Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hybrid digital twin with Bayesian synchronisation for uncertainty-aware tool-wear prognostics and predictive tool replacement in milling

Rabah BOUDIAF

rabah.boudiaf@univ-annaba.dz

University of Badji Mokhtar Annaba: Universite Badji Mokhtar Annaba https://orcid.org/0000-0002-0731-9423

Research Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11597802/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Hybrid digital twin with Bayesian synchronisation for uncertainty-aware tool-wear prognostics and predictive tool replacement in milling

Rabah Boudiaf[1*]

[1*]Electromechanical Systems Laboratory, Department of Electromechanics, Faculty of Technology, Badji Mokhtar–Annaba University, BP 12, Annaba, 23000, Algeria.

Corresponding author(s). E-mail(s): rabah.boudiaf@univ-annaba.dz;

Abstract

Tool wear in milling is usually monitored with regressors that map sensor features to flank wear. Such models rarely follow the individual tool, seldom report calibrated uncertainty and are almost never judged by the replacement decisions they support. This study presents a hybrid digital twin of a face-milling tool. Its virtual entity couples a three-stage flank-wear law, a population prior that conditions the wear parameters on feed, depth of cut and work material, and a probabilistic virtual sensor whose error is split into a persistent per-tool bias and heavy-tailed noise. A particle filter with kernel smoothing updates the wear state, tool-specific parameters and sensor bias after every cut and accepts occasional wear inspections. The predicted remaining-useful-life distribution drives cost-based replacement rules and a what-if advisory on feed and depth of cut. On the NASA Ames milling data (15 tools), under nested leave-one-tool-out validation, the twin estimated flank wear with a root-mean-square error of 63 µm, against 74 µm for the virtual sensor alone and at least 104 µm for four machine-learning regressors, with 90% of the measurements inside its nominal 95% intervals. Inspecting every third measured run cut the error on the other runs to 46 µm. The median relative error of the predicted remaining life was 24%. Replacing tools when the expected remaining life fell below a tuned multiple of the next run length lowered the cost rate by 31% relative to a condition-based schedule, for a failure costing five replacements, and avoided all failures.

Keywords: Digital twin, Predictive maintenance, Tool wear prognostics, Remaining useful life, Particle filter, Face milling

1 Introduction

Flank wear of the cutting edge sets the useful life of an indexable milling cutter. As the wear land grows, cutting forces and temperatures rise, the machined surface deteriorates and, beyond a certain point, the edge chips or breaks, which can scrap the part and load the spindle. Most shops still manage this risk with tool-life counters fixed in advance from handbook values or experience. A counter ignores the spread in life between nominally identical inserts and the effect of the cutting conditions actually used; it therefore discards part of the life of slow-wearing edges and still lets fast-wearing ones fail. Predictive maintenance replaces the counter by a decision based on the observed condition of the tool [1–3], and tool condition

monitoring (TCM) has become one of the most active topics of machining research [4, 5].

The digital twin (DT) concept gives this idea an architecture. A DT is a virtual representation of a specific physical asset that is updated with data acquired on that asset, predicts its behaviour and feeds decisions back to it [6–9]; ISO 23247-1 provides a reference framework for DTs in manufacturing [10]. In maintenance, DTs have been proposed for complex equipment [11, 12], reviewed for predictive maintenance in general [13–15] and applied to machine tools [16–18]. For the cutting tool itself, recent studies describe tool twins that combine mechanistic simulation, knowledge models or deep networks with on-line signals [19–26].

Three weaknesses recur in this literature. First, the twinning itself, that is, the way the virtual entity is kept consistent with one particular tool as evidence accumulates, is often left implicit. Many tool twins are in effect wear regressors with a visualisation layer: they map the latest features to a wear value but carry no state whose uncertainty is propagated between cuts and corrected by each new observation. Treating the DT as a probabilistic model that is updated sequentially, as argued by Kapteyn et al. [27] and examined by Ward et al. [28], makes this step explicit. Second, the uncertainty of wear and remaining useful life (RUL) estimates is seldom assessed as a forecast. Interval widths are reported without their coverage, and proper scoring rules [29] are rarely used, although an overconfident RUL is more harmful for maintenance than an imprecise but honest one [30]. Third, a prognostic model exists to improve a maintenance decision [31], yet tool-wear studies nearly always stop at an error metric. Whether a twin lowers the cost of tool replacement compared with what a shop can already do with a schedule is usually not tested [32, 33].

A more practical issue concerns validation. Public milling data sets contain only a handful of tools. When features, hyperparameters or decision thresholds are chosen with all tools in view and the model is then tested on one of them, information leaks from the test tool into the model, and the reported accuracy no longer describes a new tool. With a small fleet, this leakage can decide the ranking of methods.

This paper develops and evaluates a hybrid DT of a face-milling tool that addresses these points. The twin core is a dynamic Bayesian network whose transition model is a three-stage flank-wear law and whose observation model is a probabilistic virtual sensor built on spindle-current and acoustic-emission features. The parameters of the wear law are specific to each tool, but a new tool does not start from scratch: a population prior learnt from previously used tools conditions them on feed, depth of cut and work material. The error of the virtual sensor is split into a part that persists over the life of a tool and a heavy-tailed part that changes from cut to cut. A particle filter with kernel smoothing synchronises the wear state, the wear parameters and the sensor bias after every cut and also accepts occasional direct wear measurements. The resulting RUL distribution drives replacement rules and a what-if advisory for changes in feed and depth of cut. The contributions are the following:

- a twin core in which the wear state, the tool-specific wear parameters and the virtual-sensor bias are estimated jointly, with a condition-dependent population prior that transfers knowledge from past tools to a new one and an inspection channel for occasional wear measurements;
- a nested leave-one-tool-out protocol in which feature selection, the virtual sensor, the sensor-error model, the population prior, the noise levels of the filter and all decision thresholds are learnt without the test tool;
- an evaluation at three levels, namely wear estimation with calibration and the continuous ranked probability score, RUL prediction, and replacement decisions scored by a renewal cost rate over a range of failure-to-replacement cost ratios, together with an ablation of the twin components;
- a what-if service that recomputes the RUL distribution for modified cutting conditions from the current twin state, with an analysis of what such advice means in terms of cutting time and removed material.

Section 2 reviews related work and Section 3 describes the twin. Section 4 presents the data set and the evaluation protocol, Section 5 the results and Section 6 their implications and limitations. Section 7 concludes.

2 Related work

2.1 Digital twins for predictive maintenance

Reviews agree on the elements of a DT, namely a physical entity, a virtual entity and the data and decision flows between them, but differ on how much fidelity and autonomy the virtual entity needs [7, 8, 34, 35]. Tao et al. [36] treat the construction of the virtual model as a modelling discipline of its own, and Lu et al. [37] place the twin in a reference model for smart manufacturing. For maintenance, the five-dimension twin of Tao et al. [11] supported prognostics and health management of complex equipment, Aivaliotis et al. [12] tuned physics-based simulation models to the asset to estimate its RUL, and later reviews document a fast growth of the topic together with a shortage of validated industrial cases [13–15]. Applications include aero-engines [38], conveyor belts [39], gears [40], rolling bearings [41–45], ball-screw feed drives [46] and robot-integrated cells [47]; Ma et al. [48] review the case of rotating machinery. Twin-supported maintenance decisions have been coupled with production planning [49], and recent architectures add language models for cross-domain integration [50].

The way the virtual entity is kept in step with the asset varies widely. Some twins update model parameters at discrete instants [44, 51]; others use simulations of the asset to train or adapt a learning model [42, 43]. Kapteyn et al. [27] formalised the predictive DT as a probabilistic graphical model in which physical state, digital state, observations and control actions are linked over time, and Ward et al. [28] compared particle filtering with batch Bayesian calibration for the continuous updating of a twin. Thelen et al. [52, 53] stress that uncertainty quantification is what allows a twin to support decisions. The present study adopts this probabilistic reading of the DT.

2.2 Data-driven tool condition monitoring

Most TCM research learns a mapping from features of force, current, vibration or acoustic-emission (AE) signals to flank wear [4, 5]. Gaussian process regression [54], convolutional networks fusing several signal domains [55], recurrent and attention-based architectures [56–58] and image encodings of multichannel signals [59] have all reached low errors on their test sets, and long short-term memory networks have predicted tool RUL under variable working conditions [60]. Earlier work assessed tool health with support vector regression [61], and condition-adaptive hidden Markov models estimated wear states and RUL in micro-milling [62]. Uncertainty-aware variants have been reported, based on Bayesian inference for end milling of titanium alloy [63] and on deep learning with low-cost smartphone sensors [64]. Two limitations are common to this family. A learned map has no memory of the history of the tool beyond its input window, so consecutive estimates can contradict the monotone nature of wear; and the map needs many run-to-failure trajectories to cover the space of cutting conditions, which a shop rarely has [65].

2.3 Physics-informed and hybrid wear models

Physical knowledge of wear enters tool-wear models in several ways. Classical models relate the wear rate to contact stress, sliding speed and temperature [66], and generic semi-empirical laws have been calibrated for high-speed milling [67]. Physics-informed machine learning [68, 69] embeds such knowledge in learning models: physics-guided neural networks [70], physics-informed meta-learning [71], recurrent Gaussian processes with instance transfer [72], physics-informed deep learning [73] and hidden Markov models [74], Gaussian process regression with physical priors and quantified uncertainty [75, 76], symbolic regression [77], multi-task transfer networks [78], models for multiple machining conditions [79] and liquid neural networks [80]. Fusion frameworks combine a physics-based estimate with a data-driven one [81, 82]. These approaches generalise better than purely data-driven ones, but most remain batch predictors: the

physics shapes the model during training instead of serving as a state-transition model that is corrected cut by cut.

2.4 Bayesian filtering for tool wear and RUL

Filtering methods treat wear as a hidden state and are therefore a natural engine for a twin. Particle filters are well established in prognostics [83]. Wang et al. [84] enhanced a particle filter for tool-wear prediction; Ma et al. [85] proposed a hybrid-driven probabilistic state-space model for tool-wear monitoring; Li et al. [86] linked deep features with a Wiener degradation process to predict tool RUL; Yang et al. [87] used an interacting multiple model for wear and RUL in precision machining; and Ding et al. [88] combined a fractional-order grey model with particle filtering for milling tools. Liu et al. [89] showed how degradation can be inferred under time-varying operating conditions with a particle filter. The twin proposed here belongs to this family but differs in four respects: the wear-law parameters are drawn from a prior that is conditioned on the cutting conditions and learnt from a fleet of past tools; the persistent error of the virtual sensor is a state instead of a nuisance absorbed in white noise; the evaluation is leakage-free at the level of tools; and the twin is judged on the decisions it supports.

2.5 From prognostics to replacement decisions

Maintenance optimisation has a long tradition [31], and prognostics-driven policies have been formulated as dynamic predictive maintenance with deep-learning prognostics [32], alarm-based scheduling with imperfect RUL [33], maintenance optimisation with stochastic RUL in manufacturing systems [90], rules that account for the credibility of the prediction [91], multi-component planning [92], reinforcement learning on probabilistic RUL [93], rules based on a critical probabilistic cost [94] and frameworks that fuse several information sources under uncertainty [95]. For cutting tools, Chung et al. [96] predicted wear early in turning and regulated the cutting parameters accordingly. Decision-level validation of tool twins, with an explicit cost model and a comparison against the schedules a shop could run without a twin, remains rare. Table 1 summarises the position of the present work.

3 Hybrid digital twin of the cutting tool

3.1 Architecture and life cycle

Fig. 1 shows the architecture of the twin. Its four layers follow the domains of the ISO 23247 reference model [10]: the physical entity (the observable manufacturing element), the data and connectivity layer, the virtual entity or twin core, and the services offered to the user. The physical entity is the face-milling cutter on the machine with its sensors, here the AC and DC spindle-motor currents and the vibration and AE signals measured on the table and on the spindle. The context of the operation, $\boldsymbol{u} = (f, a_p, m)$ with feed f, axial depth of cut a_p and work material m, comes from the part programme. Flank wear V_B is measured only occasionally, when the tool is inspected.

The data layer acquires the signals of each run at the edge, validates them, extracts features from the steady part of the cut and stores the twin data, i.e. the history of the current tool and the trajectories of the tools used before it. The twin core holds four elements: the wear law, the probabilistic virtual sensor, the Bayesian synchronisation engine and the population knowledge base from which the prior of a new tool is drawn. The services are wear-state monitoring with credible bands, a probabilistic RUL, a replacement decision and a what-if advisory on feed and depth of cut. The feedback path returns a replacement order or modified cutting parameters to the machine.

A twin is instantiated when a new tool is mounted, from the population prior for its cutting context. It is updated after every run and retired when the tool is replaced; its trajectory then enters the knowledge base, so that the prior of the next tool reflects one more tool. In the validation reported below, the fleet of each fold consists of its training tools.

Table 1 Families of related work, how each uses evidence from the individual tool or asset, and the gap addressed in this study

Family	Representative studies	Use of evidence from the individual asset	Uncertainty and decisions	Addressed here
Equipment-level predictive-maintenance twins	[11, 12, 16, 38, 39, 44]	Model updating or retraining at discrete instants	Uncertainty often not quantified; maintenance planning mostly qualitative	Sequential Bayesian update after every cut
Tool digital twins	[19–26]	Simulation or learned models fed with current signals	Monitoring and alarms; little probabilistic output	Explicit twin state with posterior, RUL and decision layer
Data-driven TCM	[54–59]	Only through the input window of the regressor	Intervals only for probabilistic regressors; no decisions	Memory of the tool history; calibrated intervals
Physics-informed learning	[70, 71, 73, 75–77]	Physics constrains training; tool history mostly not carried forward	Some uncertainty quantification; no decisions	Physics used as transition model and corrected on line
Filtering and state-space prognostics	[84–88]	Recursive estimation of the state and sometimes of parameters	Probabilistic RUL; decisions rarely evaluated	Condition-dependent fleet prior; sensor-bias state; tool-level validation; decision evaluation
Prognostics-driven maintenance	[32, 33, 92, 93, 96]	Through RUL distributions supplied by a prognostic model	Policies built on RUL distributions	Policies applied to a tool twin and tuned without the test tool

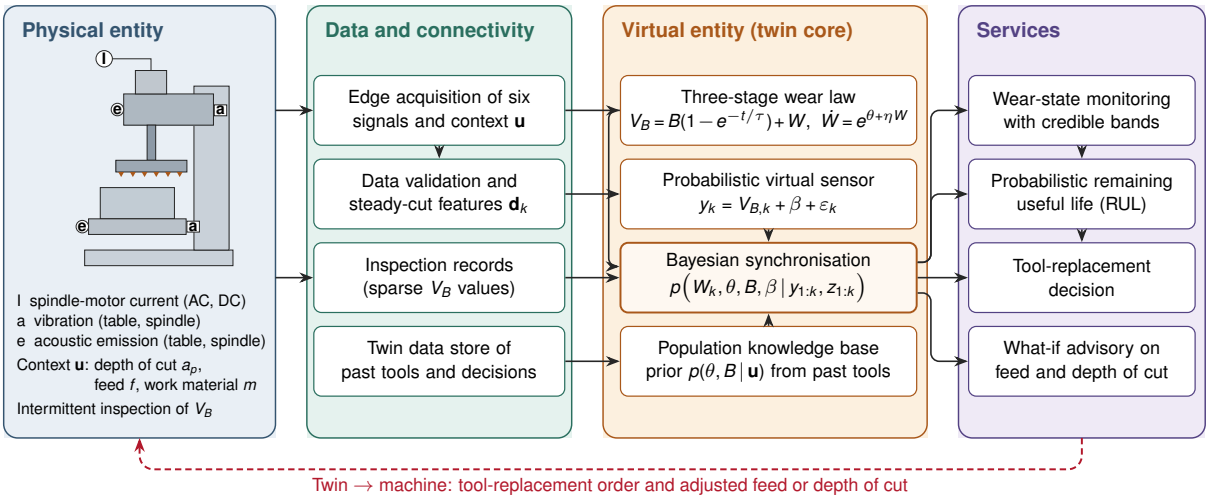


Fig. 1 Architecture of the hybrid digital twin of a face-milling tool, organised along the domains of ISO 23247: physical entity with its sensors and cutting context $\boldsymbol{u}$, data and connectivity layer, virtual entity (twin core) and services. The dashed arrow is the feedback from the twin to the machine

3.2 Probabilistic structure

The twin core is a dynamic Bayesian network (Fig. 2a). At run k, which ends at the cumulative cutting time t_k, the hidden state contains the accumulated progressive wear W_k, the log wear-rate parameter θ_k, the break-in amplitude B and the persistent sensor bias β. The flank wear $V_{B,k}$ is a deterministic function of the state. The virtual-sensor reading y_k is available after every run, whereas an inspection value z_k exists only when the tool is measured. The cutting context $\boldsymbol{u}$ and the population parameters $\psi = (\boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ of Sect. 3.5, learnt from past tools, condition the prior of θ and B. Each run triggers the six-step cycle of Fig. 2b: acquisition, validation and feature extraction, virtual-sensor reading, propagation of the particles through the wear law, weighting by the likelihoods of y_k and z_k, and computation of the posterior, the RUL and the decision.

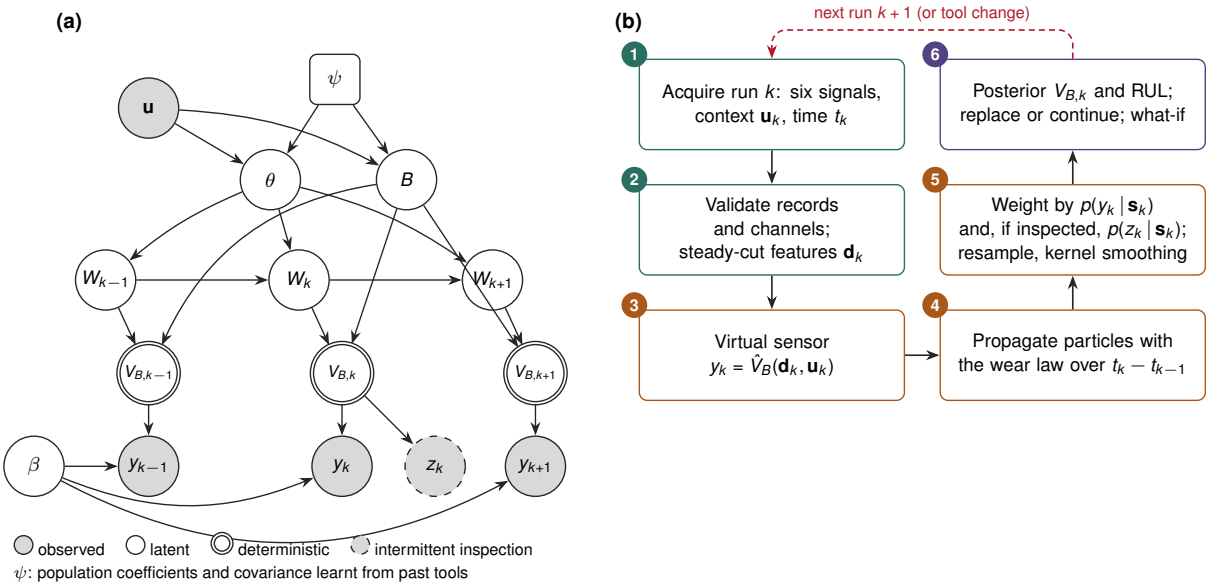


Fig. 2 Probabilistic structure of the twin. (a) Dynamic Bayesian network over three consecutive runs; shaded nodes are observed, double circles are deterministic functions of their parents and the dashed node is an intermittent inspection. (b) Update cycle executed by the twin after each run

3.3 Data layer: validation and steady-cut features

Each run yields six signals. The steady-cut segment is located on the AC spindle-current signal. After removal of the median, a moving root-mean-square (RMS) envelope over 250 samples is computed, and samples whose envelope exceeds the 5th percentile by more than half of the 5th-to-95th percentile range are marked as active. The longest active block is kept and 10% of its length is trimmed at each end to discard the entry and exit transients (Fig. 3c); if this block covers less than 15% of the record, the central 60% of the record is used instead. Six descriptors are computed for each channel over the segment: mean, RMS, standard deviation, peak-to-peak value, kurtosis and spectral centroid. The centroid is obtained from a Welch periodogram (Hann window of 512 samples, 50% overlap) and expressed in normalised frequency, which avoids any assumption on the sampling rate. A run is thus described by a vector $\boldsymbol{d}_k$ of 36 features.

Validation rules are applied before the features reach the twin and use only the past of the current tool. A record is rejected when the RMS of the AC current exceeds 10, against nominal values between about 0.9 and 4, or when any channel is flat (RMS below 10^{-4}). A dropout is declared for a vibration or AE channel when its RMS falls below one quarter of the median RMS of the previous valid runs of the same tool (or below 0.05 at the first run); the features of that channel are then replaced by its last valid values.

3.4 Three-stage wear law

The flank-wear curve of a carbide insert has three stages: a rapid break-in of the fresh edge, a long stage of nearly steady wear and a final stage in which wear accelerates [66, 67]. The twin represents them with a break-in term and a self-accelerating progressive term,

$$V_B(t) = B\left(1 - e^{-t/\tau}\right) + W(t), \qquad (1)$$

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \exp\left(\theta + \eta W\right), \qquad W(0) = 0, \qquad (2)$$

where B (mm) is the break-in wear, τ (min) its time constant, e^{θ} (mm min^{-1}) the progressive wear rate of an edge without progressive wear, and η (mm^{-1}) the growth of that rate as the wear

land widens. A wider land means a longer contact between flank and workpiece, more frictional heat and therefore faster abrasive and diffusive wear. Because temperature enters the classical wear-rate laws exponentially [66], an exponential dependence on W is a compact phenomenological surrogate for this feedback. For $\eta > 0$ the solution of Eq. (2) diverges in finite time, which reproduces the final stage.

Equation (2) integrates in closed form. Over a run of length Δ,

$$W(t+\Delta) = -\frac{1}{\eta}\ln\left[e^{-\eta W(t)} - \eta\, e^{\theta}\Delta\right], \qquad (3)$$

as long as the argument of the logarithm is positive; otherwise the edge is considered collapsed and W is capped. The time needed for W to grow from W to a level $W^* \geq W$ is

$$T(W, W^*) = \frac{e^{-\eta W} - e^{-\eta W^*}}{\eta\, e^{\theta}}. \qquad (4)$$

The law is deliberately parsimonious: four parameters, of which η and τ are shared by all tools, and closed-form expressions for both the propagation over a run and the time to reach a wear level. These two properties keep the particle filter cheap and make the RUL computation exact for each particle.

3.5 Population prior

Tools that cut under the same conditions do not wear at the same rate, but their parameters are not unrelated either. The twin encodes this with a hierarchical prior on $\boldsymbol{\varphi}_i = [\theta_i, B_i]^\top$:

$$\begin{aligned} \boldsymbol{\varphi}_i &= \boldsymbol{\Gamma}^\top \boldsymbol{x}(\boldsymbol{u}_i) + \boldsymbol{\zeta}_i, \qquad \boldsymbol{\zeta}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \\ \boldsymbol{x}(\boldsymbol{u}) &= \left[1,\ \ln f,\ \ln a_p,\ \mathbb{I}_{\text{steel}}\right]^\top. \end{aligned} \qquad (5)$$

With the logarithms, the wear rate is a power law of feed and depth of cut, similar in form to an extended Taylor tool-life equation, and the indicator shifts it for steel. The 4×2 matrix $\boldsymbol{\Gamma}$ and the covariance $\boldsymbol{\Sigma}$ are learnt from past tools in two steps. First, η and τ are set to the values that minimise the total squared misfit of Eq. (1) over all past tools, with $\tau \in \{0.5, 1, 2, 3, 4.5\}$ min and 15 values of η in $[-2, 5]\,\text{mm}^{-1}$, followed by a bounded refinement of η. Each tool contributes its labelled runs up to $V_B = 0.6$ mm; beyond this level, edge chipping makes the wear record erratic. Second, the least-squares estimates $\hat{\boldsymbol{\varphi}}_i$ are regressed on $\boldsymbol{x}(\boldsymbol{u}_i)$ with a small ridge term (10^{-3}, intercept not penalised) that keeps the regression stable with few tools, and $\hat{\boldsymbol{\Sigma}}$ is the residual covariance with $n-4$ degrees of freedom. A new tool with context $\boldsymbol{u}$ then receives the predictive prior

$$\begin{aligned} p(\boldsymbol{\varphi} \mid \boldsymbol{u}) &= \mathcal{N}\left(\hat{\boldsymbol{\Gamma}}^\top \boldsymbol{x},\ c(\boldsymbol{x})\,\hat{\boldsymbol{\Sigma}}\right), \\ c(\boldsymbol{x}) &= 1 + \boldsymbol{x}^\top \left(\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{\Lambda}\right)^{-1} \boldsymbol{x}, \end{aligned} \qquad (6)$$

where $\boldsymbol{X}$ stacks the context vectors of the past tools and $\boldsymbol{\Lambda}$ is the ridge matrix. The factor $c(\boldsymbol{x})$ widens the prior for contexts that the fleet covers poorly.

3.6 Probabilistic virtual sensor and error model

The virtual sensor converts the feature vector $\boldsymbol{d}_k$ of a run into a wear reading y_k. A Bayesian ridge regression [97] is fitted on standardised inputs made of three signal features and the three context terms $\ln f$, $\ln a_p$ and $\mathbb{I}_{\text{steel}}$, using the labelled runs of past tools with $V_B \leq 0.6$ mm. The three features are selected inside each fold: the 36 candidates are ranked by the magnitude of their Spearman correlation with V_B, and a candidate is accepted only if its Pearson correlation with every feature already accepted is below 0.95. A preliminary leave-one-tool-out screening compared Bayesian ridge regression with Gaussian process regression (isotropic and automatic-relevance kernels) and gradient boosting, each with three, five and eight features. The linear Bayesian model with three features gave the lowest error (RMSE 75.6 µm against at least 89.2 µm for the other combinations), and features normalised by the initial signature of each tool did not help. Because this screening saw all tools, the choice of model family is not entirely independent of the data; the feature subset itself is always selected without the test tool.

A reading is not the truth, and its error is not white. Inserts from different batches, the seating of the cutter and the local hardness of the stock shift the relation between signal and wear for a whole tool, so that the errors are correlated over the life of a tool. If this is ignored, successive readings

look like independent evidence and the posterior becomes overconfident. The twin therefore uses

$$y_k = V_{B,k} + \beta + \varepsilon_k, \qquad \beta \sim \mathcal{N}(b_0, s_\beta^2), \qquad \varepsilon_k \sim t_\nu(0, s_\varepsilon), \tag{7}$$

with β constant for a tool and ε_k Student-t distributed with $\nu = 5$ degrees of freedom, which prevents an occasional gross error from dragging the state. The parameters are estimated by the method of moments from out-of-sample residuals $r_{il} = y_{il} - V_{B,il}$ of the training tools, each produced by a virtual sensor that did not see tool i. With $\bar{r}_i$, S_i^2 and n_i the mean, variance and number of residuals of tool i, and n_T the number of tools,

$$b_0 = \frac{1}{n_T}\sum_i \bar{r}_i, \qquad s_\varepsilon^2 = \frac{1}{n_T}\sum_i S_i^2, \qquad s_\beta^2 = \max\left\{ v_{\bar{r}} - \frac{s_\varepsilon^2}{n_T}\sum_i \frac{1}{n_i},\ \frac{s_\varepsilon^2}{4} \right\}, \tag{8}$$

where $v_{\bar{r}}$ is the sample variance of the tool means. The lower bound keeps the bias state active when the spread between tools is masked by sampling noise, which is likely with about a dozen training tools.

3.7 Bayesian synchronisation

The state of the twin after run k is $\boldsymbol{s}_k = [W_k, \theta_k, B, \beta]$. Between runs the progressive wear follows Eq. (3) with a small diffusion, and the wear-rate parameter performs a slow random walk that lets it follow changes the context does not explain, for example a harder batch of stock:

$$W_k = \Phi_{\Delta_k}(W_{k-1}; \theta_{k-1}) + q_W\sqrt{\Delta_k}\,\xi_{1,k}, \qquad \theta_k = \theta_{k-1} + q_\theta\sqrt{\Delta_k}\,\xi_{2,k}, \tag{9}$$

where Φ_Δ denotes the right-hand side of Eq. (3), $\Delta_k = t_k - t_{k-1}$ and $\xi \sim \mathcal{N}(0,1)$; B and β are static. The reading and, when available, an inspection enter through the likelihoods

$$p(y_k \mid \boldsymbol{s}_k) = t_\nu\left(y_k;\, V_{B,k} + \beta,\, s_\varepsilon\right), \qquad p(z_k \mid \boldsymbol{s}_k) = \mathcal{N}\left(z_k;\, V_{B,k},\, \sigma_I^2\right), \tag{10}$$

with $V_{B,k} = B(1 - e^{-t_k/\tau}) + W_k$.

Algorithm 1 Synchronisation of the tool twin after run k

Require: N weighted particles; reading y_k; optional inspection z_k

$\Delta_k \leftarrow t_k - t_{k-1}$
for $j = 1, \dots, N$ **do**
  $W^{(j)} \leftarrow \Phi_{\Delta_k}(W^{(j)}; \theta^{(j)}) + q_W\sqrt{\Delta_k}\,\xi$
  $\theta^{(j)} \leftarrow \theta^{(j)} + q_\theta\sqrt{\Delta_k}\,\xi$
  $V^{(j)} \leftarrow B^{(j)}\left(1 - e^{-t_k/\tau}\right) + W^{(j)}$
  $w^{(j)} \leftarrow w^{(j)}\, t_\nu\left(y_k; V^{(j)} + \beta^{(j)}, s_\varepsilon\right)$
  if z_k available **then**
    $w^{(j)} \leftarrow w^{(j)}\,\mathcal{N}\left(z_k; V^{(j)}, \sigma_I^2\right)$
  end if
end for
Normalise the weights $w^{(j)}$
if $1/\sum_j \left(w^{(j)}\right)^2 < N/2$ **then**
  Resample systematically; move $\boldsymbol{\vartheta}^{(j)}$ by Eq. (11)
end if
$R^{(j)} \leftarrow T\left(W^{(j)}, V_{\mathrm{th}} - B^{(j)}\right)$, or 0 if $V^{(j)} \geq V_{\mathrm{th}}$
return weighted samples of $V_{B,k}$ and R_k

The posterior $p(\boldsymbol{s}_k \mid y_{1:k}, z_{1:k})$ is approximated by N weighted particles (Algorithm 1). At mounting, θ and B are drawn from the prior of Eq. (6), β from $\mathcal{N}(b_0, s_\beta^2)$, and $W_0 = 0$. After each run the particles are propagated with Eq. (9) and reweighted with Eq. (10). When the effective sample size drops below $N/2$, the particles are resampled systematically and the static or slowly varying components $\boldsymbol{\vartheta} = [\theta, B, \beta]$ are moved with the kernel-smoothing step of Liu and West [98]:

$$\boldsymbol{\vartheta}^{(j)} \leftarrow a\,\boldsymbol{\vartheta}^{(j)} + (1-a)\,\bar{\boldsymbol{\vartheta}} + \boldsymbol{h} \odot \boldsymbol{\xi}^{(j)}, \qquad \boldsymbol{h} = \max\left\{ \sqrt{1-a^2}\,\mathrm{sd}(\boldsymbol{\vartheta}),\ \boldsymbol{h}_{\min} \right\}, \tag{11}$$

with $a = 0.98$ and a floor $\boldsymbol{h}_{\min} = [0.02, 0.005, 0.005]$ (θ in log units, B and β in mm) that stops the parameter cloud from collapsing when the fleet is small. A safeguard re-anchors the 20% least likely particles on the reading when every particle lies more than five scale units away from it; it was never triggered in this study. The process-noise levels are selected inside each fold, q_W from $\{0.002, 0.005, 0.01\}\,\mathrm{mm\,min^{-1/2}}$ and q_θ from $\{0, 0.01, 0.03\}\,\mathrm{min^{-1/2}}$, by minimising the mean continuous ranked probability score (CRPS) of the inner-fold tracks.

3.8 Remaining useful life and replacement rules

The tool is considered worn out when V_B reaches the threshold V_{th}. For each particle, the RUL follows from Eq. (4) as

$$R_k^{(j)} = T\left(W_k^{(j)},\; V_{\mathrm{th}} - B^{(j)}\right), \tag{12}$$

and is set to zero if $V_{B,k}^{(j)} \geq V_{\mathrm{th}}$. Equation (12) takes the break-in as complete; while t_k is still comparable with τ the remaining break-in increment is ignored, which makes the estimate slightly conservative.

Decisions are taken at the end of each run, before the next run of length Δ_{k+1} starts, and the tool is either kept or replaced. The rules compared in this study replace the tool at t_k when

$$\begin{aligned}
&\text{expected RUL:} && \mathbb{E}[R_k] \leq \kappa\,\Delta_{k+1}, && (13)\\
&\text{risk quantile:} && \mathbb{P}(R_k \leq \Delta_{k+1}) \geq \alpha, && (14)\\
&\text{schedule:} && t_k + \Delta_{k+1} > Q_{\alpha_0}(T_f \mid \boldsymbol{u}), && (15)\\
&\text{fixed age:} && t_k + \Delta_{k+1} > T_A, && (16)
\end{aligned}$$

The first two rules use the twin; in the model-based schedule of Eq. (15), $Q_{\alpha_0}(T_f \mid \boldsymbol{u})$ is the α_0-quantile of the time to failure predicted by the population prior alone, without any sensor data. The model-based schedule therefore represents what a shop can do with condition-dependent tool-life knowledge but no monitoring, and the fixed age a single tool-life counter for all tools. Run to failure and a clairvoyant policy that knows the failure time t_f bound the comparison from both sides; the clairvoyant policy replaces the tool at the last decision epoch before t_f.

Each tool is one renewal cycle. A preventive replacement costs c_p, a failure costs $c_f = R\,c_p$, and the cycle length L_i is the replacement or failure time. The cost rate of a policy over a set of tools is

$$C = \frac{\sum_i c_i}{\sum_i L_i}, \qquad c_i \in \{c_p,\; R\,c_p\}, \tag{17}$$

reported per hour of cutting with $c_p = 1$ (cost units per hour). The parameters κ, α, α_0 and T_A are tuned by minimising C over the inner-fold tracks of the training tools (Sect. 4.3).

3.9 What-if advisory

From the current posterior, the twin can answer a question such as "how much life is left if the feed is halved from now on?". The population model of Eq. (5) gives the sensitivity of the log wear rate to the cutting conditions through the θ-column of $\hat{\boldsymbol{\Gamma}}$, $\boldsymbol{\gamma} = [\gamma_0, \gamma_f, \gamma_a, \gamma_s]^\top$. For a change from (f, a_p) to (f', a_p'), each particle keeps its wear state, break-in and bias, and its rate parameter is shifted:

$$\theta' = \theta + \gamma_f \ln\frac{f'}{f} + \gamma_a \ln\frac{a_p'}{a_p}. \tag{18}$$

The RUL distribution under the new conditions is then obtained from Eq. (12). The advisory uses the point estimates of γ_f and γ_a and is meant for conditions inside the range covered by the fleet.

4 Case study and evaluation protocol

4.1 Milling data set

The twin was evaluated on the milling data set of the NASA Ames Prognostics Data Repository, recorded at the University of California, Berkeley [99, 100]. Face milling was performed on a Matsuura MC-510V machining centre with a 70 mm face mill carrying six KC710 inserts coated with TiC/TiCN/TiN, at a cutting speed of 200 m/min. Workpieces of cast iron and J45 steel ($483 \times 178 \times 51$ mm) were machined at two depths of cut (0.75 and 1.5 mm) and two feeds (0.25 and 0.5 mm/rev, i.e. 206.5 and 413 mm/min), a full factorial design for each material. Each record contains the six signals listed in Sect. 3.1, with 9000 samples per channel (15 360 in one record), and the cumulative cutting time at which it was taken. Records are 1 to 12 min of cutting apart (median 4 min); in what follows, a run is the cutting between two records, and t_k is the time of the record that closes run k. Flank wear was measured off line between runs, but not after every run.

The data set contains 16 cases, i.e. tools, and 167 records. Case 6 has a single record and was excluded; tool Tn below corresponds to case n. The validation layer rejected one record (case 12, run 1, whose AC-current RMS is 28.9) and replaced the table AE features of two runs with dropouts (case 7, run 4, and case 8, run 3). The DC

current saturates at 9.995, the upper limit of its recorded range, in 22 records; these records were kept. The resulting set has 15 tools, 165 runs and 145 wear measurements (Table 2). Every cutting condition is represented by two tools, except steel at $f = 0.25\,\mathrm{mm/rev}$ and $a_p = 1.5\,\mathrm{mm}$, which has one. Fig. 3 shows the wear curves and the signals of one tool.

4.2 Failure threshold and scoring sets

The replacement threshold is $V_{\mathrm{th}} = 0.45\,\mathrm{mm}$. ISO 8688-1 bases tool-life testing in face milling on flank-wear criteria [101]; the value used here lies within the range of flank-wear limits commonly applied to carbide inserts and was chosen so that every tool crosses it within its record. The failure time t_f is therefore observed, by linear interpolation between the two measurements that bracket the crossing, and not extrapolated. Wear estimates are scored at every measured run after the first one, up to and including the first run at which the measured V_B reaches V_{th} (108 points); later runs, cut after the tool would have been replaced, do not belong to its service life. RUL predictions are scored at every run with $t_k < t_f$ (107 points), the true RUL being $t_f - t_k$.

4.3 Nested leave-one-tool-out protocol

Each of the 15 tools serves in turn as the test tool, and the other 14 form the fleet. For each such outer fold, the following steps are carried out without the test tool:

1. the virtual-sensor features are selected (Sect. 3.6);
2. an inner leave-one-out loop runs over the 14 training tools; for each of them, the virtual sensor and the population prior are fitted on the other 13 tools, and its out-of-sample sensor residuals and twin track are computed;
3. the sensor-error parameters are estimated from the inner residuals with Eq. (8), (q_W, q_θ) is chosen by the inner CRPS, and the parameters of the replacement rules (13)–(16) are tuned on the inner tracks for each cost ratio;
4. the virtual sensor and the population prior are refitted on the 14 training tools, and the twin is run on the test tool.

The measured wear of the test tool is used only for scoring, and, in the inspection variant of Sect. 4.4, for the inspected runs. One outer fold takes about 11 s.

4.4 Baselines and ablations

Five probabilistic variants share the data and the training of the twin:

- the population prior alone, propagated in open loop with Eqs. (1)–(2) without sensor data;
- the virtual sensor alone, i.e. the bias-corrected reading $y_k - b_0$ with a Gaussian spread of standard deviation $(s_\beta^2 + s_\varepsilon^2)^{1/2}$;
- the twin without the bias state, in which the reading error is white with scale $(s_\beta^2 + s_\varepsilon^2)^{1/2}$;
- the twin with frozen parameters, in which θ and B stay at their prior means and only W is filtered;
- the twin with sparse inspections, which receives the measured V_B of every third measured run within the service life through Eq. (10), with σ_I equal to the residual standard deviation of the wear-curve fits of the training tools (about $0.024\,\mathrm{mm}$), and is scored on the 67 runs that were not inspected.

Four regressors from scikit-learn [97] serve as data-driven baselines: a random forest (RF; 500 trees, minimum leaf size 2), gradient boosting (GBR; 300 trees of depth 2, learning rate 0.05, subsampling 0.8), support vector regression (SVR; radial kernel, $C = 3$, $\epsilon = 0.02\,\mathrm{mm}$) and a multilayer perceptron (MLP; hidden layers of 32 and 16 units, L2 penalty 0.01). They use eight features selected in the same way, the same context terms and the same training runs, and they predict each run independently. For RUL, a trend baseline fits a straight line to the last four points of the bias-corrected reading series (including the origin) and extrapolates it to V_{th}; its predictions are capped at three times the true life, which can only favour it.

4.5 Metrics

Wear estimation is scored with the RMSE, the mean absolute error (MAE) and the coefficient of determination R^2 of the posterior mean, with the prediction-interval coverage probability (PICP) and the mean width (MPIW) of the central 95% interval, and with the CRPS computed from the

Table 2 Tools of the NASA milling data set used in this study. CI cast iron, St steel; t_f time at which V_B reaches 0.45 mm (linear interpolation between measurements)

Tool	Material	f (mm/rev)	a_p (mm)	Runs	Wear measurements	Wear scoring points	RUL scoring points	t_f (min)	Largest V_B (mm)
T1	CI	0.50	1.50	17	13	10	12	41.0	0.50
T2	CI	0.50	0.75	14	13	11	11	61.8	0.55
T3	CI	0.25	0.75	14	14	13	12	75.5	0.55
T4	CI	0.25	1.50	7	7	6	5	36.8	0.49
T5	St	0.50	1.50	6	6	4	3	9.3	0.74
T7	St	0.25	0.75	8	7	6	5	18.7	0.46
T8	St	0.50	0.75	6	5	4	4	9.2	0.62
T9	CI	0.50	1.50	9	9	6	5	32.7	0.81
T10	CI	0.25	1.50	10	10	7	6	36.0	0.70
T11	CI	0.25	0.75	23	20	15	17	76.8	0.76
T12	CI	0.50	0.75	14	12	8	9	55.0	0.65
T13	St	0.25	0.75	15	13	6	6	20.9	1.53
T14	St	0.50	0.75	9	7	5	5	14.4	1.14
T15	St	0.25	1.50	7	6	4	3	11.9	0.70
T16	St	0.50	1.50	6	3	3	4	6.7	0.62
Total				165	145	108	107		

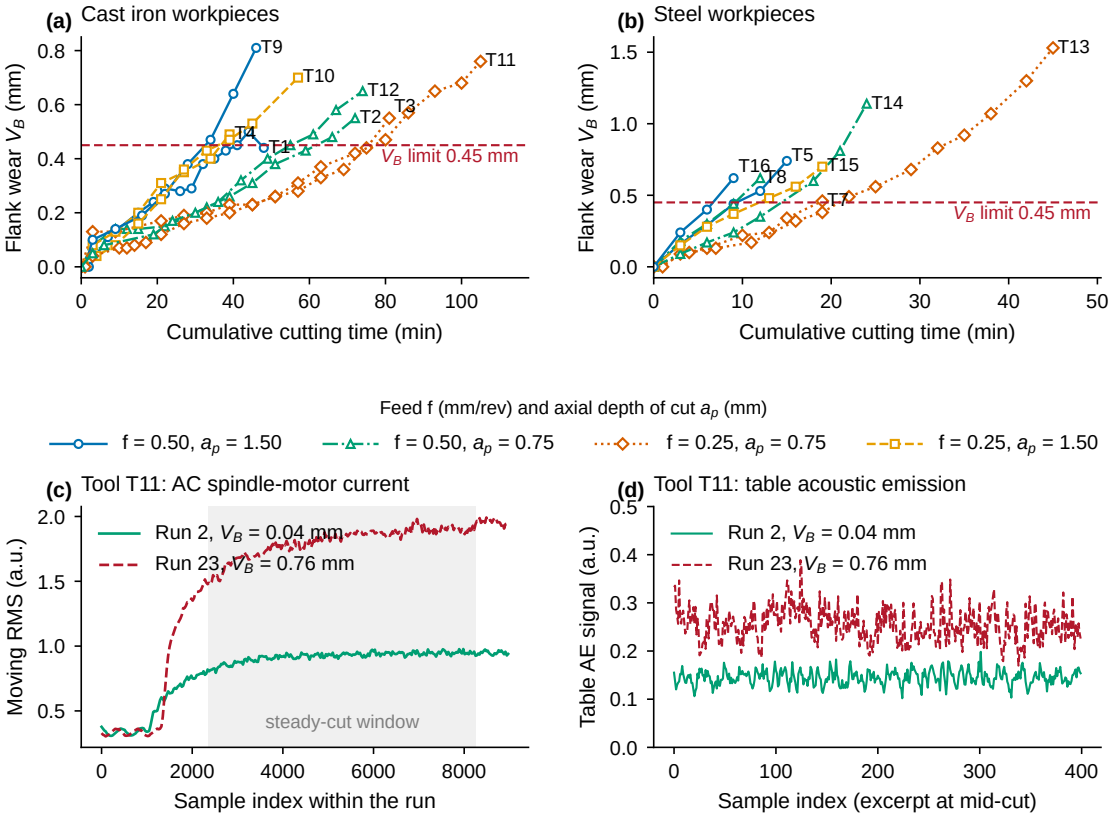


Fig. 3 Milling data set. (a), (b) Measured flank wear of the 15 tools on cast iron and steel; marker and line style identify the feed and depth of cut, and the dashed horizontal line is the 0.45 mm threshold. (c) Moving RMS of the AC spindle current of tool T11 early and late in its life, with the steady-cut window used for feature extraction (grey). (d) Excerpt of the table AE signal of the same runs

samples of each predictive distribution [29]. The CRPS rewards sharpness and calibration together and reduces to the absolute error for a point forecast. Coverage is also examined at nominal levels between 10% and 95%. Because runs of the same

tool are not independent, methods are compared through their per-tool RMSE with a one-sided Wilcoxon signed-rank test over the 15 tools.

For RUL, the median of the predictive distribution is the point estimate. The RMSE, the MAE, the median relative error $|\hat{R}-R|/R$ and the α-accuracy, i.e. the fraction of predictions within $\pm\alpha R$ of the truth with $\alpha = 0.3$, in the spirit of the α–λ metrics of Saxena et al. [102], are reported for all runs and for the second half of the tool life ($t_k/t_f \geq 0.5$, 57 points), together with the coverage of the central 90% interval and the CRPS.

Decisions are scored by the cost rate of Eq. (17) for $R \in \{2, 3, 5, 10, 20\}$, with $R = 5$ as the reference case, by the number of failures and by the fraction of the total available tool life that was used. Confidence intervals come from a paired bootstrap over tools with 10 000 resamples, in which all policies are evaluated on the same resampled set of tools.

4.6 Implementation

Table 3 lists the settings. The twin and the baselines were implemented in Python 3.13 with NumPy 2.5, SciPy 1.18 and scikit-learn 1.9 [97]; the data figures were drawn with Matplotlib 3.11 and the diagrams with TikZ. All computations ran on two cores of an Intel Xeon processor without a graphics processor. The code, the extracted features and the per-run results are available from the author on reasonable request.

Use of generative artificial intelligence. The large language model Claude (Anthropic) was used, under the author's direction, to help write and debug the analysis code, to retrieve bibliographic metadata from Crossref and to draft and edit parts of the text. The author designed the study, checked the equations, results and references against the code outputs and the original sources, and takes full responsibility for the content of the article.

Table 3 Settings of the twin and of the evaluation, and ranges of the values learnt in the 15 outer folds

Item	Value
Fixed settings	
Replacement threshold V_{th}	0.45 mm
Training range of prior and sensor	$V_B \leq 0.6$ mm
Candidate features	36 (6 channels × 6 descriptors)
Selected features	3 for the twin, 8 for the regressors; $\lvert\rho_{\mathrm{Pearson}}\rvert < 0.95$
Grid for τ and η	$\{0.5, 1, 2, 3, 4.5\}$ min; 15 values in $[-2, 5]\,\mathrm{mm}^{-1}$, then refined
Reading noise	Student-t, $\nu = 5$; $s_\beta \geq s_\varepsilon/2$
Particles	$N = 2000$ (800–1000 in inner folds)
Resampling	systematic, when ESS $< N/2$
Kernel smoothing	$a = 0.98$, $\boldsymbol{h}_{\min} = [0.02, 0.005, 0.005]$
Grid for q_W, q_θ	$\{0.002, 0.005, 0.01\}$ mm min$^{-1/2}$; $\{0, 0.01, 0.03\}$ min$^{-1/2}$
Inspections	every third measured run; $\sigma_I \approx 0.024$ mm
Costs	$c_p = 1$, $R = c_f/c_p \in \{2, 3, 5, 10, 20\}$
Rule grids	$\kappa \in \{0.25, 0.5, \ldots, 8\}$; $\alpha, \alpha_0 \in \{0.01, \ldots, 0.80\}$; $T_A \in [2, 80]$ min
Bootstrap	10 000 paired resamples of tools
Learnt in the outer folds (range)	
η; τ	1.49–2.08 mm^{-1}; 2, 3 or 4.5 min
b_0; s_ε; s_β	−0.007 to 0.002 mm; 0.076–0.091 mm; 0.038–0.046 mm
q_W; q_θ	0.002 (14 folds) or 0.005; 0 (4), 0.01 (3) or 0.03 (8)
Rules at $R = 5$	κ 2.25–3.0; α 0.02–0.19; α_0 0.03–0.23; T_A 11–63 min

5 Results

5.1 Population wear model and virtual sensor

Fig. 4 shows the three-stage law fitted to each tool together with the 90% band of the population prior for its context, and Table 4 gives the population coefficients estimated from all 15 tools. The shared parameters are $\eta = 1.90\,\mathrm{mm}^{-1}$ and $\tau = 4.5$ min, and the measured wear deviates from the fitted curves by 23.6 µm (RMS). The law follows both the long, nearly linear wear of the cast-iron tools at low feed (T3, T11) and the strongly accelerating wear of the steel tools (T13, T14). In the validation folds, η ranged from 1.49 to $2.08\,\mathrm{mm}^{-1}$ and τ took the values 2, 3 or 4.5 min. That τ often sits at the largest grid value reflects how weakly the break-in is resolved, with only one or two measurements per tool in that stage.

Work material dominates the wear rate: at equal feed and depth of cut, steel multiplies the progressive wear rate by $e^{1.197} = 3.3$ ($p < 0.001$).

Table 4 Population model of Eq. (5) estimated from the 15 tools (11 residual degrees of freedom); standard errors in parentheses

Term	θ	p	B (mm)	p
Intercept	−4.964 (0.173)	< 0.001	0.127 (0.061)	0.063
$\ln f$	0.273 (0.151)	0.097	0.064 (0.054)	0.26
$\ln a_p$	0.433 (0.151)	0.015	0.147 (0.054)	0.019
$\mathbb{I}_{\text{steel}}$	1.197 (0.104)	< 0.001	0.097 (0.037)	0.024
Residual s.d.	0.201		0.071	

Shared parameters $\eta = 1.90\,\text{mm}^{-1}$ and $\tau = 4.5\,\text{min}$; correlation of the residuals of θ and B: −0.43; RMS deviation of the measurements from the fitted curves: 23.6 µm

Doubling the depth of cut multiplies it by 1.35 ($p = 0.015$) and doubling the feed by 1.21, an effect that is not significant with 11 residual degrees of freedom ($p = 0.097$). The break-in amplitude grows with the depth of cut and is larger for steel. Tool-to-tool variation remains after conditioning on the context: the residual standard deviation of θ is 0.20, i.e. a typical factor of 1.22 on the wear rate between tools that cut under identical conditions. This is the variation that the twin has to learn on line. The residuals of θ and B are negatively correlated (−0.43), which means that tools with a large break-in tend to wear more slowly afterwards, or that the two terms partly trade off in the fit.

Fig. 5 summarises the virtual sensor. Across folds, the selection retained the RMS of the DC spindle current (14 of 15 folds), the RMS of the table AE signal (13 folds) and the spectral centroid of the AC current (11 folds). The readings follow the measured wear but scatter widely, and steel tools tend to be under-read at high wear (Fig. 5a). The error model of Eq. (8) gave s_ε between 0.076 and 0.091 mm and b_0 between −0.007 and 0.002 mm. The moment estimate of s_β^2 fell below the bound $s_\varepsilon^2/4$ in every fold, so that $s_\beta = s_\varepsilon/2$ (0.038–0.046 mm). The mean error of individual tools nevertheless ranges from −0.08 mm (T8) to +0.06 mm (T14) (Fig. 5c). With 3 to 15 measurements per tool, this spread cannot be separated statistically from the white part of the error, but, as shown in Sect. 5.2, the tools with the most negative mean error are those on which the twin performs worst.

5.2 Wear-state estimation

Fig. 6 shows the twin on four tools that span the range of conditions. At mounting, the twin starts from the population prior, whose 95% band is wide; after a few runs the posterior narrows around the measured wear and follows it through the threshold. For T2 the posterior mean stays within about 0.08 mm of the measurements over the whole record, although single readings deviate by up to 0.11 mm. For T11, the longest-lived tool, the twin tracks the steady stage closely but lags the final acceleration, and the last measurement (0.57 mm at 86 min) lies just above its 95% band. For T13 the readings increasingly under-estimate the wear in the second half of the life, and the twin, which follows them partly, ends 0.13 mm below the last measurement. T16, the most aggressive condition (steel, $f = 0.5\,\text{mm/rev}$, $a_p = 1.5\,\text{mm}$), shows the opposite situation: the first readings lie 0.1 to 0.2 mm above any plausible wear of the fresh edge. The prior and the heavy-tailed likelihood keep the state on a physically consistent path, and the bias state absorbs part of the offset (its posterior mean rises to 0.07 mm).

Table 5 and Fig. 7 summarise the 108 scoring points. The twin reached an RMSE of 63.3 µm and $R^2 = 0.80$. The virtual sensor alone had an RMSE of 73.7 µm and the population prior alone 101.7 µm; the twin was better than both on 12 of the 15 tools ($p = 0.013$ and $p = 0.002$). The two sources of information are thus complementary: the prior fixes the shape and level of the wear curve, and the readings correct it for the individual tool. The four regressors, trained on the same runs, ranged from 104.0 µm (GBR and SVR) to 131.6 µm (MLP), and the twin was better on 12 to 15 of the 15 tools ($p < 0.001$ in each case). Their errors are largely offsets between tools: run by run, the regressors react to the same feature scatter as the virtual sensor, but they have neither the context-dependent wear curve nor the memory of the twin.

Calibration separates the variants more clearly than accuracy. The twin covered 89.8% of the measurements with its nominal 95% intervals, stayed close to the diagonal of the coverage diagram at all levels (Fig. 7b) and had a CRPS of 34.2 µm, the lowest among the variants without inspections. Freezing θ and B at their prior means gave a moderate point error (74.0 µm) but

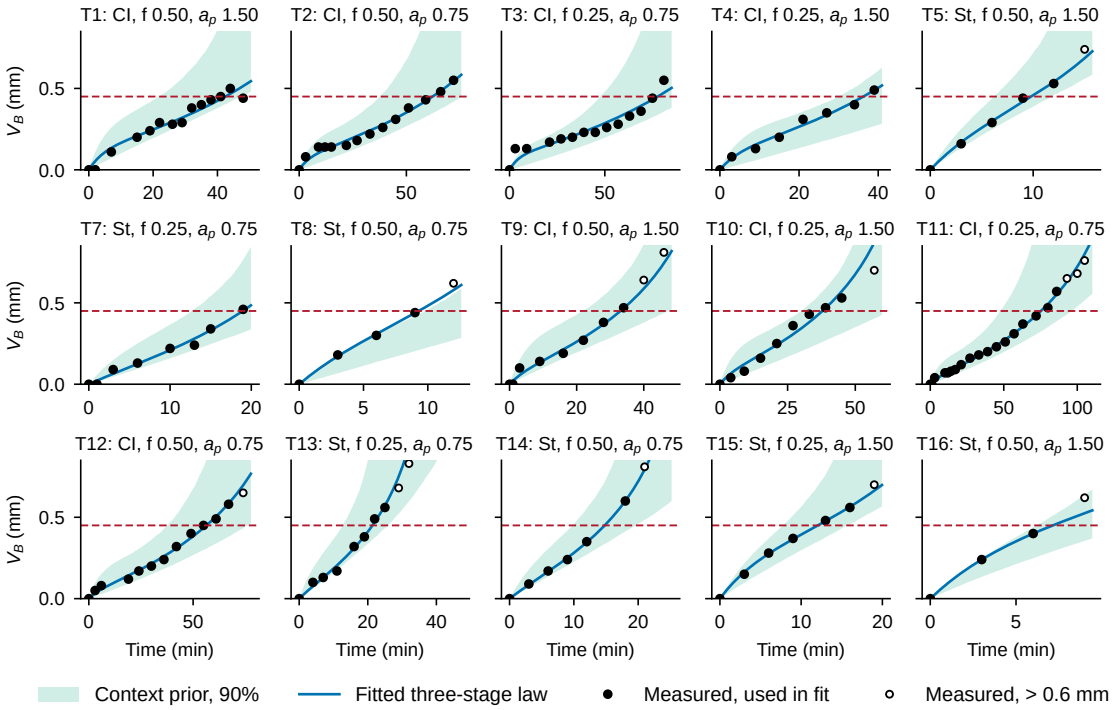


Fig. 4 Three-stage wear law fitted to each tool (solid line) and 90% band of the population prior for the context of the tool (shaded), estimated from all 15 tools for illustration. Filled circles are the measurements used in the fit and open circles those above 0.6 mm; the dashed line is the 0.45 mm threshold. In the validation, the prior of each test tool is learnt without that tool. CI cast iron, St steel; f in mm/rev, a_p in mm

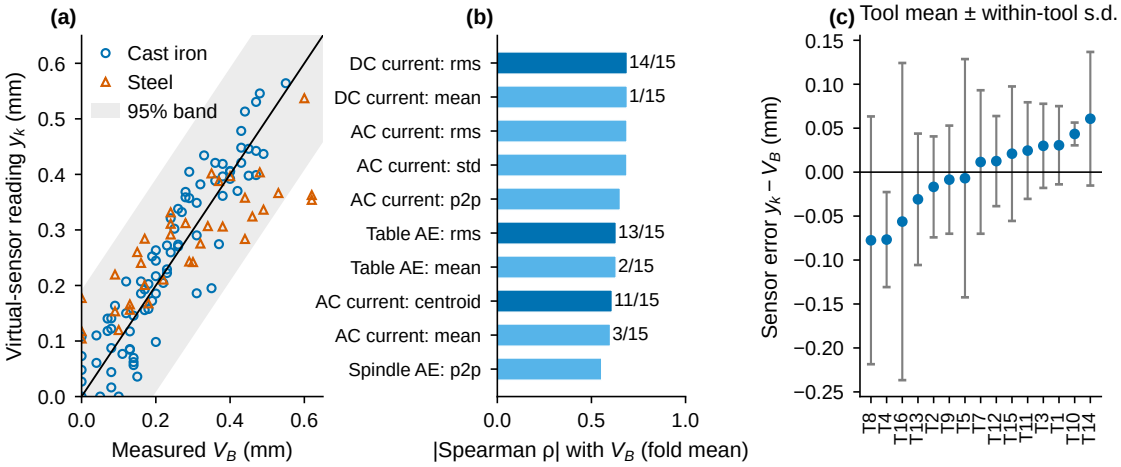


Fig. 5 Virtual sensor. (a) Out-of-sample readings against measured wear for cast-iron (circles) and steel (triangles) tools, with the band of ±1.96 total error standard deviations. (b) Features ranked by the fold-mean magnitude of their Spearman correlation with V_B; the labels give the number of folds in which a feature was selected, and darker bars mark the three features selected most often. (c) Mean error of each tool (dots) with ±1 within-tool standard deviation (bars)

a far too narrow posterior: only 32.4% of the measurements fell in its 95% intervals, and the CRPS rose to 49.9 µm. Updating the parameters is therefore what makes the uncertainty of the twin trustworthy; a twin that tracks only the state inherits the confidence of a prior that is wrong for the individual tool. Removing the bias state left the accuracy unchanged (63.0 µm, $p = 0.58$)

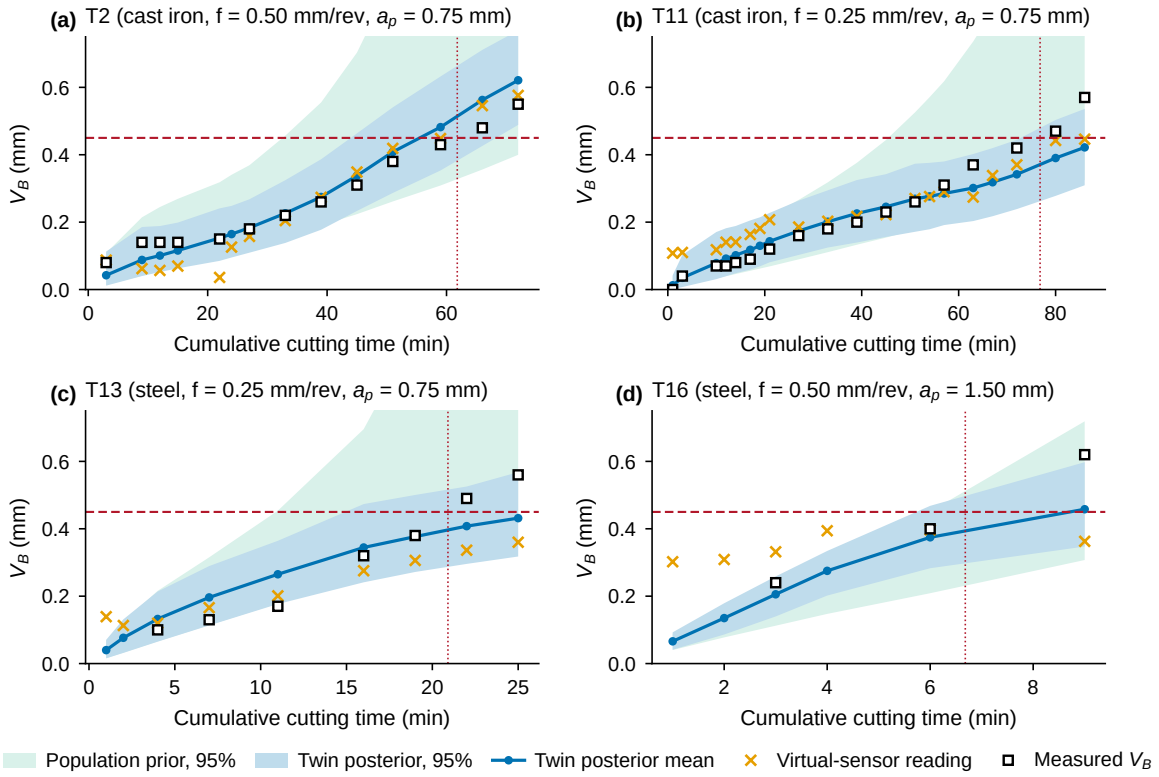


Fig. 6 Twin tracking of four test tools in leave-one-tool-out validation: 95% band of the population prior (light shading), 95% band and mean (line with dots) of the twin posterior, virtual-sensor readings (crosses) and measured wear (squares). The dashed horizontal line is the 0.45 mm threshold and the dotted vertical line the interpolated failure time

Table 5 Wear estimation on the 108 scoring points of the 15 test tools (nested leave-one-tool-out). Wins: number of tools on which the per-tool RMSE of the twin is lower; p: one-sided Wilcoxon signed-rank test on the per-tool RMSE

Method	RMSE (µm)	MAE (µm)	R^2	PICP_{95}	MPIW (µm)	CRPS (µm)	Median tool RMSE (µm)	Wins	p
Digital twin (proposed)	**63.3**	**47.9**	**0.798**	0.898	194.9	**34.2**	44.8	–	–
Twin without bias state	63.0	49.3	0.800	0.870	181.4	35.1	44.8	6/15	0.58
Twin with frozen parameters	74.0	55.8	0.724	0.324	53.4	49.9	69.1	10/15	0.032
Virtual sensor only	73.7	57.0	0.726	0.981	364.2	42.3	65.9	12/15	0.013
Population prior (open loop)	101.7	78.1	0.478	0.944	397.0	45.2	75.8	12/15	0.002
GBR	104.0	85.4	0.455	–	–	–	91.4	14/15	< 0.001
SVR	104.0	82.7	0.455	–	–	–	104.5	14/15	< 0.001
RF	116.2	95.1	0.319	–	–	–	115.7	15/15	< 0.001
MLP	131.6	100.9	0.127	–	–	–	145.4	12/15	< 0.001
Runs not inspected (67 points)									
Digital twin (proposed)	63.1	–	0.776	0.896	–	35.4	–	–	–
Twin with sparse inspections	45.6	34.8	0.883	0.836	123.4	25.5	53.0	2/15[a]	0.005[a]

but lowered the coverage to 87.0% and raised the CRPS to 35.1 µm. The virtual sensor alone is over-dispersed (98.1% coverage with an average interval of 364 µm), as expected from its total error spread.

Inspections of every third measured run gave the largest gain. On the 67 runs that were not inspected, the RMSE fell from 63.1 µm for the uninspected twin to 45.6 µm and the CRPS from 35.4 to 25.5 µm; the error decreased on 13 of the 15 tools (one-sided Wilcoxon test, $p = 0.005$). The intervals narrowed to 123 µm on average, but their coverage dropped to 83.6%, which suggests that $\sigma_I = 0.024$ mm, taken from the curve-fit residuals, understates the actual mismatch between an inspection and the model.

Errors are concentrated on a few tools (Fig. 7a). The median per-tool RMSE of the twin is 44.8 µm, and the largest values are those of T8 (176 µm), T16 (96 µm) and T4 (87 µm), the three tools with the most negative mean sensor error in Fig. 5c. For T8 the readings under-estimate the wear more and more as it grows (0.28 mm against 0.44 mm at 9 min and 0.35 against 0.62 mm at 12 min), whereas the first reading over-reads the fresh edge. The bias state, learnt from the early runs, settles at a small positive value and cannot represent an error that changes sign with the wear level.

5.3 Remaining useful life

Fig. 8a,b shows the predicted RUL of two tools. Early in life the prediction of the twin coincides with that of the prior, which under-estimates the life of both tools somewhat. As readings accumulate, the twin moves towards the true RUL, whereas the open-loop prior keeps its bias. The 90% intervals are wide, with a median width of 1.3 times the true RUL, which reflects the spread of wear acceleration between tools that the readings resolve only gradually.

Over the 107 scoring points (Table 6), the twin achieved an RMSE of 7.0 min, a median relative error of 23.7% and an α-accuracy of 0.60, against 30.0% and 0.50 for the prior alone, and 89.7% of the true values fell in its nominal 90% intervals. The predictions err on the safe side: 69% of them are shorter than the true RUL (median signed relative error -11.8%), against 82% (-20.1%) for the prior. The twin with frozen parameters has the lowest RMSE (6.6 min) but covers only 24.3% of the true values with its 90% intervals, and its CRPS (4.61 min) is the worst among the twin variants; point accuracy alone would have ranked this overconfident variant first. Extrapolating the trend of the readings fails (RMSE 45.2 min) because the noise of single readings dominates the slope over a short window.

Relative errors grow towards the end of life (Fig. 8c) because the true RUL in the denominator shrinks: the median relative error of the twin rises from 17% in the first quarter of the life to 24% in the third quarter and 76% in the last, where only a few minutes remain. Sparse inspections mainly help in the second half of the life. There, the α-accuracy rises from 0.39 to 0.44 and the RMSE falls from 6.4 to 5.4 min, and the predictions are no longer biased low (47% shorter than the truth, median signed error +5.1%), at the price of a lower interval coverage (80.4%).

5.4 Replacement decisions

Table 7 and Fig. 9 compare the policies. At $R = 5$, run to failure costs 8.88 per hour, and a single tool-life counter is even worse (9.66 per hour). An age limit that suits the slow-wearing cast-iron tools lets the steel tools fail, and a limit short enough for steel wastes most of the cast-iron life; the tuned compromise (33 min in most folds) serves neither material and lets 11 of the 15 tools fail. The model-based schedule, which uses the condition-dependent prior, brings the cost rate down to 3.46 per hour with one failure (T8). The twin with the expected-RUL rule reaches 2.38 per hour without any failure and uses 74.6% of the available life. This is 31% below the model-based schedule: the difference is 1.08 per hour, with a paired-bootstrap 95% interval of [0.15, 3.29], and only 0.25% of the resamples favour the schedule. Measured against the perfect-information bound (1.93 per hour), the twin closes 71% of the gap left by the schedule.

The risk-quantile rule was less effective than the expected-RUL rule (3.23 per hour, one failure). It acts on the lower tail of the RUL distribution, the part that is estimated least reliably from 13 inner-fold tools, and its tuned risk level varied between 0.02 and 0.19 from fold to fold. Adding inspections to this rule lowered its cost to 2.79 per hour and raised the life used to 80.7%, but one tool (T7) still failed, because the risk level was tuned on uninspected tracks whereas the inspected posterior is narrower.

Fig. 9b shows how these conclusions depend on the cost ratio. With cheap failures ($R = 2$), the schedule and the three twin-based policies lie between 2.40 and 2.74 per hour, and the schedule is not worse than the expected-RUL twin (difference -0.19 per hour, 95% interval [-0.60, 0.16]). As failures become more expensive, the cost of policies that still incur failures grows with R, whereas the expected-RUL twin, whose tuned margin κ avoided failures at every $R \geq 3$, stays between 2.38 and 2.42 per hour. At $R = 10$ and $R = 20$ the schedule costs 4.96 and 7.03 per hour, i.e. 2.54 [0.50, 7.84] and 4.61 [0.49, 15.85] per hour more than the twin.

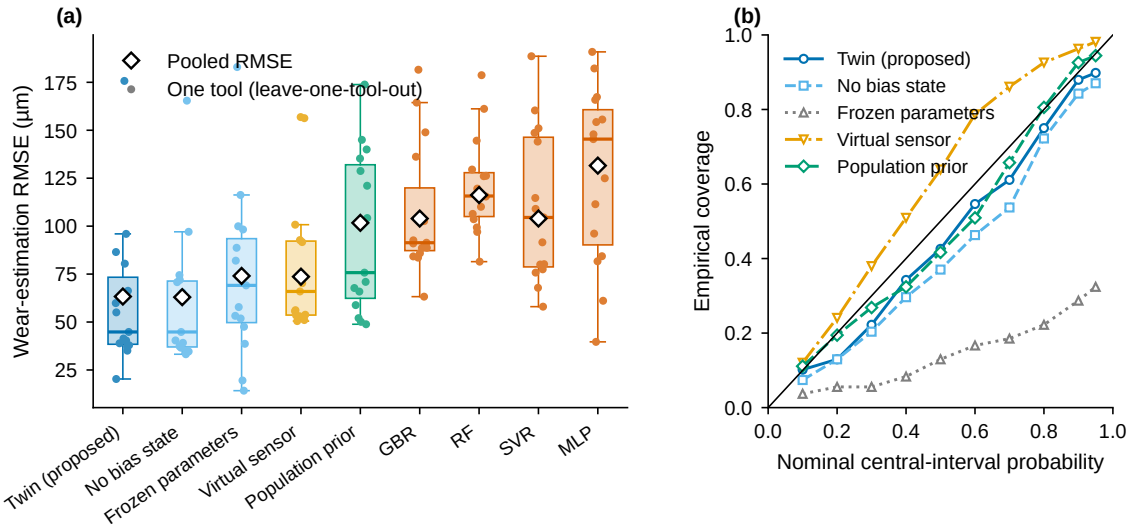


Fig. 7 Accuracy and calibration of wear estimation. (a) Per-tool RMSE of each method (dots, one per test tool; boxes show quartiles and median) and pooled RMSE (diamonds). (b) Empirical coverage of central predictive intervals against their nominal probability; the diagonal marks perfect calibration

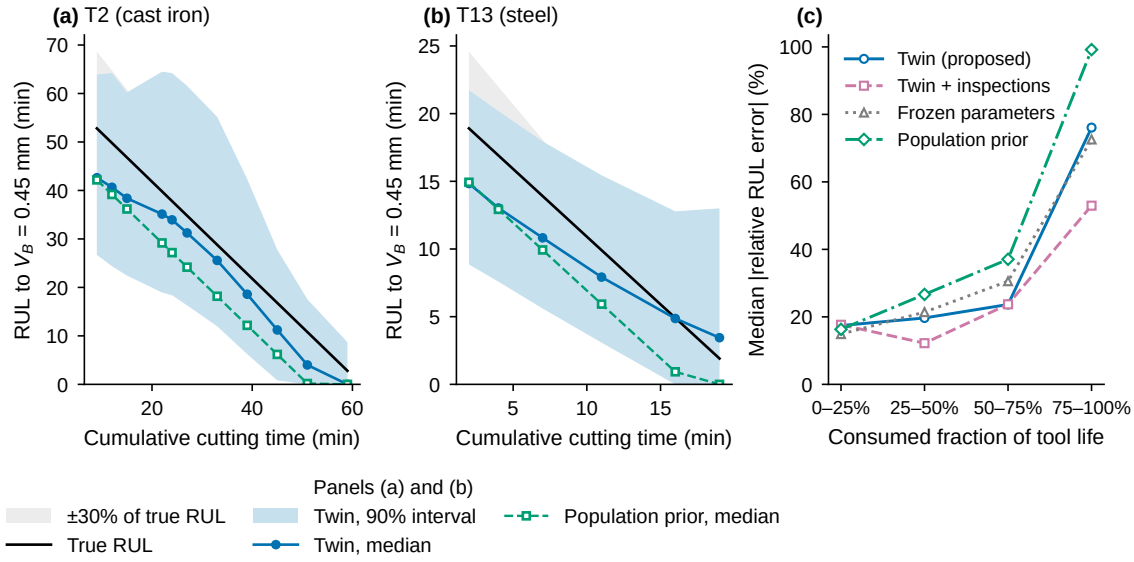


Fig. 8 RUL prognostics. (a), (b) Predicted RUL of tools T2 and T13 with the twin (median and 90% interval) and the population prior (median); the grey band is ±30% of the true RUL. (c) Median relative RUL error over the 15 test tools in four intervals of consumed tool life

5.5 What-if advisory and computational cost

Fig. 10 illustrates the what-if service for two tools past half of their life. For T16 at $t = 4$ min, the median RUL is 1.8 min at the current setting; halving the feed, the depth of cut or both raises it to 2.2, 2.5 and 3.1 min. For T4 (cast iron, $f = 0.25$ mm/rev, $a_p = 1.5$ mm) at $t = 21$ min, the corresponding values are 28.8, 36.1, 37.8 and 47.6 min. The relative gains follow from Eq. (18): halving the depth of cut multiplies the progressive wear rate by $2^{-0.433} = 0.74$, and halving the feed by $2^{-0.273} = 0.83$.

These numbers must be read together with the material removal rate, which is proportional

Table 6 RUL prediction on the 107 scoring points (median of the predictive distribution as point estimate). Late: second half of the tool life ($t_k/t_f \geq 0.5$, 57 points); α-accuracy with $\alpha = 0.3$

Method	RMSE (min)	MAE (min)	Median rel. error (%)	α-accuracy	α-accuracy late	RMSE late (min)	PICP_{90}	CRPS (min)
Digital twin (proposed)	7.00	5.65	23.7	**0.60**	0.39	6.43	0.897	3.98
Twin with sparse inspections	7.09	**5.20**	**23.1**	0.59	**0.44**	**5.41**	0.804	**3.78**
Twin without bias state	7.33	5.93	27.5	0.56	0.33	6.36	0.916	4.08
Twin with frozen parameters	**6.61**	5.21	26.2	0.58	0.40	6.26	0.243	4.61
Population prior (open loop)	7.30	6.04	30.0	0.50	0.28	6.94	0.963	4.12
Trend of virtual-sensor readings	45.23	22.18	65.8	0.23	0.14	40.83	–	–

Table 7 Replacement policies on the 15 test tools for $R = c_f/c_p = 5$. Cost rate in cost units per hour of cutting, with 95% paired-bootstrap interval

Policy	Cost rate (h^{-1})	95% interval	Failures (of 15)	Life used (%)
Run to failure	8.88	[6.55, 13.39]	15	100.0
Fixed age	9.66	[6.46, 14.48]	11	72.3
Model-based schedule	3.46	[2.09, 6.65]	1	65.0
Twin, risk quantile	3.23	[1.87, 6.47]	1	69.7
Twin + inspections, risk	2.79	[1.63, 5.32]	1	80.7
Twin, expected RUL	**2.38**	[1.66, 4.02]	0	74.6
Perfect information	1.93	[1.40, 2.98]	0	92.2

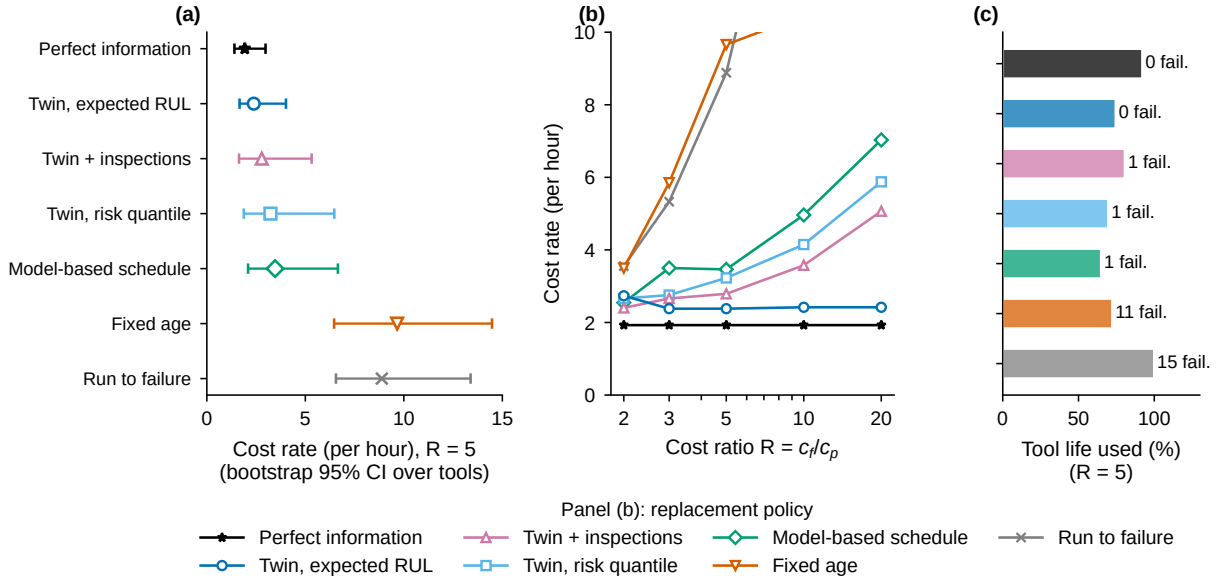


Fig. 9 Replacement decisions. (a) Cost rate of each policy at $R = 5$ with 95% paired-bootstrap interval over tools. (b) Cost rate against the failure-to-replacement cost ratio R (logarithmic axis); the legend below the figure identifies the policies. (c) Share of the total available tool life used at $R = 5$, with the number of failures among the 15 tools

to $f\,a_p$ at constant cutting speed. Halving the feed of T16 lengthens the time to the threshold by 22% but halves the removal rate, so the edge removes about 39% less material before it has to be replaced; halving both parameters lengthens the time by 72% but reduces the removed volume by 57%. Within the range of this data set, the wear rate grows more slowly than the removal rate (exponents 0.27 and 0.43, against 1 for both), so lighter cuts are a way to stretch a tool to the end of a part or of a shift, not a way to remove more material per edge. The advisory makes this trade-off explicit for the individual tool and from

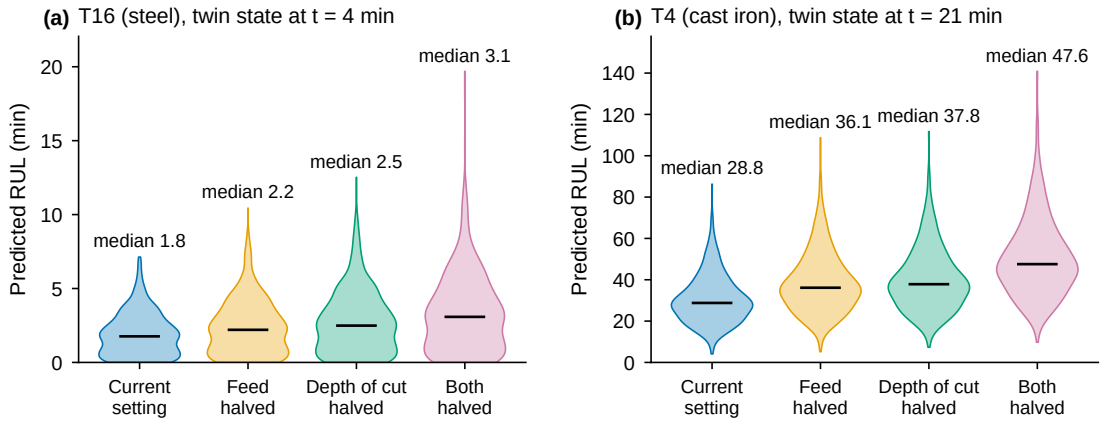


Fig. 10 What-if advisory: predicted RUL distributions (violins, with median bar and value) for the current cutting setting and for halved feed, halved depth of cut or both, computed from the twin state of (a) tool T16 at 4 min and (b) tool T4 at 21 min

Table 8 Effect of the number of particles on accuracy and on the time per update (one core). These runs use a different random seed from the main results

N	RMSE (µm)	CRPS (µm)	Median time (ms)	95th percentile time (ms)
100	64.1	35.9	0.25	0.37
250	62.9	34.4	0.28	0.38
500	62.3	34.0	0.33	0.43
1000	63.6	34.5	0.42	0.59
2000	63.5	34.3	0.59	0.85
5000	63.3	34.6	1.15	1.80

its current state, which a handbook tool-life curve cannot do.

The twin is cheap to run (Table 8). One update, comprising propagation, weighting, resampling when needed and the RUL samples, took a median of 0.62 ms (95th percentile 0.89 ms) with 2000 particles on one processor core. The accuracy hardly depends on the number of particles above a few hundred: from 250 to 5000 particles the RMSE stays between 62.3 and 63.6 µm and the CRPS between 34.0 and 34.6 µm, without any trend, and even 100 particles give 64.1 µm. The computing load is therefore negligible compared with the duration of a run and well within the capacity of an edge computer next to the machine.

6 Discussion

6.1 What each component contributes

The ablations give a consistent picture. The population prior is the most valuable single element. On its own it is not accurate for an individual tool (101.7 µm), but it gives the twin a plausible wear curve from the first run and protects it against the large scatter of single readings, as for T16 in Fig. 6. The virtual sensor supplies the evidence about the individual tool. Neither source suffices alone, and their combination in the filter beats both on 12 of the 15 tools. On-line updating of the wear parameters is what keeps the uncertainty honest: with frozen parameters, the twin is about as accurate as the virtual sensor but covers only a third of the measurements with its 95% intervals and a quarter of the true RUL values with its 90% intervals. A twin of this kind would trigger replacements with unjustified confidence.

The bias state had a smaller effect. It improved the coverage and the RUL scores (median relative error 23.7% against 27.5%, α-accuracy 0.60 against 0.56) without changing the wear RMSE, and its prior spread was set by the lower bound of Eq. (8) in every fold. Tool T8 shows its limit: a constant offset cannot describe an error that changes with the wear level, as a gain error of the virtual sensor would. A wear-dependent term in Eq. (7) is the natural extension.

The results also show why probabilistic scoring matters. Judged by the RMSE of the RUL

point estimates, the twin with frozen parameters would be the best variant (6.6 min); judged by the CRPS and the coverage, it is the worst of the twin variants. Replacement rules that act on the distribution of the RUL, such as Eq. (14), depend directly on its calibration, and the comparison of the two twin rules in Sect. 5.4 shows that a rule built on the mean is more forgiving of tail errors than one built on a small quantile when the fleet is small.

6.2 Industrial implications

The twin needs little beyond what a modern machining centre already provides. The most frequently selected signals, the spindle-motor current and an AE sensor on the table, are inexpensive or already available in the drive; the features come from the steady part of each cut; and an update costs less than a millisecond. The cutting context comes from the part programme, and the population prior grows with every tool used, so that the twin of the next tool starts from a better prior without retraining. Inspections, which gave the largest accuracy gain, can be taken at natural stops of the machine, for instance pallet changes or probing cycles, with an on-machine microscope or a camera, and the width of the posterior could be used to schedule them where they are most informative. The replacement rule of Eq. (13) is easy to explain on the shop floor: the tool is changed when its expected remaining life falls below 2.25 to 3 times the length of the next cutting interval, a margin learnt from past tools and from the cost of a failure.

Likely users are the operators of tool-management systems in batch production of costly parts, such as aerospace housings, dies and moulds and engine components, where a broken edge is expensive and tools are therefore changed early. Machine-tool builders could embed the twin in their condition-monitoring software, and cutting-tool suppliers could use the fleet prior to quantify the spread of tool life under given conditions. The what-if service supports a different kind of decision: whether to finish a part or a shift with the current edge at a lighter cut, knowing what this costs in removal rate.

6.3 Limitations

Several limitations bound these conclusions. (i) The study uses one public data set with 15 tools, one cutting speed and two levels of feed and depth of cut. The population model extrapolates poorly outside these conditions, and the cutting speed, usually the strongest driver of tool wear, does not enter the prior. (ii) The twin was run off line on recorded data; latency, data loss and integration with a machine controller were not tested. (iii) Decisions are possible only at the recorded epochs, 1 to 12 min apart, so no policy can replace a tool in the middle of a run. In one case (T5, the shortest-lived condition, with runs of 3 min) the expected-RUL rule withdrew the tool at the first epoch; a finer decision grid would reduce such losses. (iv) The 0.45 mm threshold and the costs are illustrative. The twin itself does not depend on them, but the cost comparison does, and inspections were not costed. (v) The break-in time constant often sat at the upper end of its grid and the bias spread at its lower bound, both signs that the data identify these quantities weakly. (vi) The what-if advisory uses point estimates of the condition effects, whose standard error is large for the feed (Table 4), and assumes that a change of conditions affects only the rate of progressive wear. (vii) The DC current saturates in some records, which may limit the information carried by the feature selected most often.

7 Conclusions

A hybrid digital twin of a face-milling tool was developed and evaluated with a nested leave-one-tool-out protocol on the NASA Ames milling data set. The twin combines a three-stage wear law, a population prior conditioned on feed, depth of cut and work material, a probabilistic virtual sensor with a persistent-bias error model, and particle-filter synchronisation with kernel smoothing. The main findings are the following.

1. The twin estimated flank wear with an RMSE of 63.3 µm ($R^2 = 0.80$), better than the virtual sensor alone (73.7 µm), the population prior alone (101.7 µm) and four machine-learning regressors (104.0–131.6 µm), and its 95% intervals covered 89.8% of the measurements.
2. On-line updating of the tool-specific wear parameters is essential for calibration: freezing

them reduced the coverage of the 95% intervals to 32%. Sparse inspections, one in every three measured runs, reduced the error on the other runs from 63.1 to 45.6 µm.
3. The RUL was predicted with a median relative error of 23.7% and a 90% interval coverage of 89.7%, with a tendency to err on the safe side.
4. Used for replacement, the expected-RUL rule of the twin lowered the cost rate by 31% relative to a condition-based schedule at a failure-to-replacement cost ratio of 5 and avoided all failures; its advantage grows with the cost of a failure.
5. The what-if advisory quantifies, for the individual tool, how much cutting time a lighter cut buys and how much removed material it costs.

Future work will add a wear-dependent gain to the sensor-error model, include cutting speed and tool geometry in the population prior, use the posterior to schedule inspections, and validate the twin on line on a production machine with a larger and more varied fleet of tools.

References

[1] Lei Y, Li N, Guo L, et al (2018) Machinery health prognostics: A systematic review from data acquisition to RUL prediction. Mech Syst Signal Process 104:799–834. https://doi.org/10.1016/j.ymssp.2017.11.016

[2] Zonta T, da Costa CA, da Rosa Righi R, et al (2020) Predictive maintenance in the industry 4.0: A systematic literature review. Comput Ind Eng 150:106889. https://doi.org/10.1016/j.cie.2020.106889

[3] Xia T, Dong Y, Xiao L, et al (2018) Recent advances in prognostics and health management for advanced manufacturing paradigms. Reliab Eng Syst Saf 178:255–268. https://doi.org/10.1016/j.ress.2018.06.021

[4] Serin G, Sener B, Ozbayoglu AM, et al (2020) Review of tool condition monitoring in machining and opportunities for deep learning. Int J Adv Manuf Technol 109(3-4):953–974. https://doi.org/10.1007/s00170-020-05449-w

[5] Pimenov DY, Bustillo A, Wojciechowski S, et al (2023) Artificial intelligence systems for tool condition monitoring in machining: analysis and critical review. J Intell Manuf 34(5):2079–2121. https://doi.org/10.1007/s10845-022-01923-2

[6] Tao F, Zhang H, Liu A, et al (2019) Digital twin in industry: State-of-the-art. IEEE Trans Ind Inform 15(4):2405–2415. https://doi.org/10.1109/TII.2018.2873186

[7] Jones D, Snider C, Nassehi A, et al (2020) Characterising the digital twin: A systematic literature review. CIRP J Manuf Sci Technol 29:36–52. https://doi.org/10.1016/j.cirpj.2020.02.002

[8] Liu M, Fang S, Dong H, et al (2021) Review of digital twin about concepts, technologies, and industrial applications. J Manuf Syst 58:346–361. https://doi.org/10.1016/j.jmsy.2020.06.017

[9] Tao F, Zhang H, Zhang C (2024) Advancements and challenges of digital twins in industry. Nat Comput Sci 4(3):169–177. https://doi.org/10.1038/s43588-024-00603-w

[10] ISO (2021) ISO 23247-1:2021 Automation systems and integration—Digital twin framework for manufacturing—Part 1: Overview and general principles. International Organization for Standardization, Geneva

[11] Tao F, Zhang M, Liu Y, et al (2018) Digital twin driven prognostics and health management for complex equipment. CIRP Ann 67(1):169–172. https://doi.org/10.1016/j.cirp.2018.04.055

[12] Aivaliotis P, Georgoulias K, Chryssolouris G (2019) The use of digital twin for predictive maintenance in manufacturing. Int J Comput Integr Manuf 32(11):1067–1080. https://doi.org/10.1080/0951192X.2019.1686173

[13] Errandonea I, Beltrán S, Arrizabalaga S (2020) Digital twin for maintenance: A literature review. Comput Ind 123:103316.

https://doi.org/10.1016/j.compind.2020.103316

[14] van Dinter R, Tekinerdogan B, Catal C (2022) Predictive maintenance using digital twins: A systematic literature review. Inf Softw Technol 151:107008. https://doi.org/10.1016/j.infsof.2022.107008

[15] Chen C, Fu H, Zheng Y, et al (2023) The advance of digital twin for predictive maintenance: The role and function of machine learning. J Manuf Syst 71:581–594. https://doi.org/10.1016/j.jmsy.2023.10.010

[16] Luo W, Hu T, Ye Y, et al (2020) A hybrid predictive maintenance approach for CNC machine tool driven by digital twin. Robot Comput-Integr Manuf 65:101974. https://doi.org/10.1016/j.rcim.2020.101974

[17] Lv J, Li X, Sun Y, et al (2023) A bio-inspired LIDA cognitive-based digital twin architecture for unmanned maintenance of machine tools. Robot Comput-Integr Manuf 80:102489. https://doi.org/10.1016/j.rcim.2022.102489

[18] Zhang L, Liu J, Zhuang C (2024) Digital twin modeling enabled machine tool intelligence: A review. Chin J Mech Eng 37(1):47. https://doi.org/10.1186/s10033-024-01036-2

[19] Xie Y, Lian K, Liu Q, et al (2021) Digital twin for cutting tool: Modeling, application and service strategy. J Manuf Syst 58:305–312. https://doi.org/10.1016/j.jmsy.2020.08.007

[20] Liu Z, Lang ZQ, Gui Y, et al (2024) Digital twin-based anomaly detection for real-time tool condition monitoring in machining. J Manuf Syst 75:163–173. https://doi.org/10.1016/j.jmsy.2024.06.004

[21] Sun J, Wang D, Liu Z, et al (2025) Tool digital twin based on knowledge embedding for precision CNC machine tools: Wear prediction for collaborative multi-tool. J Manuf Syst 80:157–175. https://doi.org/10.1016/j.jmsy.2025.02.021

[22] Zhou T, Zhang M, Hu T, et al (2025) Research on the online monitoring method of cutting tool wear based on the mechanism-data fusion concept of digital twin. J Manuf Process 148:386–407. https://doi.org/10.1016/j.jmapro.2025.05.028

[23] Han Z, Liang X, Luo W, et al (2025) Digital twin with dynamic mechanistic simulation core for milling tool wear prediction. J Manuf Process 149:1138–1150. https://doi.org/10.1016/j.jmapro.2025.06.023

[24] Dai W, Liu S, Lv S, et al (2026) Multi-stage tool wear prediction method based on multi-dimensional digital twin. J Intell Manuf 37(2):597–614. https://doi.org/10.1007/s10845-024-02560-7

[25] Yue C, Yue Z, Lu Y, et al (2026) Research on tool wear monitoring based on digital twin and contrastive learning. J Manuf Syst 86:1272–1289. https://doi.org/10.1016/j.jmsy.2026.04.037

[26] Lu Y, Yue C, Liu X, et al (2026) Research on an online monitoring system for tool wear status based on digital twin. Mech Syst Signal Process 252:114249. https://doi.org/10.1016/j.ymssp.2026.114249

[27] Kapteyn MG, Pretorius JVR, Willcox KE (2021) A probabilistic graphical model foundation for enabling predictive digital twins at scale. Nat Comput Sci 1(5):337–347. https://doi.org/10.1038/s43588-021-00069-0

[28] Ward R, Choudhary R, Gregory A, et al (2021) Continuous calibration of a digital twin: Comparison of particle filter and Bayesian calibration approaches. Data-Centric Eng 2:e15. https://doi.org/10.1017/dce.2021.12

[29] Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–378. https://doi.org/10.1198/016214506000001437

[30] Nemani V, Biggio L, Huan X, et al (2023) Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial. Mech Syst Signal Process 205:110796. https://doi.org/10.1016/j.ymssp.2023.110796

[31] de Jonge B, Scarf PA (2020) A review on maintenance optimization. Eur J Oper Res 285(3):805–824. https://doi.org/10.1016/j.ejor.2019.09.047

[32] Nguyen KTP, Medjaher K (2019) A new dynamic predictive maintenance framework using deep learning for failure prognostics. Reliab Eng Syst Saf 188:251–262. https://doi.org/10.1016/j.ress.2019.03.018

[33] de Pater I, Reijns A, Mitici M (2022) Alarm-based predictive maintenance scheduling for aircraft engines with imperfect remaining useful life prognostics. Reliab Eng Syst Saf 221:108341. https://doi.org/10.1016/j.ress.2022.108341

[34] Qi Q, Tao F, Hu T, et al (2021) Enabling technologies and tools for digital twin. J Manuf Syst 58:3–21. https://doi.org/10.1016/j.jmsy.2019.10.001

[35] Sharma A, Kosasih E, Zhang J, et al (2022) Digital twins: State of the art theory and practice, challenges, and open research questions. J Ind Inf Integr 30:100383. https://doi.org/10.1016/j.jii.2022.100383

[36] Tao F, Xiao B, Qi Q, et al (2022) Digital twin modeling. J Manuf Syst 64:372–389. https://doi.org/10.1016/j.jmsy.2022.06.015

[37] Lu Y, Liu C, Wang KIK, et al (2020) Digital twin-driven smart manufacturing: Connotation, reference model, applications and research issues. Robot Comput-Integr Manuf 61:101837. https://doi.org/10.1016/j.rcim.2019.101837

[38] Xiong M, Wang H, Fu Q, et al (2021) Digital twin–driven aero-engine intelligent predictive maintenance. Int J Adv Manuf Technol 114(11-12):3751–3761. https://doi.org/10.1007/s00170-021-06976-w

[39] Pulcini V, Modoni G (2024) Machine learning-based digital twin of a conveyor belt for predictive maintenance. Int J Adv Manuf Technol 133(11-12):6095–6110. https://doi.org/10.1007/s00170-024-14097-3

[40] Feng K, Ji J, Zhang Y, et al (2023) Digital twin-driven intelligent assessment of gear surface degradation. Mech Syst Signal Process 186:109896. https://doi.org/10.1016/j.ymssp.2022.109896

[41] Wang J, Ye L, Gao RX, et al (2019) Digital twin for rotating machinery fault diagnosis in smart manufacturing. Int J Prod Res 57(12):3920–3934. https://doi.org/10.1080/00207543.2018.1552032

[42] Xia M, Shao H, Williams D, et al (2021) Intelligent fault diagnosis of machinery using digital twin-assisted deep transfer learning. Reliab Eng Syst Saf 215:107938. https://doi.org/10.1016/j.ress.2021.107938

[43] Cui L, Xiao Y, Liu D, et al (2024) Digital twin-driven graph domain adaptation neural network for remaining useful life prediction of rolling bearing. Reliab Eng Syst Saf 245:109991. https://doi.org/10.1016/j.ress.2024.109991

[44] He D, Zhao J, Jin Z, et al (2025) Prediction of bearing remaining useful life based on a two-stage updated digital twin. Adv Eng Inform 65:103123. https://doi.org/10.1016/j.aei.2025.103123

[45] Gong F, Ma P, Zhang H, et al (2025) Rolling bearings remaining useful life estimation using digital twin and physics-informed methods with uncertainty quantification. Eng Appl Artif Intell 154:111070. https://doi.org/10.1016/j.engappai.2025.111070

[46] Thota HS, Saikrishna P, Vinod P, et al (2026) Remaining useful life estimation of ball-screw feed drives using a nonlinear digital twin and PINN–AIDB framework. Mech Syst Signal Process 258:114673. https://doi.org/10.1016/j.ymssp.2026.114673

[47] Faizanbasha A, Rizwan U, Hussainy ST, et al (2026) Remaining useful life aware physics-informed Bayesian digital twin for safety-constrained control in robot-integrated battery manufacturing. Robot Comput-Integr Manuf 101:103318. https://doi.org/10.1016/j.rcim.2026.103318

[48] Ma H, Fan C, Zhang Y, et al (2025) Digital twin-inspired methods for rotating machinery intelligent fault diagnosis and remaining useful life prediction: A state-of-the-art review and future challenges. Mech Syst Signal Process 232:112770. https://doi.org/10.1016/j.ymssp.2025.112770

[49] Mi S, Feng Y, Zheng H, et al (2021) Prediction maintenance integrated decision-making approach supported by digital twin-driven cooperative awareness and interconnection framework. J Manuf Syst 58:329–345. https://doi.org/10.1016/j.jmsy.2020.08.001

[50] Williams J, Jones G, Berg T, et al (2026) Cross-domain digital twin architecture for predictive maintenance via machine learning and large language models. Comput Ind Eng 215:111914. https://doi.org/10.1016/j.cie.2026.111914

[51] Xiao B, Chen L, Zhou X, et al (2026) Model updating approach for digital twin-driven industrial equipment monitoring. Adv Eng Inform 74:104678. https://doi.org/10.1016/j.aei.2026.104678

[52] Thelen A, Zhang X, Fink O, et al (2022) A comprehensive review of digital twin — part 1: modeling and twinning enabling technologies. Struct Multidiscip Optim 65(12):354. https://doi.org/10.1007/s00158-022-03425-4

[53] Thelen A, Zhang X, Fink O, et al (2023) A comprehensive review of digital twin—part 2: roles of uncertainty quantification and optimization, a battery digital twin, and perspectives. Struct Multidiscip Optim 66(1):1. https://doi.org/10.1007/s00158-022-03410-x

[54] Kong D, Chen Y, Li N (2018) Gaussian process regression for tool wear prediction. Mech Syst Signal Process 104:556–574. https://doi.org/10.1016/j.ymssp.2017.11.021

[55] Huang Z, Zhu J, Lei J, et al (2020) Tool wear predicting based on multi-domain feature fusion by deep convolutional neural network in milling operations. J Intell Manuf 31(4):953–966. https://doi.org/10.1007/s10845-019-01488-7

[56] Cheng M, Jiao L, Yan P, et al (2022) Intelligent tool wear monitoring and multi-step prediction based on deep learning model. J Manuf Syst 62:286–300. https://doi.org/10.1016/j.jmsy.2021.12.002

[57] Duan J, Zhang X, Shi T (2023) A hybrid attention-based paralleled deep learning model for tool wear prediction. Expert Syst Appl 211:118548. https://doi.org/10.1016/j.eswa.2022.118548

[58] Li W, Fu H, Han Z, et al (2022) Intelligent tool wear prediction based on Informer encoder and stacked bidirectional gated recurrent unit. Robot Comput-Integr Manuf 77:102368. https://doi.org/10.1016/j.rcim.2022.102368

[59] Cheng Y, Lu M, Gai X, et al (2024) Research on multi-signal milling tool wear prediction method based on GAF-ResNext. Robot Comput-Integr Manuf 85:102634. https://doi.org/10.1016/j.rcim.2023.102634

[60] Zhou JT, Zhao X, Gao J (2019) Tool remaining useful life prediction method based on LSTM under variable working conditions. Int J Adv Manuf Technol 104(9-12):4715–4726. https://doi.org/10.1007/s00170-019-04349-y

[61] Benkedjouh T, Medjaher K, Zerhouni N, et al (2015) Health assessment and life prediction of cutting tools based on support vector regression. J Intell Manuf 26(2):213–223. https://doi.org/10.1007/s10845-013-0774-6

[62] Li W, Liu T (2019) Time varying and condition adaptive hidden Markov model for tool wear state estimation and remaining useful life prediction in micro-milling. Mech Syst Signal Process 131:689–702. https://doi.org/10.1016/j.ymssp.2019.06.021

[63] Kim G, Yang SM, Kim DM, et al (2023) Bayesian-based uncertainty-aware tool-wear prediction model in end-milling process of titanium alloy. Appl Soft Comput 148:110922. https://doi.org/10.1016/j.asoc.2023.110922

[64] Kim G, Yang SM, Kim DM, et al (2024) Developing a deep learning-based uncertainty-aware tool wear prediction method using smartphone sensors for the turning process of Ti-6Al-4V. J Manuf Syst 76:133–157. https://doi.org/10.1016/j.jmsy.2024.07.010

[65] Fink O, Wang Q, Svensén M, et al (2020) Potential, challenges and future directions for deep learning in prognostics and health management applications. Eng Appl Artif Intell 92:103678. https://doi.org/10.1016/j.engappai.2020.103678

[66] Usui E, Shirakashi T, Kitagawa T (1984) Analytical prediction of cutting tool wear. Wear 100(1-3):129–151. https://doi.org/10.1016/0043-1648(84)90010-3

[67] Zhu K, Zhang Y (2019) A generic tool wear model and its application to force modeling and wear monitoring in high speed milling. Mech Syst Signal Process 115:147–161. https://doi.org/10.1016/j.ymssp.2018.05.045

[68] Karniadakis GE, Kevrekidis IG, Lu L, et al (2021) Physics-informed machine learning. Nat Rev Phys 3(6):422–440. https://doi.org/10.1038/s42254-021-00314-5

[69] Deng W, Nguyen KTP, Medjaher K, et al (2023) Physics-informed machine learning in prognostics and health management: State of the art and challenges. Appl Math Model 124:325–352. https://doi.org/10.1016/j.apm.2023.07.011

[70] Wang J, Li Y, Zhao R, et al (2020) Physics guided neural network for machining tool wear prediction. J Manuf Syst 57:298–310. https://doi.org/10.1016/j.jmsy.2020.09.005

[71] Li Y, Wang J, Huang Z, et al (2022) Physics-informed meta learning for machining tool wear prediction. J Manuf Syst 62:17–27. https://doi.org/10.1016/j.jmsy.2021.10.013

[72] Qiang B, Shi K, Liu N, et al (2023) Integrating physics-informed recurrent Gaussian process regression into instance transfer for predicting tool wear in milling process. J Manuf Syst 68:42–55. https://doi.org/10.1016/j.jmsy.2023.02.019

[73] Zhu K, Guo H, Li S, et al (2024) Physics-informed deep learning for tool wear monitoring. IEEE Trans Ind Inform 20(1):524–533. https://doi.org/10.1109/tii.2023.3268407

[74] Zhu K, Li X, Li S, et al (2024) Physics-informed hidden Markov model for tool wear monitoring. J Manuf Syst 72:308–322. https://doi.org/10.1016/j.jmsy.2023.11.003

[75] Sun M, Wang X, Guo K, et al (2024) Tool wear monitoring based on physics-informed Gaussian process regression. J Manuf Syst 77:40–61. https://doi.org/10.1016/j.jmsy.2024.09.001

[76] Wang Z, Zhou G, Zhang C, et al (2026) Physics-informed Gaussian process regression for tool wear monitoring with uncertainty quantification. Reliab Eng Syst Saf 268:111984. https://doi.org/10.1016/j.ress.2025.111984

[77] Han S, Awasthi U, Bollas GM (2025) Physics-informed symbolic regression for tool wear and remaining useful life predictions in manufacturing. J Manuf Syst 80:734–748. https://doi.org/10.1016/j.jmsy.2025.03.023

[78] Liu J, Wang P, Zhou K, et al (2026) Physics-informed multi-task heterogeneous transfer learning network for robust milling tool wear and RUL prediction under anomalous data

conditions. J Manuf Syst 84:403–425. https://doi.org/10.1016/j.jmsy.2025.12.011

[79] Li Z, You Z, Gu D, et al (2026) A novel physics-informed model for tool wear monitoring under multi-machining conditions. Measurement 258:119191. https://doi.org/10.1016/j.measurement.2025.119191

[80] Yan J, Bai L, Liu H, et al (2026) In-process tool wear prediction enabled by physics-informed liquid neural networks under varying cutting conditions. IEEE Trans Ind Inform 22(5):4350–4360. https://doi.org/10.1109/tii.2026.3658032

[81] Hanachi H, Yu W, Kim IY, et al (2019) Hybrid data-driven physics-based model fusion framework for tool wear prediction. Int J Adv Manuf Technol 101(9-12):2861–2872. https://doi.org/10.1007/s00170-018-3157-5

[82] Arias Chao M, Kulkarni C, Goebel K, et al (2022) Fusing physics-based and deep learning models for prognostics. Reliab Eng Syst Saf 217:107961. https://doi.org/10.1016/j.ress.2021.107961

[83] Jouin M, Gouriveau R, Hissel D, et al (2016) Particle filter-based prognostics: Review, discussion and perspectives. Mech Syst Signal Process 72-73:2–31. https://doi.org/10.1016/j.ymssp.2015.11.008

[84] Wang J, Wang P, Gao RX (2015) Enhanced particle filter for tool wear prediction. J Manuf Syst 36:35–45. https://doi.org/10.1016/j.jmsy.2015.03.005

[85] Ma Z, Zhao M, Dai X, et al (2023) A hybrid-driven probabilistic state space model for tool wear monitoring. Mech Syst Signal Process 200:110599. https://doi.org/10.1016/j.ymssp.2023.110599

[86] Li X, Liu X, Yue C, et al (2024) Data-model linkage prediction of tool remaining useful life based on deep feature fusion and Wiener process. J Manuf Syst 73:19–38. https://doi.org/10.1016/j.jmsy.2024.01.008

[87] Yang Q, Mishra D, Awasthi U, et al (2024) Tool wear and remaining useful life estimation in precision machining using interacting multiple model. J Manuf Syst 74:367–386. https://doi.org/10.1016/j.jmsy.2024.04.001

[88] Ding X, Tian W, Ma N, et al (2026) Fractional-order derivative polynomial grey particle filtering for milling tool remaining useful life prediction. Adv Eng Inform 71:104447. https://doi.org/10.1016/j.aei.2026.104447

[89] Liu X, Lunde E, Diehl F, et al (2024) Degradation modeling under time-varying operating conditions: Inference and prognosis with particle filter. Reliab Eng Syst Saf 245:109965. https://doi.org/10.1016/j.ress.2024.109965

[90] He J, Khebbache S, Anjos MF, et al (2023) Optimization of maintenance for complex manufacturing systems using stochastic remaining useful life prognostics. Comput Ind Eng 182:109348. https://doi.org/10.1016/j.cie.2023.109348

[91] Shi G, Zhang X, Zeng J, et al (2025) A predictive maintenance framework based on real-time credibility evaluation of remaining useful life prediction results. Reliab Eng Syst Saf 264:111342. https://doi.org/10.1016/j.ress.2025.111342

[92] Mitici M, de Pater I, Barros A, et al (2023) Dynamic predictive maintenance for multiple components using data-driven probabilistic RUL prognostics: The case of turbofan engines. Reliab Eng Syst Saf 234:109199. https://doi.org/10.1016/j.ress.2023.109199

[93] Lee J, Mitici M (2023) Deep reinforcement learning for predictive aircraft maintenance using probabilistic remaining-useful-life prognostics. Reliab Eng Syst Saf 230:108908. https://doi.org/10.1016/j.ress.2022.108908

[94] Zhu J, Zhu Z, Li B, et al (2026) Predictive maintenance strategy based on critical probabilistic cost. Reliab Eng Syst Saf 267:111823. https://doi.org/10.1016/j.ress.

2025.111823

[95] Xu Q, Wang Z, Jiang C, et al (2024) Data-driven predictive maintenance framework considering the multi-source information fusion and uncertainty in remaining useful life prediction. Knowl-Based Syst 303:112408. https://doi.org/10.1016/j.knosys.2024.112408

[96] Chung C, Cai BX, Chinomona B (2025) Prescriptive tool wear control: Tool wear early prediction and cutting parameter regulation for discrete turning operations. Eng Appl Artif Intell 156:111238. https://doi.org/10.1016/j.engappai.2025.111238

[97] Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830

[98] Liu J, West M (2001) Combined parameter and state estimation in simulation-based filtering. In: Doucet A, de Freitas N, Gordon N (eds) Sequential Monte Carlo Methods in Practice. Springer, New York, p 197–223

[99] Agogino A, Goebel K (2007) Milling data set. NASA Prognostics Data Repository, NASA Ames Research Center, Moffett Field, CA, URL https://data.nasa.gov/dataset/milling-wear, accessed 1 November 2026

[100] Goebel K (1996) Management of uncertainty in sensor validation, sensor fusion, and diagnosis of mechanical systems using soft computing techniques. PhD thesis, University of California at Berkeley

[101] ISO (1989) ISO 8688-1:1989 Tool life testing in milling—Part 1: Face milling. International Organization for Standardization, Geneva

[102] Saxena A, Celaya J, Saha B, et al (2010) Metrics for offline evaluation of prognostic performance. Int J Progn Health Manag 1(1). https://doi.org/10.36001/ijphm.2010.v1i1.1336

Statements and Declarations

Funding. The author declares that no funds, grants or other support were received during the preparation of this manuscript.

Competing interests. The author has no relevant financial or non-financial interests to disclose.

Author contributions. Rabah Boudiaf conceived and designed the study, implemented the methods, carried out the analysis, prepared the figures and wrote the manuscript. The author read and approved the final manuscript.

Data availability. The milling data set analysed in this study is publicly available from the NASA Ames Prognostics Data Repository [99] at https://data.nasa.gov/dataset/milling-wear. The features extracted from it and the per-run results are available from the author on reasonable request.

Code availability. The code used in this study is available from the author on reasonable request.

Ethics approval and consent. Not applicable.