Measurement Science
and Technology

PURPOSE-LED
PUBLISHING™

PAPER • **OPEN ACCESS**

A conformal virtual metrology framework for predictive uncertainty quantification: a full factorial experimental study of micro crown gear quality

To cite this article: Amine Heddoub *et al* 2026 *Meas. Sci. Technol.* **37** 445003

View the article online for updates and enhancements.

You may also like

- Remaining useful life prediction based on time-series features and conformalized quantile regression
Song Mao, Xiaofeng Li and Boyang Zhao
- Dual time-delay vibration control and stability analysis for an automotive active suspension system
Kaiwei Wu, Qing Ouyang, Hongsheng Hu et al.
- Physics-inspired semi-supervised multi-loss learning for internal defect monitoring in laser directed energy deposition
Ziqian Wu, Hao Wu, Cheng Li et al.



This content was downloaded from IP address 129.13.194.163 on 11/11/2026 at 07:42

Measurement Science and Technology

Check for updates

OPEN ACCESS

RECEIVED
25 August 2026

REVISED
5 October 2026

ACCEPTED FOR PUBLICATION
26 October 2026

PUBLISHED
4 November 2026

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

PAPER

A conformal virtual metrology framework for predictive uncertainty quantification: a full factorial experimental study of micro crown gear quality

Amine Heddoub[1,*], **Ali Bilen**[2], **Lazhar Homri**[1], **Ali Siadat**[1,3] and **Gisela Lanza**[2,4]

[1] Arts et Métiers Institute of Technology, Université de Lorraine, LCFC, F-57070 Metz, France
[2] wbk—Institute of Production Science, Karlsruhe Institute of Technology, Karlsruhe, Germany
[3] Arts et Métiers Campus of Rabat, Research Center S3I, Rabat, Morocco
[4] Global Advanced Manufacturing Institute (GAMI), Suzhou SILU Production Engineering Services Co., Ltd, Suzhou, People's Republic of China
[*] Author to whom any correspondence should be addressed.

E-mail: amine.heddoub@ensam.eu

Keywords: uncertainty quantification, conformal prediction, virtual metrology, acoustic emission, predictive quality

Abstract

Virtual metrology uses process data to estimate product quality before physical measurement results become available. However, most models return only a point prediction, which does not show whether an individual prediction is reliable enough for quality decisions. This paper develops a conformal prediction framework for virtual metrology of micro crown gear quality from acoustic emission signals. The framework combines regression pipelines selected for each quality characteristic with Split conformal prediction (Split CP), cross validation plus (CV+), and cross fitted conformalized quantile regression (CF-CQR). The framework is model agnostic and does not require a parametric form for the distribution of the data. Under exchangeability, conformal prediction provides finite sample marginal coverage guarantees for prediction intervals. Split CP, CV+ and CF-CQR are evaluated over repeated data partitions of the full-factorial experimental dataset using empirical coverage, interval width, and the Winkler score. The results show that acoustic emission signals support the prediction of the studied geometric characteristics, although predictive difficulty varies across targets. The conformal methods provide coverage close to the nominal level, while cross validation plus and CF-CQR generally produce more informative intervals than Split conformal prediction. The predicted intervals are used to formulate an interpretable quality assessment logic by distinguishing clear conformity, clear nonconformity, and ambiguous cases that require physical measurement or expert assessment. The proposed framework moves virtual metrology from isolated point predictions towards quality predictions that communicate uncertainty explicitly.

Nomenclature

Abbreviations

Abbreviation	Definition
AE	Acoustic emission
CF-CQR	Cross fitted conformalized quantile regression
CP	Conformal prediction
CV+	Cross validation plus
DOE	Design of experiments
GUM	Guide to the Expression of Uncertainty in Measurement

© 2026 The Author(s). Published by IOP Publishing Ltd

MAE	Mean absolute error
OLS	Ordinary least squares
PCA	Principal component analysis
RMSE	Root mean squared error
VM	Virtual metrology
μCMM	Micro coordinate measuring machine

Main symbols

Symbol	Definition
α	Nominal miscoverage level
$\mathbf{x}_i$	Process-derived feature vector of part i
$y_{i,t}$	Measured value of quality target t for part i
$\mathcal{D}_t$	Regression dataset associated with target t
$\widehat{f}_t(\mathbf{x})$	Point prediction model for target t
$\widehat{C}_{m,t,\alpha}(\mathbf{x})$	Prediction interval produced by method m for target t
$\widehat{L}_{m,t,\alpha}(\mathbf{x})$	Lower prediction interval bound
$\widehat{U}_{m,t,\alpha}(\mathbf{x})$	Upper prediction interval bound
$\mathcal{A}_t$	Acceptance region for target t
LSL_t	Lower specification limit for target t
USL_t	Upper specification limit for target t
N	Number of manufactured parts
p	Number of input features
T	Number of quality targets

1. Introduction

Quality assurance is essential in manufacturing because products must satisfy demanding functional and geometric requirements while production remains efficient and responsive. Many quality characteristics are still verified through measurements performed after production [1–3]. These measurements provide the reference information required for conformity assessment, but they can delay feedback to the production process. Several parts may therefore be manufactured before a quality deviation is detected [1, 4].

Industry 4.0 has expanded the use of industrial data for monitoring and decision making in manufacturing. Sensor networks, the Industrial Internet of Things, connected production systems, and artificial intelligence link physical operations with data acquisition and computational decision layers [5–7]. These technologies increase the amount and diversity of information available during production. Quality 4.0 builds on this development by using industrial data and learning methods to support monitoring, prediction, and faster quality decisions [8, 9].

Predictive quality is one important outcome of this transition. Methods for predictive quality learn relationships between manufacturing data and measured product characteristics to estimate quality outcomes from process observations [10]. These estimates can provide earlier feedback than measurements performed after production. They can support process adjustment, selective inspection, and the identification of parts that require additional assessment. VM is a specific approach to predictive quality. It uses process data and reference measurements to estimate a quality characteristic without immediately performing the corresponding physical measurement [1, 2].

This capability is particularly relevant for microscale components because small geometric deviations can affect their functional performance [11, 12]. This is the case for micro crown gears used in dental drive systems, where geometric accuracy is important for transmission performance [2]. Reference inspection may also require specialized and time consuming measurement procedures [13]. In micro crown gear hobbing, AE signals provide information about the interaction between the cutting tool and the workpiece during manufacturing. Previous studies showed that features extracted from these signals

can be related to process parameters and geometric quality characteristics [2, 14]. These findings established the basis for VM using AE signals.

VM models are commonly formulated as deterministic regression models that provide one point prediction for each quality characteristic [15, 16]. Their performance is generally assessed with the coefficient of determination R^2, RMSE, and MAE [17, 18]. These criteria summarize average prediction performance over a dataset. They do not characterize the reliability of an individual quality prediction [19]. This limitation matters when VM outputs are used for selective inspection or conformity assessment. A predicted value may lie within an acceptance region while the uncertainty associated with that prediction reaches or crosses a specification limit [2].

CP is an uncertainty quantification framework that constructs prediction sets for classification [20] and prediction intervals for regression [21] around the output of an underlying prediction model [22]. Under exchangeability between calibration observations and future observations, these prediction sets and intervals satisfy a finite sample marginal coverage guarantee. The framework is model agnostic and does not require specifying a parametric form for the data-generating distribution [22–24]. It is therefore relevant for VM applications that require predictive uncertainty to be represented explicitly. In this study, uncertainty refers to predictive uncertainty associated with the VM output. The conformal intervals capture the combined effects reflected in the observed prediction errors, but they do not separately quantify the contributions from reference measurement, AE acquisition, and model estimation.

Previous research established the feasibility of predicting microgear quality from AE signals and later investigated model reliability under changing production conditions [14, 17]. Predictive uncertainty was also studied through a framework based on the Guide to the Expression of Uncertainty in Measurement (GUM) [2]. That framework combines uncertainty associated with reference measurements, AE features, and the prediction model. It addresses uncertainty from a metrological propagation perspective. Conformal regression provides a different perspective by calibrating intervals from observed prediction errors. However, it has not yet been investigated for VM of continuous micro crown gear quality characteristics using AE signals.

The objective of this study is to develop a conformal VM framework for predicting micro crown gear quality from AE data and to evaluate it under the controlled full-factorial experimental campaign. The framework provides point predictions with prediction intervals and examines their use in quality assessment that accounts for uncertainty. The main contributions are as follows:

- A VM framework based on AE signals and CP is developed to provide point predictions and prediction intervals for continuous quality characteristics of micro crown gears.
- Split CP, cross validation plus (CV+), and CF-CQR are evaluated over repeated data partitions of the experimental dataset using point prediction criteria, empirical coverage, interval width, and the Winkler score.
- A decision logic based on prediction intervals is formulated as an intended use of the framework. It distinguishes clear conformity and nonconformity predictions from ambiguous cases that require physical measurement or expert assessment.

The remainder of this paper is organized as follows. Section 2 reviews quality prediction using AE signals, uncertainty quantification in VM, and industrial applications of CP. Section 3 presents the proposed methodology. Section 4 describes the micro crown gear case study, the experimental protocol, and the results. Section 5 summarizes the main findings and future research directions.

2. Related work

2.1. Quality prediction using AE signals and VM

Predictive quality methods use manufacturing and product data to estimate quality outcomes before conventional inspection results become available. A systematic review by Tercan and Meisen shows that these methods cover many manufacturing processes, quality characteristics, and machine learning models [10]. VM is a related concept that estimates product quality from process data and reference measurements [1]. The estimate is produced without immediately performing the corresponding physical measurement. Both research streams aim to provide earlier quality information from data generated during production.

AE has been widely studied in machining for monitoring tool condition, tool wear, and process stability [25]. More recent work has examined whether AE signals also contain information about the

quality of the manufactured part. Early applications often focused on indirect indicators such as surface roughness. Later studies extended prediction from AE signals to geometric and functional quality characteristics [2, 14].

Schiller *et al* [14] showed that features extracted from AE signals can support the supervised prediction of process parameters and microgear quality characteristics. Their study established the signal processing and feature extraction basis used in the present work. Bilen *et al* [17] later examined the reliability of these prediction models under tool changes and evolving production conditions. Their results showed that concept drift can affect model performance. More recently, Bilen *et al* [2] extended the AE based VM approach to continuous regression of geometric quality characteristics and quantified the associated predictive uncertainty using a framework based on GUM. Together, these studies demonstrate the potential of VM based on AE signals. They also show that prediction capability depends on the considered quality characteristic and the production conditions.

2.2. Uncertainty quantification in VM

Average point prediction accuracy is not sufficient when predicted quality values support measurement decisions. Several factors contribute to uncertainty in VM models. These factors include reference measurements, process and signal variability, limited training data, incomplete representation of operating conditions, and limitations of the prediction model [2]. Different methods have therefore been proposed to represent predictive uncertainty.

Probabilistic approaches estimate a predictive distribution instead of only a central value. Papananias *et al* [26] used Bayesian linear regression to estimate part quality and its associated uncertainty in a multistage manufacturing process. This approach represents uncertainty through a specified probabilistic model and the assumptions associated with that model.

Prediction intervals can also be learned directly from data. Song *et al* [27] developed an ensemble based on lower and upper bound estimation to predict intervals for steel tensile strength. The intervals were evaluated using coverage and width criteria and were then used to derive a probability of nonconformity for active sampling inspection. This probability was obtained by assuming a normal distribution within the predicted interval. Their case study used an ensemble of multilayer neural networks and a substantially larger dataset than the microgear dataset considered in the present work.

A different approach follows metrological uncertainty propagation. Bilen *et al* [2] developed a framework based on GUM for predicting micro crown gear quality from AE signals. The framework combines uncertainty associated with reference measurements, variation in the AE feature representation, and the prediction model. The resulting predictive distributions provide a metrological representation of the identified uncertainty contributions.

These studies represent predictive uncertainty from different perspectives. Bayesian regression derives uncertainty from a specified probabilistic model, lower and upper bound estimation learns interval limits directly, and metrological propagation identifies and combines individual uncertainty contributions. CP provides another perspective based on the empirical calibration of prediction errors, as presented in the following subsection.

2.3. CP in industrial applications

CP constructs uncertainty outputs by calibrating the results of an underlying prediction model. It produces prediction sets for classification and prediction intervals for regression. Under exchangeability, conformal methods can provide finite sample marginal coverage guarantees without requiring a parametric form for the data distribution [22–24]. The framework is model agnostic.

Compared with the approaches reviewed in the preceding subsection, CP is distinguished by the statistical interpretation of its output. Bayesian regression provides a predictive distribution under a specified probabilistic model, lower and upper bound estimation learns interval limits whose coverage is assessed empirically, and metrological propagation identifies and combines individual uncertainty contributions. CP instead calibrates prediction intervals to a target marginal coverage level and can be applied to different regression models. These properties make CP well aligned with the objective of the present study.

Several conformal regression strategies have been developed. Split CP uses a dedicated calibration subset and provides a simple construction with a finite sample guarantee [24]. Cross validation plus uses predictions from held out folds to use the available data more efficiently and reduce dependence on one calibration split [28]. Conformalized quantile regression combines quantile regression [29] with conformal calibration and allows the initial interval width to vary with the input [30]. Section 3 presents the mathematical constructions used in this study.

Conformal regression has been applied to several engineering problems. Akpabio and Savari [31] studied prediction intervals without distributional assumptions for line edge roughness estimation from semiconductor microscopy images. Zhang *et al* [32] developed a conformal method for uncertainty quantification in process regression with multiple outputs. Conformal intervals have also been studied for remaining useful life estimation [21] and surface roughness prediction in micromachining [18]. Laguna *et al* [33] recently applied temporal conformal calibration to wind turbine anomaly monitoring under evolving operating conditions.

Industrial applications also include conformal classification and anomaly detection. CP sets have been used to communicate ambiguity in fault diagnosis and to support decisions after prediction [20, 34]. Conformal calibration has also been applied to the control of false alarms in fault detection [35]. Diallo *et al* [36] combined conformal p values with false discovery rate control to reduce the operational burden caused by repeated false alarms.

Despite these developments, the reviewed literature does not address conformal regression for VM of continuous micro crown gear quality characteristics using AE signals. Previous work in predictive quality has produced intervals using neural-network-based lower and upper bound estimation [27] and has quantified uncertainty through metrological propagation based on GUM [2]. These approaches are relevant to the same general objective of uncertainty-aware quality prediction, but they differ in their uncertainty construction, underlying prediction models, assumptions, and data requirements. In particular, the approach of Song *et al* [27] relies on an ensemble of multilayer neural networks developed using a substantially larger dataset, whereas the present case study contains only 225 gears and uses target-specific classical regression pipelines. A direct quantitative comparison would therefore combine differences in the uncertainty quantification strategy, predictor architecture, and data regime. The GUM-based framework also addresses a different objective by identifying and propagating individual metrological uncertainty contributions. These methods are therefore not treated as like-for-like experimental baselines. The present study connects AE-based VM with conformal regression and evaluates Split CP, CV+, and CF-CQR as methods for producing prediction intervals and supporting quality assessment that accounts for uncertainty.

3. Proposed methodology

The proposed methodology combines VM based on AE signals with conformal regression. It extends each point prediction with a prediction interval. This interval represents the uncertainty associated with the predicted quality value of a newly produced gear. It can then support conformity assessment that accounts for predictive uncertainty and engineering specification limits.

3.1. Problem formulation

Consider a VM dataset containing N manufactured parts. For each part, process observations are transformed into a feature vector and paired with reference measurements of the considered quality characteristics.

For manufactured part i, let

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^p \tag{1}$$

denote its feature vector derived from the process observations. Reference metrology provides T measured quality characteristics,

$$\mathbf{y}_i = \begin{bmatrix} y_{i,1} & y_{i,2} & \cdots & y_{i,T} \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^T. \tag{2}$$

The complete dataset is therefore

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}. \tag{3}$$

A separate scalar regression problem is formulated for each quality target $t \in \{1, \ldots, T\}$:

$$\mathcal{D}_t = \{(\mathbf{x}_i, y_{i,t})\}_{i=1}^{N}. \tag{4}$$

For target t, a VM model $\widehat{f}_t$ provides the point prediction

$$\widehat{y}_t = \widehat{f}_t(\mathbf{x}). \tag{5}$$

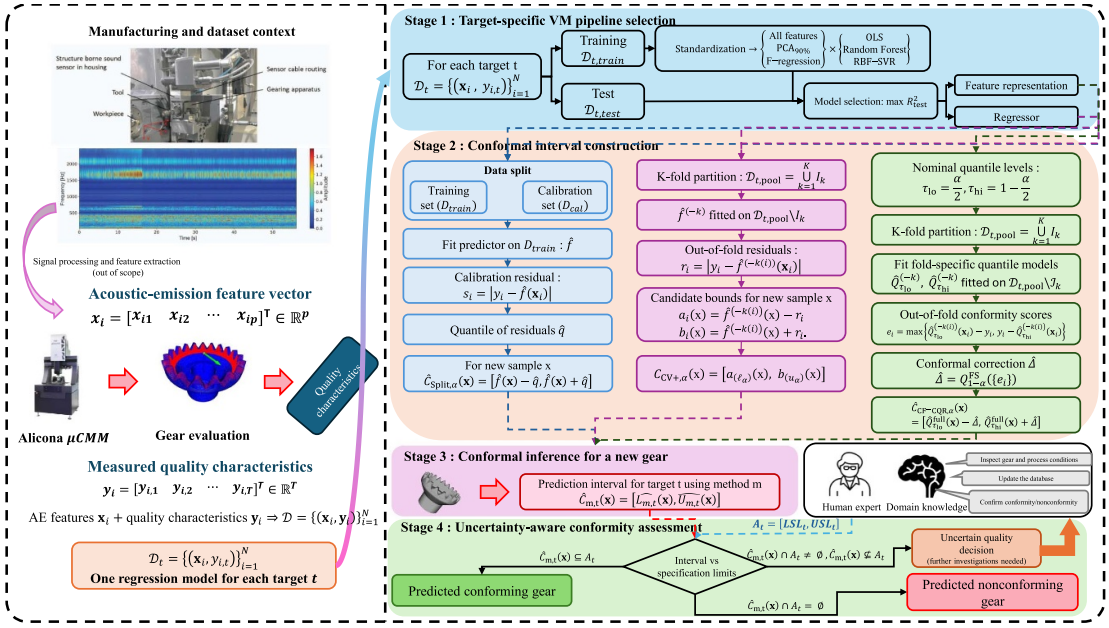


Figure 1. Proposed methodology.

A point prediction alone does not indicate the uncertainty of the predicted quality value. For method m, conformal regression therefore constructs the prediction interval

$$\widehat{C}_{m,t,\alpha}(\mathbf{x}) = \left[\widehat{L}_{m,t,\alpha}(\mathbf{x}), \widehat{U}_{m,t,\alpha}(\mathbf{x})\right], \tag{6}$$

where $m \in \{\text{Split}, \text{CV+}, \text{CF-CQR}\}$ identifies the method used to construct the interval. The value $\alpha \in (0,1)$ is the nominal miscoverage level.

The nominal marginal coverage objective is

$$\mathbb{P}\left\{Y_{\text{new},t} \in \widehat{C}_{m,t,\alpha}(\mathbf{X}_{\text{new}})\right\} \geqslant 1 - \alpha. \tag{7}$$

Conformal coverage results rely on exchangeability between the data used to construct the interval and the future observation. Exchangeability means that the joint distribution of these observations is invariant under permutations of their ordering, so that the calibration observations and a future observation are statistically interchangeable for conformal calibration [22, 23]. Equation (7) states a marginal coverage objective over future observations. It does not imply the same coverage probability for every feature vector or operating condition.

3.2. Conformal VM framework

Figure 1 summarizes the proposed methodology. The left part presents the manufacturing and dataset context. AE signals acquired during hobbing are transformed into feature vectors. These vectors are paired with reference measurements of the gear quality characteristics. A separate regression dataset is then formed for each quality target.

Stage 1 selects a VM pipeline for each target by combining a feature representation with a regression model. Stage 2 applies Split CP and CV+ to the selected point regression pipeline, while CF-CQR retains the selected feature representation and uses a quantile regression estimator. Within Stage 2, the blue, magenta, and green branches correspond to Split CP, CV+, and CF-CQR, respectively. Stage 3 represents inference for a newly produced gear. The feature vector derived from its AE signal is supplied to the calibrated method. The method returns a prediction interval for the considered quality target. Stage 4 compares this interval with the specification limits for the target. The result supports one of three outcomes: predicted conformity, predicted nonconformity, or an uncertain case that requires further investigation.

Stages 1 and 2 use historical gears with available reference measurements for offline model development and conformal calibration. Stages 3 and 4 describe the intended application to a newly produced gear.

For readability, some equations in figure 1 omit the method index m, the target index t, and the nominal miscoverage level α. The following subsections use the complete notation.

The conformity assessment logic in Stage 4 illustrates a possible operational use of the predicted intervals. It is not evaluated as a separate automated decision algorithm in this study.

3.3. Selection of the VM pipeline for each target

A separate VM pipeline is selected for each target dataset $\mathcal{D}_t$. The observations used for model development are divided into a training subset $\mathcal{D}_{t,\text{train}}$ and a test subset $\mathcal{D}_{t,\text{test}}$. The test subset is not used to fit the candidate pipelines. The pipelines are compared using their predictions on this subset.

All candidate pipelines begin with feature standardization. This step prevents features with large numerical scales from dominating the regression model or the feature reduction procedure. Three feature representations are then considered.

The first representation retains all standardized features derived from the AE signals. The second applies PCA and retains the components required to explain 90% of the variance. PCA replaces the original correlated variables with a reduced set of orthogonal components. The third representation uses feature selection based on univariate F tests. Each feature is evaluated separately according to its linear association with the measured response. The retained variables form a subset specific to the target.

Each feature representation is combined with three candidate regressors. These are OLS, random forest regression, and support vector regression with a radial basis function kernel. When required, hyperparameters are selected using only the training observations used for model development. These alternatives follow the previous development of the VM pipeline based on AE signals [2]. The feature representation and regressor are selected according to the largest coefficient of determination on $\mathcal{D}_{t,\text{test}}$:

$$\Pi_t^\star = \arg\max_{\Pi\in\mathcal{P}} R^2_{\text{test}}(\Pi), \tag{8}$$

where $\mathcal{P}$ denotes the set of candidate combinations of feature representations and regression models. RMSE and MAE are also computed as diagnostics during model development, but they do not determine the selected pipeline.

The selected pipeline specifies the feature representation, regressor class, and regressor hyperparameters. This specification is fixed before the repeated conformal evaluation, but the fitted model is not reused. All preprocessing and model parameters are estimated again using only the observations available at each fitting step. This includes standardization parameters, PCA components, selected variables, regression parameters, and the models fitted for each fold. Section 4.3 reports the exact splits used for model development, the settings used for hyperparameter search, and the pipeline selected for each target. Split CP and CV+ use the selected feature representation and point regressor. CF-CQR uses the selected feature representation but replaces the point regressor with a quantile regression estimator.

3.4. Split CP

For clarity, the following equations omit the target index t.

Split CP divides the available data into two disjoint subsets. The training set $\mathcal{I}_{\text{train}}$ has size n_{train}, and the calibration set $\mathcal{I}_{\text{cal}}$ has size n_{cal}. The selected point prediction pipeline is fitted using only the training observations. This gives the predictor $\widehat{f}_{\text{train}}$.

For each calibration observation, the absolute residual defines the nonconformity score

$$s_i = \left|y_i - \widehat{f}_{\text{train}}(\mathbf{x}_i)\right|, \qquad i \in \mathcal{I}_{\text{cal}}. \tag{9}$$

This yields one score for each calibration observation. Let

$$s_{(1)} \leqslant s_{(2)} \leqslant \cdots \leqslant s_{(n_{\text{cal}})} \tag{10}$$

denote the ordered scores, and set $s_{(n_{\text{cal}}+1)} = +\infty$. The rank with the finite sample correction is

$$r_\alpha^{\text{Split}} = \lceil (n_{\text{cal}}+1)(1-\alpha) \rceil, \tag{11}$$

and the conformal correction is

$$\widehat{q} = s_{\left(r_\alpha^{\text{Split}}\right)}. \tag{12}$$

The resulting interval for a new feature vector $\mathbf{x}$ is

$$\widehat{C}_{\text{Split},\alpha}(\mathbf{x}) = \left[\widehat{f}_{\text{train}}(\mathbf{x}) - \widehat{q}, \widehat{f}_{\text{train}}(\mathbf{x}) + \widehat{q}\right]. \tag{13}$$

The same correction $\widehat{q}$ is used for every new observation. The interval is therefore symmetric around the point prediction and has constant width $2\widehat{q}$ for a fixed training and calibration split. Under exchangeability between the calibration observations and the new observation, Split CP satisfies the marginal coverage property in equation (7) [24].

3.5. Cross validation plus

Cross Validation Plus constructs the lower and upper interval endpoints from residuals and predictions obtained across the folds [28]. Let the data pool be indexed by $\mathcal{I}_{\text{pool}}$ and have size n_{pool}. This pool is partitioned into K disjoint folds:

$$\mathcal{I}_{\text{pool}} = \bigcup_{k=1}^{K} \mathcal{I}_k. \tag{14}$$

For fold k, the selected point prediction pipeline is fitted using all pool observations outside $\mathcal{I}_k$. The resulting predictor is denoted by $\widehat{f}^{(-k)}$.

If observation i belongs to fold $k(i)$, its absolute residual is computed without using this fold:

$$r_i = \left| y_i - \widehat{f}^{(-k(i))}(\mathbf{x}_i) \right|. \tag{15}$$

Each pool observation therefore contributes one residual from a model that did not use its fold for training.

For a new feature vector $\mathbf{x}$, observation i contributes the candidate lower and upper bounds

$$a_i(\mathbf{x}) = \widehat{f}^{(-k(i))}(\mathbf{x}) - r_i, \qquad b_i(\mathbf{x}) = \widehat{f}^{(-k(i))}(\mathbf{x}) + r_i. \tag{16}$$

Let

$$a_{(1)}(\mathbf{x}) \leqslant \cdots \leqslant a_{(n_{\text{pool}})}(\mathbf{x}) \tag{17}$$

and

$$b_{(1)}(\mathbf{x}) \leqslant \cdots \leqslant b_{(n_{\text{pool}})}(\mathbf{x}) \tag{18}$$

denote the separately ordered lower and upper candidates. Define

$$\ell_\alpha = \max\left\{1, \left\lfloor \alpha\left(n_{\text{pool}}+1\right)\right\rfloor\right\}, \qquad u_\alpha = \min\left\{n_{\text{pool}}, \left\lceil (1-\alpha)\left(n_{\text{pool}}+1\right)\right\rceil\right\}. \tag{19}$$

The CV+ interval is

$$\widehat{C}_{\text{CV+},\alpha}(\mathbf{x}) = \left[a_{(\ell_\alpha)}(\mathbf{x}), b_{(u_\alpha)}(\mathbf{x})\right]. \tag{20}$$

Unlike Split CP, CV+ does not use one residual quantile around a single predictor. Each endpoint combines predictions from the models fitted across the folds with the corresponding residuals.

Under exchangeability, the standard CV+ construction has finite sample coverage bounds that do not require distributional assumptions. These bounds are weaker than the exact $1-\alpha$ marginal guarantee of Split CP [28]. Its empirical coverage is therefore assessed in the repeated evaluation.

3.6. CF-CQR

Conformalized Quantile Regression with cross fitting combines estimated conditional quantiles with scores obtained across the folds. The nominal lower and upper quantile levels are

$$\tau_{\text{lo}} = \frac{\alpha}{2}, \qquad \tau_{\text{hi}} = 1 - \frac{\alpha}{2}. \tag{21}$$

The method uses the same partition into K folds as CV+. For fold k, two quantile functions are fitted without using the observations in $\mathcal{I}_k$:

$$\widehat{Q}^{(-k)}_{\tau_{\text{lo}}}(\mathbf{x}), \qquad \widehat{Q}^{(-k)}_{\tau_{\text{hi}}}(\mathbf{x}). \tag{22}$$

For observation $i \in \mathcal{I}_{k(i)}$, the nonconformity score is computed without using its fold:

$$e_i = \max\left\{ \widehat{Q}^{(-k(i))}_{\tau_{\text{lo}}}(\mathbf{x}_i) - y_i,\ y_i - \widehat{Q}^{(-k(i))}_{\tau_{\text{hi}}}(\mathbf{x}_i) \right\}. \tag{23}$$

The score is positive when the measured response lies outside the estimated quantile interval. It may be negative when the response lies inside the interval.

Let

$$e_{(1)} \leqslant \cdots \leqslant e_{(n_{\text{pool}})} \tag{24}$$

denote the ordered scores obtained across the folds, and set $e_{(n_{\text{pool}}+1)} = +\infty$. The rank with the finite sample correction is

$$r_\alpha^{\text{CQR}} = \left\lceil \left(n_{\text{pool}} + 1\right)(1-\alpha) \right\rceil, \tag{25}$$

and the conformal correction is

$$\widehat{\Delta} = e_{\left(r_\alpha^{\text{CQR}}\right)}. \tag{26}$$

After the scores have been computed across the folds, a final quantile regression model is fitted using the complete pool. Its lower and upper predictions are

$$\widehat{Q}_{\tau_{\text{lo}}}^{\text{full}}(\mathbf{x}), \qquad \widehat{Q}_{\tau_{\text{hi}}}^{\text{full}}(\mathbf{x}). \tag{27}$$

The final CF-CQR interval is

$$\widehat{C}_{\text{CF-CQR},\alpha}(\mathbf{x}) = \left[\widehat{Q}_{\tau_{\text{lo}}}^{\text{full}}(\mathbf{x}) - \widehat{\Delta},\ \widehat{Q}_{\tau_{\text{hi}}}^{\text{full}}(\mathbf{x}) + \widehat{\Delta}\right]. \tag{28}$$

Unlike the Split CP interval, whose width is constant for a fixed split, the initial quantile width

$$\widehat{Q}_{\tau_{\text{hi}}}^{\text{full}}(\mathbf{x}) - \widehat{Q}_{\tau_{\text{lo}}}^{\text{full}}(\mathbf{x}) \tag{29}$$

depends on the input feature vector. The initial interval width can therefore vary across the feature space. The correction $\widehat{\Delta}$ calibrates these initial intervals.

The construction used here adapts standard Conformalized Quantile Regression [30]. It computes calibration scores across folds and then fits final quantile models on the complete pool. The finite sample coverage theorem for standard split conformalized quantile regression does not directly apply to this exact construction. In standard split CQR, the calibration scores and the score of a future observation are computed using the same quantile models fitted on a separate proper training set. In the present cross fitted construction, calibration scores are obtained from fold specific models, while the final prediction uses quantile models fitted on the complete pool. The reported coverage therefore provides empirical evidence for this cross fitted construction and should not be interpreted as a finite sample coverage guarantee.

3.7. Conformity assessment with prediction intervals

For a newly produced gear, the conformal VM method provides

$$\widehat{C}_{m,t,\alpha}(\mathbf{x}) = \left[\widehat{L}_{m,t,\alpha}(\mathbf{x}), \widehat{U}_{m,t,\alpha}(\mathbf{x})\right]. \tag{30}$$

Let LSL_t and USL_t denote the lower and upper specification limits for target t, respectively. The acceptable range is

$$\mathcal{A}_t = [\text{LSL}_t, \text{USL}_t], \tag{31}$$

A specification with only one finite limit can be represented by setting $\text{LSL}_t = -\infty$ or $\text{USL}_t = +\infty$.

Three situations can then be distinguished.

When the complete prediction interval lies inside the acceptance region,

$$\widehat{C}_{m,t,\alpha}(\mathbf{x}) \subseteq \mathcal{A}_t, \tag{32}$$

the gear is predicted to be conforming for target t.

When the prediction interval and the acceptance region do not overlap,

$$\widehat{C}_{m,t,\alpha}(\mathbf{x}) \cap \mathcal{A}_t = \varnothing, \tag{33}$$

the gear is predicted to be nonconforming for that target.

The remaining case is

$$\widehat{C}_{m,t,\alpha}(\mathbf{x}) \cap \mathcal{A}_t \neq \varnothing, \qquad \widehat{C}_{m,t,\alpha}(\mathbf{x}) \nsubseteq \mathcal{A}_t. \tag{34}$$

Here, the prediction interval crosses at least one specification boundary. The available information is insufficient for an unambiguous conformity decision. The gear can therefore be directed to reference metrology, inspection of the process conditions, or expert assessment.

Reference measurements obtained for such cases may later be stored for periodic dataset enrichment and model updating. Automatic online adaptation is not part of the present methodology. The decision logic in equations (32)–(34) represents an intended operational use of the prediction intervals. It is not evaluated as an independent decision algorithm in this study. The marginal coverage property of the intervals does not provide a direct guarantee on the correctness of these conformity decisions.

3.8. Experimental evaluation criteria

Figure 1 describes the intended operational workflow. The following criteria evaluate the prediction intervals separately from the conformity assessment logic.

For a fixed method and target, let $\mathcal{I}_{\text{eval}}$ denote an independent evaluation set. Its size is n_{eval}. For observation i, define

$$\widehat{L}_i = \widehat{L}_\alpha(\mathbf{x}_i), \qquad \widehat{U}_i = \widehat{U}_\alpha(\mathbf{x}_i). \tag{35}$$

The empirical marginal coverage is

$$\widehat{\text{Cov}} = \frac{1}{n_{\text{eval}}} \sum_{i \in \mathcal{I}_{\text{eval}}} \mathbf{1}\left\{ y_i \in \left[\widehat{L}_i, \widehat{U}_i\right] \right\}. \tag{36}$$

The interval width for observation i is

$$w_i = \widehat{U}_i - \widehat{L}_i, \tag{37}$$

and its mean is

$$\overline{w} = \frac{1}{n_{\text{eval}}} \sum_{i \in \mathcal{I}_{\text{eval}}} w_i. \tag{38}$$

Coverage and width must be considered jointly because excessively wide intervals can produce high coverage. Their joint assessment uses the Winkler interval score [37, 38]

$$W_i = w_i + \frac{2}{\alpha}\left(\widehat{L}_i - y_i\right)\mathbf{1}\left\{y_i < \widehat{L}_i\right\} + \frac{2}{\alpha}\left(y_i - \widehat{U}_i\right)\mathbf{1}\left\{y_i > \widehat{U}_i\right\}. \tag{39}$$

The mean Winkler score is

$$\overline{W} = \frac{1}{n_{\text{eval}}} \sum_{i \in \mathcal{I}_{\text{eval}}} W_i. \tag{40}$$

Lower values of $\overline{w}$ and $\overline{W}$ indicate more informative intervals, provided that empirical coverage remains close to the nominal level.

Point prediction performance is evaluated using

$$\text{RMSE} = \left[\frac{1}{n_{\text{eval}}} \sum_{i \in \mathcal{I}_{\text{eval}}} (y_i - \widehat{y}_i)^2\right]^{1/2}, \qquad \text{MAE} = \frac{1}{n_{\text{eval}}} \sum_{i \in \mathcal{I}_{\text{eval}}} |y_i - \widehat{y}_i|, \tag{41}$$

and

$$R^2 = 1 - \frac{\sum_{i \in \mathcal{I}_{\text{eval}}} (y_i - \widehat{y}_i)^2}{\sum_{i \in \mathcal{I}_{\text{eval}}} (y_i - \overline{y}_{\text{eval}})^2}, \tag{42}$$

where $\overline{y}_{\text{eval}}$ is the mean measured value in the evaluation set.

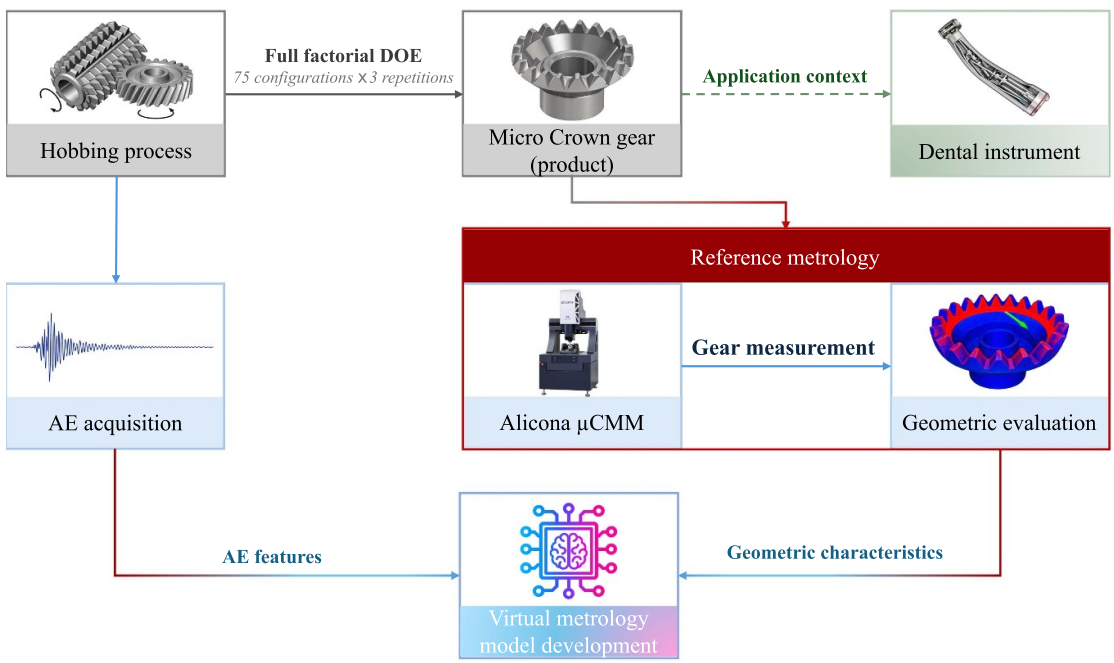


Figure 2. Case-study overview and data sources.

Let $\mathcal{S}$ denote the finite set of random data partitions used in the repeated evaluation. For method m and target t, the mean Winkler score across these partitions is

$$\overline{W}_{m,t} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \overline{W}_{s,m,t}. \tag{43}$$

Interval width, RMSE, MAE, and the Winkler score retain the physical unit of the corresponding target. Because the targets have different numerical ranges, these quantities are compared only within the same target and are not averaged directly across targets.

4. Case study and experimental evaluation

This section presents the controlled full factorial manufacturing dataset used to evaluate the proposed conformal VM methodology. Figure 2 summarizes the case-study context and the main data sources. Micro crown gears are manufactured by hobbing while AE signals are acquired during the process. The manufactured gears are measured using the Alicona μCMM, and the resulting geometric characteristics provide the reference target values for VM. The dental instrument illustrates the application context of the investigated micro crown gear. The following subsections describe the manufacturing campaign, the measured quality characteristics, the regression pipelines, and the experimental protocol.

4.1. Micro crown gear manufacturing and AE acquisition

The case study considers micro crown gears manufactured by hobbing. The manufacturing campaign followed a full factorial DOE in which tool runout, axial infeed (z-shift), and tangential displacement (x-shift) were varied. The influence of characteristic positioning and kinematic errors on the resulting crown gear geometry is investigated in detail by Bilen *et al* [39]. Table 1 reports the investigated levels. The three factors define 75 parameter combinations. Three gears were manufactured for each combination, which resulted in a balanced dataset of 225 gears [2].

A piezoelectric AE sensor was mounted close to the tool holder on a dedicated structure attached to the hobbing aggregate. This arrangement kept a constant distance between the sensor and the cutting zone. The sensor was tightened with a defined torque. The coupling condition was checked by injecting a reference signal into the sensor chain.

The machine control system triggered AE acquisition at the beginning of each machining cycle. The signals therefore shared a common temporal reference relative to the hobbing operation. Each recorded signal was associated with the corresponding manufactured gear [2].

Table 1. Manufacturing factors and levels.

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

Table 2. Predicted gear quality characteristics.

Quality characteristic	Symbol	Unit
Tooth thickness deviation	Δt	μm
Flank angle deviation	$f_{H\beta}$	μm
Profile angle deviation	$f_{H\alpha}$	μm
Profile total deviation	F_α	μm
Profile form deviation	ff_α	μm
Flank total deviation	F_β	μm
Flank form deviation	ff_β	μm

Before feature extraction, Gaussian filtering was applied to reduce noise at high frequencies. Automated outlier detection was also used to remove isolated measurement artefacts. Features were extracted from the raw signal in the time domain and from representations obtained using the short time Fourier transform, continuous wavelet transform, and discrete wavelet transform. The extraction was performed for the complete signal and for predefined segments associated with different stages of tool engagement. This signal processing and feature extraction procedure follows the pipeline introduced by Bilen *et al* [2].

After manufacturing, the gears were measured using an optical μCMM from Alicona. Section 4.2 presents the resulting feature dataset and the selected quality targets.

4.2. Dataset and quality characteristics

After signal processing and feature extraction, each manufactured gear is represented by an AE feature vector $\mathbf{x}_i \in \mathbb{R}^{8573}$. The feature set contains 3865 continuous wavelet transform features, 4650 short time Fourier transform features, 40 raw signal features, and 18 discrete wavelet transform features. The resulting feature matrix is

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_N^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{N\times p}, \qquad N = 225, \quad p = 8573. \tag{44}$$

The same feature matrix is used for all prediction targets. For each target, the response vector contains the quality values obtained from the Alicona μCMM measurements described in section 4.1.

The evaluation of the geometric quality characteristics follows the framework introduced by Bilen *et al* [40]. Seven scalar geometric quality characteristics are considered. They describe tooth thickness, profile and flank deviations as summarized in table 2. The source values were expressed in millimetres and were converted to micrometres for the analyses and results reported in this paper. The terms flank angle deviation and profile angle deviation follow the gear metrology nomenclature. Their recorded values are linear deviations and are therefore also reported in micrometres.

A separate response vector

$$\mathbf{y}_t = \begin{bmatrix} y_{1,t} & y_{2,t} & \cdots & y_{N,t} \end{bmatrix}^\mathsf{T} \in \mathbb{R}^N \tag{45}$$

is formed for each target t. The dataset associated with target t is

$$\mathcal{D}_t = \{(\mathbf{x}_i, y_{i,t})\}_{i=1}^{225}, \tag{46}$$

in accordance with the formulation introduced in equation (4).

Table 3. Selected virtual metrology pipelines.

Target	Feature representation	Regressor	Selected hyperparameters	R^2_{test}
Δt	F-regression	OLS	—	0.931
$f_{H\beta}$	F-regression	Random forest	$n_{\text{trees}} = 100$, $d_{\max} = 20$, $n_{\text{leaf}} = 1$	0.867
$f_{H\alpha}$	F-regression	Random forest	$n_{\text{trees}} = 100$, $d_{\max} = 20$, $n_{\text{leaf}} = 1$	0.863
F_{α}	F-regression	Random forest	$n_{\text{trees}} = 100$, $d_{\max} = 10$, $n_{\text{leaf}} = 5$	0.558
ff_{α}	All features	Random forest	$n_{\text{trees}} = 100$, $d_{\max} = 20$, $n_{\text{leaf}} = 1$	0.687
F_{β}	F-regression	Random forest	$n_{\text{trees}} = 100$, $d_{\max} = 20$, $n_{\text{leaf}} = 2$	0.567
ff_{β}	F-regression	Random forest	$n_{\text{trees}} = 100$, $d_{\max} = 10$, $n_{\text{leaf}} = 5$	0.397

4.3. VM model development

The model selection procedure introduced in section 3.3 was applied separately to the seven quality characteristics. The same fixed indices were used for every target. Model development used 155 observations for training and 30 observations as a fixed test set for pipeline selection. The test observations were not used to fit the candidate pipelines. The remaining 40 observations were not used during this selection stage.

Three feature representations were evaluated. The first retained all 8573 standardized AE features. The second applied PCA after standardization and retained the components required to explain 90% of the variance in the training data. The third applied univariate feature selection based on F tests after standardization. A test that controls the false discovery rate at level 0.05 determined the retained features. This procedure follows the feature reduction approach used in the preceding study of VM based on AE signals [2].

Each feature representation was combined with three regressors. These regressors were OLS, random forest regression, and support vector regression with a radial basis function kernel. The resulting comparison contained nine candidate pipelines for each target.

OLS was fitted without hyperparameter tuning. Random forest regression and support vector regression were tuned using randomized search and cross validation with three folds on the training set. The coefficient of determination was used as the validation criterion. The selected hyperparameter configuration was then refitted using the complete training set.

The nine candidate pipelines were evaluated on the independent test set according to the selection rule in equation (8). Here, $\mathcal{P}$ contains the nine combinations of feature representation and regressor. The retained pipeline achieved the highest test R^2 among the nine candidates for the corresponding target. The test RMSE and MAE were also computed as diagnostic measures, but they were not used for model selection.

Table 3 reports the selected pipelines. For the random forest models, n_{trees} denotes the number of trees, $d_{\max}$ denotes the maximum tree depth, and n_{leaf} denotes the minimum number of observations in a terminal leaf.

The feature representation, regressor class, and regressor hyperparameters selected for each target were fixed before the repeated conformal experiment. The fitted preprocessing transformations and regression models were not reused. At each fitting step, the available observations were used to estimate the standardization parameters, PCA components, selected features, regression coefficients, and model parameters. The same rule was applied within every cross validation fold. The repeated experiment therefore evaluates fixed pipeline specifications and does not include variability caused by model selection.

Split CP and CV+ used the selected point regression pipeline. CF-CQR retained the selected feature representation but replaced the point regressor with the quantile regression estimator described in section 4.4.

4.4. Experimental protocol

The three conformal methods were evaluated over ten random data partitions with a nominal miscoverage level of $\alpha = 0.05$. This level corresponds to a nominal marginal coverage of 95%. For each partition, 20 of the 225 gears were reserved for independent validation. The remaining 205 observations were used

Table 4. Point prediction performance.

	Split CP			CV+			CF-CQR		
Target	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Δt	0.912	15.395	12.442	0.923	14.386	11.559	0.910	15.656	9.121
$f_{H\beta}$	0.899	0.977	0.756	0.909	0.932	0.705	0.908	0.934	0.608
$f_{H\alpha}$	0.884	1.213	0.945	0.900	1.140	0.874	0.899	1.138	0.797
F_α	0.600	0.557	0.472	0.616	0.547	0.457	0.611	0.547	0.447
ff_α	0.549	0.404	0.309	0.577	0.385	0.286	0.572	0.380	0.278
F_β	0.475	0.662	0.523	0.538	0.617	0.487	0.523	0.632	0.459
ff_β	0.491	0.139	0.104	0.540	0.133	0.097	0.562	0.128	0.093

for model fitting and conformal calibration. Within each repetition, all targets and methods used the same validation observations.

The pipeline specifications reported in table 3 remained fixed across the ten repetitions. However, every preprocessing transformation and model parameter was estimated again using only the observations available in the corresponding training subset or cross validation fold.

For Split CP, the pool of 205 observations was divided into 155 training observations, 20 calibration observations, and 30 observations used for point prediction diagnostics. The regression model was fitted using the training observations. Only the calibration observations were used to determine the conformal correction. With $n_{\text{cal}} = 20$ and $\alpha = 0.05$, the finite sample correction corresponded to the largest calibration residual.

CV+ used the complete pool of 205 observations. The pool was divided into $K = 10$ folds. Each observation received a prediction from a model fitted without that observation. The interval endpoints were then obtained from the empirical quantiles of the candidate lower and upper bounds defined in section 3.5.

CF-CQR used the same pool and fold structure. For each target, its selected feature representation was combined with a random forest quantile regression model. The lower, median, and upper quantile levels were 0.025, 0.5, and 0.975, respectively. The fold-specific quantile models used 200 trees. Quantile predictions obtained without using the corresponding observation were used for conformal calibration. The conformal correction was determined from the resulting cross fitted conformity scores. A final quantile model with 300 trees was then fitted using the complete pool.

The repeated design comprised

$$10 \text{ partitions} \times 7 \text{ targets} \times 3 \text{ methods} = 210 \tag{47}$$

combinations of partition, target, and conformal method. For each combination, the prediction intervals were evaluated on the independent validation set using the criteria defined in section 3.8.

The validation observations were excluded from all model fitting and conformal calibration steps. For CV+ and CF-CQR, calibration predictions were generated out of fold, so that each observation was predicted by a model fitted without that observation.

4.5. Results and discussion

4.5.1. Point prediction performance

The point prediction performance is first examined to determine whether the AE features contain sufficient information to predict each gear quality characteristic. This analysis provides context for interpreting the prediction intervals. However, point prediction accuracy, empirical coverage, and interval informativeness describe different properties of the prediction pipeline and must be evaluated separately.

The point predictor associated with each conformal method depends on how the available data are used. For Split CP, the prediction is provided by the model fitted on the proper training set. For CV+, the point prediction is obtained from a model fitted on the complete pool that excludes the validation set. For CF-CQR, the median predicted by the final quantile regression model is used. The reported point prediction metrics are therefore not systematically computed from the midpoint of the prediction intervals.

Table 4 reports the mean validation performance over the ten data partitions. The RMSE and MAE are reported in μm.

The predictive performance depends strongly on the quality characteristic. The mean R^2 values for tooth thickness deviation and the two angle deviation targets were close to or above 0.90. The total

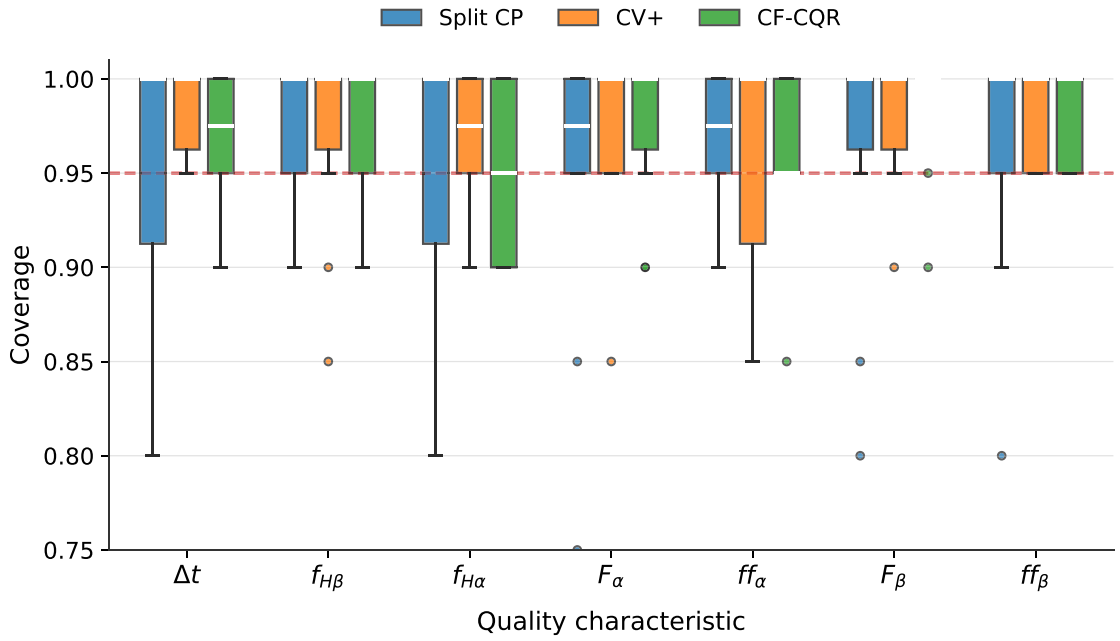


Figure 3. Empirical coverage across repeated partitions.

and form deviations were more difficult to predict. For these targets, the mean R^2 values ranged from approximately 0.48–0.62.

The small differences between the three methods partly result from their different training sets and central predictors. The purpose of this comparison is therefore not to rank the conformal methods according to point prediction alone. Instead, it shows that the underlying VM models have different levels of accuracy across the seven targets. The next analysis examines whether the prediction intervals reliably represent the uncertainty that remains after point prediction.

4.5.2. Coverage and stability

The reliability of the prediction intervals is first assessed through empirical coverage. This criterion represents the proportion of measured quality values contained within the predicted intervals. For the selected miscoverage level $\alpha = 0.05$, the nominal coverage level is 0.95.

Figure 3 shows the empirical coverage over the ten data partitions. The dashed line marks the nominal coverage level of 0.95. Across all targets and partitions, the mean coverage was 0.958 for Split CP, 0.974 for CV+, and 0.970 for CF-CQR. For CF-CQR, the reported coverage is an empirical result of the cross fitted construction used in this study and is not interpreted as a finite sample coverage guarantee.

At the target level, the mean coverage for Split CP ranged from 0.945 to 0.970. The corresponding ranges were 0.960 to 0.985 for CV+ and 0.950 to 0.985 for CF-CQR. CV+ and CF-CQR reached or exceeded 0.95 for all seven targets. Split CP also remained close to this level, although its mean coverage for F_α was 0.945.

As shown in figure 3, coverage varied more across the repetitions for Split CP. This behaviour is consistent with the use of a dedicated calibration set containing only 20 observations. By contrast, CV+ and CF-CQR construct their calibration information from predictions generated across the complete pool that excludes the validation set. This reduces their dependence on one calibration subset.

The variability across repetitions must also be interpreted in relation to the size of the validation set. Each validation set contains 20 gears, so one additional uncovered observation changes the empirical coverage by 0.05. An empirical coverage below 0.95 in one partition does not therefore establish systematic undercoverage. The repeated evaluation provides a more representative assessment than any single validation split.

Overall, the three methods produced prediction intervals with coverage close to the intended level. Coverage alone is nevertheless insufficient because a method can attain high coverage by producing very wide intervals. The interval width and Winkler score are therefore examined next.

Table 5. Interval performance (Bold values show the lowest mean interval width and Winkler score among the three methods for each characteristic).

	Split CP		CV+		CF-CQR	
Target	Width	Winkler	Width	Winkler	Width	Winkler
Δt	67.176	78.835	64.357	**65.547**	**58.319**	65.897
$f_{H\beta}$	4.818	5.144	4.695	4.974	**4.235**	**4.605**
$f_{H\alpha}$	5.690	7.110	**4.961**	**5.466**	5.156	5.686
F_α	2.413	2.953	**2.181**	**2.334**	2.308	2.519
ff_α	2.242	2.683	1.762	2.312	**1.754**	**2.263**
F_β	3.248	3.680	3.109	3.353	**2.856**	**2.950**
ff_β	0.926	1.177	**0.614**	**0.805**	0.744	0.936

4.5.3. Interval informativeness

The practical usefulness of an interval depends on both its coverage and its width. The Winkler score combines these two properties by adding a penalty when the measured value falls outside the interval. Since the seven quality characteristics have different numerical ranges, widths are compared between methods within each target and not across targets.

Table 5 reports the mean interval width and Winkler score over the ten data partitions. All values are reported in μm, and the lowest value for each target and criterion is shown in bold.

Split CP produced the widest mean intervals for all seven targets. Its partition-level mean widths also showed greater variability across the ten data partitions. This result is consistent with its dependence on one calibration subset containing only 20 observations. CV+ and CF-CQR used calibration information from the complete pool that excludes the validation set. In this experiment, they produced narrower and more stable intervals.

CF-CQR obtained the lowest mean width for Δt, $f_{H\beta}$, ff_α, and F_β. CV+ obtained the lowest mean width for $f_{H\alpha}$, F_α, and ff_β. Width must, however, be interpreted together with coverage. For Δt, CF-CQR produced the narrowest mean intervals, but CV+ achieved a slightly lower Winkler score because it provided higher mean coverage.

CV+ obtained the lowest Winkler score for four targets, while CF-CQR obtained the lowest score for three targets. CF-CQR was particularly effective for F_β, where it achieved the highest mean coverage, the lowest mean width, and the lowest Winkler score. No method was uniformly best for every quality characteristic. Nevertheless, both CV+ and CF-CQR provided more informative intervals than Split CP in this case study.

To complement the aggregate comparison, figure 4 shows individual predictions from one additional randomly generated partition. This partition was generated once with the fixed pipelines selected for each target and was not included in the repeated evaluation. The figure presents Δt, which had the highest point prediction accuracy, and F_β, which had the lowest average point prediction accuracy across the three methods. The validation gears are ordered by their measured values to improve readability. This ordering has no temporal or process meaning.

For Δt, Split CP produced a mean width of 45.47 μm, which was approximately 29% smaller than the widths produced by CV+ and CF-CQR in this partition. This narrower interval came with lower empirical coverage. Split CP covered 17 of the 20 measured values, and all three misses occurred below the lower bound. CV+ covered 19 observations, while CF-CQR covered all 20. The coverage of Split CP was therefore below the nominal level in this randomly selected partition. This single result does not indicate systematic undercoverage because each observation changes the coverage by 0.05. It instead illustrates why the repeated analysis is required. It also contrasts with the aggregate results, in which Split CP produced the widest intervals on average.

For F_β, Split CP covered all measured values but produced visibly wider intervals. Its mean width was 4.20 μm, compared with 2.90 μm for CV+ and 2.53 μm for CF-CQR. The Split CP intervals were therefore approximately 45% wider than the CV+ intervals and 66% wider than the CF-CQR intervals in this partition. CV+ produced one miss below the lower bound. CF-CQR combined complete coverage with the smallest mean width. This observation agrees with the repeated evaluation for F_β, where CF-CQR also achieved the lowest Winkler score.

The individual intervals in figure 4 show that coverage and width must be interpreted jointly. For Δt, the narrowest intervals were associated with three misses. For F_β, complete coverage from Split CP was associated with considerably wider intervals. The additional partition is used only to illustrate

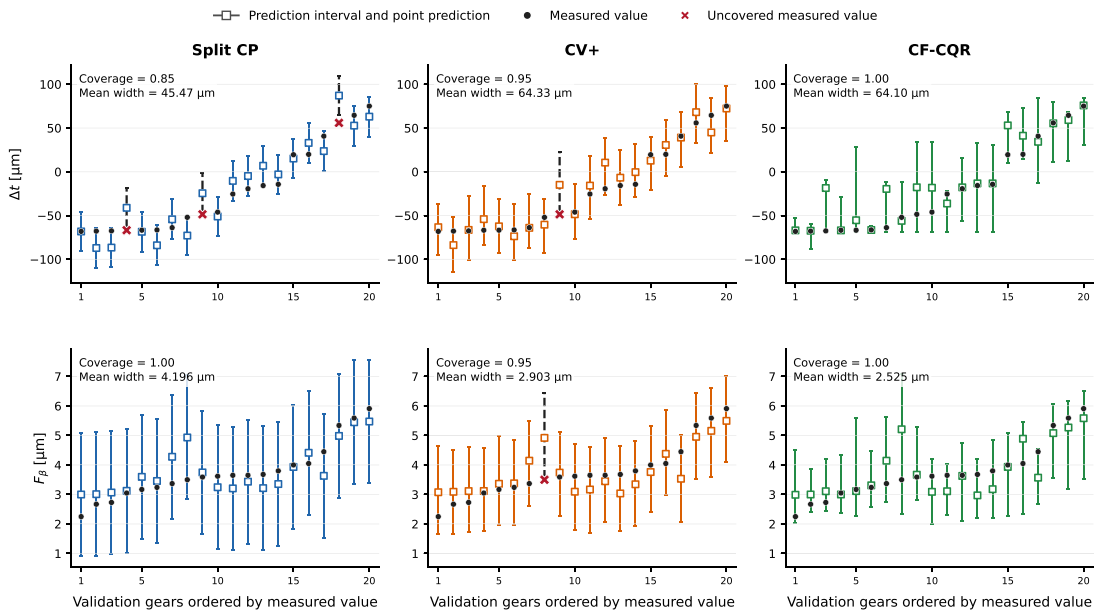


Figure 4. Illustrative conformal prediction intervals.

individual predictions and intervals. All comparisons between the methods remain based on the repeated evaluation across ten partitions.

These results show that the conformal method affects both empirical coverage and interval informativeness.

4.5.4. Implications for uncertainty quantification in VM

The point prediction results confirm that the AE signals contain useful information for predicting the seven gear quality characteristics. The level of predictive accuracy nevertheless differs across the targets. A point prediction alone does not quantify the reliability of an individual prediction. This limitation is especially relevant for the total and form deviations, whose regression models showed lower predictive performance.

The conformal methods complement each point prediction with a prediction interval. The empirical results show that this information can be obtained with coverage close to the nominal 95% level. The intervals therefore express predictive uncertainty that is not available from conventional VM based only on point predictions.

CV+ achieved the lowest Winkler score for four targets, and CF-CQR achieved the lowest score for the other three targets. Both methods therefore provided a better balance between coverage and width than Split CP for this dataset. Split CP remained the simplest method, but it produced wider intervals and greater variability in the partition-level mean widths than CV+ and CF-CQR for all seven targets. This result is consistent with its dependence on a dedicated calibration set containing only 20 observations, although the methods also differ in how they use the available data and construct the intervals. The results do not support one method as universally preferable because the best choice depended on the quality characteristic.

In practice, a prediction interval can be compared with the acceptance region of a quality characteristic. An interval contained entirely within the acceptance region supports a predicted conforming decision. An interval that is completely disjoint from the acceptance region supports a predicted nonconforming decision. An interval that overlaps a specification limit indicates an ambiguous case. The corresponding gear can then be directed to physical measurement or expert assessment.

This logic supports selective quality control. The interval identifies cases in which the available AE information does not support a clear decision. Metrology resources can then be directed towards these uncertain cases instead of treating all point predictions as equally reliable. The prediction interval reports uncertainty explicitly, but it does not by itself guarantee the correctness of the final conformity decision.

The conformity assessment stage was not evaluated experimentally because specification limits and operational decision costs were not available for the seven targets. The present evaluation characterizes the prediction intervals and their empirical coverage. The industrial decision strategy remains an intended application that requires further validation with actual specifications and reference inspections.

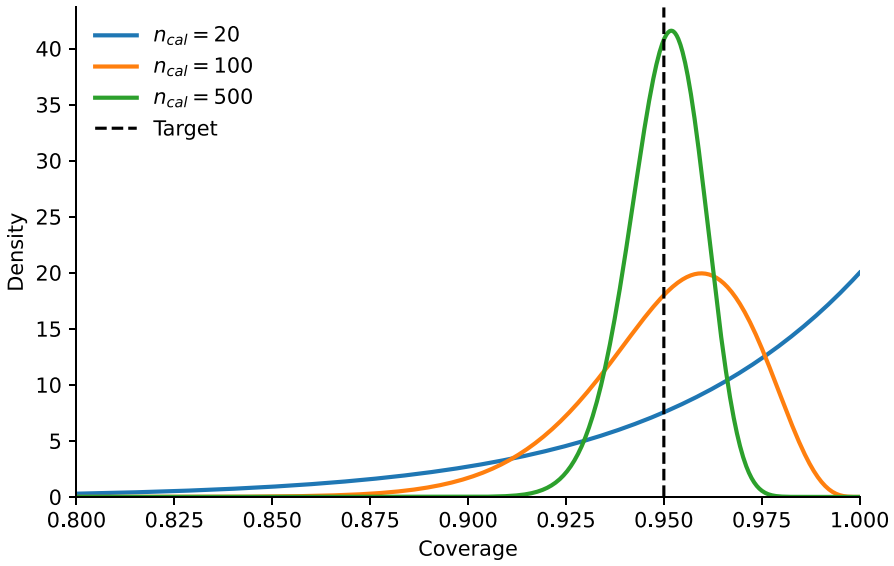


Figure 5. Theoretical effect of calibration set size on conditional coverage for Split CP at $\alpha = 0.05$.

Overall, the results show a progression from VM based only on point predictions towards quality prediction that includes uncertainty information. The point prediction analysis confirms that AE features contain useful but target dependent information about gear geometry. The conformal evaluation shows that the remaining uncertainty can be represented by intervals with coverage close to the selected nominal level. CV+ and CF-CQR produced the most informative intervals, while Split CP offered a simpler approach that used the available data less efficiently. These findings support CP as a practical layer for uncertainty quantification in VM based on AE signals.

4.5.5. Limitations and implications for production

The present evaluation is based on 225 gears produced through a balanced DOE. This design covers the selected combinations of tool runout, z-shift, and x-shift in a controlled way. However, it does not represent all conditions that may occur during routine production. A production process may remain close to one operating region for a long period. Tool wear and process drift also develop over time and were not reproduced in the present experiment. The reported results should therefore be interpreted within the investigated experimental domain.

The limited number of gears also affects conformal calibration. For Split CP, only $n_{\text{cal}} = 20$ gears are used for calibration in each repetition. To illustrate the effect of calibration size, consider a fixed proper training set $\mathcal{D}_{\text{train}}$ and its fitted predictor. For this calculation, calibration observations and a future observation are assumed to be independent draws from the same distribution and independent of the proper training set. The nonconformity scores are assumed to have a continuous distribution. The coverage conditional on the training and calibration sets is

$$\Theta_{\text{cal}} = \mathbb{P}\left(Y_{\text{new}} \in \widehat{C}_{\text{Split},\alpha}\left(\mathbf{X}_{\text{new}}\right) \mid \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{cal}}\right). \tag{48}$$

With the proper training set and predictor fixed, its distribution across calibration samples is, for the calibration sizes considered here, [22]

$$\Theta_{\text{cal}} \sim \text{Beta}\left(n_{\text{cal}} + 1 - \ell, \ell\right), \qquad \ell = \lfloor (n_{\text{cal}} + 1)\,\alpha \rfloor. \tag{49}$$

For the value $\alpha = 0.05$ used in this study, the standard deviation of this distribution is approximately 0.045 for $n_{\text{cal}} = 20$, 0.021 for $n_{\text{cal}} = 100$, and 0.010 for $n_{\text{cal}} = 500$. These values illustrate how a larger calibration set can reduce variability without changing the nominal coverage objective $1 - \alpha$. Figure 5 shows the corresponding distributions.

This calculation isolates the effect of calibration sample variability under the stated assumptions. It does not describe the full distribution of empirical coverage across the DOE partitions, which also involves model refitting and a finite validation set. Each repetition uses 20 validation gears. The empirical coverage therefore changes by $1/20 = 0.05$ when one additional gear is not covered. Table 6 shows how this effect decreases when the validation set becomes larger.

Table 6. Effect of validation set size on empirical coverage resolution.

Validation gears	Coverage with one miss	Change in coverage
20	0.950	0.050
50	0.980	0.020
100	0.990	0.010
200	0.995	0.005

This explains why small differences between mean coverage values should be interpreted carefully. The repeated partitions reduce the dependence on one validation split and show the variability across repetitions, but they do not replace validation on a larger independent dataset.

The distribution of operating conditions is also different between the DOE and routine production. The DOE gives balanced representation to the selected factor combinations. Production data may contain many observations near normal operating conditions and only a few observations in rare regions. Since the conformal guarantee considered here is marginal, good overall coverage does not guarantee the same coverage in every operating region. A larger production dataset may therefore improve the stability of the global calibration while still providing limited information in rarely observed regions.

Finally, the present evaluation does not include temporal changes in the process. Tool wear, machine interventions, or process drift can change the relationship between the AE features and gear quality. Calibration observations and future gears may then no longer be exchangeable. In this case, the standard conformal coverage results do not directly apply. Adaptive conformal calibration can be used to address such changes [41]. Its evaluation requires temporally ordered production data and is left for future work.

5. Conclusion

This study developed a CP framework for VM and evaluated it on the prediction of continuous micro crown gear quality characteristics from AE data. The framework retains the regression pipeline selected for each quality characteristic and extends each point prediction with a prediction interval. Split CP, CV+, and CF-CQR were evaluated to determine how conformal regression can represent the uncertainty that remains after point prediction. The framework therefore moves this application from isolated quality predictions towards predictions accompanied by calibrated uncertainty information.

The results confirmed that predictive performance depends strongly on the quality characteristic. The conformal methods provided empirical coverage close to the selected nominal level. Their interval informativeness nevertheless varied across methods and targets. CV+ and CF-CQR generally offered a better balance between coverage and interval width than Split CP. However, no method was uniformly preferable for every target. These findings show that point prediction accuracy, empirical coverage, and interval informativeness describe different properties and should be evaluated separately.

The prediction intervals also provide an interpretable basis for quality assessment. An interval contained within an acceptance region supports a predicted conforming decision. An interval that is disjoint from that region supports a predicted nonconforming decision. An interval that crosses a specification limit indicates an ambiguous case that should be directed to physical measurement or expert assessment. This logic could support selective inspection. Metrology resources could be directed towards predictions for which the available AE information does not support a clear decision. However, this decision stage was not evaluated experimentally because specification limits and operational decision costs were not available for the studied characteristics.

The experimental evaluation is based on a controlled full factorial DOE. Producing gears and obtaining accurate reference measurements are costly and time consuming, which limits the amount of data available for model development and validation. The limited number of validation observations also restricts the precision with which empirical coverage can be assessed. The conformal coverage guarantee is marginal and does not ensure the same coverage within every process condition or region of the prediction space. In addition, the present experiment does not evaluate temporal changes in the manufacturing process. Tool wear, process drift, or other changes may violate exchangeability and affect interval validity.

Future work should evaluate the framework over successive manufacturing campaigns and under evolving production conditions. Adaptive or online conformal approaches could update the calibration procedure when process drift changes the relationship between AE signals and gear quality. Another

direction is to assess and improve coverage within sufficiently represented groups. These groups could represent process configurations or observable operating regions. This analysis would complement the reported marginal coverage. Coverage should also be examined retrospectively across different ranges of the quality characteristics to identify regions in which the intervals are less reliable. Finally, the proposed quality assessment logic should be validated using actual specification limits, reference inspections, and operational decision costs.

Overall, this work establishes a practical CP layer for VM based on AE signals. It supports quality predictions that communicate uncertainty explicitly. It also provides a foundation for more reliable use of VM models in industrial quality assessment.

Data availability statement

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

Funding

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 55 868 2077. The authors gratefully acknowledge the financial support provided by the DFG.

Author contributions

Amine Heddoub 0009-0009-4350-5512
Conceptualization (equal), Data curation (equal), Formal analysis (equal), Funding acquisition (equal), Investigation (equal), Methodology (equal), Project administration (equal), Software (equal), Validation (equal), Visualization (equal), Writing – original draft (equal)

Ali Bilen 0009-0002-8571-4653
Data curation (equal), Formal analysis (equal), Funding acquisition (equal), Investigation (equal), Validation (equal), Writing – original draft (equal)

Lazhar Homri 0000-0001-6754-844X
Conceptualization (equal), Formal analysis (equal), Funding acquisition (equal), Investigation (equal), Methodology (equal), Supervision (equal), Validation (equal), Writing – original draft (equal)

Ali Siadat 0000-0002-8406-1892
Funding acquisition (equal), Project administration (equal), Resources (equal), Supervision (equal)

Gisela Lanza
Funding acquisition (equal), Project administration (equal), Resources (equal), Supervision (equal)

References

[1] Dreyfus P A, Psarommatis F, May G and Kiritsis D 2022 Virtual metrology as an approach for product quality estimation in Industry 4.0: a systematic review and integrative conceptual framework *Int. J. Prod. Res.* **60** 742–65
[2] Bilen A, Decman M, Stamer F and Lanza G 2026 A method for uncertainty quantification in virtual metrology models: acoustic-emission-based quality prediction in micro crown gear manufacturing *Meas. Sci. Technol.* **37** 265005
[3] Catalucci S, Thompson A, Piano S, Branson D T and Leach R 2022 Optical metrology for digital manufacturing: a review *Int. J. Adv. Manuf. Technol.* **120** 4271–90
[4] Catalucci S, Thompson A, Eastwood J, Zhang Z M, Branson D T, Leach R and Piano S 2022 Smart optical coordinate and surface metrology *Meas. Sci. Technol.* **34** 012001
[5] Lee J, Bagheri B and Kao H A 2015 A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems *Manuf. Lett.* **3** 18–23
[6] Lu Y 2017 Industry 4.0: a survey on technologies, applications and open research issues *J. Ind. Inf. Integr.* **6** 1–10
[7] Hamed O and Hamlich M 2024 Navigation method for autonomous mobile robots based on ROS and multi-robot improved Q-learning *Prog. Artif. Intell.* (https://doi.org/10.1007/s13748-024-00320-5)
[8] Escobar C A, McGovern M E and Morales-Menendez R 2021 Quality 4.0: a review of big data challenges in manufacturing *J. Intell. Manuf.* **32** 2319–34
[9] Zouhri W, Homri L and Dantan J Y 2022 Identification of the key manufacturing parameters impacting the prediction accuracy of support vector machine (SVM) model for quality assessment *Int. J. Interact. Des. Manuf. (IJIDeM)* **16** 177–96

[10] Tercan H and Meisen T 2022 Machine learning and deep learning based predictive quality in manufacturing: a systematic review *J. Intell. Manuf.* **33** 1879–905
[11] Bilen A, Decman M, Stamer F and Lanza G 2025 Minimizing optical measurement time of micro spur gears through point cloud completion techniques *Meas. Sci. Technol.* **36** 125202
[12] Gauder D, Gölz J, Jung N and Lanza G 2023 Development of an adaptive quality control loop in micro-production using machine learning, analytical gear simulation and inline focus variation metrology for zero defect manufacturing *Comput. Ind.* **144** 103799
[13] Goch G, Guenther A, Peng Y and Ni K 2023 Gear metrology–An update *CIRP Ann.* **72** 725–51
[14] Schiller V, Klaus S, Bilen A and Lanza G 2023 In-process monitoring of hobbing process using an acoustic emission sensor and supervised machine learning *Algorithms* **16** 183
[15] Maitra V, Su Y and Shi J 2024 Virtual metrology in semiconductor manufacturing: current status and future prospects *Expert Syst. Appl.* **249** 123559
[16] Cai H, Feng J, Yang Q, Li F, Li X and Lee J 2021 Reference-based virtual metrology method with uncertainty evaluation for material removal rate prediction based on Gaussian process regression *Int. J. Adv. Manuf. Technol.* **116** 1199–211
[17] Bilen A, Grommes A, Stamer F and Lanza G 2025 Coping with concept drift in a virtual metrology application to predict part quality in micro gear manufacturing *Meas.: Sens.* **38** 101797
[18] He Z, Sun Y, Yuan S, Qin F, Zhang X, Cheung C F, Cao H and Wang C 2026 A hybrid intelligence framework with uncertainty quantification for reliable surface roughness prediction in micro-milling *Expert Syst. Appl.* **326** 132774
[19] Cai H, Feng J, Yang Q, Li W, Li X and Lee J 2020 A virtual metrology method with prediction uncertainty based on Gaussian process for chemical mechanical planarization *Comput. Ind.* **119** 103228
[20] Heddoub A, Diallo A R, Homri L, Dantan J Y and Siadat A 2025 Uncertainty-aware fault diagnosis with conformal prediction *IFAC-PapersOnLine* **59** 536–41
[21] Javanmardi A and Hüllermeier E 2023 Conformal prediction intervals for remaining useful lifetime estimation *Int. J. Progn. Health Manag.* **14** 2
[22] Angelopoulos A N and Bates S 2023 Conformal prediction: a gentle introduction *Found. Trends Mach. Learn.* **16** 494–591
[23] Vovk V, Gammerman A and Shafer G 2005 *Algorithmic Learning in a Random World*edn 1 (Springer) pp I–XV, 1–324
[24] Lei J, G'Sell M, Rinaldo A, Tibshirani R J and Wasserman L 2018 Distribution-free predictive inference for regression *J. Am. Stat. Assoc.* **113** 1094–111
[25] Teti R, Jemielniak K, O'Donnell G and Dornfeld D 2010 Advanced monitoring of machining operations *CIRP Ann.* **59** 717–39
[26] Papananias M, McLeay T E, Mahfouf M and Kadirkamanathan V 2019 A Bayesian framework to estimate part quality and associated uncertainties in multistage manufacturing *Comput. Ind.* **105** 35–47
[27] Song Y, Li F, Wang Z, Zhang B and Zhang B 2024 Uncertainty quantification of data-driven quality prediction model for realizing the active sampling inspection of mechanical properties in steel production *Int. J. Comput. Intell. Syst.* **17** 74
[28] Barber R F, Candès E J, Ramdas A and Tibshirani R J 2021 Predictive inference with the jackknife+ *Ann. Stat.* **49** 486–507
[29] Koenker R and Bassett G 1978 Regression Quantiles *Econometrica* **46** 33–50
[30] Romano Y, Patterson E and Candès E J 2019 Conformalized quantile regression *Proc. 33rd Int. Conf. on Neural Information Processing Systems (Red Hook, NY, USA)* vol 318 (Curran Associates Inc) pp 3543–53
[31] Akpabio I I and Savari S A 2022 On the construction of distribution-free prediction intervals for an image regression problem in semiconductor manufacturing (arXiv:2203.03150 [cs.CV])
[32] Zhang R, Zhou P and Chai T 2023 Improved Copula-based conformal prediction for uncertainty quantification of multi-output regression *J. Process Control* **129** 103036
[33] Laguna J, Janal Y, Reder M and Melero J J 2026 Temporal conformal prediction for wind turbine anomaly detection: a robust distribution-free framework with conditional clustering *J. Phys.: Conf. Ser.* **3224** 062054
[34] Heddoub A, Homri L and Siadat A 2026 Class-conditional conformal prediction with discriminant-analysis backbones for reliable open-set fault diagnosis in safety-critical industrial systems *J. Process Control* **161** 103701
[35] Diallo A R, Homri L and Dantan J Y 2025 Reducing false alarms in fault detection: A comparative analysis between conformal prediction and classical methods applied to PCA and autoencoders *J. Process Control* **152** 103495
[36] Diallo A R, Homri L, Boeuf T, Dantan J Y and Bonnet F 2026 Quantifying and mitigating alarm fatigue caused by fault detection systems *Reliab. Eng. Syst. Saf.* **267** 111890
[37] Winkler R L 1972 A decision-theoretic approach to interval estimation *J. Am. Stat. Assoc.* **67** 187–91
[38] Gneiting T and Raftery A E 2007 Strictly proper scoring rules, prediction and estimation *J. Am. Stat. Assoc.* **102** 359–78
[39] Bilen A, Paul Braunschweiger N, Karabegović R, Felix Kircher T, Eduardo Waggershauser F, Malte Wübbena P, Antonia Weiland L, Robert Reister L, Gindin E and Lanza G 2026 A kinematics-based simulation framework for face gear hobbing and manufacturing error analysis *Meas. Sci. Technol.* **37** 266004
[40] Bilen A, Ernst L, Gindin E, Schubert J, Rauwolf M, Leon Rauschenberg E, Michael Schnaith S, Edelmann T, Benfer M and Lanza G 2026 Quality features for Holistic evaluation and quality control of micro face gears *Meas. Sci. Technol.* **37** 275005
[41] Barber R F, Candes E J, Ramdas A and Tibshirani R J 2023 Conformal prediction beyond exchangeability (arXiv:2202.13415 [stat:ME])