Zhang and Gao *Chin. J. Mech. Eng.* (2021) 34:71
https://doi.org/10.1186/s10033-021-00587-y

Chinese Journal of Mechanical Engineering

REVIEW

Open Access

Deep Learning-Driven Data Curation and Model Interpretation for Smart Manufacturing

Jianjing Zhang and Robert X. Gao*

Abstract

Characterized by self-monitoring and agile adaptation to fast changing dynamics in complex production environments, smart manufacturing as envisioned under Industry 4.0 aims to improve the throughput and reliability of production beyond the state-of-the-art. While the widespread application of deep learning (DL) has opened up new opportunities to accomplish the goal, data quality and model interpretability have continued to present a roadblock for the widespread acceptance of DL for real-world applications. This has motivated research on two fronts: data curation, which aims to provide quality data as input for meaningful DL-based analysis, and model interpretation, which intends to reveal the physical reasoning underlying DL model outputs and promote trust from the users. This paper summarizes several key techniques in data curation where breakthroughs in data denoising, outlier detection, imputation, balancing, and semantic annotation have demonstrated the effectiveness in information extraction from noisy, incomplete, insufficient, and/or unannotated data. Also highlighted are model interpretation methods that address the "black-box" nature of DL towards model transparency.

Keywords: Deep learning, Data curation, Model interpretation

1 Introduction

Throughout the modern history of humankind, manufacturing has been of central importance to economic advancement. According to statistics from the World Bank, the manufacturing industry contributed 16.8% to international gross domestic product (GDP) in 2018. The contribution to national GDP is as high as 27% in China, representing one of the highest proportions among all nations [1]. Driven by the evolving demands from "mass production" and "mass customization" to "mass personalization" [2, 3], manufacturing has evolved from mechanization and manual operation (Industry 1.0, 18th century) to today's Industry 4.0, where operations take place in complex, digitized cyber-physical production systems (CPPS) that are characterized by sensor-rich monitoring and Internet-enabled edge/cloud computing for in-situ failure root cause diagnosis and future performance prognosis [4–7]. Accompanied by the increasing availability of abundant sensor data [6] as illustrated in Figure 1, analytical and numerical models, and computational infrastructure, the state-of-the-art in real-time condition monitoring, failure root cause diagnosis, and machine remaining useful life (RUL) prognosis has enabled a higher level of automation, robustness, and adaptivity of networked and optimized manufacturing systems [8, 9].

The current wave of innovation in manufacturing, characterized by the concept of smart manufacturing and digital transformation of the factory, is witnessing the convergence of big data [6, 8], artificial intelligence (AI, e.g., machine learning, ML, and deep learning, DL) [10, 11], and the expansion of communication and computational capabilities (e.g., industrial internet of things, IIoT, cloud and edge computing, and graphic processing unit,

*Correspondence: robert.gao@case.edu
Department of Mechanical and Aerospace Engineering, Case Wester Reserve University, Cleveland, OH 44106-7222, USA

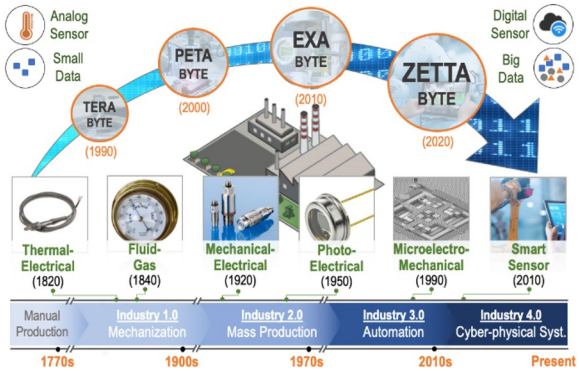

Figure 1 Evolution of manufacturing, sensing technology and big data, adapted from Ref. [6]

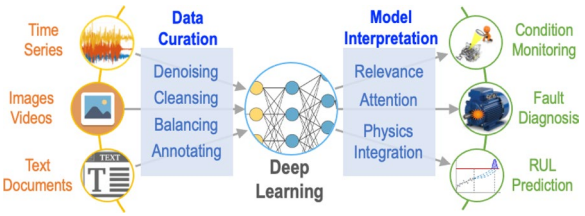

Figure 2 DL-driven pipeline for monitoring, diagnosis and prognosis

GPU) [12–14]. The convergence has transformed a variety of manufacturing practices [15, 16], with condition monitoring, fault diagnosis and RUL prognosis among the most significant beneficiaries [17]. The increased availability of data due to massive deployment of sensors and the rapid advancement of DL have made it possible to gain insight into the mechanism underlying manufacturing operations, leading to enhanced observability of machines and processes [18]. It also made it feasible to associate data with condition-related parameters (e.g., fault and RUL) with unprecedented accuracy [19]. Also, advancements in communication and computational infrastructure have made it possible to carry out data transmission and computation with low latency to satisfy the requirement for real-time operation [20].

The past decade has seen a fast growth in the number of papers on DL-enabled condition monitoring, diagnosis and prognosis, which are comprehensively summarized in several review articles [21–23]. While contributions from DL have been well highlighted, several limitations have also been identified. First, the datasets investigated to evaluate DL algorithms are generally error-free (e.g., no outliers or missing values) and well balanced, with each having sufficient number of samples to fully optimize the DL model parameters [10]. However, in real-world manufacturing scenarios, data errors can occur due to sensing or communication errors. Collecting a large amount of data from faulty equipment for algorithm training is often times infeasible, due to both safety and economic reasons. The consequence is that datasets with error or imbalance can potentially degrade the performance of otherwise high-performing DL models if the level of error or imbalance is high [23, 24]. Second, the datasets used in the reported studies did not require a posteriori (and therefore error-prone) labeling. This is due to the fact that commonly investigated scenarios, such as faults in the inner or outer of a rolling bearing are usually pre-labeled and seeded into the testing equipment before the data is collected. However, in realistic manufacturing scenarios, structural faults or anomalies are not "pre-labeled" because they are not known a priori, and therefore have to be interpreted from the collected data a posteriori. As an example, in additive manufacturing (AM), part surface defects are observed from images acquired after the completion of the AM process, and automated defect annotation (data labelling) is crucial to supporting the relevant diagnostic and/or predictive tasks. Third, the prediction logic of many DL models is generally not interpretable (or transparent) to the users in a physical sense [25]. Without a clear understanding of the data patterns that a DL-based method uses to carry out specific analysis tasks, it is difficult for the readers to establish trust in the performance and outcome of the algorithms.

To tackle these limitations, research on the topics of (1) data curation [26], which aims to improve data quality and provide semantic annotation, and (2) model interpretation, which aims to decipher DL model prediction logic [27], has become an indispensable step before and after the execution of DL algorithms (Figure 2), and thus is gaining attention over the past years. These include: (1) data denoising and cleansing methods that remove data pollution [28]; (2) generative models that recognize patterns underlying the data and synthesize samples to resolve problems arising from small or unbalanced datasets [29]; (3) semantic data annotation that automates the data labeling and contextualization process [30]; (4) relevance analysis methods, such as layer-wise relevance propagation (LRP), which trace the feature extraction processes utilized by neural networks to reveal salient information from the input data for decision-making [31]; (5) attention mechanisms that enable the incorporation of interpretable prediction logic at the design stage of neural networks for enhanced model interpretability [32], and (6) integration of DL and physics to ensure consistency between DL discoveries and existing domain physical knowledge [33].

This paper is motivated by the above identified limitations and aims to fill this knowledge gap by analyzing research outcomes on data curation and model interpretability enabled by DL for operation monitoring, fault diagnosis, and RUL prognosis in manufacturing. As illustrated in Figure 3, data provides the material basis for DL algorithms, and DL algorithms advance the state of technology for monitoring, diagnosis, and prognosis.

The rest of this paper is organized as follows: Section 2 reviews the latest development in data quality assurance to support effective data curation. Section 3 highlights several methods that improve the outcome interpretability of DL models by relating the propagation of reasoning logics through the neural network structure to the physical laws, thereby improving the model interpretability. Section 4 examines several manufacturing applications that have benefited from these techniques. Conclusions and future directions are described in Section 5.

2 Data Curation

As a co-product of manufacturing, data encodes critical information underlying the dynamical behaviors of manufacturing machines and processes, providing the foundation for DL-driven algorithms. Advancements in sensing technologies have resulted in an ever-increasing amount of data acquired on factory floors [6, 34]. The increasing diversity and complexity of data not only pose new challenges for handling data quality issues such as noise, but also amplify additional problems such as data imbalance, outliers, unannotated data, or data with missing values. These become more prominent with the widespread usage of DL. To help ensure success of DL analysis, low-quality data need to be properly curated first [8, 26]. Several representative data curation techniques are summarized in Table 1 and are discussed in detail below.

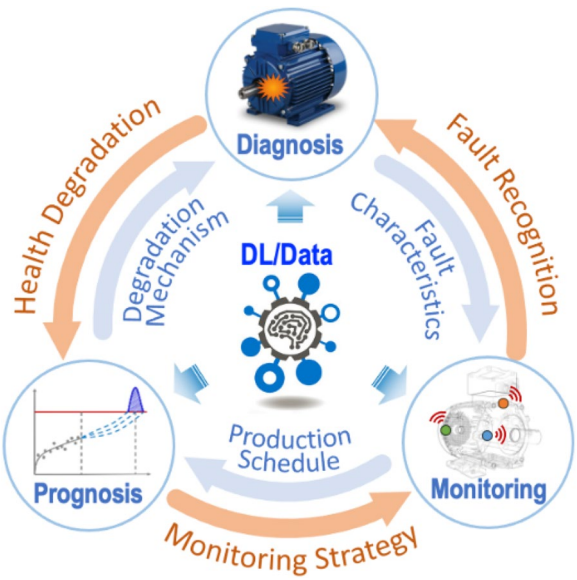

Figure 3 Interactions in an DL-enabled monitoring, diagnosis, and prognosis paradigm

2.1 Data Denoising

The purpose of data denoising is to extract pertinent information (e.g., process and machine dynamics, fault characteristic) from occluding background noise, thereby improving the effectiveness of data analysis [35]. The most adopted approach is to denoise by increasing the signal-to-noise ratio (SNR). Relevant techniques include projection-based method, such as local geometric projection (LGP) [36], and frequency or time-frequency analysis, such as empirical model decomposition (EMD) [37] and wavelet transform [38].

The idea of LGP is that once the data is mapped into a high-dimensional phase space, useful information and noise embedded in data can be decomposed by orthogonal projection into different subspaces. By reconstructing the data from the subspace occupied by the useful information, noise can be removed. In practice [36], the phase space is first segmented into local regions. Within each of the regions, the orthogonal projection matrix is computed by the method of singular value decomposition (SVD). Specifically, only the largest eigenvectors in SVD are used to form the projection matrix, which contains the majority of data variance in the phase space and is likely to capture the useful information. In experimental evaluations, an SNR improvement of 10 dB has been reported. Since LGP does not require prior knowledge about the frequency range of the noise components, it is more convenient to use than filtering-based methods.

The EMD algorithm decomposes data (commonly time series) into a sum of intrinsic mode functions (IMFs). The first IMF represents the highest dominant frequency in the data and the frequency decreases as decomposition proceeds [37]. As a result, EMD represents the data as a sum of frequency bands, and noise removal can be achieved by reconstructing the data from the IMFs that only contains the useful information (e.g., critical frequency components). In practice, suitable IMF range can be determined based on metrics such as mutual information ratio (MIR). For example, the cutoff point can be chosen as the one that leads to the largest increase in MIR, representing the threshold when useful data information is captured by IMF [37].

One of the time-frequency techniques for data denoising is wavelet transform, which is based on thresholding small wavelet coefficients and reconstructing the data using inverse wavelet transform. This is because large

Table 1 Representative techniques for data curation

Data denoising		Ref.	Outlier detection		Ref.	Data imputation		Ref.
Projection-based	Local geometric projection	[36]	Data-level	Autoencoder	[48, 55]	Time series	Recurrent neural network	[62, 65, 66]
Frequency-based	Empirical model decomposition; Wavelet transform	[37–40]						
Noise-assisted	Stochastic resonance	[41, 42]	Model-level	Probabilistic neural network; Temperature scaling; Input perturbation	[56–61]	Image	Convolutional neural network; Hybrid approach	[67, 68]
Data-driven/hybrid	Generative prior; Unrolled optimization	[46, 47]						
Data balancing		**Ref.**	**Data annotation**		**Ref.**			
Data interpolation	Synthetic minority over-sampling technique	[75]	Image annotation	Fully convolutional network; U-Net; Mask region-based CNN	[30] [79, 80]			
Generative model	Variational autoencoder; Generative adversarial network	[76] [29]	Natural language processing	Word embedding; Transformer; BERT	[83–88]			

wavelet coefficients usually contain dominant data components [38]. The threshold value is commonly determined based on the estimated data variance [39]. Using a customized wavelet developed out of the impulse response of a sensor-embedded rolling bearing, the SNR of the bearing's vibration data could be improved by up to eight times as compared to a standard wavelet [40].

An alternative approach to minimizing the effect of noise is stochastic resonance (SR) [41]. The idea is to amplify the critical frequency (e.g., fault characteristic frequency) through the interactions between the data (e.g., time series) and a bistable system [41]. Specifically, when the time series is added to the governing equation of the bistable system, it can be shown mathematically that the critical frequency will be amplified at the system output if the "switch" frequency of the system is tuned to match the critical frequency [41]. In Ref. [42], an adaptive SR strategy is introduced that overcomes the limitation of requiring prior knowledge about the critical frequency as standard SR method does and can accurately pinpoint the fault-related frequency from noisy vibration data that is otherwise undetectable. The technique is applicable to a wide range of critical frequency extraction applications.

More recently, hybrid denoising methods have been reported that take advantage of both the pattern recognition capability of DL and the physical understanding of noise contamination [43, 44]. Specifically, the methods first establish a contamination model $\boldsymbol{y}=G(\boldsymbol{x})$ based on the knowledge of the contaminants, where $\boldsymbol{x}$ is the ideal, clean data representing the physical phenomenon and $\boldsymbol{y}$ represents the measured data, contaminated with noise and determined by G. The contamination model serves as the guidance for data denoising as only the clean data that satisfies the model will be recovered (e.g., improving SNR in a physically meaningful way). Since solving $\boldsymbol{x}$ from $\boldsymbol{y}$ is a generally ill-posed problem (i.e., a large number of $\boldsymbol{x}$ can satisfy the model), the solution $\boldsymbol{x}$ must also be regularized to be consistent with prior knowledge about the data [45]. For this purpose, the method follows the Bayesian theory by iteratively minimizing the inconsistency between the solution $\boldsymbol{x}$ and the contamination model $G(\boldsymbol{x})$, as well as the inconsistency with the prior knowledge, $R(\boldsymbol{x})$. Mathematically, the denoising problem is expressed as:

$$\boldsymbol{x}=\mathrm{argmin}_{\boldsymbol{x}}\left[||\boldsymbol{y}-G(\boldsymbol{x})||_2^2+R(\boldsymbol{x})\right], \tag{1}$$

The outcome of the process is termed the maximum-a-posteriori (MAP) estimation and is graphically illustrated in Figure 4. Essentially, $\boldsymbol{x}$ is iteratively recovered by alternately projecting the intermediate outcome onto the cluster (orange line) that satisfies the contamination model and the cluster (blue line) that satisfies the prior knowledge. At the end, the joint distance between $\boldsymbol{x}$ and the clusters are minimized, leading to the denoised data that is most consistent with the physical contamination knowledge as well as the prior knowledge.

A major challenge to solving Eq. (1) is to analytically formulate $R(\boldsymbol{x})$, given the limitation in prior knowledge to characterize $\boldsymbol{x}$ [46]. To solve this problem, DL-based prior characterization has been developed, and two representative algorithms are described below.

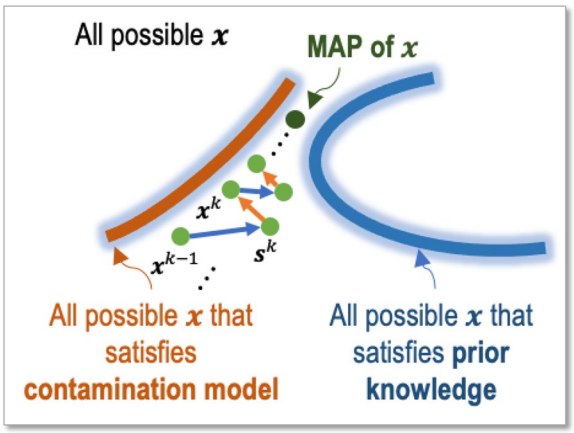

Figure 4 Illustration of MAP estimation

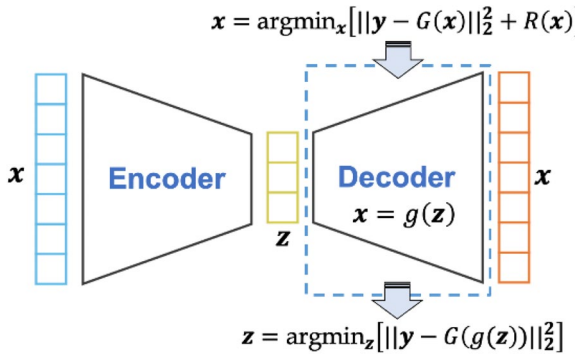

Figure 5 Generative prior for simplification of Eq. (1)

Unrolled optimization. This is motivated by the iterative process of solving Eq. (1) in which each iteration can be concatenated as a layered structure. At each layer (iteration), two steps are carried out: (1) optimization with respect to prior $R(\boldsymbol{x}^{k-1})$ to obtain the intermediate outcome $\boldsymbol{s}^k$, and (2) optimization with respect to contamination model $G(\boldsymbol{s}^k)$ to obtain the new estimation for the next iteration $\boldsymbol{x}^k$. To formulate $R(\boldsymbol{x})$ under limited physical knowledge for characterizing $\boldsymbol{x}$, optimization w.r.t the prior is modeled by a neural network, or $\boldsymbol{s}^k \leftarrow \text{NN}(\boldsymbol{x}^{k-1})$. The complete layered structure is trained in an end-to-end manner [46]. As a result of the iteration process, the underlying structural pattern of the image is learned and improvement of image peak signal-to-noise ratio (PSNR) of 3 dB is reported as compared to other techniques. This denoising technique is also computationally efficient with the processing time per image of around 0.1 s, making it suited for real-time applications.

Generative prior. Sensing data in the form of time series or images often contain information that can be represented in a sparse way under certain representation domain (e.g., the Fourier or wavelet domain) [45]. As a result, such information sparsity in data can enable the formulation of $R(\boldsymbol{x})$ to regularize the solution for Eq. (1), for example, by setting $R(\boldsymbol{x}) = \|x\|_1$. However, such algorithms suffer from low efficiency as the computational cost increases with the square of data dimension [47]. To resolve this limitation, the method of generative prior has been developed [47]. Specifically, it first uses a generative DL model, such as the variational auto-encoder (VAE), to obtain the sparse representation of $\boldsymbol{x}$. Subsequently, by replacing $\boldsymbol{x}$ with its sparse representation $\boldsymbol{z}$ based on the decoder of VAE, $g(\boldsymbol{z})$, the prior term in Eq. (1) can be neglected (as the VAE already enforces the sparsity) and $G(\boldsymbol{x})$ becomes $G(g(\boldsymbol{z}))$, resulting in a low-dimensional problem, as shown in Figure 5, that can be more efficiently solved [47].

A significant advantage of generative prior is its high computational efficiency, since it reduces the computational cost from a quadratic increase with data dimension to a linear increase while achieving comparable denoising results. Table 2 summarizes the advantages and disadvantages of unrolled optimization and generative prior.

2.2 Data Cleansing

The widespread deployment of sensors and increasing complexity of machines and processes have also increased the vulnerability of in-situ data to problems, such as outliers [48] or missing values [49]. Data cleansing addresses these problems by detecting, removing, or correcting outliers and/or missing values among normal samples to improve data quality.

Outlier detection. An outlier (also known as an out-of-distribution, or OOD, sample) generally refers to a data sample that significantly deviates from the expected data pattern associated with the physical phenomenon that it represents [50]. Common outlier detection methods can be classified into the following categories [48]: (1) statistical methods, which detect outliers based on the likelihood of seeing the sample under the assumed data distribution (e.g., Gaussian) [51]; (2) distance-based methods, which assume that within-distribution samples are located in a dense region in the data space while outliers are located further away [52]; (3) density-based methods, which are based on the assumption that the data distribution should be similar around within-distribution samples and significantly different around outliers [53]; and (4) cluster-based methods, for which clustering techniques are applied and outliers are detected as samples not in the neighborhood of any clusters [54]. Despite ongoing progress, these methods are limited when handling high-dimensionality and nonlinearity [48]. Most

Table 2 Comparison between hybrid denoising techniques

Technique	Advantage	Disadvantage
Unrolled optimization	Suited for data with sparse or non-sparse structure; End-to-end learning	Denoising progression not accessible; Intermediate tweak not possible
Generative prior	Complete denoising progression accessible; Allow intermediate tweak	May not perform well for data with non-sparse structure

recently, DL-based methods have been reported, which focus primarily on enhancing the separability between within-distribution samples and outliers to facilitate the determination of outlier detection threshold.

One DL-based method for outlier detection is to use autoencoders (AE) to find latent features by projecting data into network layers with progressively reduced dimensions. The basic idea is that data reconstruction error is expected to be small for within-distribution samples and large for outliers [55]. This is because within-distribution samples are generally well clustered in the data space while outliers are likely to be randomly scattered. Therefore, the gradients induced by outliers during AE training are likely to be mitigated by the within-distribution sample gradients. Accordingly, the weights of the related layers of the network will be updated using predominantly those of within-distribution samples. Consequently, data reconstruction error will be minimized primarily for within-distribution samples rather than outliers [55]. It is reported in Ref. [48] that by training an AE using only within-distribution samples, the reconstruction error serves as a good indicator for outlier detection based on a simple 3-standard deviation threshold.

For datasets in which within-distribution samples are polluted with a comparatively small number of outliers (which is not ideal as compared to the scenario in Ref. [48]), an iterative approach has been developed in Ref. [55], where two steps are involved at each iteration: (1) discriminative labeling, which estimates within-distribution samples from the mixed data based on the reconstruction errors at that iteration, and (2) reconstruction learning, which updates the AE to reduce the reconstruction error for the identified within-distribution samples. This iterative approach has shown to gradually converge to the level of performance comparable to that of the AE trained using within-distribution samples only. The advantage of AE-based approaches is that the outlier detection is done completely at the data level and is therefore task-agnostic.

Besides AE, probabilistic neural networks have also been investigated for outlier detection. The idea is to quantify the uncertainty associated with the predicted outcome (e.g., type of a structural fault), for which a high uncertainty suggests that the input can potentially be an outlier. Commonly used techniques for neural network probabilistic formulation include ensemble learning [56] and Monte Carlo (MC) methods [57, 58], which assume that multiple DL diagnostic/prognostic models exist for any task, and uncertainty is quantified based on the prediction entropy across all model outputs.

One of the latest developments is a multi-head network [59] (Figure 6), which consists of shared layers at lower part of the network structure, while diverging to multiple classifiers at the upper part. In this work, a distributed gradient is developed for DL network training, in which only a fraction of the gradient is used to update the classifier with the best performance at each iteration to increase accuracy, while the remaining fraction of gradient flows through the other classifiers to improve generality for unseen data. As the multi-head network is trained only on within-distribution samples, the probability that the network can produce similar uncertainty for both within-distribution samples and outliers is expected to drop quickly as M becomes large [59]. The result is that it facilitates the determination of an uncertainty-based threshold for outlier detection. Table 3 summarizes probabilistic network techniques.

DL-based outlier detection methods, e.g., based on temperature scaling and input perturbation, have been developed for pretrained DL classifiers [60]. In this example, temperature scaling refers to calibrating the scaling factor in the *softmax* function in the classifier. Mathematically, larger scaling factors lead to larger *softmax* scores for within-distribution samples than outliers. Perturbation refers to preprocessing the input by adding a perturbation term calculated based on the gradient of the *softmax* function with respect to the input, which tends

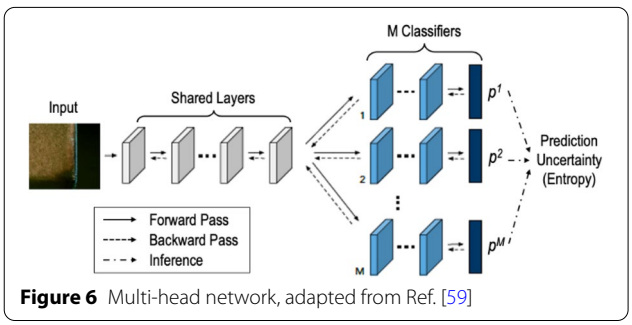

Figure 6 Multi-head network, adapted from Ref. [59]

Table 3 Comparison of probabilistic network techniques

Technique	Advantage	Disadvantage
Deep ensemble	Highest model diversity; Easy to parallelize	Multi network structure; Most parameter tuning
MC dropout	Single network structure; Easy to parallelize	Potentially unstable when applied throughout large network
Multi-head network	Single network structure; Fewest parameter tuning	Lowest model diversity

to have a larger value for normal samples than outliers based on experimental observations. Therefore, by adding this gradient as perturbation to the input, the *softmax* scores with respect to within-distribution samples are likely to be greater than the outliers. Collectively, temperature scaling and perturbation enhance the separability of the *softmax* scores between the within-distribution samples and outliers, allowing a threshold to be set up flexibly for outlier detection.

The work reported is further extended in Ref. [61]. Rather than relying on a single score at the final network layer for outlier detection, a Mahalanobis distance-based confidence score is calculated in this work for all layers based on the layer-wise features, and the final score is represented by a weighted sum of all scores.

Data imputation. Besides outliers, another frequently encountered data quality issue is missing values, which is commonly caused by sensing or communication errors [62]. Imputation of missing values improves data quality by filling data gaps. Typical data imputation methods are based on statistics, such as linear interpolation or autoregressive modeling [63, 64]. However, these methods often require strong assumptions (e.g., linearity) with respect to the data generation process, making them less effective to more complex data in which the assumptions do not hold.

One of the actively researched topics recently is time series data imputation based on recurrent neural networks (RNN) and their variants, such as long short-term memory (LSTM) and gated recurrent units (GRU) [62, 65, 66]. These techniques capture the non-linear time evolution pattern underlying the data for estimation of missing values. The main idea is to use a rolling window that progressively predicts the missing values by analyzing the data sequences that immediately precede these gaps.

The newly developed methods mainly differ in the approach with which the missing value is handled at the network input. In Refs. [65, 66], the input to the network is designed as a weighted sum of the observed value at the current step and the predicted value from the previous time step, when the observed value is available at the current step (Figure 7). When data is missing at the current step, these methods use the predicted value from the previous time step directly as network input at the current step.

In contrast to Refs. [65, 66] in which the missing values are replaced by the predicted values of the same sequence, a different approach has been developed in Ref. [62] that aligns a separate, auxiliary "source" sequence to the "target" sequence with missing values. Then, it replaces the missing values with the corresponding values in the source sequence that are adjusted by the mean values of both sequences.

The development of DL, especially the convolutional neural networks (CNNs) that are specialized in image processing, has advanced image imputation that was previously considered challenging [67]. One strategy of image imputation is to train an end-to-end CNN that learns the direct mapping between the images with missing pixels to the corresponding, complete images.

However, this approach has shown to produce unsatisfactory results that tend to blur out the regions of missing pixels rather than learning the underlying structural pattern. This is mainly because the network training is guided by the reconstruction error that is averaged over all pixels [67]. One remedy is to treat image imputation as a special "denoising" process and implement the hybrid approach as described in Section 2.1 [68]. The other approach, which has attracted increasing attention, is to

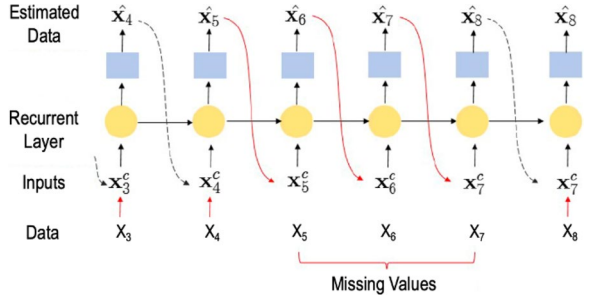

Figure 7 Time series imputation with RNN, adapted from Ref. [66]

add an "adversarial" training loss that penalizes the CNN when the recovered image does not resemble the original image [67]. Adversarial approaches are gaining popularity for high-fidelity data generation. In the context of data curation, one of its most significant applications is in data balancing, which will be described in the next Section.

2.3 Data Balancing

Data balancing addresses the need for having sufficient faulty data to produce a balanced dataset when training neural networks to minimize learning bias [69]. This is particularly important for tasks involving data classification, e.g., in fault diagnosis [70]. However, operating machines under faulty conditions just for the purpose of collecting faulty data for algorithm training is not feasible in real-world applications, since machines are required to maintain normal operating conditions to ensure product quality. One approach to remedy this issue is transfer learning, which allows to transfer the diagnostic model or feature from a source domain in which faulty data is sufficient to the target domain which lacks sufficient data for network training. The latest development of transfer learning has been summarized by a series of review and technical papers [71–74]. The other research direction has been typically relying on high fidelity synthetic data to augment the number of samples and improve data balancing, which is the primary focus of this section.

Early works of data synthesis mainly relied on data interpolation, e.g., synthetic minority over-sampling technique or SMOTE [75]. The idea is to first select randomly a minority class sample and one of its neighbors. Then, a synthetic example is generated as a convex combination of the two chosen samples. While the method works well for low dimensional data such as process parameters and machine settings, SMOTE and related techniques cannot capture the complex characteristics as commonly shown in high dimensional data, such as high-speed time series or images [24].

A more systematic approach is made available by the AE-based generative model [76]. The idea is to learn a latent representation of the existing data and its underlying distribution using the AE and generate synthetic data from that distribution via data sampling. However, as the generation process is not guided by any supervision with regards to the quality of the outcome, the result of the AE can be dissatisfactory. A major breakthrough came with the development of the generative adversarial networks (GAN) [29], which is a specialized DL architecture that allows for high fidelity data synthesis with supervision.

The main structure of GAN is composed of a generator and a discriminator, as shown in Figure 8. The GAN operates on the premise that the generator can be trained to convert a random noise vector into synthetic data (e.g., time series or image) that closely resembles the real data. The performance of the generator is evaluated by a discriminator, which aims to correctly classify an input as either "real" or "generated". Specifically, the discriminator randomly takes as input either the real data or the synthetic data produced by the generator and outputs a scalar representing the probability that the input data is "real". Conversely, the objective of the generator is to generate synthetic data that are indistinguishable from the real ones and deceive the discriminator. This is realized through the training of the GAN, in which the generator and discriminator play a minimax game: the generator will try to minimize the discriminator's accuracy, while the discriminator tries to maximize it. The final training outcome will be an equilibrium point, at which the discriminator will no longer be able to distinguish the generated data from the real ones, and the generator can no longer synthesize "better" data as the discriminator no longer provides useful feedback for further improvement. At this point, the generator is capable of synthesizing data with high-fidelity to augment the number of samples in the minority classes and reduce dataset imbalance.

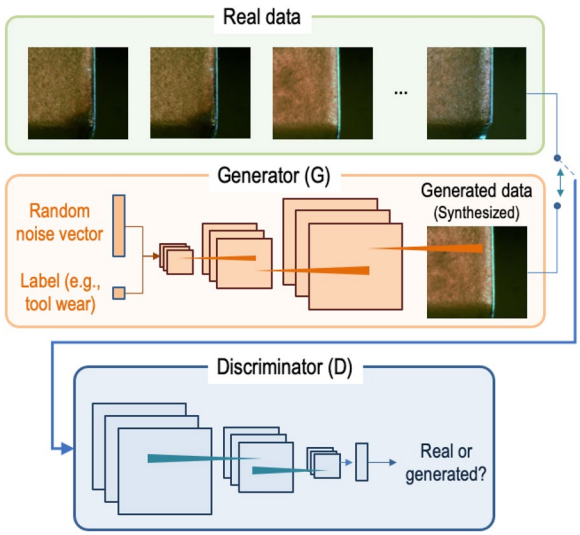

Figure 8 GAN for image data synthesis, adapted from Ref. [29]

2.4 Data Annotation

Data annotation is about associating data with proper contextual information under which it is acquired by using proper semantic format. The use of image data, which contain rich spatial information that are not captured in time series signals, has become an important aspect of DL-based techniques [6]. At the same time, annotation and labeling of the image regions of interest (ROIs) semantically, which are indicative of critical

information on the condition of the machine and process of interest, has become a challenge due to the lack of techniques to effectively parse abstract image patterns. Traditionally, the method of thresholding has been extensively applied [77, 78] with the goal to set a pixel intensity threshold value that separates the ROI from the remaining regions. However, the technique assumes that: (1) all pixels with intensity values within an established range belong to the same ROI, and (2) ranges for different ROIs are non-overlapping. In reality, both of the assumptions are often times invalid.

Built upon the image analysis capability of the CNNs, fully convolutional networks (FCNs) have been developed for the purpose of image semantic annotation [30]. A typical FCN is constructed by a pair of CNNs: an encoder and a decoder. The encoder, consisting of convolutional layers and pooling layers, distills essential information from the input image that is most relevant to semantic annotation. The decoder, which consists of upsampling operations and a classification layer, generates the annotated images.

At the classification layer, instead of producing a single probability distribution indicating the probabilities with which the image belongs to different categories (e.g., fault types) as in standard CNNs, the FCNs utilizes the *softmax* function at the pixel level, generating for each pixel a probability distribution that the corresponding pixel belongs to different ROIs (e.g., defect, tool wear) or non-ROI. Each pixel is then classified as the ROI or non-ROI with the highest probability.

Over the years, various semantic annotation methods built upon FCNs have been developed. The two widely used techniques are U-Net [79] and mask region-based CNN (RCNN) [80]. The basic structure of U-Net is shown in Figure 9. Compared to FCN, U-Net is designed in a symmetric fashion with the progressive upsampling layers in the decoder that match the encoder layers. In addition, the corresponding layers in the encoder and decoder are connected via skip connections for network training [82].

In mask RCNNs, instead of analyzing the image as a whole, a region proposal network (RPN) is attached before the FCN to allow the network to first focus on small regions that potentially contain ROI, before carrying out FCN-based annotation [80]. Once annotated, the image can be not only used for direct diagnosis purpose (e.g., surface defect diagnosis, tool wear evaluation), the information extracted from the ROIs, such as area and geometry features, can also serve as the input to the DL model for other predictive tasks.

Besides image data, semantic annotation and labeling of text is also attracting increasing attention, as reflected in the development of natural language process (NLP)

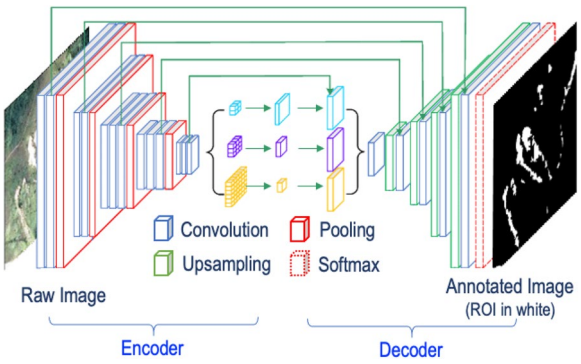

Figure 9 Structure of U-Net, adapted from Ref. [81]

techniques [83]. The fundamental problem of annotating text in manufacturing, such as maintenance logs and inspection reports, is to convert text into computable representations while maintaining their semantic information. One of the most widely investigated techniques is embedding, which refers to the mapping of words to their representations in a high-dimensional space [84]. To establish domain-specific embedding and annotate manufacturing context, a key step is to train the embedding mapping in order to maximize the consistency (quantified as inner product) between an individual word and its existing manufacturing context while minimizing the consistency with the non-existing contexts. This allows word semantics to be implicitly encoded in their respective representations based on the number and frequency of the shared contexts, and the semantically similar words are expected to have similar representations.

Once the embedding is established, DL-based language models can be trained to decompose interested texts into interpretable labels for diagnosis and prognosis [85]. Common DL-based models include 1-D CNN and RNN (and its variants), both of which allow the analysis of sequential patterns, which is a prerequisite for language understanding [86]. Recently, more dedicated, and pre-trained language models have emerged, such as transformers [87] and bidirectional encoder representations from transformers, or BERT [88]. These models generally consist of a stack of encoders and self-attention modules, which allow efficient analysis of relationship among different words in the inputs and outputs. For example, the key element in the transformer is a {query, key, value} tuple that is computed for each word, and the association among different words is quantified as the inner product of their corresponding tuple values. These pre-trained models provide a backbone for general language analysis, which can be adapted for specific purpose through

fine-tuning with a task model [89]. In Figure 10, the general flowchart for text annotation/labeling is shown.

3 Model Interpretation

The capability of DL algorithms to automatically learn characteristic features from data to minimize errors in diagnosis or prognosis and reduce the need for extensive human knowledge is often credited as the advantage of DL [10]. However, the prediction logic of DL-based algorithms is generally not clearly interpretable in a physical sense, thus making it difficult to establish trust from the users in the model performance. To address the need for understanding the working mechanisms of DL and facilitate its broad acceptance, several representative techniques that improves the interpretability of DL models are highlighted in this section, and are summarized in Table 4.

3.1 Relevance Analysis

One major research activity towards improving DL model interpretability is to determine the association, or relevance of each input with the output. For tasks in which different regions of the input have semantic meanings, such as images or frequency spectrums, the prediction logic of the DL model can then be evaluated and verified against human knowledge by means of relevance analysis. Representative techniques include saliency maps [90, 91], deconvnet [92], layer-wise relevance propagation (LRP) [31].

Saliency maps. Saliency maps provide the ranking of the individual inputs based on their influence on the network decision [90]. The idea is to approximate the network in the neighborhood of the inputs using a Taylor expansion and quantify the sensitivity of the decision relative to changes in each input. Once the sensitivity of each input is computed, a sensitivity heatmap can then be generated to visualize the input regions that most influence the network decision, as shown in Figure 11.

Deconvnet. An approach similar to the saliency maps is deconvnet [92]. Intuitively, a deconvnet is a network that uses the same kernels and pooling operations as the standard CNN that carries out the decision making, but in a reversed direction. For example, instead of computing a weighted sum of image pixels based on convolutional operations to generate features, it distributes the features backwards to the individual pixels. In practice, the deconvnet is attached to its corresponding CNN, forming a U-shape structure as shown in Figure 12. At each convolutional layer, each individual neuron is evaluated by first setting all other neurons in the layer to zero. Next, the generated feature maps are passed as input to the attached deconvnet layer to reconstruct its association with the layer beneath that produces the output of the selected neuron. This process is then repeated until the associations from individual image pixels are obtained.

LRP. Different from saliency maps and deconvnet in which the relevance is determined via network weights, the concept of LRP is to redistribute the network's outcome backwards using local distribution rule based on

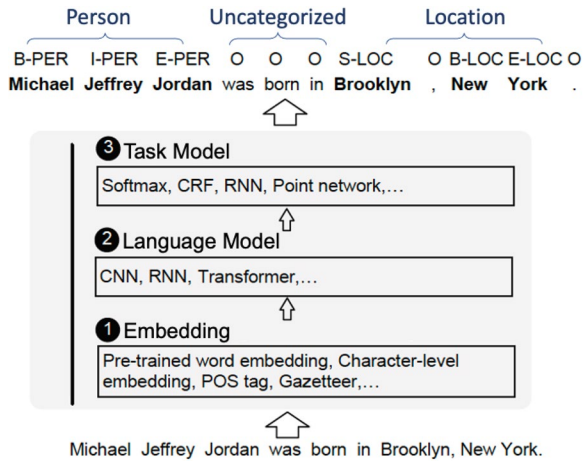

Figure 10 Flowchart of NLP in text annotation, adapted from Ref. [86]

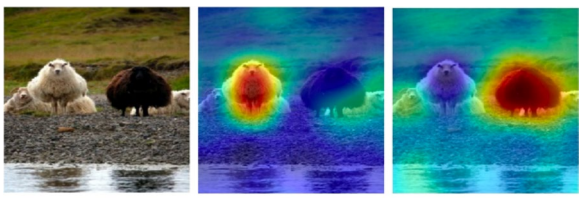

Figure 11 Example of saliency maps, adapted from Ref. [91]

Table 4 Representative techniques for DL model interpretation

Relevance analysis		Ref.	Interpretable structure		Ref.
Weight-based	Saliency maps; Deconvnet	[90–92]	Attention mechanism	Bahdanau attention; Luong attention	[32, 94]
Weight and activation-based	Layer-wise relevance propagation	[31]	Physics-integrated	Physical model compensation; numerical parameter calibration; physics-informed training	[99, 101–104]

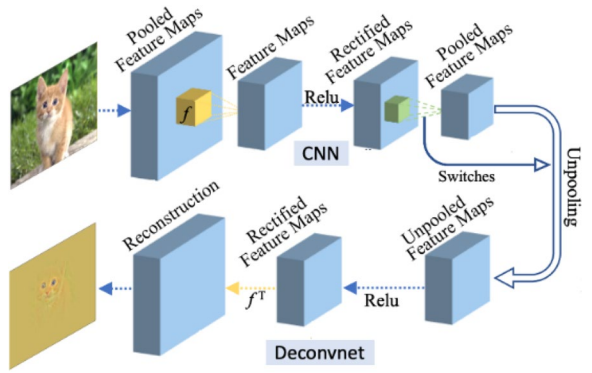

Figure 12 Structure of deconvnet, adapted from Ref. [93]

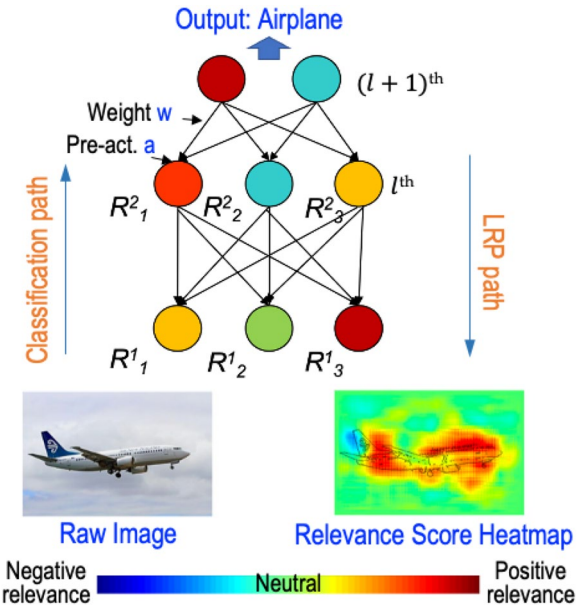

Figure 13 Illustration of LRP, adapted from Ref. [31]

both network weights and activations, until it assigns a relevance score to each individual input [31]. Specifically, the relevance scores are propagated starting with the final layer in the neural network. The scores are then propagated to the early layers in a way that the sum of the score is preserved at each layer. Mathematically, the local distribution rule is expressed as:

$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j} + \epsilon \text{sign}(\sum_{i'} z_{i'j})} R_j^{(l+1)}, \tag{2}$$

where $R_i^{(l)}$ is the relevance score for the ith neuron in the lth layer. sign() represents the sign function, ϵ is a numerical stabilizer, and z_{ij} is the contribution (neuron pre-activation value times the corresponding weight) of the ith neuron in the lth layer to the jth neuron in the $(l+1)$ th layer. Eq. (2) indicates that the relevance scores can be positive or negative.

Using the diagnosis of machine fault types as an example, a positive-valued relevance score represents the evidence for the diagnostic decision, while a negative-valued score indicates the evidence against the diagnostic decision. By analyzing relevance scores propagated to the input of the DL model, the input regions assigned with high positive scores can be interpreted as an indication that the corresponding regions significantly contribute to the diagnostic decision, and vice-versa for regions with high negative scores. An illustrative example of recognizing major structural features of an airplane image using LRP is shown in Figure 13.

3.2 Attention for Interpretable Structure

Different from the relevance analysis for interpreting a trained DL model, the attention mechanism is a structure incorporated into the network design to establish the prediction logic that is inherently interpretable [32, 94]. The design of the attention mechanism comes from domain knowledge and is most suited for capturing the dynamic relationship of the processes. For example, in the sequential printing process of additive manufacturing (AM), the layer-wise influence on the final part property induced by thermal activities can be different for different parts, for the same number of printed layers. In addition, the number of total printed layers of a part also varies as the setting of layer height changes. This means that the prediction logic of the DL model is required to be adaptive to the variations. This poses a challenge for the standard neural network, since it uses network weights to encode the relationship that become fixed values after training and are independent of the input. Therefore, they cannot capture the dynamic relationships.

Attention mechanisms provide a means to alleviate this limitation by enabling dynamic weight generation based on the specific context in the process. Specifically, the weights are generated by a separate context network that takes the relevant context as the input. For example, to compute the thermal influence of a particular printed layer i to the part property in AM, the context can include the machine settings, material property and the thermal activities of the adjacent printed layers. The related adaptive weights w_i, i=1, 2,..., N, are computed with the corresponding unnormalized weights being first generated through a dense layer in the context network. Subsequently, to ensure that the relative influences of all printed layers add up to one for interpretability, a *softmax* layer is incorporated to normalize

the generated weights. Mathematically, the process can be expressed as:

$$w_i^{'} = W_{\text{attn,dense}} x_{\text{context}}, \tag{3}$$

$$w_i = \frac{\exp(w_i^{'})}{\sum_{j=1,2,\ldots,N} \exp(w_j^{'})}, \tag{4}$$

in which $W_{\text{attn,dense}}$ represents the weights of the dense layer in the context network, x_{context} are the inputs to the context network, and $w_i^{'}$ is the unnormalized weights. This version of attention mechanism is called Bahdanau attention, named after its creator [32]. Luong et al. [94] later improved the Bahdanau version by replacing the dense layer in the context network with inner product to improve computational efficiency.

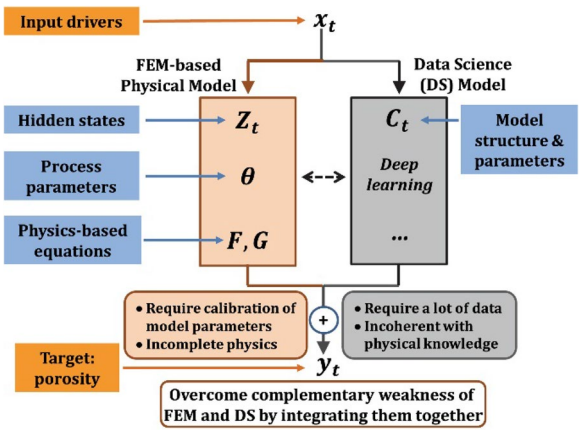

Figure 14 Integration of physical model and neural network [99]

3.3 Integrating Neural Network with Physics

Although the attention-based network structure provides a pathway to capturing interpretable relations, it does not guarantee that the discovered relation is consistent with the underlying physics of the machine or process. This is because during network training, the update of the network weights, which determine the exact relationship between the inputs and the outputs that the network represents, is guided by the prediction error only. Therefore, there is no guarantee that the network will converge to the relation that is also physically meaningful. The "spurious" relations discovered by the neural networks often cannot generalize to unseen scenarios and can be detrimental for critical tasks such as machine fault diagnosis and performance prognosis [95]. In an effort to remedy this issue, the integration of neural networks and physics is attracting increasing attention in recent years. Three representative approaches are described in this section.

The first approach is based on the fact that physical models underlying machines and processes often involve assumptions and simplifications [96]. For example, physical predictive models for machining processes such as grinding and milling often include the effects from major process parameters only, such as depth of cut, while having limited capability to incorporate other factors, such as the operating conditions [97, 98]. Therefore, while these models can generalize well, their predictive accuracy is often lacking due to the incompleteness of physical phenomena that are accounted for. On the other hand, neural networks have the advantage of learning the deviation of the physical model from real-world observations by leveraging the in-situ sensing data that reflects the operating conditions. Therefore, by using a neural network to compensate for the deviation of the physical model (Figure 14), the complementary strength of the two can be synergistically integrated [99], leading to improved predictive accuracy as compared to physical models alone, and enhanced network capability to generalize as compared to pure data-driven methods.

The second approach leverages neural networks to numerically calibrate the unknown parameters in the physical models that are time-consuming or difficult to calibrate experimentally. As an example, Paris's law for fatigue crack propagation is expressed as: $\mathrm{d}a/\mathrm{d}t = C\Delta K^m$, in which both C and m are unknown parameters that require experimental testing to determine [100]. In addition, the stress intensity range ΔK also depends on the parameter that is related to the part geometry [100]. In this scenario, a neural network can be used to calibrate these unknown model parameters by associating them to the in-situ sensing inputs, thereby preserving the physical intuition of the model while alleviating the requirement for extensive experiment parameter calibration [101].

The third approach involves adding physical constraints during the network training process, such that the relation discovered by the network will be consistent with the physical domain knowledge. The physical constraints can be in the form of analytical equations or experimentally verified trends. For example, machine performance degradation should be monotonic, therefore, the performance predicted by the neural network should be monotonically decreasing as the operation cycle increases [102]. With the physical constraint, the network is forced to follow the physical equation or trend imposed by the constraint and can generalize well outside of the range of training data [103, 104]. Table 5 summarizes the comparison of these three approaches.

Table 5 Comparison of physical model and neural network integration approach

Approach	Suited for	Not suited for
Physical model compensation	Explicit physical model with known parameters	Implicit physical model; Empirical and experimental trend
Model parameter numerical calibration	Explicit physical model with unknown parameters	Implicit physical model; Empirical and experimental trend
Physics-informed network training	Explicit and implicit physical model; Empirical or experimental trend	Other form of domain physical knowledge

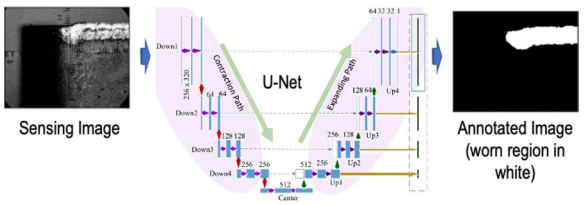

Figure 15 U-Net for tool wear image annotation, adapted from Ref. [106]

4 Application Highlights

The ultimate goal of data curation and model interpretation is to improve data quality to ensure effectiveness and reliability of DL-based analysis and improve interpretability of DL-based methods. In this section, several applications in manufacturing that have benefited from data curation and model interpretation techniques are highlighted.

4.1 Condition Monitoring of Machines and Processes

Condition monitoring refers to monitoring the quality and performance-related variables of a machine or process to identify significant deviation that is indicative of potential faults or anomalies [105]. One prerequisite for effective condition monitoring is the assurance of data quality such that the important variables can be faithfully reflected. Recent developments in data curation have contributed to data quality improvement in condition monitoring; two representative examples based on data annotation and data imputation are presented as follows.

In many condition monitoring scenarios, although part quality related information is captured in the sensing data, it is often not directly computable and requires significant manual examination. One example is machine tool wear images. While a human worker can delineate regions of tool wear and consequently estimate the tool's condition from images, automated annotation that saves time has long been missing until the recent development of DL-based methods. In Ref. [106], Miao et al. presented a U-Net based approach for tool wear annotation in cutting process, as shown in Figure 15. Considering that the worn region of the tool typically covers only a small portion of the image, making the numbers of the worn region and normal region pixels unbalanced, a Matthews Correlation Coefficient (MCC)-based loss function is designed to alleviate the effect of data imbalance during the U-Net training. Effectiveness of the developed method has been confirmed in the experimental evaluation, achieving over 95% accuracy in tool wear ROI annotation.

With the increasing variety of data sources, data with missing values has become a frequent phenomenon, which negatively impacts the effectiveness of condition monitoring and potentially lead to faults or anomalies going undetected. DL-based data imputation has provided an effective mean of dealing with this issue. In Ref. [62], a bi-directional LSTM-based method has been developed for time-series imputation in energy consumption monitoring. In addition to the rolling-window strategy and auxiliary sequence alignment as described in Section 2.2, this work also features a bi-directional strategy that allows the missing values to be estimated based on two estimators in order to further improve the accuracy and robustness [107]. Experimental evaluation demonstrated a clear advantage of the developed method over traditional techniques in terms of imputation accuracy (root mean squared error reduced from 170.8 W to 90.3 W), especially in the situation of continuous missing values.

While common machine and process variables such as temperature can be measured in real-time, other important variables may not be directly measurable in-situ. They often only become available at the end of the process through post-process inspection. Therefore, predictive models are required to infer these variables from in-situ sensing data for timely detection of faults or anomalies. While DL-based predictive modeling for condition monitoring has been an active research field, the recent development of interpretable DL models has the potential to facilitate their widespread acceptance.

In Ref. [102], a bi-directional GRU with physics-informed network training has been developed for tool wear monitoring in milling. The input to the network at each step consists of statistical, frequency, and time-frequency features extracted from real-time force and vibration sensing data. The tool wear prediction at the

network output is regularized by a physics-based loss function, which penalizes the network training when any pair of predicted tool wear values do not monotonically increase with the increasing cycle number. This penalty guides the network weight update to achieve maximum physical consistency. Experimental evaluation has demonstrated that the integration of neural network with physics not only eliminated the physical inconsistency in tool wear prediction, but also consistently achieved higher predictive accuracy as compared to the networks without physics-informed training.

In Ref. [108], an attention-based AM process monitoring and part predictive modeling method has been developed as shown in Figure 16. The attention mechanism is designed to capture the dynamic layer-wise thermal influence on the AM condition and part property. To generate the dynamic weights for each printed layer, the machine settings, material properties and the thermal activities up to that printed layer are selected as context for the attention mechanism. Evaluation results have shown that larger weights are generated for the layers printed later in the AM process as compared to the earlier layers and the trend is consistent under different AM machine settings.

4.2 Diagnosis of Fault and Anomaly

DL-enabled diagnosis requires associating condition-related features extracted from sensing data to the corresponding fault or anomaly root cause. To handle a large number of fault types with multiple fault severity levels, DL models often require a large number of training samples in order to fully optimize. In real-world applications, the collection of faulty data is often limited by production and safety constraints.

Recently, the method of data synthesis based on GAN to alleviate the lack of high-fidelity data for model training has shown great potential. As an example, the effectiveness of GAN in synthesizing sensing data features related to a faulty motor is presented in Ref. [109]. Specifically, the features evaluated are the IMFs from the EMD. The evaluated motor conditions include normal condition, inner race and outer race faults of a motor bearing, and broken rotor bar. In addition, different data imbalance ratios (from 2:1 to 16:1) between the normal and faulty datasets are considered. For data synthesis, both the generator and the discriminator of the GAN are formulated as fully connected networks, with the synthesized features serving as an input to another fully connected network for motor condition diagnosis. It has shown that the GAN-based method has consistently outperformed SMOTE-based approach in terms of fault recognition accuracy.

In Ref. [110], an auxiliary classifier GAN, or ACGAN, has been presented to incorporate the classification capability for the fault types directly into the discriminator, as show in Figure 17. In this work, vibration signal from the motor is used as the target for data synthesis. To analyze the temporal pattern embedded in the time series data, both the generator and the discriminator are constructed as stacked 1-D CNNs. Evaluated on a set of six different motor conditions, including normal, stator winding defect, unbalanced rotor, inner race bearing fault, broken rotor bar and bowed rotor, the GAN-based method has shown significant improvement in diagnosis accuracy for the dataset with a 2:1 imbalance ratio. A similar work has been reported by Wang et al., in which they investigated synthetic vibration signals for gearbox fault diagnosis [111].

In addition to machine fault diagnosis, GAN has also been investigated for non-compliant tool condition detection. In Ref. [112], synthesis of wavelet time-frequency spectrums using GAN has been investigated for non-compliant tool detection in milling. Different from the previous works in which the classifier is either

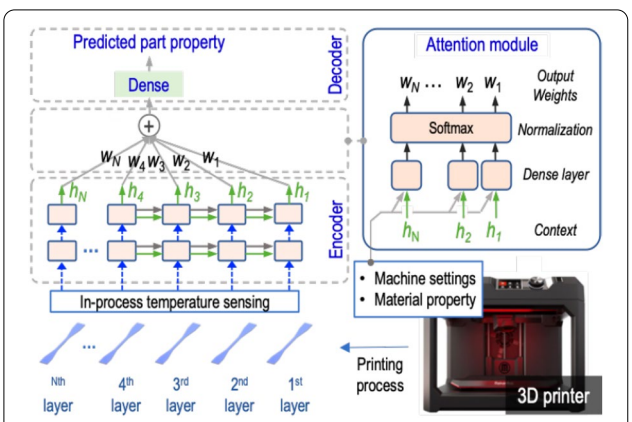

Figure 16 Attention-based AM monitoring and part property prediction [108]

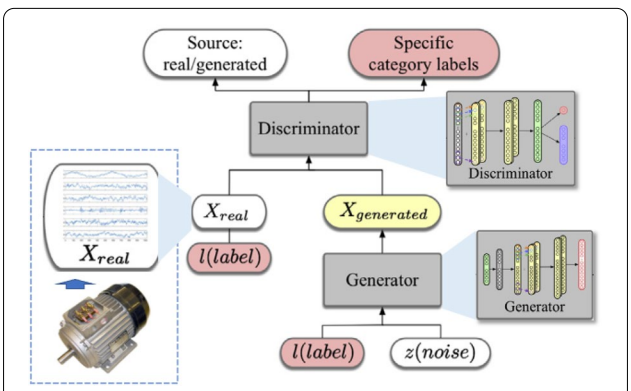

Figure 17 Auxiliary classifier GAN, adapted from Ref. [110]

constructed separately or incorporated with the discriminator, the generator of the GAN is inverted to perform non-compliance detection in this work, resulting in a 25% improvement in accuracy for the dataset with 2:1 imbalance ratio.

Besides data balancing, the capability of image semantic annotation based on FCN has also enabled process condition monitoring and anomaly detection that would otherwise require significant human intervention. One of the successful applications is AM. In Ref. [113], a comprehensive investigation of layer-wise anomaly detection and evaluation based on image semantic annotation has been reported for three AM technologies: laser fusion, binder jetting and electron beam fusion. A total of 12 common surface anomalies, such as spatter and recoater streaking, have been evaluated. The network structure of the developed annotation method is built upon the U-Net while being enhanced by multi-stream analysis of the image at multiple scales. Evaluation of the developed method has shown that it can be executed in real-time on image data of resolution up to 3672×5496 pixels. The performance of the developed method is also shown to be superior to that of previous state-of-the-art in terms of anomaly ROI segmentation accuracy. A similar work has been reported for over-extrusion detection in the fused filament fabrication process [114].

In addition to the AM processes, Wu et al. developed a solder joint annotation method based on mask RCNN, as shown in Figure 18, which allows to locate, segment, and classify solder joint regions at the same time, which is critical for quality assurance in printed circuit board (PCB) manufacturing [115]. Due to the limitation in training images for solder joint, the method of transfer learning has also been investigated, which transfers a pre-trained network using the large-scale "common objects in context" dataset (by Microsoft) for the purpose of solder joint annotation. Four defective joint conditions are evaluated. The mask RCNN-based method has shown to achieve 100% condition recognition accuracy and 97.4% ROI segmentation accuracy.

Recent development of text annotation has also contributed to the field of fault diagnosis. In Ref. [89], a text semantic decomposition method based on BERT has been described that extracts fault-related information, such as equipment, fault, cause, and solution directly from text documents. In addition to BERT as the pre-trained backbone language model, the developed method also features a stacked LSTM and conditional random field (CRF) [116] as a task model. The whole network structure is shown in Figure 19. In experimental evaluation, the developed method has achieved state-of-the-art performance, outperforming the method based on LSTM + CRF in terms of annotation accuracy for "equipment"

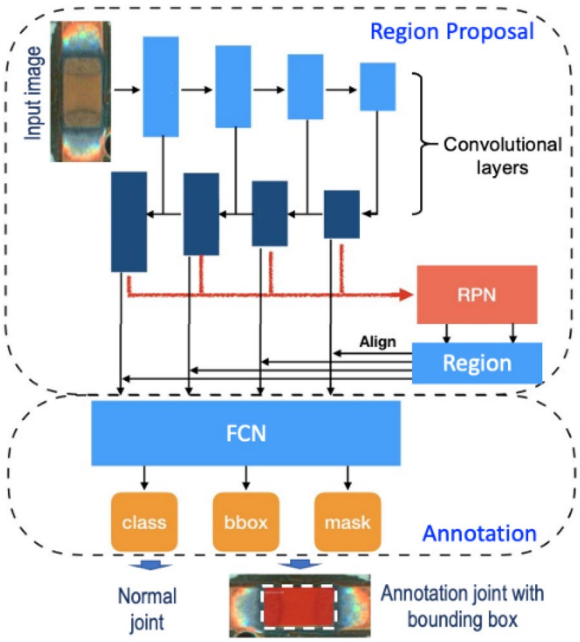

Figure 18 Mask RCNN for solder joint annotation and defect detection, adapted from Ref. [115]

(from 89.7% to 83.7%), "fault" (from 61.6% to 53.8%), "cause" (from 77.4% to 76.3%) and "solution" (from 44.2% to 36.7%). On the other hand, it is noted that the absolute performance is still far from satisfactory. For example, the accuracy of extracting "fault", "cause" and "solution" information is significantly worse than that of the "equipment" information, indicating that NLP for fault diagnosis is still at its early development stage and has a long way to go before reaching its full potential.

Beyond data curation, research on improving the interpretability of the DL-based diagnostic models has also been reported. Grezmak et al. investigated LRP to determine which regions of the wavelet time-frequency spectrum of the vibration signal that contribute the most to the motor fault diagnosis performance [117]. The diagnostic model is first constructed as a CNN, which takes the time-frequency spectrums as the input and determine which of the four conditions the corresponding motor belongs to—normal, bowed rotor, broken rotor bar and unbalanced rotor. Subsequently, LRP is investigated to determine how the fault-related information embedded in the input are learned by network to recognize different fault types. Figure 20 shows the corresponding flowchart of the developed method. Experimental evaluation confirms that the CNN learns to distinguish different fault types through different frequency bands in the wavelet spectrums such that the patterns are consistent for the same motor conditions, while being robust to the initial

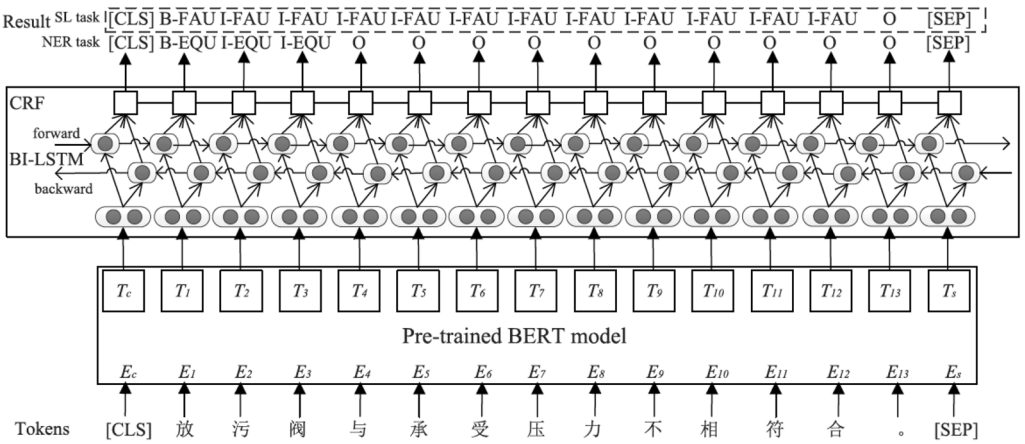

Figure 19 Flowchart for fault-related information extraction from text documents based on NLP [89] (The input (in Chinese) means "the drain valve does not match the pressure applied")

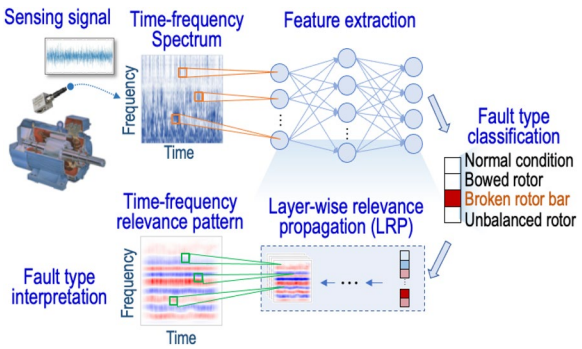

Figure 20 Interpretable CNN using LRP for motor fault diagnosis, adapted from Ref. [117]

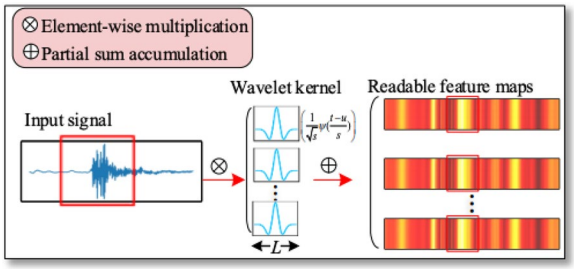

Figure 21 Operation of CWConv layer [122]

network weights during the training process. A similar work has been reported for gearbox fault diagnosis based on frequency spectrums of vibration signal [118]. In Ref. [119], different relevance-based DL interpretation methods are compared under the context of LCD panel inspection, in which LRP has shown to produce the most desirable relevance heatmap for defect detection.

In addition to relevance analysis, attention mechanism has also been increasingly investigated to determine how input elements are being associated by DL models to machine and process conditions. Li et al. designed an attention-incorporated network to determine the influence of different segments of bearing vibration time series signals on the decision-making process of fault recognition [120]. In this work, for each evaluated vibration signal segment, the context of the attention mechanism includes the segments within a time interval in its vicinity. Experimental evaluation has shown that the attention mechanism tends to assign large weights to the segments that contain or are located closer to fault-related impulses and smaller weights to the remaining regions, which is consistent with human logic. A similar work is reported in Ref. [121].

Recently, new interpretable neural networks structures have been reported. For example, Li et al. developed WaveletKernelNet [122], in which a continuous wavelet convolutional layer (CWConv), as shown in Figure 21, is designed to replace the standard convolutional layer in the CNN to discover interpretable filters. By parameterizing the filter using a scaling and a translation parameter [123], the network is shown to generate highly customized wavelet filters by learning from the raw time series signals, which are shown to be effective for bearing and gearbox fault diagnosis.

4.3 Prognosis of Remaining Useful Life

Prognosis aims at predicting the temporal evolution of machine performance from the current time into the future, and possibly until its functional failure. Accurate RUL prediction provides the technological basis for

predictive maintenance and contributes directly to the reduction of unexpected downtime in manufacturing [7].

In general, DL-based prognosis consists of establishing a machine performance evolution model that parses the sequential pattern embedded in the performance evolution and forecasts its future progression based on its historical trajectory. Relevant DL models are usually trained with a set of run-to-failure sequences, which can be time-consuming to obtain. This limitation can be well addressed by GAN through run-to-failure data synthesis. Khan et al. investigated this approach for bearing degradation prognosis [124]. Bearing health degradation is represented as the evolution of the root mean square (RMS) value of the vibration signal over time. The degradation trajectories generated by GAN have shown to highly resemble the run-to-failure data collected from the real experiment and provides the foundation for the degradation model training.

Hou et al. developed an integrated method based on GAN and LSTM for RUL prediction [125]. The network structure is shown in Figure 22.

Different from the conventional GAN-based method, the data synthesis function of the GAN is utilized in this work to improve the quality of the feature extracted from the sequential data to support RUL prediction, rather than generating more training samples. Specifically, the generator is constructed as an AE, and is supervised by a 1-D CNN-based discriminator. In addition, the latent feature extracted by the AE is associated to the RUL by a LSTM. As a result, the training process is guided by two objectives: improving the capability of data synthesis from the latent features (which is implicit) and the RUL predictive accuracy based on these latent features. Once trained, the encoder in the generator and the LSTM are directly used to take an on-going sequence as input and predict its RUL. The authors demonstrated that the developed method has reduced the RUL prediction error for aircraft engine by up to 15% as compared to the previous state-of-the-art.

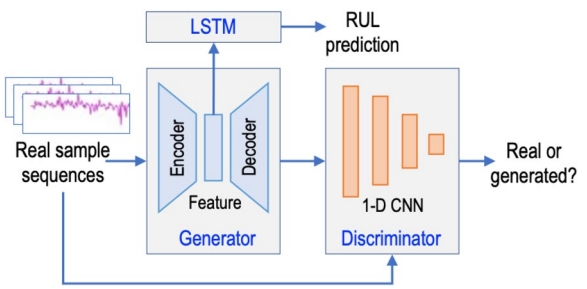

Figure 22 RUL prediction based on GAN and LSTM, adapted from Ref. [125]

Similar to the application in diagnosis, attention mechanism has also been increasingly investigated in RUL prognosis. The objective is to determine the importance of individual features from the sequential data as well as the relevance of individual time steps in the past trajectory to the machine's RUL. Chen et al. developed an attention-based DL model that not only fuses the temporal features learned using a LSTM from different time steps, but also with the handcrafted features, such as the mean and the coefficient of a regression model based on the historical sequential data [126]. The context of the attention mechanism at each time step is represented by the sequential evaluation pattern up to that time step as generated by the LSTM. The developed network structure is shown in Figure 23.

In the experimental evaluation for aircraft engine RUL prediction, the weights generated by the attention mechanism indicates that the features from the more recent time steps have larger influence on the RUL prediction and the importance decreases for the earlier time steps. This is consistent with the logic used by human for prediction. An attention-based method has also been investigated for bearing RUL prediction [127]. In this work, an encoder-decoder structure based on gated recurrent unit (GRU) network has been developed. The encoder first distills the essential information from the temporal features and stores them in hidden states. Then, the attention-incorporated decoder analyzes the hidden states and adaptively determine the information to be used for RUL prediction.

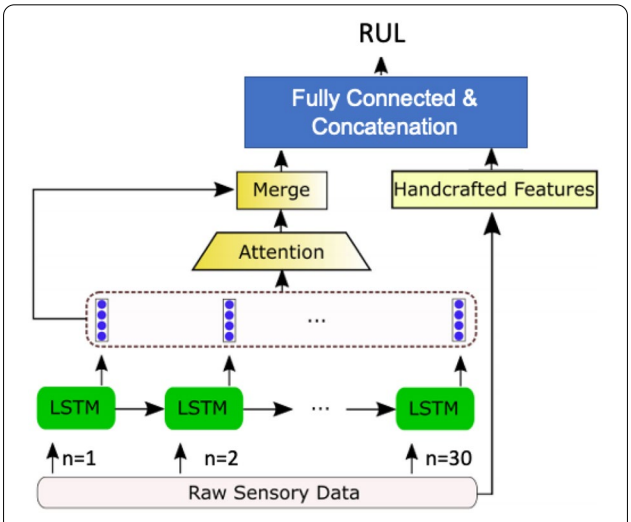

Figure 23 Attention-based RUL prediction, adapted from Ref. [126]

5 Conclusions and Future Work

While the convergence of big data, DL and computation has provided an unprecedented opportunity to advance the state-of-the-art in machine condition monitoring, fault diagnosis and RUL prognosis, uncertainty associated with data and the resulting low data quality, as well as the general "black-box" nature of DL algorithms, have posed significant challenges to the effectiveness and broad acceptance of DL-based methods in manufacturing. To improve data quality and promote user trust in DL, two critically related topics—data curation and model interpretability, have been comprehensively reviewed in this paper. Major techniques covered include: (1) data denoising that utilizes both physical modeling and data-driven characterization for contamination removal; (2) data cleansing that detects, corrects or removes outliers and missing values to ensure data completeness and validity; (3) data synthesis that resolves biases caused by insufficient and unbalanced dataset to reduce bias in model learning; (4) semantic annotation that provides condition-related contextualization to the sensing data; (5) relevance analysis that quantifies the contribution from different inputs in the decision-making process of the neural networks; (6) attention mechanism that allows the capturing of process dynamic relationships for improved model interpretation and performance, and (7) integration of DL with physics to ensure the consistency of DL findings with domain physical knowledge. To explain how these techniques are utilized in practical scenarios, typical manufacturing applications that were enabled by these techniques are highlighted.

As research on DL-enabled manufacturing continues to accelerate, several topics that closely relate to data curation and model interpretation are summarized here, as recommendations for future study:

Uncertainty quantification. Uncertainty quantification is critical to ensuring the robustness of DL models. DL algorithms do not natively incorporate data uncertainty into the analysis, and few reported research on DL-enabled monitoring, fault diagnosis and RUL prognosis has discussed uncertainty quantification [128]. This makes it difficult to translate algorithms developed in academic laboratories into critical applications on the factory floor, where analysis and prediction results without uncertainty quantification cannot be considered realistic and trustworthy. Several uncertainty quantification techniques have been proposed recently for DL models, such as Bayesian deep learning [56, 57, 129, 130]. Still, more rigorous and general approaches need to be developed.

Physics-informed learning. While researchers have started to explore the integration of nerual network with physics by incorporating relevant physical knowledge directly into DL models to ensure the consistency between the DL findings and physical laws [99, 101, 102], physics-informed learning is still in its infancy. Manufacturing is characterized by rich physical-domain knowledge that has been accumulated over the past century. However, most of the knowledge still cannot be incorporated into the existing physics-informed learning framework. A broad, systematic approach is needed for transforming physical knowledge in various forms into elements that can be recognized and operated on by the DL algorithms.

Mitigating false discovery. One of the most compelling aspects of DL is the discovery of potential new knowledge, such as the unknown associations between machine settings and process parameters and the resulting material characteristics of the product. A common limitation associated with the current DL techniques is that they are generally not capable of controlling false discovery rate (FDR). This leads to significant waste of resources in the verification of the DL algorithms' findings. Researchers have started to develop techniques that integrate analytical rigor into DL algorithms to control FDR [131]. Successful adaptation of these techniques into DL-based analysis in manufacturing continues to present an exciting future research direction.

Acknowledgements
The authors would like to thank Clayton Cooper from Case Western Reserve University for editorial assistance.

Authors' Contributions
Both authors are responsible for conceptualization, material, visualization, writing and editing. Both authors read and approved the final manuscript.

Authors' Information
Jianjing Zhang is currently a PhD candidate at *Department of Mechanical and Aerospace Engineering, Case Wester Reserve University, Cleveland, OH, USA*. He received his Master degree from *Case Wester Reserve University, Cleveland, USA*, in 2012, and worked for four years in the industry as an engineer for product quality control. His research interests include the integration of physics with machine learning for modeling of cyber physical systems, such as manufacturing machines and processes, and human–robot collaboration in smart manufacturing.

Robert X. Gao is the Cady Staley Professor of Engineering and Department Chair of Mechanical and Aerospace Engineering at *Case Western Reserve University, Cleveland, OH, USA*. Since receiving his Ph.D. degree from the *Technical University of Berlin, Germany*, in 1991, he has been working in the areas of signal transduction mechanisms for process-embedded sensing, multiresolution data analysis, stochastic modeling, and physics-informed machine learning for improving the observability of cyber physical systems, with applications in manufacturing.

Funding
Not applicable.

Competing Interests
The authors declare no competing financial interests.

Received: 10 December 2020 Revised: 25 May 2021 Accepted: 21 June 2021
Published online: 16 July 2021

References

[1] The World Bank. Manufacturing, value added (% of GDP), 2019.
[2] A Kumar. From mass customization to mass personalization: A strategic transformation. *International Journal of Flexible Manufacturing Systems*, 2007, 19(4): 533-547.
[3] S J Hu. Evolving paradigms of manufacturing: From mass production to mass customization and personalization. *Procedia CIRP*, 2013, 7: 3-8.
[4] L Monostori, B Kádár, T Bauernhansl, et al. Cyber-physical systems in manufacturing. *CIRP Annals*, 2016, 65(2): 621-641.
[5] R Y Zhong, X Xu, E Klotz, et al. Intelligent manufacturing in the context of Industry 4.0: A review. *Engineering*, 2017, 3(5): 616-630.
[6] R Gao, L Wang, M Helu, et al. Big data analytics for smart factories of the future. *CIRP Annals*, 2020, 69(2): 668-692.
[7] R Gao, L Wang, R Teti, et al. Cloud-enabled prognosis for manufacturing. *CIRP Annals*, 2015, 64(2): 749-772.
[8] A Kusiak. Smart manufacturing must embrace big data. *Nature News*, 2017, 544(7648): 23.
[9] F Tao, Q Qi, A Liu, et al. Data-driven smart manufacturing. *Journal of Manufacturing Systems*, 2018, 48: 157-169.
[10] Y LeCun, Y Bengio, G Hinton. Deep learning. *Nature*, 2015, 521(7553): 436-444.
[11] M Sharp, R Ak, T Hedberg Jr. A survey of the advancing use and development of machine learning in smart manufacturing. *Journal of Manufacturing Systems*, 2018, 48: 170-179.
[12] H Yang, S Kumara, S T Bukkapatnam, et al. The internet of things for smart manufacturing: A review. *IISE Transactions*, 2019, 51(11): 1190-1216.
[13] P Wang, R Gao, Z Fan. Cloud computing for cloud manufacturing: Benefits and limitations. *Journal of Manufacturing Science and Engineering*, 2015, 137(4).
[14] A Cano. A survey on graphic processing unit computing for large-scale data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018, 8(1): e1232.
[15] L Wang, R Gao, J Váncza, et al. Symbiotic human-robot collaborative assembly. *CIRP Annals*, 2019, 68(2): 701-726.
[16] C Wang, X P Tan, S B Tor, et al. Machine learning in additive manufacturing: State-of-the-art and perspectives. *Additive Manufacturing*, 2020: 101538.
[17] S Khan, T Yairi. A review on the application of deep learning in system health management. *Mechanical Systems and Signal Processing*, 2018, 107: 241-265.
[18] J Cao, E Brinksmeier, M Fu, et al. Manufacturing of advanced smart tooling for metal forming. *CIRP Annals*, 2019, 68(2): 605-628.
[19] Z Zhao, T Li, J Wu, et al. Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study. *ISA Transactions*, 2020, 107: 224-255.
[20] G Qian, S Lu, D Pan, et al. Edge computing: A promising framework for real-time fault diagnosis and dynamic control of rotating machines using multi-sensor data. *IEEE Sensors Journal*, 2019, 19(11): 4211-4220.
[21] L Zhang, J Lin, B Liu, et al. A review on deep learning applications in prognostics and health management. *IEEE Access*, 2019, 7: 162415-162438.
[22] J Wang, Y Ma, L Zhang, et al. Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems*, 2018, 48: 144-156.
[23] R Zhao, R Yan, Z Chen, et al. Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, 2019, 115: 213-237.
[24] D Kozjek, D Kralj, P Butala. Interpretative identification of the faulty conditions in a cyclic manufacturing process. *Journal of Manufacturing Systems*, 2017, 43: 214-224.
[25] W Samek, T Wiegand, K R Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. 2017, arXiv preprint arXiv:1708.08296.
[26] A Freitas, E Curry. Big data curation. In: *New horizons for a data-driven economy*. Springer, Cham, 2016: 87-118.
[27] R Roscher, B Bohn, M F Duarte, et al. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 2020, 8: 42200-42216.
[28] Y Wang, X Sun, J Fleischer. When deep denoising meets iterative phase retrieval. *International Conference on Machine Learning*, 2020: 10007–10017.
[29] I Goodfellow, J Pouget-Abadie, M Mirza, et al. Generative adversarial nets. *Neural Information Processing Systems*, 2014, 3: 2672-2680.
[30] J Long, E Shelhamer, T Darrell. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015: 3431-3440.
[31] S Bach, A Binder, G Montavon, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One*, 2015, 10(7): e0130140.
[32] D Bahdanau, K Cho, B Y Engio. Neural machine translation by jointly learning to align and translate. 2014: arXiv preprint arXiv:1409.0473.
[33] M Raissi, P Perdikaris, G E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 2019, 378: 686-707.
[34] General Electric Intelligent Platforms. The rise of industrial big data. silo.tips/download/the-rise-of-industrial-big-data-2, 2012.
[35] L Song, F Wang, S Li, et al. Phase congruency melt pool edge extraction for laser additive manufacturing. *Journal of Materials Processing Technology*, 2017, 250: 261-269.
[36] R Yan, R Gao. A nonlinear noise reduction approach to vibration analysis for bearing health diagnosis. *Journal of Computational and Nonlinear Dynamics*, 2012, 7(2).
[37] S Liu, R Gao, D John, et al. Tissue artifact removal from respiratory signals based on empirical mode decomposition. *Annals of Biomedical Engineering*, 2013, 41(5): 1003-1015.
[38] A M Wink, J B Roerdink. Denoising functional MR images: A comparison of wavelet denoising and Gaussian smoothing. *IEEE Transactions on Medical Imaging*, 2004, 23(3): 374-387.
[39] J Gao, H Sultan, J Hu, et al. Denoising nonlinear time series by adaptive filtering and wavelet shrinkage: a comparison. *IEEE Signal Processing Letters*, 2009, 17(3): 237-240.
[40] B Holm-Hansen, R Gao, L Zhang. Customized wavelet for bearing defect detection. *Journal of Dynamic Systems, Measurement, and Control*, 2004, 126(4): 740-745.
[41] J Collins, C Chow, T Imhoff. Stochastic resonance without tuning. *Nature*, 1995, 376(6537): 236-238.
[42] R Zhao, R Yan, R Gao. Dual-scale cascaded adaptive stochastic resonance for rotary machine health monitoring. *Journal of Manufacturing Systems*, 2013, 32(4): 529-535.
[43] C Wang, F A Cheikh, M Kaaniche, et al. Variational based smoke removal in laparoscopic images. *Biomedical Engineering Online*, 2018, 17(1): 1-18.
[44] C Tian, L Fei, W Zheng, et al. Deep learning on image denoising: An overview. *Neural Networks*, 2020, https://doi.org/10.1016/j.neunet.2020.07.025.
[45] M Elad, M Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 2006, 15(12): 3736-3745.
[46] S Diamond, V Sitzmann, F Heide, et al. Unrolled optimization with deep priors. 2017: arXiv preprint arXiv:1705.08041.
[47] P Hand, O Leong, V Voroninski. Phase retrieval under a generative prior. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018: 9154–9164.
[48] F Wan, G Guo, C Zhang, et al. Outlier detection for monitoring data using stacked autoencoder. *IEEE Access*, 2019, 7: 173827-173837.
[49] W Lin, C Tsai. Missing value imputation: A review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 2020, 53(2): 1487-1509.

[50] H Wang, M J Bah, M Hammad. Progress in outlier detection techniques: A survey. *IEEE Access*, 2019, 7: 107964-108000.

[51] M Ahsan, M Mashuri, H Kuswanto, et al. Outlier detection using PCA mix based T 2 control chart for continuous and categorical data. *Communications in Statistics-Simulation and Computation*, 2019: 1-28.

[52] J Ahn, M H Lee, J A Lee. Distance-based outlier detection for high dimension, low sample size data. *Journal of Applied Statistics*, 2019, 46(1): 13-29.

[53] G Bhattacharya, K Ghosh, A S Chowdhury. Outlier detection using neighborhood rank difference. *Pattern Recognition Letters*, 2015, 60: 24-31.

[54] G Gan, M K P Ng. K-means clustering with outlier removal. *Pattern Recognition Letters*, 2017, 90: 8-14.

[55] Y Xia, X Cao, F Wen, et al. Learning discriminative reconstructions for unsupervised outlier removal. *Proceedings of the IEEE International Conference on Computer Vision*, 2015: 1511–1519.

[56] B Lakshminarayanan, A Pritzel, C Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017: 6405–6416.

[57] Y Gal, Z Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, 2016: 1050-1059.

[58] A Kendall, Y Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017: 5580–5590.

[59] J Linmans, J van der Laak, G Litjens. Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks. *Medical Imaging with Deep Learning*, 2020: 465–478.

[60] S Liang, Y Li, R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *International Conference on Learning Representations*, 2018: 1-15.

[61] K Lee, K Lee, H Lee, et al. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018: 7167–7177.

[62] J Ma, J C Cheng, F Jiang, et al. A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data. *Energy and Buildings*, 2020, 216: 109941.

[63] T Ouyang, X Zha, L Qin. A combined multivariate model for wind power prediction. *Energy Conversion and Management*, 2017, 144: 361-373.

[64] A A Kasam, B D Lee, C J Paredis. Statistical methods for interpolating missing meteorological data for use in building simulation. *Building Simulation*, 2014, 7(5): 455-465.

[65] Z Che, S Purushotham, K Cho, et al. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 2018, 8(1): 1-12.

[66] W Cao, D Wang, J Li, et al. BRITS: bidirectional recurrent imputation for time series. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018: 6776–6786.

[67] Y Zhuang, R Ke, Y Wang. Innovative method for traffic data imputation based on convolutional neural network. *IET Intelligent Transport Systems*, 2018, 13(4): 605-613.

[68] D Ulyanov, A Vedaldi, V Lempitsky. Deep image prior. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018: 9446–9454.

[69] Y Zhang, X Li, L Gao, et al. Imbalanced data fault diagnosis of rotating machinery using synthetic oversampling and feature learning. *Journal of Manufacturing Systems*, 2018, 48: 34-50.

[70] P Santos, J Maudes, B A ustillo. Identifying maximum imbalance in datasets for fault diagnosis of gearboxes. *Journal of Intelligent Manufacturing*, 2018, 29(2): 333-351.

[71] R Yan, F Shen, C Sun, et al. Knowledge transfer for rotary machine fault diagnosis. *IEEE Sensors Journal*, 2019, 20(15): 8374-8393.

[72] C Li, S Zhang, Y Qin, et al. A systematic review of deep transfer learning for machinery fault diagnosis. *Neurocomputing*, 2020, 407: 121-135.

[73] B Yang, Y Lei, F Jia, et al. An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings. *Mechanical Systems and Signal Processing*, 2019, 122: 692-706.

[74] S Xing, Y Lei, S Wang, et al. Distribution-invariant deep belief network for intelligent fault diagnosis of machines under new working conditions. *IEEE Transactions on Industrial Electronics*, 2020, 68(3): 2617-2625.

[75] N V Chawla, K W Bowyer, L O Hall, et al. SMOTE: synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 2002, 16: 321-357.

[76] D P Kingma, M Welling. Auto-encoding variational bayes.2013: arXiv preprint arXiv:1312.6114.

[77] M Grasso, A G Demir, B Previtali, et al. In situ monitoring of selective laser melting of zinc powder via infrared imaging of the process plume. *Robotics and Computer-Integrated Manufacturing*, 2018, 49: 229-239.

[78] S Clijsters, T Craeghs, S Buls, et al. In situ quality control of the selective laser melting process using a high-speed, real-time melt pool monitoring system. *The International Journal of Advanced Manufacturing Technology*, 2014, 75(5-8): 1089-1101.

[79] O Ronneberger, P Fischer, T Brox. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015: 234–241.

[80] K He, G Gkioxari, P Dollár, et al. Mask r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, 2017: 2961–2969.

[81] T Lei, Q Zhang, D Xue, et al. End-to-end change detection using a symmetric fully convolutional network for landslide mapping. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019: 3027–3031.

[82] K He, X Zhang, S Ren, et al. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770–778.

[83] D W Otter, J R Medina, J K Kalita. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 2021, 32(2): 604-624.

[84] T Mikolov, K Chen, G Corrado, et al. Efficient estimation of word representations in vector space. 2013: arXiv preprint arXiv:1301.3781.

[85] T Sexton, M P Brundage, M Hoffman, et al. Hybrid datafication of maintenance logs from ai-assisted human tags. *2017 IEEE International Conference on Big Data*, 2017: 1769–1777.

[86] A Thomas, S Sangeetha. Deep learning architectures for named entity recognition: A survey. In: *Advanced computing and intelligent engineering*, 2020: 215–225.

[87] A Vaswani, N Shazeer, N Parmar, et al. Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017: 6000–6010.

[88] J Devlin, M W Chang, K Lee, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019: 4171–4180.

[89] T Chen, J Zhu, Z Zeng, et al. Compressor fault diagnosis knowledge: A benchmark dataset for knowledge extraction from maintenance log sheets based on sequence labeling. *IEEE Access*, 2021: 59394–59405.

[90] K Simonyan, A Vedaldi, A Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. 2013: arXiv preprint arXiv:1312.6034.

[91] B Dickson. Deep learning doesn't need to be a black box. https://bdtechtalks.com/2021/01/11/concept-whitening-interpretable-neural-networks/.

[92] M D Zeiler, R Fergus. Visualizing and understanding convolutional networks. *European Conference on Computer Vision*, 2014: 818–833.

[93] Z Qin, F Yu, C Liu, et al. How convolutional neural network see the world-A survey of convolutional neural network visualization methods. 2018: arXiv preprint arXiv:1804.11191.

[94] M T Luong, H Pham, C D Manning. Effective approaches to attention-based neural machine translation. 2015: arXiv preprint arXiv:1508.04025.

[95] A Karpatne, W Watkins, J Read, et al. How can physics inform deep learning methods in scientific problems? Recent Progress and Future Prospects. *31st Conference on Neural Information Processing Systems (NeurIPS)*, 2017: 1–5.

[96] T J Choi, N Subrahmanya, H Li, et al. Generalized practical models of cylindrical plunge grinding processes. *International Journal of Machine Tools and Manufacture*, 2008, 48(1): 61-72.

[97] G Xiao, S Malkin. On-line optimization for internal plunge grinding. *CIRP Annals*, 1996, 45(1): 287-292.

[98] A Mansour, H Abdalla. Surface roughness model for end milling: a semi-free cutting carbon casehardening steel (EN32) in dry condition. *Journal of Materials Processing Technology*, 2002, 124(1-2): 183-191.

[99] Q Tian, S Guo, Y Guo. A physics-driven deep learning model for process-porosity causal relationship and porosity prediction with interpretability in laser metal deposition. *CIRP Annals*, 2020, 69(1): 205-208.

[100] P C Paris, F A Erdogan. Critical analysis of crack propagation laws. *Journal of Basic Engineering*, 1963, D85(4): 528-534.

[101] R G Nascimento, F A Viana. Fleet prognosis with physics-informed recurrent neural networks. 2019: arXiv preprint arXiv:1901.05512.

[102] J Wang, Y Li, R Zhao, et al. Physics guided neural network for machining tool wear prediction. *Journal of Manufacturing Systems*, 2020, 57: 298-310.

[103] X Jia, J Willard, A Karpatne, et al. Physics guided RNNs for modeling dynamical systems: A case study in simulating lake temperature profiles. *Proceedings of the 2019 SIAM International Conference on Data Mining*, 2019: 558–566.

[104] A Karpatne, W Watkins, J Read, et al. Physics-guided neural networks (pgnn): An application in lake temperature modeling. 2017: arXiv preprint arXiv:1710.11431.

[105] ISO 17359: Condition monitoring and diagnostics of machines – General guidelines.

[106] H Miao, Z Zhao, C Sun, et al. A U-Net-Based approach for tool wear area detection and identification. *IEEE Transactions on Instrumentation and Measurement*, 2020, 70: 1-10.

[107] S M chuster, K K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 1997, 45(11): 2673-2681.

[108] J Zhang, P Wang, R X Gao. Attention mechanism-incorporated deep learning for AM part quality prediction. *Procedia CIRP*, 2020, 93: 96-101.

[109] Y O Lee, J Jo, J Hwang. Application of deep neural network and generative adversarial network to industrial maintenance: A case study of induction motor fault detection. *2017 IEEE International Conference on Big Data*, 2017: 3248-3253.

[110] S Shao, P Wang, R Yan. Generative adversarial networks for data augmentation in machine fault diagnosis. *Computers in Industry*, 2019, 106: 85-93.

[111] Z Wang, J Wang, Y Wang. An intelligent diagnosis scheme based on generative adversarial learning deep neural networks and its application to planetary gearbox fault pattern recognition. *Neurocomputing*, 2018, 310: 213-222.

[112] C Cooper, J Zhang, R X Gao, et al. Anomaly detection in milling tools using acoustic signals and generative adversarial networks. *Procedia Manufacturing*, 2020, 48: 372-378.

[113] L Scime, D Siddel, S Baird, et al. Layer-wise anomaly detection and classification for powder bed additive manufacturing processes: A machine-agnostic algorithm for real-time pixel-wise semantic segmentation. *Additive Manufacturing*, 2020, 36: 101453.

[114] Z Jin, Z Z hang, J Ott, et al. Precise localization and semantic segmentation detection of printing conditions in fused filament fabrication technologies using machine learning. *Additive Manufacturing*, 2021, 37: 101696.

[115] H Wu, W Gao, X Xu. Solder joint recognition using mask R-CNN method. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 2019, 10(3): 525-530.

[116] Z Huang, W Xu, K Yu. Bidirectional LSTM-CRF models for sequence tagging. 2015: arXiv preprint arXiv:1508.01991.

[117] J Grezmak, J Zhang, P Wang, et al. Interpretable convolutional neural network through layer-wise relevance propagation for machine fault diagnosis. *IEEE Sensors Journal*, 2019, 20(6): 3172-3181.

[118] J Grezmak, P Wang, C Sun, et al. Explainable convolutional neural network for gearbox fault diagnosis. *Procedia CIRP*, 2019, 80: 476-481.

[119] M Lee, J Jeon, H Lee. Explainable AI for domain experts: A post Hoc analysis of deep learning for defect classification of TFT–LCD panels. *Journal of Intelligent Manufacturing*, 2021: 1-13.

[120] X Li, W Zhang, Q Ding. Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism. *Signal Processing*, 2019, 161: 136-154.

[121] Z B Yang, J P Zhang, Z B Zhao, et al. Interpreting network knowledge with attention mechanism for bearing fault diagnosis. *Applied Soft Computing*, 2020, 97: 106829.

[122] T Li, Z Zhao, C Sun, et al. WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021: 1–11.

[123] R Gao, R Yan. *Wavelets: Theory and applications for manufacturing*. Springer Science & Business Media, 2010.

[124] S A Khan, A E Prosvirin, J M Kim. Towards bearing health prognosis using generative adversarial networks: Modeling bearing degradation. *2018 International Conference on Advancements in Computational Sciences (ICACS)*, 2018: 1–6.

[125] G Hou, S Xu, N Zhou, et al. Remaining useful life estimation using deep convolutional generative adversarial networks based on an autoencoder scheme. *Computational Intelligence and Neuroscience*, 2020.

[126] Z Chen, M Wu, R Zhao, et al. Machine remaining useful life prediction via an attention-based deep learning approach. *IEEE Transactions on Industrial Electronics*, 2020, 68(3): 2521-2531.

[127] Y Chen, G Peng, Z Zhu, et al. A novel deep learning method based on attention mechanism for bearing remaining useful life prediction. *Applied Soft Computing*, 2020, 86: 105919.

[128] M Fujishima, K Ohno, S Nishikawa, et al. Study of sensing technologies for machine tools. *CIRP Journal of Manufacturing Science and Technology*, 2016, 14, 71-75.

[129] H Wang, D Y Yeung. Towards Bayesian deep learning: A framework and some existing methods. *IEEE Transactions on Knowledge and Data Engineering*, 2016, 28(12): 3395-3408.

[130] S Depeweg, J M Hernandez-Lobato, F Doshi-Velez, et al. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. *International Conference on Machine Learning*, 2018: 1184–1193.

[131] R F Barber, E J Candès. Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 2015, 43(5): 2055-2085.