Technical Language Supervision for Intelligent Fault Diagnosis in Process Industry

Karl Löwenmark, Cees Taal, Stephan Schnabel, Marcus Liwicki, and Fredrik Sandin

Abstract

In the process industry, condition monitoring systems with automated fault diagnosis methods assist human experts and thereby improve maintenance efficiency, process sustainability, and workplace safety. Improving the automated fault diagnosis methods using data and machine learning-based models is a central aspect of intelligent fault diagnosis (IFD). A major challenge in IFD is to develop realistic datasets with accurate labels needed to train and validate models, and to transfer models trained with labeled lab data to heterogeneous process industry environments. However, fault descriptions and work-orders written by domain experts are increasingly digitized in modern condition monitoring systems, for example in the context of rotating equipment monitoring. Thus, domain-specific knowledge about fault characteristics and severities exists as technical language annotations in industrial datasets. Furthermore, recent advances in natural language processing enable weakly supervised model optimization using natural language annotations, most notably in the form of natural language supervision (NLS). This creates a timely opportunity to develop technical language supervision (TLS) solutions for IFD systems grounded in industrial data, for example as a complement to pre-training with lab data to address problems like overfitting and inaccurate out-of-sample generalisation. We surveyed the literature and identify a considerable improvement in the maturity of NLS over the last two years, facilitating applications beyond natural language; a rapid development of weak supervision methods; and transfer learning as a current trend in IFD which can benefit from these developments. Finally, we describe a framework for integration of TLS in IFD which is inspired by recent NLS innovations.

Keywords: Intelligent Fault Diagnosis, Fault Severity Estimation, Condition Monitoring, Natural Language Processing, Natural Language Supervision, Technical Language Supervision, Weak Supervision

1. Introduction

Condition-monitoring (CM) based fault diagnosis of rotating machinery is widely used in industry to optimize equipment availability, uniformity of product characteristics and safety in the work environment, and to minimize production losses and material waste. In process industry, this typically requires human expert analysts with years of training and detailed knowledge about the operational states, functional roles and contexts of the machines being monitored. Due to growing demands on production efficiency and the vast amounts of data consequently generated in modern CM systems, automated fault diagnosis systems are required to assist human analysis through alarms and policy recommendation. Important tasks for the automated system are fault detection and classification to generate alarms and filter data, and fault severity estimation to predict remaining useful life and recommend policy options. Existing automated systems are mainly based on expert systems, with a knowledge-base derived from physical properties of analysed components, and a rule-based inference engine with local thresholds set by experts. In the case of

*Corresponding author: karl.ekstrom@ltu.se

1Current address: Bräuhausgasse 33, 84130 Dingolfing Tel. +49 (0)871 – 506 8342.
vibration measurements of rotating machinery, signal processing and kinematics based condition indicators are commonly used as knowledge-bases [6, 7, 8]. Intelligent fault diagnosis (IFD) [9] has been proposed to enhance the automated systems by inferring fault characteristics directly from process or lab data through learning based methods. Improving existing models is vital to meet the increasing demands on CM systems to improve production and equipment life cycle efficiency in process industry [10, 11], and the machine CM market is estimated at $2.6 billion with a compound annual growth rate estimation of 7.1%. For example, improved IFD algorithms can contribute to: reducing the number of unnecessary interventions; facilitating remanufacturing of components [12]; optimising maintenance schedules; and enabling analysts to focus on qualified preventive tasks.

However, it is difficult to develop realistic datasets with accurate labels needed to train and validate IFD models, and such data are expected to generalize poorly between process-industry plants due to their heterogeneous nature. Recent innovations in natural language processing offer a timely opportunity to address this challenge with methods used in natural language supervision (NLS) using digitalised technical language fault descriptions and work-orders available in many process industry datasets. Processing technical language poses unique challenges different from natural language, promoting the need for Technical Language Processing (TLP) and a technical version of NLS in Technical Language Supervision (TLS). Therefore we survey the state of the art in IFD, NLS and TLP, and discuss how TLS can be applied to IFD in a process industry context.

1.1. Background

Fault Diagnosis (FD) deals with the mapping of measured signal features to component conditions. The most basic condition is whether a fault is present or not, but more complex estimations such as fault class, fault severity, remaining useful life (RUL) and root cause analysis (RCA) can also be required. Table 1 describes these five major subtasks of FD, ordered in rising complexity based on interviews with condition monitoring experts from process industry. Fault detection and classification are tasks that are frequently automated in process industry through signal processing [3, 4, 5], and e.g. model-based thresholding. Fault severity estimation, a vital tool in maintenance decisions, is next in line to be automated, but is challenging due to nonlinear relationships between signal features and fault evolution [13]. RUL depends on the evolution of estimated fault severity over time, and predicts the remaining time until a fault is so severe that a component is no longer useful [14, 15]. RCA is a complex task that may be challenging to automate, but will indirectly be improved if simpler tasks are automated and human experts can invest more time in preventive policies.

The upper part of Figure 1 illustrates an example of a typical FD system (labeled Pipeline 1) implemented in process industry, see for instance [5, 16, 17]. The system requires no fault history data to learn from, but requires process information and kinematic models for the extraction of condition indicators [18]. Faults are detected and classified using signal processing [2], for instance root mean square, peak-to-peak and time synchronous average in the time domain [19]; spectral density, enveloping and Hilbert transform in the frequency domain; and dictionaries, wavelets and the Wigner-Ville distribution in the time-frequency domain; as well as kinematics based condition indicators, for instance the frequency intensity in the ball pass frequency of the outer race in ballpoint bearings. The decomposed signal is then analysed with typically simple rules based on indicator magnitude defined by experienced analysts. Once a fault is detected by the

[Table 1: Fault Diagnosis Tasks]

| Task       | Question addressed | Output      | Added value                  | Automated? |
|------------|--------------------|-------------|------------------------------|------------|
| Detection  | Is there a fault present? | Yes/No      | Alert human analysis        | Yes        |
| Classification | What type of fault? | Class       | Guide human analysis        | Partially  |
| Severity   | How severe is the fault? | Magnitude   | Motivate maintenance        | No         |
| RUL        | Time until maintenance is needed? | Risk vs Time | Maintenance planning        | No         |
| Root Cause | What caused the fault? | Description | Preventive policies         | No         |

[https://www.marketsandmarkets.com/Market-Reports/machine-health-monitoring-market-29627363.html]
model, a human analyst is alerted for in-depth diagnosis. The analyst decides whether to further investigate alarms or not, describes eventual faults in the form of natural-language annotations and makes work orders. Thus, the automated FD model acts like a filter between the massive amount of sensor data that is constantly generated, and the accurate but resource-constrained analysis of human experts. Based on cases from two industry collaborations with major process industry actors in Northern Sweden, analysts monitor around 5000 alarms per analyst per year, after filtering, where at most 20% of generated alarms point to component faults and the rest are due to temporary or constant signal malfunctions.

With improved automated FD, analysts could focus on more advanced fault diagnosis tasks beyond the current capabilities of IFD. Considerable research has been invested in automated FD, and many learning-based methods have shown promising results on test datasets [20, 21, 22]. However, the accurate deep learning models used in many IFD publications require vast amounts of training data in the form of labelled datasets, sets that typically do not exist in process industry cases [23]. Instead, training and test datasets are created in lab environments with artificial or accelerated fault development, such as the Case Western Reserve University bearing dataset [24], the Intelligent Maintenance System (IMS) by the University of Cincinnati dataset [25], and the Machinery Failure Prevention Technology (MFPT) dataset [26], but typically generalize poorly to heterogeneous environments [27, 28] such as process industries. Thus, despite the maturity of IFD methods in terms of literature, supervised IFD lacks wide-spread implementation in industry.

Industry datasets suitable for IFD can in some cases potentially be created, but it is difficult and costly to define high-quality labels that are accurately connected to relevant data. Therefore, transfer learning [29], illustrated in Pipeline 2 in Figure 1, has become an increasingly popular approach to develop IFD methods without requiring a large labelled dataset in the target domain [9]. Ideally, a model could be developed/trained with data from a lab environment, then transferred to similar components in an industrial environment. However, this remains a challenging goal due to differences between developing faults, heterogeneous environments, varying sensor and signal-to-noise conditions, and complex coupling of signal components. Thus, a method for the extraction of labels for industry data would be valuable and can facilitate implementations of current IFD models, as well as transfer learning by providing access to labels in the

![Figure 1: An overview of a typical process industry fault diagnosis pipeline (1), possible transfer learning IFD pipeline additions (2), and our suggested natural language supervision pipeline (3). Both (2) and (3) can provide considerable contributions to (1), with the strongest contributions coming from both pipelines implemented in symbiosis.](image)
target domain.

While labels are lacking in realistic CM datasets, technical language fault descriptions are written by analysts when documenting and monitoring the development of for example bearing faults over long periods of time (several months). Thus, the text-annotations produced as outputs of Pipeline 1 in Figure 1 contain valuable albeit noisy information about fault development characteristics and severities. This motivates the question, can such domain-specific annotations and related knowledge be used for training and finetuning of IFD methods as a substitute for regular labels?

Language has been used to train machine learning models for image recognition and object detection through recent breakthroughs in natural language supervision [30, 31]. Can a similar approach be used to train IFD models on industry data using annotations and work orders as zero-shot labels?

1.2. Contribution

We propose the usage of TLS on technical language fault descriptions to overcome the lack of labels in industry CM datasets. TLS is grounded in three fields, IFD, TLP and NLS, and we briefly survey all three to motivate the purpose and benefits of TLS. Potential contributions from implementing natural language supervision are improved support for human analysts and automation of simpler tasks by augmenting the label domain for transfer learning and weakly natural language supervised learning.

Pipeline 3 in Figure 1 illustrates the concept of a technical language supervision framework for process industry data. Unlabelled CM sensor data and process data are used to extract features through methods already used in IFD models, and the features are mapped to annotation-embeddings using natural language supervision and processing. Upon implementation, an unannotated time-series would thereby be mapped to the closest natural language descriptions in the embedding space, and through a query function, the fault class and severity can be estimated and described. Besides alarms and work orders, a language based model could also produce new annotations and natural language descriptions of detected faults for supervision, decision support and analyst training purposes.

Figure 2: Trends of publications between 1967 and 2020, obtained through Scopus queries looking for publications with the targeted keywords in the article title, the abstract or the keywords. For instance, a query for fault diagnosis related keywords and transfer learning is designed as follows: ("condition monitoring" OR "fault diagnosis" OR "fault classification" OR "fault detection") AND "transfer learning" The annual number of articles about the application of machine learning (ML) to condition monitoring (CM) and fault diagnosis (FD) increases exponentially. That is also the case for the annual number of natural language processing (NLP) articles, which now equates the total annual number of FD-related articles. A total of 15 articles that utilize NLP on work orders (WO) were found, but no implementations of natural language supervision on fault diagnosis problems were identified. Weak supervision, or weakly supervised learning, is also not yet commonly used, with 4 articles in 2020 and 5 articles so far in 2021.
### 1.3. Research Trends

We also surveyed the fault diagnosis literature and recent publications on language-based learning in the context of natural language supervision and image captioning to identify the trends of publications that combine these concepts. Figure 2 shows the number of published articles per year according to Scopus for search queries containing keywords related to fault diagnosis and machine learning. We present the publication trends of natural language processing (NLP), fault diagnosis (FD), fault diagnosis with machine learning (FD + ML), fault diagnosis with transfer learning (FD + Transfer Learning), image captioning, work orders with natural language processing (WO + NLP), and finally fault diagnosis with weak supervision. For fault diagnosis, a query including "condition monitoring" OR "fault diagnosis" OR "fault detection" OR "fault classification" was used. For machine learning, "machine learning" OR "data driven" OR "deep learning" OR "artificial intelligence" were used. The queries "transfer learning", "word order", "natural language processing" and "image captioning" were used explicitly as is. Weak supervision was queried as "weak supervision" OR "weakly supervised".

The trends show that machine learning is increasingly applied in the FD literature, and that transfer learning has become increasingly popular, going from 2 publications in 2016 to 178 publications in 2020. NLP is a rapidly evolving field of research, with significant practical advancements in the last decade. This is also reflected in the swift growth of image captioning publications starting in 2015, increasing from 11 to 322 publications in four years. 15 publications that utilize natural language processing with work orders were found, but NLP was employed for information retrieval, and no publications combining natural language supervision with IFD were found. Weak supervision only appeared nine times (with one valid article scheduled for 2022 not counted) in our queries, but notably three articles were cited more than ten times; Li et al (2020) with 64 [32] and 30 [33] citations, and Yu et al (2021) with 12 [34], showing that the interest far outweighs the current publication number. Articles citing weak supervision articles were mainly focused on transfer learning, but we predict an increase in direct mentions of weak supervision methods.

### 1.4. Outline of article

In Section II, we describe the application of FD in process industry, which is subject to constraints related to the high cost of unplanned stops that can affect the whole production process. Five principal FD tasks are described, and the related methods and algorithms used for automated FD are also presented. In section III we briefly review natural language supervision and related fields such as image captioning, and discuss how natural language can be integrated in an IFD framework. We focus on rotating machinery in process industry, but in principle the framework of natural language supervision is expected to generalise to other fault diagnosis applications.

### 2. Deep Learning in Intelligent Fault Diagnosis

Table 2 summarises different data-driven methods used for IFD, besides the kinematic rule-based method already discussed in the background. The methods are ordered roughly by maturity and data requirements. Unsupervised learning applies directly to unlabelled CM data, and it is partially implemented in process industries [35, 36, 37, 38]. Supervised learning requires a labelled dataset in the application environment, and is widely investigated in the literature [39, 23, 20, 40, 21, 41], but not in process industry. Transfer learning requires a labelled dataset for pretraining, and data from the application environment, ideally labelled, for finetuning. The number of articles on transfer learning have increased rapidly in the last decade, but although transfer between lab environments show great results, we find no articles that apply transfer learning.
methods directly on process industry data. Finally, natural language supervision based learning only requires unlabelled CM data with associated annotations, but this method remains to be adapted and investigated for fault diagnosis tasks. The first mentions of natural language processing for supervised IFD was in 2020 in the name of "technical language processing", though natural language supervision is yet to be introduced to IFD.

2.1. Unsupervised Learning

Unsupervised learning, i.e learning patterns without labels, is connected to the modelling module and is primarily used for clustering, encoding, feature extraction and anomaly detection fault detection [42]. Models commonly used for clustering are K-Nearest Neighbours (KNN), principal component analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE). (Variational) Auto-Encoders [43, 44] and Dictionary Learning [45, 46] are commonly used for Encodings and Anomaly Detection. Virtually all models can be used to reduce dimensionality and extract features depending on the data, with PCA and T-SNE being more direct dimensionality reductions and auto-encoders serving as a more complex reconstructing model, often with encoders/decoders based on convolutions, recurrence or transformers.

Clustering, encodings and feature extraction can be valuable ways of understanding, simplifying or visualising data. A CM dataset with healthy and unhealthy data might be divisible into two clusters, which can then be manually labelled healthy and unhealthy, thus detecting faults [47]. Likewise, encodings or extracted features can serve as values in a simple rule-based system for fault detection or classification, and extracted features in particular can give valuable insight in feature importance. Regardless, the lack of a supervision signal necessitates a human in the last step to validate or assign meaning to clusters, encodings or features, before the model is ready to automatically detect faults. Anomaly detection can work more autonomously by learning the healthy state of a signal, then classifying deviations from this state as detected faults [48] or into fault classes [49].

However, healthy states that lack sufficient presence in the training set might be considered unhealthy at deployment, and unhealthy states that are present during training might be considered healthy, which is difficult to detect due to the lack of labels in the dataset. Furthermore, deviations might occur due to healthy states, or in directions relatively orthogonal to previous deviations. Such issues fall within the scope of zero-shot learning, wherein a model is required to observe and predict samples from a previously unseen class or distribution [50]. For zero-shot learning to work, there has to be a distinct characteristic of faults and healthy states that is true for previously unseen faults or healthy states, which can be leveraged to assign these distributions to the correct class. NLS is sometimes discussed in the scope of zero-shot learning, and zero-shot learning techniques are often used in NLS. Likewise, zero-shot learning can be used to augment supervised learning methods beyond classes present in the supervision signal, but it is best described under the umbrella term of unsupervised learning or through the lens of weak supervision, as discussed in 2.2. Supervised Learning.

2.2. Supervised Learning

Supervised Learning can be employed for any FD task, as long as sufficient data and good labels are present. Transfer Learning, Weak Supervision and Language Supervision are all arguably subgroups of supervised learning explicitly designed to circumvent the limitation of requiring good labels. Architectures used in supervised learning are thus also employed in its derivatives, though with different learning procedures, just as how for instance auto-encoders from unsupervised learning can be used together with an output layer in a supervised paradigm.

Supervised learning architectures used in IFD range from "shallow" models such as tree-based models such as random forest [51]; support vector machines [39, 52]; probabilistic models such as Bayesian statistics [53, 54]; and deep architectures such as fully connected feed forward deep neural networks [55]; (variational) auto-encoders with classification layers [43, 44]; convolutional neural networks [57, 58], commonly used in image analysis; recurrent neural networks [59, 60, 61], commonly used in language analysis but applicable on sequential data in general. Importantly, supervised learning has been employed for fault severity estimation [13] and RUL prediction [62, 63, 64, 65, 14, 15].

Labelling industry datasets for supervised learning can facilitate implementations in that industry environment, but the labelling process is costly, and requires analyst efforts. Furthermore, some faults have stochastic features, for example due to the varying nature of the source geometry or signal transfer function, and are thus difficult to generalise with supervised classifiers. In general, faults are undesirable and
therefore relatively scarce in industrial datasets, but they are required in training datasets for supervised learning. Consequently, producing a labelled industry dataset for supervised learning would require considerable resources and potentially occupy analyst time necessary for condition monitoring. Therefore, fault classification models described in the literature are typically trained on labelled data from lab environments, where faults are generally either artificially induced or provoked through intense loads, as it might take several years until faults develop naturally. The development of the fault is then accelerated by e.g. high loads or starved lubrication, which increase fault development per revolution, and high speeds to increase revolutions per minute (RPM). High RPM also produce higher signal-to-noise ratios as some noise is stationary and fault features increase more in magnitude than noise features.

Ideally, a model supervised on a component in a lab environment would then be deployable in an industry environment, but there are two issues that makes this difficult. Firstly, artificial faults or accelerated fault developments result in fault characteristics that are different compared to faults in industry environments. Therefore, the decision boundaries do not necessarily generalise well from lab to industry environments, and the feature space can differ due to different fault development processes. Secondly, signals generated in a lab setting differ greatly from signals in an industry environment where a component is connected to several other components in a larger system, and signal components are combined and masked by noise. The signal to noise-ratio will consequently be lower in the industry environment, and the coupling with surrounding components can shift the true feature space as well. Thus, direct supervised learning works best in the environment where it has been trained, and generalisation can be difficult unless labels are preserved in the target space.

2.3. Transfer Learning

Recently, the research focus in IFD has shifted to include methods to overcome the lack of labels in industry datasets such as transfer learning and weak supervision.

Transfer learning seeks to develop methods for training of a model in one environment, then fine-tuning the feature space and decision boundaries to suit implementation in another environment [66]. In situations with sparse data optimization limits, transfer learning can utilize domains with rich data, such as lab datasets, to infer necessary knowledge [67]. The research on transfer learning in fault diagnosis applications has increased rapidly over the last few years, with many successful transfers between different lab datasets [9]. As models improve, transfer learning can enable broader implementation of these models in process industry with a lower demand for labeled instances compared to supervised learning [68]. Tasks solvable through transfer learning are the same as in supervised learning, with the added benefit of generalising from one environment to another, thereby reducing the number of label instances needed in the target domain. Methods used in transfer learning vary; many publications use transferrable convolutional neural networks [69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79], occasionally employed with adversarial networks [80, 81, 82]; some use recurrent neural networks [83, 84, 85]; auto-encoders are also used [86], and recently weak supervision [87] and digital twin-based transfer learning [88] have been successfully implemented.

Transferring knowledge from one environment to another adds an additional benefit to symbol-feature relation graphs besides illustrating the process of the reasoning module. Humans learn concepts in a highly transferable manner, and it is for instance highly feasible that an experienced analyst could diagnose faults in a previously unseen environment with good accuracy, while a learning based model would certainly fail at adapting unless optimized through transfer learning. The underlying concepts of fault developments are likely the same in both environments, which is what humans utilize to generalise knowledge. Optimizing not only direct mappings, but symbol-feature relation graphs as well, can thus create models with stronger generalisability by mimicking human knowledge [89].

2.4. Weak Supervision

Weak supervision is an umbrella term for a set of methods developed to perform supervised tasks on data where labels are insufficient for regular supervised learning [90]. It can work in conjunction with transfer learning to enhance the fine-tuning on the target dataset, or stand-alone to facilitate direct optimization in the target environment. Table 3 illustrates three major ways in which labels can be insufficient, the cause, and proposed methods to amend the issue. Incomplete labels are characterised by a dataset where most data points are unlabelled. In a CM dataset, faults that have not been discovered yet are a cause for incompleteness, as this prevents the assumption that all unlabelled data is healthy data. *Inexact labels*
Table 3: Different weak supervision challenges, causes and solutions

| Weak Supervision group | Cause                                      | Solution                      |
|------------------------|--------------------------------------------|-------------------------------|
| Incomplete labels      | Missing Labels from datapoints             | Active Learning               |
|                        |                                            | Semi-Supervised Learning     |
|                        |                                            | Few-shot Learning             |
|                        |                                            | Zero-shot Learning            |
| Inexact labels         | Multiple datapoints per label              | Multi-instance learning       |
|                        |                                            | Contrastive Learning          |
| Inaccurate labels      | Label is wrong                             | Regularization                |
|                        |                                            | Re-labelling                  |

Coarsely describe some aspects of the ground truth for a set of features, but do not accurately define it. In general, symbols like labels can not fully represent physical processes of unknown dimensions. Instead, labels define semantics at a certain level of approximation and scale. Thus, labels of physical processes are by nature incomplete semantical descriptions of reality. For CM annotations, this is the case when annotations for example describe fault class, and not severity, or severity, but not severity development. Inaccurate labels occur when analysts make fault diagnosis mistakes. This is unlikely to occur with fault classification, but possible with fault severity due to the higher complexity of that task. An analyst can for example assume that a fault may be severe and order a replacement of the component to avoid failure, while the fault actually is minor.

The main strategy for dealing with incomplete datasets is called semi-supervised learning [91, 92, 93], which aims to create clusters of features that correspond to the available labels, and to estimate the probability that an unseen feature belongs to one of the identified clusters. Semi-supervised learning has been employed in IFD settings on lab datasets with partial [94] or limited labels [95]. By implementing semi-supervised learning on a CM dataset with natural language supervision, it is possible to include all time series data for a prediction, where particularly noisy samples would be less likely to affect the model optimization process, as they are likely distributed far away from the cluster centres. The diagnosis of faults in unlabelled samples also belong to the domain of semi-supervised learning, albeit with the additional challenge associated with many unique components and features. This challenge can necessitate active learning [96, 93], in which a model identifies selected unlabelled datapoints and alerts a human expert to label them. Active learning requires human intervention, but aims to make use of human efforts as efficiently as possible to improve the model accuracy [67]. Another scheme utilized to overcome incomplete labels is few-shot learning [97], where a model is optimized to perform supervision tasks with insufficient data for normal supervision training [98, 99, 100]. Few-shot learning provides an interesting opportunity to learn fault features with only a few instances in a training datasets, as can be the case for many rare faults or components. In the case where no labels exist, supervision algorithms might still be applicable through zero-shot learning [50]. In zero-shot learning, the model seeks to generalise knowledge from seen classes to unseen classes with similar behaviour, much like how humans can see images of house-cats and dogs and then correctly categorize lions to felines and wolves to canines [101, 102].

The challenge of inexact labels has been proposed to be overcome through multi-instance learning [103, 104, 105] and contrastive learning. In multi-instance learning, the optimization algorithms seeks to find the common denominators in the label "bags" that are present for learning. By learning from which components were replaced and which were not, correlations in underlying features such as fault severity or deterioration speed can be associated as parts of the bag and used for predictions.

Inaccurate labels are characterised by not conforming to the ground truth, in other words being wrong. To learn with noisy or inaccurate labels, a model seeks to identify and potentially correct incorrect labels [106]. Thus, the model maintains some trust in its predictions, capable of deeming the label inaccurate when confidence in prediction is high and label features deviate from similar labels [107].
3. Technical Language Supervision

The direction of research in IFD points towards finding ways to transfer the success on lab datasets to successful applications on industry datasets [9, 108]. Both transfer learning and weak supervision can create the opportunity to implement successful algorithms on new datasets without requiring an expensive labelling process. Inspired by recent innovations in TLP and NLS, TLS present a third, yet unutilized direction to integrate the annotations present in CM datasets as labels, learning directly from technical language.

The potential effects of TLS can be summarized as

- **Opportunities**
  - Facilitates direct optimization on heterogeneous industry data
  - Methods are available and developed in other research areas
  - Language data is commonly associated with condition monitoring data-bases

- **Challenges**
  - Language annotations are uncertain, and require weak supervision techniques to utilize
  - NLS methods for technical language is a novel area of research yet to be developed
  - Rapid progress requires open industry data sets containing potentially sensitive information

TLS faces two major implementation challenges requiring further research and development - implementation of big data NLS techniques on more limited industry data, and language processing of technical language as compared to the natural language present in NLS datasets. In this section, we briefly describe the state of TLP and NLS, then combine these into an outline of how TLP can be implemented.

3.1. Natural Language Supervision

Natural Language Supervision is a recent term introduced to describe machine learning optimization based on natural language processing rather than predefined labels. Interactions between human language and machine learning occur in many domains; in pure NLP, Labutov et al. [109] trained semantic parsers that interpret questions and feedback from user natural language responses. Hancock et al. [110], used natural language explanations of human labelling decision to create BabbleLabble, which converts explanations to noisy labels through a semantic parser. Murty et al. [111] introduced ExpBERT, which is a BERT variation that forms representations using BERT with natural language explanations of the inputs.

Text-encoding is a crucial part of NLP and has seen rapid development recent years. Language models [112] based on the transformer have increased the representational powers of text encoders drastically [113, 114, 115, 116, 117, 118]. Early examples of text-image pairings used simpler encoding methods, such as bag of words and TF-IDF, or recursive encodings derived from the word2vec model [119], the predecessor of current transformer-based language models. The choice of text-encoder depends on data size and computational power; a larger model can produce better representations, but requires more data and computational power to train. Pre-trained language models with general natural language representational capacity, such as BERT [114], have successfully been fine-tuned on specific tasks with significantly smaller datasets, based on the assumption that the target language and source language has similar underlying distributions.

Optimizing mappings between natural language and images has been done before natural language supervision was introduced; for example, image captioning [120, 121, 122] and visual question answering [123] have both trained mappings between images and text through top-down or bottom-up mappings [124] and semantic attention [125, 126]. Knowledge and concepts can also be integrated using language as a supervision tool through neuro-symbolic concept learning [127], where visual concepts, word representations, and semantic parsing of sentences are jointly learned.

Image recognition generally utilizes image-text pairs available from online data crawling to train mappings between text and images. Learning directly from the text can also facilitate zero-shot classifiers from language descriptions. Elhoseiny et al. [128] used text-based descriptions to create a zero-shot image classifier. Text features extracted through TF-IDF followed by Clustered Latent Semantic Indexing. Lu et al. [129] introduced ViLBERT, a Vision-and-Language version of BERT, that learns image recognition and
Table 4: Annotations associated with data from Figure 3.

| Case       | Months after fault detection | Annotation (translated from Swedish)                          |
|------------|-----------------------------|---------------------------------------------------------------|
| BPFO indication | 4                           | BPFO Env low                                                  |
| BPFO       | 10                          | BPFO visible in mm/s as overtones high up in the spectrum between 1000 and 2000 Hz. WO written on BPFO |
| Feedback   | 12                          | Bearing replaced YYYYYMMDD levels of BPFO low again            |

Language understanding in a two-stream model with interactions between image and text to improve performance compared to single-stream models. Zhang et al. [130] classified medical images by utilizing text-image pairs through contrastive visual representation learning (ConVIRT) to learn pairings between images and texts. Desai & Johnson [131] introduced Virtex, which utilizes captions to enhance pretraining of an image recognition CNN. Sariyildiz et al. [132] mask words in image-annotations pairs to create image-conditioned masked language modelling (ICMLM) for image classification.

In a recent publication, Radford et al. at OpenAI [30] presented CLIP, Contrastive Language–Image Pre-training, which is the most extensive paper on natural language supervision yet. They used transformers [133] for both text and image encodings [134], and a contrastive [135] bag-of-words prediction objective to connect text label to image features in a vector quantized encoding space [136, 137]. FILIP by Yao et al. [138] utilizes a fine-grained word-patch image alignment to detect and classify objects based on text descriptions, obtaining finer level-alignment in image-text comprehension through unsupervised natural language supervision. Jia et al. [139] scaled natural language supervision further by training directly on un-filtered images and annotations with over one billion image-text pairs. Wang et al. [140] introduced unsupervised data generation to synthesize labels for downstream tasks and thus achieve SOTA results on SuperGLUE [141].

In earlier models, Ramanathan et al. [142] used natural language supervision to train a video event understanding model in 2013 through a rule-based bag-of-words-like model, and Williams et al. [143] used language as reward functions for training robots.

3.2 Technical Language Processing

TLP concerns the application of NLP techniques and pipelines on technical language [144]. The processing of technical language requires natural language processing methods with additional considerations related to the characteristics of technical language, which is characterised by a higher frequency of information-rich key-words, more abbreviations, and considerably less data than natural language. TLP can be utilized as a basis for TLS, but can also directly enhance CM practices by offering insights into key performance indicators from work order features [145].

Language-based models require a mathematical representation of language. This is achieved through pre-processing and an embedding algorithm. The pre-processing step involves tokenisation, cleaning and spell-checking, stop-words removal, stemming/lemmatisation, and fundamental language analysis such as part of speech tagging and named entity recognition. The embedding algorithm can be as simple as one-hot encoding or a complex massive transformers based architecture.

Figure 3 and Table 4 showcase an example of technical language annotations and condition monitoring signals from a craft liner production plant in northern Sweden. The figure shows three different envelope-filtered measurements associated with the annotations shown in the table. The first annotation indicates that there is a fault of class Ball-Pass Frequency Outer ring (BPFO) with a low severity, which is related to the low-intensity peaks at characteristic kinematically based order frequencies in the spectrum. The second annotation describes that the corresponding overtones have increased in magnitude and that a work order has been written. At that point the fault is estimated to be more severe and at the end of its RUL, so the component (bearing) has to be replaced. Finally, the third annotation is a note that the bearing has been replaced and that the vibration levels are low, indicating a healthy component.

Pre-processing of technical language faces several difficulties, as use of technical language can vary even in the same field, and there is no uniformly defined list of stems/lemmas, stop-words or correct spellings. For instance, if a CM dataset contains faults of class "Ball-Pass Frequency Outer" (BPFO) and "Ball-Pass
Figure 3: Order analysis results for a vibration signal at a) 4 months; b) 10 months; and c) 12 months after the first indication of a fault in a drying cylinder bearing of a paper machine. Included are also the corresponding text annotations written by experienced condition monitoring analysts employed at the factory. The annotations have been translated from Swedish to English to improve clarity. BPFO peaks are clearly visible in panel a) four months after the first indication of the bearing fault. After ten months, the amplitude of the BPFO peaks in panel b) have increased and a work order (WO) has been written by the analysts. Two months later the bearing has been replaced and no BPFO signature can be seen in panel c).

Frequency Inner” (BPFI), but one is considerably more common than the other, an automated spell-checker might assume that one is a spelling error. Likewise, there is no defined dictionary for stemming of technical words such as BPFO or BPFI, and reducing both words to ”BPF” naturally loses critical information. Therefore it is necessary with a ”human-in-the-loop” system until a level of language processing maturity which accurately covers the heterogeneous field of technical language is achieved. One dictionary of technical stop words has been produced [146], though it is not necessarily the case that this list is accurate for industries besides those covered in the article.

Encoding technical language to vectors faces a major challenge in that many technical words specific to industries are not in the vocabulary of NLP models trained on natural language. Addressing this directly with NLP methods is thus related to handling out of vocabulary (OOV) words. A common method to deal with OOV words, used in for instance BERT [114] and GPT [147, 148, 149], is to input subwords and byte-pair encodings rather than the words themselves to the model. Other models try to learn to predict the meaning of an unknown word based on surrounding words, individual characters, or a combination of both [150]. Implementing an OOV solution which allows transfer learning of a pre-trained deep learning NLP encoder could potentiate more semantically accurate representations of technical language word embeddings, which in turn would improve the potential for TLS.

Another method to encode technical language is through human designed expert systems - essentially a
set of rules describing the keywords for faults, actions, severities etc \[151\]. The annotation "High BPFO in env3, WO on bearing replacement" would thus be decomposed into

\[
\text{class} - \text{BPFO}; \text{severity} - \text{high}; \text{detected in} - \text{env3}; \text{action} - \text{WO replacement}; \text{action target} - \text{bearing}.
\]

These keywords can then serve as targets for annotation prediction or language based supervision, acting as less noisy labels than learned embeddings for language representations. However, such a system is difficult to scale and vulnerable to new keywords being introduced, essentially requiring tailored engineering and maintenance for each unique industry. It is also vulnerable to oversights from the engineers of the expert system, for instance missing negations in statements, unforeseen keyword usage or a lack of context due to the removal of semantics.

### 3.3. Outline of Technical Language Supervision concepts and model

In the infant stage of TLP, classical NLP methods such as stop-word removal, lemmatisation, stemming and bag-of-words-analysis have been utilized. A potential improvement is to apply more recent innovations in preprocessing and analysis, such as a word embedding algorithm coupled with manual tagging of industry-specific technical language.

A TLS model consists of four parts, as shown in Figures 4 and 5, which we modified and adapted from \[30\]. In the pre-training step, a technical language encoder and a fault diagnosis encoder are used to produce fault and text features. A mapping between fault and text encodings is learned through contrastive learning \[135, 130\]. In the inference phase, the same encoders are used, but additionally there exists a label query mechanism that maps an input signal to the annotation-based label that is closest to the query in the joint data and language embedding space.

In the case of IFD of rotating machinery, the input is typically sensor data in time-, frequency- and time-frequency-domains. IFD data encoding methods are described in section II, and typically consist of variations of CNNs. Recently, the Transformer \[133\], an architecture introduced to model long-range dependencies and training inefficiencies in NLP, has been successfully used for image recognition without any convolutions in the model \[134, 152, 153\].

In order to train classification or regression models using language, and not just an annotation generator, a language based labelling method is required. Based on current state-of-the-art methods, some human

![Figure 4: Example illustrating the pretraining step of a natural language supervision model. Annotations and time-frequency domain signal features are encoded, and the model is optimized to connect the correct text-feature pair in the batch, here marked with dark green colour, through contrastive learning.](image)
Figure 5: Example illustrating how inference can be generated with the natural-language supervised model outlined in Figure 6. Given a model that connects signal features to annotations, an unannotated set of features is connected to the closest matching term, which serves as a label through a query process. The matching labels are the prediction outputs, such as fault class and severity.

intervention is required in this step to pre-define the label-space, so that annotations can be matched to the closest label semantically. In [30], a Bag-of-words method is implemented to complete pre-defined sentence structures by inserting the correct term chosen from the bag. A similar model could be used in IFD, with more than one degree of freedom in the query to label both fault class and severity. Potentially, further degrees of freedom also enables labelling time-aspects of fault evolution. With a large text dataset and access to well defined labels in parallel with the annotations, a mapping between a more feature-rich encoding and the label space can be learned and implemented to produce labels in a weakly supervised manner for data-annotation pairs where labelled data are not available.

In the case of CM data, the volume and density of text data is low compared to web-crawl results for captioned images on the Internet, or extensively annotated datasets such as COCO [154]. The language is also domain specific, and annotations are connected to time-frequency data instances in the dataset, while the semantics of an annotation can be based on analysis of trends over many measurement instances. This motivates the use of pre-trained models, in combination with feature-engineering and fine-tuning to adapt the model to the domain-specific terms used in process industry. Weak supervision will also be required to deal with unannotated faults, time-delays, a lack of annotations in healthy data, and noise in the annotations resulting from domain-specific language, spelling errors, and grounding noise due to subjective interpretations.

4. Conclusion

The fault descriptions and maintenance records commonly stored in modern process industry CM systems are unexploited information source for training IFD systems. The language present in CM datasets can in principle be used as labels for natural language supervision of IFD models [155, 156] to facilitate automation of routine FD tasks and develop more accurate decision support for more complex tasks. Since language based labels are intrinsically uncertain, weakly supervised learning methods need to be developed, which can also support transfer learning of IFD pretrained models with labels generated from industry datasets.

The flow of data and knowledge in the envisioned language-integrated IFD system is shown in Figure 6. Intelligent fault diagnosis constitutes a layer between the analysts and the monitored process industry equipment. With the integration and processing of annotations, data originates both from the process layer and the analyst layer, with learning (constrained optimisation) and reasoning (rational inference and decision making) enabling or enhancing the communication. This way the fault-detection knowledge is gradually updated by a reasoning process (artificial intelligence algorithms), which integrates feedback from analysts as well as relations learned or prescribed by models, for example in terms of pretraining or kinematics. The data originates from the process equipment and sensors layer, which represents the physical reality. The IFD
Supervision and Analysis

Fault severity, remaining useful life, root cause

Intelligent Fault Diagnosis

Reasoning

Probabilistic and logic based inference

Modeling

Physical and data-driven models

Decisions

Maintenance plan, preventive measures

Supervision and Analysis Experts

Fault severity, remaining useful life, root cause

Intelligent Fault Diagnosis

Data

Symbol-feature relations graph

Knowledge-base

Process Equipment and Sensors

Sensors

Actuators

Controllers

Database

Figure 6: Illustration of the interplay between the equipment, analysts and cognitive layer in the envisioned IFD system with technical language supervision capabilities. The supervision layer provides analysis and decisions in the form of annotations and maintenance work orders. The Process layer provides sensor and process data to the IFD database. In the IFD layer, the technical language knowledge from the supervision layer is used for model optimization, inference based reasoning and formation of a graph-like knowledge-base between predictions, language and process layer data. The IFD layer provides decision support, modelling updates and aligns the integrated knowledge with feedback from the analysts.

The modules in the IFD layer represent different aspects of data processing. The database module contains sensor data, with associated process parameters, annotations and work order history from process components. The models used to analyse data based on existing knowledge and model parameters are maintained by the models module. For example, a rule-based model can be defined in terms of specific signal processing algorithms and kinematic frequencies of a monitored bearing, and a data-driven model for bearing fault development analysis can be defined in terms of historic sensor data and annotations. The models depend on the physical properties of the process layer. For example, a change of bearing that results in a different number of rolling elements in the bearing will severely affect the accuracy of models that are not updated accordingly. The reasoning module fuses the output from the models with existing knowledge to provide decision support and improve the inferences made via knowledge and model improvements based on feedback from the analysts. The reasoning module answers questions like: "given the following data, what is the fault presence/class/severity/remaining useful life?". The knowledge module integrates known relationships, for example between data, context, models and annotations, and thus integrates features with interpretable concepts to enhance the generalisation capacity and transparency of the IDF system. Here, "context" can for example include aspects like the placement of a component in the process, slowly varying process variables, such as the product characteristics, the component age, or the next planned maintenance stop.

Further research is required to adapt and integrate natural language processing methods and models to enable weak supervision and transfer learning of IFD models using technical language annotations. In particular, overcoming the challenges in TLP can directly result in improved supervision. Improvements of TLP can occur both through an enhancement of the standard NLP pipeline as implemented on technical language, or through augmented integration of deep embedding models with a strong capacity to handle out of vocabulary words. However, a major challenge in TLP research is the requirements of annotated industry data, which can be difficult for researchers to obtain rights to use. Furthermore, the assistance of an industry expert might be required to fully understand the annotation language and how annotations were motivated...
by signals. Therefore, we encourage further collaboration between industry experts and academia, and that annotated datasets be made public when possible, with clearly described features.

References

[1] E. Peter Carden and Paul Fanning. Vibration based condition monitoring: A review. *Structural Health Monitoring*, 3(4):355–377, 2004.

[2] Andrew K.S. Jardine, Daming Lin, and Dragan Banjevic. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, 20(7):1483 – 1510, 2006.

[3] Ranganath Kothamasu, Samuel H. Huang, and William H. VerDuin. System health monitoring and prognostics — a review of current paradigms and practices. *The International Journal of Advanced Manufacturing Technology*, 28(9):1012–1024, Jul 2006.

[4] Cen Nan, Faisal Khan, and M. Tariq Iqbal. Real-time fault diagnosis using knowledge-based expert system. *Process Safety and Environmental Protection*, 86(1):55–71, 2008.

[5] SKF. Skf @ptitude observer user manual.

[6] A.K.S. Jardine, D. Lin, and D. Banjevic. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, 20(7):1483–1510, 2006, cited By 2690.

[7] A. Rai and S.H. Upadhyay. A review on signal processing techniques utilized in the fault diagnosis of rolling element bearings. *Tribology International*, 96:289–306, 2016. cited By 236.

[8] Robert B. Randall and Jérôme Antoni. Rolling element bearing diagnostics—a tutorial. *Mechanical Systems and Signal Processing*, 25(2):485 – 520, 2011.

[9] Yaguo Lei, Bin Yang, Xinwei Jiang, Feng Jia, Naipeng Li, and Asoke K. Nandi. Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mechanical Systems and Signal Processing*, 138:106587, 2020.

[10] Processit.eu european roadmap for process industrial automation (2018). second version.

[11] Jong-Ho Shin and Hong-Bae Jun. On condition based maintenance policy. *Journal of Computational Design and Engineering*, 2(2):119 – 127, 2015.

[12] AB SKF and Solberg Kommunikation. Skf annual report 2020. [https://investors.skf.com/sites/default/files/pr/202103032688-1.pdf](https://investors.skf.com/sites/default/files/pr/202103032688-1.pdf), March 2020.

[13] Mariela Cerrada, René-Vinicio Sánchez, Chuan Li, Fannia Pacheco, Diego Cabrera, José [Valente de Oliveira], and Rafael E. Vásquez. A review on data-driven fault severity assessment in rolling bearings. *Mechanical Systems and Signal Processing*, 99:169 – 196, 2018.

[14] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin. Machinery health prognostics: A systematic review from data acquisition to rul prediction. *Mechanical Systems and Signal Processing*, 104:799–834, 2018, cited By 284.

[15] D. Wang, K.-L. Tsui, and Q. Miao. Prognostics and health management: A review of vibration based bearing and gear health indicators. *IEEE Access*, 6:665–676, 2017. cited By 117.

[16] PdM. Pdm services vibration analysis monitoring.

[17] Jim Cahill. Improving subsurface models to reduce drilling uncertainty.

[18] V. Sharma and A. Parey. A review of gear fault diagnosis using various condition indicators. volume 144, pages 253–263, 2016. cited By 119.

[19] Tak chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

[20] R. Liu, B. Yang, E. Zio, and X. Chen. Artificial intelligence for fault diagnosis of rotating machinery: A review. *Mechanical Systems and Signal Processing*, 103:33–47, 2018. cited By 422.

[21] A. Stetco, F. Dinmohammadi, X. Zhao, V. Robu, D. Flynn, M. Barnes, J. Keane, and G. Nenadic. Machine learning methods for wind turbine condition monitoring: A review. *Renewable Energy*, 133:620–635, 2019. cited By 111.

[22] D.-T. Hoang and H.-J. Kang. A survey on deep learning based bearing fault diagnosis. *Neurocomputing*, 335:327–335, 2019. cited By 87.

[23] S. Khan and T. Yairi. A review on the application of deep learning in system health management. *Mechanical Systems and Signal Processing*, 107:241–265, 2018. cited By 264.
[24] Case western reserve university bearing data center website. https://csegroups.case.edu/bearingdatacenter/pages/welcome-case-western-reserve-university-bearing-data-center-website. Accessed: 2020-01-30.

[25] Nasa prognostic data repository. https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/. Accessed: 2021-01-30.

[26] Condition based maintenance fault database for testing of diagnostic and prognostics algorithms. https://www.mfpt.org/fault-data-sets/. Accessed: 2021-09-30.

[27] W.A. Smith and R.B. Randall. Rolling element bearing diagnostics using the case western reserve university data: A benchmark study. Mechanical Systems and Signal Processing, 64-65:100–131, 2015. cited By 352.

[28] Shen Zhang, Fei Ye, Bingnan Wang, and Thomas G. Habetler. Semi-supervised learning of bearing anomaly detection via deep variational autoencoders, 2019.

[29] Sebastian Schwendemann, Zubair Amjad, and Axel Sikora. Bearing fault diagnosis with intermediate domain based layered maximum mean discrepancy: A new transfer learning approach. Engineering Applications of Artificial Intelligence, 105:104415, 2021.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[31] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

[32] X. Li, W. Zhang, Q. Ding, and X. Li. Diagnosing rotating machines with weakly supervised data using deep transfer learning. IEEE Transactions on Industrial Informatics, 16(3):1688–1697, 2020. cited By 28.

[33] Xiang Li, Xu Li, and Hui Ma. Deep representation clustering-based fault diagnosis method with unsupervised data applied to rotating machinery. Mechanical Systems and Signal Processing, 143:106825, 2020.

[34] Kun Yu, Qiang Fu, Hui Ma, Terry Lin, and Xiang Li. Simulation data driven weakly supervised adversarial domain adaptation approach for intelligent cross-machine fault diagnosis. Structural Health Monitoring, 20, 07 2021.

[35] SKF. Skf enlight ai.

[36] Amazon Monitron. Detect abnormal machine behavior and enable predictive maintenance.

[37] Julien Simon. Amazon monitron, a simple and cost-effective service enabling predictive maintenance.

[38] Emerson. Featured technologies/machine learning.

[39] S. Yin, S.X. Ding, X. Xie, and H. Luo. A review on basic data-driven approaches for industrial process monitoring. IEEE Transactions on Industrial Electronics, 61(11):6418–6428, 2014. cited By 841.

[40] Georg Helbing and Matthias Ritter. Deep learning for fault detection in wind turbines. Renewable and Sustainable Energy Reviews, 98:189 – 198, 2018.

[41] S. Zhang, S. Zhang, B. Wang, and T. G. Habetler. Deep learning algorithms for bearing fault diagnostics—a comprehensive review. IEEE Access, 8:29857–29881, 2020.

[42] Y. Lei, F. Jia, J. Lin, S. Xing, and S.X. Ding. An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data. IEEE Transactions on Industrial Electronics, 63(5):3137–3147, 2016. cited By 456.

[43] G. Jiang, P. Xie, H. He, and J. Yan. Wind turbine fault detection using a denoising autoencoder with temporal information. IEEE/ASME Transactions on Mechatronics, 23(1):89–100, 2018.

[44] S. Haidong, J. Hongkai, L. Xingqiu, and W. Shuaipeng. Intelligent fault diagnosis of rolling bearing using deep wavelet auto-encoder with extreme learning machine. Knowledge-Based Systems, 140:1–14, 2018. cited By 119.

[45] Vardan Papyan, Yaniv Romano, Jeremias Sulam, and Michael Elad. Theoretical foundations of deep learning via sparse representations: A multilayer sparse model and its connection to convolutional neural networks. IEEE Signal Processing Magazine, 35(4):72–89, 2018.

[46] Haining Liu, Chengliang Liu, and Yixiang Huang. Adaptive feature extraction using sparse coding for machinery fault diagnosis. Mechanical Systems and Signal Processing, 25(2):558 – 574, 2011.

[47] C.T. Yiakopoulos, K.C. Gryllias, and I.A. Antoniadis. Rolling element bearing fault detection in industrial environments based on a k-means clustering approach. Expert Systems with Applications,
S. Martin del Campo and F. Sandin. Online feature learning for condition monitoring of rotating machinery. *Engineering Applications of Artificial Intelligence*, 64:187 – 196, 2017.

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

T. Zhang, J. Chen, F. Li, K. Zhang, H. Lv, S. He, and E. Xu. Intelligent fault diagnosis of machines with small & imbalanced data: A state-of-the-art review and possible extensions. *ISA Transactions*, 2021. cited By 5.

D. Zhang, L. Qian, B. Mao, C. Huang, B. Huang, and Y. Si. A data-driven design for fault detection of wind turbines using random forests and xgboost. *IEEE Access*, 6:21020–21031, 2018. cited By 168.

S.J. Qin. Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control*, 36(2):220–234, 2012. cited By 672.

A. Stief, J.R. Ottewill, J. Baranowski, and M. Orkisz. A pca and two-stage bayesian sensor fusion approach for diagnosing electrical and mechanical faults in induction motors. *IEEE Transactions on Industrial Electronics*, 66(12):9510–9520, 2019. cited By 33.

H. Zhang, Q. Zhang, J. Liu, and H. Guo. Fault detection and repairing for intelligent connected vehicles based on dynamic bayesian network model. *IEEE Internet of Things Journal*, 5(4):2431–2440, 2018. cited By 30.

F. Jia, Y. Lei, J. Lin, X. Zhou, and N. Lu. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. *Mechanical Systems and Signal Processing*, 72-73:303–315, 2016. cited By 710.

Xiaoan Yan, Daoming She, Yadong Xu, and Minping Jia. Deep regularized variational autoencoder for intelligent fault diagnosis of rotor bearing system within entire life-cycle process. *Knowledge-Based Systems*, 226:107142, 2021.

F. Jia, Y. Lei, N. Lu, and S. Xing. Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization. *Mechanical Systems and Signal Processing*, 110:349–367, 2018. cited By 144.

J. Pan, Y. Zi, J. Chen, Z. Zhou, and B. Wang. Liftingnet: A novel deep learning network with layerwise feature learning from noisy mechanical data for fault classification. *IEEE Transactions on Industrial Electronics*, 65(6):4973–4982, 2018. cited By 80.

H. Liu, J. Zhou, Y. Zheng, W. Jiang, and Y. Zhang. Fault diagnosis of rolling bearings with recurrent neural network-based autoencoders. *ISA Transactions*, 77:167–178, 2018. cited By 173.

M. Qiao, S. Yan, X. Tang, and C. Xu. Deep convolutional and lstm recurrent neural networks for rolling bearing fault diagnosis under strong noises and variable loads. *IEEE Access*, 8:66257–66269, 2020. cited By 25.

X. Chen, B. Zhang, and D. Gao. Bearing fault diagnosis base on multi-scale cnn and lstm model. *Journal of Intelligent Manufacturing*, 32(4):971–987, 2021. cited By 16.

G.S. Babu, P. Zhao, and X.-L. Li. Deep convolutional neural network based regression approach for estimation of remaining useful life. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9642:214–228, 2016. cited By 150.

X. Li, Q. Ding, and J.-Q. Sun. Remaining useful life estimation in prognostics using deep convolutional neural networks. *Reliability Engineering and System Safety*, 172:1–11, 2018. cited By 183.

J. Ben Ali, B. Chebel-Morello, L. Saïdi, S. Malinowski, and F. Fnaiech. Accurate bearing remaining useful life prediction based on weibull distribution and artificial neural network. *Mechanical Systems and Signal Processing*, 56:150–172, 2015. cited By 195.

L. Guo, N. Li, F. Jia, Y. Lei, and J. Lin. A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing*, 240:98–109, 2017. cited By 266.

Chuan Li, Shaozhi Zhang, Yi Qin, and Edgar Estupinan. A systematic review of deep transfer learning for machinery fault diagnosis. *Neurocomputing*, 407:121 – 135, 2020.

Qian Zhang, Jie Lu, and Yaochu Jin. Artificial intelligence in recommender systems. *Complex & Intelligent Systems*, 7(1):439–457, Feb 2021.

P. Cao, S. Zhang, and J. Tang. Preprocessing-free gear fault diagnosis using small datasets with deep
[69] P. Cao, S. Zhang, and J. Tang. Preprocessing-free gear fault diagnosis using small datasets with deep convolutional neural network-based transfer learning. *IEEE Access*, 6:26241–26253, 2018. cited By 156.

[70] S. Shao, S. McAleer, R. Yan, and P. Baldi. Highly accurate machine fault diagnosis using deep transfer learning. *IEEE Transactions on Industrial Informatics*, 15(4):2446–2455, 2019.

[71] Bin Yang, Yaguo Lei, Feng Jia, and Saibo Xing. An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings. *Mechanical Systems and Signal Processing*, 122:692 – 706, 2019.

[72] L. Guo, Y. Lei, S. Xing, T. Yan, and N. Li. Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data. *IEEE Transactions on Industrial Electronics*, 66(9):7316–7325, 2019.

[73] S.-S. Zhong, S. Fu, and L. Lin. A novel gas turbine fault diagnosis method based on transfer learning with cnn. *Measurement: Journal of the International Measurement Confederation*, 137:435–453, 2019. cited By 76.

[74] G. Xu, M. Liu, Z. Jiang, W. Shen, and C. Huang. Online fault diagnosis method based on transfer convolutional neural networks. *IEEE Transactions on Instrumentation and Measurement*, 69(2):509–520, 2020. cited By 81.

[75] Te Han, Chao Liu, Wenguang Yang, and Dongxiang Jiang. Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application. *ISA Transactions*, 97:269 – 281, 2020.

[76] Z. He, H. Shao, X. Zhong, and X. Zhao. Ensemble transfer cnns driven by multi-channel signals for fault diagnosis of rotating machinery cross working conditions. *Knowledge-Based Systems*, 207, 2020. cited By 43.

[77] L. Wen, X. Li, and L. Gao. A transfer convolutional neural network for fault diagnosis based on resnet-50. *Neural Computing and Applications*, 32(10):6111–6124, 2020. cited By 38.

[78] Z. Chen, K. Gryllias, and W. Li. Intelligent fault diagnosis for rotary machinery using transferable convolutional neural network. *IEEE Transactions on Industrial Informatics*, 16(1):339–349, 2020. cited By 64.

[79] H. Shao, M. Xia, G. Han, Y. Zhang, and J. Wan. Intelligent fault diagnosis of rotor-bearing system under varying working conditions with modified transfer convolutional neural network and thermal images. *IEEE Transactions on Industrial Informatics*, 17(5):3488–3496, 2021. cited By 42.

[80] Q. Wang, G. Michau, and O. Fink. Domain adaptive transfer learning for fault diagnosis. pages 279–285, 2019. cited By 38.

[81] Te Han, Chao Liu, Wenguang Yang, and Dongxiang Jiang. A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults. *Knowledge-Based Systems*, 165:474–487, 2019.

[82] X. Li, W. Zhang, N.-X. Xu, and Q. Ding. Deep learning-based machinery fault diagnostics with domain adaptation across sensors at different places. *IEEE Transactions on Industrial Electronics*, 67(8):6785–6794, 2020. cited By 47.

[83] A. Zhang, H. Wang, S. Li, Y. Cui, Z. Liu, G. Yang, and J. Hu. Transfer learning with deep recurrent neural networks for remaining useful life estimation. *Applied Sciences (Switzerland)*, 8(12), 2018. cited By 63.

[84] Z. An, S. Li, Y. Xin, K. Xu, and H. Ma. An intelligent fault diagnosis framework dealing with arbitrary length inputs under different working conditions. *Measurement Science and Technology*, 30(12), 2019. cited By 16.

[85] K. Zhao, H. Jiang, Z. Wu, and T. Lu. A novel transfer learning fault diagnosis method based on manifold embedded distribution alignment with a little labeled data. *Journal of Intelligent Manufacturing*, 2020. cited By 12.

[86] L. Wen, L. Gao, and X. Li. A new deep transfer learning based on sparse auto-encoder for fault diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(1):136–144, 2019.

[87] X. Li, W. Zhang, Q. Ding, and X. Li. Diagnosing rotating machines with weakly supervised data using deep transfer learning. *IEEE Transactions on Industrial Informatics*, 16(3):1688–1697, 2020.
Y. Xu, Y. Sun, X. Liu, and Y. Zheng. A digital-twin-assisted fault diagnosis using deep transfer learning. *IEEE Access*, 7:19990–19999, 2019.

Yunzhu Li, Toru Lin, Kexin Yi, Daniel M. Bear, Daniel L. K. Yamins, Jiajun Wu, Joshua B. Tenenbaum, and Antonio Torralba. Visual grounding of learned physical models, 2020.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 08 2017.

Jesper E. van Engelen and Holger H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, Feb 2020.

Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning, 2019.

Chuanxiao Jian, Kajjun Yang, and Yinhui Ao. Industrial fault diagnosis based on active learning and semi-supervised learning using small training set. *Engineering Applications of Artificial Intelligence*, 104:104365, 2021.

Roozbeh Razavi-Far, Ehsan Hallaji, Maryam Farajzadeh-Zanjani, and Mehrdad Saif. A semi-supervised diagnostic framework based on the surface estimation of faulty distributions. *IEEE Transactions on Industrial Informatics*, 15(3):1277–1286, 2019.

Kun Yu, Tian Ran Lin, Hui Ma, Xiang Li, and Xu Li. A multi-stage semi-supervised learning approach for intelligent fault diagnosis of rolling bearing using data augmentation and metric learning. *Mechanical Systems and Signal Processing*, 146:107043, 2021.

Hamed H. Aghdam, Abel Gonzalez-Garcia, Joost van de Weijer, and Antonio M. López. Active learning for deep detection neural networks, 2019.

Yaqing Wang, Quanming Yao, James Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning, 2020.

D. Zhou, J. He, H. Yang, and W. Fan. Sparc: Self-paced network representation for few-shot rare category characterization. pages 2807–2816, 2018. cited By 30.

A. Zhang, S. Li, Y. Cui, W. Yang, R. Dong, and J. Hu. Limited data rolling bearing fault diagnosis with few-shot learning. *IEEE Access*, 7:110895–110904, 2019. cited By 60.

Z. Ren, Y. Zhu, K. Yan, K. Chen, W. Kang, Y. Yue, and D. Gao. A novel model with the ability of few-shot learning and quick updating for intelligent fault diagnosis. *Mechanical Systems and Signal Processing*, 138, 2020. cited By 23.

Y. Gao, L. Gao, X. Li, and Y. Zheng. A zero-shot learning method for fault diagnosis under unknown working loads. *Journal of Intelligent Manufacturing*, 31(4):899–909, 2020. cited By 17.

L. Feng and C. Zhao. Fault description based attribute transfer for zero-sample industrial fault diagnosis. *IEEE Transactions on Industrial Informatics*, 17(3):1852–1862, 2021. cited By 21.

Thomas G. Dietterich, Richard H. Lathrop, and TomÁ¡s Lozano-PÁ©rez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71, 1997.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 541–550, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa. Joint optimization framework for learning with noisy labels. pages 5552–5560, 2018. cited By 156.

J. Li, Y. Wong, Q. Zhao, and M.S. Kankanhalli. Learning to learn from noisy labeled data. volume 2019-June, pages 5046–5054, 2019. cited By 50.

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

Igor Labutov, Bishan Yang, and Tom Mitchell. Learning to learn semantic parsers from natural language supervision, 2019.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher
RÉ. Training classifiers with natural language explanations, 2018.

[111] Shikhar Murty, Pang Wei Koh, and Percy Liang. Expert: Representation engineering with natural language explanations, 2020.

[112] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018.

[113] A. Radford. Improving language understanding by generative pre-training. 2018.

[114] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[115] A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[116] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2020.

[117] Microsoft. Turing-nlg: A 17-billion parameter language model by microsoft, 2020.

[118] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[119] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2013.

[120] M.D. Zakir Hossain, F. Sohel, M.F. Shiratuddin, and H. Laga. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys, 51(6), 2019. cited By 67.

[121] X. Lu, B. Wang, X. Zheng, and X. Li. Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing, 56(4):2183–2195, 2018. cited By 91.

[122] Sen He, Wentong Liao, Hamed R. Tavakoli, Michael Yang, Bodo Rosenhahn, and Nicolas Pugeault. Image captioning through image transformer, 2020.

[123] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), December 2015.

[124] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering, 2018.

[125] Z. Zhang, Q. Wu, Y. Wang, and F. Chen. High-quality image captioning with fine-grained and semantic-guided visual attention. IEEE Transactions on Multimedia, 21(7):1681–1693, 2019.

[126] S. Ding, S. Qu, Y. Xi, and S. Wan. Stimulus-driven and concept-driven analysis for image caption generation. Neurocomputing, 398:520–530, 2020. cited By 17.

[127] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision, 2019.

[128] M. Eloiseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In 2013 IEEE International Conference on Computer Vision, pages 2584–2591, 2013.

[129] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.

[130] Yuhao Zhang, Hang Jiang, Yusuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text, 2020.

[131] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations, 2020.

[132] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations, 2020.

[133] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[134] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.
[135] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2020.
[136] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018.
[137] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2, 2019.
[138] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training, 2021.
[139] Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. Towards zero-label language learning, 2021.
[140] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems, 2020.
[141] E. C. Williams, N. Gopalan, M. Rhee, and S. Tellex. Learning to parse natural language to grounded reward functions with weak supervision. In 2018 IEEE International Conference on Robotics and Automation (ICRA), pages 4430–4436, 2018.
[142] Michael P. Brundage, Thurston Sexton, Melinda Hodkiewicz, Alden Dima, and Sarah Lukens. Technical language processing: Unlocking maintenance knowledge. Manufacturing Letters, 27:42–46, 2021.
[143] Karl Ekström and Fredrik Sandin. Fault severity estimation using weak supervision with language based labels and condition monitoring data, 2020.