PAPER

A Method for Uncertainty Quantification in Virtual Metrology Models: Acoustic-Emission-Based Quality Prediction in Micro Crown Gear Manufacturing

Ali Bilen[1,*], Max Decman[1], Prof.Dr.-Ing. Florian Stamer[2] and Prof.Dr.-Ing. Gisela Lanza[1,3]

[1]wbk - Institute of Production Science, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Institute of Production Technology and Systems (IPTS), Leuphana University of Lüneburg, Lüneburg, Germany
[3]Global Advanced Manufacturing Institute (GAMI), Suzhou SILU Production Engineering Services Co., Ltd., Suzhou, P. R. China
[*]Author to whom any correspondence should be addressed.

E-mail: ali.bilen@kit.edu

Keywords: uncertainty quantification, virtual metrology, acoustic emission

Abstract

This paper proposes a method to quantify the uncertainty of machine learning prediction-based measurement chains building on a feasibility investigation of AE-based virtual metrology for micro crown gear manufacturing. Providing reliable uncertainty statements is essential if regression models are to support conformity decisions or partially replace physical measurements. The established metrological framework of the *Guide to the Expression of Uncertainty in Measurement* (GUM) and GUM Supplement 1 provides a basis for uncertainty propagation. However, it treats a learned model typically as deterministic and epistemic uncertainty is not represented. This limitation of application to data-driven models is discussed. Building up on the discussion, we formulate an uncertainty model for virtual metrology that explicitly incorporates epistemic model uncertainty alongside stochastic input and label uncertainties considering current approaches of the state of the art. These contributions are combined within a Monte Carlo–based propagation framework. The resulting methodology yields predictive distributions and coverage intervals for VM outputs, enabling traceable and decision-relevant uncertainty reporting for AE-based quality prediction in micro-machining.

1 Introduction

Micro-manufactured functional components impose extremely high demands on geometric accuracy and quality assurance. In particular, micro gears are key elements in medical devices, microrobotics, and precision drive systems, where tolerances in the single-digit micrometer range are functionally critical (Härtig, Kniel, and Rost 2009; Goch et al. 2023). At these scales, conventional quality assurance relies on complex and time-consuming measurement chains that are spatially and temporally decoupled from production. As a result, quality information is only available with significant delay, limiting timely process control and increasing scrap rates (Lanza et al. 2019; R. Schmitt and Dietrich 2023).

These challenges are particularly pronounced for micro crown gears used in dental drive systems, which must transmit rotational speeds of up to 200 000 rpm at nearly constant torque (VDI 2731 2009). Even minor geometric deviations can severely affect functional performance, while dedicated standards and established evaluation strategies are largely lacking for this gear type (VDI 2731 2009; Tao et al. 2023). Consequently, there is a strong need for process-integrated quality monitoring concepts that enable rapid feedback without compromising cycle time.

Virtual metrology (VM) addresses this need by predicting quality-relevant characteristics directly from process data rather than relying exclusively on physical measurements (Chang et al. 2006). This is particularly attractive in micro-manufacturing, where direct measurements are slow and costly. In contrast to conventional approaches, VM enables the generation of quality information significantly earlier, either inline or even in-process (Lanza et al. 2019).

As illustrated in Figure 1, this shift in information availability reduces the control dead time within the quality control loop. Instead of reacting to deviations only after delayed offline measurements, quality information becomes available during production, enabling faster and more targeted process adjustments. This is especially beneficial for processes with short cycle times and high sensitivity to parameter variations, such as micro gear hobbing.

Among various process signals, acoustic emission (AE) sensing has emerged as a promising technology for machining applications, as it is highly sensitive to cutting dynamics and tool–workpiece interactions. Previous studies have demonstrated a strong correlation between AE features and functional geometry in micro gear hobbing, highlighting the potential of AE-based predictive quality models (Schiller et al. 2023). However, by replacing direct measurement with data-driven inference, VM fundamentally shifts the problem from measurement accuracy to predictive uncertainty.

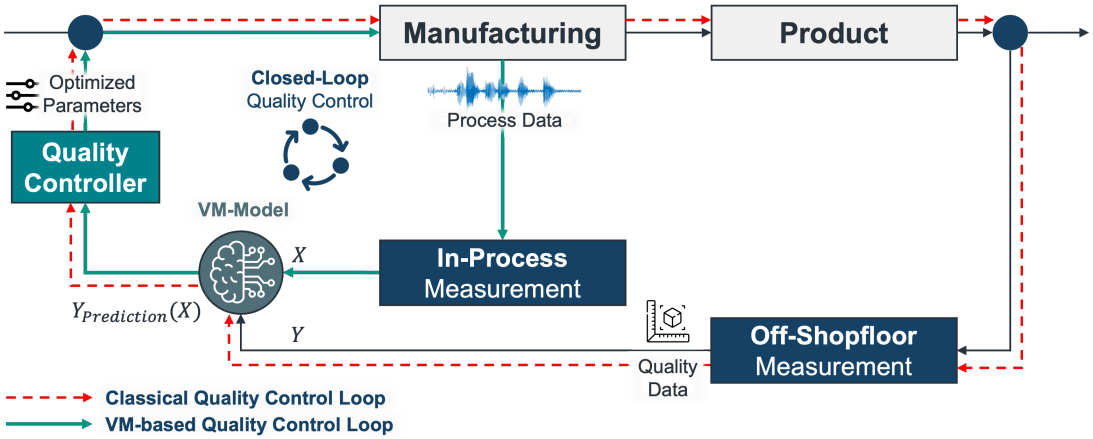


Figure 1: Comparison of quality control loops with and without virtual metrology. Virtual metrology enables earlier access to quality information by shifting measurement from offline to inline or in-process stages, thereby reducing control dead time (Bilen, Skade, et al. 2025; Lanza et al. 2019).

If such models are to support conformity decisions or partially replace physical measurements, their predictions must be accompanied by reliable and interpretable uncertainty statements. While the *Guide to the Expression of Uncertainty in Measurement* (GUM) provides a well-established framework, its direct application to data-driven regression models is limited, particularly with respect to epistemic uncertainty arising from limited data and model imperfections.

This paper therefore makes two contributions. First, it investigates the feasibility of acoustic-emission-based virtual metrology for predicting geometric quality characteristics of micro crown gears. Second, it proposes a GUM-inspired methodology to quantify the uncertainty of such prediction-based measurement chains by jointly considering sensor-, label-, and model-related uncertainty con-

tributions. The proposed approach enables traceable and decision-relevant uncertainty statements for virtual metrology in micro-manufacturing.

2 Fundamentals

2.1 Fundamentals of the GUM Uncertainty Framework

In high-precision manufacturing and quality assurance, such as the production of micro crown gears, reporting a measured value alone is insufficient. A complete measurement result must always be accompanied by an associated measurement uncertainty, which characterizes the dispersion of values that can reasonably be attributed to the measurand. Measurement uncertainty reflects the quality of the measurement process and directly affects conformity assessment, as it effectively reduces the usable tolerance range in accordance with ISO 14253-1.

GUM framework The internationally accepted basis for evaluating measurement uncertainty is the *Guide to the Expression of Uncertainty in Measurement* (GUM, ISO/IEC Guide 98-3). Within this framework, the measurement process is described by a mathematical model in which the output quantity Y is expressed as a function of N input quantities X_i:

$$Y = f(X_1, X_2, \ldots, X_N). \tag{1}$$

Uncertainties of the input quantities are classified as Type A, obtained from statistical analysis of repeated measurements, or Type B, derived from external information such as calibration certificates, manufacturer specifications, or prior knowledge. In the classical GUM approach, uncertainty propagation is performed by linearizing the measurement model using a first-order Taylor expansion, resulting in the combined standard uncertainty of the output quantity.

GUM Supplement 1 and Monte Carlo propagation The classical GUM approach relies on approximate linearity of the measurement model and an approximately normal distribution of the output quantity. These assumptions are often violated in complex, nonlinear models or when asymmetric input distributions are present, as is frequently the case for optical measurement systems and micro-manufacturing processes. To address such situations, GUM Supplement 1 introduces a Monte Carlo based propagation of probability distributions.

In this approach, probability density functions are assigned to all input quantities and repeatedly sampled. For each Monte Carlo run, the sampled inputs are propagated through the measurement model, yielding a numerical approximation of the output distribution. This method avoids model linearization and normality assumptions, providing robust uncertainty estimates even for strongly nonlinear models and non-Gaussian distributions.

Coverage intervals A key result of uncertainty evaluation is the coverage interval $[Y_\mathrm{L}, Y_\mathrm{U}]$, which contains the true value of the measurand with a specified coverage probability p:

$$\Pr(Y_\mathrm{L} \leq Y \leq Y_\mathrm{U}) = p.$$

While classical GUM commonly determines coverage intervals using an expanded uncertainty with a coverage factor k, GUM Supplement 1 derives coverage intervals directly from the simulated output distribution. Appropriate quantiles of the Monte Carlo sample are used to obtain coverage intervals for a given probability level (e.g. $p = 95\,\%$ or $p = 99\,\%$), enabling a consistent treatment of asymmetric and non-normal output distributions.

2.2 Uncertainty in Machine Learning

While the GUM framework and its Supplement 1 provide a robust foundation for traditional metrology, their direct application to data-driven virtual metrology and machine learning (ML) models reveals fundamental limitations. In classical GUM, the measurement model f is assumed to be an explicitly known physical or analytical relationship. In contrast, ML algorithms approximate this relationship based on finite and potentially noisy training data, resulting in an empirical model f_ML.

Treating the learned model f_ML merely as a fixed, deterministic "black box" during Monte Carlo propagation primarily captures *aleatoric uncertainty*—that is, the variance originating from stochastic inputs such as sensor noise or feature repeatability. However, this approach systematically neglects *epistemic uncertainty*, which arises from model imperfections, limited generalization capabilities, and biases in the training data (Hora 1996; Kiureghian and Ditlevsen 2009; Hüllermeier and Waegeman

2021). Consequently, applying standard GUM propagation directly to machine learning models often yields computed coverage intervals that are unrealistically narrow or systematically shifted with respect to the true measurand.

In statistical learning theory, epistemic uncertainty is closely related to the bias–variance tradeoff and reflects the incomplete knowledge of the true underlying data-generating process (Geman, Bienenstock, and Doursat 1992). Unlike aleatoric uncertainty, which represents irreducible stochastic variability, epistemic uncertainty can in principle be reduced through additional training data, improved model structures, or expanded coverage of the input space. From a metrological perspective, neglecting this contribution contradicts the fundamental requirement that all relevant uncertainty sources must be explicitly identified and quantified in the uncertainty budget.

To formulate a metrologically sound and traceable uncertainty statement for virtual metrology, the standard GUM framework must therefore be methodologically extended to explicitly quantify and incorporate these epistemic, model-related uncertainty contributions. In recent years, several conceptual approaches have been proposed to approximate epistemic uncertainty in machine learning models. One pragmatic strategy is to estimate the dispersion of prediction residuals on an independent holdout dataset that was not used during model training. The resulting residual distribution provides an empirical estimate of model-form deviations within the explored feature space and can therefore serve as a data-driven approximation of epistemic uncertainty (Kendall and Gal 2017; Hüllermeier and Waegeman 2021).

A complementary interpretation treats epistemic model deviation analogously to systematic effects in classical metrology. In this view, the deterministic model output $f_{\mathrm{ML}}(X)$ is augmented by an additive stochastic term representing the discrepancy between the learned model and the unknown physical relationship. This formulation allows model-form uncertainty to be integrated into the probabilistic measurement model as an additional uncertainty contributor, thereby extending the GUM-compatible uncertainty propagation framework.

In practice, such model deviations are commonly represented as random variables parameterized by empirical statistics of prediction errors, for example assuming Gaussian distributions with parameters estimated from residual analysis. While this representation is necessarily an approximation, it provides a tractable means of incorporating epistemic effects into uncertainty propagation schemes (Kendall and Gal 2017; Hüllermeier and Waegeman 2021).

Finally, the validity of the resulting uncertainty model must be assessed independently of the estimation procedure. A common strategy is to evaluate the empirical coverage of predicted uncertainty intervals on a fully unseen validation dataset. If the proportion of observations falling within the predicted intervals approximately matches the nominal coverage probability, the uncertainty model can be considered statistically calibrated within the explored feature space (Hüllermeier and Waegeman 2021). Such independent consistency checks are essential to ensure that the combined treatment of aleatoric and epistemic uncertainty yields reliable and interpretable uncertainty statements for data-driven virtual metrology systems.

3 State of Research

3.1 Acoustic-Emission-Based Quality Prediction

The use of acoustic emission (AE) signals in manufacturing has evolved significantly over the past decades. Early research predominantly focused on tool condition monitoring and process stability, aiming at the detection of tool wear, tool breakage, or unstable cutting conditions. In these studies, AE signals were primarily interpreted as indirect indicators of process health rather than as carriers of explicit workpiece-related information.

More recent research increasingly targets the prediction of product-related quality characteristics, reflecting a paradigm shift toward predictive quality and virtual metrology concepts. Initial approaches in this direction mainly addressed indirect quality indicators, most notably surface roughness. For example, Beggan et al. (1999) analyzed deviations between measured AE signals and theoretical models to infer surface quality during turning, while Albers et al. (2017) identified AE-based indicators correlating with workpiece diameter and surface roughness in serial production environments.

Beyond surface-related metrics, recent studies have demonstrated that AE signals can encode information about more complex geometric characteristics. In the context of gear manufacturing, Schiller et al. (2023) showed that supervised machine learning models enable the prediction of discrete quality classes for profile deviations (F_α) and the overall transmission error during micro gear hobbing. Complementarily, Han et al. (2022) proposed a physics-inspired approach in which vibration signals are transformed into displacement representations, allowing the estimation of explicit geometric deviations such as profile, pitch, and flank line errors.

A related line of research focuses on the detection of discrete defects rather than continuous geometric quantities. Examples include the identification of pores in cast components during machining (Gauder et al. 2022) or the detection of gear faults such as pitting and tooth breakage (Erkaya and Ulus 2016). While these approaches demonstrate the sensitivity of AE signals to process anomalies, they do not directly provide quantitative information about the resulting workpiece geometry.

Overall, the state of research indicates that acoustic emission signals possess substantial potential for quality prediction in machining processes. While early work emphasized indirect indicators and tool-related states, recent contributions increasingly demonstrate that AE-based models can be linked to explicit, functionally relevant geometric quality characteristics. However, existing approaches differ significantly in terms of target quantities, feature representations, and modeling strategies, and a systematic assessment of their applicability for quality-critical virtual metrology remains an open research challenge.

3.2 Uncertainty Quantification in Virtual Metrology

As motivated in Section 2.2, the use of machine learning in measurement-related applications requires predictive uncertainty to be addressed explicitly. This is particularly relevant for virtual metrology (VM), where quality characteristics are inferred from process data rather than obtained by direct physical measurement. Recent review papers show that VM and predictive quality methods are already widely applied across manufacturing domains, while uncertainty treatment is increasingly recognized as a key requirement for industrial applicability. (Tercan and Meisen 2022; Dreyfus et al. 2022)

From a methodological perspective, the literature on uncertainty quantification in VM and related predictive-quality settings can be structured into several streams. A first group comprises probabilistic and Bayesian regression approaches, which directly yield predictive distributions instead of point estimates. For example, Papananias et al. (2019) apply Bayesian linear regression in a multistage manufacturing setting and derive credible intervals for predicted part quality from in-process data. Likewise, Cramer, Huber, and R. H. Schmitt (2022) use Bayesian neural networks for predictive quality and show how aleatoric and epistemic contributions can be captured jointly within a probabilistic model. These approaches are particularly relevant for VM because they align well with the need to express prediction results together with an associated uncertainty statement.

A second stream focuses on approximate uncertainty estimation techniques for data-driven models. A prominent example is Monte Carlo dropout, in which stochastic forward passes are used at inference time to approximate predictive uncertainty. Such approaches have already been demonstrated in manufacturing-related regression tasks, for instance for tool flank wear prediction, where interval estimates are derived from repeated stochastic evaluations of the same network. (Dey 2023)

Although these methods are computationally attractive and easy to integrate into existing deep-learning pipelines, their uncertainty statements are only indirectly connected to metrological concepts of traceability and measurement uncertainty.

A third stream comprises interval-based methods that aim at statistically calibrated prediction intervals. In particular, conformal prediction and conformalized quantile regression have gained attention because they provide coverage guarantees under comparatively weak assumptions. Their relevance for manufacturing-related quality prediction has already been shown for semiconductor applications, where conformal methods were used to quantify uncertainty in line-edge-roughness estimation and to stabilize interval predictions under heteroscedastic conditions. (Akpabio and Savari 2022)

Overall, the state of research indicates that uncertainty quantification for VM is no longer limited to conceptual discussions, but is already addressed through Bayesian models, sampling-based approximations, and interval-calibration methods in adjacent manufacturing applications. At the same time, the literature also shows that many existing approaches primarily assess predictive reliability from a machine-learning perspective (addressing epistemic uncertainty), whereas a fully metrologically consistent treatment (aleatoric and epistemic uncertainty) still requires an explicit linkage between process data, learned prediction model, reference measurements, and coverage-based validation. Consequently, a systematic and metrologically traceable integration of epistemic and aleatoric uncertainty for virtual metrology remains largely an open research challenge. (Hüllermeier and Waegeman 2021; Hora 1996; Kiureghian and Ditlevsen 2009)

3.3 Intermediate Summary

In summary, existing research demonstrates that acoustic emission signals can encode relevant information about geometric quality characteristics, while parallel developments in uncertainty quantification provide methodological tools for assessing predictive reliability. However, these streams

remain largely disconnected. Current AE-based approaches rarely provide uncertainty statements that are traceable in a metrological sense, whereas uncertainty-focused methods are seldom applied to process-integrated sensing scenarios with direct relevance to geometric quality. As a result, there is a lack of approaches that enable quantitative, uncertainty-aware quality prediction suitable for conformity decisions.

4 Methodological Framework for Uncertainty-Aware Virtual Metrology

To provide a structured overview, the proposed methodology integrates two tightly coupled core components: (i) the systematic development of feature-based virtual metrology (VM) models from acoustic emission (AE) signals, and (ii) a metrologically consistent quantification of their predictive uncertainty. Both aspects are treated as integral parts of a unified processing pipeline rather than independent steps.

Previous research has demonstrated that AE signals contain sufficient information to predict process parameters as well as geometric and functional quality characteristics in micro gear hobbing (Schiller et al. 2023). This forms the basis for the model development in this work. Reliable reference data are enabled by the analytical description of crown gear flank geometry according to Gindin, Bilen, and Lanza (2025), while the extraction of quantitative quality characteristics from optical measurements follows the feature-based evaluation approach introduced by Bilen, Ernst, et al. (2025). Simulation-based investigations (Bilen, Braunschweiger, et al. 2025) further support the physical interpretation of the identified signal–quality relationships.

Building on these foundations, the overall methodology follows a two-stage approach with strong interdependencies between model development and uncertainty analysis.

(1) Feature-based model development In the first stage, virtual metrology models are developed using a structured data processing pipeline (cf. Section 4). AE signals are preprocessed, segmented, and transformed into a high-dimensional feature space capturing time-domain and time–frequency characteristics. Through feature selection and dimensionality reduction, a compact and informative representation is obtained and linked to optically derived geometric quality characteristics.

This step is not only aimed at maximizing predictive performance, but also at ensuring a physically meaningful and reproducible feature representation. In particular, the segmentation of signals and the use of interpretable feature descriptors enable a direct link between process phenomena and model inputs, which is essential for the subsequent uncertainty analysis.

(2) Uncertainty-aware prediction and quantification In the second stage, the predictive uncertainty of the developed VM models is systematically quantified (cf. Section 6). In contrast to purely performance-driven evaluations, the approach explicitly incorporates uncertainty contributions arising from three sources: (i) stochastic variability in AE-based feature extraction, (ii) uncertainty of the reference metrology used for label generation, and (iii) epistemic uncertainty of the data-driven model itself.

To achieve this, a GUM-S1-inspired Monte Carlo framework is extended by an empirical estimation of epistemic model uncertainty based on independent holdout data. This enables the propagation of both aleatoric and epistemic contributions through the learned regression model, resulting in predictive distributions instead of point estimates.

Integration of both stages A key contribution of this work lies in the explicit coupling of model development and uncertainty quantification. The structure of the feature space, the design of experiments, and the quality of the reference data directly influence the magnitude and composition of the resulting predictive uncertainty. Conversely, the uncertainty analysis provides a quantitative assessment of model reliability and highlights limitations in data coverage and model generalization.

The resulting workflow enables the transition from purely data-driven point predictions to metrologically interpretable quality estimates with associated uncertainty statements. This is a prerequisite for the integration of VM models into closed-loop quality control systems, where reliable confidence information is required for decision-making.

An overview of the complete processing and evaluation pipeline is illustrated in Fig. 2.

5 Data Processing and Model Development

This section describes the development of the data processing pipeline and the corresponding model training strategy for acoustic-emission-based virtual metrology. The approach follows a structured,

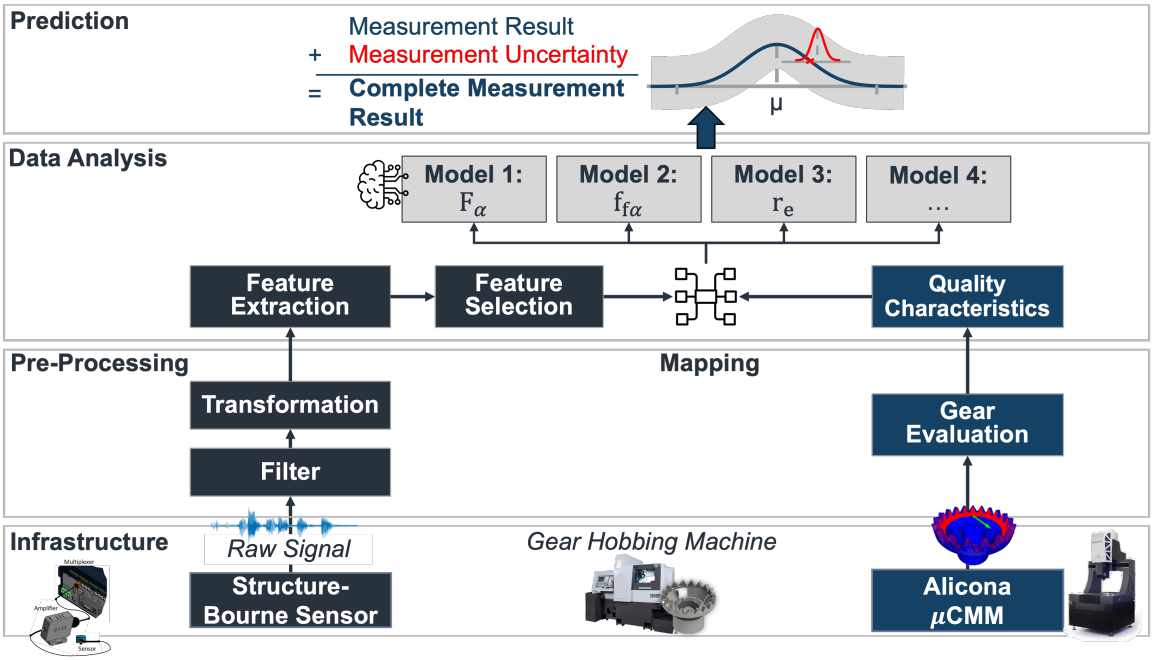


Figure 2: Conceptual overview of the coupled model development and uncertainty quantification pipeline for virtual metrology.

feature-based methodology in which process-related information is extracted from raw AE signals and linked to geometry-based quality characteristics obtained from optical measurements.

The pipeline comprises sensor integration and synchronized data acquisition, signal preprocessing, feature extraction and selection, as well as the generation of labeled datasets based on a controlled design of experiments. These data are subsequently used to train supervised learning models for both process parameter reconstruction and the prediction of geometric quality features. The modular structure of the pipeline ensures reproducibility and provides the basis for the subsequent uncertainty quantification.

5.1 Sensor Integration and Data Acquisition

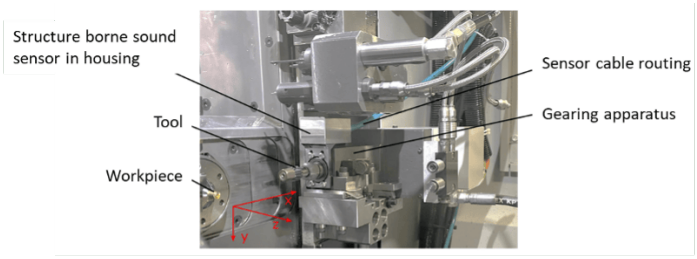
The acoustic emission (AE) sensor chain in Figure 3 was integrated into the gear hobbing machine based on prior work by Schiller et al. (2023), where the in-process monitoring of micro gear cutting was investigated in collaboration with *Dentsply Sirona*. A piezoelectric AE sensor is mounted close to the tool holder, ensuring a constant distance to the cutting zone while minimizing dynamic damping effects. Due to the small dimensions of the tool and workpiece, direct mounting at the cutting interface is not feasible.

The sensor is fixed to a dedicated mounting structure attached to the hobbing aggregate and tightened with a defined torque to ensure reproducible coupling conditions. Coupling quality is continuously monitored by injecting a reference signal into the sensor and evaluating the correlation of the returned signal. According to manufacturer specifications, correlation values above 80 % indicate stable and low-loss coupling conditions.

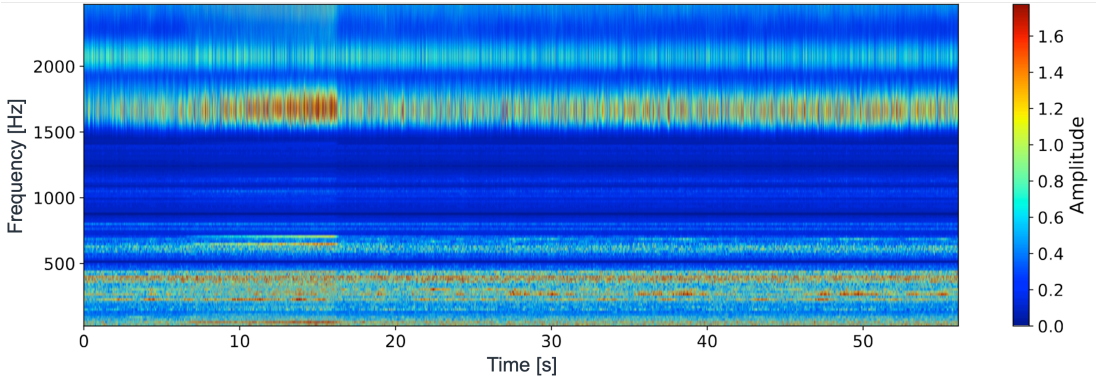
Signal acquisition and digitization are performed using a high-frequency measurement system with a maximum sampling rate of up to 100 MHz. To ensure comparability across measurements, data acquisition is synchronously triggered by the machine control system at the beginning of each machining cycle. This guarantees a consistent temporal reference for all recorded AE signals.

5.2 Signal Preprocessing

Prior to further processing, the recorded AE signals undergo basic preprocessing. High-frequency noise components and isolated outliers caused by measurement artifacts or transmission errors are reduced using Gaussian filtering and automated outlier detection. This step ensures robust feature extraction while preserving the physically relevant signal characteristics associated with the cutting process.



(a) Sensor setup integrated into the gear hobbing machine (Schiller et al. 2023)



(b) Resulting acoustic emission signal represented as CWT

Figure 3: Acoustic emission-based monitoring in gear hobbing: sensor integration (top) and resulting signal representation (bottom)

5.3 Feature Extraction

Virtual metrology model performance strongly depends on the quality of the extracted features. Therefore, a comprehensive feature set is derived from the AE signals to capture both global and localized process characteristics.

Features are extracted from the raw time-domain signal as well as from time–frequency representations obtained via short-time Fourier transform (STFT), continuous wavelet transform (CWT), and discrete wavelet transform (DWT). In addition to classical statistical descriptors such as mean, standard deviation, root mean square (RMS), signal energy, and extrema, frequency-resolved statistics are computed from the time–frequency representations. These features have been shown to be sensitive to tooth engagement events, process instabilities, and micro-scale disturbances in comparable machining applications.

To capture process-phase-specific effects, feature extraction is performed on multiple predefined time segments corresponding to different stages of tool engagement. This segmented analysis allows localized phenomena to be identified that would be obscured in purely global feature evaluations. Feature naming follows a structured convention combining signal source, transformation type, statistical descriptor, and time segment to ensure traceability and reproducibility.

5.4 Feature Reduction and Selection

Given the high dimensionality of the extracted feature space, both multivariate and univariate reduction strategies are employed. First, principal component analysis (PCA) (Runkler 2015) is used to obtain a compact, orthogonal representation of the feature space while preserving the majority of variance and removing redundant correlations. All features are standardized prior to transformation.

In parallel, explicit feature selection is performed using univariate F-regression (GuyonIsabelle and ElisseeffAndré 2003). Each feature is individually evaluated with respect to its linear relationship to the target variable, and statistically significant features are retained based on false-discovery-rate-controlled hypothesis testing. Feature selection is conducted separately for each target quantity, resulting in task-specific feature subsets.

5.5 Design of Experiments and Reference Measurements

Model training data are generated using a full-factorial design of experiments (DoE) covering relevant variations of key process parameters, namely tool runout, axial infeed (z-shift), and tangential

tool displacement (x-shift). The selected factor levels represent both stable and boundary process conditions encountered in industrial micro gear manufacturing.

Each parameter combination is manufactured repeatedly, yielding a balanced dataset for training and evaluation. During machining, AE signals are recorded inline and associated with the corresponding process parameters. Subsequently, all produced micro crown gears are measured using an optical micro-coordinate measuring system to obtain reference values for geometric quality metrics.

Table 1: Factors and levels of the design of experiments (DoE)

Factor	**Unit**	**Levels**
Tool runout	µm	0 — 5 — 9
Axial infeed (z-shift)	mm	−0.10 | − 0.05 | 0 | 0.05 | 0.10
Tangential displacement (x-shift)	mm	−0.10 | − 0.05 | 0 | 0.05 | 0.10

5.6 Model Training and Evaluation

Virtual metrology models were developed following a unified training and evaluation procedure to ensure comparability across model types and feature representations. Three regression approaches with increasing model complexity were considered. Linear regression serves as an interpretable baseline model (Hastie, Tibshirani, Friedman, et al. 2009). Random forest regression provides a robust ensemble-based approach (Breiman 2001), while support vector regression (SVR) enables the modeling of nonlinear relationships (Cortes and Vapnik 1995).

Model development was conducted in two stages. First, the reconstruction of process parameters from AE features was evaluated to verify process sensitivity and the physical consistency of the sensor signal (Schiller et al. 2023). This step establishes that the acoustic emission signal contains sufficient information to represent underlying process variations. Second, regression models were trained to predict geometric quality characteristics solely from AE data, forming the core of the virtual metrology approach.

Separate models were trained for each target variable. Model training followed a fixed split strategy. Of 225 manufactured and measured parts, 155 were used for training, 30 for model selection on a fixed test set, 20 as an independent holdout for epistemic uncertainty estimation and 20 as a fully disjoint validation holdout for coverage assessment . All subsets are strictly disjoint. For each target variable, linear regression, random forest, and SVR were evaluated across multiple feature representations. Model selection was based on performance on a fixed test set. Cross-validation on the training subset was used as an auxiliary indicator of model stability (Kohavi et al. 1995). Independent holdout sets were reserved for epistemic uncertainty estimation and final coverage validation.

Hyperparameters of nonlinear and ensemble models were optimized using randomized search. Long-term drift effects were not explicitly considered, as all experiments were conducted within a temporally stable process window. Consequently, the derived models reflect stationary relationships between process conditions, acoustic emission signals, and resulting quality characteristics.

6 Method for Uncertainty Quantification

This section introduces a GUM-inspired methodology to quantify the predictive uncertainty of virtual metrology (VM) models. The objective is to quantify uncertainty contributions and derive metrologically consistent, decision-relevant uncertainty statements.

In VM, quality characteristics are inferred from process data using data-driven models, which introduces additional uncertainty contributions beyond classical measurement uncertainty, particularly due to limited data and model-form deviations.

GUM Supplement 1 provides a framework for propagating probability distributions through deterministic models using Monte Carlo simulation (BIPM et al. 2008b). While this is, in principle, applicable to VM, a direct application reveals a key limitation: when the learned model is treated as deterministic, the propagation captures primarily *aleatoric* uncertainty, whereas *epistemic* uncertainty arising from model limitations is not represented.

To address this, the proposed methodology extends GUM-based uncertainty propagation by explicitly incorporating epistemic model uncertainty as an additional contribution and by separating *uncertainty estimation* from *uncertainty validation*.

Two independent holdout datasets are used for this purpose: (i) a first holdout dataset for estimating epistemic uncertainty based on prediction residuals, and (ii) a second, fully unseen validation holdout for assessing the empirical coverage of the predicted uncertainty intervals.

This separation avoids optimistic bias and enables an independent consistency check of the uncertainty model.

6.1 Overview and Identification of Uncertainty Sources

Following the principles of the *Guide to the Expression of Uncertainty in Measurement* (GUM) and GUM Supplement 1 (BIPM et al. 2008a; BIPM et al. 2008b), the first step is a structured identification of uncertainty sources along the complete VM pipeline. This includes (i) process-integrated AE sensing and feature extraction, (ii) reference metrology used for label generation, and (iii) the data-driven model itself.

To systematically capture these contributions, an Ishikawa diagram is used as an organizing instrument (Fig. 4). It combines classical metrological categories (instrument, workpiece, environment) with VM-specific sources such as data quality, learning bias, and limited coverage of the process space.

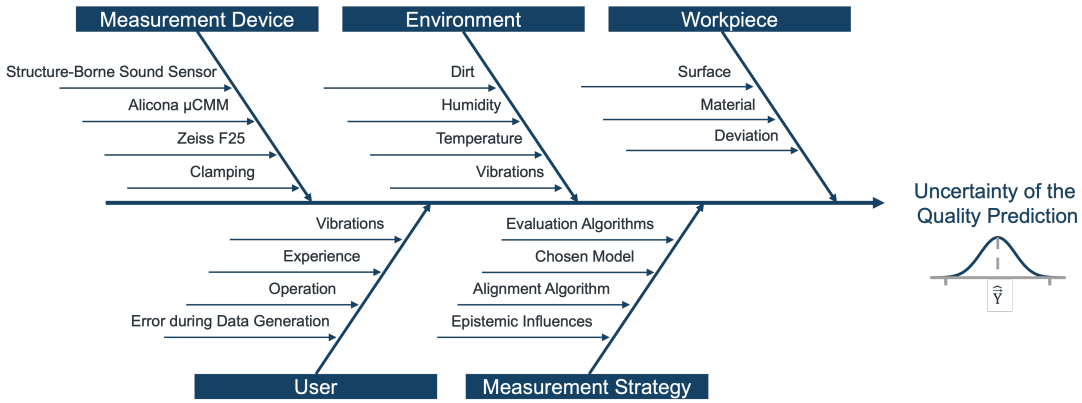


Figure 4: Ishikawa diagram for structured identification of uncertainty sources along the virtual metrology pipeline.

The identified sources are grouped into *aleatoric* and *epistemic* uncertainty. Aleatoric uncertainty describes irreducible stochastic variability, in this work mainly originating from AE sensing, signal processing, and the repeatability of reference measurements. Epistemic uncertainty reflects incomplete knowledge, for example due to limited or uncertain training data, restricted process coverage, model-form limitations, and systematic effects in the reference measurements. Figure 5 summarizes the considered uncertainty components and their combination within the proposed framework.

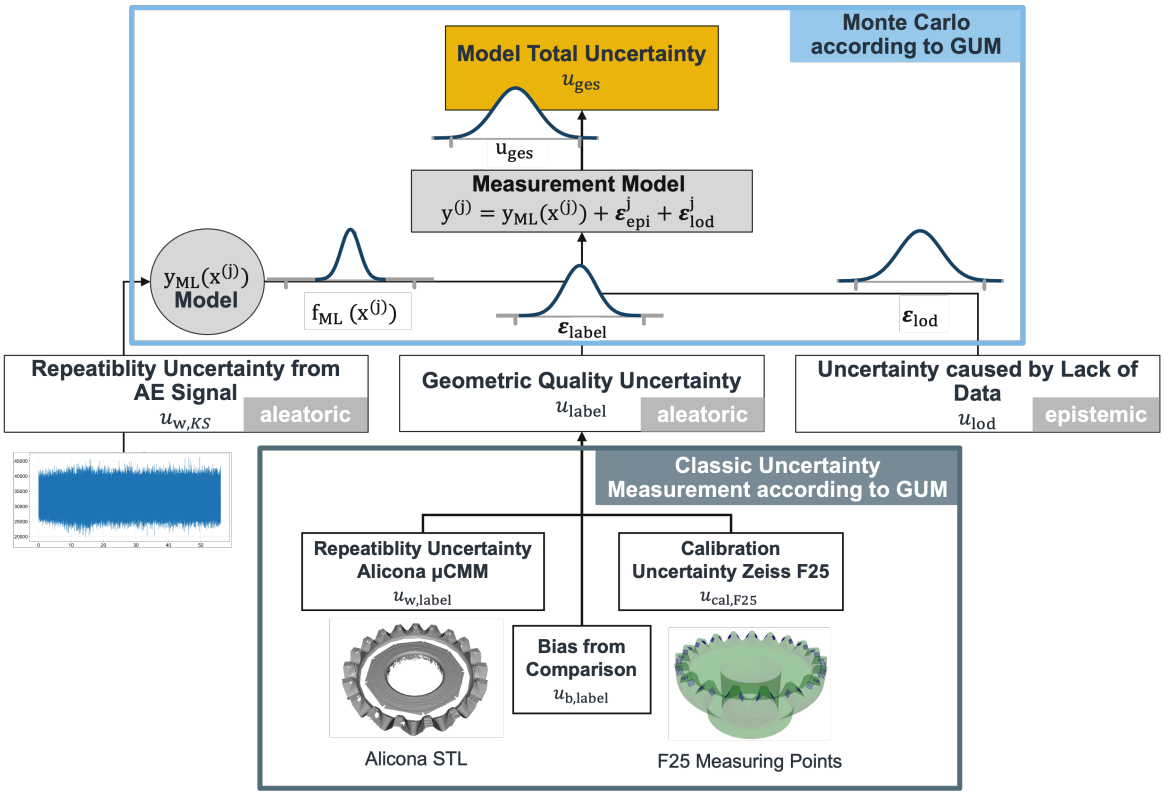


Figure 5: Conceptual overview of the total predictive uncertainty and its combination within the Monte-Carlo simulation.

6.2 Uncertainty Model Formulation

In the formulation phase, the output quantity Y is defined as a geometric quality characteristic derived from optical micro-metrology. The VM model maps an AE feature vector

$$\mathbf{X} = (X_1, \ldots, X_N)^{\mathrm{T}} \tag{2}$$

to a predicted output via a trained regression model

$$Y_{\mathrm{ML}} = f_{\mathrm{ML}}(\mathbf{X}). \tag{3}$$

To represent all relevant uncertainty contributions within a single probabilistic framework, the VM prediction is extended by additive stochastic terms:

$$Y = f_{\mathrm{ML}}(\mathbf{X}) + \varepsilon_{\mathrm{label}} + \varepsilon_{\mathrm{epi}}. \tag{4}$$

Here, $\varepsilon_{\mathrm{label}}$ accounts for uncertainty in the reference measurements, while $\varepsilon_{\mathrm{epi}}$ represents epistemic model deviation.

Epistemic model uncertainty from holdout residuals Epistemic uncertainty is estimated using an independent holdout dataset $\mathcal{D}_{\mathrm{HO}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_{\mathrm{HO}}}$, which is not used for training or model selection. Residuals are computed as

$$r_i = y_i - f_{\mathrm{ML}}(\mathbf{x}_i). \tag{5}$$

Assuming that residuals consist of epistemic and label contributions,

$$r_i = \varepsilon_{\mathrm{epi},i} + \varepsilon_{\mathrm{label},i}, \tag{6}$$

and that both are independent, the variance decomposes as

$$\mathbb{V}(r) = \mathbb{V}(\varepsilon_{\mathrm{epi}}) + u^2_{\mathrm{label}}. \tag{7}$$

The epistemic variance is then estimated by subtracting the known label contribution:

$$\hat{\sigma}^2_{\mathrm{epi}} = \max\left(0,\, \hat{\sigma}^2_r - u^2_{\mathrm{label}}\right). \tag{8}$$

The residual mean $\bar{r}$ is interpreted as a potential systematic offset μ_{epi}. For propagation, epistemic uncertainty is modeled as

$$\varepsilon_{\text{epi}} \sim \mathcal{N}(\mu_{\text{epi}}, \sigma^2_{\text{epi}}), \tag{9}$$

with $\mu_{\text{epi}} = \bar{r}$ and $\sigma^2_{\text{epi}} = \hat{\sigma}^2_{\text{epi}}$.

6.3 Estimating Aleatoric Input and Label Uncertainty

Label uncertainty The combined standard uncertainty of the labels is obtained by aggregating calibration, repeatability, and systematic deviation:

$$u_{\text{label}} = \sqrt{u^2_{\text{cal,F25}} + u^2_{\text{p,label}} + u^2_{\text{bias,label}}}. \tag{10}$$

AE feature uncertainty The uncertainty of the AE feature vector is estimated as a Type A repeatability contribution. A global relative uncertainty measure $u_{p,KS}$ is derived and used to parameterize the input distribution $p(\mathbf{X})$.

6.4 Monte-Carlo Propagation and Uncertainty Summaries

Monte Carlo simulation is used to propagate aleatoric uncertainties (input and label uncertainty). Epistemic uncertainty, estimated from holdout residuals, is incorporated as an additional stochastic model discrepancy term:

$$\mathbf{x}^{(j)} \sim p(\mathbf{X}), \quad \varepsilon^{(j)}_{\text{label}} \sim p(\varepsilon_{\text{label}}), \quad \varepsilon^{(j)}_{\text{epi}} \sim p(\varepsilon_{\text{epi}}), \tag{11}$$

$$y^{(j)} = f_{\text{ML}}\left(\mathbf{x}^{(j)}\right) + \varepsilon^{(j)}_{\text{label}} + \varepsilon^{(j)}_{\text{epi}}. \tag{12}$$

The resulting sample approximates the predictive distribution $p(Y)$. The estimator and standard uncertainty are

$$\hat{y} = \frac{1}{M}\sum_{j=1}^{M} y^{(j)}, \qquad u(\hat{y}) = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}\left(y^{(j)} - \hat{y}\right)^2}. \tag{13}$$

Coverage intervals are obtained from empirical quantiles.

Global uncertainty indicator For each part i, a combined uncertainty measure is defined as

$$\sigma_{\text{tot},i} = \sqrt{\sigma_i^2 + e_i^2}. \tag{14}$$

A global uncertainty metric is obtained as

$$U_m^{95} = \mathrm{P}_{95}(\sigma_{\text{tot},i})\,. \tag{15}$$

6.5 Validation of Predictive Uncertainty

An independent validation dataset $\mathcal{D}_{\text{VAL}} = \{(\mathbf{x}_k, y_k)\}_{k=1}^{n_{\text{VAL}}}$ is used to assess the consistency of the predicted uncertainty intervals.

Prediction intervals are derived from empirical quantiles:

$$[Y_{\text{L},k},\, Y_{\text{U},k}] = \left[\mathrm{Q}_{(1-p)/2},\ \mathrm{Q}_{(1+p)/2}\right]. \tag{16}$$

A prediction is considered consistent if

$$Y_{\text{L},k} \le y_k \le Y_{\text{U},k}. \tag{17}$$

The empirical coverage is computed as the fraction of consistent predictions and compared to the nominal coverage probability. Agreement between both indicates a statistically consistent and calibrated uncertainty model.

7 Results

This section presents the results of the proposed virtual metrology framework in a structured manner, following the methodological pipeline introduced in Section 4. First, the performance of the feature-based models is evaluated, including the reconstruction of process parameters and the prediction of geometric quality characteristics. These results establish the empirical basis and reveal target-dependent model behavior.

Subsequently, the different contributions to predictive uncertainty are analyzed individually. This includes the uncertainty of the optical reference labels, the variability of the acoustic emission features, and the epistemic uncertainty of the trained models. Finally, these contributions are combined to obtain a global uncertainty estimate for the virtual metrology predictions.

7.1 Results of Model Development

This section presents the key results of the acoustic-emission-based virtual metrology models. Both the reconstruction of process parameters and the prediction of geometric quality characteristics are evaluated to establish the empirical basis for the subsequent uncertainty analysis.

7.1.1 Reconstruction of Process Parameters The reconstruction of the manufacturing parameters reveals a pronounced difference in how strongly individual parameters are encoded in the acoustic emission (AE) signal. Table 2 summarizes the achieved coefficients of determination for the three investigated parameters.

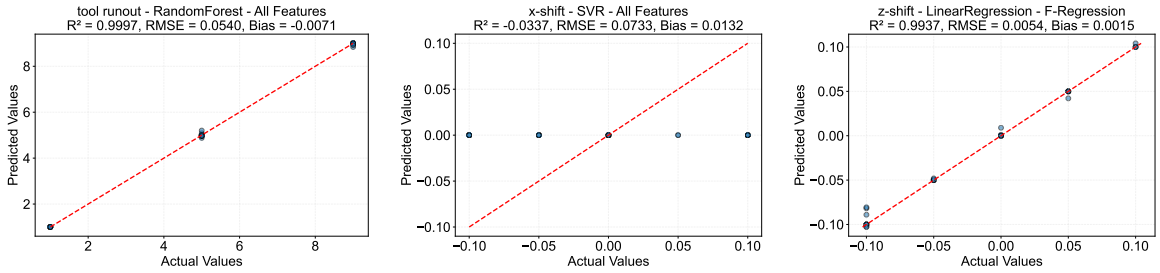
Table 2: Reconstruction performance of process parameters based on AE features.

Target	**Model**	R^2_{Test}
Tool runout	Linear regression	0.991
Tool runout	Random forest	0.996
Tool runout	SVR	0.939
x-shift	Linear regression	-0.045
x-shift	Random forest	-0.030
x-shift	SVR	-0.001
z-shift	Linear regression	0.843
z-shift	Random forest	0.829
z-shift	SVR	-0.005

Tool runout is reconstructed with very high accuracy across all model classes. Even simple linear regression achieves coefficients of determination close to unity in both cross-validation and test datasets, indicating a strong and predominantly linear relationship between AE features and runout-induced process effects. Random forest models achieve comparable accuracy with slightly increased variance, while SVR shows reduced but still substantial performance.

The axial infeed (*z-shift*) is also encoded in the AE signal, albeit with lower robustness. Linear regression and random forest models yield high but less stable R^2 values, whereas SVR fails to generalize. This suggests increased sensitivity to nonlinearities and stochastic process variations.

In contrast, the tangential displacement (*x-shift*) cannot be reconstructed. All models yield negative or near-zero R^2 values, indicating the absence of an exploitable relationship between AE features and this parameter. This behavior is further illustrated by the predicted-versus-actual plots in Fig. 6.



(a) Tool runout (b) x-shift (c) z-shift

Figure 6: Predicted versus actual values for the reconstructed process parameters.

7.1.2 Results of Feature Selection The feature selection reveals a clear mapping of relevant features to specific signal intervals, as well as characteristic differences in signal representation across the investigated process parameters.

For *tilt*, significant features are distributed across multiple intervals (mainly 2, 3, and 4), indicating a global influence over the entire engagement. RMS-based features and STFT metrics dominate, suggesting consistent changes in signal energy and spectral amplitudes. Thus, tilt appears as a global process deviation with a stable signature across intervals.

In contrast, *x-shift* shows a strong localization of relevant features, mainly concentrated in interval 3. These are predominantly STFT-based mean and median values, indicating that x-shift primarily

Table 3: Top 10 most significant features for tilt, x-shift and z-shift, based on F-regression ranking.

Rank	**Feature**
1	STFT_interval2_median_freq_27
2	CWT_interval2_median_scale_107
3	CWT_interval2_median_scale_108
4	CWT_interval2_median_scale_106
5	CWT_interval2_median_scale_109
6	CWT_interval2_median_scale_105
7	STFT_interval2_median_freq_35
8	STFT_interval2_median_freq_28
9	STFT_interval2_median_freq_44
10	CWT_interval2_mean_scale_107

affects spectral positioning rather than overall signal energy. Despite this localized sensitivity, robust reconstruction remains limited, suggesting weak or inconsistent signal signatures.

For *z-shift*, relevant features are clearly concentrated in interval 2, directly after initial engagement. Both STFT- and CWT-based features are dominant, indicating that z-shift is encoded in both frequency components and localized, scale-dependent structures. This points to a localized effect in the early engagement phase.

Overall, the results highlight distinct spatio-temporal signatures in the acoustic emission signal: tilt acts globally across intervals, while x- and z-shift are localized, with x-shift mainly affecting spectral position and z-shift showing additional multiscale characteristics.

7.1.3 Prediction of Geometric Quality Characteristics Building on the demonstrated sensitivity of acoustic emission signals to process parameters and the identified feature–process relationships, the following analysis focuses on the core objective of virtual metrology, namely the prediction of geometric quality characteristics from process-integrated data. The selected geometric quality characteristics cover both locally defined and globally aggregated deviations that are relevant for functional performance and quality assurance of micro gears. Local characteristics, such as angular and form deviations, are directly influenced by instantaneous cutting interactions and are therefore expected to be reflected in the acoustic emission signal. In contrast, global characteristics represent cumulative effects along the tool path and pose a more challenging prediction task. This selection allows assessing the limits of AE-based virtual metrology with respect to physical observability. The prediction performance for geometric quality characteristics is summarized in Table 4. Compared to process parameters, overall performance is reduced but strongly target-dependent.

Table 4: Prediction performance for geometric quality characteristics (best model per target).

Target	R^2_{Test}
Tooth thickness deviation Δt	0.936
Flank angle deviation $f_{H\beta}$	0.924
Profile angle deviation $f_{H\alpha}$	0.922
Profile total deviation F_α	0.671
Profile form deviation $f_{f\alpha}$	0.777
Flank total deviation F_β	0.641
Flank form deviation $f_{f\beta}$	0.704
Pitch deviation F_p	0.125
Eccentricity radius r_e	-0.006

Angle-related deviations and tooth thickness deviation achieve the highest predictive accuracy, reflecting their strong physical coupling to the cutting process. In contrast, global form and positional deviations show limited or no predictive capability, indicating that these effects are only weakly encoded in the AE signal or dominated by measurement and process variability.

7.1.4 Relevance for Uncertainty Analysis The presented results demonstrate that predictive performance varies substantially across target variables and model types. Even for targets with high R^2 values, residual scatter and systematic deviations remain, as evident from the predicted-versus-actual distributions. These effects arise from stochastic process variability, uncertainty in the reference measurements, and limited generalization of the data-driven models.

Consequently, performance metrics alone are insufficient for assessing the suitability of AE-based virtual metrology in quality assurance. A rigorous uncertainty quantification is required to determine the reliability of individual predictions and to enable metrologically meaningful coverage intervals. The following section therefore introduces a comprehensive uncertainty modeling and propagation methodology building directly on these results.

7.2 Results of Uncertainty Analysis

This section presents the key quantitative results that directly underpin the subsequent uncertainty quantification of the virtual metrology models. In contrast to a purely performance-oriented evaluation, the focus is placed on identifying and quantifying the dominant uncertainty contributions arising from reference measurements, acoustic emission (AE) input data, and model-related epistemic effects.

7.2.1 Uncertainty of Optical Reference Labels The uncertainty of the optically acquired reference data represents a fundamental lower bound for the achievable accuracy of any data-driven prediction model. Table 5 summarizes the repeatability uncertainty u_p and the systematic bias u_b of the optical measurements relative to the tactile reference.

Table 5: Repeatability uncertainty u_p, systematic bias u_b, and resulting combined standard uncertainty u_{label} of optical reference measurements.

Geometric characteristic	u_p	u_b	u_{label}
Tooth thickness deviation Δt [μm]	0.15	0.81	0.87
Flank angle deviation $f_{H\beta}$ [°]	0.00009	0.00005	0.00017
Profile angle deviation $f_{H\alpha}$ [°]	0.00014	0.0011	0.00015
Profile total deviation F_α [μm]	0.014	-0.14	0.33
Profile form deviation $f_{f\alpha}$ [μm]	0.018	-1.79	1.81
Flank total deviation F_β [μm]	0.023	1.34	1.37
Flank form deviation $f_{f\beta}$ [μm]	0.013	0.37	0.47
Pitch deviation F_p [°]	0.005	0.034	0.0015
Eccentricity radius r_e [μm]	6.93	5.34	8.76

For most profile- and flank-related characteristics, the repeatability uncertainty is well below 0.1 µm, indicating a highly stable optical measurement process. However, systematic bias dominates the uncertainty budget for several features and therefore must be explicitly accounted for in the uncertainty model.

The eccentricity radius r_e constitutes a clear outlier. Both repeatability and bias uncertainties are more than one order of magnitude larger than for all other characteristics, indicating numerical instabilities in the evaluation pipeline rather than physical measurement limitations.

The resulting combined standard uncertainties of the label data, summarized in Table 5, define a hard baseline that cannot be undercut by any prediction model.

7.2.2 Uncertainty of the Acoustic Emission Measurement Chain The aleatoric uncertainty of the AE measurement chain is quantified via the repeatability scatter of extracted spectral features under constant process conditions. An excerpt of the resulting statistics, including Shapiro–Wilk normality tests, is shown in Table 6.

The distribution of relative standard deviations across all features and experimental series is depicted in Fig. 7. The 95th percentile of this distribution is selected as a conservative perturbation level for Monte-Carlo propagation.

7.2.3 Epistemic Uncertainty of the Prediction Models Epistemic uncertainty is estimated using an independent holdout dataset and reflects model limitations due to finite training data and restricted

Table 6: Excerpt of repeatability statistics of selected AE spectral features with Shapiro–Wilk test results.

Feature	**Series**	**Std. [%]**	***p*-value**
STFT RMS @ 1334 Hz	1	4.09	0.006
STFT Std. @ 1286 Hz	1	4.24	0.302
STFT Std. @ 5335 Hz	16	3.03	0.466
STFT RMS @ 6050 Hz	16	4.43	0.016
STFT Median @ 762 Hz	16	2.57	0.159

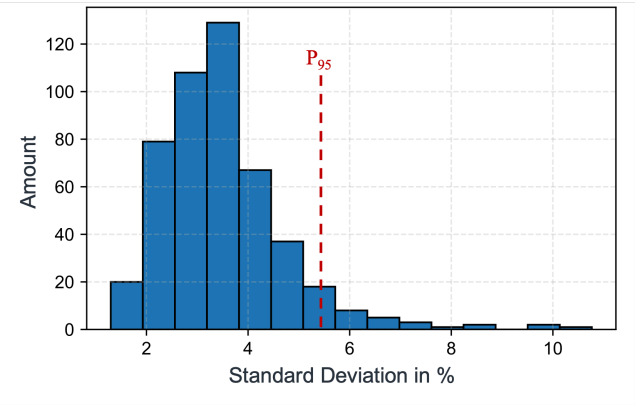


Figure 7: Histogram of relative standard deviations of AE spectral features under constant process conditions, with the 95th percentile P_{95} highlighted.

generalization capability. Table 7 summarizes the resulting epistemic uncertainty estimates for the best-performing model of each target quantity.

Table 7: Estimated epistemic uncertainty ε_{epi} for the investigated target quantities.

Target quantity	ε_{epi}
Tooth thickness deviation Δt [μm]	14.69
Flank angle deviation $f_{H\beta}$ [°]	0.00088
Profile angle deviation $f_{H\alpha}$ [°]	0.00180
Profile total deviation F_α [μm]	0.63
Profile form deviation $f_{f\alpha}$ [μm]	0
Flank total deviation F_β [μm]	0
Flank form deviation $f_{f\beta}$ [μm]	0
Tool runout [μm]	229.29
z-shift [μm]	24.90

A strong dependence on the target quantity is observed. Aggregated quantities such as tool runout, axial infeed, and tooth thickness deviation exhibit substantially higher epistemic uncertainty than local geometric form or angle deviations.

7.2.4 Global Uncertainty of Virtual Metrology Predictions The combined effect of label uncertainty, AE input variability, and epistemic model uncertainty is evaluated via Monte-Carlo simulation. The resulting global uncertainty metric U_m^{95} for each geometric characteristic is summarized in Table 8.

Local form and angle deviations exhibit low global uncertainty levels, indicating that these quantities can be predicted reliably using acoustic-emission-based virtual metrology within the investigated process window. In contrast, aggregated quantities such as tooth thickness, tool runout, and axial infeed show substantially higher uncertainty levels, limiting their applicability for precise quality control.

The empirical coverage obtained on the independent validation holdout is reported in Table 9.

Table 8: Global uncertainty U_m^{95} of the virtual metrology predictions.

Geometric characteristic	U_m^{95}
Tooth thickness deviation Δt [μm]	23.02
Flank angle deviation $f_{H\beta}$ [°]	0.00147
Profile angle deviation $f_{H\alpha}$ [°]	0.00189
Profile total deviation F_α [μm]	0.87
Profile form deviation $f_{f\alpha}$ [μm]	1.89
Flank total deviation F_β [μm]	0.87
Flank form deviation $f_{f\beta}$ [μm]	0.50
Tool runout [μm]	305.52
z-shift [μm]	25.92

For a nominal coverage probability of 95%, the prediction intervals achieve adequate coverage for most geometric characteristics. Profile and flank form deviations as well as pitch deviation and flank total deviation reach full empirical coverage, indicating that the combined uncertainty model tends toward overestimation for these targets. Profile angle deviation and profile total deviation yield 90% coverage, which is close to the nominal level given the limited sample size of $n = 20$.

Table 9: Empirical coverage of the investigated target quantities on 20 entries.

Target quantity	Empirical coverage
Tooth thickness deviation Δt [μm]	0.65
Flank angle deviation $f_{H\beta}$ [°]	0.85
Profile angle deviation $f_{H\alpha}$ [°]	0.90
Profile total deviation F_α [μm]	0.90
Profile form deviation $f_{f\alpha}$ [μm]	1.00
Flank total deviation F_β [μm]	1.00
Flank form deviation $f_{f\beta}$ [μm]	1.00
Pitch deviation F_p [°]	1.00
Eccentricity radius r_e [μm]	0.95

However, tooth thickness deviation Δt exhibits a substantial coverage deficit at 0.65, suggesting that the modeled uncertainty contributions do not fully capture the prediction variability for this target. Despite the comparatively large epistemic uncertainty already accounted for, the resulting prediction intervals remain insufficient, which may indicate additional unmodeled systematic effects or distributional deviations not captured. Flank angle deviation $f_{H\beta}$ also falls below the nominal level at 0.85.

It should be noted that with only 20 validation samples, the empirical coverage estimates carry considerable statistical uncertainty themselves. A single misclassified sample shifts the coverage by 5 percentage points, limiting the granularity of any calibration assessment. Nevertheless, the results provide a useful plausibility check.

These results imply that the suitability of virtual metrology strongly depends on the target quantity. While high coefficients of determination may indicate strong predictive relationships, they do not necessarily guarantee reliable predictions in a metrological sense. Only the combined consideration of prediction performance and associated uncertainty enables a meaningful assessment of model applicability in production environments.

8 Discussion of Results

The results of this study demonstrate that acoustic-emission-based virtual metrology (VM) can provide meaningful in-process information about part quality, while also highlighting current limitations that are primarily of an engineering rather than a fundamental nature. The discussion therefore focuses on (i) the conditions under which AE-based VM yields reliable results, (ii) the implications of the quantified predictive uncertainty, and (iii) opportunities for systematic improvement.

8.1 Predictive Capability of Acoustic Emission Signals

The model-building results show that process-integrated acoustic emission (AE) signals encode substantial information about both manufacturing parameters and resulting geometric quality characteristics. High predictive performance was achieved for several target quantities, most notably for tooth thickness deviation Δt ($R^2_{\text{test}} = 0.887$) and angular deviations such as $fH\alpha$ and $fH\beta$ ($R^2_{\text{test}} \approx 0.86$–$0.88$). These values indicate a strong and robust relationship between AE features and local geometric characteristics.

In contrast, other target quantities such as pitch deviation F_p and eccentricity radius r_e show little to no predictive capability ($R^2_{\text{test}} \approx 0.1$ or lower), indicating that these characteristics are not sufficiently encoded in the AE signal under the given conditions.

A more differentiated interpretation reveals that predictive capability is strongly dependent on the physical nature of the target variable. Local geometric characteristics, such as angular deviations and tooth thickness, exhibit consistently high prediction accuracy and robustness. These quantities are directly linked to instantaneous tool–workpiece interactions and are therefore well represented in the AE signal.

In contrast, more global or aggregated quantities are influenced by cumulative effects along the tool path, machine kinematics, and additional sources of variability that are not fully observable in the local AE signal. This indicates that the predictive capability of AE-based VM is inherently bounded by the observability of the underlying physical phenomena.

At the same time, these limitations are not purely intrinsic to the sensing principle. The present study is based on a finite design of experiments with controlled parameter variations. Expanding the dataset in terms of process variability, long-term effects such as tool wear, and broader operating conditions is expected to improve model generalization, particularly for global quality characteristics.

8.2 Implications of Prediction Uncertainty

The uncertainty analysis provides a more nuanced assessment of model applicability than predictive performance alone. While several target quantities exhibit high predictive accuracy, the associated uncertainties are non-negligible and strongly target-dependent.

For locally defined characteristics, uncertainty levels remain comparatively low. For example, angular deviations such as $fH\alpha$ and $fH\beta$ exhibit global uncertainty levels on the order of $U_{95} \approx 0.002°$, while local form deviations such as $ff\beta$ remain below $1\,\mu m$. In these cases, the uncertainty is small relative to typical tolerance ranges, indicating that VM predictions can be considered reliable for process monitoring and potentially for certain quality control tasks.

In contrast, aggregated quantities show significantly higher uncertainty levels. Tooth thickness deviation Δt, despite its high predictive performance ($R^2_{\text{test}} = 0.887$), exhibits a global uncertainty of approximately $U_{95} \approx 28\,\mu m$. Even more pronounced, tool runout shows uncertainty levels exceeding $300\,\mu m$. These values indicate that high correlation alone does not imply decision-relevant accuracy, as uncertainty may exceed acceptable tolerance limits.

Importantly, the current uncertainty estimates represent a deliberately conservative scenario. The variability of AE features is incorporated using a worst-case-oriented aggregation (e.g., high-percentile statistics) and applied uniformly across all prediction instances. While this ensures robustness and prevents underestimation, it likely leads to systematic overestimation of uncertainty in many regions of the process space.

This observation highlights a key opportunity for methodological refinement. More differentiated uncertainty modeling—such as feature-specific variability, operating-point-dependent uncertainty, or heteroscedastic prediction models—could significantly reduce conservatism while maintaining metrological consistency.

From an application perspective, the results indicate that VM is already well suited for process monitoring, trend detection, and early-stage quality assessment. For conformity decisions, applicability depends on the ratio between uncertainty and tolerance limits. Reducing conservative overestimation of uncertainty therefore represents a critical step toward broader industrial use.

8.3 Limitations and Opportunities for Improvement

The proposed uncertainty quantification approach provides a structured and metrologically interpretable framework for assessing predictive uncertainty. At the same time, several limitations indicate clear opportunities for further improvement.

First, epistemic uncertainty is estimated from a finite holdout dataset and is therefore limited to the explored region of the process space. Increasing the amount and diversity of training data is expected to reduce epistemic uncertainty and improve model robustness.

Second, the current uncertainty propagation follows a conservative design principle by combining uncertainty contributions in a global and worst-case-oriented manner. While appropriate for an initial implementation, this approach does not exploit the full potential of data-driven modeling. More granular representations—such as feature-wise uncertainty, local residual modeling, or probabilistic learning approaches—offer significant potential for reducing uncertainty bounds.

Third, the experimental setup assumes a temporally stable process. In industrial environments, additional variability due to tool wear, thermal drift, or environmental effects will introduce further uncertainty contributions that are not yet explicitly modeled.

Finally, the achievable performance and uncertainty are directly linked to the quality of the reference metrology. Systematic biases and measurement uncertainty in the reference data define a lower bound for achievable prediction accuracy and must be carefully controlled.

Overall, the results indicate that the presented methodology constitutes a reliable and conservative baseline for uncertainty-aware virtual metrology. The identified limitations do not represent fundamental barriers, but rather define a clear and technically feasible pathway toward improved accuracy, reduced uncertainty, and broader industrial applicability.

9 Summary and Outlook

High-precision micro-manufacturing, such as the production of micro-scale crown gears for dental applications, places stringent demands on quality assurance. Conventional measurement chains are often time-consuming and technologically limited, leading to long feedback loops and restricted process control capabilities. Virtual metrology approaches promise to overcome these limitations by predicting quality-relevant characteristics directly from process-integrated sensor data.

In this work, acoustic emission signals acquired during the machining process were investigated as an information source for virtual metrology. The results demonstrate pronounced correlations between acoustic emission features and both manufacturing parameters and resulting geometric quality characteristics. In particular, high predictive machine learning model performance was achieved for locally defined characteristics such as angular deviations and tooth thickness, confirming the fundamental potential of acoustic-emission-based models for predictive quality in micro-gear manufacturing. At the same time, this paper demonstrated that predictive machine learning model performance alone is not sufficient to assess model applicability.

For a rigorous uncertainty quantification, a GUM-oriented uncertainty quantification framework was proposed. It covers and combines aleatoric and epistemic uncertainty. The Monte-Carlo-based propagation of all identified uncertainty contributions allows consistent aleatoric uncertainty statements for nonlinear, data-driven models. By explicitly accounting for label uncertainty, stochastic variability of the acoustic emission feature space, and epistemic model uncertainty, the method enables a conservative and metrologically interpretable estimation of uncertainty.

The quantified uncertainties are substantial and strongly target-dependent. While locally defined characteristics exhibit comparatively low uncertainty levels, aggregated quantities show significantly higher uncertainty, in some cases exceeding practically relevant tolerance ranges. These findings underline that uncertainty quantification is a prerequisite for the meaningful use of data-driven virtual metrology in quality-critical contexts.

The presented results indicate that acoustic-emission-based virtual metrology is already well suited for applications such as process monitoring, trend analysis, and early-stage quality assessment. At the same time, the current uncertainty levels limit its direct use for conformity assessment across all target quantities.

Future work should therefore focus on three key directions. First, increasing the amount and diversity of training data is expected to reduce epistemic uncertainty and improve model generalization, particularly for globally aggregated quality characteristics. Second, the current uncertainty modeling approach can be refined by moving from a conservative, worst-case-oriented aggregation toward more differentiated representations, such as feature-specific or operating-point-dependent uncertainty models, as well as probabilistic learning approaches. Third, the extension of the methodology to more realistic industrial conditions, including tool wear, process drift, and varying environmental influences, is essential to ensure robustness and transferability.

Overall, the presented methodology provides a reliable and transparent baseline for uncertainty-aware virtual metrology. The identified limitations define a clear and technically feasible pathway toward more accurate, less conservative, and industrially applicable VM systems, thereby supporting the transition from predictive quality estimation toward decision-relevant, uncertainty-aware process control.

Data availability

The underlying component data supporting the findings of this study are confidential due to industrial cooperation agreements and therefore cannot be made publicly available.

References

Akpabio, Inimfon I. and Serap A. Savari (2022). "On an application of denoising to the uncertainty quantification of line edge roughness estimation". In: *2022 33rd annual SEMI advanced semiconductor manufacturing conference (ASMC)*, pp. 1–6. DOI: 10.1109/ASMC54647.2022.9792521.

Albers, Albert et al. (Jan. 2017). "Prediction of the Product Quality of Turned Parts by Real-time Acoustic Emission Indicators". In: *Procedia CIRP* 63. Num Pages: 6, pp. 348–353. ISSN: 22128271. DOI: 10.1016/j.procir.2017.03.173.

Beggan, C. et al. (Sept. 1999). "Using Acoustic Emission to Predict Surface Quality". en. In: *The International Journal of Advanced Manufacturing Technology* 15.10, pp. 737–742. ISSN: 1433-3015. DOI: 10.1007/s001700050126. URL: https://doi.org/10.1007/s001700050126 (visited on 11/30/2025).

Bilen, Ali, Nikolas Paul Braunschweiger, et al. (2025). "A Simulation-Based Error Analysis Approach for the Crown Gear Hobbing Process". In: Manuscript under review.

Bilen, Ali, Lennart Ernst, et al. (2025). "Quality Features for Holistic Evaluation and Quality Control of Micro Face Gears". In: *Measurement: Sensors*. Manuscript under review.

Bilen, Ali, Kim Laura Skade, et al. (2025). "Towards a Comprehensive Virtual Metrology Framework: Integrating AutoML, Data Integration, Uncertainty Quantification & Model Maintenance". In: *IFAC-PapersOnLine* 59.30, pp. 395–400. ISSN: 2405-8963. DOI: https://doi.org/10.1016/j.ifacol.2025.12.269. URL: https://www.sciencedirect.com/science/article/pii/S2405896325029799.

BIPM et al. (2008a). *Evaluation of measurement data — Guide to the expression of uncertainty in measurement*. tex.howpublished: Joint Committee for Guides in Metrology, JCGM 100:2008. DOI: https://doi.org/10.59161/JCGM100-2008E.

— (2008b). *Evaluation of measurement data — Supplement 1 to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions using a Monte Carlo method*. tex.howpublished: Joint Committee for Guides in Metrology, JCGM 101:2008. DOI: https://doi.org/10.59161/JCGM101-2008.

Breiman, Leo (2001). "Random forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324.

Chang, Yaw-Jen et al. (Jan. 2006). "Virtual metrology technique for semiconductor manufacturing". In: *The 2006 IEEE International joint conference on neural network proceedings*, p. 52895293. DOI: 10.1109/IJCNN.2006.247284.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine learning* 20.3, pp. 273–297.

Cramer, Simon, Maximilian Huber, and Robert H. Schmitt (2022). "Uncertainty quantification based on bayesian neural networks for predictive quality". In: *Artificial intelligence, big data and data science in statistics*. Ed. by Ansgar Steland and Kwok-Leung Tsui. Cham: Springer International Publishing, pp. 253–268. DOI: 10.1007/978-3-031-07155-3_10.

Dey, A. (2023). "Addressing uncertainty in tool wear prediction with dropout-based neural network". In: *Computers* 12.9, p. 187. DOI: 10.3390/computers12090187.

Dreyfus, Pierre-Alexandre et al. (2022). "Virtual metrology as an approach for product quality estimation in Industry 4.0: a systematic review and integrative conceptual framework". In: *International Journal of Production Research* 60.3, pp. 742–765. DOI: 10.1080/00207543.2021.1976433.

Erkaya, Selçuk and Şaban Ulus (Jan. 2016). "An Experimental Study on Gear Diagnosis by Using Acoustic Emission Technique". In: *The International Journal of Acoustics and Vibration* 21.1. Num Pages: 9. DOI: 10.20855/ijav.2016.21.1400.

Gauder, Daniel et al. (Jan. 2022). "In-process acoustic pore detection in milling using deep learning". In: *CIRP Journal of Manufacturing Science and Technology* 37, pp. 125–133. ISSN: 1755-5817. DOI: 10.1016/j.cirpj.2022.01.008. URL: https://www.sciencedirect.com/science/article/pii/S1755581722000141.

Geman, Stuart, Elie Bienenstock, and René Doursat (1992). "Neural networks and the bias/variance dilemma". In: *Neural Computation* 4.1, pp. 1–58. DOI: 10.1162/neco.1992.4.1.1.

Gindin, Edouard, Ali Bilen, and Gisela Lanza (2025). "Eine mathematische Beschreibung der Flankengeometrie geradverzahnter Kronenräder". In: *tm - Technisches Messen*. ISSN: 0171-8096, 2196-7113. DOI: 10.1515/teme-2025-0033.

Goch, Gert et al. (Jan. 2023). "Gear metrology – An update". In: *CIRP Annals* 72.2. Num Pages: 27, pp. 725–751. ISSN: 00078506. DOI: 10.1016/j.cirp.2023.05.008.

GuyonIsabelle and ElisseeffAndré (Mar. 2003). "An introduction to variable and feature selection". EN. In: *The Journal of Machine Learning Research*. DOI: 10.5555/944919.944968. URL: https://dl.acm.org/doi/10.5555/944919.944968 (visited on 01/25/2026).

Han, Jiang et al. (Mar. 2022). "Online gear hobbing error estimation based on shaft vibration signal analysis". In: *Mechanical Systems and Signal Processing* 167, p. 108559. ISSN: 0888-3270. DOI: 10.1016/j.ymssp.2021.108559. URL: https://www.sciencedirect.com/science/article/pii/S0888327021008979 (visited on 12/02/2025).

Härtig, F., K. Kniel, and K. Rost (2009). *Messung von Mikroverzahnung. Studie zum Bedarf und den Möglichkeiten der Messung von kleinen Verzahnungen*. Tech. rep. 5671. FVA-Forschungsvorhaben.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman, et al. (2009). *The elements of statistical learning*.

Hora, Stephen C. (1996). "Aleatory and epistemic uncertainty in probability elicitation". In: *Reliability Engineering & System Safety* 54.2–3, pp. 217–223. DOI: 10.1016/S0951-8320(96)00077-4.

Hüllermeier, Eyke and Willem Waegeman (2021). "Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods". In: *Machine Learning* 110.3, pp. 457–506. DOI: 10.1007/s10994-021-05946-3.

Kendall, Alex and Yarin Gal (2017). "What uncertainties do we need in Bayesian deep learning for computer vision?" In: *Advances in Neural Information Processing Systems*. DOI: 10.48550/arXiv.1703.04977.

Kiureghian, Armen Der and Ove Ditlevsen (2009). "Aleatory or epistemic? Does it matter?" In: *Structural Safety* 31.2, pp. 105–112. DOI: 10.1016/j.strusafe.2008.06.020.

Kohavi, Ron et al. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Ijcai*. Vol. 14. Number: 2. Montreal, Canada, pp. 1137–1145.

Lanza, Gisela et al. (Jan. 2019). "In-Line Measurement Technology and Quality Control". In: *Metrology*. Ed. by Wei Gao. Singapore: Springer Singapore, p. 399433. ISBN: 978-981-10-4938-5. URL: 10.1007/978-981-10-4938-5_14.

Papananias, Michail et al. (2019). "A Bayesian framework to estimate part quality and associated uncertainties in multistage manufacturing processes". In: *Computers in Industry* 105, pp. 40–50. DOI: 10.1016/j.compind.2018.10.008.

Runkler, Thomas A. (2015). *Data mining. Modelle und algorithmen intelligenter datenanalyse*. 2nd ed. Computational intelligence. Pages: XII + 145 tex.isbn_electronic: 978-3-8348-2171-3 tex.issn_electronic: 2522-0527. Springer Vieweg Wiesbaden. ISBN: 978-3-8348-1694-8. DOI: 10.1007/978-3-8348-2171-3.

Schiller, Vivian et al. (2023). "In-Process Monitoring of Hobbing Process Using an Acoustic Emission Sensor and Supervised Machine Learning". In: *Algorithms* 16.4, p. 183. ISSN: 1999-4893. DOI: 10.3390/a16040183.

Schmitt, Robert and Edgar Dietrich (Jan. 2023). *Handbuch Messtechnik in der industriellen Produktion: Valide Messergebnisse planen, erhalten, auswerten und verteilen*. Carl Hanser Verlag GmbH Co KG. ISBN: 978-3-446-46559-6.

Tao, Jinyang et al. (Jan. 2023). "An efficient and accurate measurement method of tooth flank variations for face gears". In: *Measurement* 221, p. 113486. ISSN: 0263-2241.

Tercan, Hasan and Tobias Meisen (2022). "Machine learning and deep learning based predictive quality in manufacturing: a systematic review". In: *Journal of Intelligent Manufacturing* 33, pp. 1879–1905. DOI: 10.1007/s10845-022-01963-8.

VDI 2731 (2009). *VDI 2731 Blatt 1:2009-04 – Mikrogetriebe– Grundlagen*. Tech. rep. Düsseldorf: Verein Deutscher Ingenieure (VDI). URL: https://www.vdi.de/richtlinien/details/vdi-2731-blatt-1-mikrogetriebe-grundlagen.