Journal of Intelligent Manufacturing
https://doi.org/10.1007/s10845-025-02596-3

Explainable neural network for time series-based condition monitoring in sheet metal shearing

Marco Becker[1] · Philipp Niemietz[1] · Thomas Bergs[1,2]

Received: 26 November 2024 / Accepted: 18 March 2025

Abstract

Research indicates the effectiveness of machine learning for condition monitoring in sheet metal shearing. However, existing studies primarily focused on model accuracy while neglecting model explainability. In consequence, potential biases and novel insights captured by the models remained concealed. This work contributes to the state of the art by exploring the intersection of deep learning and causal inference to obtain an explainable condition monitoring model. A causal representation learning framework based on the variational autoencoder architecture is adapted to derive a latent variable model of punch force signals from a fine blanking process. The latent variable model serves two purposes. First, it identifies latent factors explaining variations in observed force signals. Second, it provides a generative model that translates manipulations of latent factors into corresponding force signal changes, thereby, enabling interpretation of the factors. The latent variable model is integrated with a neural network that estimates punch wear. The importance of the latent factors with respect to the network's wear predictions is analyzed to understand how the model arrives at its predictions. Experimental findings indicate that the latent variable model successfully discovered factors which correspond to real-world mechanisms affecting the punch force. One latent factor isolated a bias from measurement interventions, while another captured force variations which are attributed to punch wear itself. Furthermore, the approach demonstrated effectiveness in detecting biased prediction models, contributing to more reliable condition monitoring systems.

Keywords Explainable AI · XAI · Fine blanking · Sheet metal shearing · Condition monitoring · Time series

Introduction

Sheet metal forming and shearing processes are fundamental for the series production of components, e.g., in the automotive, the medical, or the household appliances industries (Klocke, 2013). Many researchers studied data-driven methods, in particular supervised and unsupervised machine learning, for condition monitoring in these processes (e.g., Asahi et al, 2021, Molitor et al. 2022). Machine learning enables to indirectly infer the state of tool components, which are not directly observable in industrial processes, from process signals like acoustic emission data (Unterberg et al., 2024). Ultimately, the aim of these scientific efforts is to provide manufacturers with a transparency about their processes that allows them to optimize their maintenance protocols and thereby reduce machine downtimes and extend machine life (Kubik et al., 2022b).

Many publications indicate that machine learning models achieved promising predictive accuracies in the context of condition monitoring in sheet metal processing (see Section "State of the art of condition monitoring in sheet metal shearing"). However, existing research often neglects how the models arrived at their predictions. Models are often solely validated based on their outputs, neglecting their explainability and essentially treating them as black boxes. Consequently, these models only provide limited insights into the relationship between, e.g., wear mechanisms and pro-

✉ Marco Becker
m.becker@mti.rwth-aachen.de

Philipp Niemietz
p.niemietz@mti.rwth-aachen.de

Thomas Bergs
t.bergs@mti.rwth-aachen.de

[1] Manufacturing Technology Institute MTI, RWTH Aachen University, Campus-Boulevard 30, 52074 Aachen, Germany

[2] Fraunhofer Institute for Production Technology IPT, Steinbachstraße 17, 52074 Aachen, Germany

Published online: 28 March 2025

cess signals. Moreover, biases in the data and models remain concealed.

The field of explainable AI (xAI) is concerned with methods that enhance the explainability of machine learning models and thereby addresses the aforementioned research gap. As pointed out by Liewald et al. (2022), xAI methods are not yet prevalent in in the field of metal forming and shearing but bear the potential to gain insights about previously unknown relationships between features of process data and the physical state of the process. Explainable AI methods not only contribute to an increased acceptance of data-driven models, but may also facilitate conclusions for improved process design.

The effectiveness of a machine learning model heavily relies on how the input data is represented through features (Goodfellow et al., 2016). Since the relationship between wear and process signals is not yet fully understood, manual feature engineering bears the risk of neglecting relevant but so far undiscovered features. Thus, a machine learning-based wear monitoring system would ideally receive the whole process signal and identify relevant features autonomously.

As Section "State of the art of condition monitoring in sheet metal shearing" will show, process signals used for wear monitoring in sheet metal shearing are typically recorded in the form of time series data (e.g., force signals). Compared to image or tabular data, time series data pose a unique challenge for explainable AI. Well established xAI methods like SHAP (Lundberg & Lee, 2017) effectively explain model decisions for images and tabular data by quantifying the importance of input pixels or features. Important regions in images are visually interpretable. For instance, it would be possible to check whether a vision-based monitoring model focuses on worn areas in an image or on another spuriously correlated feature. Similarly, engineered features have a clear meaning to their human developers. In contrast, when data points in time series are identified as important, it remains ambiguous whether the model is responding to the absolute values, the slope, the shape or other characteristics. Hence, an xAI model for process signals from sheet metal shearing processes requires more expressive explanations that facilitate interpretation by engineers.

Concluding, there are two interrelated main challenges, which this paper aims to address: On the one hand, potentially novel and relevant features must be identified from raw process signals. On the other hand, explanations must be suitable for time series data from sheet metal shearing processes.

This article advances the current state of the art of condition monitoring in sheet metal shearing by making the following three contributions:

1. This research presents an approach for learning explainable condition monitoring models directly from raw process data. Through a generative model the method provides clear and expressive illustrations of process signal characteristics, which contribute to wear predictions. The approach is validated using force signals from a fine blanking machine to estimate punch wear.
2. The learned model is used to identify a specific feature of the raw force signals which may serve as a valuable indicator for wear monitoring in fine blanking.
3. The article highlights the importance of xAI by uncovering a bias in a predictive model trained in the course of this research.

The following Section "State of the art of condition monitoring in sheet metal shearing" reviews related scientific literature and highlights current research gaps that this article addresses. Based on these gaps, Section "Research objectives and approach" establishes the research objectives and questions of this work.

State of the art of condition monitoring in sheet metal shearing

Lee et al. (1997) fitted an autoregressive model on force peak values from a blanking process and used the model's coefficients as input features for a linear discriminant analysis to classify punch wear states as sharp or worn. Explanations for the model's predictions were not considered. Moreover, the approach is restrictive in that it only takes into account peak values of the blanking force.

Jin and Shi (2000) applied principal component analysis (PCA) to extract features from force signals of a stamping process. The features were related to (binary) states of different process variables, e.g., normal vs. abnormal lubrication, using a subsequent regression analysis. The authors derived a hierarchy indicating the significance of the association between the force signal variations and the process variables. However, it remained unknown what features of the force signals related to which process variables, as the force signals were represented by abstract principal components of the PCA.

Ge et al. (2004) proposed the use of support vector machines (SVM) to classify whether strokes of sheet metal stamping operations belong to a certain fault type based on time series signals from strain sensors. Their black box approach did not provide any explanations as to which features of the time series signals were relevant for the SVM-based classification.

Griffin et al. (2021) extracted features from acoustic emission signals to classify galling wear in sheet metal stamping. They fitted a decision tree, a neural network as well as a fuzzy clustering model and concluded from their results that the two former models outperform their unsupervised alternative. While decision trees are inherently able to explain their predictions, explanations for neither of the models were considered. Moreover, the authors restricted the available information from the acoustic emission signals to a small number of extracted features like minimum and maximum amplitude.

Kubik et al. (2021) utilized correlation analysis and linear regression to investigate the relationship between force displacement curves, process parameters and resulting component quality in blanking. While the authors emphasized the importance of identifying suitable features of process signals, their analysis is limited to a set of known relevant features. In another work, Kubik et al. (2022a) applied an SVM with a linear kernel to classify different wear states in blanking based on force signals. The authors again emphasized the importance of identifying suitable features. Their model was tested on three different feature sets. The first consisted of two knowledge-based features, the second of two PCA-based features and the third of one feature from each of the latter two sets. While the accuracy was 100% in all cases, Kubik et al. argued that the distance between the classes varies with different feature sets. In a third work, Kubik et al. (2022b) studied multiple machine learning models, including SVMs, random forest, and k-nearest neighbors, to classify different states of abrasive wear in blanking. In this study, they again compared different approaches for feature extraction, either based on engineering knowledge or based on data-driven methods like PCA or an autoencoder. While they emphasized the performance of the latter, both approaches yielded models with above 99.99% accuracy. In yet another publication, Kubik et al. (2023a) again considered engineered features as well as PCA-based features for monitoring of abrasive wear in blanking and in roll forming. From initially eight considered regression models they selected an artificial neural network as a particularly suitable model. They found that the best model performance is achieved when a combination of engineered features and principal components is used as model input. Neither of these publications incorporated xAI methods to investigate how the single features contributed to the models' predictions. Potentially novel and interesting features extracted by the autoencoder (Kubik et al., 2022b) also remain concealed.

Unterberg et al. (2021) investigated the use of acoustic emission (AE) data from a fine blanking machine for tool condition monitoring. They first extracted features from the time series signals, before applying PCA and UMAP to project the data into a lower dimensional space. They found patterns in the projections that seemed to correlate with the present punch wear. In a later study, Unterberg et al. (2024) fitted a Lasso regression model and an XGBoost model on features from AE data to estimate punch wear in fine blanking. Their models utilized a broad range of statistical, temporal and spectral features. Studying SHAP values and the coefficients of the Lasso model respectively, Unterberg et al. identified a range of spectral features, in particular several FFT coefficients, which had a high importance for the models' wear predictions. Concluding, Unterberg et al. (2024) took into account model explainability and a broad range of descriptive features. A limitation of the study is that it relied on generic time series features extracted via the library TSFEL (Barandas et al., 2020), which may only be connected to real-world phenomena via a complex function and therefore complicate the model interpretation from an engineering perspective.

Asahi et al. (2021) argued for the use of deep learning to avoid feature extraction. They proposed a convolutional autoencoder to estimate punch wear from raw time series data like pressure, force and sound signals. Asahi et al. fitted their model on data recorded with a new punch. Subsequently, the fitted model was applied to new data acquired with different punch wear states. They report that the reconstruction error, i.e. the difference between the encoder input and the decoder output, corresponded to the punch wear state and hence serves as a suitable indicator for wear estimation. Niemietz et al. (2021) followed the same idea. In their work, they applied convolutional autoencoders to acoustic emission and force signals from a fine blanking machine also indicating that the reconstruction error may serve as a metric for wear condition monitoring. Antoher related work was published by Biegel et al. (2022), who applied autoencoders and used their learned data representations for statistical process control in sheet metal forming. Neither of these three works provided explanations for the learned models or features.

In another study, Niemietz et al. (2022) projected force signals from a fine blanking process to lower dimensional representations using PCA, UMAP and an autoencoder. While the autoencoder learned from raw data, the other two approaches received a diverse set of extracted features from the statistical, spectral and time domain. The authors derived indicators (e.g., cumulative rolling variance) that aimed to relate variations in the low-dimensional representations to changes in punch wear states. In a related work, Niemietz et al. (2023) again compared different methods, including PCA, discrete Fourier transform, and an autoencoder to transform fine blanking punch force signals to lower dimensional representations. Their results indicated a relationship between changes in the low dimensional representation and the punch wear state. However, the authors acknowledged themselves that changes in the process signals may originate from diverse, interdependent causes that are not explained by their analysis, but should be isolated and understood.

Using a convolutional neural network, Huang and Dzulfikri (2021) classified tool wear in a stamping process based on vibration measurements. Their approach first converts the signals to the frequency domain via fast Fourier transform. The network then categorizes the processed signals into seven states, representing mild or heavy wear at three different positions or no wear. Model explainability was not considered.

While the majority of work dealing with data-driven approaches for condition monitoring in sheet metal shearing is concerned with time series process signals like force or acoustic emission data, Molitor et al. (2022) presented a study where convolutional neural networks classified abrasive punch wear states based on images of produced workpieces. The authors did not employ any methods to explain the neural network's prediction. In a recent study, Schlegel et al. (2024) explored U-Net, a convolutional neural network for image segmentation, to assess wear in blanking processes. They recorded images of the tool at the top dead center point at 600 strokes per minute. The neural network segmented the images into regions corresponding to different wear types (e.g., adhesive wear or grooves). However, as pointed out by Kubik et al. (2023b), optical systems are not yet used in practice and difficult to integrate in blanking tool systems. Kubik et al. (2023b) fitted machine learning models to estimate tool wear in blanking based on quality parameters of the blanked parts. Their study did not consider the explainabiltiy of the wear estimation models.

Concluding, the scientific literature suggests that data-driven methods represent an effective approach for condition monitoring in sheet metal shearing. However, existing work typically evaluated models solely based on performance metrics while treating the models as black boxes. In consequence, potentially novel and useful knowledge captured by the models remained hidden. Additionally, neglecting model explainability bears the risk of biased models going unnoticed.

Moreover, many studies relied on manually engineered features rather than letting an algorithm learn relevant features directly from raw data. As already argued by Liewald et al. (2022) feature extraction results in loss of information and impairs the interpretability of the data from an engineering point of view. When features are engineered based on current domain knowledge, new unknown features and novel patterns may be overlooked in the data. When using generic feature extraction libraries to obtain a large number of descriptive features (e.g. spectral or statistical), the connections between considered features and the underlying physical phenomena of a dataset likely become more complicated.

Research objectives and approach

The previous section highlighted two interrelated research gaps. First, existing work usually omits model explainability. Second, there is a lack of methods to identify novel features of process signals that are suitable for condition monitoring and allow for meaningful explanations of the trained models.

In this paper, we investigate how to learn an explainable condition monitoring model from raw process data of fine blanking operations using deep learning. Our research pursues two main objectives:

1. Provide model explanations, allowing to detect biases as well as generating new insights.
2. Automatically learn features, which reflect relevant real-world phenomena explaining variations in the measured data, directly from raw process signals.

In a recent publication, we explored Variational Autoencoders (VAE) for learning an explainable wear estimation model in fine blanking (Becker et al., 2024), building on O'Shaughnessy et al.'s (2020) argument that deep generative models provide a rich and flexible vocabulary for xAI. While the approach generally showed potential, one main limitation remained: The VAE provided explanations through learned factors which only correlated with real-world phenomena. However, there were no direct one-to-one mappings, where single factors represented individual real-world causes of variation in the dataset. Instead, factors blended effects from several phenomena and single phenomena were spread over more than one factor. This complicated the domain-specific interpretation of the model explanations.

Schölkopf et al. (2021) suggested that the integration of concepts from the field of causal inference into the machine learning process allows to learn data representations that better reflect the underlying physical causal mechanisms of a dataset. Hence, we hypothesize that combining deep learning and concepts from causal inference will enable the identification of generative factors from raw fine blanking force signals, that more accurately correspond to real-world physical mechanisms and thereby enable more meaningful explanations. To test this hypothesis, we utilize an algorithm developed by Locatello et al. (2020), which is motivated by the *independent causal mechanisms* (ICM) principle from the field of causal inference. Our study addresses three research questions (RQ):

1. How does incorporating the ICM principle affect the latent representation of fine blanking force signals learned by a variational autoencoder?
2. Can the resulting latent representation be used to provide interpretable explanations for a wear estimation model?
3. To what extent can the model accurately identify the underlying data-generating mechanisms from a fine blanking force dataset?

Dataset and methods

Section "Fine blanking punch force dataset" provides background on the dataset acquisition and preparation for the experiments in this work. Section "Model architecture" elaborates on the neural network architecture and training regime explored in this study to fit a data-driven model which (a) explains its predictions and (b) allows to gain insights regarding hidden biases and relevant features reflecting real-world phenomena in a dataset.

Fine blanking punch force dataset

The data used in this work are available at the Harvard Dataverse (Niemietz, 2022) and consist of punch force signals recorded with piezoelectric sensors from a *Feintool XFT 2500 speed* fine blanking machine. The fine blanking machine was operated at 50 strokes per minute and the force signals were acquired with a sampling rate of 10 kHz. The dataset comprises four experiments E_1, E_2, E_3, and E_4 with 3,334, 3,204, 2,058, and 3,340 strokes respectively. At the start of every experiment the punch of the fine blanking tool was in pristine condition. The progressing wear of the punches was measured twice during each experiment (approximately after 1000 and after 2000 strokes) as well as at the end of each experiment. The wear was measured via scanning electron microscopy (SEM) with 100-fold magnification. For each measurement, SEM images were recorded at four characteristic edges depicted in Fig. 1a. Subsequently, the relative amounts of pixels corresponding to the damaged punch surface were quantified using the *OpenCV* Python library. This wear quantification process was described by Niemietz et al. (2022) and included the following steps: First, the contrast was increased and the images were smoothed through application of OpenCV's median filter. Next, the images were binarized via thresholding. Following, the contour detection algorithm implemented by OpenCV was used to detect contours of the worn area. The detected worn area was completely filled in white and the proportion of white pixels in the processed SEM images was used as a wear estimate. Figure 1b summarizes the image processing steps and depicts exemplary images. The low frequency of wear measurements (approximately every 1,000$^{\text{th}}$ stroke) is due to the fact that a measurement requires to stop the process and disassemble the punch from the tool. All four experiments were carried out with X5CrNi18-10 steel.

Due to irregularities concerning an error with the lubrication in E_1 and missing data of 1000 strokes in E_3 (see Niemietz et al., 2022), this work focused on the experiments E_2 and E_4. Since the findings from E_2 were also representative of E_4, only results from E_2 are discussed below. The results from E_4 are presented in the appendix (see Appendix A and B).

As illustrated by Bergs et al. (2020), force signals of single fine blanking strokes comprise different process phases, including the sheet metal insertion into the tool, the clamping of the sheet metal, the actual shearing phase, and a phase during which the sheet metal is stripped from the punch and the sheared part is ejected. From a tribological perspective, the shearing and the stripping segments of fine blanking punch force signals are most relevant for wear monitoring.

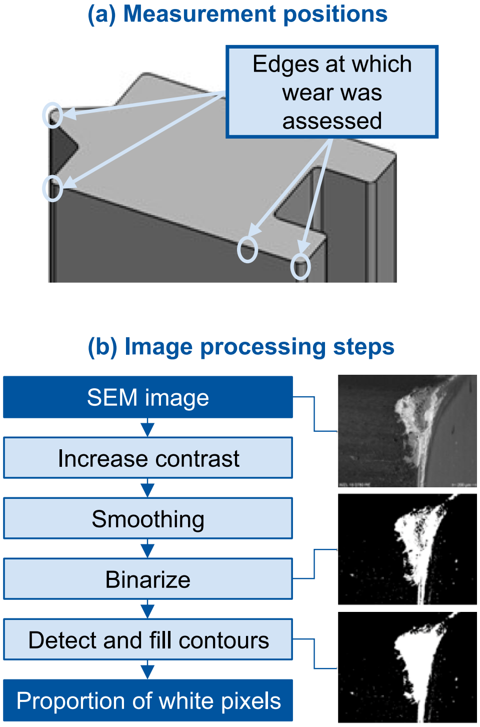

(a) Measurement positions
Edges at which wear was assessed
(b) Image processing steps
SEM image
Increase contrast
Smoothing
Binarize
Detect and fill contours
Proportion of white pixels

Fig. 1 **a** Punch geometry and wear assessment positions and **b** illustration of SEM image processing

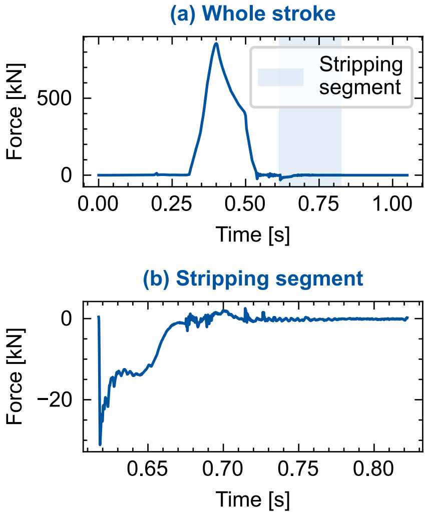


Fig. 2 **a** Exemplary punch force signal and **b** isolated stripping segment

Prior research indicated that especially the stripping segment is useful for data-driven wear monitoring (Niemietz et al., 2022). This is plausible, as friction resulting from wear might be less prominent during the shearing segment, during which the force required to separate the material is more dominant. Hence, the focus of this work is on the stripping segment. Figure 2 depicts a complete punch force signal of a single stroke as well as the isolated stripping segment of the signal.

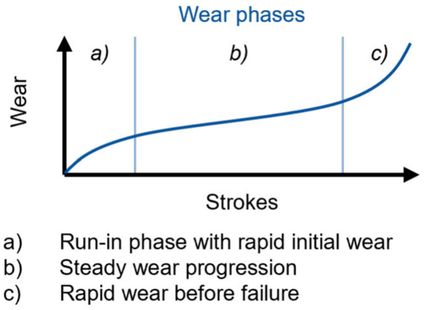

Wear phases
a)
b)
c)
Wear
Strokes
a) Run-in phase with rapid initial wear
b) Steady wear progression
c) Rapid wear before failure

Fig. 3 Theoretical wear progression in three phases (Behrens et al., 2016)

To train a model which relates features of punch force measurements to punch wear, a wear label was assigned to each force signal. As previously described for each experiment only four wear measurements (including the initial wear state) were available. Hence, the values in between the measurements were interpolated using SciPy's (Virtanen et al., 2020) Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) interpolation. The PCHIP interpolation was chosen as it allows to approximate the shape of the theoretical wear progression (see Fig. 3).

The resulting dataset was split into training, validation and test datasets with a ratio of 70:15:15. A global min-max-normalization was applied to the force signals as well as the wear labels, such that all values fall in a range between zero and one. The parameters for the scaling (i.e., global maximum and minimum) were determined based on the training data only to avoid inducing a bias to validation and test data.

Model architecture

Figure 4 illustrates the general model architecture that is proposed in this work. It consists of three neural networks. The *encoder* network and the *decoder* network jointly form a latent variable model (LVM), which will be explained in more detail in Section "Variational autoencoder as latent variable model". The LVM serves two purposes. First, it aims to identify latent variables which explain the observed process signals. For instance, lubrication, wear (Voss et al., 2017) or varying material properties (Schenek et al., 2022) could be latent variables that are not directly observed but cause variation in the measured force signals. Second, the LVM provides a generative model which allows to visualize, how an observed force signal would change if the value of a latent variable was different. Thereby, it provides the means to interpret the learned latent variables. The *predictor* network estimates punch wear using the learned latent variables as input features. Besides, it is used to derive the importance of latent variables for the wear estimation task. Details of the predictor network are elaborated in Section "Explainable wear predictor". While existing research has already utilized encoder-decoder architectures to obtain features for wear monitoring, the approach presented here differs in two aspects. First, it combines a *probabilistic* (generative) decoder with the predictor network, enabling both visualization of features for interpretation as well as analysis of their importance for wear predictions. This is illustrated in detail in Sections "Explainable wear predictor" and "Model constraints for causal disentanglement" and Figs. 5 and 6, respectively. This analysis goes beyond the autoencoder-based feature extraction described in existing research on wear monitoring for sheet metal shearing processes. Second, the latent variable model is subjected to special constraints, inspired by the field of causal inference. These constraints were first presented by Locatello et al. (2020). Section "Model constraints for causal disentanglement" describes these constraints in detail and proposes a strategy to structure training data in pairs of process signals required for adapting Locatello et al.'s framework. Through this approach, we aim to identify latent variables that align with actual causal mechanisms underlying the observed variations in measured force signals, rather than generic variables that only correlate with the process signals.

Variational autoencoder as latent variable model

This work uses a variational autoencoder (VAE) to identify latent variables that explain variations in the observed punch force signals as well as the overall model's (see Fig. 4) wear predictions. VAEs jointly optimize an encoder and a decoder. The encoder $q_\phi(z|x)$ models how likely values of the latent variable z are for given observations of a process signal x, where ϕ denotes the parameters (i.e., weights and biases) of the encoder. The decoder $p_\theta(x|z)$ is also a probabilistic model producing the distribution of x for given values of z. Here, θ denotes the parameters of the decoder. The generative model

$$p_\theta(x, z) = p_\theta(x|z) \cdot p_\theta(z) \tag{1}$$

also requires a prior $p_\theta(z)$. In this work, the prior is a multivariate Gaussian $p_\theta(z) = \mathcal{N}(z; 0, I)$. The optimization criterion of the VAE is the *evidence lower bound* (ELBO). The ELBO is a lower bound on the marginal log-likelihood $\log p_\theta(x^{(i)})$ of an observed datapoint $x^{(i)}$. In other words, the model parameters are optimized, such that according to the model it is likely to observe the actually observed datapoints.

Kingma and Welling (2014) derived the (Monte Carlo) estimator of Eq. (2) for the ELBO. The first (right hand side) term of the equation is the expected negative reconstruction error for input values $x^{(i)}$. The second term can be interpreted as a regularization term. The Kullback–Leibler divergence D_{KL} pushes the modelled posterior $q_\phi(z|x^{(i)})$ to be close to the prior $p_\theta(z)$. L denotes the number of samples $z^{(i,l)}$ drawn

Fig. 4 Model architecture

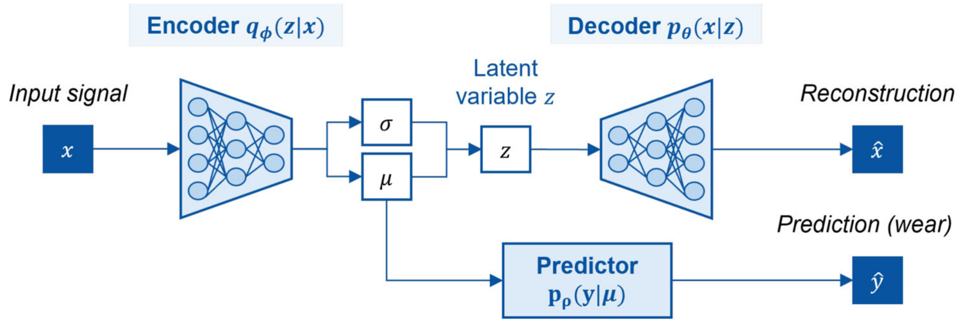


Fig. 5 To interpret the effect of a change in μ_i, a value Δ is added to μ_i. The manipulated vector is used as input for the probabilistic decoder to generate a force signal with the corresponding effect. The standard deviations σ_i are set to zero to obtain the expected value of z. The resulting force signal is plotted together with the force signal corresponding to the unchanged vector μ to analyze the manipulation effect

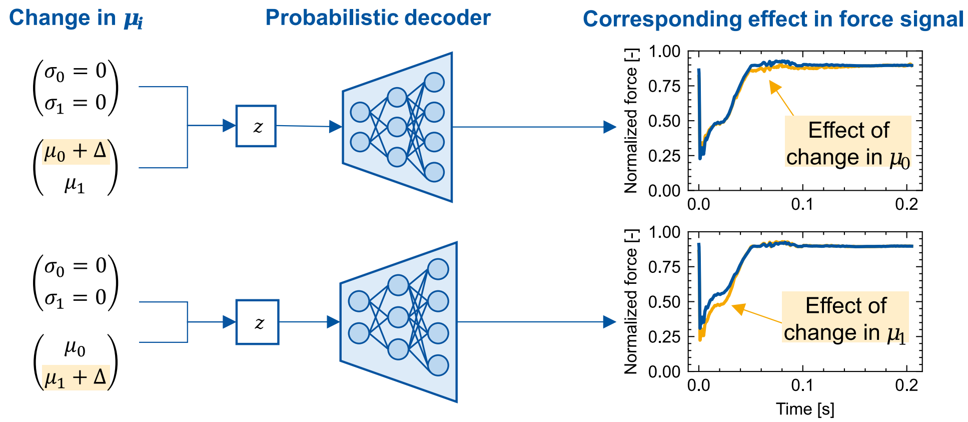
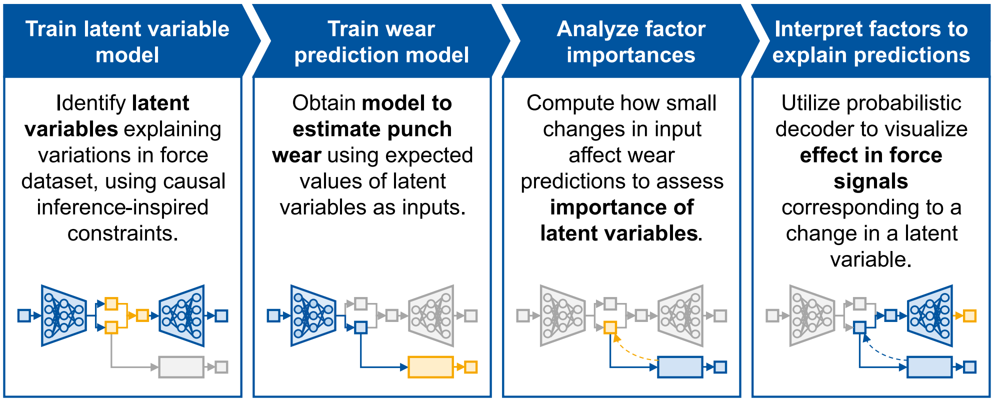


Fig. 6 Interplay of model components for explainable wear monitoring system learned from raw process signals



Train latent variable model

Identify **latent variables** explaining variations in force dataset, using causal inference-inspired constraints.

Train wear prediction model

Obtain **model to estimate punch wear** using expected values of latent variables as inputs.

Analyze factor importances

Compute how small changes in input affect wear predictions to assess **importance of latent variables**.

Interpret factors to explain predictions

Utilize probabilistic decoder to visualize **effect in force signals** corresponding to a change in a latent variable.

per data point $x^{(i)}$. As in the original VAE paper (Kingma & Welling, 2014), $L = 1$ was used for the experiments in this paper.

$$\mathcal{L}(\theta, \phi; x^{(i)}) \simeq \frac{1}{L}\sum_{l=1}^{L}(\log p_\theta(x^{(i)}|z^{(i,l)})) - D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z)) \qquad (2)$$

When $p_\theta(z) = \mathcal{N}(z; 0, I)$ and the approximate posterior is a multivariate Gaussian of the form $q_\theta(z|x) = \mathcal{N}(z; \mu^{(i)}, \sigma^{2(i)}I)$, where $\mu^{(i)}$ and $\sigma^{(i)}$ are outputs of the encoder, the KL divergence D_{KL} from Eq. (2) can be directly computed from $\mu^{(i)}$ and $\sigma^{(i)}$ and no estimation is required. Even when $p_\theta(z)$ and $q_\theta(z|x)$ are Gaussian, deep learning models allow to approximate complex distributions of the observed data x (Kingma & Welling, 2019).

To optimize the model through gradient descent and backpropagation, the random variable z must be reparametrized to get a differentiable version of the estimator. For the Gaussian case described above, the differentiable transformation of Eq. (3) is used, where ϵ is an auxiliary noise variable $\epsilon^{(l)} \sim \mathcal{N}(0, I)$ and $\odot$ denotes an elementwise product.

$$z^{(i,l)} = \mu^{(i)} + \sigma^{(i)} \odot \epsilon^{(l)} \qquad (3)$$

In this work, a variation of the VAE called β-VAE (Higgins et al., 2017) is used, where β is a hyperparameter and factor that is used to control the weight of the KL divergence term of the ELBO estimator. We implemented the β-VAE with six convolutional layers in the encoder and matching transposed convolutional layers in the decoder. All of these layers used 64 filters, a kernel size of 4, a stride of 2 and the tanh activation function, except for the decoder output layer, which used a single filter and the sigmoid activation. Linear layers were used to transform the flattened output of the last convolutional layer into two two-dimensional vectors $\mu^{(i)}$ and $\sigma^{(i)}$. The optimal value for the hyperparameter β differs with the dataset under investigation as well as the model architecture (Higgins et al., 2017). According to Higgins et al. (2017), higher values of β lead to better and interpretable latent factors, but may also impair the model's capability to preserve information and produce reconstructions with high fidelity. The results reported in this paper were achieved with $\beta = 0.1$. The value was chosen as higher values in initial tests resulted in a collapse of the latent space and low reconstruction fidelity.

Explainable wear predictor

As a wear estimation model a simple multilayer perceptron (Goodfellow et al., 2016) is used. The model uses the expected values μ of the latent vector (see Fig. 4) as inputs to produce a wear estimation as output. In principle, the *predictor* model can be any model suitable to solve the desired downstream task (here wear estimation). However, by making it a neural network it can be seamlessly trained jointly with the latent variable model using a gradient descent algorithm and backpropagation. The wear estimation model is optimized by minimizing the mean squared error between predicted wear values $\hat{y}_i$ and the target values y_i, according to Eq. (4).

$$MSE = \frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)^2 \tag{4}$$

In the simplest case, the predictor network is a one-layer neural network without non-linear activation. In this case, the network resembles a linear regression (trained with backpropagation) and the layer weights correspond to the coefficients of the linear regression model. Hence, the layer weights may be directly interpreted as feature importances, assuming that the latent vector has a similar scale across all dimensions due to the KL divergence term from Eq. (2).

By computing the gradients of the predictive model's wear estimates with respect to its input vector μ using backpropagation, local explanations of the model's predictions can be derived for the case of a nonlinear predictor. The gradient vector expresses how much a prediction would change, given an infinitesimal change in the respective input values. Therefore, the gradients are a measure of importance, in that they reflect the model's sensitivity to its individual input features. Using the generative model $p_\theta(x, z)$ (see Section “Variational autoencoder as latent variable model") it is possible to interpret the meaning of a small change in an input value μ_i of the predictor network. This is illustrated in Fig. 5.

In this study, we report results for a predictor network that consisted of three ReLU activated fully connected layers with 25, 50, and 25 neurons respectively. Another fully connected layer with sigmoid activation was used as an output layer to produce the wear predictions.

Model constraints for causal disentanglement

The ELBO optimization criterion described in Section “Variational autoencoder as latent variable model" encourages the VAE to (a) produce a low dimensional latent representation z of the original force signals that allows to reproduce the original data as accurately as possible and (b) arrange the latent space in such a way that it becomes possible to generate realistic force signals with new values of z. However, the latent factors z_i, i.e. the dimensions of z, do not necessarily coincide with the actual data generating, causal factors (e.g. wear or material fluctuations), that cause the variation in the force signals.

We utilize an approach proposed by Locatello et al. (2020) to address this issue. Their approach is inspired by the *independent causal mechanisms (ICM)* principle (Peters et al., 2017; Schölkopf et al., 2021). Peters et al. (2017) illustrate the ICM principle by example of a joint density $p(a, t)$ of altitudes a and average annual temperatures t of a set of cities. Both, $p(a|t) \cdot (t)$ and $p(t|a) \cdot (a)$ are valid decompositions of $p(a, t)$. The conditional probability $p(t|a)$ gives the probability of a temperature t, given an observed altitude a. Hence, temperature values follow from the altitude. This corresponds to the causal direction. The conditional $p(a|t)$ corresponds to the opposite direction, where amplitude follows from temperature. If data were recorded in cities from different countries, e.g., Austria and Switzerland, the distribution of altitudes $p(a)$ might differ between the two countries. However, the physical mechanism responsible for temperature changes in dependence of the altitude is likely the same in both countries. In an idealized setting, the conditional $p(t|a)$ is therefore not affected by a change in $p(a)$ (hence, “independent causal mechanisms”). Conversely, the other conditional $p(a|t)$ is not invariant to a change in $p(t)$.

More genereally, the ICM principle states that “the causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other” (Peters et al., 2017). While every change in observed data is due to a change in at least one causal mechanism, the ICM principle implies that changes in causal mechanisms are usually sparse (i.e., only a few mechanisms change at a time) because of the independence of the mechanisms (Schölkopf et al., 2021).

Following this idea, Locatello et al. considered a learning regime where the model observes pairs of input instances for which some causal mechanisms differ, while others remain invariant due to the ICM principle. To identify latent variables corresponding to causal mechanisms z_i, they proposed a VAE-based model, where each input data pair (x_1, x_2) shares a set S of mechanisms or latent factors respectively. This is expressed by Eq. (5).

$$p(z_i|x_1) = p(z_i|x_2) \quad \forall i \in S \tag{5}$$

To enforce input data pairs will have shared factors, Locatello et al. proposed to constrain the encoder outputs $q_\phi(\hat{z}_i|x_1)$ and $q_\phi(\hat{z}_i|x_2)$, where $\hat{z}_i$ is an estimation of the true data generating mechanisms z_i. For k of the model's latent dimensions i, the encoder outputs $q_\phi(\hat{z}_i|x_1)$ and $q_\phi(\hat{z}_i|x_2)$ will be replaced by the average of the two (see Eqs. 6 and 7).

$$\tilde{q}_\phi(\hat{z}_i|x_1) = \frac{q_\phi(\hat{z}_i|x_1) + q_\phi(\hat{z}_i|x_2)}{2} \tag{6}$$

$$\tilde{q}_\phi(\hat{z}_i|x_2) = \frac{q_\phi(\hat{z}_i|x_1) + q_\phi(\hat{z}_i|x_2)}{2} \tag{7}$$

The averaging is done for those k factors z_i for which $q_\phi(\hat{z}_i|x_1)$ and $q_\phi(\hat{z}_i|x_2)$ are closest to each other measured by their KL divergence.

Typically, the number of shared factors k is unknown. Therefore, Locatello et al. introduced a heuristic to determine k. When k is unknown, they average all coordinates of z where $\delta_i < \tau$. The threshold value τ is computed according to Eq. (8) and δ_i according to Eq. (9).

$$\tau = \frac{1}{2}(\max_i \delta_i + \min_i \delta_i) \tag{8}$$

$$\delta_i = D_{KL}(q_\phi(\hat{z}_i|x_1)||q_\phi(\hat{z}_i|x_2)) \tag{9}$$

Locatello et al. (2020) established the term Ada-GVAE to describe a VAE trained with the constraint described above. The Ada-GVAE is fitted by optimizing the sum of the β-VAE ELBOs (Higgins et al., 2017) of both inputs x_1 and x_2.

If points in time of interventions are known, these could be used to form pairs of input datapoints for the Ada-GVAE. For example, if an influence of material property variations between different coils is suspected, the data can be grouped within coils such that input data pairs (x_1, x_2) are from the same coil. This way, the model will presumably isolate the variation of the coil in distinct latent dimensions shared by input data pairs.

Whenever these concrete intervention times are not known, other assumptions must be made. We propose to build pairs of data instances that are close to each other in time. More precisely, we define a model hyperparameter that specifies the number of consecutive fine blanking strokes that form a group within which force signals are randomly paired each training epoch. The underlying assumption is that in short periods of time (i.e., within a small number of consecutive strokes), the (causal) data generating mechanisms will only change sparsely. Specifically, we assume that it is likely that punch wear and other potential external influencing factors will not all simultaneously and constantly change significantly within the time span of four consecutive fine blanking strokes.

Figure 6 summarizes how components described in Section “Model architecture" work together to obtain an explainable wear monitoring model.

Results

The results are structured into three subsections. First, Section “Latent factors explaining variation in fine blanking force data" compares latent factors learned by a β-VAE and the previously described Ada-GVAE respectively and shows that the factors learned by the Ada-GVAE seem to indeed better align with actual real-world phenomena (RQ 1). Second, Section “Explainable predictor" presents results illustrating how a learned wear estimation model based on the Ada-GVAE explains its predictions. The results emphasize the importance of model explainability by unveiling a bias the model relied on (RQ 2). Finally, Section “Disentanglement of synthetically injected features" further validates the suitability of the learning framework discussed in this paper through an experiment, where additional generative factors were synthetically induced into the fine blanking force dataset to evaluate whether these factors will be correctly identified by the model (RQ 3).

Latent factors explaining variation in fine blanking force data

Figure 7 depicts two two-dimensional latent representations of the force dataset. Each marker in the scatter plots corresponds to one of the original force signals. The representation of Fig. 7a was learned by training a conventional β-VAE without the ICM-inspired constraint explained in Section “Dataset and methods". In contrast, the representation depicted in Fig. 7b was learned by training the Ada-GVAE with the additional constraint.

Both representations display four clusters. However, while in Fig. 7a the clusters are arbitrarily rotated in the latent space, the clusters in Fig. 7b better align with the latent factors. The latent factor z_1 presumably encoded the unobserved, actual cause of the clusters as the cluster boundaries are orthogonal to z_1.

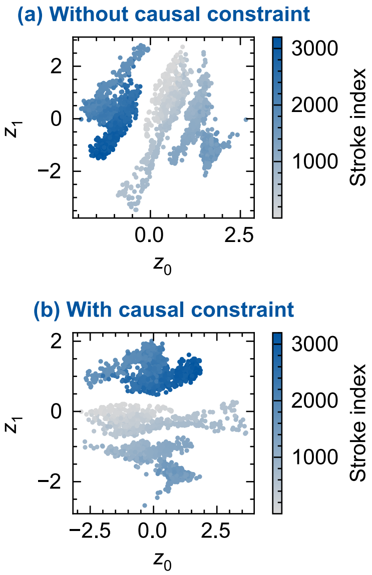


Fig. 7 **a** Latent space learned with model without ICM-based constraint. **b** Latent space learned with model with ICM-based constraint

Figure 8 illustrates the variation that is controlled by the factor z_1 of the Ada-GVAE. Figure 8a depicts four exemplary strokes that were selected from the latent space, with one stroke selected per cluster. The arrows in the latent space visualize the manipulations that were applied to each of these data points to unveil the effect of changes in z_1. For example, at stroke *A1*, the value in z_0 was kept constant, while the value of z_1 was changed from the value at the root of the arrow to the value at the arrowhead by adding a value Δ (as described in Section "Dataset and methods", Fig. 5). For visualization purposes the magnitude Δ of the manipulation was deliberately chosen high. All effects visualized in this section are still present for smaller manipulations, but less discernible in graphical representations.

The blue line in Fig. 8b represents the stripping force signal that the Ada-GVAE reconstructed from the latent values corresponding to stroke A1 back to the original (raw signal) space. The latent values (at the arrowhead) received after the manipulation in z_1 are also decoded into the original space. The difference between the two resulting stripping force signals is represented by an orange filled area. Figure 8b exemplarily shows that a change in z_1 *within* a cluster has almost no effect on the resulting stripping force signals.

Conversely, Fig. 8c and d illustrate that the slope and shape of the stripping force signals change when moving in z_1-direction *between* the three large clusters in the latent space. The stroke numbers that separate these clusters coincide with the points in time when the wear measurements were conducted. Hence, the factor z_1 appears to encode variations in the force signal that were caused by the wear measurement procedure (including the disassembly of the tool for measurement purposes). As z_1 identifies force signal changes clearly attributable to the wear measurement procedure itself, it isolated a bias in the dataset.

Lastly, there is another effect encoded in z_1. Moving from the fourth smaller cluster to its neighboring larger cluster, the jagged part at the top of the signal changes (see Fig. 8e). The cause of this effect is unclear. During this part of the time series, the shearing and the off-stripping of the sheet metal is already finished. Hence, this variation is presumably not related to tool wear.

The effect of the second factor z_0 is explained by Fig. 9. Again, four exemplary strokes were selected as depicted by Fig. 9a. The factor z_0 encodes variations in the height of the "valley" at the beginning of the force signals (see Fig. 9b–e). This feature is consistent across all clusters and may potentially be caused by punch wear. Increased friction due to wear represents a plausible reason for the shift in the valley height.

The feature is related to a specific area and shape in the stripping force segment and differs from simple, manually engineered features like maximum force, mean force or the area under the force curve, which have been used for wear monitoring in related research studying fine blanking punch force data. Besides, other studies have investigated generic statistical, spectral and temporal features, e.g. obtained with the TSFEL library, and methods like PCA or UMAP resulting in abstract and less interpretable features. Related research which utilized deep learning, treated models as black boxes in which the learned features remained hidden. Against this background, z_0 may represent a novel feature for effective punch wear monitoring in fine blanking.

Concluding, the results indicate that the Ada-GVAE learned latent factors that indeed isolate real-world mechanisms causing the variation in the data. In particular, the model isolated force signal variations likely caused by human interventions during the measurement process and a remaining variation which is possibly caused by punch wear. Conversely, the normal β-VAE learned latent factors that are composed of a mix of the real-world causes. Results shown in appendix A and B) underline these findings.

Unlike related work, the machine learning approach presented in this article neither requires manual feature engineering or extraction nor obscures learned features with a black box model. It allows to detect new relevant features, but also biases in datasets, which could go unnoticed when only relying on a small set of knowledge-based, engineered features. Moreover, the features learned in this study are expressive in that they can be visualized through a generative model and appear to isolate real-world mechanisms (e.g., wear and measurement intervention). This may facilitate the interpretation of data and models by engineers, compared to the large sets of spectral, temporal and statistical features, including for example percentiles, autocorrelation, or param-

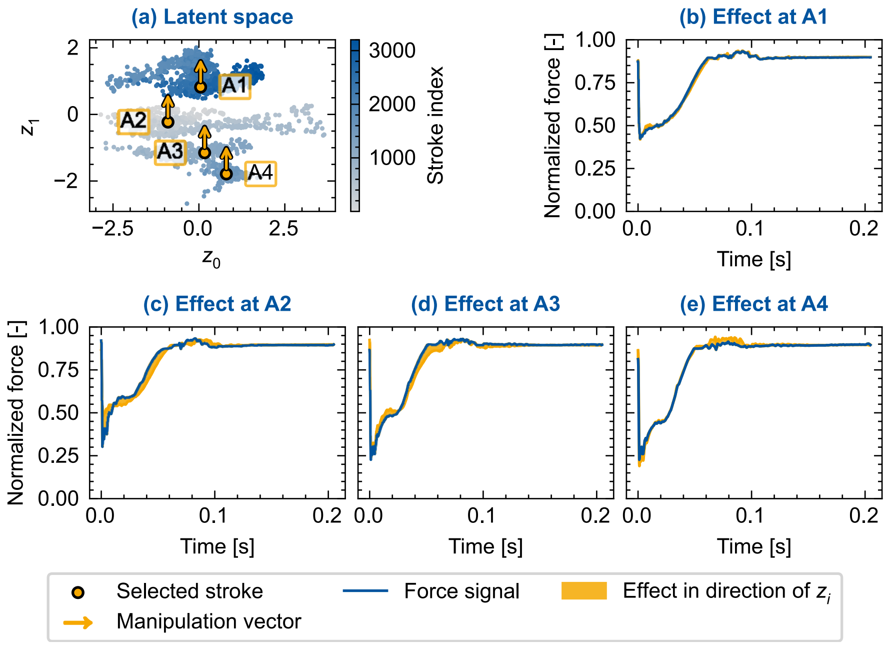


Fig. 8 Explanation of factor z_1: **a** manipulations in latent space, **b** minor effect of manipulation within cluster in direction of z_1, **c**–**e** effect of manipulation between clusters

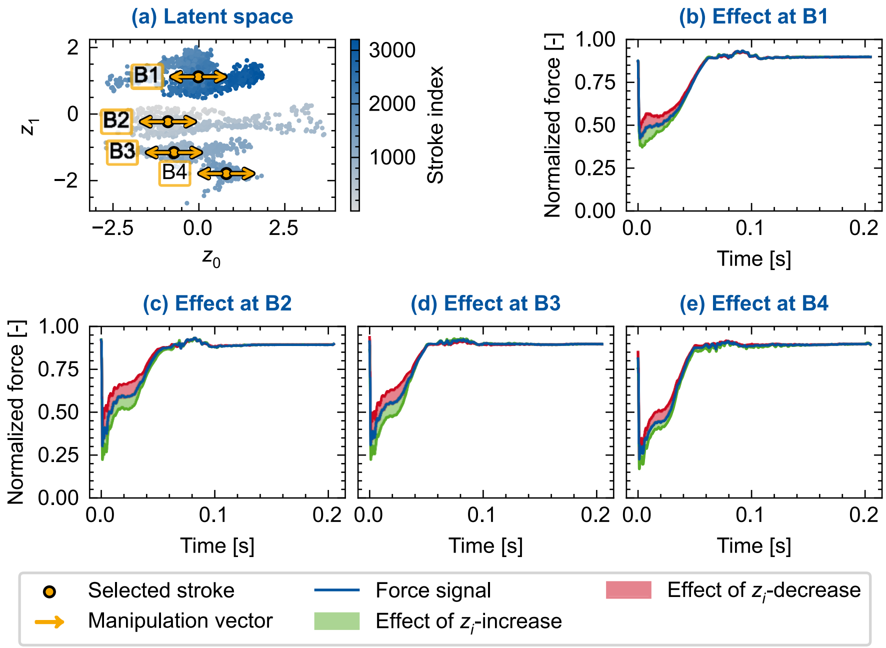


Fig. 9 Explanation of factor z_0: **a** manipulations in latent space, **b**–**e** consistent effect of manipulation within clusters in direction of z_0

eters of wavelet transforms, that some studies obtained with feature extraction libraries.

Explainable predictor

As outlined in Section "Dataset and methods", we combine the latent variable model (Ada-GVAE) with a predictive model that is optimized using supervised learning. Figure 10a shows the model predictions in comparison to the actual target values for hold-out test data. While locally predictions significantly deviate from their target values, the model generally captures the progression of the wear values well. With the coefficient of determination being $R^2 = 0.97$ on test data, the model would typically be considered to be a good model if only judged on its predictive performance.

Figure 10b depicts exemplary local feature importances, i.e. the extents to which features contributed to the model's wear prediction for a specific stroke. The signs of the importance values signify the direction of their impact (i.e., increase vs. decrease in predicted wear), while their absolute values quantify the magnitude of their contribution. The importance values were derived based on the gradient com-

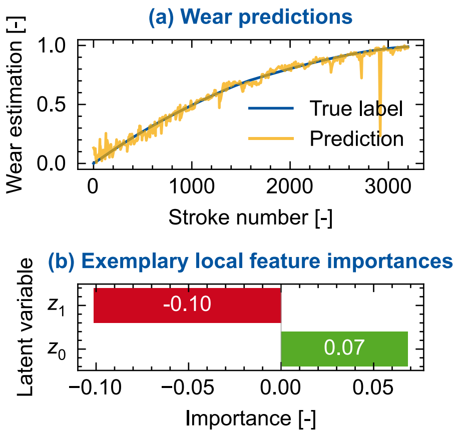


Fig. 10 Wear predictions and exemplary local feature importances for a randomly selected stroke

Table 1 Means and standard deviations of (absolute) feature importances for z_0 and z_1 over all test data

	Mean	SD
Importance of z_0	0.070	0.038
Importance of z_1	0.217	0.369

putation outlined in Section "Dataset and methods". By way of example, the depicted importance values highlight that the model's prediction significantly relies on the factor z_1. As discussed in Section "Latent factors explaining variation in fine blanking force data", z_1 encodes a bias induced by the wear measurement procedure. Table 1 shows the mean feature importances of factor z_0 and z_1 for the whole test dataset (with 480 strokes). On average, factor z_1 contributes approximately three times as much to the model predictions, indicating that the model significantly relies on the bias in general. Since, the executions of the wear measurements as well as the wear itself both correlate with time, it is plausible that the model exploited the bias. Nevertheless, the results underline that it is crucial to evaluate data-driven models not just by their predictive performance but also by considering xAI methods - at least when models are used that are not inherently interpretable like, e.g., linear regression models.

As a second analysis to study the importance of the two features to the model, an ablation study was conducted. This was done by averaging out the values of one latent factor, to isolate the contribution of the other factor to the model's wear estimation. The resulting model predictions are depicted in Fig. 11. The figure shows that the model's predictions largely rely on z_1 which encodes the bias originating from the wear measurement procedure. When the factor z_0 is averaged out, the model's predictions are still close to the original labels (see Fig. 11a). Conversely, the predictions are far off the true values when the factor z_1 is averaged out (see Fig. 11b).

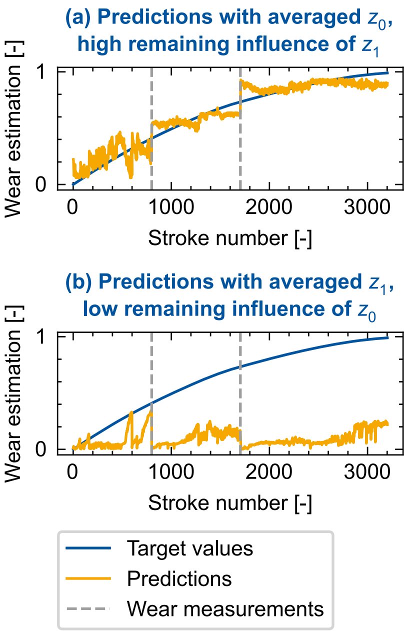


Fig. 11 Importance of feature derived from ablating single neurons

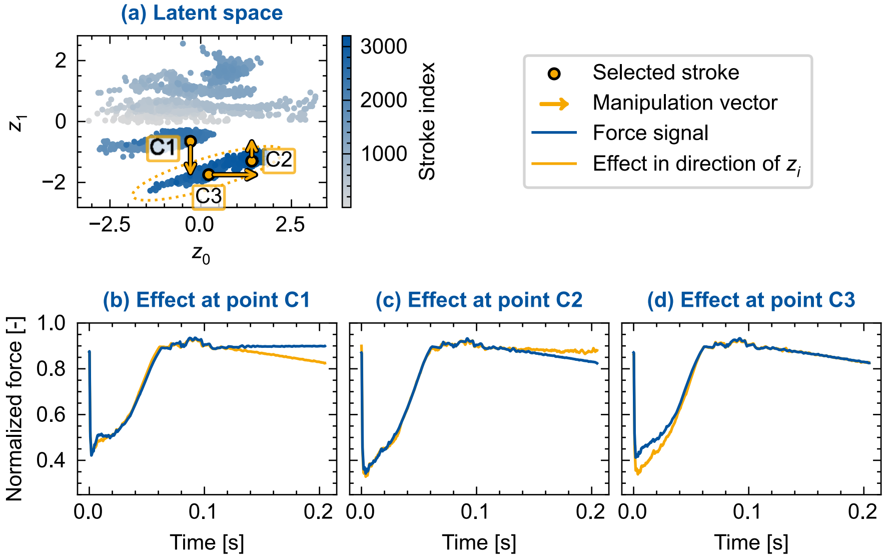
Figure 11**a** further underlines that z_1 encodes a bias from the measurement procedure. The wear predictions exhibit sudden jumps after each measurement, but (after the first measurement) remain rather stable between measurements when averaging out z_0. This suggests, that the model primarily uses z_1 to simply differentiate whether a stroke took place after the first measurement or after the second measurement. Before the the first measurement, the estimated wear steadily grows, suggesting that in this region z_1 also contains time-dependent information that is independent of the measurement interventions.

When the factor z_1 is averaged out, the estimated wear remains stable for some time before it peaks twice shortly before the first wear measurement (see Fig. 11b). Afterwards the estimated wear steadily grows between measurements. Potentially, this behaviour is caused by rapid initial wear increase (peaks before the first measurement) followed by a steady wear progression.

Disentanglement of synthetically injected features

The results presented so far indicate that the model successfully learned factors from the raw dataset, that align with real-world phenomena and allow to explain the observed data variation. However, the ground truth of the dataset is unknown. Consequently, it is unclear whether in reality there were additional phenomena that the model did not isolate from the previously presented factors and therefore remain hidden. To study whether the model would identify additional latent factors if they were present in the dataset, we introduced synthetic data generating factors to the dataset.

Fig. 12 Results for dataset with synthetically induced discrete feature; **a** manipulations in latent space, **b–d** effects of manipulations in the original space



First, we introduced a discrete change in all force signals after stroke number 2500 by manually adding a slope to the horizontal segment at the end the force signal. In practice, a discrete change in the signal might for example occur from events like tool changes. Figure 12 shows that the model successfully learned to isolate the feature. The group of synthetically changed signals form a new cluster in the latent space (see Fig. 12a, inside the dashed orange ellipse). When moving between this cluster and its neighboring cluster, the synthetically injected feature becomes present or goes absent (see Fig. 12b, c). When moving within the new cluster in direction of z_0, it becomes apparent that z_0 still only isolates the change in height in the left valley of the time series (see Fig. 12d). Moving in z_1-direction within the new cluster (correctly) has no effect.

Second, we injected the same factor synthetically into the dataset but as a continuous feature that starts at stroke 1,301 and linearly grows until stroke 2,500 where it ends with a magnitude that equals the previous discrete feature. Figure 13b exemplifies how the model isolated the continuous feature in factor z_0.

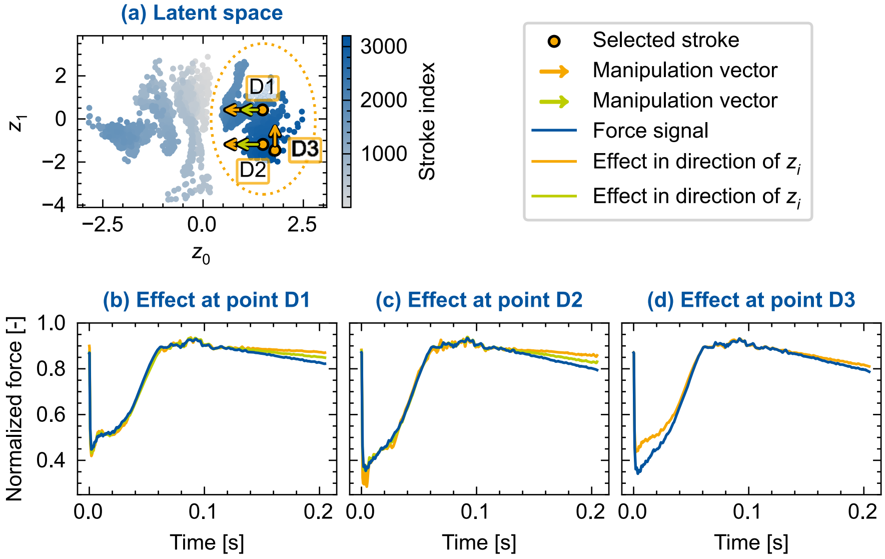
When a change in z_0 also leads to a transition between clusters, the synthetically induced factor starts to be superimposed with the force signal alteration caused by the wear measurement procedure (see larger manipulation in Fig. 13c). Figure 13d illustrates that a change parallel to the cluster boundaries (here in z_1-direction) still controls the phenomenon, which was previously hypothesized to be a result of wear. However, in contrast to Fig. 12d the factor also includes some variation of the synthetically induced feature this time.

Concluding, the experiments with the synthetic data further indicate that the model is capable of learning factors z_i that align with the actual, latent mechanisms that caused variation in the observed data. Especially, the discrete, synthetic feature was isolated well from the previously identified mechanisms. The continuously changing, synthetic feature was isolated as well, but two limitations became apparent. First, when the chosen number of factors z_i in the model is smaller than the actual number of data generating mechanisms, the mechanisms will locally overlap in a single factor z_i. Second, the disentanglement of different mechanisms became less accurate when two mechanisms changed continuously and simultaneously (see Fig. 12d vs. 13d). In practice, this scenario, where two causal mechanisms that correlate with time appear at the same time, might for example occur when wear grows steadily while at the same time material characteristics vary continuously along a coil (Ortjohann et al., 2024). This is also related to a more fundamental limitation affecting the results in general: the true data-generating mechanisms underlying the measurements are (inevitably) not fully known. For instance, it is unclear whether the force measurements used in this study contain any material-related effects, as well.

Conclusion

This study explored a potential solution to learn expressive xAI models directly from raw force signals of a fine blanking process for wear condition monitoring. We found that incorporating concepts from the field of causal inference into machine learning, as proposed by Locatello et al. (2020), allows to learn a model from raw fine blanking force signals that explains the variation in the measured data through latent factors that correspond to underlying real-world phenomena (RQ 1). For example, the model identified one factor that explains variation in the dataset, which is related to a

Fig. 13 Results for dataset with synthetically induced continuous feature; **a** manipulations in latent space, **b–d** effects of manipulations in the original space



bias caused by human interventions into the fine blanking machine to measure punch wear. Another factor explains an effect in the force data that is independent of the measurements and is potentially explained by the punch wear itself. We demonstrated that the learned latent variable model can be used to produce explanations for predictions of a downstream regression model for wear estimation (RQ 2). This approach enabled the detection of biases and may serve as a tool to derive novel insights regarding relationships between process signals and phenomena like tool wear. Lastly, we created semi-synthetic datasets, in which we synthetically introduced effects into the fine blanking force dataset, and tested whether the data-driven approach will correctly isolate them (RQ 3). We found that the model indeed identified the synthetic effects. In conclusion, the data-driven approach studied in this work provides a potential tool to (a) uncover biases in data and models and (b) derive potentially novel insights, e.g., regarding suitable indicators to monitor wear.

The experiments with semi-synthetic data also revealed limitations. When two mechanisms affecting the process data change simultaneously and continuously their separation into distinct latent factors became less accurate. In practice, this scenario might for example appear, when material properties steadily change along a sheet metal coil while tool wear is also continuously increasing. Moreover, the disentanglement of factors is limited when the number of latent dimensions of the model is chosen too low to capture all real-world mechanisms. A general limitation of the research presented in this paper is that the true mechanisms influencing the measured process signals are not fully known. Hence, it cannot be ruled out that the data contained any significant effects from unknown mechanisms. If there had been any other time-correlated mechanism, it could have overshadowed the effect of wear and thereby lead to a misinterpretation of the latent factor that was attributed to tool wear in this study. An example of such a mechanism could be a wear-independent monotonic change in process temperature. Another limitation originates from the closed tool design of the fine blanking process, which prevented wear measurements for each stroke. This necessitated the use of interpolation to estimate wear progression between available measurements. The interpolated values likely differ from the actual wear progression to some degree, introducing potential deviations between the wear prediction model and real-world behavior. However, this does not affect the general approach for obtaining an explainable wear monitoring model presented in this article.

More research is required with additional datasets to further confirm that the approach reliably identifies factors representing relevant real-world phenomena, suitable for, e.g., indirect data-driven wear condition or quality monitoring. Another interesting direction for future research is to investigate whether the generative latent variable model may also serve as a tool to remove known external influences or biases from datasets. In other words, is it possible to produce a dataset that reflects how the data would have looked like if the cause of a bias would not have been present to build more robust models. Taking the experiment studied in this article as an example, an idea could be to keep the values of z_1 (bias) constant while extrapolating the values of z_0 (remaining force variance) to estimate how the force signals would have progressed had the bias not been present. Further research is needed to investigate whether and how the effect encoded by latent factors z_0 (see figure 9) can be utilized for robust, feature-based wear monitoring.

The findings presented in this article suggest several practical implications for condition monitoring of industrial sheet metal shearing processes. First, considering model explainability during model evaluation is important even when models show good performance on hold-out test data. Methods of xAI reveal both spurious correlations that may lead to model failure as well as new insights regarding potentially useful wear indicators. Second, the results indicate that wear monitoring systems may be more effective when built upon features identified through machine learning, rather than being built upon the machine learning model itself. Since wear increases with time, other time-dependent phenomena (e.g., temperature increase) or interventions (e.g., machine stops or coil changes) affecting the process signal introduce potentially misleading patterns. By providing seemingly useful information to distinguish different points in time - and thus wear levels - they introduce spurious correlations misleading learning algorithms. A simple system that monitors the evolution of actually relevant features identified through machine learning may be more robust and interpretable. Third, for the fine blanking process examined in this study, features related to the punch force "valley" during the stripping segment (see figures 9) offer a promising foundation for developing reliable wear monitoring systems.

Appendix A

Results for fine blanking experiment *E4* (cf. Section "Fine blanking punch force dataset") also indicate that latent factors learned with the Ada-GVAE (see Fig. 14b) are better aligned with actual causal factors than those learned by the β-VAE (see Fig. 14a).

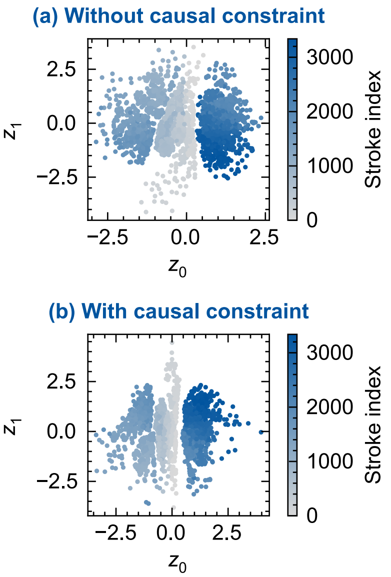


Fig. 14 **a** Latent space learned with model without ICM-based constraint. **b** Latent space learned with model with ICM-based constraint

Data points are separated along the z_0-axis. The gaps between the resulting clusters coincide with the points in time when the tool was disassembled to measure the punch wear. In Fig. 14b the gaps between the clusters are perpendicular to the z_0 axis indicating the the factor isolated a bias induced by the measurement procedure. Within each cluster an effect is isolated along factor z_1 that resembles the one also observed in experiment *E2* (cf. Appendix B and Section "Explainable predictor").

Appendix B

Figures 15, 16, 17, and 18 depict exemplary effects of manipulations along z_0 or z_1 in the latent space shown in Fig. 14b in Appendix A.

A manipulation in z_0-direction across clusters causes a change in the signal's overall shape as shown by way of example in Fig. 15. This effect was also observed in the other fine blanking experiment (cf. Section "Explainable predictor", Fig. 8) and likely originates from the tool disassembly performed to measure punch wear.

Manipulations within clusters in z_1-direction result in a change in the height of the valley in the force signal (see Figs. 16 and 17). This effect, independent from the data variation caused by the measurement procedure, was also observed in the other fine blanking experiment (cf. Section "Explainable predictor", Fig. 9).

The data points on the left-hand side in Fig. 14b are separated along the z_1-axis into two smaller clusters. Moving from one of these two clusters to the other in z_1-direction causes a shift in the jagged part after the upwards slope succeeding the valley in the signal (see Fig. 18). A z_1-manipulation within these clusters isolates the change in the valley as already illustrated in Fig. 17. Variance in the jagged part of the signal was also isolated by the model trained

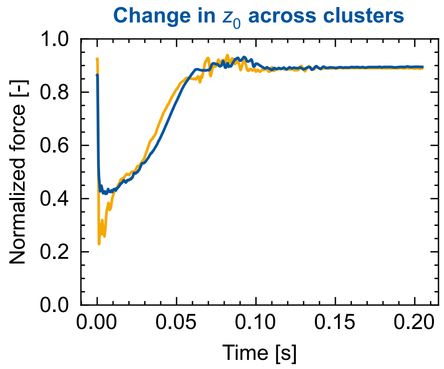


Fig. 15 Exemplary visualization of effect from changing between two clusters in z_0-direction (here from $\mathbf{z} = \begin{pmatrix}1.0 & 0.0\end{pmatrix}$ to $\mathbf{z} = \begin{pmatrix}0.0 & 0.0\end{pmatrix}$)

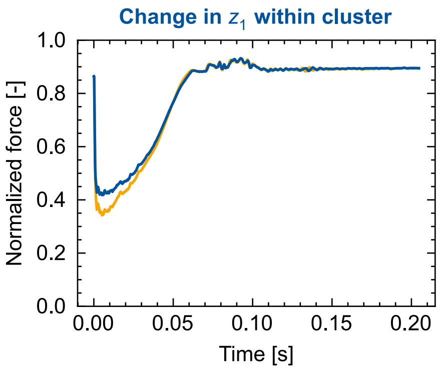


Fig. 16 Exemplary visualization of effect of manipulation within cluster in z_1-direction (here from $\mathbf{z} = \begin{pmatrix}1.0 & 0.0\end{pmatrix}$ to $\mathbf{z} = \begin{pmatrix}1.0 & 1.0\end{pmatrix}$)

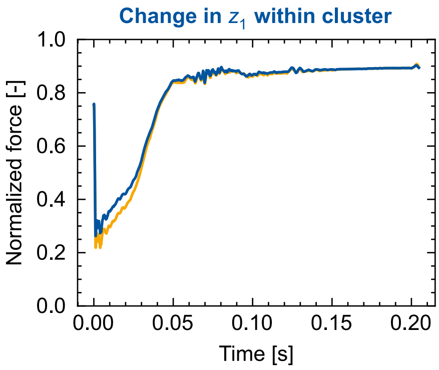


Fig. 17 Exemplary visualization of effect from changing between two clusters in z_0-direction (here from $\mathbf{z} = \begin{pmatrix}-1.0 & 0.5\end{pmatrix}$ to $\mathbf{z} = \begin{pmatrix}-1.0 & 2.0\end{pmatrix}$)

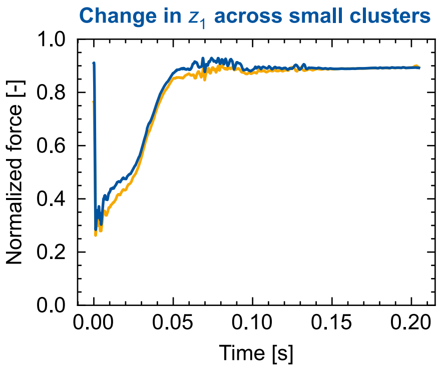


Fig. 18 Exemplary visualization of effect from changing between two clusters in z_0-direction (here from $\mathbf{z} = \begin{pmatrix}-1.0 & -0.25\end{pmatrix}$ to $\mathbf{z} = \begin{pmatrix}-1.0 & 0.25\end{pmatrix}$)

on data from the other fine blanking experiment (see Section "Explainable predictor", Fig. 8e),

Acknowledgements Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen der Exzellenzstrategie des Bundes und der Länder - EXC-2023 Internet of Production—390621612. [Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC-2023 Internet of Production—390621612.] Computational work was performed with computing resources granted by RWTH Aachen University under project rwth1559.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The fine blanking punch force data and punch wear measurements used in this work are available for download from the *Harvard Dataverse* repository at https://doi.org/10.7910/DVN/OYNDZO (Niemietz, 2022).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Asahi, S., Karadogan, C., Tamura, S., et al. (2021). Process data based estimation of tool wear on punching machines using TCN-Autoencoder from raw time-series information. *IOP Conference Series: Materials Science and Engineering, 1157*(1), 012078. https://doi.org/10.1088/1757-899x/1157/1/012078

Barandas, M., Folgado, D., Fernandes, L., et al. (2020). Tsfel: Time series feature extraction library. *SoftwareX, 11*, 100456.

Becker, M., Niemietz, P., & Bergs, T. (2024). Study on the explainability of deep learning models for time series analysis in sheet metal forming. *Procedia CIRP, 126*, 727–732. https://doi.org/10.1016/j.procir.2024.08.298

Behrens, B. A., Bouguecha, A., Vucetic, M., et al. (2016). Advanced Wear Simulation for Bulk Metal Forming Processes. *MATEC Web of Conferences, 80*, 04003. https://doi.org/10.1051/matecconf/20168004003

Bergs, T., Niemietz, P., Kaufman, T., et al. (2020). Punch-to-punch variations in stamping processes. In *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)* (pp. 000213–000218). https://doi.org/10.1109/sami48414.2020.9108761

Biegel, T., Jourdan, N., Hernandez, C., et al. (2022). Deep learning for multivariate statistical in-process control in discrete manufacturing: A case study in a sheet metal forming process. *Procedia CIRP, 107*, 422–427. https://doi.org/10.1016/j.procir.2022.05.002

Ge, M., Du, R., Zhang, G., et al. (2004). Fault diagnosis using support vector machine with an application in sheet metal stamping operations. *Mechanical Systems and Signal Processing, 18*(1), 143–159. https://doi.org/10.1016/s0888-3270(03)00071-2

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Griffin, J. M., Shanbhag, V. V., Pereira, M. P., et al. (2021). Application of machine learning for acoustic emissions waveform to classify galling wear on sheet metal stamping tools. *The International Journal of Advanced Manufacturing Technology, 116*(1–2), 579–596. https://doi.org/10.1007/s00170-021-07408-5

Higgins, I., Matthey, L., Pal, A., et al. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

Huang, C. Y., & Dzulfikri, Z. (2021). Stamping monitoring by using an adaptive 1d convolutional neural network. *Sensors, 21*(1), 262. https://doi.org/10.3390/s21010262

Jin, J., & Shi, J. (2000). Diagnostic feature extraction from stamping tonnage signals based on design of experiments. *Journal of Manufacturing Science and Engineering, 122*(2), 360–369. https://doi.org/10.1115/1.538926

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In *2nd International Conference on Learning Representations (ICLR 2014)*

Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends in Machine Learning, 12*(4), 307–392. https://doi.org/10.1561/2200000056. arXiv:1906.02691.

Klocke, F. (2013). *Manufacturing Processes 4: Forming*. Springer. https://doi.org/10.1007/978-3-642-36772-4

Kubik, C., Becker, M., Molitor, D. A., et al. (2023). Towards a systematical approach for wear detection in sheet metal forming using machine learning. *Production Engineering, 17*(1), 21–36. https://doi.org/10.1007/s11740-022-01150-x

Kubik, C., Hohmann, J., & Groche, P. (2021). Exploitation of force displacement curves in blanking feature engineering beyond defect detection. *The International Journal of Advanced Manufacturing Technology, 113*(1–2), 261–278. https://doi.org/10.1007/s00170-020-06450-z

Kubik, C., Knauer, S. M., & Groche, P. (2022). Smart sheet metal forming: importance of data acquisition, preprocessing and transformation on the performance of a multiclass support vector machine for predicting wear states during blanking. *Journal of Intelligent Manufacturing, 33*(1), 259–282. https://doi.org/10.1007/s10845-021-01789-w

Kubik, C., Molitor, D. A., Rojahn, M., et al. (2022). Towards a realtime tool state detection in sheet metal forming processes validated by wear classification during blanking. *IOP Conference Series: Materials Science and Engineering, 1238*(1), 012067. https://doi.org/10.1088/1757-899x/1238/1/012067

Kubik, C., Molitor, D. A., Varchmin, S., et al. (2023). Image-based feature extraction for inline quality assurance and wear classification in high-speed blanking processes. *The International Journal of Advanced Manufacturing Technology, 129*(11–12), 4883–4897. https://doi.org/10.1007/s00170-023-12653-x

Lee, W., Cheung, C., Chiu, W., et al. (1997). Automatic supervision of blanking tool wear using pattern recognition analysis. *International Journal of Machine Tools and Manufacture, 37*(8), 1079–1095. https://doi.org/10.1016/s0890-6955(97)88104-7

Liewald, M., Bergs, T., Groche, P., et al. (2022). Perspectives on data-driven models and its potentials in metal forming and blanking technologies. *Production Engineering, 16*(5), 607–625. https://doi.org/10.1007/s11740-022-01115-0

Locatello, F., Poole, B., Rätsch, G., et al. (2020). Weakly-Supervised Disentanglement Without Compromises. In Daumé, H., & Singh, A. (Eds.), *ICML'20: Proceedings of the 37th International Conference on Machine Learning* (Vol. 119, pp. 6348–6359). PMLR

Lundberg, S. M., & Lee, S. I., et al. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, & S. Bengio (Eds.), *Advances in neural information processing systems*. (Vol. 30). Curran Associates Inc.

Molitor, D. A., Kubik, C., Hetfleisch, R. H., et al. (2022). Workpiece image-based tool wear classification in blanking processes using deep convolutional neural networks. *Production Engineering, 16*(4), 481–492. https://doi.org/10.1007/s11740-022-01113-2

Niemietz, P. (2022). *Series of Time Series representing Fine-blanking Punch Force Strokes with Wear assessment*. https://doi.org/10.7910/DVN/OYNDZO

Niemietz, P., Fencl, M., & Bergs, T. (2023). Study on learning efficient stroke representations in clocked sheet metal processing: theoretical and practical evaluation. *Production Engineering, 17*(2), 279–289. https://doi.org/10.1007/s11740-023-01182-x

Niemietz, P., Kornely, M. J. K., Trauth, D., et al. (2022). Relating wear stages in sheet metal forming based on short- and long-term force signal variations. *Journal of Intelligent Manufacturing, 33*(7), 2143–2155. https://doi.org/10.1007/s10845-022-01979-0

Niemietz, P., Unterberg, M., Trauth, D., et al. (2021). Autoencoder based Wear Assessment in Sheet Metal Forming. *IOP Conference Series: Materials Science and Engineering, 1157*(1), 012082. https://doi.org/10.1088/1757-899x/1157/1/012082

Ortjohann, L., Peters, A., Gerhard, J., et al. (2024). Study on material-data-driven process parameterization in fine blanking. *Procedia CIRP, 126*, 733–738. https://doi.org/10.1016/j.procir.2024.08.300

O'Shaughnessy, M., Canal, G., Connor, M., et al. (2020). Generative causal explanations of black-box classifiers. In H. Larochelle, M. Ranzato, R. Hadsell, et al. (Eds.), *Advances in neural information processing systems* (pp. 5453–5467). Springer.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference. Foundations and learning algorithms. Adaptive computation and machine learning series*. The MIT Press.

Schenek, A., Görz, M., Liewald, M., et al. (2022). Data-driven derivation of sheet metal properties gained from punching forces using an artificial neural network. *Key Engineering Materials, 926*, 2174–2182.

Schlegel, C., Molitor, D. A., Kubik, C., et al. (2024). Tool wear segmentation in blanking processes with fully convolutional networks based digital image processing. *Journal of Materials Processing Technology, 324*, 118270. https://doi.org/10.1016/j.jmatprotec.2023.118270

Schölkopf, B., Locatello, F., Bauer, S., et al. (2021). Toward Causal Representation Learning. *Proceedings of the IEEE, 109*(5), 612–634.

Unterberg, M., Becker, M., Niemietz, P., et al. (2024). Data-driven indirect punch wear monitoring in sheet-metal stamping processes. *Journal of Intelligent Manufacturing, 35*(4), 1721–1735. https://doi.org/10.1007/s10845-023-02129-w

Unterberg, M., Voigts, H., Weiser, I. F., et al. (2021). Wear monitoring in fine blanking processes using feature based analysis of acoustic emission signals. *Procedia CIRP, 104*, 164–169. https://doi.org/10.1016/j.procir.2021.11.028

Virtanen, P., Gommers, R., Oliphant, T. E., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Voss, B. M., Pereira, M. P., Rolfe, B. F., et al. (2017). Using stamping punch force variation for the identification of changes in lubrication and wear mechanism. *Journal of Physics: Conference Series, 896*(1), 012028. https://doi.org/10.1088/1742-6596/896/1/012028

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.