *domains 0,2* and *3*, *FaultSignatureGAN* results in a considerably better performance on *domain 1*, where the domain gap is large (as indicated by the low *Baseline* performance). In all settings, there is a substantial relative gain. On *domains 0,2,* and *3*, an average improvement of 3.82% was achieved by *FaultSignatureGAN* compared to the *Baseline*. On *domain 1*, the relative improvement is even 23.76%. The absolute performance differs between the different domain shift experiments: If *domain 1* is the target domain, the absolute performance of all approaches is still rather low ($< 50\%$) despite the relatively high improvement. In the opposite direction, when *domain 1* is the source domain, higher absolute results were achieved (average performance of the three domain shift experiments with *FaultSignatureGAN* is 79.72%). Although the domain gap should be the same in both directions (domain as source or target), this difference in the performance could potentially be explained if the fault data in *domain 1* shows more variability compared to the other domains. This leads to better generalization on tasks from *domain 1* to other domains. Especially in these instances, *FaultSignatureGAN* shows a superior performance.

Domain Shift	Baseline	Unilateral*	*FaultSignatureGAN*
$0 \rightarrow 2$	**99.78±0.06**	99.92±0.02	99.76±0.24
$0 \rightarrow 3$	69.49±0.73	69.98±1.73	**72.08±1.24**
$2 \rightarrow 0$	99.42±0.01	**99.67±0.08**	99.54±0.16
$2 \rightarrow 3$	74.66±0.40	**75.71±1.69**	75.49±0.23
$3 \rightarrow 0$	67.43±0.6	71.29±1.61	**78.87±0.65**
$3 \rightarrow 2$	68.37±1.5	**77.14±1.42**	76.35±0.42
Mean	79.86	82.29	**83.68**
$0 \rightarrow 1$	22.88±1.51	29.37±1.20	**45.88±2.21**
$1 \rightarrow 0$	58.66±1.71	74.54±0.53	**84.34±0.21**
$1 \rightarrow 2$	63.28±1.77	75.16±3.56	**86.34±0.77**
$1 \rightarrow 3$	47.99±0.22	61.87±1.77	**68.50±0.51**
$2 \rightarrow 1$	21.77±0.53	29.96±1.83	**47.59±1.66**
$3 \rightarrow 1$	22.77±1.31	26.47±0.24	**45.30±0.61**
Mean	39.23	49.56	**62.99**

Table 5.6: Extreme *Partial* DA results on the Paderborn dataset (6-class classification). In the upper part all results with domain shifts including *domains 0,2* and *3* are shown and in the lower part all domain shifts including *domain 1*.

5.6.2 *Open-Partial* Domain Experiments

(a) Real datasets. (b) Real and generated training datasets. (c) Real test dataset.

Figure 5.8: Example of an extreme experimental *Open-Partial* DA Settings on the Paderborn case study. In the Figure 5.7a the real datasets are shown where only the healthy class is shared between the source and the target dataset and the source dataset has five private classes. In Figure 5.8b The training dataset is shown where the missing fault classes in the target domain are synthetically generated and in Figure 5.8c), the two test datasets are shown where one consists of the real source and the other of the real target data.

To showcase the versatility of our framework, we conduct *Open-Partial* DA experiments in addition to the *Partial* DA experiments (Section 5.6.1). These experiments are only conducted on the Paderborn dataset since the domain gaps are larger compared to the CWRU dataset. The other DA methods used for comparison for the *Partial* DA setup are not directly applicable for the *Open-Partial* setup. Therefore, we only report the baseline results for comparison. Figure 5.8 depicts an example for the the experimental setup: For the *Open-Partial* DA experiments, we assume that in each of the two domains, one fault type occurred with different severities. The outer race fault with severity 1 and 2 (OR1 and OR2) occurred in the target domain, whereas the inner race fault (severity 1, 2 and 3; IR1, IR2 and IR3) occurred in the source domain. Hence, in a first step, two generative models are trained. The fault signature of the outer race fault is trained on the target data, the fault signature of the inner race fault on the source data (see **Training Phase** in Section 5.3). In a second step, the missing fault data is generated: In the example of Figure 5.8, the outer race fault is generated for the source domain and the inner race fault classes with severity 1,2 and 3 for the target domain. This generated and real data composes the training dataset. Usually, only the performance on the target dataset is evaluated. However, in the experimental setup for *Open-Partial* DA, there is missing data in each of the domains. Therefore, we evaluate the performance on two test datasets - the source $\mathbb{S}$ and the target $\mathbb{T}$. The test datasets comprise the real missing fault data as well as of a 30% of known health conditions. The results on the 6-class classification task are reported in Table 5.7. The baseline was trained on all available real data from the source and target domain.

The experiments show that the synthetically generated data enables to achieve a good classification performance on all settings ($> 90\%$) excluding *domain 1*, by far exceeding the performance of the *Baseline* method of 83.37% on average. On the DA task related to *domain 1*, the absolute performance of the classifier is considerably lower, however, it still results in a large relative improvement in all instances compared to the *Baseline* method.

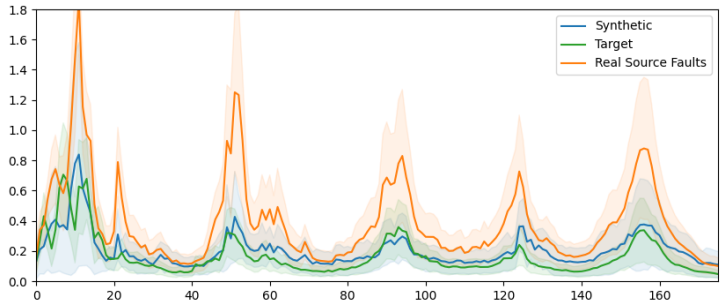
	Domain Shift		Baseline	Proposed			Baseline	Proposed
Source (IR) ⇌ Target (OR)	0 ⇌ 2	$\mathbb{S}$	99.41±0.15	**99.81±0.02**	Source (OR) ⇌ Target (IR)	$\mathbb{S}$	99.12±0.42	**99.96±0.05**
		$\mathbb{T}$	99.88±0.01	**99.97±0.02**		$\mathbb{T}$	99.53±0.14	**99.85±0.10**
	0 ⇌ 3	$\mathbb{S}$	73.65±0.13	**91.74±0.05**		$\mathbb{S}$	78.99±0.69	**91.59±0.18**
		$\mathbb{T}$	72.80±0.51	**95.33±0.01**		$\mathbb{T}$	75.41±0.48	**96.18±0.75**
	2 ⇌ 3	$\mathbb{S}$	75.30±1.04	**93.99±0.61**		$\mathbb{S}$	78.10±1.17	**97.13±0.70**
		$\mathbb{T}$	72.93±1.16	**94.82±0.23**		$\mathbb{T}$	75.31±0.81	**91.67±0.77**
Mean		$\mathbb{S}$	82.79	**95.18**		$\mathbb{S}$	85.40	**96.22**
		$\mathbb{T}$	81.87	**96.71**		$\mathbb{T}$	83.42	**95.90**
Source (IR) ⇌ Target (OR)	0 ⇌ 1	$\mathbb{S}$	56.91±1.44	**66.55±0.61**	Source (OR) ⇌ Target (IR)	$\mathbb{S}$	51.53±1.43	**53.90±1.33**
		$\mathbb{T}$	65.53±0.09	**71.04±0.22**		$\mathbb{T}$	69.06±1.28	**81.48±1.53**
	1 ⇌ 2	$\mathbb{S}$	53.62±0.23	**56.21±1.18**		$\mathbb{S}$	51.93±1.21	**65.65±1.10**
		$\mathbb{T}$	69.71±0.16	**76.08±1.41**		$\mathbb{T}$	65.76±0.26	**67.54±0.34**
	1 ⇌ 3	$\mathbb{S}$	51.83±1.66	**66.10±1.40**		$\mathbb{S}$	63.66±0.56	**67.64±0.15**
		$\mathbb{T}$	66.50±0.02	**74.75±0.65**		$\mathbb{T}$	65.25±0.06	**71.14±0.62**
Mean		$\mathbb{S}$	54.12	**62.95**		$\mathbb{S}$	55.71	**62.39**
		$\mathbb{T}$	67.24	**73.96**		$\mathbb{T}$	66.69	**73.39**

Table 5.7: *Open-Partial* DA results on the Paderborn dataset (6-class classification). In the upper part all results with domain shifts including domains 0,2 and 3 are shown and in the lower part all domain shifts including domain 1. As shown in Figure 5.8, the trained classification model is evaluated on two datasets: The source test dataset ($\mathbb{S}$) and the target test dataset ($\mathbb{T}$).

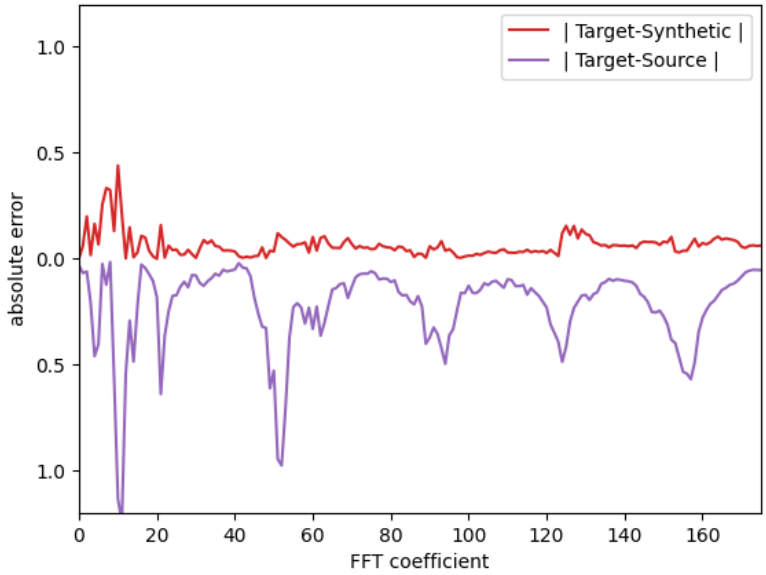
5.6.3 Qualitative evaluation

To evaluate the physical plausibility qualitatively, we visualise the mean of the generated signals (blue line in Figure 5.9a), of the true faults in the target domain (orange line in Figure 5.9a) and of the true faults in the source domain (green line in Figure 5.9a). A batch of 1000 data samples was used for the illustration including its standard deviation. Exemplary, we chose the *fault type OR* and *severity 1* on the domain shift 0 → 1 on the Paderborn

dataset. To better visualize the differences, the residual of the generated target to the true target mean signal is visualized as well as the residual of the true source to the true target (which can be considered as the baseline) - see Figure 5.9b. The proposed framework appears to generate the true target data considerably well. It performs substantially better compared to the baseline (just using the source faults without any adaptation for the target domain). Especially in the higher frequency range, it represents the true target faults noticeably better than the true source faults.



(a) Mean Fault Signal OR Severity 1 of the synthetic data (blue line), the true target fault class data (orange line) and the true source fault class data (green line).



(b) Mean Absolute Residual Signal OR Severity 1 of the Target to Synthetic (red line) and Target to Source (purple line).

Figure 5.9: Paderborn data visualization of the OR severity 1 fault comparing real fault data with generated fault data.

5.7 Discussion

The experiments performed in this research demonstrate the validity of the proposed framework *FaultSignatureGAN* to generate previously unobserved fault data, that can be used for DA with different types of extreme label discrepancies, where only the healthy class is shared between the domains. The obtained results open interesting points for discussion.

***FaultSignatureGAN* for DA with label space discrepancies:** Given small domain gaps, *FaultSignatureGAN* outperforms most of the comparison methods, especially when comparing the results to *GenAlign*, the other generative approach . This particularly sup-

ports the assumption that the unsupervised generation of unseen data should not rely on extrapolation abilities of the generative model (as it does for the comparison methods). Instead, our approach enables the generation of unseen faults building on the hypothesis that domain-specific fault data can be disentangled in domain-specific characteristics and class specific ones and therefore, requires no extrapolation ability of the generative model. The comparison method *BA3US* outperforms *FaultSignatureGAN* solely on domain shift $0 \longrightarrow 3$ (by 1.83%), where target data has been used to tune the hyperparameters of *BA3US*. In practical safety-critical applications, target data to tune the hyperparameters of a model is typically not available. Therefore, *FaultsignatureGAN* did not rely on this information and thus, satisfies more realistic requirements for PHM applications.The improvement that *BA3US* provided on that one domain shift could not be translated to other domain shifts. This, once again, showcases the importance of hyperparameter tuning for DA methods based on feature alignment and in particular the importance of the access to fault data for hyperparameter tuning. Only the feature alignment approach *Unilateral** provides a similar performance as *FaultSignatureGAN* under small domain gaps in *Partial* DA settings (see all results on CWRU in Section 5.6.1 and on Paderborn with domains 0,2 and 3). This is not surprising since synthetic data generation is never perfect. Therefore, when the domain gap is small, the source data represents the target data already quite well and one would expect little benefits in generating synthetic target specific data. On large domain gaps (those including domain 1 on the Paderborn dataset), however, the performance of the feature alignment method *Unilateral* drops. These are the scenarios where the proposed generative approach *FaultSignatureGAN* outperforms other approaches (see Section 5.6.1 and Section 5.6.2). Therefore, if the size of the domain gap is unknown, *FaultSignatureGAN* is the best option to choose in safety-critical systems since it provides a comparable performance under small domain gaps but a considerably better performance under large domain gaps.

Versatility of *FaultSignatureGAN*: Many different scenarios of label discrepancies are possible in real operations as exemplified in Section 5.1. Having one versatile method that can be applied in multiple of these scenarios is, therefore, utterly important for practical applications. The versatility of the proposed approach is demonstrated by applying it successfully to DA experiments with different types of label discrepancies (*Partial* and *Open-Partial*) (if the respective labels are known in the source domain), where it consistently outperforms other comparison methods under large domain gaps. We consider the versatility of the proposed approach as a one of the key benefits for practical PHM applications.

Plausibility of Unsupervised Data Generation and Validity of the Underlying Hypothesis: The generation of unseen target data is unsupervised and the plausibility of the target data cannot be directly imposed while training the generative model. Therefore, it is required to evaluate how realistic the target data generated by *FaultSignatureGAN* is; to which extent it can be used as a surrogate of real target data. The data visualization (see Section 5.6.3) shows that the generated data represents real target data well. In particular, it represents the target fault data substantially better compared to the source data. This finding is also supported by the findings in the DA experiments (both *Partial* and *Open-Partial*) where *FaultSignatureGAN* consistently outperforms the **Baseline** method. This supports the validity of the underlying hypothesis to enable controlled generation of previously unobserved data. We can draw the following conclusions: (1) Equation 5.1 serves as a good approximation of real fault data, (2) the disentanglement of domain-specific and fault-specific characteristics was successful and (3) that domain-invariant fault signatures can be extracted by *FaultSignatureGAN* given only one source domain. However, our assumption about the structure of domain-specific and fault-specific components composing real fault data as defined in Equation 5.1 could be extended and further refined in future work, in particular, the assumption that the OCs impact the fault-specific components linearly. Moreover, the DA experiments show that the generative process succeeds in preserving the semantic meaning

of the generated data. If this would not be the case, the generated data would introduce label noise to the training data and, quite likely, result in a performance drop in the target domain.

Synthetic Data for Hyperparameter Tuning: Our results have shown that the classification performance of the *Baseline* method in the target domain is highly dependent on the chosen classifier architecture (see evaluation in Section 5.5). This is also observed in the literature, where different *Baseline* results on exact same tasks are reported with different hyperparameters of classifier architectures. For example, Wang et al. (2020a) reported a mean performance 99.78% on the domain shift 2 → 3, whereas (Li et al., 2020b) reported a baseline performance of 92.2% using a different classifier architecture. The classification improvements in the target domain gained by an appropriate choice of hyperparameters is even larger compared to improvements gained by other DA methods. This emphasizes the importance of hyperparameter tuning including the choice of the network architecture for the task of DA with extreme label space discrepancies. Previous publications, therefore, rely, for example, on the availability of target data and labels for one domain shift experiment. This availability of any target faults is not realistic in real safety-critical applications, where faults did not occur. In absence of fault target data, there is no possibility to tune these hyperparameters with respect to the classification task in the target domain, which can pose a major risk in safety-critical assets. If, however, synthetic data is available that represents the real target data well, the data can be used for validation. In this study, we showed that synthetic fault data generated by *FaultSignatureGAN* can support selecting the optimal architecture without relying on real target fault data that is usually not available in real satefy-critical applications. Herein lays one major benefit of the proposed data generative approach *FaultSignatureGAN*. Although a proof of optimality is impossible (as the real target data has not been observed), the synthetic data provides a better indication of which hyperparameters to choose compared to the hyperparameter choice based on the source dataset performance or even a random choice.

Decreasing Data Acquisition Time: In practice, a short data acquisition phase is essential to enable to start monitoring the condition of a new asset within a short period of time. However, faults are extremely rare in complex (safety-critical) systems. This lack of real fault data is a major limitation to applying data-driven solutions for fault diagnostics. *FaultSignatureGAN* allows to transfer fault patterns to a new target domain. Once a fault occurred in one domain providing sufficient fault data to train a generative model, the fault signature can be learned, which then can be used to generate new fault data for any newly emerging domain. This ultimately can speed up the data collection process significantly, enabling the application of data-driven solutions within a shorter time span.

5.8 Conclusion

In this research, we proposed the *FaultSignatureGAN* framework for controlled generation of unseen faults in the target domain. The resulting generated fault data is (1) specific to a desired domain and (2) specific to a certain fault type and the severity level of the fault in that domain. Therefore, *FaultSignatureGAN* enables to start monitoring the condition of new assets without any faults observed in the target domain since plausible faulty data can be generated for all future target domains. While we considered different operating conditions as domains in this research, the proposed framework is also applicable to generate synthetic faults in new units of a fleet.

We demonstrated the potential of the *FaultSignatureGAN* to complement partial label spaces in different DA experiments - *Partial* as well as *Open-Partial* DA settings. The results show that the generated data represents true faults in the target domain considerably better than the source fault data, leading to an improved classification performance on the target domain. Our proposed method excels particularly on large domain gaps. *FaultSigna-*

tureGAN also enabled hyperparameter tuning for unseen target domain which can be applied in combination with any other DA approach. Without any access to target faults, tuning existing methods optimally is not possible. This demonstrates one of the benefits of plausible data generation in the evaluated tasks.

For future work, an additional step integrating real but unlabeled target data in addition to the synthetically generated data is an interesting direction to explore. Additional unsupervised or semi-supervised DA approaches could be employed to bridge the synthetic to real gap. Furthermore, the transferability of the generated fault signatures between different bearing types is of high interest for future research. One further direction of future research would be to investigate the source data demand for *FaultSignatureGAN*, evaluating how many samples and how diverse they need to be in order to train a representative generative model. On a bigger scale, the integration of novel or evolving fault detection (those that have not been observed neither in the source nor in the target domain) in addition to the performed fault classification would be of a significant practical relevance.

5.9 Appendix

Unless stated otherwise, the following model architectures were used:

Generation Model: The first layer of the generation model is a single neuron. The activation of this neuron is sampled from a categorical distribution corresponding to the number of fault classes (fault type severities). The second fully connected layer is the sampling layer (mean and variance), containing three units each, activated by LeakyReLu (α=0.001).

The following fully connected layers successively increase the dimensionalty to the desired final output shape. Each layer is activated by LeakyReLu (α=0.001), using no bias and followed by a BatchNormalization (BN) layer.

Three 1D convolutional layers follow, each layer is activated by LeakyReLu (α=0.001), and followed by a BatchNormalization (BN) layer.

At last the generated signal is added to a randomly drawn data point from the base dataset.

The Adam optimizer used with a learning rate of 0.0001, $beta_1 = 0.5$ and $beta_2 = 0.999$.

Triplet Encoder Model: The triplet encoder model consists of 6 fully connected layers, each activated with Leaky ReLu (alpha=0.1) and followed by a dropout layer (rate=0.4). The final layer is 4 dimensional and is L2 normalized.

The Adam optimizer used with a learning rate of 0.0001, $beta_1 = 0.5$ and $beta_2 = 0.999$.

Discriminator Model: The discriminator model consists of six fully connected layers, each activated with Leaky ReLu (alpha=0.1) and followed by a dropout layer (rate=0.1). The final layer is 1-dimensional. The Adam optimizer used with a learning rate of 0.0001, $beta_1 = 0.5$ and $beta_2 = 0.999$.

Classification Model for Early Stopping: It consists of 4 1-D convolutions layers (8 filters in each layer and kernel size is 3), each activated with Leaky ReLu (alpha=0.1) and followed by a dropout layer (rate=0.1). Followed by a flattening layer and a fully connected layer with the appropriate number of units according to the number of classes in the dataset.

The Adam optimizer used with default parameters.

Classification Model for Evaluation: The classification model for evaluation is inspired by Wang et al. (2020a) . It consists of three 1D convolutional layers (10 filters in each

layer, activated by ReLu) and dropout layers (0.4). The Adam optimizer used with default parameters.

For training the CWRU classification models a batch size of 64 is chosen, for Paderborn 2000.

Model Architecture for the comparison method BA3US: The comparison method *BA3US* has not yet been applied to any timeseries data. Without using target fault data, the methodology could not be tuned to give satisfying results. Therefore, we followed the procedure of Wang et al. (2020a) and tuned *BA3US* on a validation task 0⟶3. We started using the exact same generator, discriminator and classifier architecture as well as optimizer setting as proposed by Wang et al. (2020a). All hyperparameters (model architecture and weighting of the different loss terms) are then consecutively optimized on the validation task. Ultimately, the following architecture was used: The **feature extractor** consists of a 3 layer 1D-convolutional layer with kernel size 3 and 10 filters per layer. Each layer is batch normalized and activated by the sigmoid function, followed by a dropout layer (rate= 0.5). Last, based on the flattened activations, a fully connected layer is added with 256 units. The **classifier** model consists of two fully connected layers. The first with 256 units is activated with the ReLU activation function and followed by a dropout layer (rate 0.5). The second contains 10 units (corresponding to the number of classes) and is activated by the softmax function. The **discriminator** contains three fully connected layer, each with 256 ReLU activated units. Only the last layer contains only one unit and is Sigmoid activated. The model is optimized on batches of 64 samples in the target and the source domain using the StochasticGradientDescent algorithm with a learning rate of 0.005. The initial ratio of augmented source samples is set to 1.0 ($\rho_0 = 1$), the test interval N_u is set to 50. The loss conditional entropy loss is weighted with a factor of 10^{-3} and the transfer loss with a factor of 10^{-1}. The weighted complement entropy loss is not considered since it did not lead to satisfying results ($w = 0$).

6 Improving generalization of deep fault detection models in the presence of mislabeled data

This chapter corresponds to the published article:[1]

Rombach, Katharina, Gabriel Michau, and Olga Fink (2020). "Improving generalization of deep fault detection models in the presence of mislabeled data". In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, pp. 3103–3110. DOI: https://doi.org/10.1109/SMC42975.2020.9283002.

Abstract: Mislabeled samples are ubiquitous in real-world datasets as rule-based or expert labeling is usually based on incorrect assumptions or subject to biased opinions. Neural networks can "memorize" these mislabeled samples and, as a result, exhibit poor generalization. This poses a critical issue in fault detection applications, where not only the training but also the validation datasets are prone to contain mislabeled samples. In this work, we propose a novel two-step framework for robust training with label noise. In the first step, we identify outliers (including the mislabeled samples) based on the update in the hypothesis space. In the second step, we propose different approaches to modifying the training data based on the identified outliers and a data augmentation technique. Contrary to previous approaches, we aim at finding a robust solution that is suitable for real-world applications, such as fault detection, where no clean, "noise-free" validation dataset is available. Under an approximate assumption about the upper limit of the label noise, we significantly improve the generalization ability of the model trained under massive label noise.

6.1 Introduction

Many real-world datasets exhibit label noise (Krishna et al., 2016). In practical applications of fault detection, labels for distinguishing between healthy and faulty conditions are often generated by predefined rules or else based on assumptions. For example, a system is considered to be healthy within a defined period of time after a performed maintenance action. However, this assumption does not always hold, which results in mislabeled samples in both the training and validation datasets. While deep neural networks (NN) have been applied successfully in the field of Prognostics and Health Management (PHM) (Abdeljaber et al., 2017; Krummenacher et al., 2017), their performance is heavily impacted if trained on a dataset with label noise. As universal approximators, NNs are capable of fitting to any labels (Zhang et al., 2017a). This ability is referred to as "memorization" (Arpit et al., 2017) and leads to poor generalization of the resulting models. In the absence of a clean validation dataset, this lack of generalization cannot be detected since the model might exhibit good performance on a validation dataset that is impacted by the same label bias as the training dataset. However, the model may not have learned the true relationship between input and

[1] Please note, this is the author's version of the manuscript published in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. Changes resulting from the publishing process, namely editing, corrections, final formatting for printed or online publication, and other modifications resulting from quality control procedures may have been subsequently added. The final publication is available at https://doi.org/10.1109/SMC42975.2020.9283002.

output. This is especially problematic in the context of fault detection in industrial assets, where faults are safety-critical.

In this work, we tackle the challenge of training a robust model despite the presence of label noise and without exact knowledge about the label noise or a clean dataset with which to perform hyperparameter (HP) tuning. We propose a two-step framework that first identifies outliers based on the samples' consistency with the hypothesis update and second, modifies the training dataset based on the identified outlier samples. An adaptation of the data augmentation technique called *mixup* (Zhang et al., 2017b) is introduced for the data modification. We aim at providing universally applicable recommendations for learning under label noise.

To the best of our knowledge, this is the first study to tackle the inability to tune certain HP if no reliable ground truth information is available. Our proposed solution relies only on a rough assumption regarding the level of label noise. Ultimately, we significantly improve the generalization ability of the trained models under massive label noise on an image dataset and a time series dataset for fault detection.

After reviewing the existing relevant literature in Section 6.2, the proposed framework is introduced in Section 6.3. The methodology is evaluated on the experimental setup as defined in Section 6.4 and the results are shown in Section 6.5. Based on the discussion in Section 6.6, final recommendations about training deep models with label noise are given in Section 6.7.

6.2 Related Work

Robust learning on noisy datasets has attracted increasing attention - especially in the context of deep learning. In the literature review, we aim at giving a brief overview of different approaches and their limitations with respect to the scenario relevant for fault detection. We elaborate in more detail only closely related methods. For a detailed survey on classification under label noise, the reader is referred to (Frénay and Verleysen, 2013).

Direct approaches aim at detecting mislabeled samples explicitly. E.g. all samples are ranked by their probability of being assigned to the original label (based on the current model's prediction) and a fraction α of this ranked list is then presented to experts for relabeling (Müller and Markert, 2019). Hence, this approach i.a. relies on human intervention.

Other approaches based on the model's prediction aim at automatic relabeling (Tanaka et al., 2018; Reed et al., 2014). However, these tend to favor trivial solutions to the classification task. To counteract this, they require prior knowledge, such as the prior distribution over all classes (Tanaka et al., 2018). In addition to the model's prediction, logits (Pleiss et al., 2020) and the training loss (Shen and Sanghavi, 2018) were also considered to identify mislabeled samples.

As an alternative to relabeling, several researchers have proposed altering the current model's prediction to match the label noise as in (Vahdat, 2017; Patrini et al., 2017; Sukhbaatar et al., 2015). Yet, some approaches presuppose e.g. the ground truth noise model (Vahdat, 2017; Patrini et al., 2017). Others aim to learn this model but do not carefully focus on the memorization ability of DNNs (Tanaka et al., 2018). We argue that a clean validation dataset is required to tune crucial HP such as when to start updating the noise model Q (Sukhbaatar et al., 2015).

Learning to reweight samples of a noisy dataset has been proposed in the field of meta-learning (Jiang et al., 2017; Ren et al., 2018). For example, a meta-gradient step was proposed in (Ren et al., 2018) to reweight samples by evaluating the gradient directions based on a noise-free dataset before the network is updated. These methodologies usually rely on a small, noise-free validation dataset (Jiang et al., 2017; Ren et al., 2018) and are therefore not applicable in our setting.

Indirect approaches deal with mislabeled data only implicitly. They aim at robust optimization in general, resulting in good generalization of the model despite the presence of

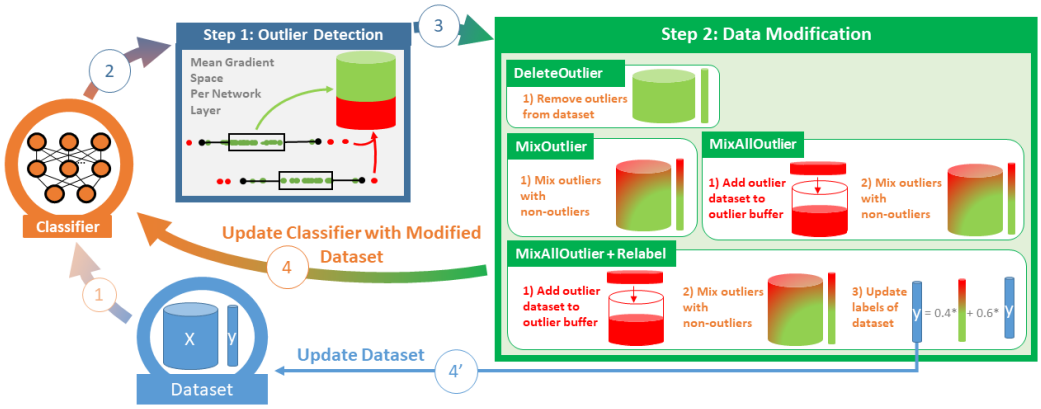


Figure 6.1: Process Flow.

mislabeled samples. Modified loss functions have been proposed, including generalized cross entropy (Zhang and Sabuncu, 2018) and an information-theoretic loss function (Xu et al., 2019). Also, different regularization techniques have been shown to yield good generalization (Hu et al., 2020; Arpit et al., 2017). Each of the proposed approaches comes with a set of specific HP. We argue that tuning these optimally requires a clean validation dataset, which is not available in the scenario considered here.

Vicinal Risk Minimization (VRM) (Chapelle et al., 2001) has been proposed as an alternative to Empirical Risk Minimization (ERM) (Vapnik, 1998), and not only in the context of label noise. It relies on the assumption that the true density function of the data is smooth in the vicinity of any data point and therefore opts to represent it by a vicinity distribution instead of the empirical one as in ERM (Vapnik, 2013). Recently, VRM has been shown to stabilize the training of NNs on noisy data with *mixup* (Zhang et al., 2017b). The data is augmented by drawing samples from a generic vicinal distribution, resulting in convex combinations of the datapoints and their respective labels. Thus, linear behaviour between the classes is favored, which has been shown to prevent the model from overfitting to individual mislabeled samples.

In this work, we tackle the problem in a more general context compared to previous works, by relaxing two strong assumptions that do not hold for many practical applications: 1) we rely neither on a clean dataset for tuning the methodology nor 2) on the exact knowledge of the label noise. We only assume a rough upper estimation of the noise level. The proposed end-to-end approach does not require any human intervention. It combines elements of both direct approaches (outlier detection (OD)) and indirect approaches since the detected outliers are used to adapt the model training. Since *mixup* has shown good performance in other contexts and introduces only one additional HP that can be related to the noise level estimation, we also evaluate its performance and compare it to the proposed approaches.

6.3 Methodology

6.3.1 Problem Formulation

In a classification task, we aim at finding a function h that captures the true relationship between a variable X and a label Y, which follow the joint distribution $P(X, Y)$. In reality, only a finite number of samples of this joint distribution are available - a finite dataset $\mathcal{D}$. The set of functions $h \in \mathcal{A}(\mathcal{D})$ that can be reached depends i.a. on the provided dataset $\mathcal{D}$ (Arpit et al., 2017). Given a training dataset $\mathcal{D}'$ with unknown label noise, we aim at finding a deep model h that performs well on the underlying true but unknown data distribution $P(X, Y)$. Since no ground truth is available, all HP can only be tuned on $\mathcal{D}'$.

6.3.2 Proposed Framework

We introduce a novel two-step framework for robust training with label noise (see Figure 6.1). In a first step, we identify a set of outliers including mislabeled samples (see Section 6.3.4 and **Step 1** in Figure 6.1). The novel algorithm aims at **early detection** based on the gradient update in the hypothesis space, i.e. mislabeled samples are identified before they have been considered for the model update in order to prevent the NN from overfitting to these samples.

In a second step, the training data is modified based on the previously identified outliers. A new adaptation of the data augmentation technique called *mixup* is proposed. In Equation 6.1,

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda)x_j \\ \tilde{y} &= \lambda y_i + (1-\lambda)y_j \end{aligned} \tag{6.1}$$

the original *mixup* augmentation of (x_i, y_i) is defined as proposed in (Zhang et al., 2017b), where $\lambda \sim Beta(\alpha, \alpha)$ and the sample-label pairs (x_j, y_j) are randomly chosen. The novel adaptation of the data augmentation is explained in Section 6.3.4. Multiple data modification techniques, including the automatic relabeling of the training dataset, are proposed in Section 6.3.5 (see**Step 2** in Figure 6.1).

6.3.3 Assumptions

We aim at developing a universal approach for the scenario in which label noise is suspected but no ground truth information is available. In such situations, necessity compels us to make only rough assumptions. Hence, we make the following rough assumptions regarding the label noise:

1. the number of mislabeled samples is less than 50% of the entire dataset
2. we use a rough estimate of the upper limit of label noise (little, medium, massive) to set a maximum threshold of possibly detected outliers and to set the HP

We argue that this is a very loose assumption compared to those required in previous studies where, e.g. the exact label noise must be known (Tanaka et al., 2018). We also evaluate extreme scenarios in which the estimated upper noise limit is 10-20% above the actual noise ratio. Furthermore, we assess the sensitivity of the proposed framework to these assumptions. To evaluate potential limitations, the scenarios are evaluated for cases in which the assumptions concerning the noise level are wrong (see Section 5.7).

6.3.4 Outlier Detection

Outliers are defined as samples that are inconsistent with the update in the hypothesis space, i.e. samples with gradients surpassing certain thresholds as defined in Algorithm 2. The gradient ∂_i^k in the parameter space for all samples i and nodes k is calculated (see line 7 in Algorithm 2), representing the gradient distribution in the parameter space. The thresholds are set based on the confidence interval of the gradient distribution: wh_{lower}^l and wh_{upper}^l are set to be below the 25th percentile and above the 75th percentile by a factor of 1.5 of the Interquartile Range (IQR) (see lines 9 - 13 in Algorithm 2). Furthermore, to enable convergence, a minimal threshold is set for the IQR. The gradients are calculated with respect to the cross-entropy loss as it emphasizes difficult samples (Zhang and Sabuncu, 2018). For all proposed approaches, hard labels (the maximum-a-posterior (MAP) estimates) are used for the gradient calculation. Furthermore, we use assumptions about the upper limit of the label noise to set a maximum number of outliers that can be detected. This limit is defined in Table 6.1. All samples that are not detected as outliers are considered consistent in updating the hypothesis.

Computational Complexity: Since the parameter space of NNs can be very high-dimensional, the proposed outlier selection algorithm is computationally expensive. Thus, we propose to represent the gradient distribution over all samples in a compact way by only considering the mean gradient per layer $l \in L$ of the NN, defined as

$$\partial^l_{i,mean} = \frac{1}{K} \sum_{k \in K} \left(\frac{\partial f_i(w)}{\partial w_{k,l}} \right), \tag{6.2}$$

where f is the loss function, $k \in K$ the nodes in the respective layers $l \in L$ and w the weights of the NN.

Outliers are defined as samples whose mean gradient surpasses the defined thresholds in any of the layers. Furthermore, we only consider the weights of the NN and neglect the biases. In this research, the outlier selection is performed for each class individually. It is important to note that this is not necessary but rather a design choice. The per-class detection enables the approach to be applied to imbalanced datasets as well. This makes the proposed approach more universally applicable.

Algorithm 2 Outlier Detection

procedure DetectOutlier($\mathcal{D}^{'}$, h)
  C: # classes $\in \mathcal{D}^{'}$
  L: # layers $\in h$
  for all $c \in C$ **do**
    for all $l \in L$ **do**
      for all $(x_i, y_i) \in c$ **do**
        $\partial^l_{i,mean} = \frac{1}{K} \sum_{k \in K} \left(\frac{\partial f_i(w)}{\partial w_{k,l}} \right)$
      end for
      $p^l_{25} = \text{Percentile}_{25}(\{\partial^l_{i,mean}\}_{i \in c})$
      $p^l_{75} = \text{Percentile}_{75}(\{\partial^l_{i,mean}\}_{i \in c})$
      $IQR^l = p^l_{75} - p^l_{25}$
      $wh^l_{c,low} = p^l_{25} - 1.5 * IQR^l$
      $wh^l_{c,up} = p^l_{75} + 1.5 * IQR^l$
    end for
  end for
  $\mathcal{O} = \{i \mid i \in c; \exists l \in L; \partial^l_{i,mean} \notin [wh^l_{c,low}, wh^l_{c,up}]\}$
  $\mathcal{D}^* = \mathcal{D}^{'} \setminus \mathcal{O}$
  return $\mathcal{O}, \mathcal{D}^*$
end procedure

6.3.5 Data Modification

We propose different approaches to stabilizing the training through data modification using the detected set of outliers. These range from ERM on the non-outliers to VRM on the complete dataset. All approaches are visualized in **Step 2** of Figure 6.1.

More concretely, we propose to enforce different degrees of data augmentation, i.e. of linear interpolation between outliers and non-outliers, by using different values of α in Equation 6.1. Furthermore, all samples are only mixed with non-outliers (x_j in Equation 6.1). We refer to this as *Adapted MixUp*. If no outliers are detected, we optimize based on the empirical distribution, i.e. $\alpha = 0$. Thereby, memorization of outliers (incl. mislabeled samples) is prevented while, simultaneously, the model is still able to learn nonlinear relationships based on the dataset that is consitent with the hypothesis. The different approaches for the data

augmentation step are introduced below: The identified outliers are removed from the training set for the next model update, i.e. for the next ERM optimization step. Yet the dataset for the OD remains unaltered, i.e. the OD is always performed on the original dataset. All samples are augmented with *Adapted MixUp*. A higher value of α is used on the currently detected set of outliers and a lower one for the current set of non-outliers. Again, the dataset for the OD is left unaltered.

This approach is equivalent to **MixOutlier** except that samples that were detected as an outlier in any of the iterations are treated as outliers in each subsequent augmentation step.

This is based on **MixAllOutlier**. However, the labels in the dataset are permanently altered by building convex combinations with a factor of 0.6 from the label and the current prediction of the model - similar to Reed et al. (Reed et al., 2014).

6.4 Experimental Setup

The proposed two-step framework is evaluated based on two different datasets with different characteristics. We aim at evaluating the following properties: (1) Exp.1: different training dynamics on two datasets given symmetric label noise; (2) Exp.2: feasibility of *Adapted MixUp*; (3) Exp.3: performance of the proposed OD approach and (4) Exp.4: the generalization capabilities of the resulting deep models evaluated on a clean test dataset. All experiments are repeated 5 times and the mean and standard deviation are reported. Symmetric label noise is added to the inherently clean training datasets as described in (Chen et al., 2019; Jiang et al., 2017; Ma et al., 2018).

For evaluation purposes only, the test dataset is not corrupted by label noise. Our evaluation focuses on binary classification tasks as this is a relevant setup for fault detection. However, the proposed framework is also applicable to multi-class classification tasks. We compare the proposed framework, comprising the OD combined with the different variants of data augmentation, to the two baseline methods ERM and *mixup*.

6.4.1 Dataset

We evaluate the approaches on the MNIST dataset (LeCun et al., 1990) containing images of handwritten digits from 0 to 9. We reformulate it as a binary classification task for demonstration purposes - i.e. digits 0-3 are grouped into class 0 (30596 training samples, 5139 test samples) and digits 4-9 are grouped into class 1 (29404 training samples, 4862 test samples) - to simulate defect detection in PHM where fault types, as well as the healthy states, can have multiple patterns (which can be regarded as different operating conditions). The dataset contains 60000 training samples and 10000 test samples in total and has previously been used as a benchmark dataset for anomaly detection in PHM applications (Ducoffe et al., 2019). The images are normalized.

Furthermore, we apply the proposed framework to a simulated time series dataset - the Building Defect Detection dataset (BDD dataset) - to detect faults in buildings (Granderson et al., 2020). We conducted the experiments on the multi-zone variable air volume AHU dataset (MZVAV-2-2). The dataset contains measurements in a healthy state as well as measurements from three different fault patterns (leaking valve of heating coil, stuck valve of cooling coil, and stuck outdoor air damper). The sensor readings are recorded once per minute over 26 days. Half-hour time windows are selected for the classification as this is sufficient to distinguish healthy from faulty conditions. In total, this resulted in 1247 sample-label pairs, of which 623 are considered healthy and 624 are considered faulty. 20% of the data is randomly selected for the test dataset. In total, we consider 15 sensor readings, i.e. all besides one set point and one control signal (AHU: Supply Air Temperature Set Point, AHU: Exhaust Air Damper Control Signal). Therefore, one sample consists of 450 measurements (30 minutes x 15 sensors). The measurements for each sensor are standardized.

Assumption	Noise Ratio	Upper Threshold	α
Little Noise	$0- < 10\%$	10%	0.4
Medium Noise	$10- < 30\%$	30%	8
Massive Noise	$30- < 50\%$	50%	32

Table 6.1: Setting of α under different noise assumptions based on proposed values in (Zhang et al., 2017b).

6.4.2 Hyperparameter Settings

As we assume that no clean, reliable validation dataset is available for HP tuning, we can only rely on metrics based on the training dataset and on the aforementioned assumptions. All parameters regarding the model and the standard optimization algorithm were chosen such that a training accuracy of at least 75% is achieved for all label noise ratios for each dataset under *ERM*.

The model used for MNIST dataset is a four-layer fully connected NN with 128, 32, 10 nodes activated by ReLU, and 2 nodes activated by the sigmoid function. It is updated with Adam (Kingma and Ba, 2014) (initial learning rate of 0.001). The batch size is set to 64. The model used for the two BDD datasets is a five-layer fully connected NN with 256, 128, 64, 16 nodes activated by ReLU, and 2 nodes activated by the sigmoid function. The optimizer Adam (Kingma and Ba, 2014) is used (initial learning rate of 0.0001) and the batch size is set to 16. Both models are trained by minimizing the cross-entropy loss as well as the entropy calculated based on the model's prediction as an additional regularization loss (Tanaka et al., 2018).

The minimal IQR for the OD is set heuristically to 0.0001 and is kept constant over all experiments - see Section 6.3.4. For *mixup*, the default values for α proposed in the original paper are used (Zhang et al., 2017b) based on a basic assumption about the label noise as defined in Table 6.1. Unless stated otherwise, for our proposed approaches, we set $\alpha = 0.4$ for the set of non-outliers $\mathcal{D}^*$ (as it corresponds to the assumption of *Little Noise*) and $\alpha = 32$ for the outlier dataset (corresponding to the *Massive Noise* setting). As mentioned above, if no outliers are detected at all during the training process, we revert to *ERM* by using a value of $\alpha = 0$.

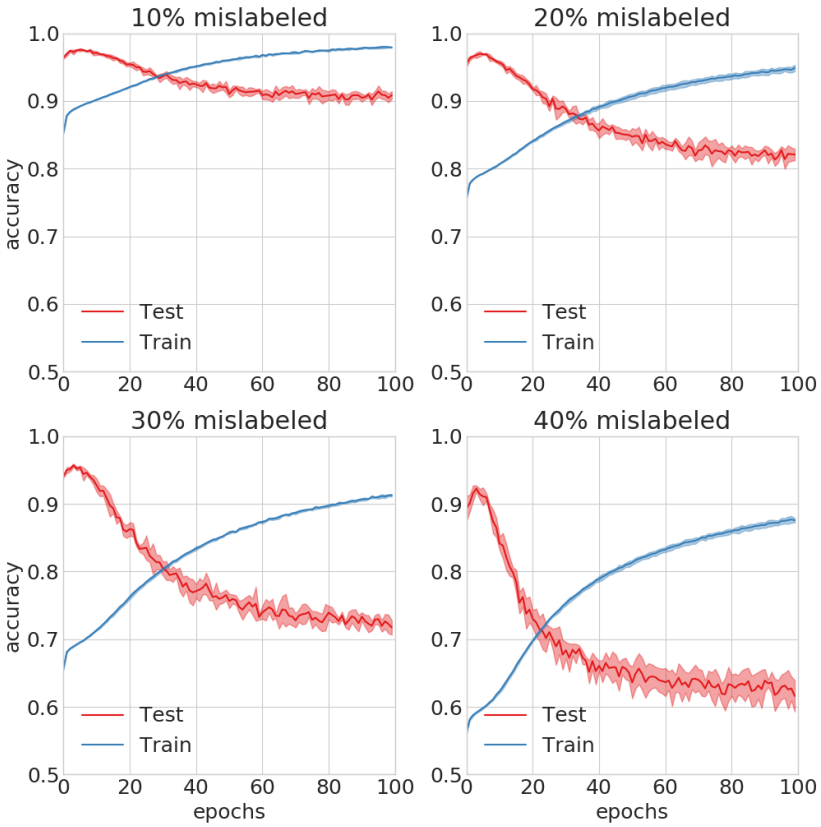
6.5 Results

6.5.1 Experiment 1 - Training Dynamics on Mislabeled Data

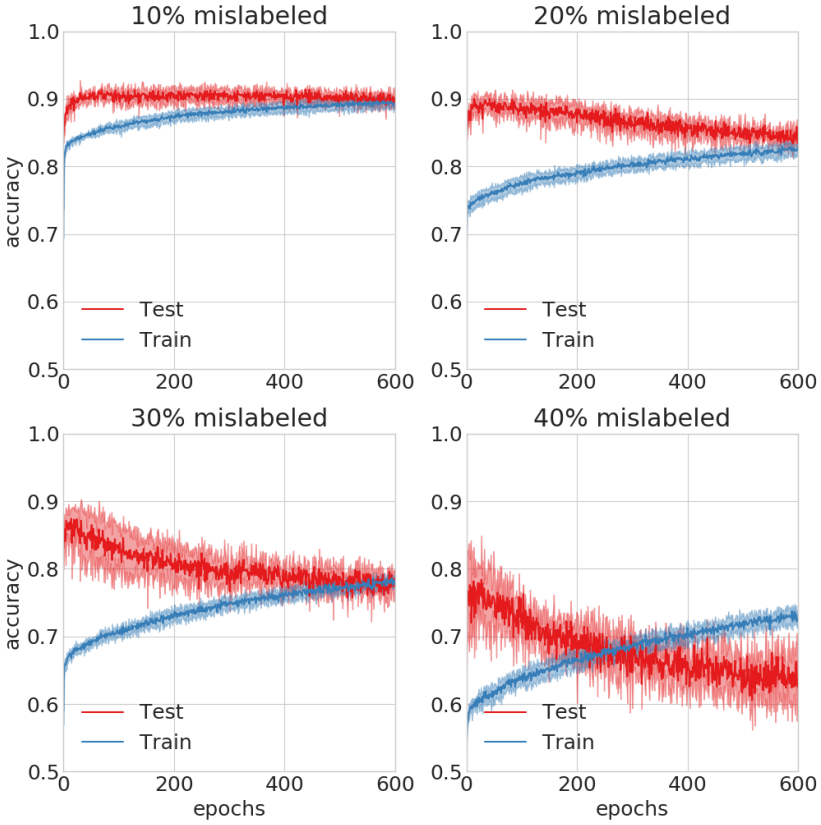
As preliminary results, we demonstrate the overfitting behaviour leading to poor generalization as described in the introduction. For this purpose, we train a model with ERM on a noisy dataset and plot the accuracy on the mislabeled training dataset and the test accuracy on a "clean" test dataset. In Figure 6.2a and Figure 6.2b, results with different noise ratios are plotted. This demonstrates how the models overfit to noisy labels.

6.5.2 Experiment 2 - Adapted MixUp

As a preliminary exploration, we evaluate the feasibility of enforcing different degrees of interpolation as described in Section 6.3.5 to demonstrate the rationale behind the proposed methodology. Therefore, we assume that we know the ground truth labels of the noisy training set and can thus identify the mislabeled samples. The results are compared to the original *mixup* augmentation (Zhang et al., 2017b). While α for the original *mixup* is set as stated in Section 6.4.2, for the *Adapted MixUp* a fixed value of $\alpha = 32$ is set for the mislabeled samples and a value of $\alpha = 0$ is set for the non-outliers as it is known to be noise-free given the ground truth information.



(a) MNIST dataset.

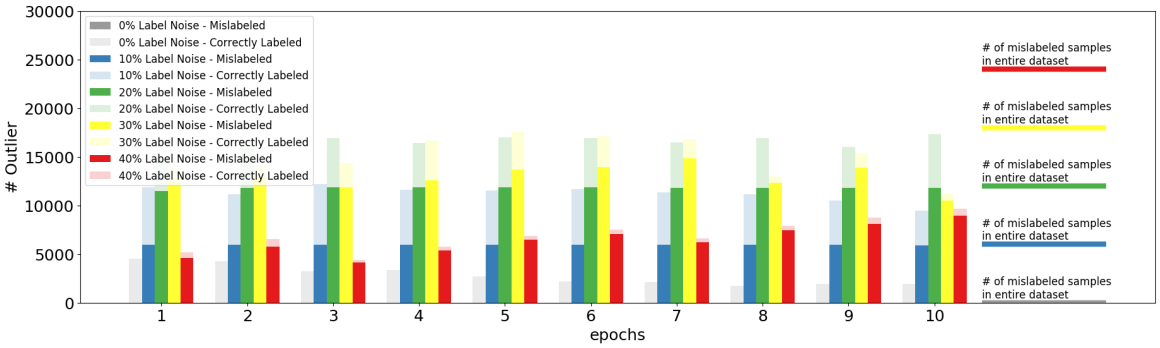


(b) Building Defect Detection dataset.

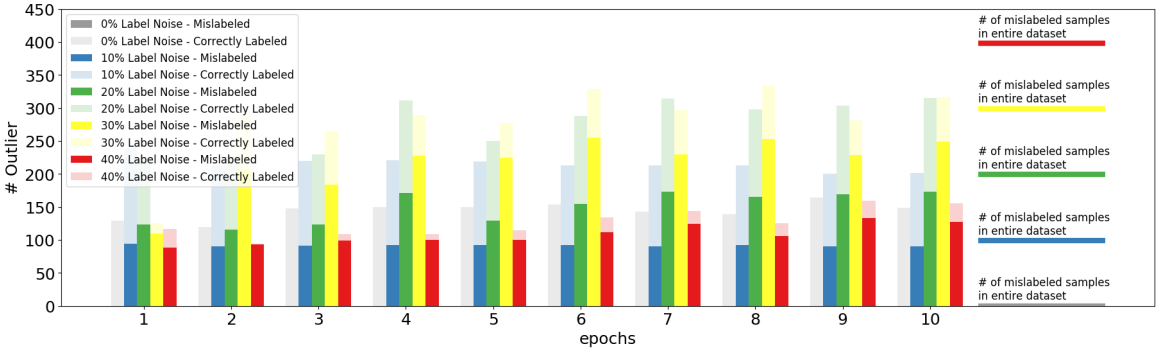
Figure 6.2: Training and Test Accuracy.

		Label Noise			
Dataset	**Method**	**10%**	**20%**	**30%**	**40%**
MNIST	*mixup*	$97 \pm 0\,\%$	$95 \pm 0\,\%$	$90 \pm 1\,\%$	$80 \pm 1\,\%$
	MixOutlier	$\mathbf{98 \pm 0\%}$	$\mathbf{98 \pm 0\%}$	$\mathbf{98 \pm 0\%}$	$\mathbf{98 \pm 0\%}$
BDD	*mixup*	$\mathbf{93 \pm 0\%}$	$90 \pm 1\,\%$	$87 \pm 1\,\%$	$78 \pm 1\,\%$
	MixOutlier	$\mathbf{93 \pm 1\%}$	$\mathbf{93 \pm 0\%}$	$\mathbf{93 \pm 2\%}$	$\mathbf{94 \pm 0\%}$

Table 6.2: Final Accuracy on Clean Test Dataset with Ground Truth Information.



(a) MNIST



(b) Building Defect Detection Dataset

Figure 6.3: No. of mislabeled and correctly labeled samples $\in \mathcal{O}$ for *MixOutlier*.

The results for $10\% - 40\%$ label noise are shown in Table 6.2. For 0% label noise, the approach is equivalent to *ERM* under the assumption that ground truth information is available. Therefore, this evaluation is not listed in the table. The *Adapted MixUp* augmentation, given ground truth information, leads to better generalization capabilities compared to default *mixup* - especially under massive noise. The final performance of *Adapted MixUp* is independent of the label noise ratio for both datasets.

6.5.3 Experiment 3 - Outlier Detection

We evaluate the proposed algorithm for OD on the first 10 epochs on one iteration of the *MixOutlier* approach, which is representative for all other approaches - see Figure 6.3a and Figure 6.3b, where mislabeled samples in the outlier dataset are distinguished from those that are correctly labeled.

Given no label noise, only correctly labeled samples are identified as outliers. While the number of outliers quickly decreases on the MNIST dataset from 4521 (7.5% of the entire dataset) to 281 (0.5% of the entire dataset) within the first 10 epochs, it stays rather constant over the first 273 epochs with 141 ± 18 outlier samples on the BDD dataset. In the subsequent

epochs, considerably fewer outliers are detected (64 ± 11).

Given medium noise levels, the detection of mislabeled samples is very efficient. 99% of the mislabeled samples are identified within the first few epochs on the MNIST dataset for both noise settings. On the BDD dataset, 95% of mislabeled samples are initially identified on the dataset with 10% label noise and 87% on the dataset with 20% label noise. Yet the detection approach is lacking precision at the medium noise level as the set of outliers also includes correctly labeled samples. For example, on the BDD dataset 144 (on the dataset with 10% label noise) and 141 (on the dataset with 20% label noise) correctly labeled samples are initially in the set of outliers. This corresponds to 16% and 18% of all correctly labeled samples in the respective datasets. Yet, as the training continues, the number of mislabeled samples in the set of outliers stays approximately constant over all epochs ($88\% \pm 4\%$ and $81\% \pm 6\%$ of the truly mislabeled samples), whereas in the last training epoch, the number of correctly labeled samples decreases to 8% and 9% of the correctly labeled samples in the respective training datasets.

Given a massive noise level (40% label noise ratio), the algorithm is more precise but less effective on both datasets. Initially on MNIST, 13% of all mislabeled samples are detected and 2% of all correctly labeled ones. As the training continues, up to 67% of mislabeled samples are detected and 20% of the correctly labeled ones.

6.5.4 Experiment 4 - Binary Classification

All introduced approaches combining the proposed OD with *Adapted MixUp*, as introduced in Section 6.3.5, are applied to the MNIST dataset and to the BDD dataset mislabeled by different ratios as described in Section 6.4. The approaches are compared to ERM and *mixup*. The results are shown in Table 6.3.

While all proposed approaches show similar performance to the baseline method *mixup* on MNIST with 0 and 10% label noise, they outperform *mixup* on higher noise ratios. The performance gain compared to the baseline methods is most visible when using the *MixAll-Outlier+Relabel* approach on the MNIST dataset with a label noise ratio of 40% (accuracy gain of 31% accuracy compared to *ERM* and 14% compared to *mixup*).

On the BDD dataset, the best-performing approach depends on the label noise ratio. On the datasets with little to medium noise levels, *mixup* performs about as well as *MixOutlier*, whereas all of the other proposed approaches perform worse. At massive noise levels, most of the proposed approaches outperform both baseline methods (*ERM* and *mixup*). Again, the model trained with *MixAllOutlier+Relabel* on 40% label noise achieves the biggest gain in accuracy compared to the models trained on the baseline methods.

6.6 Discussion

Training Dynamics on Two Datasets: The models trained on both datasets provide a suitable testbed for evaluating the universality and sensitivity of the proposed approaches as they show different dynamics under label noise.

Outlier Detection: Given the ground truth information about all labels, *Adapted MixUp* augmentation leads to a good generalization capability for both datasets and all label noise settings. However, the detection of mislabeled samples is a challenging task. While the OD shows similar behaviour on both datasets, its precision and effectiveness depends on the label noise ratio: it is more effective and less precise at medium noise levels and vice versa - more precise but less effective at massive noise levels. The lower precision at medium noise levels does not negatively impact the performance of the models trained on the larger MNIST dataset. It does, however, decrease the model's final performance on the smaller BDD dataset compared to the baseline method *mixup* (e.g. with the *DeleteOutlier*, *MixAllOutlier*, or *MixAllOutlier+Relabel* approach). The low effectiveness at massive noise levels especially affects the performance of models that are able to "memorize" the mislabeled samples faster

Method - Label Noise	0%	10%	20%	30%	40%
	MNIST				
ERM	**99 ± 0%**	91 ± 1 %	82 ± 1 %	72 ± 1 %	62 ± 2 %
mixup	98 ± 0 %	97 ± 0 %	94 ± 0 %	91 ± 1 %	79 ± 1 %
DeleteOutlier	98 ± 0 %	**98 ± 0%**	**97 ± 0%**	94 ± 0 %	80 ± 1 %
MixOutlier	98 ± 0 %	**98 ± 0%**	**97 ± 0%**	93 ± 0 %	80 ± 2 %
MixAllOutlier	98 ± 0 %	97 ± 0 %	**97 ± 0%**	95 ± 0 %	88 ± 1 %
MixAllOutlier+Relabel	98 ± 0 %	97 ± 0 %	**97 ± 0%**	**97 ± 0%**	**93 ± 2%**
	BDD				
ERM	94 ± 1 %	90 ± 2 %	83 ± 1 %	79 ± 1 %	63 ± 4 %
mixup	94 ± 0 %	**92 ± 2%**	**89 ± 1%**	83 ± 2 %	73 ± 5 %
DeleteOutlier	92 ± 1 %	87 ± 2 %	84 ± 1 %	**85 ± 3%**	77 ± 3 %
MixOutlier	**95 ± 1%**	91 ± 1 %	**89 ± 2%**	**85 ± 3%**	72 ± 2 %
MixAllOutlier	94 ± 1 %	85 ± 3 %	87 ± 3 %	84 ± 3 %	79 ± 2 %
MixAllOutlier+Relabel	94 ± 1 %	86 ± 2 %	86 ± 2 %	84 ± 2 %	**83 ± 3%**

Table 6.3: Final Accuracy on Clean Test Dataset Trained on Mislabeled Datasets for Various Label Noise.

than they are detected. This becomes particularly evident looking at the final performance of the models trained with *DeleteOutlier* or *MixOutlier* on MNIST with 40% label noise.

Generalization Capabilities of the Trained Models: Contrary to the previous study of Zhang et al. (Zhang et al., 2017b), *mixup* does not outperform *ERM* in our experiments if no label noise (0%) is present. This is most likely due to the fact that α has not been tuned but rather set based on assumptions about the upper limit of the label noise as defined in Section 6.4.2. *MixOutlier* slightly outperforms *ERM* on the clean BDD training dataset. This might hint towards findings in the literature on curriculum learning where a curriculum that sorts the training dataset can guide optimization towards a preferable optimum (Hacohen and Weinshall, 2019). However, the performance gain in our experiments is not significant and the results on MNIST do not support the hypothesis that the proposed approach acts as a curriculum that is beneficial for optimization. Therefore, this is left for future research.

If the dataset is truly mislabeled (label noise ratio ¿ 0%), all of our proposed approaches along with *mixup* outperform the baseline method *ERM*. *mixup* results in a satisfactory performance at medium noise levels on both datasets. However, *MixOutlier* yields a superior performance on MNIST and a comparable performance on the BDD dataset. Most of the other proposed approaches in Section 6.3.5 suffer from the insufficient precision of the OD on the smaller BDD dataset, as described above at medium noise levels. In the case of massive noise, the performance of the baseline method *mixup* drops compared to the other noise levels. While some of the proposed approaches suffer from the initially low effectiveness of the OD (as decribed above), they still perform as well as or better than baseline method *mixup*. Moreover, *MixAllOutlier* and *MixAllOutlier+Relabel* outperform *mixup* significantly on both datasets at high noise levels (accuracy gain of 6-14%).

Based on the above evaluations, we recommend choosing the optimization methodology based on the assumption regarding the label noise. We recommend using *MixOutlier* or the baseline *mixup* for scenarios in which little to medium noise is suspected and the *MixAllOutlier* approach for massive noise.

Sensitivity Analysis of the Recommendations: To evaluate the sensitivity of these recommendations, the behaviour of the proposed approaches is evaluated for cases in which the assumption regarding the upper limit of label noise is incorrect. We assess only the best-performing approaches per noise level, i.e. **MixAllOutlier** and **MixOutlier+Relabel**

Dataset	Assumed Noise Limit	Method	Actual Label Noise 0%	10%	20%
MNIST	*Massive* 50%	*mixup*	**98 ± 0%**	**97 ± 0%**	95 ± 0 %
		MixAllOutlier	**98 ± 0%**	**97 ± 0%**	96 ± 0 %
		MixAllOutlier +Relabel	97 ± 0 %	**97 ± 0%**	**97 ± 0%**
BDD	*Massive* 50%	*mixup*	**94 ± 0%**	**93 ± 0%**	**89 ± 2%**
		MixAllOutlier	92 ± 1 %	89 ± 2 %	87 ± 2 %
		MixAllOutlier +Relabel	86 ± 0 %	85 ± 1 %	83 ± 2 %

Table 6.4: Final Accuracy with Overestimated Noise Level.

Dataset	Assumed Noise Limit	Method	Actual Label Noise 20%	30%	40%
MNIST	*Little* 10%	*mixup*	**89 ± 1%**	**81 ± 1%**	**69 ± 1%**
		MixOutlier	82 ± 1 %	76 ± 3 %	**69 ± 1%**
	Medium 30%	*mixup*	94 ± 0 %	89 ± 1 %	78 ± 1 %
		MixOutlier	**97 ± 0%**	**93 ± 2%**	**79 ± 2%**
BDD	*Little* 10%	*mixup*	**90 ± 2%**	**82 ± 2%**	**69 ± 3%**
		MixOutlier	86 ± 2 %	80 ± 2 %	**69 ± 2%**
	Medium 30%	*mixup*	**89 ± 1%**	**86 ± 1%**	**76 ± 1%**
		MixOutlier	**89 ± 2%**	84 ± 4 %	74 ± 3 %

Table 6.5: Final Accuracy with Underestimated Noise Level.

(Table 6.4) for a massive assumed noise level, and **MixOutlier** (Table 6.5) for a little to medium assumed noise level. The performance is compared to the baseline method *mixup* given the same assumptions, i.e. the HP setting of α as described in Table 6.1.

The sensitivity analysis reveals that the baseline *mixup* approach is less sensitive to erroneous assumptions of the label noise. This is particularly true if the noise level is underestimated, as the number of detected outliers surpasses the threshold and, therefore, hardly any outliers are considered. However, if the noise level is overestimated, the best-performing approaches depend on the dataset. Yet the baseline *mixup* shows, on average, the best performance over all settings and on both datasets. Hence, one drawback of the proposed framework is its sensitivity to inaccurate assumptions - especially if the noise is underestimated. However, these false assumptions can be easily identified: For example, if the number of detected outliers constantly surpasses the estimated upper noise level, it is a clear indication of faulty assumptions.

6.7 Conclusion

In this study, we proposed multiple variations of a two-step framework to train fault detection models that are robust to label noise. The framework first identifies outliers based on the hypothesis update and then modifies the training dataset accordingly. The framework's hyperparameters are set on the basis of a rough assumption regarding the label noise. Ultimately, we demonstrate that the different strategies handle the level of noise and the uncertainty of this level very differently. Our proposed approaches outperform other approaches from the literature when the level of noise is expected to be high. In practical applications, it is considered realistic to obtain a good estimate regarding an upper noise limit. This is

particularly the case for technical systems, where a rough assumption can be made as to how noisy the labels are expected to be (representing how unsure the domain experts are about the labels). In addition, our extensive evaluation can be used by practitioners to choose the appropriate approach based on a rough estimation of the upper limit of the label noise level. Lastly, our framework enables the early detection of outliers in the optimization process. This can provide additional information about the dataset that can be used for further evaluation. For example, analyzing the set of outliers already in the first training iteration provides information on the label noise in the dataset. Still, in real applications, the success of the optimization cannot be evaluated if no clean test dataset is available to measure the performance of the resulting model. However, the analysis of the detected outliers could be used to validate the model instead. Evaluating the proposed methodology on more datasets is left for future research, particularly in terms of assessing the robustness of the approaches with respect to intra-class variability. Furthermore, evaluating the proposed OD in terms of its ability to act as a curriculum for efficient learning will be subject to further research.

7 Contrastive Feature Learning for Fault Detection and Diagnostics in Railway Applications

This chapter corresponds to the published article:[1]

Rombach, Katharina and Michau, Gabriel and Ratnasabapathy, Kajan and Ancu, Lucian-Stefan and Bürzle, Wilfried and Koller, Stefan and Fink, Olga. "Contrastive Feature Learning for Fault Detection and Diagnostics in Railway Applications." *under review in scientific journal.* 2022.

Abstract: Recently, the application of data-driven algorithms for fault detection and diagnostics in railway applications has been increasing, for both infrastructure and rolling stock applications. In practice, the performance of data-driven models can be compromised if the training dataset is not representative of all possible future conditions. We propose to approach this challenge by learning a feature representation that is, on the one hand, invariant to operating or environmental factors but, on the other hand, sensitive to changes in the asset's health condition. We evaluate the proposed contrastive learning framework under different degrees of label availability and representativeness of training data on two different case studies from railway infrastructure and rolling stock. First, we evaluate the performance of supervised contrastive feature learning on a railway sleeper defect classification task given a labeled image dataset that is collected by a diagnostic vehicle. Second, we evaluate the performance of unsupervised contrastive feature learning without access to faulty samples on an anomaly detection task on a railway wheel dataset that is collected by wayside monitoring systems (equipped with strain gauge sensors). Here, we test the hypothesis of whether a feature encoder's sensitivity to degradation is also sensitive to novel fault patterns in the data. Our results demonstrate that contrastive feature learning improves the performance on the supervised classification task regarding sleepers compared to a state-of-the-art method by 13.3%. Moreover, on the anomaly detection task on railway wheels, the detection of shelling defects is improved compared to state-of-the-art methods.

7.1 Introduction

Train travel is expected to be safe, affordable, and punctual. Hence, railway operators need to ensure safe and reliable operations while being economically efficient. There is an increasing investment in condition monitoring (CM) devices in the railway system. For example, in Switzerland, sleepers are monitored by cameras installed on diagnostic vehicles and railway wheels are monitored by wayside monitoring (WSM) systems that are installed on the railway tracks (Krummenacher et al., 2017). The collected CM data enables the implementation of data-driven solutions for fault detection and diagnostics (Asplund, 2016; Xie et al., 2020; Ghofrani et al., 2018). Especially solutions based on machine learning algorithms have been increasingly applied for fault detection and diagnostics of different assets within the railway

[1] Please note, this is the author's version of the manuscript that is currently under review in a scientific journal. Changes resulting from the future publishing process, namely editing, corrections, final formatting for printed or online publication, and other modifications resulting from quality control procedures may have been subsequently added.

system (Tang et al., 2022). While the captured CM data differ substantially between different assets, fault detection and diagnostics solutions based on real CM data from in-service assets are often facing similar challenges. These challenges include variability of operating conditions, missing or scarce fault data at training time, missing ground truth information, as well as partial observability.

In-service assets are constantly subject to changing operating and environmental conditions, causing domain shifts in the CM data. These domain shifts can significantly decrease the performance of data-driven models. In literature, domain adaptation methods were proposed to mitigate the negative effect of domain shifts (Li et al., 2021b). However, these solutions require the identification of distinct domains. In systems that operate in the wild such as railway systems, it is often impossible to identify all factors that cause variations in the CM data and therefore, it is impossible to identify each distinct and discrete domain shift, especially since these factors of variations change continuously. In the following, we will refer to data variations that are caused by changes of the operational and environmental conditions as non-informative variations as they are not informative on the asset's health condition.

While domain shifts negatively impact the performance of data-driven models per se, the challenges are especially pronounced if the available fault data is either scarce or not available. This is typically the case in CM datasets since (safety) critical assets are reliable by design and therefore, faults occur very rarely which results in either scarcity of fault data or, more commonly, no available fault data in the training dataset. In the first case, when fault data is available but scarce, the available fault data might not be representative of the entire fault class. Equally so, the data of the healthy class might not be fully representative of all possible future operational or environmental conditions (non-informative variations). This low representativeness of the healthy and faulty classes impacts the performance of diagnostics models negatively, especially if fault patterns in the data resemble normal healthy variations in the data closely and it is inherently difficult to distinguish between different health conditions. This is often the case in fault diagnostics tasks within a railway system where early indications of faults might not impact the CM data significantly but need to be classified correctly to enable safe and efficient operations. The second case, when no fault data is available for training a model for CM, is even more common. Since practical CM solutions need to be taken into operation within a short period of time, it is often not feasible to wait until any fault has occurred. In this case, fault detection models are typically trained, to detect any novel variation in the data. As elaborated above, the CM data is not only impacted by a change in the asset's health condition, but also by domain shifts. If the CM data is impacted by novel data variations caused by domain shifts, reliable fault detection can be very challenging. Sensitive fault detection models might raise many false alarms if domain shifts cause novel variations in the data (Michau et al., 2020). False alarms significantly reducing the usability of the implemented model to real operations. On the contrary, if an anomaly detection (AD) model is trained to be robust to domain shifts, e.g. by implementing a domain generalization method (Li et al., 2020c), it is potentially also robust to detecting novel data variations that are caused by faults as well. The tradeoff between reducing the amount of false alarms and increasing the sensitivity to defect is especially challenging since it is important to detect faults in an early stage to operate an asset reliably and efficiently where the data might be impacted only minimally. Therefore, a reliable solution to fault detection needs to be able to be sensitive to early signs of unseen faults while being insensitive to novel non-informative variations in the data due to domain shifts.

Furthermore, expert annotation is not always reliable. This raises the challenge of noisy labels in CM datasets and increases the challenge of implementing a reliable fault detection and diagnostics solution.

Moreover, some CM systems are only capable of monitoring parts of the assets. For exam-

ple, the WSM systems in Switzerland with strain gauge sensors monitor only limited sections of the entire wheel circumference. The available CM data, therefore, only holds partial information on the condition of the assets, which makes reliable data-driven fault detection and diagnostics difficult.

The challenges elaborated above limit the application of data-driven solutions in railway applications. This can be exemplified on two case studies of railway applications considered in this paper: (1) Monitoring of railway sleepers with a camera system. (2) Monitoring of railway wheels with a WSM system.

For some assets within the railway system a meticulous fault annotation process has been established. For these assets, a labeled training dataset is available. For example, railway sleepers are monitored with a camera system by the Swiss Federal Railways and the recorded images are labeled by domain experts. Also, since the entire railway network is monitored, a sufficient number of faulty samples can be collected. However, certain faults in this dataset resemble normal variations in healthy conditions. Stones on the sleeper, for example, can easily be mistaken for spalling defects or a shadow can easily be mistaken for a crack defect. As elaborated above, if all possible variations are not sufficiently represented in the training dataset, they can harm the performance of data-driven fault detection and diagnostics models (Rombach et al., 2021). This is especially the case if the distinction between normal variations and faults in the data is difficult. Therefore, sleeper defect diagnostics based on an image dataset collected under varying environmental conditions is a challenging task. Previous works on sleeper fault diagnostics mainly worked with more expensive sensor installations and conducted only in-workshop experiments (Jing et al., 2021). As elaborated above, in-workshop experiments are not representative of real conditions.

WSM systems with strain gauge sensors have been proposed for monitoring the condition of railway wheels in literature before. Alemi et al. (2017), for example, investigated how many strain gauge sensors are required to monitor the wheel circumference to enable reliable fault detection of wheel flats in workshop experiments. However, in real operations, the number of sensor needs to be kept low as each additional sensor increases the maintenance cost. In Switzerland, for example, only eight widely spread sensors are installed providing only partial observable CM data. Furthermore, experiments conducted in workshop experiments do not face the challenges real CM data from in-service assets such as data variations caused by external factors in real operations. Additionally, in reality not only wheel flats occur but also other fault types with less distinct fault patterns can occur and need to be detected. In summary, the work does neither address the challenge of real in-service conditions, nor the partial observability of real WSM systems, nor the fact that multiple wheel defects can possibly emerge.

Ni and Zhang (2021) considered real CM data from one WSM system of in-service wheels. However, experiments were conducted with a fixed train speed. The acquired dataset is, therefore, not representative of real operating conditions where data is captured under different speed conditions and multiple WSM systems might be used. Further, 21 densely deployed sensors are used in the conducted experiments to have a full coverage of the wheel circumference. In our considered case study from an operational WSM system in Switzerland installed in the entire railway network, the amount of available sensors is lower. While the measurement setup in (Krummenacher et al., 2017) is similar to our case study: Krummenacher et al. (2017) worked on real CM data from WSM systems with only eight widely spaced sensors. However their proposed approach requires a fully labeled dataset, which is often not available in real applications. Moreover, they did not consider the challenge of a limited label reliability. Also the challenge of the partial observability of the available CM data has not been addressed.

In this work, we propose to improve the robustness of fault detection and diagnostics by inducing a robust encoding in the feature space based on contrastive learning. We evaluate

how this can be achieved in supervised and unsupervised tasks for real railway applications. The main novelty of the proposed framework is in the unsupervised setup without any observed faults. We evaluate the performance of the proposed method on two CM datasets of railway applications. First, we conduct experiments in a supervised setting on a sleeper defect classification task on a labeled image dataset of different sleeper conditions. Second, we conduct experiments on a CM dataset of railway wheels (recorded on WSM systems) where presumably no fault data is available for training. In this task, we use the feature representation learned with unsupervised contrastive learning to detect anomalies (shelling and cracks) and monitor the evolution of the fault condition over time.

Our results demonstrate that contrastive feature learning achieves a higher performance on the supervised task of sleeper defect classification and improves the performance on the AD task regarding the railway wheels as compared to state-of-the-art methods.

The remainder of the paper is organized as follows: in Section 7.2, the related work is reviewed, followed by the introduction to the case studies in Section 7.3. The proposed methodology is introduced in Section 7.4, the performed experiments are detailed in Section 7.5 and the results are reported in Section 7.6. The findings are discussed in Section 7.7 and conclusions are drawn in Section 7.8.

7.2 Related Work

Detecting and preventing failures is essential to improving the railway's system efficiency and safety (Davari et al., 2021; Yamashita et al., 2022). Data-driven approaches have been proposed to diagnose the condition of infrastructure and vehicle components (Hu et al., 2017; Tatarinov et al., 2019). However, to date, human inspection remains the most common assessment method for **sleepers** (Jing et al., 2021), whereby human inspectors walk along the tracks to find defects. This process is time- and resource-consuming and its quality depends on the experience and attention to detail of the individual inspector. To automate this process, approaches based on different CM devices, such as acoustic emission sensors (Janeliukstis et al., 2019) or laser speckle imaging sensors (Pang et al., 2020), have been proposed. (Rui et al., 2020; Tatarinov et al., 2019; Mori et al., 2010). However, most of the proposed approaches are limited to laboratory applications that may be challenging to adapt for real applications. For example, analysis based on acoustic emission of in-workshop experiments (Janeliukstis et al., 2019) might not be transferable to real operating settings, where additional sources of acoustic emission exist. A detailed overview of current approaches to sleeper monitoring and their limitations is given by Jing et al. (2021). Moreover, none of the existing publications focused on making the approaches robust to changes in real environmental conditions. In this study, we utilize data from cameras mounted on the diagnostic vehicles for crack and spalling classification. Therefore, the utilized data is easy to acquire and recorded under various real environmental conditions.

Also for **railway wheels**, different approaches have been proposed to monitor their condition (Alemi et al., 2017). Several studies have focused on using on-board health monitoring systems, which enable continuous monitoring (Baasch et al., 2021; Li et al., 2017; Bosso et al., 2018; Wang et al., 2020b). However, with the ongoing increase in rolling stock in European rail networks (Mosleh et al., 2021), on-board systems do not scale well due to cost- and time-intensive installations and maintenance. Moreover, they are not particularly well suited for freight transportation. An alternative to on-board systems are WSM systems (Asplund, 2016). These are permanently installed in the railway tracks and therefore scale better as they can be frequented by an increasing quantity of rolling stock. Previous studies developed approaches on simulated (Mosleh et al., 2021; Song et al., 2013) or test rig data (Alemi et al., 2017). These may not be transferable to real conditions. For example, the approach proposed by (Alemi et al., 2017) requires a reference point on the wheels that is not known in real operations. Furthermore, previous studies have mainly focused on the fault

type 'flat', which is a rather easily detected fault type due to its high impact force on the wheel–rail surface (Alemi et al., 2017; Bian et al., 2013). A characteristic sensor response to a flat defect is shown in the paper of Krummenacher et al. (2017). Only a few studies have considered other common types of faults such as shelling or cracks. Krummenacher et al. (2017), for example, trained binary classification models to detect out-of-roundness and shelling defects in addition to flats. The proposed approach, however, needs to be trained on a fully labeled dataset. Faults rarely occur on safety-critical assets in general and accessing the true condition of railway wheels is difficult. Therefore, acquiring a sufficient amount of fault data is time-consuming and the labeling requires additional effort. Solutions that can be applied on unlabeled data are therefore preferable.

Feature learning, both supervised and unsupervised, has attracted a lot of attention in recent years (Fink et al., 2020; Zhong et al., 2016). The learned feature representation has been typically combined with classification models in supervised setups (Rombach et al., 2021), clustering methods in unsupervised settings (Chao et al., 2021), or One-Class Classification (OCC) models in unsupervised settings without fault data (Michau et al., 2020). Different types of autoencoders (AE) have been proposed for feature learning in unsupervised settings (Chao et al., 2021; Michau et al., 2020). In the context of fault detection given only healthy data, the hierarchical extreme learning machine (HELM) was proposed, whereby each layer of an AE was trained separately by solving a single-variable convex optimisation problem (Michau et al., 2020). The objective of training any AE is to compress all information contained in the data, not distinguishing based on the type of information - whether it is descriptive of a health condition or not. It was demonstrated that AEs are not robust when applied to varying OCs (Rombach et al., 2021). Therefore, these approaches might not be suitable if the data sample is affected by large variations that are non-informative of health conditions.

Contrastive feature learning has been shown to achieve robustness against changing operating conditions (Rombach et al., 2021) and is, therefore, a promising feature learning paradigm. The learning objective of the encoder model is to group data with similar semantic meaning close to each other in the feature space and spread dissimilar data far apart (Chopra et al., 2005). In supervised settings, the semantic similarity has typically been determined by the sample's label (Rombach et al., 2021). However, neither full supervision nor full representation of possible data classes (such as different fault types) is available in real CM datasets (Fink et al., 2020). Hence, unsupervised implementations of contrastive feature learning have been proposed (Chen et al., 2020). Franceschi et al. (2019), for example, proposed employing time-based negative sampling for contrastive feature learning for time-series data in an unsupervised setup. It is assumed that a randomly picked sample that is far in time is highly dissimilar to the currently observed sample. This assumption would not hold for time series with seasonalities (if the sample would coincide with the same periodicity). Equally so, if the semantic meaning of the timeseries data is not changing over time, time-based negative sampling would not result in a semantically reasonable feature space as the objective functions would push semantically similar patterns apart in the feature space. CM data is recorded either continuously or in discrete time intervals and the condition of an asset typically changes over time due to normal degradation processes. Therefore, measuring the similarity of different data samples in time, presents a very promising direction to form data pairs for semantically similar and dissimilar conditions for contrastive learning. In order to achieve robustness to the various operating and external factors of variations of real CM datasets, it needs to be ensured that semantically similar pairs are recorded under different and diverse conditions. A selection criteria as proposed by Franceschi et al. (2019), where the positive pair consists of subsamples of the same measurement, is not sufficient as one measurement of CM data is typically only recorded under the same operating and environmental conditions. Furthermore, workshop visits need to be taken into account to

reset the time criteria that is used for selecting negative pairs to prevent the above mentioned problem of seasonality when employing time-based negative sampling. Previous approaches to unsupervised feature learning of time-series data have also not yet been evaluated in settings where the data is not only unlabeled but also in which, presumably, only data from one class (healthy data) is available during the learning process. In absence of different health conditions in the training dataset, it is impossible to directly impose sensitivity to different health conditions. In other words, in absence of fault data, the encoder model can not be trained to be sensitive to patterns in the data that correspond to faulty conditions. Therefore, in this study, we aim to evaluate if a model trained to be sensitive to degradation of railway wheels is also sensitive to fault patterns of different types. Contrastive learning has not yet been applied to the real data collected under real operating conditions in railway applications.

In this research, we propose to utilize contrastive feature learning for learning a compact representation of the CM data that is, on the one hand, invariant to data variations caused by non-informative factors and, on the other hand, sensitive to changing health conditions. We showcase how this can be achieved for tasks with different degrees of data and label availability within the railway system. A special focus is put on how contrastive learning can be used in an unsupervised setting where, presumably, only normal degradation data is available for training but no fault data. We test the hypothesis that selecting negative samples based on time may be better suited for fault types that are more similar to degradation. Based on the learned feature representation, in a second step, the condition of the assets is monitored by implementing a fault detection or diagnostic solution.

7.3 Case Studies

Two real CM datasets from the Swiss national railways were used in this research: 1) an image dataset for infrastructure and 2) a time-series dataset for vehicles. Both datasets are affected by variability caused by factors that are non-informative of health conditions.

7.3.1 Sleeper Defect Classification (supervised)

The sleeper dataset comprises images of concrete sleepers and slab tracks partitioned into three classes: healthy conditions, cracks, and spalling. The images were collected by three line scanners mounted on diagnostic vehicles positioned to get a view of the left, right, and middle part of the sleepers (see Fig. 7.1). Each image was labeled by domain experts. The initially single-channel, large-scale images (3000x1024 pixels) are split into images of 1024x1024 pixels with an offset of 1024 pixels. Only if a bounding box of a defect is split into two separate images, the split is shifted such that the entire bounding box is included in the image. Furthermore, the single channel is replicated to three channels. The final training dataset comprises 1209 images with no defects, 1209 images with crack defects, and 964 images with spalling defects. 20% of the training dataset is chosen for the validation dataset. The test dataset contains 241 images of healthy sleepers, 229 images with cracks, and 199 images with spalling. Difficulties in the dataset arise from the similarity of defects to normal variability in the dataset. For example, a stone on the sleeper is difficult to distinguish from a true spalling defect (see Figure 7.2). **Pre-Processing:** All images are downscaled to images with a 512x512 pixel resolution using the area interpolation method from the OpenCV library and are normalized.

7.3.2 Railway Wheel Monitoring (unsupervised)

For the wheel defect detection dataset, the data was collected from WSM systems, referred to as Wheel Load Checkpoints (WLCs). They are permanently installed in the railway tracks and equipped with eight strain gauge sensors for each side of the train that measure the vertical force at a frequency of $10kHz$ when a train passes. The exact setup is described in

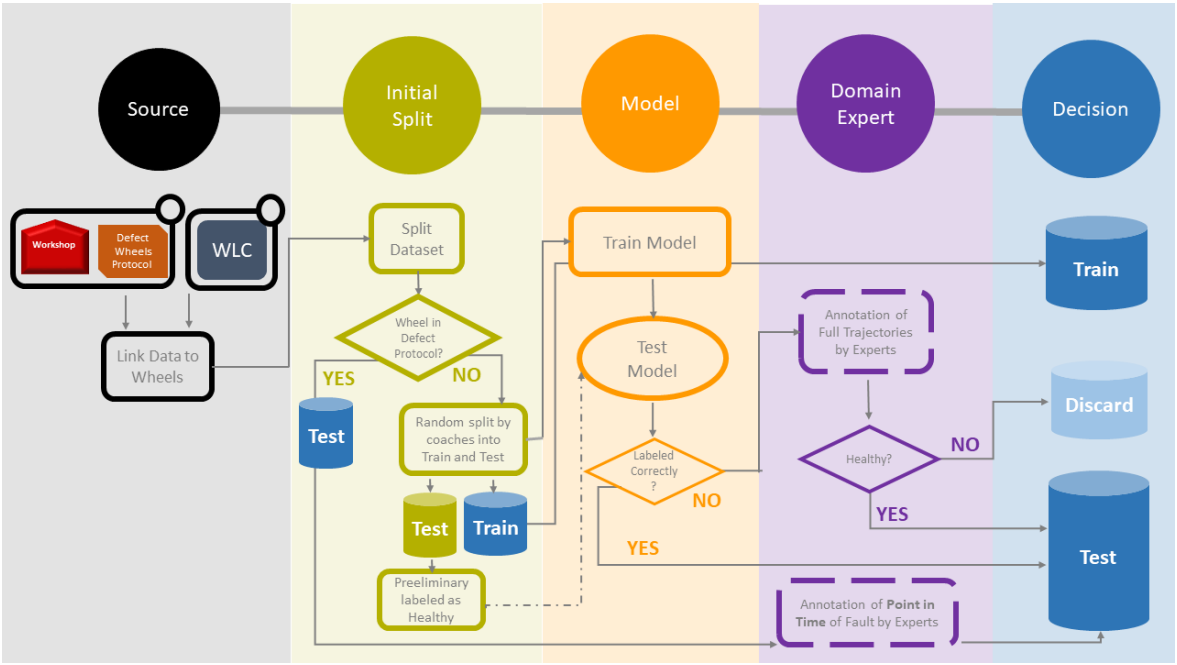
Figure 7.1: Diagnostic vehicle and illustration of monitoring system

Figure 7.2: Examples of sleeper defects

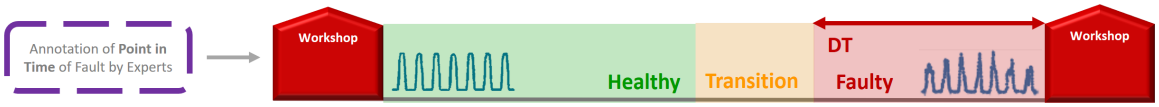
Krummenacher et al. (2017). WLCs are distributed over the entire infrastructure network in Switzerland and trains pass them approximately five times a day, although the exact value can differ significantly depending on the travel plan of individual compositions. Apart from changing health conditions of the wheels, variability in the data can be caused, for example, by different measurement locations and changing environmental factors. In this study, we monitor one fleet of passenger trains. These are less prone to wheel flats but rather to other defect types. We extract all signals for each wheel from each of the eight sensors and concatenate these signals to one measurement, such that each sensor measurement covers only parts of the wheel. The length of each signal depends on the speed of the train while passing.

For training, three trains comprising 20 coaches were considered. However, since coaches with reported defect wheels were excluded, data from 16 coaches were considered. Hence, no defect was documented for any of the measurements in the training dataset. However, the supervision process of the railway wheels is prone to error due to human inattention, different definitions or perceptions of defects and also errors in reporting. Due to the lenient supervision process, there is the possibility that some wheels are affected by faults. Still, building on the lenient supervision process, we consider the training dataset as presumably healthy. Due to the assumed absence of fault data and the lacking supervision in the training dataset, supervised learning as proposed in other studies – e.g. (Krummenacher et al., 2017) – cannot be applied.

For the test dataset, supervision of the wheel condition is provided during the workshop visit (see data sources in Figure 7.3a), where the wheels are inspected and re-profiled. A protocol is maintained to report wheels that have obvious tread defects. During the monitoring phase of one year, two defect types occurred: wheels with cracks (26 wheels) and with shelling (53 wheels). Examples of the defect types are shown in Figure 7.4. An initial test dataset split is performed as shown in Figure 7.3a (see 'Initial Split'), whereby all data from



(a) Process Flow



(b) Point in Time Annotation

Figure 7.3: Data acquisition for the railway wheel dataset: First, all data sources are linked to individual wheels ('Source'), resulting in a first data split into train and test ('Initial Split'). For the defective wheels, the time of defect initiation is provided by domain experts, as shown in Figure 7.3b. The preliminary healthy label of the wheels in the test dataset is challenged by fault detection models and evaluated by domain expert feedback.

defective wheels and randomly chosen healthy wheels comprise the test dataset.

In general, it is difficult to obtain a ground truth health condition of the wheels. On the one hand, wheels are not inspected between workshop visits. Therefore, we do not have any ground truth information on defect initiation. Instead, we rely on the labeling provided by domain experts who investigated the corresponding wheel data as shown in Figure 7.3a ('Domain Expert') and Figure 7.3b. The domain experts were asked to label the test dataset: The green zone corresponds to the timespan in which the fault is not yet manifested in the data. The orange zone corresponds to the transition phase in which the fault manifests itself in the data. The domain experts were asked to mark the first possible point in time they would be able to detect the fault. The time period after that is marked as red (red zone). Wheels that have not been reported as having defects might nevertheless be defective. Different degrees of attention or detail of reporting by the inspectors, as well as non-obvious defect types, can quickly lead to overlooked defects - also in the test dataset. Since the exact condition of the presumably healthy wheels in the test dataset is not known, the measurements from healthy wheels that are labeled as faults by the fault detection model (see 'Model' in Figure 7.3a) are discussed with domain experts (see 'Domain Expert' in Figure 7.3a). Only those data samples of the wheels that are labeled as healthy by the domain experts are added to the test dataset (see 'Decision' in Figure 7.3a).

Pre-Processing: First, we resample the concatenated signals via linear interpolation such that each signal has a length of 1024 to compensate for the different speeds of the train. Next, we normalize each recorded signal to compensate for the different loads of the train.

7.4 Methodology

The proposed methodology comprises two parts: 1) learning a contrastive feature representation (see Section 7.4.1) and 2) health monitoring based on the feature representation (fault detection or diagnostic - see Section 7.4.2). The entire methodology in its supervised and unsupervised application is illustrated in Figure 7.5.

7.4.1 Contrastive Feature Learning

The core of the proposed methodology is the contrastive loss function with which we train an encoder model to impose invariance to non-informative factors and sensitivity to changing health conditions. More concretely, the **semi-hard triplet loss** is used as defined in Equation 7.1, whereby x_a is the anchor sample, x_p the positive sample (that shares the semantic meaning resp. health condition with the anchor), and x_n is the negative sample with a different semantic meaning resp. health condition (Schroff et al., 2015), $f(.)$ is the encoded sample, $||.||$ is a distance metric, and ϵ is a margin parameter. By minimizing the distance between the anchor and positive sample pair, the invariance to non-infomrative variations is imposed and by increasing the distance between the anchor and negative pair, the sensitivity to faults is increased. To emphasize on the invariance within the same health condition, the sum over all positive samples within a batch is used. To enable smooth learning, a semi-hard negative sample is chosen as defined in (Schroff et al., 2015). This loss function is the core of the proposed approach. However, given different data and label availabilities of different tasks within complex systems such as the railway system, the implementation of the loss function needs to be adapted. More concretely, it needs to be adapted on how the data triplet are selected for calculating the loss function. Furthermore, the feature space needs to be regularized according to the respective data setting.

$$L(x_a, x_p, x_n) = \sum_{x_a \in X} max(0, \sum_{x_p \in P} (||f(x_a) - f(x_p)|| - ||f(x_a) - f(x_n)|| + \epsilon) \tag{7.1}$$

Supervised implementation of contrastive feature learning: If the available training dataset contains data from different health conditions, and this data is even labeled, the implementation of the triplet loss resp. the choice of the data triplets is straightforward: The positive samples are selected as those that share the same label as the anchor sample and the negative sample is selected as one sample with a different label as illustated in the upper half of Figure 7.5. Therefore the invariance to variations within one health condition as well as the sensitivity to other health conditions can be directly imposed.

Unsupervised implementation of contrastive feature learning without faults: If the available training dataset presumably does not contain data from faulty conditions, it is neither possible to directly impose sensitivity to faults, nor to select the data triplets based on the labels. In this work, we exploit the fact that (a) the condition of industrial

(a) Crack (b) Shelling

Figure 7.4: Examples of railway wheel defects

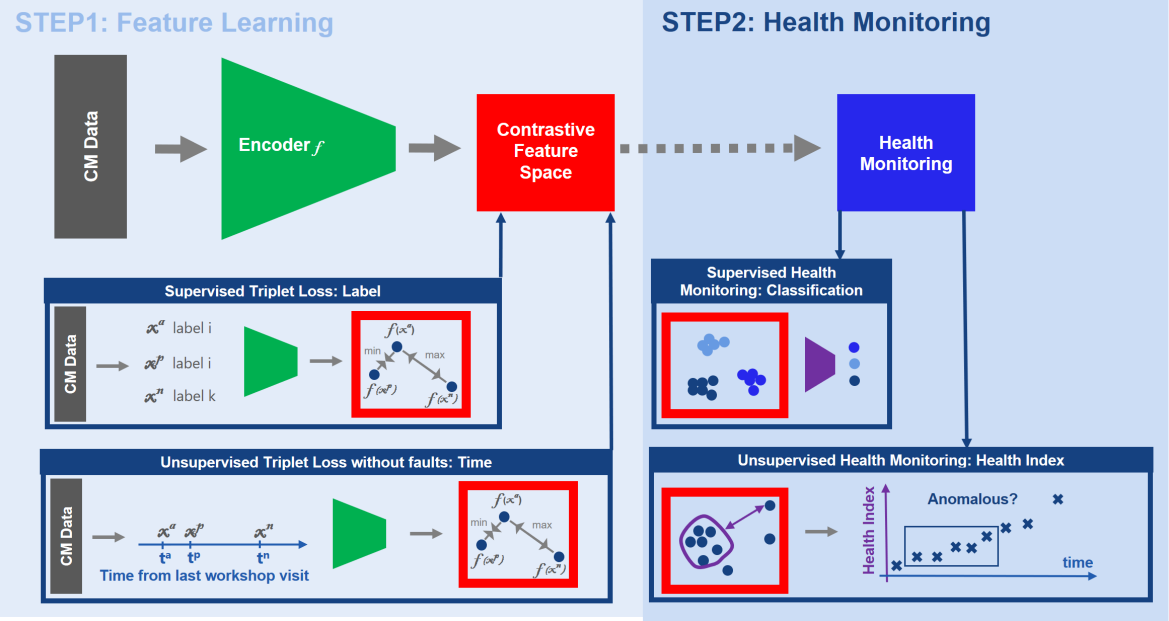


Figure 7.5: Illustration of the proposed framework: In a firts step, an encoder model f is trained on with the triplet loss function $L(x_a, x_p, x_n)$, whereby the selection of the data triplets x_a, x_p and x_n depends on the data availability. In the supervised setting (upper half), the triplets are chosen based on the label. In the unsupervised setting without fault data, the triplets are chosen with respect to the time that has passed since the last maintenance action. In a second step, the trained feature space is exploited to perform health monitoring. In the supervised setting a fault classification model is trained. In the unsupervised setting, a health index is extracted to detect anomalies.

assets is typically monitored in discrete time periods and (b) the health condition of each asset typically changes over time due to normal degradation. Therefore, even though we may not have faulty data available in the training dataset, it contains data from various non-informative variations as well as data from various states of degradation. The different states of degradation can be considered as different health states within the healthy class. This is exploited in our proposed implementation of the triplet loss where we aim to train a feature space that is sensitive to degradation and evaluate if this is transferable to sensitivity to faults. However, the dataset is not labeled with the different states of degradation either. Therefore, training a feature space sensitive to degradation is not straightforward. We propose to use the time passed after the last workshop visit as a proxy for the different states of degradation. This proxy only holds under the assumption that assets degrade similarly over time. Building on that assumption, we propose to choose the data triplets as illustrated in the lower half of Figure 7.5 to impose this sensitivity to degradation and invariance to other non-informative factors. We track the time that has passed since the last maintenance action of the asset. Starting from the maintenance actions, we consider the degradation process of the individual assets to increase in time. Therefore, measurements that are recorded close in time relative to the last workshop visit are considered to have a similar health condition while being recorded under different non-informative factors (positive pair). Whereas a data sample that is recorded at a more distant point in time is considered being dissimilar (negative sample). In other words, samples that are far in time have a different degradation state. The selection of the data triplets, therefore, is similar to the one proposed by Franceschi et al. (2019) with the difference, that they are just sampling the negative sample randomly from distant time points and the positive sample as a subsample from the anchor without considering the operational context as in our case. We propose to choose data samples that are recorded within a fixed timeperiod after a workshop visit as the positive pair (x_a as an anchor sample and x_p as the

positive sample) to ensure variability in the different operating and environmental conditions. All other samples are considered being dissimilar (x_n).

We evaluate the hypothesis that a model that is sensitive to normal degradation, is also sensitive to different fault types. However, certain fault types may be very dissimilar to degraded system conditions. It is not guaranteed that the sensitivity of a model to degradation patterns transfers to sensitivity to all fault types. If the fault data is very dissimilar to the the degradation, it may not provide any benefit. If however the data variations caused by a fault type resembles extreme degradation processes, the learned feature space will not only provide high sensitivity to this fault type but also the distances in the feature space can correlate to the severity of a fault.

To provide a solution to detecting all fault types, those that are similar to extreme degradation processes and those that are not, a combination of the contrastive model with other AD techniques is also possible as long as they are set not to be too sensitive to domain shifts in data. In this study we combine the contrastive model with one of the comparison method called HELM as described in Section 7.4.5 (Michau et al., 2020).

The **encoder model f** that is trained with the above loss function implementations depending on the data and label availability of the specific task. It, therefore, provides the desired feature representation. The model used in the paper is a deep convolutional model. Depending on the task and the type of the available dataset, the encoder model needs to be adapted. For the image dataset (railway sleepers as described in Section 7.3), a 2D convolutional model is used. For the timeseries dataset (railway wheel as described in Section 7.3), a 1D convolutional model is used. Details on the exact architecture are provided in Section 7.5. The last layer of the encoder model spans the feature space. The size of the feature space is set as described in Section 7.5. Depending on the data and label availability, this feature space is regularized differently.

Feature Space Regularization in a Supervised Setting: If this space is trained for a supervised task, where the goal is to distinguish between health conditions that are already represented in the training dataset, we regularize the feature space with a l2-normalization.

Feature Space Regularization in a Unsupervised Setting without faults: If this space is supposed to be used for anomaly detection in a subsequent step, it is beneficial not to have a restricted feature space (e.g. through l2 regularization). Since we aim to distinguish healthy from faulty conditions, we assume that a fault will be even more dissimilar to the degraded conditions.

7.4.2 Health Monitoring

In the second step, the learned feature representation is used for health monitoring of different assets within the railway system. Different methods are used depending on the data availabilities and characteristics of the railway case studies.

Supervised Health Monitoring: In the supervised case where a labeled dataset with different health conditions is available, a fully connected classification model is trained based on the feature representation using the cross entropy loss to distinguish between the different health conditions.

Unsupervised Health Monitoring without faults: In the more common case, where presumably only healthy data is available to train a model for health monitoring, we extract a health index that allows to monitor the health condition of an asset over time and ultimately to detect anomalies. We first train a OC-SVM on the learned feature representation. Once trained, the health index is extracted. At test time, the distance of the encoded data samples to the decision boundary of the OC-SVM in the feature space is measured. The feature space is trained to be semantically feasible in Section 7.4.1. I.e. we aimed to group similar health conditions close to each other, data from slightly different health conditions slightly further apart and data from substantially different health conditions far apart in the feature

space. If this is the case, the distance from the healthy class in the feature space can be representative of the severity of a defect. Therefore, the distance in the feature space to the decision boundary of the OC-SVM can represent an health index that is used to detect anomalies and monitor the health condition over time.

7.4.3 Health Monitoring with Partial Observable Railway Wheels

If the CM data is only capable of observing parts of the asset, reliable fault detection and diagnostics is challenging. WLC measurement sites with strain gauge sensors in Switzerland (see Section 7.3), for example, are only capable of covering parts of the railway wheel i.e. the wheel's condition is only observed partially by the measurement system. This is illustrated in Figure 7.6 whereas a lower limit of the observed region on the wheel per sensor is 28 cm.

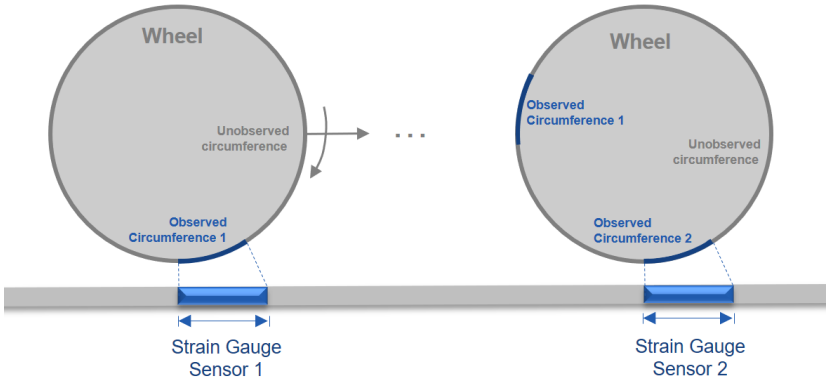


Figure 7.6: Illustration of the partial observation of railway wheels provided by a strain gauge sensor in a WSM measurement site.

If individual measurements from the measurement sites monitor only parts of the asset, data samples from defective wheels might not show any signs of faults in the data. Therefore, multiple consecutive measurements are required to decide on the asset's condition since this increases the probability that a defect is observed by the CM data. Furthermore, we aim to desensitise the fault detection model to variations caused by domain shifts e.g. due to measurement site calibration. To account for that, we choose to model the probability of a fault being sufficiently represented in the data with a binomial distribution, whereby we assume that the defect should be represented in at least K out of N measurements. The value N is chosen such that monitoring within a reasonable time span is possible, K is chosen by setting a desired probability threshold T as defined in Equation 7.2, whereby p corresponds to the probability that a defect is represented in an individual measurement from the measurement site.

$$T < \sum_{k=K}^{N} \binom{N}{k} p^k (1-p)^{(N-k)} \tag{7.2}$$

For the concrete application of railway wheel monitoring with WSM by strain gauge sensors, the probability of a fault being represented in an individual measurement from the WSM site corresponds to the percentage of the wheel circumference that is covered by the strain gauge sensors. However, this probability value depends on the current diameter of the railway wheel. This is shown in Figure 7.7 where the colored regions correspond to the parts on the circumference being monitored by the eight different sensors. The figure was produced with the lower limit of possible circumference coverage of the measurement site (28 cm of the circumference being observed by an individual sensor) and provides an lower limit of the probability of a defect being represented in a measurement. Moreover, wheel diameters

are used that are within the specifications of the monitored fleet. In the application case of railway wheel monitoring, where approximately five measurement sites are frequented within a day, we set N to five since we want to have a decision within a day of monitoring. Given that value, the lower limit of probability that an existing defect is represented in a measurement is shown in Figure 7.8, where the detection probability is shown in dependency of K and the diameter. In this study, we choose a value of 3 for K. Although the probability of detection is rather low for some diameter settings with $K = 3$, we consider this value to be an appropriate trade-off to prevent many false alarms due to measurement site failures. Furthermore, we would like to emphasize that the calculated probability corresponds to a lower limit (see above). To implement this rule, a wheel is labeled defective if the median of n consecutive health indices is above the defined threshold.

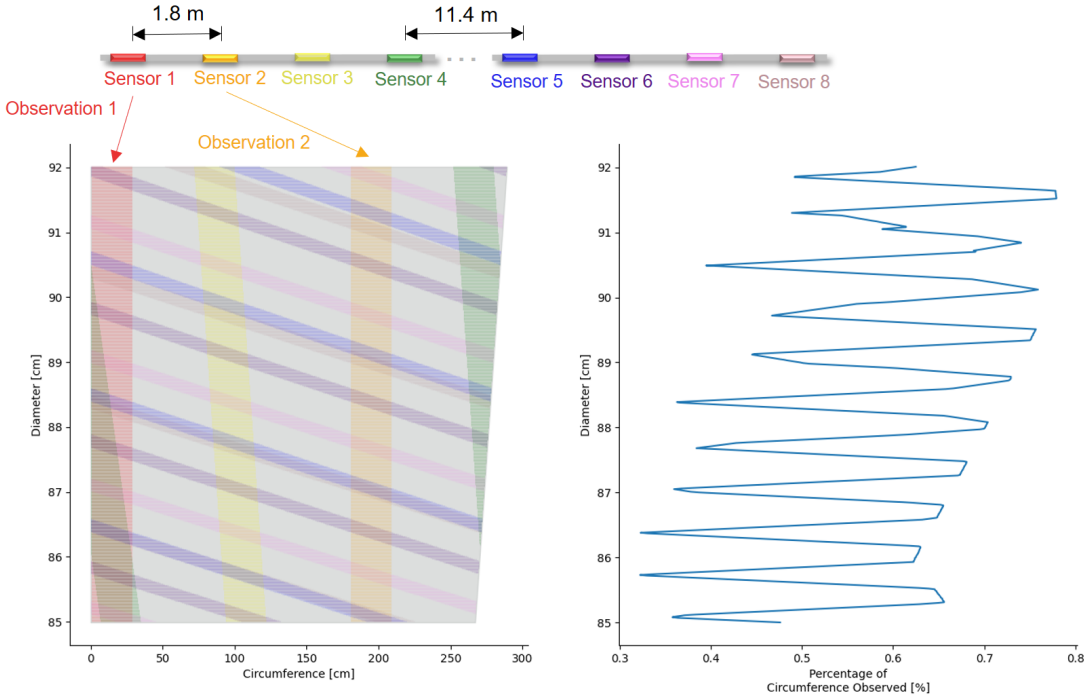


Figure 7.7: Wheel circumference regions monitored by the eight strain gauge sensors (blue regions) in dependency of different diameters compared to the entire wheel circumference.

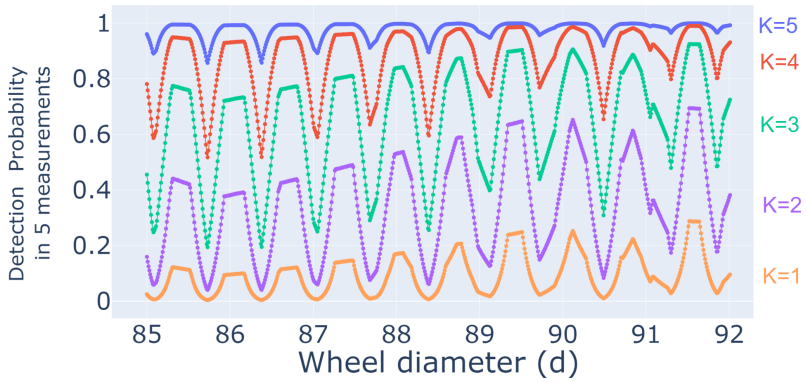


Figure 7.8: Probability of a defect being detected in at least K measurements out of 5 in dependency of the wheel diameter.

7.4.4 Performance Evaluation of Railway Applications

Supervised Health Monitoring of Railway Sleepers: We report the balanced classification accuracy of the trained classification model.

Unsupervised Health Monitoring of Railway Wheels: We report the balanced detection accuracy of the wheels in the test dataset. Since it is also important to detect the defect at an appropriate point in time, we also evaluate the detection time of the defects (see Figure 7.3b). For defective wheels, the time interval dt (number of days) of the detection before or after wheel defect manifested itself in the CM data according to expert labeling is reported (see Figure 7.3b). A negative value ($dt < 0$) corresponds to a detection in the green zone (see Figure 7.3b), a positive value ($dt > 0$) corresponds to a detection in the red zone, and a value of 0 corresponds to a detection in the orange zone ($dt = 0$). Since detections in the red zone can potentially be critical and the length of the red zone differ quite substantially between the different wheels, we additionally evaluate those with a relative measure dr of the total delay time interval $dt > 0$ to the entire time interval DT in which the defect is present (see Figure 7.3b). An early detection (in the green zone) could be considered favourable, as it allows for early maintenance planning. However, it can also indicate that the model is too sensitive to be used in real operations. Therefore, we consider detection in the orange zone as desirable.

7.4.5 Alternative Methods for Comparison

For the supervised sleeper dataset, a supervised comparison method is used. An end-to-end classification model is trained with cross entropy loss. The same CNN architecture is used for the proposed method and the comparison method; only the loss functions differ.

For the unsupervised AD task, unsupervised methods are used for comparison. First, we compare our results to a statistical measure called the dynamic coefficient (*dynCoeff*) - see Equation 7.3. This coefficient describes the ratio of the maximum dynamic to the static wheel load within each sensor measurement x. It is currently used in operations with a threshold of 1.8.

$$dynCoeff = \frac{max(x)}{mean(x)} \tag{7.3}$$

Furthermore, as a comparison method for feature learning, HELM is used that was used in previous case studies in similar setups(Michau et al., 2020). This method is not only suited for AD but also allows to track the evolving condition over time in the form of a health index. As the feature encoder model of HELM is trained to reconstruct the signal, a different model architecture is chosen compared to that of the proposed method.

7.5 Experimental Setup

Details on the exact experimental set up are provided in the following.

7.5.1 Sleeper Defect Classification

For all experiments on the sleeper dataset, a ResNet50 model (He et al., 2016) is used as the backbone architecture for the feature encoder model. On top of that, nine ReLu-activated fully connected layers are stacked (2048, 1024, 512, 256, 128, 64, 32, 16, 8 units). The architecture is chosen based on the performance of a validation dataset. The feature space is chosen to be eight-dimensional and the triplet loss is applied on the eight-dimensional features. The last fully connected layer has three units (three-class classification), that are trained in the second step with the cross-entropy loss. All models are updated over 50 epochs on batches of 48 samples using Stochastic Gradient Descent (SGD) with a learning rate of 0.001. The comparison model is trained using the identical architecture and hyperparameter setting, although the cross-entropy loss is applied on the last layer directly.

7.5.2 Health Monitoring Algorithm for Railway Wheels

A five-layer 1D convolutional model (with ReLu activation with a degree of 0.1) is used with 10 filters and a kernel size of 16 in each layer. Last, a fully connected layer is added with four nodes. The architecture is chosen based on a validation dataset (10% of the entire training dataset), whereby the feature space dimensionality of four is chosen to be as small as possible to encourage the encoder model to focus only on the relevant data variations. The model is trained with the Adam optimizer with default settings. To calculate the contrastive loss function, positive pairs are defined as data measurements that are recorded within the same month after the workshop visit. This timeframe is chosen based on domain knowledge. An OC-SVM is applied to the extracted features with a Radial Basis Function (RBF) kernel function and a threshold of 0.88. The threshold is set rather low due to the characteristics of the real data: First, the exact condition of the training dataset is not known. Second, individual sensors can be calibrated poorly, leading to anomalies in the training dataset. The health index is calculated at test time as the distance to the decision boundary of the OC-SVM. The comparison method HELM is trained using a single layer AE with 30 neurons, and a one-class classifier with 100 neurons. The multiplicative factor for determining the threshold is set to 1.0 and the threshold is set based on 88% of the training data (same setting as for the contrastive model), whereby an ensemble of five models were trained and ran. Other values were chosen to be the standard values for HELM ($C = 1e-5$, $\lambda = 1e-3$).

7.6 Results

The results obtained by the conducted experiments are reported below. The result on the supervised case of railway Sleeper classification is reported in Section 7.6.1, the results on the unsupervised case of railway wheel monitoring is reported in Section 7.6.2.

7.6.1 Classification for Railway Sleepers (supervised)

The confusion matrices of the three-class classification on the sleeper dataset are shown in Table 7.1. A balanced accuracy gain of 13.3% was achieved by employing contrastive feature learning. Spalling defects were misclassified more often by both models. They appear to provide a bigger challenge in terms of fault detection.

		Predicted					
		Cross-Entropy			Contrastive		
		H	C	S	H	C	S
Actual	H	171	12	58	**235**	3	3
	C	4	231	12	3	**226**	0
	S	16	26	165	11	10	**178**
	Balanced Accuracy	81.1%			**94.4%**		

Table 7.1: Confusion matrix of three-class classification of the sleeper dataset including healthy conditions (H), spalling defects, (S) and cracks (C).

7.6.2 Railway Wheel Monitoring (unsupervised)

The results from the railway wheel case study are presented based on the extracted health index that is monitored over time and the decision rule for partial observable measurements. First, the AD results are displayed and second, the detection time is evaluated (see Section 7.4.4). In Table 7.2 the AD results are shown. The *dynCoeff* appears to be the least sensitive model to faults. It recognizes all healthy wheels correctly. However, it detects only four defective wheels. By contrast, both feature learning methods perform better than the *dynCoeff* and equally well in detecting 63 out of 79 faults. Additionally, we report the results

of an ensemble of the HELM and the Contrastive+OC-SVM model. The decision rule of the ensemble is the following: Any wheel that is detected as having a defect by either of the models (HELM or the Contrastive+OC-SVM) is labeled as defective.

		Predicted							
		dynCoeff		HELM		Contrastive + OC-SVM		HELM + Contrastive	
		Defect	Healthy	Defect	Healthy	Defect	Healthy	Defect	Healthy
Actual	Defect	4	75	63	16	63	16	71	8
	Healthy	0	16	1	15	1	15	2	14
	Balanced Accuracy	52.5%		87.7%		86.7%		**88.7%**	

Table 7.2: AD results on the railway wheel dataset.

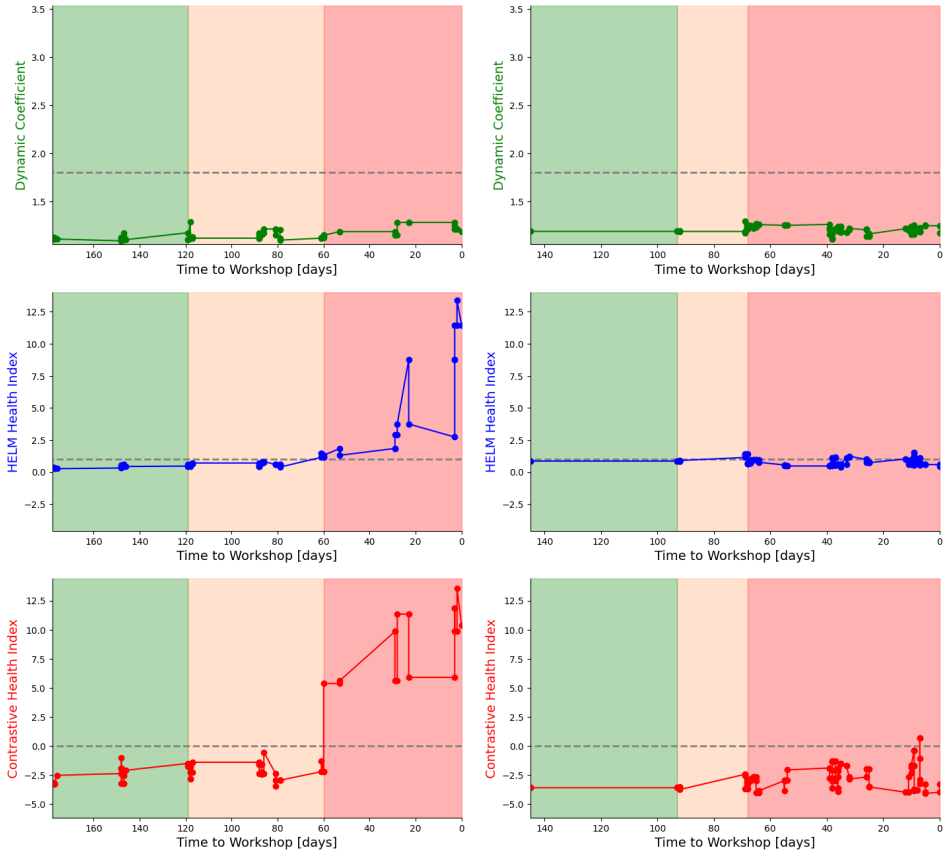
Two examples of health monitoring over time are shown for one wheel with spalling defects (see Figure 7.9a) and one with crack defects (see Figure 7.9b), whereby the background color indicates the ground truth label from the domain experts as defined in Figure 7.3b and the x-axis shows the number of days left until the wheel was inspected and the defect was identified. The HELM health index is scaled by the threshold as proposed in (Michau et al., 2020) and the contrastive health index is scaled by a constant value of 50. On top, the *dynCoeff* is plotted over time (green line), in the middle, the HELM health index is plotted (blue line) and on the bottom the health index extracted from the contrastive feature space is shown (red line). The *dynCoeff* is neither able to detect the shelling defect nor the crack defect. But both other methodologies are capable of detecting the shelling defect - even at the same point in time. However, for the contrastive model (on the bottom) a clear jump is visible at an early point in time, suggesting that the learned feature representation is sensitive to variation in the data caused by shelling defects. For crack defect, both models show less sensitivity. The HELM model, however, shows a higher sensitivity since the defect is detected considerably earlier (in the orange zone).

Shelling Defect Detection Time: The detection time of each model is evaluated in Table 7.3, where the model's detection time is compared with the domain experts' annotation as described in Section 7.5. The model based on the *dynCoeff* detected 3 out of 26 shelling defects (see TP column) and all of these were detected after the fault became obvious in the data (in the red zone), shortly before the next workshop visit (dr close to 1). HELM is more sensitive and detected 21 of the 26 shelling defects, most of them (10 wheels) in the green zone, before the defect became obvious in the data. The two wheels detected in the red zone were still detected close to the expert's label ($dr < 0.5$). The contrastive model detected most shelling defects (23 out of 26 defective wheels), of which the majority were detected in the orange zone (15 wheels).

Crack Defect Detection: The results for the cracks are shown in the lower half of Table 7.3. The *dynCoeff* detected 1 out of 53 crack defects and it was detected late, shortly before the next workshop visit (dr close to value 1). HELM detected most of the crack defects (42 of the 53 wheels), most of which were detected in the orange zone, i.e. exactly when the fault manifested in the data. The contrastive model detected 40 crack defects in total, most of which were detected late (17 wheels).

7.7 Discussion

In this work, we proposed contrastive learning to improve the robustness of fault detection and diagnostics and induce a robust encoding in the feature space. We proposed how this can be achieved in supervised and unsupervised tasks for real railway applications and evaluated the suitability of the learned feature representations for health monitoring of railway assets.



(a) Shelling Defect (b) Crack Defect

Figure 7.9: Health Index of Wheel Trajectories before the Workshop visit. The *dynCoeff* is plotted in green on top, HELM is shown in blue in the middle and the contrastive model is shown in red at the bottom.

Supervised vs. unsupervised contrastive learning: On the sleepers fault diagnostics task based on the image dataset, if the features were trained in a supervised way, the classification performance was improved considerably. This showcases the benefits of contrastive learning in cases when faulty conditions and normal variations in the data are inherently hard to distinguish based on the available CM data. On the wheel health monitoring dataset, in the unsupervised setup with presumably no faults, the result depends on the fault type. While both feature extraction methods performed considerably better compared to the *dynCoeff*, both feature learning methods performed equally well on the AD task over all the defect types. HELM performed better on the crack defects and the contrastive learning model performed better on the shelling defects. Combining the two methods resulted in the best detection performance with the HELM being more sensitive to cracks and the contrastive model being more sensitive to shelling.

Learning Sensitivity to Degradation without Ground Truth Information: To learn a contrastive feature representation in the unsupervised setup without any observed faults, we aimed to train a feature encoder that is sensitive to the normal degradation process of a wheel. However, training a feature space that is sensitive to normal degradation is challenging given an unlabeled dataset. Due to the lack of labels for degraded conditions during training, it was not known which data samples would form a suitable triplet pair. A suitable positive pair needs to represent the same degradation condition under different operating conditions. Whereas a suitable negative pair needs to have different degrees of degradation represented in the CM data. This is challenging without any ground truth

Method	FN	TP	Total Time dt [days]			Relative Time $dr = \frac{dt}{DT}$		
			$dt < 0$	$dt = 0$	$dt > 0$	$dr < 0.1$	$dr <= 0.5$	$dr > 0.5$
	Shelling							
dynCoeff.	23	3	0	0	3	0	0	3
HELM	5	21	10	9	2	1	1	0
Contrastive+OCSVM	3	**23**	4	15	4	1	1	2
	Cracks							
dynCoeff	52	1	0	0	1	0	0	1
HELM	11	**42**	15	23	4	0	0	4
Contrastive+OCSVM	13	40	7	16	17	1	2	14

Table 7.3: Time of railway wheel fault detection on shelling and crack defects. Column 1 shows the defects that were falsely labeled as healthy (FN) and the true positive (TP) detected defective wheels. Column 2 shows the total number of correctly labeled defective (TP) wheels that were labeled in the green ($dt < 0$), orange ($dt = 0$), or red ($dt > 0$) zone in Figure 7.3b. Column 3 shows the relative value of the time difference to the total time of the defect (DT) of the wheels detected in the red zone.

information on the true degree of degradation, especially since degradation of railway wheels does not affect the CM data substantially. Without access to the ground truth information, we used the time passed from the last workshop visit as a proxy for the degree of degradation relying on the hypothesis that coaches of a fleet are operated in a similar way and therefore, the degradation process can be assumed to be comparable in time between the different wheels. One of the interesting lessons of the case study was that the choice of a suitable time interval to define similar and dissimilar degradation states is essential to train a feature space that is sensitive to degradation. If the degree of degradation in the positive pair would differ substantially, we would impose invariance to different degrees of degradation. Vica versa, if the condition in the negative pair would not be in different degradation states, the contrastive loss function would impose sensitivity to other factors of variations in the data instead of sensitivity to a change in the health condition. To circumvent this challenge, domain knowledge it required to find an appropriate time interval.

From Degradation Sensitivity to Fault Detection: The goal of training a feature space based on degradation was to evaluate to which extent an encoder model that is sensitive to degradation is also sensitive to faults. However, different fault types induce different patterns in the CM data and the model might only be sensitive to certain patterns and display insensitivity to others. This observation was also found in our experiments. The contrastive encoder's sensitivity to degradation transfers well to sensitivity to shelling defects but not to crack defects. For shelling defects, the contrastive model not only detected most of the defects but also determined the time of fault occurrence best. For cracks, however, this good performance could not be replicated. The proposed model detected fewer cracks compared to HELM. One possible explanation for this could be that shelling defects may be more similar to the inherently occurring degradation than cracks. Another possible explanation is that due to the lack of labels in the training dataset, it is not known if faulty data is also present in the training dataset. If indeed faulty data may have been present in the training dataset, faulty and healthy data of various degradation states may have formed a positive pair. This would inevitably result in insensitivity of the encoder model to the respective fault pattern. Therefore, if many crack defects were present in the training dataset, it would explain the poor performance of the contrastive model to detect crack defects. This is plausible since crack defects are generally less visible compared to shelling defects, increasing the probability that a crack defect is not identified and reported by the workshop inspectors. Furthermore, from qualitative visual evaluations of the test dataset, it is apparent that crack defects impact the strain gauge signals less compared to the shelling defects,

making the cracks more difficult to detect with strain gauge sensors. This might be another explanation why a model trained to be invariant to certain variations in the data (caused by operating or environmental factors) showed lower sensitivity to cracks. When combining the two approaches the highest balanced accuracy of 88.7% is achieved since the ensemble of both approaches benefits from the sensitivity of HELM to cracks and the sensitivity from the contrastive model to shelling.

Detection Time: The evaluation of the detection time showed that HELM is the best in performing early detection (green zone in Figure 7.3b). Early detection might be considered desirable as it enables early maintenance planning. However, premature detections can also lead to additional work and wasted resources if the wheel is sent to the workshop too early. Surprisingly, HELM is not the most sensitive model with respect to all fault types as it detected fewer shelling defects compared to the contrastive model. Further, the contrastive model detected most of the shelling defects in the same timespan as the domain experts - 15 out of 26 wheels detected in the orange zone. However, the contrastive learning model is less sensitive to cracks, where it often resulted in late detections (17 out of 53). In general, the sensitivity of the models can be adapted to the requirements of the users. If, for example, the user immediately stops operating the train given a detected defect, then early detection would result in machine downtime. If, however, the model is used to make long-term maintenance plans, then early detection is desirable. Furthermore, early detection of faults and the corresponding health index as a severity measure can provide additional information since fault severity has not been defined yet or tracked before, early detection could provide additional information on how faults evolve. This information can be verified in the future when the trains are entering the depot or workshop. Thus, the severity evolution of faults can be verified in the future and the model will enable to monitor the severity evolution.

Fault Occurrence: A surprising finding in the labeling process depicted in Figure 7.3a of the wheel defect dataset is that many of the healthy wheels in the preliminary test dataset were identified as anomalous by the domain experts, which resulted in a small number of healthy wheels in the test dataset. It should be noted that the domain experts who evaluated the data did not have access to the real condition of the wheel but only to the data. In contrast, the maintenance technicians in the workshops primarily use visual inspection information to label the health condition of the wheel. In the future, it would be interesting to investigate whether the wheels that have been detected as defective by the domain experts and the contrastive model, show a different type of defect that the maintenance technicians in the workshop might not be familiar with and may not be used to detecting through visual inspection.

7.8 Conclusion

In this work, we proposed to use contrastive learning to improve the robustness of fault detection and diagnostics by learning a robust feature encoding. We proposed how this can be achieved and implemented in a supervised and a unsupervised tasks without faults for two real railway applications: Supervised railway sleeper diagnostics and health monitoring of railway wheels, where no labeled fault data was available for training the model (unsupervised) but the training dataset was presumed to be mainly healthy. Although the tasks of classifying sleeper conditions and monitoring railway wheels and detecting defects differ in many aspects, the conducted experiments demonstrated that contrastive learning improves the performance of different fault detection and diagnostics tasks in the railway system in both supervised and unsupervised setups as compared to state-of-the-art methods. This supports our initial assumption that contrastive learning is a suitable learning paradigm for different applications in railway systems. In future work, we will integrate a monoticity constraint for the health index and will explore the suitability of the feature space for prognostics tasks. One potentially promising direction could be to incorporate some observed faults in feature learning

and investigate semi-supervised setups rather than applying solely unsupervised approaches. Generalization to other fleets will also be investigated in the future.

8 Discussions

Throughout the dissertation, we addressed different types of data scarcity and label quality challenges. Methods were developed for different maturity levels of assessing the health condition of a complex industrial system and detailed discussions on the individual contributions were presented in the respective papers. The major motivation was to address label and data scarcity settings that are relevant in monitoring the condition of a complex industrial asset. Additionally, constraints that previously limited the application of deep neural networks to fault detection and diagnostics in changing operational environments of industrial assets were relaxed and alleviated. This opens many interesting points of discussion. We will elaborate on each of the modules whereby we first discuss the key findings, followed by a discussion of the mitigated limitations as well as the applicability of the proposed modules in the real world. Lastly we will summarize the contribution of the entire framework with respect to the limitations mitigated and the application of deep learning to PHM tasks.

8.1 The modules

The key findings of the four modules are summarized from the respective chapters.

8.1.1 Module 1: Contrastive Learning for Fault Detection and Diagnostics in the Context of Changing Operating Conditions and Novel Fault Types

In Chapter 4, we proposed a contrastive learning model that is able to achieve the following two objectives simultaneously: (1) robustness towards domain shifts and (2) sensitivity towards novel faults.

Contrastive Learning for Domain Generalization under Small Domain Gaps: First, we evaluated if the proposed model can generalize well to unforeseen domain shifts. The proposed contrastive learning model performed best in classifying known health conditions under small previously unknown domain shifts (accuracy of 100%). One of the comparison methods (normal classification model optimized with the cross-entropy loss) performed only slightly worse but within a comparable range. However, both, the silhouette score of 0.81 and the visualization of the respective feature embeddings in Chapter 4, showed that the class clusters of the contrastive model are considerably more cohesive within the same class cluster and separated towards other class clusters compared to the best performing comparison method where only a silhouette score of 0.38 was reached. This suggests that the contrastive model could be more suited if the domain gaps increased. Evaluating this, however, is subject of future work. Considering that real faults tend to evolve in a continuous manner rather then in a discrete manner, it would be of high interest to evaluate the physical plausibility of the learned feature representation. To this end, two key aspects of the learned feature representation should be evaluated: (1) whether the different severities from the same fault type are grouped close to each other in the feature space, in comparison to other fault types and (2) whether the feature representations of faults with intermediate severities are positioned between the discrete severity levels (i.e., lower and higher) of the same fault type. Due to the limitations of the dataset, it was not possible to conduct any further analysis on domain generalization in this particular case study. However, in Chapter 7, we evaluated the distance in the contrastive feature space with respect to the fault evolution over time.

Contrastive Learning for Novel Fault Detection: Second, we evaluated if the proposed model is suited to detect novel health condition as outliers by clustering the feature space.

The constrastive learning model performed best in clustering the learned feature space and identifying outliers by a large margin. An adjusted mutual information (AMI) of 73.5% on average was reached with the proposed model. In comparison, the next best performing comparison method reached only an AMI value of 45.5% on average. Moreover, 94% of the outliers detected by the proposed contrastive learning model actually corresponded to real new fault types and severity levels. The compactness of the known class clusters in the contrastive feature space enables the identification of outliers in the feature space more reliably compared to the comparison methods, where the class clusters are less compact. Thus, the invariance to small domain shifts of the contrastive model does not impact the model's sensitivity to novel faults i.e. does not translate to invariance towards changes in the health condition.

Limitations Mitigated: Our proposed model is able to generalize to variations in the data caused by domain shifts that are similar to those that it has seen before while staying sensitive to variations in the data that are caused by a change in the health condition. Previous methods were only able to achieve either one of the objectives: either generalization or sensitivity. This makes the proposed methodology more applicable in real applications.

Applicability in Real Applications: Chapter 4 represents a mature and robust solution for condition assessment as (1) undetected domain shifts do not pose a safety risk; (2) detected domain shifts do not require immediate retraining of the health monitoring model and (3) safe operation can still be enabled as novel faults can be detected and investigated.

8.1.2 Module 2: Controlled Generation of Unseen Faults for *Partial* and *Open-Partial* Domain Adaptation

In Chapter 5, we proposed a data generative approach that enables domain adaptation, also under extreme cases of label space discrepancies when the domain gap is large.

Domain Adaptation with Different Sizes of Domain Gaps: While one comparison method achieved comparable results to our proposed method under small domain gaps, the performance gain of the comparison method with respect to the baseline method dropped considerably under a large domain gap compared to our method. For example, the performance gain achieved with the best comparison method based on feature alignment on large domain gaps was 10.32% under the extreme *Partial* domain adaptation setup. In comparison, the performance gain achieved with our proposed method was 23.46%. This demonstrates that our proposed data generative approach performs better compared to feature alignment approaches on large domain gaps.

Domain Adaptation with Different Types of Extreme Label Space Discrepancies: As demonstrated in Chapter 5, our proposed method is not limited to one type of label space discrepancy. Instead, it can be universally applied to *Partial* (performance gain of 23.46% on large domain gaps with respect to the baseline method) as well as to *Open-Partial* domain adaptation tasks (performance gain of 6.21% on average with respect to the baseline method). Previous methods are typically only applicable in either one of the settings. While this is not a hard constraint in applying previously proposed methods, our solution satisfies the requirements of industrial applications by providing one solution that is applicable in many possible scenarios without major adaptions. We also want to point out that, although we have only demonstrated the transfer of knowledge between two domains, the proposed method can be easily adapted to distribute the experience on the fault types between many domains. Once the fault signatures are trained on one domain, they can be transferred to

each new domain that emerges in the future. This makes the proposed method particularly flexible.

Unsupervised and Controlled Data Generation: In Chapter 5, we proposed unsupervised generation of domain-specific fault data that was not observed beforehand. The generated data needed to satisfy two requirements to be usable for domain adaption: (1) it needed to satisfy the specifities of the particular target domain and (2) it needed to be specific to a certain fault type and severity level. In absence of any real domain-specific fault data, however, it is impossible to know how the target fault data should look like, making it impossible to learn a generative model directly. To overcome this limitation, we assumed that we can disentangle fault data in the frequency space into domain-specific components that are independent of the fault class and into class-specific fault signatures. The validity of this assumption was implicitly demonstrated in the domain adaptation experiments. The improved performance (see above for concrete performance gains) achieved with the synthetically generated data samples, suggests that the synthetic data in fact resembled true target data better than the initially available source faults, even when the domain gap is larger. Therefore, the proposed method enables the controlled generation of previously unseen data and thus, also provides a significant step forward in the data generation literature where the generation is typically limited to known classes or approximations between known domains.

Limitations Mitigated: In Chapter 5, we demonstrated the effect that even small changes in the hyperparameter settings can have on the performance of deep neural networks in the target domain. By generating physically plausible target fault data, a synthetic validation dataset can be generated to tune the hyperparameters of any applied domain adaptation method. We showed how this can lead to a more optimal setting of the hyperparameters. Therefore, our proposed method can be employed when target data is not available in contrast to other methods that rely on an available target validation dataset that includes target faults that have not yet been observed or validation domain shift tasks to tune its hyperparameters.

Applicability in Real Applications: The developed methodology lends itself especially to the extreme but highly relevant case for PHM applications, where only one class is shared between the domains and the size of the domain gap is unknown but potentially large. Furthermore, the method is applicable if no validation tasks are available and, thus, satisfies real requirements. Moreover, it is applicable to different types of label space discrepancies (*Partial* and *Open-Partial*). Thus, Chapter 5 is a pivotal pillar in transferring the models between different domains under realistic conditions.

8.1.3 Module 3: Improving generalization of deep fault detection models in the presence of mislabeled data

In Chapter 7, we proposed a method that can stabilize the optimization process of a deep learning model in presence of label noise if neither exact ground truth information is available on the type and amount of label noise, nor a clean validation dataset. We proposed to identify mislabeled samples in the gradient space of the deep learning model's optimization.

Detection of Mislabeled Samples in the Model's Gradient Space By investigating the gradient space of the classification models in Chapter 6, we prevent 'memorization' of mislabeled samples before they were even used for the actual model update step. The results of our proposed method and also the comparison methods differed per noise level in the data (percentage of mislabeled samples). However, our proposed method outperformed almost consistently the comparison methods. Only in two experimental setups on the smaller

dataset (998 data samples for training) with a low label noise level, one comparison method slightly outperformed our proposed method (+8.5%). In all other experiments, however, our proposed method was the best performing one. The benefits of our proposed method is particularly pronounced under a severe label noise level (40% label noise). A performance gain of +20% resp. +10% was achieved with respect to the two comparison methods on the smaller timeseries dataset and on the bigger image dataset a performance gain of +31% resp. +14% was reached. Thus, the proposed method demonstrates to be effective in enabling robust classification despite the presence of label noise.

Limitations Mitigated In Chapter 6, we relaxed the assumption about requiring ground truth information on the exact type and amount of label noise. We developed a method that can perform well under just a rough estimate of label noise. As a rough estimate we defined three different label noise levels, each with an upper threshold for the percentage of mislabeled samples. We consider such a rough estimate to be easily accessible in real operations as it could , for example, represent how unsure the domain experts are about the labels. This makes this method more applicable under real-world constraints where such rough estimates can typically be provided.

Applicability in Real Applications: Chapter 6 is particularly relevant for systems that are relatively low on the safety-criticality scale, have been operational for an extended period, and possess a dataset containing information on typical faults. Accurate labeling, however, is impossible to obtain as the system is not being monitored by experts constantly. An example of such a scenario is provided in Section 8.3. If fault data is available but the labeling is unreliable, the robust model developed during the condition monitoring phase can then be used to collect a sufficiently large and reliably labeled dataset within a particular domain. This dataset can facilitatedomain adaptation in the subsequent monitoring phase, even in presence of unreliable labels.

8.1.4 Module 4: Contrastive Feature Learning for Fault Detection and Diagnostics in Railway Applications

In Chapter 6, we proposed a contrastive model for fault diagnostics and fault detection on two real condition monitoring datasets within a railway system. Contrary to conducting experiments on datasets that were acquired under laboratory conditions, in Chapter 7 we conducted experiments on condition monitoring datasets that were acquired under real in-service conditions. Datasets of controlled in-workshop conditions typically only partially represent the complexities and challenges of real in-service conditions of systems that operate in an open environment. Thus, models developed for datasets recorded under real condition face bigger challenges such as a larger variability within the healthy class or a less pronounced fault patter per fault type.

Fault diagnostics for railway sleepers: On the fault diagnostics task of railway sleeper defect classification, the proposed methods achieved a considerable performance gain compared to a normal classification model (+13%). This showcases that the contrastive learning idea is especially suited for tasks where the fault pattern in the condition monitoring data is less pronounced and resembles normal variations in the healthy class closely.

Fault detection for railway wheels: While contrastive learning has shown to be successful on fault diagnostics tasks in Chapter 4, its application is not directly transferable to an anomaly detection setting where presumably only healthy data is available. Typically, to induce sensitivity to faults, the contrastive learning paradigm requires the availability of fault

data (labeled or unlabeled). This is typically not available when a new system starts to be monitored or when the condition monitoring system is newly installed. To alleviate this, in Chapter 7, we show how contrastive learning can be applied in this anomaly detection setting where no fault data is available. We proposed to train a feature encoding that is sensitive to degradation processes and evaluate if this sensitivity translates well to sensitivity towards faults. The experiments conducted on the task of detecting railway wheel faults show that this is partially true. The contrastive model has shown a higher detection rate for one defect type (shelling), where two more defective wheels were detected. However, for a second defect type (cracks) two defective wheels were not detected by the proposed method but by the comparison method. This suggests that either the sensitivity to degradation translates only well to some fault types that resemble the characteristics of degradation or that the imposed invariance to operational conditions translates to insensitivity to other fault types. The exact reason, unfortunately, cannot be determined due to the lack of ground truth system conditions. However, the results show that the proposed method is a valuable addition to other fault detection methods as it provides more sensitivity for certain fault types.

Limitations mitigated: Although contrastive learning has shown to be very useful in the context of PHM tasks, its application is limited when no fault data is available. In this module we alleviated this limitation.

Applicability in Real Applications: This module is applicable in the beginning of the monitoring of an industrial asset, where either no faults have occurred yet or where fault data is available but it is unknown how representative it is. It therefore provides a first step of a condition assessment solution.

8.2 The framework

Lastly, we discuss the contribution and effectiveness of the proposed framework.

Reaching a Mature Solution for Condition Assessment: The solution proposed in 'Domain Generalization module' represents a mature level of condition assessment solutions due to the following two reasons: (a) domain shifts do not decrease the performance of the deployed fault diagnostics model without any further adaptation of the model and (b) novel faults can be detected as such and distinguished from known faults. Reaching such a mature solution, however, cannot be achieved within a short period of time after taking a condition monitoring system into operation as the initially recorded data will not be very representative of future conditions. A long data acquisition time to increase the representativeness is often not acceptable in practice since an industrial asset needs to be operated safely from the very beginning. Furthermore, providing accurate supervision e.g. by domain experts during a long data acquisition time may not be accepted in practice. Instead, data-driven solutions can be deployed at different phases of monitoring an asset in order to (a) enable safe and efficient operation at earlier phases of the life cycle of an asset and (b) make the models better adapted to the realistic data scarcity settings by overcoming the realistic lack of information in each phase. The four modules developed in this dissertation provide more reliable solutions for different phases of monitoring complex assets, where the data availability and the constraints differ and they enable the condition assessment solution to reach a higher maturity level within a shorter amount of time.

Progressively Increase the Maturity Level of the Condition Assessment Solution: Each of the four developed modules can be applied in different phases of monitoring the

condition of a system. Thus, the proposed framework comprising four modules enables to progressively and efficiently increase the maturity level of the condition assessment solution.

When a condition monitoring system is taken into operation, the '*No Fault Label module*' has shown to detect certain fault types more reliably (see Chapter 7) than other anomaly detection methods. The module can be deployed in an operational system such that gradually the condition monitoring data is assessed and potential faults are detected. Thus, the module helps the operators to investigate, document and label the ground truth status of the asset e.g. about the fault type and the severity in a more targeted way. If a labeled dataset of different health conditions is available, a fault diagnostics model can be trained. If however the labels in the datasets are noisy, the '*Label Noise module*' can be applied. The module has demonstrated to provide more robust classification of known health conditions if the training dataset is subjected to label noise. Thus, the module can provide more reliable labels during operations on the recorded condition monitoring for future developments. If one domain (e.g. one operating condition or one unit of a fleet) can be monitored reliably, it is desirable to transfer the model to new domains. The '*Extreme Domain Adaptation module*' demonstrated that it can outperform other domain adaptation methods even under extreme label space discrepancy settings and large domain gaps. The module, therefore, enables the transfer of fault diagnostics models to new domains within a shorter period of time. We want to point out that the data generative method developed in this module provides synthetic data that could be sufficient to directly implement the method proposed in the *'Domain Generalization module'*: Once the synthetic dataset is generated for some domains where only healthy data is available, the contrastive learning model as proposed in Chapter 5 can be trained on the synthetically generated data. This will enable us to make the fault diagnostics even more robust and may require the generated data to be less precise. Furthermore, it enables to reach a higher maturity level for the condition assessment solution within a shorter period of time, as no fault data needs to be available in most of the source domains. However, evaluating this is subject to future work. Lastly, a mature solution to condition assessment should be robust towards domain shifts while being sensitive to novel fault types. The '*Domain Generalization module*' demonstrated that it can achieve both objectives and this provides a mature solution to condition assessment. It neither poses a safety risk if a domain shift stays undetected nor requires retraining once a domain shift was detected. Moreover, it still is capable of detecting novel health conditions that can be investigated by operators and then consecutively be added to the method.

Common Constraints Mitigated in Data-Driven Condition Monitoring: One major motivation behind each methodological development in this dissertation was to further relax or alleviate non-realistic requirements and constraints of previously proposed methods.

One limitation that is very important from a practical point of view and that we overcame in this research is the dependency on ground truth information when validating the developed deep learning methods. In previous research, most of the methods proposed for classification with label noise, for example, rely on either a clean validation dataset or some prior knowledge about the label noise. Similarly, methods proposed for domain adaptation with label space discrepancies required a validation task to find optimal parameters of the proposed methodology. Despite the impressive advances in both fields, in practical applications, such assumptions cannot be fulfilled. It is unrealistic to assume the availability of ground truth information for methods that aim to overcome the lack of exactly that ground truth information. The methodological developments in this dissertation relaxed these limitations in two ways. Firstly, the method proposed in Chapter 6 enables robust classification under only an approximate estimation of the magnitude of the label noise. Secondly, the method proposed in Chapter 5 enables the generation of a synthetic fault dataset of previously unobserved

faults. We have demonstrated that the synthetically generated data can be used for hyperparameter tuning, alleviating the requirement of a validation task that includes real previously unobserved fault data. The methodological developments in this dissertation advance the applicability of deep learning models under more realistic settings and are therefore flexible to be applied in many different applications that are not limited to fault diagnostics.

In addition to addressing the lack of ground truth information, we also aimed to tackle extreme setups of data scarcity that are common in PHM tasks and have not been sufficiently addressed in the current literature. Most notably, we have achieved this in the *'Domain Adaptation Module'*, where we enabled domain adaptation under extreme label discrepancy settings. Moreover, also in the other modules (*'No Fault Label Module'* and *'Domain Generalization Module'*), we developed methods that are applicable when neither all fault types were represented in the training dataset, nor all non-informative factors. These scenarios are especially important in the context of PHM tasks. The superior performance of our proposed methodology demonstrates that this thesis has advanced deep learning techniques to be applicable also in extreme cases of data scarcity.

8.3 Datasets

In this thesis, we conducted experiments on benchmark datasets that are recorded under test rig conditions and commonly used in research studies as well as datasets that are recorded under real operating conditions from real applications.

Although real datasets provide a realistic test bed as they represent real challenges (such as the railway sleeper dataset or the railway wheel dataset in Chapter 7), they may not allow for concrete evaluations of specific challenges such as domain shifts. If not all factors of variation in the data are known or can be controlled, the identification of distinct domains is not possible. This makes these datasets less suitable candidates to demonstrate the effectiveness of domain adaptation or generalization approaches. Furthermore, data collected under real conditions are often not available in open source making reproducibility and comparison of the results impossible. Therefore, it is often impossible to use those as benchmark datasets. The lack of accurately labeled data is another challenge of datasets from real applications as addressed in Chapter 5. One example for this is the railway wheel case study presented in cha:5 where label noise can be introduced by rule-based labeling: Defective wheels are typically only detected during workshop visits that are not very frequent. During these workshop visits, it is impossible to determine when exactly the fault has been initiated in hindsight and accurate labeling of the respective data around initiation time is impossible. To circumvent this problem, basic labeling rules have been applied in this application previously to train binary fault diagnostics models (Krummenacher et al., 2017) whereby wheels that just left the workshop (after maintenance) have been labeled as healthy within a certain time span and detected defects are labeled as faults within a certain time span before the workshop visit. If the considered time span used for labeling is small, these fixed rules can result in biased models as they are only trained on freshly maintained wheels (not considering normal degradation processes) and pronounced faults (not considering the initiation phase of a fault). If the considered time span for labeling is larger, label noise is introduced as real defects might have been undetected and thus, are falsely labeled as healthy or detected defects might have occurred later in time as the fixed time span considered for labeling and thus, healthy conditions are falsely labeled as faulty. This scenario, that faults can only be detected in non-frequent workshop visits and fixed rules need to be applied to label the real time data, introducing label noise, is common in PHM applications.

Contrary to datasets from real applications, the first kind of datasets (those that are recorded under test rig conditions) allow for benchmarking and detailed analysis under domain shifts since typically all parameters under which the data is recorded are controlled and known. Thus, these datasets allow to demonstrate the effectiveness of the proposed

methods concretely with respect to e.g. domain adaptation or domain generalization and are often used in several research studies as benchmarks to compare to state-of-the-art methods. One of the benchmark datasets used in this thesis (CWRU dataset) has been reported to be affected by artifacts (Smith and Randall, 2015) and its suitability for domain adaptation experiments has been questioned recently as the transferrability to other physical bearings cannot be tested (Hendriks et al., 2022).

In this thesis, we still used the CWRU dataset as a benchmark dataset as it has been used extensively in the literature and thus, provides a good comparison to current state-of-the-art methods. First, the CWRU dataset is used in the domain generalization experiments outlined in Chapter 4, within an experimental setup that differs from that of conventional fault diagnostics (as described by Smith and Randall (2015)). The goal of the experiments in this thesis is not to determine whether the proposed algorithm is particularly suited to extract features that are unique to bearing faults for diagnostics purposes. Rather, the aim is to assess if the learned features for fault diagnostics are robust to domain shifts while being sensitive to changes in the health condition. Consequently, the artefacts present in the CWRU data do not diminish the validity of the conducted experiments with respect to robustness to domain shifts. Second, the CWRU dataset is used for benchmarking in the domain adaptation experiments outlined Chapter 6. In addition to the experiments on the CWRU dataset, we have conducted experiments using a second dataset, the Paderborn dataset. Interestingly, the performance improvement of our proposed method is more evident in the case of the Paderborn dataset.

Furthermore, datasets recorded under test rig conditions might lack expressiveness. To name one example, typically only discrete severities (e.g. discrete sizes of defects) of the same fault type are represented in the datasets as the faults are artificially imposed and do not evolve naturally. Therefore, continuously evolving fault severities cannot be evaluated on these datasets. Identifying not only different fault types but also severity levels could either be defined as an ordinal regression problem for the different severity levels or as a classification problem where the different severity levels from the same fault type are considered as individual and discrete classes. In this thesis, we adopt the problem formulation commonly used in the field of fault diagnostics, as described in (Neupane and Seok, 2020). Specifically, we assume that fault severities can be categorized in discrete classes, which we believe is a reasonable and realistic assumption. However, we also recognize the importance of addressing the continuous evolution of fault severities, in addition to the discrete classification problem discussed in Chapter 4. Hence, we explore the continuous evolution of fault severities in Chapter 7, in order to provide a more comprehensive analysis.

9 Conclusions

9.1 Research Objectives Revisited

The overarching aim of this research was *to develop methods that can efficiently deal with the different types of data and label scarcity in different phases of condition monitoring under real-world constraints.*

This aim was accomplished by focusing on four objectives: (1) The development of a fault diagnostics and detection method which relaxes the need to retrain the used models given each domain shift but rather is capable of generalizing well to unknown domains while being sensitive to novel faults (Chapter 4); (2) The development of an approach which can transfer fault diagnostics models from one domain to another under extreme label space discrepancies and large domain gaps (Chapter 5); (3) The development of a classification method that enables robust fault diagnostics in the presence of label noise (Chapter 6); (4) The development of a fault detection method that particularly takes the challenge of large non-informative variations within the healthy class into account (Chapter 7).

9.2 Summary

Based on the main research aim, this dissertation proposes a framework consisting of four modules. Each of the four modules can be applied at different phases of monitoring the health condition of an industrial asset and aims to progressively increase the maturity level of the condition assessment solution within a short period of time: Starting from enabling the detection of faults under large variability in the healthy class to being able to distinguish different fault types and severities in known and unknown domains while being able to identify the the occurrence of novel previously unobserved fault types. In each of the modules previously existing challenges, constraints and limitations are addressed and mitigated or relaxed. This dissertation pushes the boundary of state-of-the-art research in the field of fault detection and diagnostics by answering the following four research questions:

How can we train a fault diagnostics model that is both, able to perform well on known and unknown domains as well as able to detect novel fault types? Chapter 4 enables fault diagnostics that is robust towards small domain shifts and, thus, provides a reliable solution for fault diagnostics in a real operational context by providing robustness towards domain shifts. Additionally, the module also enables the detection of novel fault types. The module can be considered as a stable and mature solution for condition assessment.

How can we enable universal domain adaptation for fault diagnostics models with extreme label space discrepancies and large domain gaps? In Chapter 5, the importance of enabling the transfer of knowledge between domains is addressed in the relevant case where only one class (the healthy one) is shared between the domains and the domain gap can be large. By enabling the generation of previously unobserved faults that are domain- and class-specific, we can transform a *Partial* or *Open-Partial* domain adaptation task into a *ClosedSet* one. Experiments have not only resulted in superior results under large domain gaps but also, the proposed method alleviates the requirements of real target data to tune the method.

How can effective fault diagnostics be enabled in the presence of label noise if no preliminary knowledge about the amount of label noise and no clean validation dataset is available? In Chapter 6, the harmful effect of label noise has been demonstrated. To relax the requirements of previous methods, a method was developed that can deal with only a rough estimation of the label noise. By identifying mislabeled data samples already in the gradient space, they can be disregarded before preforming the actual model update step. The experiments on the a image dataset (MNIST) and one condition monitoring dataset validated that a rough estimate of the label noise is sufficient to considerably improve the classification performance of the proposed method.

How to concurrently achieve invariance to non-informative factors and sensitivity to fault types for fault diagnostics but also for fault detection, where only healthy data and no fault data is available? In Chapter 6, contrastive learning was applied to fault diagnostics and also to fault detection on in-service assets within different railway systems. Namely, fault diagnostics for in-service railway sleepers based on image data and fault detection for in-service railway wheels based on timeseries data. Thus, we evaluated the proposed method under challenges that arise when working on real datasets recorded under in-service conditions as opposed to datasets recorded under laboratory conditions. For fault detection under in-service conditions, a multitude of factors might not be sufficiently represented in the training dataset that can cause variability in the healthy class. For the task of fault diagnostics under real in-service conditions, the fault pattern might not be as pronounced in the condition monitoring data and thus, it can be hard to distinguish early faults from normal variations in the healthy class. The superiority of the proposed method was demonstrated in both tasks. On the fault diagnostics task, a considerable performance gain could be achieved. On the fault detection task, the proposed method has achieved higher sensitivity to the detection of certain fault types. This lets us conclude that sensitivity to degradation can only translate to some fault types.

In summary, the four modules address specific challenges of data and label scarcity in different phases of monitoring the condition of an industrial asset. Each of the modules can be used and combined flexibly to meet the given requirements towards a condition assessment solution. Chapter 4 can be seen as a final adaption of the fault detection and diagnostics solution as it generalizes to novel domains without the need to retrain while maintaining sensitivity to novel faults to ensure safe operations; Chapter 5 then enables the transfer of models from one domain to other distinct domains and thus, is a pivotal step in extending the automated monitoring; Chapter 6 can effectively decrease the negative effect of label noise and enable the robust distinction between different fault types and severities; Chapter 7 enables more reliable fault detection and diagnostics in the context of real operational conditions.

9.3 Limitations and Outlook

Although this dissertation has extended the applicability of deep learning methods for data scarcity settings that are common in PHM tasks, there are still ways to extend the framework and address some potential remaining limitations. This section points out some of these remaining limitations, and discusses possible future directions.

Integration of Unlabeled Data: In this thesis, we have developed methods that can perform well under label noise as well as extreme data scarcity. We consider it to be realistic that in safety-critical systems, the condition of an asset is documented if it enters a workshop or if a fault has occurred. In less safety critical assets, this might not be the case. Instead, there might be a plethora of unlabeled data available that could be utilized for developing

methods. The integration of unlabeled data has not yet been considered in some of the developed modules and thus, provides an interesting and promising direction to explore.

Domain Generalization for Larger Domain Gaps: While we have tested the domain adaptation method in Chapter 5 on small domain gaps as well as on large ones, the suitability for the '*Domain Generalization module*' proposed in Chapter 4 still has to be tested in the context of large domain gaps. It needs to be evaluated to which extent the two competing objectives of achieving robustness towards domain shifts and sensitivity towards novel faults can be achieved even under larger domain gaps.

Regression Tasks: A natural extension of our work is to advance or apply the developed methods to regression tasks such as the prediction of the remaining useful life. While we did not approach the task of remaining useful life prediction in this thesis, several of the proposed approaches can be applied or adapted to this task. For example, the proposed idea on learning a contrastive feature representation with time as a proxy for the asset's health condition (presented in Chapter 7) lends itself to be tested in the context of regression tasks. Testing contrastive feature learning as proposed in Chapter 7 in the context of domain generalization (as in Chapter 4) for regression tasks would be a natural extension and an interesting future direction to explore.

Continuous Domain Shifts: In this thesis, discrete domain gaps were investigated for domain adaptation and domain generalization. While distinct domain shifts can be a good approximation for some applications, there exist other scenarios where, for example, environmental or operational conditions evolve continuously, causing continuous domain shifts. The methods developed in Chapter 7 for domain generalization or in Chapter 7 for domain adaptation can be applied and tested on continuous domain shifts without any methodological adaption.

Prior Knowledge for Domain Generalization: In recent works, the progress of domain generalization has been evaluated on image datasets (Wiles et al., 2021; Gulrajani and Lopez-Paz, 2020). One recommendation for future developments to achieve domain generalization is to integrate previous knowledge if possible (Wiles et al., 2021). While this has remained a rather vague recommendation, exploring the integration of prior knowledge in the context of PHM tasks presents a very interesting direction to explore. One possible way to integrate prior knowledge is proposed in Chapter 5, where fault signatures are learned that can be made transferable across domains. Exploring further ways of integrating knowledge e.g. by physics induced deep learning could hold much potential.

Generative Adversarial Networks Architecture Search for Fault Generation: The architecture of the generative adversarial model can affect the quality of the generated data immensely. Finding the correct architecture can be a tedious process. Architecture search algorithms for generative adversarial networks have already been applied in computer vision. Adapting these algorithms to condition monitoring datasets could enhance the performance of generative approaches for PHM applications.

Meta-Learning: Meta-learning studies aims to extract higher order knowledge (meta-knowledge) on a variety of tasks such that new tasks can be achieved more quickly or that existing models can be adapted to new environments with only very little data available. This paradigm provides an opportunity to tackle many conventional challenges of deep learning, including data and computation bottlenecks, as well as generalization. So far, we have solved

problems on the individual components' levels. Meta-learning could help the extraction of higher order knowledge on individual components of a system and enable the transfer of knowledge to either similar components that are operated in a new machines or in a drastically different environment, or to a new task that needs to fulfilled.

Bibliography

Abdeljaber, Osama, Onur Avci, Serkan Kiranyaz, Moncef Gabbouj, and Daniel J Inman (2017). "Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks". In: *Journal of Sound and Vibration* 388, pp. 154–170.

Abid, Anam, Muhammad Tahir Khan, and Javaid Iqbal (2021). "A review on fault detection and diagnosis techniques: basics and beyond". In: *Artificial Intelligence Review* 54.5, pp. 3639–3664.

Abid, Firas Ben, Marwen Sallem, and Ahmed Braham (2019). "Robust interpretable deep learning for intelligent fault diagnosis of induction motors". In: *IEEE Transactions on Instrumentation and Measurement* 69.6, pp. 3506–3515.

Alemi, Alireza, Francesco Corman, and Gabriel Lodewijks (2017). "Condition monitoring approaches for the detection of railway wheel defects". In: *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit* 231.8, pp. 961–981.

Algan, Görkem and Ilkay Ulusoy (2021). "Image classification with deep learning in the presence of noisy labels: A survey". In: *Knowledge-Based Systems* 215, p. 106771.

Ankerst, Mihael, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander (1999). "OPTICS: Ordering points to identify the clustering structure". In: *ACM Sigmod record* 28.2, pp. 49–60.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017). "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR, pp. 214–223.

Arpit, Devansh, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. (2017). "A closer look at memorization in deep networks". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 233–242.

Asplund, Matthias (2016). "Wayside condition monitoring system for railway wheel profiles: Applications and performance assessment". PhD thesis. Luleå tekniska universitet.

Baasch, Benjamin, Judith Heusel, Michael Roth, and Thorsten Neumann (2021). "Train wheel condition monitoring via cepstral analysis of axle box accelerations". In: *Applied Sciences* 11.4, p. 1432.

Bachman, Philip, R Devon Hjelm, and William Buchwalter (2019). "Learning representations by maximizing mutual information across views". In: *Advances in neural information processing systems* 32.

Balaji, Yogesh, Swami Sankaranarayanan, and Rama Chellappa (2018). "Metareg: Towards domain generalization using meta-regularization". In: *Advances in neural information processing systems* 31.

Bian, Jian, Yuantong Gu, and Martin Howard Murray (2013). "A dynamic wheel–rail impact analysis of railway track under wheel flat by finite element analysis". In: *Vehicle System Dynamics* 51.6, pp. 784–797.

Biggio, Luca and Iason Kastanis (2020). "Prognostics and health management of industrial assets: Current progress and road ahead". In: *Frontiers in Artificial Intelligence* 3, p. 578613.

Boris, Chidlovskii, Assem Sadek, and Christian Wolf (2021). "Universal Domain Adaptation in Ordinal Regression". In: *arXiv preprint arXiv:2106.11576*.

Bosso, Nicola, Antonio Gugliotta, and Nicolò Zampieri (2018). "Wheel flat detection algorithm for onboard diagnostic". In: *Measurement* 123, pp. 193–202.

Bousmalis, Konstantinos, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan (2017). "Unsupervised pixel-level domain adaptation with generative adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3722–3731.

Buitinck, Lars, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux (2013). "API design for machine learning software: experiences from the scikit-learn project". In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122.

Cao, Zhangjie, Lijia Ma, Mingsheng Long, and Jianmin Wang (2018). "Partial adversarial domain adaptation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 135–150.

Cao, Zhangjie, Kaichao You, Mingsheng Long, Jianmin Wang, and Qiang Yang (2019). "Learning to transfer examples for partial domain adaptation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2985–2994.

Chao, Manuel Arias, Bryan T Adey, and Olga Fink (2021). "Implicit supervision for fault detection and segmentation of emerging fault types with Deep Variational Autoencoders". In: *Neurocomputing* 454, pp. 324–338.

Chapelle, Olivier, Jason Weston, Léon Bottou, and Vladimir Vapnik (2001). "Vicinal risk minimization". In: *Advances in neural information processing systems*, pp. 416–422.

Chen, Pengfei, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang (2019). "Understanding and Utilizing Deep Neural Networks Trained with Noisy Labels". In: *International Conference on Machine Learning*, pp. 1062–1070.

Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton (2020). "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR, pp. 1597–1607.

Chen, Yuanhang, Gaoliang Peng, Chaohao Xie, Wei Zhang, Chuanhao Li, and Shaohui Liu (2018). "ACDIN: Bridging the gap between artificial and real bearing damages for bearing fault diagnosis". In: *Neurocomputing* 294, pp. 61–71.

Chien, Chen-Fu and Chia-Cheng Chen (2020). "Data-driven framework for tool health monitoring and maintenance strategy for smart manufacturing". In: *IEEE Transactions on Semiconductor Manufacturing* 33.4, pp. 644–652.

Chopra, Sumit, Raia Hadsell, and Yann LeCun (2005). "Learning a similarity metric discriminatively, with application to face verification". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. IEEE, pp. 539–546.

Cooley, James W and John W Tukey (1965). "An algorithm for the machine calculation of complex Fourier series". In: *Mathematics of computation* 19.90, pp. 297–301.

Creswell, Antonia, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath (2018). "Generative adversarial networks: An overview". In: *IEEE signal processing magazine* 35.1, pp. 53–65.

Cubillo, Adrian, Suresh Perinpanayagam, and Manuel Esperon-Miguez (2016). "A review of physics-based models in prognostics: Application to gears and bearings of rotating machinery". In: *Advances in Mechanical Engineering* 8.8, p. 1687814016664660.

Davari, Narjes, Bruno Veloso, Gustavo de Assis Costa, Pedro Mota Pereira, Rita P Ribeiro, and João Gama (2021). "A Survey on Data-Driven Predictive Maintenance for the Railway Industry". In: *Sensors* 21.17, p. 5739.

Deng, Minqiang, Aidong Deng, Yaowei Shi, Yang Liu, and Meng Xu (2022). "A novel sub-label learning mechanism for enhanced cross-domain fault diagnosis of rotating machinery". In: *Reliability Engineering & System Safety*, p. 108589.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ding, Yifei, Minping Jia, Jichao Zhuang, Yudong Cao, Xiaoli Zhao, and Chi-Guhn Lee (2023). "Deep imbalanced domain adaptation for transfer learning fault diagnosis of bearings under multiple working conditions". In: *Reliability Engineering & System Safety* 230, p. 108890.

Dosovitskiy, Alexey, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox (2014). "Discriminative unsupervised feature learning with convolutional neural networks". In: Citeseer.

Ducoffe, Mélanie, Ilyass Haloui, Jayant Sen Gupta, and ISAE Supaero (2019). "Anomaly Detection on Time Series with Wasserstein GAN applied to PHM". In: *International Journal of Prognostics and Health Management*.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *Kdd*. Vol. 96, pp. 226–231.

Farahani, Abolfazl, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia (2021). "A brief review of domain adaptation". In: *Advances in data science and information engineering*, pp. 877–894.

Fink, Olga (2020). "Data-driven intelligent predictive maintenance of industrial assets". In: *Women in Industrial and Systems Engineering*. Springer, pp. 589–605.

Fink, Olga, Qin Wang, Markus Svensén, Pierre Dersin, Wan-Jui Lee, and Melanie Ducoffe (2020). "Potential, challenges and future directions for deep learning in prognostics and health management applications". In: *Engineering Applications of Artificial Intelligence* 92, p. 103678.

Franceschi, Jean-Yves, Aymeric Dieuleveut, and Martin Jaggi (2019). "Unsupervised scalable representation learning for multivariate time series". In: *Advances in neural information processing systems* 32.

Frénay, Benoıt and Michel Verleysen (2013). "Classification in the presence of label noise: a survey". In: *IEEE transactions on neural networks and learning systems* 25.5, pp. 845–869.

Ghofrani, Faeze, Qing He, Rob MP Goverde, and Xiang Liu (2018). "Recent applications of big data analytics in railway transportation systems: A survey". In: *Transportation Research Part C: Emerging Technologies* 90, pp. 226–246.

Gomez, Raul, Lluis Gomez, Jaume Gibert, and Dimosthenis Karatzas (2018). "Learning to learn from web data through deep semantic embeddings". In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2020). "Generative adversarial networks". In: *Communications of the ACM* 63.11, pp. 139–144.

Goyal, Priya, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. (2021). "Self-supervised pretraining of visual features in the wild". In: *arXiv preprint arXiv:2103.01988*.

Granderson, Jessica, Guanjing Lin, Ari Harding, Piljae Im, and Yan Chen (2020). "Building fault detection data to aid diagnostic algorithm creation and performance testing". In: *Scientific Data* 7.1, pp. 1–14.

Guan, Yang, Zong Meng, Dengyun Sun, Jingbo Liu, and Fengjie Fan (2021). "2MNet: Multi-sensor and multi-scale model toward accurate fault diagnosis of rolling bearing". In: *Reliability Engineering & System Safety* 216, p. 108017.

Gui, Jie, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye (2021). "A review on generative adversarial networks: Algorithms, theory, and applications". In: *IEEE Transactions on Knowledge and Data Engineering*.

Gulrajani, Ishaan, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville (2017). "Improved training of wasserstein gans". In: *Advances in neural information processing systems* 30.

Gulrajani, Ishaan and David Lopez-Paz (2020). "In search of lost domain generalization". In: *arXiv preprint arXiv:2007.01434*.

Hacohen, Guy and Daphna Weinshall (2019). "On The Power of Curriculum Learning in Training Deep Networks". In: *International Conference on Machine Learning*, pp. 2535–2544.

Hadsell, Raia, Sumit Chopra, and Yann LeCun (2006). "Dimensionality reduction by learning an invariant mapping". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 2. IEEE, pp. 1735–1742.

Han, Te, Yan-Fu Li, and Min Qian (2021). "A hybrid generalization network for intelligent fault diagnosis of rotating machinery under unseen working conditions". In: *IEEE Transactions on Instrumentation and Measurement* 70, pp. 1–11.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Heideman, Michael, Don Johnson, and Charles Burrus (1984). "Gauss and the history of the fast Fourier transform". In: *IEEE ASSP Magazine* 1.4, pp. 14–21.

Hendriks, Jacob, Patrick Dumond, and DA Knox (2022). "Towards better benchmarking using the CWRU bearing fault dataset". In: *Mechanical Systems and Signal Processing* 169, p. 108732.

Hermans, Alexander, Lucas Beyer, and Bastian Leibe (2017). "In defense of the triplet loss for person re-identification". In: *arXiv preprint arXiv:1703.07737*.

Hoang, Duy-Tang and Hee-Jun Kang (2019). "A survey on deep learning based bearing fault diagnosis". In: *Neurocomputing* 335, pp. 327–335.

Hong, Weixiang, Zhenzhen Wang, Ming Yang, and Junsong Yuan (2018). "Conditional generative adversarial network for structured domain adaptation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1335–1344.

Hu, Hexuan, Bo Tang, Xuejiao Gong, Wei Wei, and Huihui Wang (2017). "Intelligent fault diagnosis of the high-speed train with big data based on deep neural networks". In: *IEEE Transactions on Industrial Informatics* 13.4, pp. 2106–2116.

Hu, W, Z Li, and D Yu (2020). "Simple and effective regularization methods for training on noisily labeled data with generalization guarantee". In: *International Conference on Learning Representations*.

Jaiswal, Ashish, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon (2021). "A survey on contrastive self-supervised learning". In: *Technologies* 9.1, p. 2.

Janeliukstis, Rims, Sandris Ručevskis, and Sakdirat Kaewunruen (2019). "Mode shape curvature squares method for crack detection in railway prestressed concrete sleepers". In: *Engineering Failure Analysis* 105, pp. 386–401.

Jardine, Andrew KS, Daming Lin, and Dragan Banjevic (2006). "A review on machinery diagnostics and prognostics implementing condition-based maintenance". In: *Mechanical systems and signal processing* 20.7, pp. 1483–1510.

Jiang, Lu, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei (2017). "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels". In: *CoRR, abs/1712.05055*.

Jiao, Jinyang, Ming Zhao, Jing Lin, and Chuancang Ding (2019). "Classifier inconsistency-based domain adaptation network for partial transfer intelligent diagnosis". In: *IEEE Transactions on Industrial Informatics* 16.9, pp. 5965–5974.

Jing, Guoqing, Mohammad Siahkouhi, J Riley Edwards, Marcus S Dersch, and NA Hoult (2021). "Smart railway sleepers-a review of recent developments, challenges, and future prospects". In: *Construction and Building Materials* 271, p. 121533.

Khan, Samir and Takehisa Yairi (2018). "A review on the application of deep learning in system health management". In: *Mechanical Systems and Signal Processing* 107, pp. 241–265.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kingma, Diederik P and Max Welling (2013). "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114*.

Krishna, Ranjay A, Kenji Hata, Stephanie Chen, Joshua Kravitz, David A Shamma, Li Fei-Fei, and Michael S Bernstein (2016). "Embracing error to enable rapid crowdsourcing". In: *Proceedings of the 2016 CHI conference on human factors in computing systems*, pp. 3167–3179.

Krummenacher, Gabriel, Cheng Soon Ong, Stefan Koller, Seijin Kobayashi, and Joachim M Buhmann (2017). "Wheel defect detection with machine learning". In: *IEEE Transactions on Intelligent Transportation Systems* 19.4, pp. 1176–1187.

LeCun, Yann, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel (1990). "Handwritten digit recognition with a back-propagation network". In: *Advances in neural information processing systems*, pp. 396–404.

Lee, Jinwook, Myungyon Kim, Jin Uk Ko, Joon Ha Jung, Kyung Ho Sun, and Byeng D Youn (2022). "Asymmetric inter-intra domain alignments (AIIDA) method for intelligent fault diagnosis of rotating machinery". In: *Reliability Engineering & System Safety* 218, p. 108186.

Lessmeier, Christian, James Kuria Kimotho, Detmar Zimmer, and Walter Sextro (2016). "Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification". In: *PHM Society European Conference*. Vol. 3. 1.

Li, Baojie, Claude Delpha, Demba Diallo, and A Migan-Dubois (2021a). "Application of Artificial Neural Networks to photovoltaic fault detection and diagnosis: A review". In: *Renewable and Sustainable Energy Reviews* 138, p. 110512.

Li, Chuan, Shaohui Zhang, Yi Qin, and Edgar Estupinan (2020a). "A systematic review of deep transfer learning for machinery fault diagnosis". In: *Neurocomputing* 407, pp. 121–135.

Li, Chunsheng, Shihui Luo, Colin Cole, and Maksym Spiryagin (2017). "An overview: modern techniques for railway vehicle on-board health monitoring systems". In: *Vehicle system dynamics* 55.7, pp. 1045–1070.

Li, Guangrui, Guoliang Kang, Yi Zhu, Yunchao Wei, and Yi Yang (2021b). "Domain consensus clustering for universal domain adaptation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9757–9766.

Li, Haoliang, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot (2018a). "Domain generalization with adversarial feature learning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5400–5409.

Li, Jie, Yu Wang, Yanyang Zi, Haijun Zhang, and Zhiguo Wan (2021c). "Causal Disentanglement: A Generalized Bearing Fault Diagnostic Framework in Continuous Degradation Mode". In: *IEEE Transactions on Neural Networks and Learning Systems*.

Li, Weihua, Ruyi Huang, Jipu Li, Yixiao Liao, Zhuyun Chen, Guolin He, Ruqiang Yan, and Konstantinos Gryllias (2022). "A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: Theories, applications and challenges". In: *Mechanical Systems and Signal Processing* 167, p. 108487.

Li, Xiang and Wei Zhang (2020). "Deep learning-based partial domain adaptation method on intelligent machinery fault diagnostics". In: *IEEE Transactions on Industrial Electronics* 68.5, pp. 4351–4361.

Li, Xiang, Wei Zhang, and Qian Ding (2018b). "Cross-domain fault diagnosis of rolling element bearings using deep generative neural networks". In: *IEEE Transactions on Industrial Electronics* 66.7, pp. 5525–5534.

Li, Xiang, Wei Zhang, Hui Ma, Zhong Luo, and Xu Li (2020b). "Deep learning-based adversarial multi-classifier optimization for cross-domain machinery fault diagnostics". In: *Journal of Manufacturing Systems* 55, pp. 334–347.

Li, Xiang, Wei Zhang, Hui Ma, Zhong Luo, and Xu Li (2020c). "Domain generalization in rotating machinery fault diagnostics using deep neural networks". In: *Neurocomputing* 403, pp. 409–420.

Li, Xiang, Wei Zhang, Hui Ma, Zhong Luo, and Xu Li (2020d). "Partial transfer learning in machinery cross-domain fault diagnostics using class-weighted adversarial networks". In: *Neural Networks* 129, pp. 313–322.

Liang, Jian, Yunbo Wang, Dapeng Hu, Ran He, and Jiashi Feng (2020). "A balanced and uncertainty-aware approach for partial domain adaptation". In: *European Conference on Computer Vision*. Springer, pp. 123–140.

Liao, Yixiao, Ruyi Huang, Jipu Li, Zhuyun Chen, and Weihua Li (2020). "Deep semisupervised domain generalization network for rotary machinery fault diagnosis under variable speed". In: *IEEE Transactions on Instrumentation and Measurement* 69.10, pp. 8064–8075.

Lloyd, Stuart (1982). "Least squares quantization in PCM". In: *IEEE transactions on information theory* 28.2, pp. 129–137.

Locatello, Francesco, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem (2019). "Challenging common assumptions in the unsupervised learning of disentangled representations". In: *international conference on machine learning*. PMLR, pp. 4114–4124.

Lu, Chen, Zhen-Ya Wang, Wei-Li Qin, and Jian Ma (2017). "Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification". In: *Signal Processing* 130, pp. 377–388.

Lu, Weining, Bin Liang, Yu Cheng, Deshan Meng, Jun Yang, and Tao Zhang (2016). "Deep model based domain adaptation for fault diagnosis". In: *IEEE Transactions on Industrial Electronics* 64.3, pp. 2296–2305.

Luo, Jia, Jinying Huang, and Hongmei Li (2021). "A case study of conditional deep convolutional generative adversarial networks in machine fault diagnosis". In: *Journal of Intelligent Manufacturing* 32.2, pp. 407–425.

Luo, Jianhui, Madhavi Namburu, Krishna Pattipati, Liu Qiao, Masayuki Kawamoto, and SACS Chigusa (2003). "Model-based prognostic techniques [maintenance applications]". In: *Proceedings AUTOTESTCON 2003. IEEE Systems Readiness Technology Conference.* Ieee, pp. 330–340.

Ma, Xingjun, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah Erfani, Shutao Xia, Sudanthi Wijewickrema, and James Bailey (2018). "Dimensionality-Driven Learning with Noisy Labels". In: *International Conference on Machine Learning*, pp. 3355–3364.

MacQueen, James (1967). "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. Oakland, CA, USA, pp. 281–297.

McKinnon, Conor, James Carroll, Alasdair McDonald, Sofia Koukoura, David Infield, and Conaill Soraghan (2020). "Comparison of new anomaly detection technique for wind turbine condition monitoring using gearbox SCADA data". In: *Energies* 13.19, p. 5152.

Menke, Maximilian, Thomas Wenzel, and Andreas Schwung (2022). "Improving gan-based domain adaptation for object detection". In: *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, pp. 3880–3885.

Miao, Mengqi, Jianbo Yu, and Zhihong Zhao (2022). "A sparse domain adaption network for remaining useful life prediction of rolling bearings under different working conditions". In: *Reliability Engineering & System Safety* 219, p. 108259.

Michau, Gabriel and Olga Fink (2019). "Domain Adaptation for One-Class Classification: Monitoring the Health of Critical Systems Under Limited Information". In: *International Journal of Prognostics and Health Management* 10.4.

Michau, Gabriel and Olga Fink (2021). "Unsupervised transfer learning for anomaly detection: Application to complementary operating condition transfer". In: *Knowledge-Based Systems* 216, p. 106816.

Michau, Gabriel, Yang Hu, Thomas Palmé, and Olga Fink (2020). "Feature learning for fault detection in high-dimensional condition monitoring signals". In: *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability* 234.1, pp. 104–115.

Mirza, Mehdi and Simon Osindero (2014). "Conditional generative adversarial nets". In: *arXiv preprint arXiv:1411.1784*.

Mori, Hirotaka, Hitoshi Tsunashima, Takashi Kojima, Akira Matsumoto, and Takeshi Mizuma (2010). "Condition monitoring of railway track using in-service vehicle". In: *Journal of Mechanical Systems for Transportation and Logistics* 3.1, pp. 154–165.

Mosleh, Araliya, Pedro Montenegro, Pedro Alves Costa, and Rui Calçada (2021). "An approach for wheel flat detection of railway train wheels using envelope spectrum analysis". In: *Structure and Infrastructure Engineering* 17.12, pp. 1710–1729.

Müller, Nicolas M and Karla Markert (2019). "Identifying Mislabeled Instances in Classification Datasets". In: *2019 International Joint Conference on Neural Networks*. IEEE, pp. 1–8.

Nejjar, Ismail, Jean Meunier-Pion, Gaetan Frusque, Olga Fink, and Gif-sur-Yvette CentraleSupélec (2022). "DG-MIX: Domain generalization for anomolous sound detection based on self-supervised learning". In: *Detection and Classification of Acoustic Scenes and Events 2022*.

Neupane, Dhiraj and Jongwon Seok (2020). "Bearing Fault Detection and Diagnosis Using Case Western Reserve University Dataset With Deep Learning Approaches: A Review". In: *IEEE Access* 8, pp. 93155–93178.

Nguyen, Thi Phuong Khanh, Amor Khlaief, Kamal Medjaher, Antoine Picot, Pascal Maussion, Diego Tobon, Bertrand Chauchat, and Regis Cheron (2018). "Analysis and comparison of multiple features for fault detection and prognostic in ball bearings". In: *Fourth european conference of the prognostics and health management society 2018*, pp. 1–9.

Ni, Yi-Qing and Qiu-Hu Zhang (2021). "A Bayesian machine learning approach for online detection of railway wheel defects using track-side monitoring". In: *Structural Health Monitoring* 20.4, pp. 1536–1550.

Oord, Aaron van den, Yazhe Li, and Oriol Vinyals (2018). "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748*.

Pan, Sinno Jialin and Qiang Yang (2009). "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10, pp. 1345–1359.

Pandhare, Vibhor, Jaskaran Singh, and Jay Lee (2019). "Convolutional neural network based rolling-element bearing fault diagnosis for naturally occurring and progressing defects using time-frequency domain features". In: *2019 Prognostics and System Health Management Conference (PHM-Paris)*. IEEE, pp. 320–326.

Pang, Yong, Siva N Lingamanaik, Bernard K Chen, and Siu Fung Yu (2020). "Measurement of deformation of the concrete sleepers under different support conditions using non-contact laser speckle imaging sensor". In: *Engineering Structures* 205, p. 110054.

Patel, Raj Kumar and VK Giri (2016). "Feature selection and classification of mechanical fault of an induction motor using random forest classifier". In: *Perspectives in Science* 8, pp. 334–337.

Patrini, Giorgio, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu (2017). "Making deep neural networks robust to label noise: A loss correction approach". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1944–1952.

Pleiss, Geoff, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger (2020). "Identifying mislabeled data using the area under the margin ranking". In: *Advances in Neural Information Processing Systems* 33, pp. 17044–17056.

Ragab, Mohamed, Zhenghua Chen, Wenyu Zhang, Emadeldeen Eldele, Min Wu, Chee-Keong Kwoh, and Xiaoli Li (2022). "Conditional Contrastive Domain Generalization for Fault Diagnosis". In: *IEEE Transactions on Instrumentation and Measurement* 71, pp. 1–12.

Ramesh, Aditya, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever (2021). "Zero-shot text-to-image generation". In: *International Conference on Machine Learning*. PMLR, pp. 8821–8831.

Reed, Scott, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich (2014). "Training deep neural networks on noisy labels with bootstrapping". In: *CoRR, abs/1412.6596,*

Ren, Mengye, Wenyuan Zeng, Bin Yang, and Raquel Urtasun (2018). "Learning to reweight examples for robust deep learning". In: *International Conference on Machine Learning*.

Rombach, Katharina, Gabriel Michau, and Olga Fink (2020). "Improving generalization of deep fault detection models in the presence of mislabeled data". In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, pp. 3103–3110. DOI: https://doi.org/10.1109/SMC42975.2020.9283002.

Rombach, Katharina, Gabriel Michau, and Olga Fink (2021). "Contrastive Learning for Fault Detection and Diagnostics in the Context of Changing Operating Conditions and Novel Fault Types". In: *Sensors* 21.10, p. 3550. DOI: https://doi.org/10.3390/s21103550.

Rombach, Katharina, Gabriel Michau, and Olga Fink (2023). "Controlled generation of unseen faults for Partial and Open-Partial domain adaptation". In: *Reliability Engineering & System Safety* 230, p. 108857. DOI: https://doi.org/10.1016/j.ress.2022.108857.

Rombach, Katharina, Gabriel Michau, Kajan Ratnasabapathy, Lucian-Stefan Ancu, Wilfried Bürzle, Stefan Koller, and Olga Fink (2022). "Contrastive feature learning for railway infrastructure fault diagnostic". In: *32nd European Safety and Reliability Conference (ESREL 2022)*.

Rosenberg, Andrew and Julia Hirschberg (2007). "V-measure: A conditional entropy-based external cluster evaluation measure". In: *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 410–420.

Rousseeuw, Peter J (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20, pp. 53–65.

Rui, Liu, Emanuele Zappa, and Andrea Collina (2020). "Vision-based measurement of crack generation and evolution during static testing of concrete sleepers". In: *Engineering Fracture Mechanics* 224, p. 106715.

Saito, Kuniaki and Kate Saenko (2021). "Ovanet: One-vs-all network for universal domain adaptation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9000–9009.

Saunshi, Nikunj, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar (2019). "A theoretical analysis of contrastive unsupervised representation learning". In: *International Conference on Machine Learning*. PMLR, pp. 5628–5637.

Saxe, Andrew M, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox (2019). "On the information bottleneck theory of deep learning". In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.12, p. 124020.

Schott, Lukas, Julius Von Kügelgen, Frederik Träuble, Peter Gehler, Chris Russell, Matthias Bethge, Bernhard Schölkopf, Francesco Locatello, and Wieland Brendel (2021). "Visual representation learning does not generalize strongly within the same domain". In: *International Conference on Learning Representations*.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin (2015). "Facenet: A unified embedding for face recognition and clustering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823.

Serradilla, Oscar, Ekhi Zugasti, Jon Rodriguez, and Urko Zurutuza (2022). "Deep learning models for predictive maintenance: a survey, comparison, challenges and prospects". In: *Applied Intelligence*, pp. 1–31.

Shankar, Shiv, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi (2018). "Generalizing across domains via cross-gradient training". In: *International Conference on Learning Representations*.

Shen, Changqing, Yumei Qi, Jun Wang, Gaigai Cai, and Zhongkui Zhu (2018). "An automatic and robust features learning method for rotating machinery fault diagnosis based on contractive autoencoder". In: *Engineering Applications of Artificial Intelligence* 76, pp. 170–184.

Shen, Yanyao and Sujay Sanghavi (2018). "Learning with bad training data via iterative trimmed loss minimization". In: *International Conference on Machine Learning*.

Shenfield, Alex and Martin Howarth (2020). "A novel deep learning model for the detection and identification of rolling element-bearing faults". In: *Sensors* 20.18, p. 5112.

Shi, Yongjie, Xianghua Ying, and Jinfa Yang (2022). "Deep unsupervised domain adaptation with time series sensor data: A survey". In: *Sensors* 22.15, p. 5507.

Shwartz-Ziv, Ravid and Naftali Tishby (2017). "Opening the black box of deep neural networks via information". In: *arXiv preprint arXiv:1703.00810*.

Siahpour, Shahin, Xiang Li, and Jay Lee (2020). "Deep learning-based cross-sensor domain adaptation for fault diagnosis of electro-mechanical actuators". In: *International Journal of Dynamics and Control* 8.4, pp. 1054–1062.

Smith, Wade A and Robert B Randall (2015). "Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study". In: *Mechanical Systems and Signal Processing* 64, pp. 100–131.

Song, Ying, Zhichen Wang, and Yingming Shen (2013). "Research on the Control of Out-of-Round Wheel Profiles of High-Speed Railway Derived from Numerical Simulations." In: *International Journal of Online Engineering* 9.

Sukhbaatar, Sainbayar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus (2015). "Training convolutional networks with noisy labels". In: *Proc. Int. Conf. Learn. Represent. Workshop, 2015*, pp. 1–11.

Tan, Chuanqi, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu (2018). "A survey on deep transfer learning". In: *International conference on artificial neural networks*. Springer, pp. 270–279.

Tanaka, Daiki, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa (2018). "Joint optimization framework for learning with noisy labels". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5552–5560.

Tang, Ruifan, Lorenzo De Donato, Nikola Besinović, Francesco Flammini, Rob MP Goverde, Zhiyuan Lin, Ronghui Liu, Tianli Tang, Valeria Vittorini, and Ziyulong Wang (2022). "A literature review of Artificial Intelligence applications in railway systems". In: *Transportation Research Part C: Emerging Technologies* 140, p. 103679.

Tatarinov, Alexey, Aleksandrs Rumjancevs, and Viktors Mironovs (2019). "Assessment of cracks in pre-stressed concrete railway sleepers by ultrasonic testing". In: *Procedia Computer Science* 149, pp. 324–330.

Torrey, Lisa and Jude Shavlik (2010). "Transfer learning". In: *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, pp. 242–264.

Vahdat, Arash (2017). "Toward robustness against label noise in training deep discriminative neural networks". In: *Advances in Neural Information Processing Systems*, pp. 5596–5605.

Van der Maaten, Laurens and Geoffrey Hinton (2008). "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11.

Vapnik, Vladimir (1998). "Statistical learning theory." In: *Wiley. Wang, K., Tsung, F.(2007). Run-to-run Process Adjust. using Categ. Obs. J. Qual. Technol.* 39.4, p. 312.

Vapnik, Vladimir (2013). *The nature of statistical learning theory*. Springer science & business media.

Vinh, Nguyen Xuan, Julien Epps, and James Bailey (2010). "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance". In: *The Journal of Machine Learning Research* 11, pp. 2837–2854.

Vrignat, Pascal, Frédéric Kratz, and Manuel Avila (2022). "Sustainable manufacturing, maintenance policies, prognostics and health management: A literature review". In: *Reliability Engineering & System Safety* 218, p. 108140.

Wang, Jindong, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu (2022). "Generalizing to unseen domains: A survey on domain generalization". In: *IEEE Transactions on Knowledge and Data Engineering*.

Wang, Mei and Weihong Deng (2018). "Deep visual domain adaptation: A survey". In: *Neurocomputing* 312, pp. 135–153.

Wang, Qin, Gabriel Michau, and Olga Fink (2019). "Domain adaptive transfer learning for fault diagnosis". In: *2019 Prognostics and System Health Management Conference (PHM-Paris)*. IEEE, pp. 279–285.

Wang, Qin, Gabriel Michau, and Olga Fink (2020a). "Missing-class-robust domain adaptation by unilateral alignment". In: *IEEE Transactions on Industrial Electronics* 68.1, pp. 663–671.

Wang, Qin, Cees Taal, and Olga Fink (2021). "Integrating Expert Knowledge with Domain Adaptation for Unsupervised Fault Diagnosis". In: *IEEE Transactions on Instrumentation and Measurement*.

Wang, Xiaodong and Feng Liu (2020). "Triplet loss guided adversarial domain adaptation for bearing fault diagnosis". In: *Sensors* 20.1, p. 320.

Wang, YW, YQ Ni, and X Wang (2020b). "Real-time defect detection of high-speed train wheels by using Bayesian forecasting and dynamic model". In: *Mechanical Systems and Signal Processing* 139, p. 106654.

Weinberger, Kilian Q and Lawrence K Saul (2009). "Distance metric learning for large margin nearest neighbor classification." In: *Journal of machine learning research* 10.2.

Widodo, Achmad and Bo-Suk Yang (2007). "Support vector machine in machine condition monitoring and fault diagnosis". In: *Mechanical systems and signal processing* 21.6, pp. 2560–2574.

Wiles, Olivia, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Taylan Cemgil (2021). "A fine-grained analysis on distribution shift". In: *International Conference on Learning Representations*.

Williams, John H, Alan Davies, and Paul R Drake (1994). *Condition-based maintenance and machine diagnostics*. Springer Science & Business Media.

Wilson, Garrett and Diane J Cook (2020). "A survey of unsupervised deep domain adaptation". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 11.5, pp. 1–46.

Xie, Jiawei, Jinsong Huang, Cheng Zeng, Shui-Hua Jiang, and Nathan Podlich (2020). "Systematic literature review on data-driven models for predictive maintenance of railway track: Implications in geotechnical engineering". In: *Geosciences* 10.11, p. 425.

Xu, Yilun, Peng Cao, Yuqing Kong, and Yizhou Wang (2019). "$\mathcal{L}_{DMI}$: A Novel Information-theoretic Loss Function for Training Deep Nets Robust to Label Noise". In: *Advances in Neural Information Processing Systems*, pp. 6222–6233.

Yacob, Syamilah, Azlan Shah Ali, and Cheong Peng Au-Yong (2022). "Building Condition Monitoring and Assessment". In: *Managing Building Deterioration*, pp. 65–93.

Yamashita, Hiroki, Christofer Feldmeier, Yosuke Yamazaki, Takanori Kato, Takahiro Fujimoto, Osamu Kondo, and Hiroyuki Sugiyama (2022). "Wheel profile optimization procedure to minimize flange wear considering profile wear evolution". In: *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit* 236.6, pp. 672–683.

Yang, Yanchao and Stefano Soatto (2020). "Fda: Fourier domain adaptation for semantic segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4085–4095.

Yao, Siya, Qi Kang, MengChu Zhou, Muhyaddin J Rawa, and Abdullah Abusorrah (2022). "A survey of transfer learning for machinery diagnostics and prognostics". In: *Artificial Intelligence Review*, pp. 1–52.

Yoon, Andre S, Taehoon Lee, Yongsub Lim, Deokwoo Jung, Philgyun Kang, Dongwon Kim, Keuntae Park, and Yongjin Choi (2017). "Semi-supervised learning with deep generative models for asset failure prediction". In: *arXiv preprint arXiv:1709.00845*.

Yu, Leijian, Erfu Yang, Peng Ren, Cai Luo, Gordon Dobie, Dongbing Gu, and Xiutian Yan (2019). "Inspection robots in oil and gas industry: a review of current solutions and future trends". In: *2019 25th International Conference on Automation and Computing (ICAC)*. IEEE, pp. 1–6.

Zareapoor, Masoumeh, Pourya Shamsolmoali, and Jie Yang (2021). "Oversampling adversarial network for class-imbalanced fault diagnosis". In: *Mechanical Systems and Signal Processing* 149, p. 107175.

Zhang, Bo, Wei Li, Jie Hao, Xiao-Li Li, and Meng Zhang (2018). "Adversarial adaptive 1-D convolutional neural networks for bearing fault diagnosis under varying working condition". In: *arXiv preprint arXiv:1805.00778*.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals (2017a). "Understanding deep learning requires rethinking generalization". In: *International Conference on Learning Representations*.

Zhang, Hongyi, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz (2017b). "mixup: Beyond empirical risk minimization". In: *International Conference on Learning Representations*.

Zhang, Qiyang, Zhibin Zhao, Xingwu Zhang, Yilong Liu, Chuang Sun, Ming Li, Shibin Wang, and Xuefeng Chen (2021a). "Conditional adversarial domain generalization with a single discriminator for bearing fault diagnosis". In: *IEEE Transactions on Instrumentation and Measurement* 70, pp. 1–15.

Zhang, Shen, Shibo Zhang, Bingnan Wang, and Thomas G Habetler (2020). "Deep learning algorithms for bearing fault diagnostics—A comprehensive review". In: *IEEE Access* 8, pp. 29857–29881.

Zhang, Wei, Xiang Li, Hui Ma, Zhong Luo, and Xu Li (2021b). "Open-set domain adaptation in machinery fault diagnostics using instance-level weighted adversarial learning". In: *IEEE Transactions on Industrial Informatics* 17.11, pp. 7445–7455.

Zhang, Wei, Xiang Li, Hui Ma, Zhong Luo, and Xu Li (2021c). "Universal domain adaptation in fault diagnostics with hybrid weighted deep adversarial learning". In: *IEEE Transactions on Industrial Informatics* 17.12, pp. 7957–7967.

Zhang, Weiting, Dong Yang, and Hongchao Wang (2019). "Data-driven methods for predictive maintenance of industrial equipment: A survey". In: *IEEE Systems Journal* 13.3, pp. 2213–2227.

Zhang, Wen, Lingfei Deng, Lei Zhang, and Dongrui Wu (2022). "A survey on negative transfer". In: *IEEE/CAA Journal of Automatica Sinica*.

Zhang, Zhilu and Mert Sabuncu (2018). "Generalized cross entropy loss for training deep neural networks with noisy labels". In: *Advances in neural information processing systems*, pp. 8778–8788.

Zhao, Chao, Guokai Liu, and Weiming Shen (2022). "A balanced and weighted alignment network for partial transfer fault diagnosis". In: *ISA transactions*.

Zhao, Chao and Weiming Shen (2022a). "A domain generalization network combing invariance and specificity towards real-time intelligent fault diagnosis". In: *Mechanical Systems and Signal Processing* 173, p. 108990.

Zhao, Chao and Weiming Shen (2022b). "Adaptive Open Set Domain Generalization Network: Learning to Diagnose Unknown Faults under Unknown Working Conditions". In: *Reliability Engineering & System Safety*, p. 108672.

Zhao, Chao and Weiming Shen (2022c). "Dual adversarial network for cross-domain open set fault diagnosis". In: *Reliability Engineering & System Safety* 221, p. 108358.

Zhao, Hongshan, Huihai Liu, Wenjing Hu, and Xihui Yan (2018). "Anomaly detection and fault analysis of wind turbine components based on deep learning network". In: *Renewable energy* 127, pp. 825–834.

Zhao, Zhibin, Tianfu Li, Jingyao Wu, Chuang Sun, Shibin Wang, Ruqiang Yan, and Xuefeng Chen (2020). "Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study". In: *ISA transactions* 107, pp. 224–255.

Zheng, Huailiang, Rixin Wang, Yuantao Yang, Yuqing Li, and Minqiang Xu (2019). "Intelligent fault identification based on multisource domain generalization towards actual diagnosis scenario". In: *IEEE Transactions on Industrial Electronics* 67.2, pp. 1293–1304.

Zhong, Guoqiang, Li-Na Wang, Xiao Ling, and Junyu Dong (2016). "An overview on data representation learning: From traditional feature learning to recent deep learning". In: *The Journal of Finance and Data Science* 2.4, pp. 265–278.

Zhou, Huafeng, Peiyuan Cheng, Siyu Shao, Yuwei Zhao, and Xinyu Yang (2022a). "Bearing fault diagnosis based on partial domain adaptation adversarial network". In: *Measurement Science and Technology* 33.12, p. 124003.

Zhou, Kaiyang, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy (2022b). "Domain Generalization: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20. DOI: 10.1109/TPAMI.2022.3195549.

Zhou, Kaiyang, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy (2022c). "Domain generalization: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, Kaiyang, Yongxin Yang, Timothy Hospedales, and Tao Xiang (2020a). "Deep domain-adversarial image generation for domain generalisation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07, pp. 13025–13032.

Zhou, Kaiyang, Yongxin Yang, Timothy Hospedales, and Tao Xiang (2020b). "Learning to generate novel domains for domain generalization". In: *European Conference on Computer Vision*. Springer, pp. 561–578.

Zhou, Taotao, Te Han, and Enrique Lopez Droguett (2022d). "Towards trustworthy machine fault diagnosis: A probabilistic Bayesian deep learning framework". In: *Reliability Engineering & System Safety* 224, p. 108525.

Zhuang, Fuzhen, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He (2020). "A comprehensive survey on transfer learning". In: *Proceedings of the IEEE* 109.1, pp. 43–76.