Automated process monitoring in injection molding via representation learning and setpoint regression

Peng Yan*¶ Ahmed Abdulkadir*¶, Giulia Aguzzi†, Gerrit A. Schatte†,
Benjamin F. Grewe‡, and Thilo Stadelmann*§
{yanp, abdk, stdm}@zhaw.ch, {giulia.aguzzi,gerrit.schatte}@kistler.com, bgrewe@ethz.ch
*Centre for Artificial Intelligence, ZHAW School of Engineering, Winterthur, Switzerland
†Innovation Lab, Kistler Instrumente AG, Winterthur, Switzerland
‡ETH AI Center, ETH Zurich, Zurich, Switzerland
§Fellow, ECLT (European Centre for Living Technology, Venice, Italy) and Senior Member, IEEE
¶These authors contributed equally to this work and share the first authorship

***Abstract*—Online process monitoring is essential to detect failures and respond promptly in automated industrial processes such as injection molding. Traditional systems rely on experienced operators manually defining operational boundaries around a reference signal. We propose a data-driven representation that auto-tunes the sensitivity to a pre-set specificity threshold and automatically detects anomalies alongside interpretable indices that help identify root causes. Our automated system achieved an average AUC of 0.998 and detected 100 percent of the anomalies with the proposed dynamic calibration of the data-driven embedding method. The dynamic calibration, which accounted for drift, boosts the average specificity from 0.362 to 0.869. The outputs also indicate the direction and relative magnitude of characteristic deviations caused by machine parameters, including holding pressure, mold temperature, and injection speed. The AI-derived process boundaries are superior to manual annotation in tested real-world production environments.**

***Index Terms*—anomaly detection, time series, variational autoencoder, root-cause analysis, explainable AI, transfer learning**

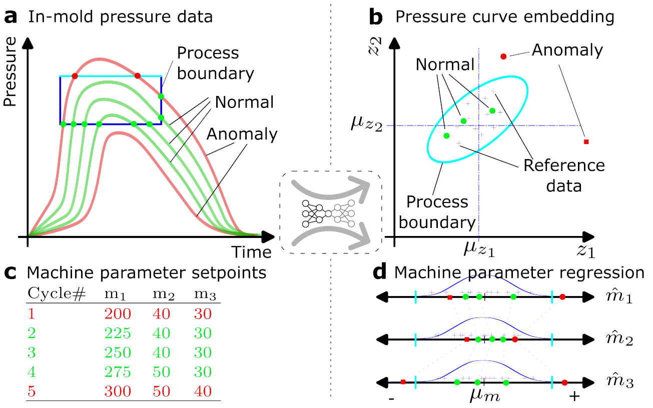

Fig. 1: Overview of this work: Manual process boundaries in the traditional approach (a) that depend on machine settings (c) and other factors and proposed data-driven process boundaries (b and d). (**a**) Examples of pressure curves are traditionally classified as *normal* (green) or *abnormal* (red) using evaluation labels (the rectangle with three blue and one cyan edge on the top). A normal curve must pass through at least one blue line and not go through a cyan line. A pressure curve is defined as an *anomaly* if it fails to meet any specified requirement. (**b**) Our deep learning-based approach maps the pressure curves into a two-dimensional embedding for automated analysis. Given a set of reference embedded curves (purple crosses), the adaptive process boundaries (cyan ellipse in b) are computed automatically. (**c**) Machine parameter setpoints control the injection molding process, and thus impact the part quality. (**d**) The neural network is also used to estimate machine parameters and the direction and magnitude of the standardized calibrated data distribution assist in detecting root causes.

I. Introduction

Injection molding is a cornerstone in mass-producing plastic parts with high precision [1]. This process involves injecting molten plastic into molds, forming the final products. Monitoring time series from in-mold sensors can help detect abnormal behaviors in real time [2]. For example, the transient temperature curve in each mold's cavity measured via an infrared sensor or thermocouple is expected not to change in stable operations [3]. Thus, the time series of sensors acquired under normal operation represents a reference for detecting deviations or anomalies. Early detection of anomalies or deviations can trigger timely intervention that lowers production costs [4]. Traditionally, anomaly detection has relied on the expertise of operators when analyzing complex data like pressure curves. However, this manual approach has limitations: it's labor-intensive and inconsistent due to varying operator expertise. It needs to be improved. Compounding the need for improvement is the declining number of apprenticeships for shopfloor workers in the European and US manufacturing ecosystems [5]. This is expected to result in fewer personnel operating more machines on the shop floor in the following decades. Increasing automation in the monitoring process aims to improve efficiency, accuracy, and product quality [6], [7].

In the injection molding process, sensors are installed in the mold to capture process signals during production. Among those signals, the pressure time series is a fingerprint of the injection molding process as it determines the crystallization

This work is supported by Innosuisse grant 62174.1 IP-ENG "DISTRAL".

of the molten substrate in the mold [8]. Its monitoring contains information about the process stability and, implicitly, the part quality [9] (or its deviation from the desired state) without measuring the dimensions, weight, and appearance of every part. It is, therefore, suitable for processes that need monitoring but not a 100%-part inspection. In industrial solutions like ComoNeo [10], operators traditionally set so-called evaluation objects (EOs) on these curves manually (see Fig. 1a) to define process boundaries. The EOs are set before the production nominally starts for each specific process and part and depend on the operator's experience with different materials, shapes, cycle times, or typical process parameters. They are hence highly subjective and their setup and any intervention, once an anomaly is detected, relies on the individual highly skilled and experienced professional. The location and extent of the boundaries determine the sensitivity and specificity of the anomaly detection process concerning certain types of anomalies. For example, a malfunction of the heating system of the substrate will impact the pressure curves differently than a change in the substrate. In theory, setting multiple EOs allows for fine-tuning the detection precisely, but in practice, possible failure scenarios with their impact are unknown. Thus, deeper automation is required to remove the manual definition of process boundaries and help identify root causes.

Machine learning, deep learning (DL), and statistical modeling offer transformative potential in this field [11]. By adopting a data-centric approach, we can surpass limitations imposed by manual tools [12], [13], allowing operators to focus on timely interventions and quality control. DL algorithms, trained on extensive injection molding data, can effectively represent pressure curves. Especially, variational autoencoder (VAE) [14] is a powerful DL framework for learning low-dimensional interpretable latent representations in an unsupervised manner [15] due to the unavailability of sufficient labeled data in the industrial setting [16]. Applying such DL models and using statistical methods for anomaly detection in new processes promises tuneable sensitivity, high accuracy in detecting anomalies, and a more efficient, consistent, and predictive monitoring strategy [17] while triggering appropriate interventions to maintain product quality and process efficiency [18]. Furthermore, utilizing a DL model that incorporates domain-specific knowledge can be instrumental in identifying these anomalies and infer the underlying causes.

In this paper, our main contribution is a practical solution for automated anomaly detection and machine parameter prediction for injection molding processes without feature engineering: (1) a novel VAE-based model that represents transient 1D sensor signals in a 2D latent space with robust out-of-distribution embedding capabilities which is used to detect deviations (anomalies) from reference curves representative of a stable process, and utilizes additional neurons that capture changes in machine parameter setpoints for interpretable outputs; (2) a dynamic calibration process that automatically determines the operational bounds of the process for anomaly detection with the ability to account for drifts; (3) prediction of set machine parameters to assist the root-cause identification.

We assess our approach's effectiveness with real-world data from injection molding production in which data originated from different production environments, achieving a remarkable AUC score of 0.998.

II. Related work

DL-based methods have been widely used for the analysis of industrial time series data as recently surveyed by Yan et al. [19]. These methods can be divided into reconstruction-based, forecasting-based, and other statistical methods. In deep reconstruction-based methods, the anomalies are recognized by comparing the deviation of the reconstructed time series sequence and the actual sequence using a defined threshold [20]. In contrast, for forecasting-based anomaly detection, only the forecasted sequence is compared with the ground truth sequence [21]. Most of the related anomaly detection methods assume that testing and training data follow the same data distribution, although by design, the "normal" state is unknown *a priori*. Regarding model architecture, Tuggener et al. have shown that the best-performing model architecture depends much more on the actual data set at hand than on the current state-of-the-art method on public benchmarks [22]. In this paper, we build on simple but well-tuned reconstruction-based models based on preliminary experiments.

Real-world industrial processes are dynamic and can be affected by a variety of production conditions [23]. Taking the injection molding process as an example, ambient environment, machine wear and tear, and other variables can exert an unknown but sometimes high influence on the manufacturing process. These conditions can then lead to a drift of the observed control data during production [24], [25]. The aforementioned anomaly detection methods might not be applicable. Various approaches have been proposed to deal with domain shift issues, such as transfer learning and domain adaptation. For instance, Saurav et al. [26] proposed a temporal model based on recurrent neural networks for time series anomaly detection to account for sudden or regular changes in normal behavior. The model is updated incrementally as new data becomes available and is capable of adapting to the changes in the data distribution. Instead of adapting directly to the new domain, Yang et al. seek to learn a domain invariant representation [27]. They train a DL model to extract a domain-invariant representation from normal data from the source domain and a limited number of normal data in the target domain. The domain-invariant representation is achieved by adversarial learning. Apart from that, transfer learning approaches have been applied in injection molding to transfer knowledge from one or more source setups to solve tasks in new setups [28], [29]. However, such approaches analyze machine parameters instead of sensory data collected directly from the manufacturing process, hence the transfer learning's effectiveness needs to be further validated. In this paper, we develop an auto-calibration mechanism based on statistical methods to mitigate data drift from sensor data while adding machine parameters for root cause analysis.

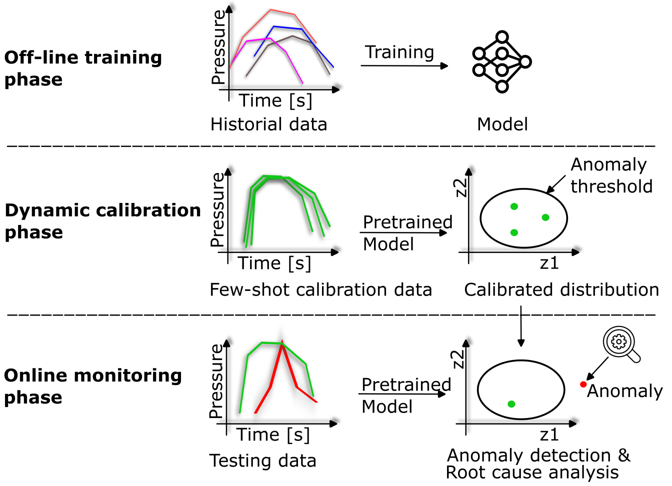

Fig. 2: Proposed data driven anomaly detection. A large data set of pressure curves is used to train a model to map pressure curves to a low-dimensional representation z (Here we only show the 2D representations of the pressure curve). Once the molding process is stable, a small sample of reference curves is mapped into the latent space to calibrate the process boundary. During online monitoring, an anomaly event is triggered if a data point is mapped to a position outside the anomaly threshold (calibrated process boundary).

Explainability and interpretability of DL models enable the quantification of the contribution of input features to the prediction results [30]. Alternatively, handcrafted features from the injection molding process have been investigated in [31]. An important factor analysis and a random forest model is used to identify the most influential features in cavity pressure and temperature data.

However, the relevant features may differ for different production conditions. Determining them requires expert knowledge and manual intervention and could miss other relevant information from the manufacturing processes. In this paper, we use the rich in-mold pressure signals without manual feature engineering, adding machine parameters to assist root cause analysis.

III. Proposed Methods

A. Overall pipeline

Fig. 2 depicts the overall pipeline of the proposed methods for data-driven anomaly detection. It consists of off-production representation learning, in-production dynamic calibration, and online monitoring. In the off-production training phase, historic pressure curves from various batches are used to train a useful low-dimensional representation of pressure curves. This is possible because the characteristic shape of the relevant curves is similar across productions (rapid ascent and a slower descent) but varies in the location of the peak and slopes.

Once the process is stable and produces the parts of the desired quality, the latent representation $\mathbf{z}_p$ of a small number $N_C = \{10 \dots 100\}$ of reference cycles is extracted using the pre-trained model. The embedded low-dimensional reference data are then used to estimate their distribution parameters. Under stable conditions, normal data from subsequent cycles are expected to follow the estimated distribution, while that of the anomalies is expected to deviate from it.

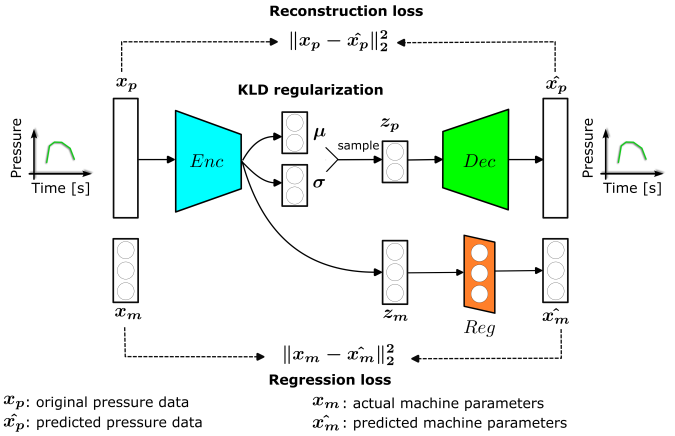
Reconstruction loss
$\|x_p - \hat{x_p}\|_2^2$
x_p
KLD regularization
$\hat{x_p}$
Pressure
Time [s]
Enc
μ
σ
sample
z_p
Dec
Pressure
Time [s]
x_m
z_m
Reg
$\hat{x_m}$
$\|x_m - \hat{x_m}\|_2^2$
Regression loss
x_p: original pressure data
$\hat{x_p}$: predicted pressure data
x_m: actual machine parameters
$\hat{x_m}$: predicted machine parameters

Fig. 3: Model architecture of the autoencoder with machine parameter prediction. The model consists of encoder Enc, decoder Dec, and regressor Reg. Training is driven by reconstruction loss of the pressure curves, regression loss of machine parameters, and distribution regularizer over $\boldsymbol{z_p}$. See text for further details.

Analogous and in parallel, a predictor for machine parameter setpoints $\mathbf{z}_m$ is trained (see Figure 3). Deviations from the calibrated reference distribution of these predictions can be used for root cause identification.

B. Deep learning architecture

We employ an auto-encoder that reconstructs the input through a bottleneck layer [32] to leverage unlabeled data and learn a useful representation. The network is implemented as a VAE with convolutional layers and a two-dimensional variational bottleneck without tied weights in the four hidden layers of the encoder and decoder. Fig. 3 shows our proposed network architecture, which consists of an encoder, a decoder, and a regressor.

Encoder. The encoder receives a pressure time series $\boldsymbol{p} = (p_i, \dots, p_D)$ and maps it into $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{z_m}$, where $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ represent the mean and variance of the estimated 2D Gaussian distribution, and $\boldsymbol{z_m}$ denotes the latent representation for 3 machine parameters. In summary, we can express the mapping as $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{z_m} = Enc(\boldsymbol{p})$. Moreover, $\boldsymbol{z_p}$ is sampled from the learned Gaussian distribution, given by $\boldsymbol{z_p} = \boldsymbol{\mu} + \varepsilon \cdot \boldsymbol{\sigma}$, where $\varepsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

Decoder. After encoding the pressure signal into a low-dimensional latent representation, we use the decoder Dec to reconstruct the pressure signal with the original dimension from $\boldsymbol{z_p}$, which can be denoted as $\hat{\boldsymbol{p}} = Dec(\boldsymbol{z_p})$.

Machine parameter regression. Three non-variational neurons are added to the VAE and connected to the bottleneck layer. They predict setpoints of holding pressure, injection speed, and molding temperature via an extra layer $\hat{\boldsymbol{m}}_i = Reg_i(\boldsymbol{z}_{m_i})$. This extra linear mapping between the intermediate representation $\boldsymbol{z}_m$ and the output $\hat{\boldsymbol{m}}$ allows production-batch-specific scaling and shift factors during training. For new batches (e.g. using different materials), the mapping between the pressure curve and machine parameters is expected to be proportional. For our application of detecting outliers and the direction of deviations in the machine parameter predictions, we use $\boldsymbol{z}_m$ for the analysis.

C. Loss function

The VAE minimizes the mean squared reconstruction error between the predicted and original pressure data. The two-dimensional latent variables are regularized to follow a bivariate standard Gaussian distribution using the KL divergence. Reconstruction loss and regularization are formulated as

$$L_{rec} = \frac{1}{ND} \sum_{n=1}^{N} \sum_{i=1}^{D} (p_i^n - \hat{p}_i^n)^2 \tag{1}$$

$$L_{kld} = -\frac{1}{2N} \sum_{n=1}^{N} \sum_{i=1}^{2} \left[1 + \log\left({\sigma_{n,i}}^2\right) - {\mu_{n,i}}^2 - {\sigma_{n,i}}^2\right] \tag{2}$$

where N and D are the sample size and the number of time stamps in the pressure time series, and p_i^n and $\hat{p}_i^n$ are the original and predicted pressure signal at time step i for n-th time series sample. In addition, we apply L2 loss between the predicted and original machine parameters:

$$L_{reg} = \frac{1}{NC} \sum_{n=1}^{N} \sum_{i=1}^{C} (m_i^n - \widehat{m}_i^n)^2 \tag{3}$$

where C is the number of machine parameters, and m_i^n and $\widehat{m}_i^n$ are the original and predicted machine parameters for n-th time series sample. The overall objective is to minimize

$$Loss = L_{rec} + \omega_{kld} \cdot L_{kld} + \omega_{reg} \cdot L_{reg} \tag{4}$$

where ω_{kld} and ω_{reg} are weights to balance losses. We also apply L2 regularization for the model weights during training.

D. Calibration and setting of process boundaries

Calibration is necessary to set meaningful process boundaries of an unknown process automatically. Using a few numbers (N_C) of reference curves from a stable process, the system fits a multivariate Gaussian distribution on the bottleneck node and outputs means $\boldsymbol{\mu}$. Additionally, it estimates an univariate Gaussian distribution for each unscaled machine-parameter neuron output z_{m_i}.

Given the empirical mean of the reference sample distribution, the distance of a new sample to the center of the reference distribution mean can be interpreted as an anomaly indicator. Specifically, the larger this distance, the more abnormal the new sample is considered. However, the scale of the Euclidean distance, especially in the multi-dimensional latent space, is meaningless. We overcome this by incorporating the variances and relating the resulting statistical measures analytically to the cumulative distribution function as follows: The distance standardized by either the covariance matrix in the multivariate distribution or the variance in the univariate distribution yields the Mahalanobis distance [33], respectively. Based on the Mahalanobis distance to the reference distribution d, a chi-squared distribution $\chi^2(d, k)$ is employed to express the fraction of samples i from the distribution where $d_i < d$, with $k = |\boldsymbol{\mu}|$ representing the degree of freedom. An operational bound B can thus be set such that a sample is considered an anomaly if $\chi^2(d, k) > B$, where $(1 - B)$ is the expected rate of false positives given a stationary process.

Since the machine parameters express physical properties that impact the process and are characterized by one-dimensional distributions, the system retains the sign of the relative deviation from the mean of the distribution. This facilitates root cause analysis, suggesting the direction of a parameter change as an intervention to recover a stray process.

When a process is non-stationary, meaning that the mean of the distribution gradually shifts over time, or if the noise exhibits heteroscedastic behavior (variance changes with time), the parameters of the initial calibration eventually lose their validity. This can happen due to external factors such as changes in environmental conditions like room temperature or humidity. We propose a dynamic update of the distribution parameters based on the latest N_C cycles.

IV. Experimental setup and evaluation

In this section, we describe the experimental setup and evaluation results, including the dataset, training procedure, and metrics.

A. Data: collection, handling, and preprocessing

IMM datasets. The injection molding machine (IMM) data is pooled from non-systematically acquired and incompletely labeled historical data provided by three vendors of injection molding systems. The resulting datasets were classified based on the availability of anomaly labels and machine parameters (see D0–D4 in Tab. I). Each mold has one to 8 Kistler pressure sensors attached to each cavity. The sensors measure the pressure in the cavity during each cycle and each sensor reports its pressure reading independently. The data are acquired at 16kHz for a cycle period between 0.6 and 20 seconds, depending on the size and shape of the products. After the acquisition, the time series signals are downsampled to a maximum of $1,000$ measurement points with irregular time intervals. The time series signals (per cavity/sensor) are stored in a database with various meta-data such as the time of day, mold ID, cavity ID, and a binary quality label. The labels are derived from EO, visual inspection, or physical measurements. The 5 datasets D0–D4 represent common production data but cover only a small fraction of the possible variability in mold, machine, part size, substrate, etc. In addition, a total of 540 pressure readings contained in two datasets (D5–D6) have been acquired within a controlled experimental design where we systematically varied the 3 machine parameters using 3

TABLE I: IMM datasets characteristics used for training, calibration, and testing.

Dataset	Sample size (train/val)	Cavity count	Cycle duration (s)	Anomaly rate	Different machine settings count (train/val)
D0	$5,208$ / $1,240$	8	6	assumed 0%	18 / 5
	Sample size (cal/test)				**Different machine settings count (cal/test)**
D1	70 / $2,905$	7	1.5	5.50%	unknown
D2	80 / $5,104$	8	2	0.78%	unknown
D3	80 / $4,240$	8	2.5	2.08%	unknown
D4	80 / $3,024$	8	1.5	1.59%	unknown
D5	10 / 260	1	10	assumed 0%	1 / 26
D6	10 / 260	1	10	assumed 0%	1 / 26

distinct values for each, yielding 10 cycles for each of the $3^3 = 27$ possible combinations.

Training and validation data. Out of a larger collection, we construct the unlabeled dataset D0 by pre-selecting pressure curves that visually appear similar to remove any environmental influence on the data apart from the known machine parameter settings. This yields $6,448$ pressure data with 23 machine settings out of 44 in the overall collection. Then, data from random 18 machine settings is used for training and the remaining 5 for validation.

Calibration and testing data. We have two groups of testing data for evaluating anomaly detection and machine parameter indication, respectively, using D1–D4 to evaluate time series anomaly detection performance and D5–D6 to evaluate the performance of machine parameter indication. For D1–D4, we only keep the data with valid EO labels, i.e., we keep the data generated when a specific number of consecutive normal samples are obtained. In D1, we identified an invalid cavity that recorded noise and excluded the corresponding data in the study. In all experiments, we use a calibration window size (number of consecutive normal pressure curves) of $N_C = 10$. For the dynamic calibration, we re-calculate the distribution parameters for each cycle with the previous $N_C = 10$ cycles.

Data preprocessing. Since the original time series data is sampled with irregular time intervals, we interpolate the pressure time series data into the same length (200) with equidistant time intervals. More specifically, the irregularly sampled time series data are resampled in 200 samples with regular intervals between 0 and 10 seconds. Additionally, we normalize pressure data by dividing by $1,000$ and subtracting 0.5. This roughly brings the observed data between -0.5 and 0.5. The inverse operation is applied to the network output to recover the original location and scale. This ensures a uniform scale across training, validation, calibration, and testing.

B. Training process

We train the model using stochastic gradient descent with batch size 16 for 200 epochs. Gradient updates are performed with the Adam optimization procedure [34]. The coefficients for the losses in Eq. (4) are set as $\omega_{kld} = 100, \omega_{reg} = 10$, balancing the terms such that they are of roughly equal importance. We train the model with early stopping when the validation loss does not decrease for 5 consecutive epochs.

C. Evaluation metrics

To evaluate the anomaly detection performance, we obtain the confusion matrix of the anomaly detection based on the actual (EO-based) and predicted (using process boundaries estimated from calibration data) labels of the testing dataset. Based on the confusion matrix, we then calculate the recall/true positive rate (TPR) and specificity. TPR measures the ratio of correctly classified normal samples to all actual normal samples. Specificity measures the ratio of the predicted negatives to the actual negatives. Furthermore, leveraging the Mahalanobis distance between the testing data and the estimated Gaussian distribution, we compute the area under the receiver operating characteristic curve (AUC score) as the main metric to assess the performance of anomaly detection.

D. Results

Overview. In the experiments, D5 and D6 share similar experimental settings across different products. For brevity, we only show one arbitrary dataset per figure as both datasets yield the same conclusion. Fig. 4 shows a comprehensive illustrative analysis of data from D6. The pressure data are mapped into the latent space to obtain μ and z_m. The green color represents calibration data (assumed normal cases), and the ellipses represent different estimated distributions for different confidence levels. A suitable operational bound is set to classify abnormal samples (we assume data from machine settings other than calibration data are abnormal). We also analyze the machine parameter indication by comparing the latent representation for the machine parameters. Note that the reconstructed signals and the proportionally scaled machine parameters are not of interest and, therefore, not reported here.

Anomaly detection. We provide qualitative and quantitative analyses for anomaly detection based on the deviation from the calibrated reference distribution of $\mathbf{z}_p$. Quantitative results are summarized in Tab. II. AUC scores are reported as average over all cavities of each dataset. In addition, we compute recall, specificity, and odds ratio at the operational bound of B=0.95. The odds ratio is computed as the ratio between precision and *a priori* anomaly rate. We compare results between one-time calibration and dynamic calibration. The results show that dynamic calibration outperforms one-time calibration in most instances since it consistently adapts to the data drift over time. As each cavity is calibrated separately and the anomaly value is expressed as the Mahalanobis distance to

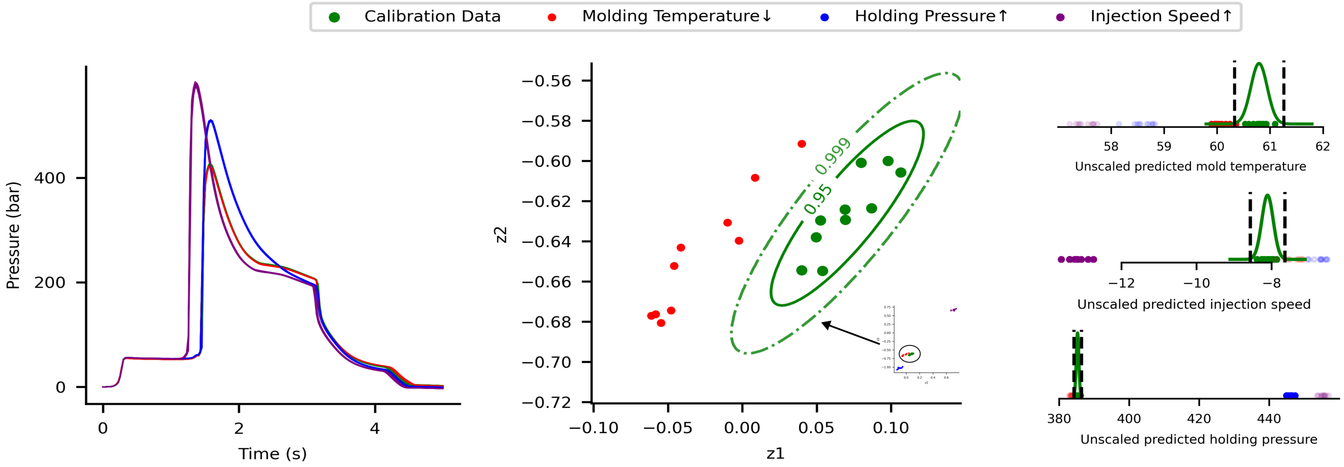

Fig. 4: **Left**: Pressure curves of 4 machine settings from D6, using 10 reference cycles from one machine setting as calibration data (green) and the rest for evaluation, where only a single machine parameter has been changed. **Middle**: Process boundaries at B=0.95 (solid line) and B=0.999 (dashed line). The small zoomed-out figure shows that the latent representations from increasing injection speed and holding pressure significantly deviate from the calibrated distribution. **Right**: Machine parameter predictions superimposed on the calibrated distribution (dotted lines indicate a deviation of 99.9% from the mean). It is sensitive to changes in the set values (e.g., for the red curves with set lower molding temperature, the molding temperature is indeed predicted lower, pointing at the anomaly's root cause as well as at an intervention). It is however less specific (i.e., a lower molding temperature is also predicted for the blue and purple curves, although temperature was kept constant there). This lack of disentanglement is partly due to the extreme nature of the parameter variations here as well as to actual physical couplings.

the cavity-specific distribution's center, the values are directly comparable across cavities.

Machine parameter prediction. Datasets D5–D6 are included for the evaluation of the prediction of the machine parameters because of their $3 \times 3 \times 3$ design in which each of the 3 machine parameters is changed threefold while keeping the others constant. The prediction of the holding pressure and the injection speed are in a positively linear relation, as shown in Fig. 5. The effect of the mold temperature is smaller than that of holding pressure and injection speed. For example, at 400 bar holding pressure and 30 mm/s injection speed, the prediction of the temperature is unreliable.

V. Discussion and conclusions

Summary. This study marks significant progress in the practical automation of injection molding process monitoring. It introduces a straightforward and effective method for domain adaptation and root cause identification, addressing prior challenges in transfer learning and unrealistic data requirements. We built on the state of the art in representation learning, employed statistical modeling, and injected domain knowledge to overcome three main limitations in the automation of injection molding anomaly detection and root cause analysis with a novel model and process. Specifically, the dependence on experienced personnel and manual configuration of the detection process has been reduced by using a variational auto-encoder-based representation that is dynamically calibrated with reference data during production. An objective choice of acceptable false positive rate replaces the subjective EOs. The designed machine-parameter-sensitive activations facilitate rule-based root cause identification that increases the practical value of the output. As such, the proposed system meets the needs of a self-calibrating system for anomaly detection. Our approach is implemented in the context of injection molding processes, but it also applies to other transient sensor data in which a deviation from a "reference" indicates an anomaly. The solution automatically sets process boundaries in a low-dimensional and explainable representation of pressure curves in a data-centric way. The detection process can be dynamically re-calibrated if desired. The principle is not limited to 1D time-series signals, but the representation-learning based framework generalizes to multi-dimensional data if the reference data is consistent and enough data is available. The same dynamic calibration process can be applied elsewhere as long as smooth data drift occurs. This conceptual advance has been filed for a European patent [35].

Discussion. We evaluated the feasibility of automatically setting meaningful process boundaries in unseen process data based on a small number of reference cycles. Data were acquired with sensors of a single modality, but the principle applies to other transient signals. The test data were distinct from the training data, most notably in cycle duration, which was 6 seconds in the training data but ranged from 1.5 to 10 seconds in the test data. The inductive bias of the unsupervised VAE made the representation useful because it was sensitive to changes in the input signals and detected the anomalies with high sensitivity when the operational borders were set to produce a false positive rate of 5 percent. At this threshold,

TABLE II: Evaluation of anomaly detection performance on four different datasets.

Dataset	**AUC**		**Recall** ($B = 0.95$)		**Specificity** ($B = 0.95$)		**Odds Ratio** ($B = 0.95$)	
	One-batch calibration	Dynamic calibration	One-batch calibration	Dynamic calibration	One-batch calibration	Dynamic calibration	One-batch calibration	Dynamic calibration
D1	0.9039	**0.9974**	**1.0**	**1.0**	0.2266	**0.8729**	1.4002	**5.8545**
D2	0.9620	**0.9974**	**1.0**	**1.0**	0.3146	**0.8593**	1.5769	**6.9487**
D3	**0.9980**	**0.9980**	**1.0**	**1.0**	0.6823	**0.8614**	**7.4856**	3.1971
D4	0.9425	**0.9998**	**1.0**	**1.0**	0.2245	**0.8809**	1.4151	**7.7862**
Avg	0.9516	**0.9982**	**1.0**	**1.0**	0.3620	**0.8686**	2.9695	**5.9466**

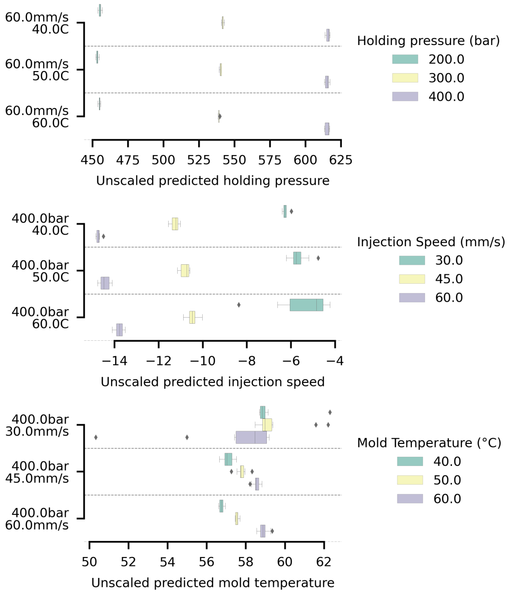

Fig. 5: Predicted machine parameters of a subset of D5. Each triplet of boxplots has 2 machine parameters fixed as indicated in the tick labels of the horizontal axis. The aim of the modeling is that the mini-batches produce correctly ordered predicted parameters $m_1 < m_2 < m_3$ or vice versa for correspondingly ordered set machine parameters. This is clearly achieved for holding pressure and injection speed. The scale and sign of the unscaled predictions are arbitrary.

the specificity was much lower than the expected percentage. By continuously re-calibrating the distribution parameters, the specificity increased above 80% but never reached the expected 95%. This is indicative of the data exhibiting slow drifts and that our assumption about the normality of the latent variables does not fully hold. We also observed that the estimated variance is not constant. In D3, for example, the variance of the first calibration batch was larger than throughout the remaining cycles. Here, the dynamic calibration reduced precision, triggering more false alarms.

Holding pressure and injection speed were proportionally predicted, yet the model did not fully disentangle them. This is not surprising as both parameters lead to a higher pressure peak. Changes in the mold temperature caused relatively little change in the pressure curves and therefore the system was not sensitive to those changes.

Limitations. Although the test set consisted of six batches (datasets), it only represents a small fraction of possible production data. A wider study is needed to explore the limits of the method in terms of generalization performance.

Some molds produce multiple parts and host one pressure sensor for each part. The system evaluation in this paper is limited to analyzing each cavity in isolation. This ignores potential information from interactions between channels. For example, when a single cavity is clogged, this would affect mainly the signals of that cavity. If the inlet were to be clogged, the pressure reading of all the cavities would be affected, triggering anomalies across all channels. Such scenarios would require updated training and/or infusion of domain expert knowledge at the level of root cause detection.

Another limitation is that when the reference cycles are very similar, the variance of the distribution is underestimated. This reduced the specificity. Currently, the model can adapt to slow data drift. However, the industrial process still needs to be supervised. When data constantly drifts in an unexpected direction, leading to an anomaly, the operator should intervene.

Conclusion and outlook. This study demonstrates progress of practical importance in the automation of injection molding process monitoring. The anomaly detection is sensitive and reliable if the process is stable. The dynamic calibration improves the reliability if there is a slow drift in the representation. The estimated distribution generally underestimates the variance which leads to a specificity that is lower than if the observed data would follow the theoretical model.

The proposed system can be improved in several ways without changing the architecture. If controlling the false positive rate is critical, the density estimation must be improved in future work. This could be achieved by either finding a distribution with few parameters (such as the Student distribution) that fits the observations better or fitting a more complex distribution such as a Gaussian mixture [36]. The latter case would require more training data, and in both cases, the appeal of easy-to-interpret elliptical process boundaries would be lost. Enlarging the training data set would help make the network generalize better. Readily available process data without machine parameters or quality labels could be used for pre-training. However, more variability in the data may entail the need to increase the network capacity by adding depth or breadth to the network. In our experiments, we found that insufficient capacity caused the network to get trapped in local minima where the posterior collapses on the prior [37].

The proposed system underestimated the variance as indicated by a specificity lower than 0.95 at an anomaly threshold corresponding to 0.05. The posterior variance σ of the VAE may be leveraged to estimate the batch-specific posterior better than with the distribution of the posterior mean μ alone.

Improving the disentanglement of the latent factors [38] is necessary to improve the system's utility further and could be achieved without a fundamental change in the architecture. On the other hand, transformer-based self-supervised models with dimensionality reduction capabilities [39], [40] may produce more accurate reconstructions and, therefore, a more expressive latent representation.

References

[1] M. Czepiel, M. Bańkosz, and A. Sobczak-Kupiec, "Advanced injection molding methods: Review," *Materials*, vol. 16, no. 17, 2023.

[2] S. Farahani, N. Brown, J. Loftis, C. Krick, F. Pichl, R. Vaculik, and S. Pilla, "Evaluation of in-mold sensors and machine data towards enhancing product quality and process monitoring via industry 4.0," *Int J Adv Manuf Technol*, vol. 105, pp. 1371–1389, 2019.

[3] T. Ageyeva, S. Horváth, and J. G. Kovács, "In-mold sensors for injection molding: On the way to industry 4.0," *Sensors*, vol. 19, no. 16, p. 3551, 2019.

[4] V. Rousopoulou, A. Nizamis, T. Vafeiadis, D. Ioannidis, and D. Tzovaras, "Predictive Maintenance for Injection Molding Machines Enabled by Cognitive Analytics for Industry 4.0," *Front Artif Intell*, vol. 3, 2020.

[5] R. Olsen, "Ausbildung: Interesse an Lehrberufen im Kunststoffbereich nimmt weiter ab," Kunststoff Information, Bad Homburg. Available online: https://www.kiweb.de/Default.aspx?pageid=199&docid=252599, 2023.

[6] M. Javaid, A. Haleem, R. P. Singh, and R. Suman, "Artificial intelligence applications for industry 4.0: A literature-based study," *J Ind Integr*, vol. 7, no. 01, pp. 83–111, 2022.

[7] L. Waltersmann, S. Kiemel, J. Stuhlsatz, A. Sauer, and R. Miehe, "Artificial intelligence applications for increasing resource efficiency in manufacturing companies—a comprehensive review," *Sustainability*, vol. 13, no. 12, p. 6689, 2021.

[8] M. C. Le, S. Belhabib, C. Nicolazo, P. Vachot, P. Mousseau, A. Sarda, and R. Deterre, "Pressure influence on crystallization kinetics during injection molding," *J Mater Process Technol*, vol. 211, no. 11, pp. 1757–1763, 2011.

[9] D. C. Angstadt and J. P. Coulter, "Cavity pressure and part quality in the injection molding process," in *Proc. ASME IMECE*, vol. 16622. American Society of Mechanical Engineers, 1999, pp. 7–17.

[10] "Process monitoring and control system ComoNeo for injection molding," Available online: https://www.kistler.com/DE/en/cp/process-monitoring-and-control-system-comoneo-5887a/P0000382, 2023.

[11] P. Waibel, S. Haltmeier, R. Vaculik, T. Wuhrmann, and N. Pascher, "Cavity pressure-based machine learning service for advanced injection molding processes," White paper, avaiable online: https://www.kistler.com/INT/en/monitoring-and-control-of-injection-molding-processes/C00000041, 2021.

[12] T. Stadelmann, T. Klamt, and P. H. Merkt, "Data centrism and the core of data science as a scientific discipline," *Archives of Data Science, Series A*, vol. 8, no. 2, 2022.

[13] P.-P. Luley, J. M. Deriu, P. Yan, G. A. Schatte, and T. Stadelmann, "From concept to implementation: The data-centric development process for AI in industry," in *Proc. SDS*, 2023, pp. 73–76.

[14] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *Stat*, vol. 1050, p. 1, 2014.

[15] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "Beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. ICLR*, 2016.

[16] T. Stadelmann, M. Amirian, I. Arabaci, M. Arnold, G. F. Duivesteijn, I. Elezi, M. Geiger, S. Lörwald, B. B. Meier, K. Rombach *et al.*, "Deep learning in the wild," in *Proc. ANNPR*, 2018, pp. 17–38.

[17] V. Ketonen and J. O. Blech, "Anomaly detection for injection molding using probabilistic deep learning," in *Proc. ICPS*, 2021, pp. 70–77.

[18] J. Gim and B. Rhee, "Novel analysis methodology of cavity pressure profiles in injection-molding processes using interpretation of machine learning model," *Polymers*, vol. 13, no. 19, p. 3297, Jan. 2021.

[19] P. Yan, A. Abdulkadir, P.-P. Luley, M. Rosenthal, G. A. Schatte, B. F. Grewe, and T. Stadelmann, "A comprehensive survey of deep transfer learning for anomaly detection in industrial time series: Methods, applications, and directions," *IEEE Access*, 2024.

[20] T. Stadelmann, V. Tolkachev, B. Sick, J. Stampfli, and O. Dürr, "Beyond ImageNet: deep learning in industrial practice," *Applied data science: lessons learned for the data-driven business*, pp. 205–232, 2019.

[21] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM CSUR*, vol. 54, no. 2, pp. 1–38, 2021.

[22] L. Tuggener, J. Schmidhuber, and T. Stadelmann, "Is it enough to optimize CNN architectures on ImageNet?" *Front Comput Sci*, vol. 4, p. 1041703, 2022.

[23] B. Maschler, H. Vietz, H. Tercan, C. Bitter, T. Meisen, and M. Weyrich, "Insights and example use cases on industrial transfer learning," *Proc. CIRP*, vol. 107, pp. 511–516, 2022.

[24] C. H. Tan, V. C. Lee, M. Salehi, S. Marusic, S. Jayawardena, and D. Lucke, "A fully unsupervised and efficient anomaly detection approach with drift detection capability," in *Proc. ICDMW*, 2021, pp. 312–321.

[25] D. Wu, D. Zhou, and M. Chen, "Probabilistic stationary subspace analysis for monitoring nonstationary industrial processes with uncertainty," *IEEE TII*, vol. 18, no. 5, pp. 3114–3125, 2022.

[26] S. Saurav, P. Malhotra, V. TV, N. Gugulothu, L. Vig, P. Agarwal, and G. Shroff, "Online anomaly detection with concept drift adaptation using recurrent neural networks," in *Proc. CODS-COMAD*, 2018, pp. 78–87.

[27] Z. Yang, I. Soltani, and E. Darve, "Anomaly detection with domain adaptation," in *Proc. CVPR*, 2023, pp. 2957–2966.

[28] Y. Lockner, C. Hopmann, and W. Zhao, "Transfer learning with artificial neural networks between injection molding processes and different polymer materials," *J Manuf Process*, vol. 73, pp. 395–408, 2022.

[29] H. Tercan, A. Guajardo, and T. Meisen, "Industrial transfer learning: Boosting machine learning in production," in *Proc. INDIN*, vol. 1, 2019, pp. 274–279.

[30] T. Chen, V. Sampath, M. C. May, S. Shan, O. J. Jorg, J. J. Aguilar Martín, F. Stamer, G. Fantoni, G. Tosello, and M. Calaon, "Machine learning in manufacturing towards Industry 4.0: From 'for now' to 'four-know'," *Applied Sciences*, vol. 13, no. 3, 2023.

[31] Z.-H. Wang, F.-C. Wen, Y.-T. Li, and H.-H. Tsou, "A novel sensing feature extraction based on mold temperature and melt pressure for plastic injection molding quality assessment," *IEEE Sens. J.*, vol. 23, no. 7, pp. 7451–7459, 2023.

[32] N. Simmler, P. Sager, P. Andermatt, R. Chavarriaga, F.-P. Schilling, M. Rosenthal, and T. Stadelmann, "A survey of un-, weakly-, and semi-supervised learning methods for noisy, missing and partial labels in industrial vision applications," in *Proc. SDS*, 2021, pp. 26–31.

[33] P. C. Mahalanobis, "On the generalized distance in statistics," *Proceedings of the National Institute of Sciences of India*, pp. 49–55, 1936.

[34] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, San Diega, CA, USA, 2015.

[35] Schatte, Gerrit A. and Abdulkadir, Ahmed and Yan Peng and Rosenthal, Matthias and Schwizer, Simone and Wildmann, Damian and Vaculik, Robert and Aguzzi, Giulia and Schirmer, Mathias and Thilo Stadelmann, "Verfahren und Vorrichtung zur Überwachung eines zyklischen Herstellprozesses," Swiss Patent Request EP23 212 020, Nov 4, 2023.

[36] T. Stadelmann and B. Freisleben, "Dimension-decoupled Gaussian mixture model for short utterance speaker recognition," in *Proc. ICPR*, 2010, pp. 1602–1605.

[37] Y. Wang, D. Blei, and J. P. Cunningham, "Posterior collapse and latent variable non-identifiability," *Proc. NeuRIPS*, vol. 34, pp. 5443–5455, 2021.

[38] S. Hahn and H. Choi, "Disentangling latent factors of variational autoencoder with whitening," in *Proc. ICANN*. Springer, 2019, pp. 590–603.

[39] R. Ran, T. Gao, and B. Fang, "Transformer-based dimensionality reduction," *arXiv:2210.08288*, 2022.

[40] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," in *Proc. ICCV*, October 2021, pp. 11 936–11 945.