symmetry

MDPI

Article

Pre-Rolling Forming-Quality Prediction in 2219 Aluminum Alloy Ring Rolling Using Physics-Informed Heterogeneous Stacking

Jianqiao Liu [1,2], Ruining Xie [1,2], Zehui Li [1,2], Xiaojie Jia [2,3], Jie Hou [1,2], Xiaoxuan Huang [1,2], Jianping Tan [1,2,3], Liang Zhou [4] and Li Li [1,2,3,*]

[1] College of Mechanical and Electrical Engineering, Central South University, Changsha 410083, China; 203701032@csu.edu.cn (J.L.); 223701019@csu.edu.cn (R.X.); 213701020@csu.edu.cn (Z.L.); 233701010@csu.edu.cn (J.H.); 233702037@csu.edu.cn (X.H.); jptan@csu.edu.cn (J.T.)
[2] State Key Laboratory of Precision Manufacturing for Extreme Service Performance, Central South University, Changsha 410083, China; 233802032@csu.edu.cn
[3] Light Alloys Research Institute, Central South University, Changsha 410083, China
[4] Hunan Zhongchuang Aerospace New Materials Co., Ltd., Yueyang 414000, China; zhouliang@innochina.ltd
* Correspondence: tonylili@csu.edu.cn

Abstract

Industrial 2219 aluminum alloy ring production requires quality assessment before a prescribed rolling plan is executed. At the investigated site, independent geometric inspection is incomplete and forming-quality feedback is not used for in-process plan adjustment. This study establishes a pre-rolling prediction framework by separating quality supervision derived from historical process signals from information available for a new workpiece. A terminal radial-fluctuation root-mean-square (TRF-RMS) indicator is constructed from the follower-roll terminal response and mapped to three enterprise-defined grades. Physics-informed heterogeneous stacking (PI-HStack) predicts these process-based grades using blank and target geometry, derived geometric relationships, initial temperature, and prescribed rolling time. Engineering descriptors represent the planned transformation; out-of-fold probability fusion combines learners, and class-specific decision adjustment addresses three-grade assignment. Five stratified 80/20 hold-out tests on 2224 industrial records yielded mean accuracy of 0.6769, macro-averaged F1-score (Macro-F1) of 0.6436, and macro-averaged one-vs-rest area under the receiver operating characteristic curve (Macro-AUC) of 0.8316. Basic geometry already provides substantial predictive information, and additional descriptors do not improve all objectives. PI-HStack has the highest reported mean overall metrics, whereas extreme gradient boosting (XGBoost) performs better on poor-grade screening measures. Ranking by the PI-HStack poor-grade score captures 56.28% of poor-grade records within the highest-ranked 20% of test records. The findings support objective-specific feature and model selection for pre-rolling plan review. The evaluation is retrospective and limited to the observed production distribution; independent geometric-tolerance validation and generalization to unseen specifications remain unestablished.

Keywords: 2219 aluminum alloy; ring rolling; pre-rolling forming-quality prediction; rotational symmetry; terminal radial fluctuation; physics-informed features; heterogeneous stacking

Check for updates

Academic Editor: Yunqing Tang

Received: 23 September 2026
Revised: 11 October 2026
Accepted: 13 October 2026
Published: 16 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Ring rolling is an incremental forming process used to produce seamless rings for bearing, aircraft, energy, and automotive applications [1]. Allwood et al. reviewed the development of ring-rolling technology [2] and investigated process behavior and production equipment [3]. The resulting annular components are designed with rotational symmetry and must meet specified geometric requirements. Previous studies have examined extremum parameters in ring rolling [4] and the influence of roll size through coupled thermo-mechanical simulation [5]. Forming defects in large aluminum alloy rings have also been investigated [6], while ring-stiffness conditions have been analyzed for radial–axial ring rolling [7].

For each specification of 2219 aluminum alloy ring, the blank dimensions, target geometry, initial temperature, and prescribed rolling time are defined before rolling begins. At the production site considered here, the rolling plan is not adjusted in response to in-process forming-quality feedback, so the plan must be assessed before execution. Rings rejected after rolling are returned for remelting without further dedicated geometric inspection. Consequently, the production database does not contain independent geometric measurements for every workpiece. These conditions create two related needs: a workpiece-level quality target derived from the retained production signals and a model that can predict that target from planning-stage information. The decision of interest is whether a prescribed plan warrants closer engineering review before execution.

Process-plan review commonly relies on engineering experience and numerical analysis. Guo and Yang investigated conditions for steady forming in radial–axial ring rolling [8]. Berti et al. developed a process worksheet and a method for predicting ring-geometry expansion during process setup [9]. Although numerical evaluation can describe ring growth and deformation behavior, repeated evaluation of alternative plans may require case-specific preparation and solution of process models. Jenkouk et al. studied finite-element analysis with integrated closed-loop control for ring rolling [1], whereas Lafarge et al. investigated data-based model predictive control [10]. Measurement-guided correction has also been explored for other circular components, including pipe-end calibration based on size detection and ovality recognition [11]. These studies address process modeling or feedback-based adjustment, whereas the present application requires an initial quality estimate from planning information available before rolling.

Data-driven smart manufacturing exploits manufacturing data to generate information and intelligence for manufacturing decisions [12]. Predictive-quality methods use machine-learning or deep-learning models to estimate product-related quality from process and product data [13], and recent reviews have summarized artificial-intelligence applications in metal forming [14]. In radial–axial ring rolling, Fahle et al. investigated time-series quality classification from rolling-process data, with the form-error target for each ring obtained from outer-shape measurements using a line-laser unit [15]. Transfer-learning approaches have also been investigated for radial–axial ring rolling [16], while Wang et al. outlined a complementary database-generation approach combining experimental and simulation methods for cold ring rolling [17]. The present study addresses a different information setting: assessment is required before rolling, while independent geometric inspection is unavailable for every historical workpiece. The terminal follower-roll response is used only to define historical supervision, and a new workpiece is assessed from its prescribed plan. This changes both the permissible inputs and the interpretation of the output: the predicted quantity is an enterprise-defined process-quality grade available for planning-stage review. The technical question is how to use existing production records for this earlier assessment and determine which representations and predictors are useful for it.

https://doi.org/10.3390/sym18101725

Three linked questions organize the study: whether a consistently defined terminal-response grade can be predicted from planning information; whether explicit blank-to-target descriptors improve on basic dimensions; and whether a model suitable for three-grade prediction also serves limited-budget poor-grade screening. The quality of manufacturing data affects predictor development and assessment [18], while class imbalance requires attention to class-sensitive learning and evaluation [19]. PI-HStack addresses these requirements through engineering feature construction, heterogeneous probability fusion, and training-based grade adjustment. Comparisons with individual learners and alternative configurations assess the value of these choices for the stated task.

The contributions are summarized as follows:

1. A workpiece-level pre-rolling prediction task is established by linking terminal radial-fluctuation root-mean-square (TRF-RMS) grades with planning records. Enterprise rules define historical grades, while only planning variables are used to predict a new workpiece. This specifies both an operational quality target and the information boundary required for assessment before rolling.
2. A task-specific physics-informed heterogeneous stacking (PI-HStack) framework combines an explicit engineering representation with learned probability fusion and class-sensitive grade assignment. Five nested feature sets (FS1–FS5), out-of-fold (OOF) meta-learning, and training-selected decision offsets provide a structured way to evaluate representation and prediction choices on the same industrial task.
3. Evaluation on 2224 industrial records yields two engineering findings: basic geometry already discriminates the process-based grades, and the best reported overall grading configuration is not the best poor-grade screening configuration. The feature and model comparisons, together with fixed-budget ranking, establish why representation and model selection should follow the intended plan-review objective.

2. Materials and Methods

2.1. Industrial Production Setting and Dataset

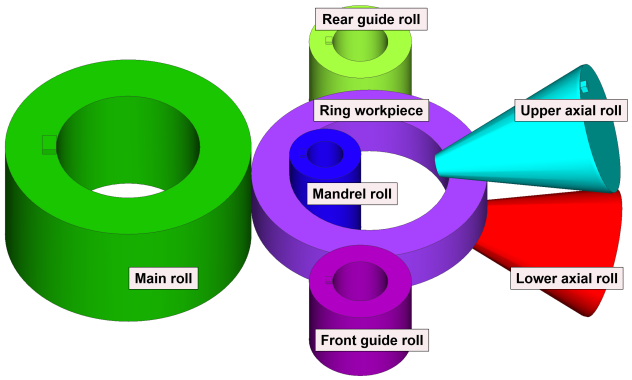
All 2224 records were collected from industrial radial–axial ring rolling of 2219 aluminum alloy, with one modeling record per workpiece. The modeling dataset contains no finite-element or synthetically generated ring-rolling samples. Workpiece-specific metallographic observations and hardness measurements were not retained in the historical production database.

The workpieces were produced using a nominal 5.5 m class industrial mill that can process larger-diameter rings under specific production configurations. The nominal radial and axial rolling-force capacities are approximately 6.9 and 3.9 MN, respectively; the rolling line speed is 0.4–1.6 m/s, and the nominal aluminum-alloy ring-mass capacity is approximately 8 t. The manufacturer, exact equipment model, production-line identifiers, product codes, and customer information are withheld for enterprise confidentiality.

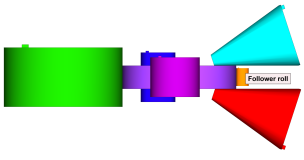
The industrial production process considered in this study comprises ring-blank preparation, pre-rolling process planning, radial–axial ring rolling, and post-rolling treatment. The 2219 aluminum alloy ring blanks are prepared through preceding forging operations. Before rolling, the blank and target dimensions, initial temperature, and prescribed rolling time are specified in the process plan. The prepared ring blank is then loaded into the mill, where coordinated radial and axial deformation progressively reduces the ring-wall thickness and height while increasing its diameter. After ring rolling, rings accepted for further processing undergo solution treatment and subsequent processing to the T8 temper, whereas rejected rings are returned for remelting without further dedicated geometric inspection.

During rolling, the follower roll remains in contact with the outer surface of the expanding ring, and its radial response is continuously recorded by the industrial control system. For historical production records, the terminal portion of this signal is used to construct the TRF-RMS-based forming-quality grades. For new workpieces, the prediction model uses only information available before rolling. Since the production line does not adjust the prescribed process plan during rolling based on forming-quality feedback, pre-rolling predictions can provide a basis for reviewing the planned process conditions before execution.

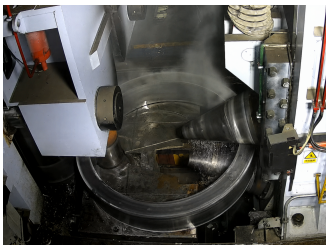
Within the production sequence described above, Figure 1a shows the roll configuration for the rolling stage, Figure 1b identifies the follower-roll measurement arrangement, and Figure 1c shows the industrial operation. During rolling, the main roll is actively driven and transfers circumferential motion to the ring through friction. The mandrel roll contacts the inner surface of the ring and moves radially toward the main roll, progressively reducing the wall thickness of the ring while rotating with the workpiece. The upper and lower axial rolls contact the upper and lower end faces of the ring and control the axial deformation; the upper axial roll moves toward the lower axial roll to reduce the ring height, while the axial-roll assembly adjusts its position as the ring diameter increases. The coordinated radial and axial deformation therefore reduces the ring cross-section and produces continuous diameter growth over successive revolutions. The front and rear guide rolls contact the outer circumference and provide lateral guidance, helping to stabilize ring motion and limit excessive center deviation during rolling. Previous analyses have examined conditions for steady radial–axial ring rolling [8], and the role of ring stiffness in the forming process has also been investigated [7].



(**a**) Overall radial–axial configuration.



(**b**) Follower-roll arrangement.



(**c**) Industrial process view.

Figure 1. Radial–axial ring rolling process used for the industrial dataset. (**a**) Overall roll configuration. (**b**) Side-view arrangement of the follower roll. (**c**) Anonymized field photograph of the industrial process.

As shown in Figure 1b, the follower roll is positioned between the upper and lower axial rolls. It rotates with the ring and follows the radial motion of the outer contour as the ring diameter increases. Such contact-based roll displacement provides a means of tracking the changing radial position of the outer surface [20]. The recorded follower-roll radial

response is denoted by $g_i(t)$ in Section 2.2. Because the response may also contain eccentric motion, contact effects, and process dynamics, it is treated as a process-level response rather than as a direct measurement of roundness or geometric tolerance.

The dataset contains 1029 distinct combinations of target outer diameter, target inner diameter, and target height. These are dimensional combinations rather than formal product model numbers. The modeling table contains no missing entries in the retained predictors and quality labels and includes 431 poor-, 755 intermediate-, and 1038 good-quality records. The predictor set contains no workpiece-specific hardness value or microstructural measurement. Table 1 summarizes the material state, equipment, geometry ranges, and quality-grade distribution.

Table 1. Anonymized industrial process and dataset summary.

Item	Description
Material	2219 aluminum alloy for all workpieces
Material state before ring rolling	Forged 2219 aluminum alloy ring blanks; solution treatment and T8 aging are performed after ring rolling
Industrial equipment	Nominal 5.5 m class radial–axial ring rolling mill; larger-diameter processing is available under specific production configurations
Nominal rolling-force capacity	Radial approximately 6.9 MN; axial approximately 3.9 MN
Rolling line speed	0.4–1.6 m/s
Modeling records	2224 workpiece-level production records
Distinct target-geometry combinations	1029 combinations of target outer diameter, inner diameter, and height
Target geometry range	Outer diameter 1260–6017 mm; inner diameter 618–5721 mm; height 80–677 mm
Blank geometry range	Outer diameter 687–4838 mm; inner diameter 236–4409 mm; height 90–705 mm
Quality-grade distribution	Poor 431 (19.38%); intermediate 755 (33.95%); good 1038 (46.67%)

Exact equipment model, product codes, and customer-specific specifications are withheld because of enterprise confidentiality.

2.2. TRF-RMS-Based Forming-Quality Target and Pre-Rolling Prediction Task

During dataset construction, TRF-RMS was defined from the terminal follower-roll response to quantify the radial fluctuation of each workpiece. For rings designed to be rotationally symmetric, the response provides process-level information associated with terminal forming quality. It may include eccentric motion, contact effects, and process dynamics. Throughout this paper, forming-quality grade denotes the enterprise-rule classification of TRF-RMS: a process-based grade of terminal forming quality, not a direct measurement of roundness or geometric tolerance.

Let $g_i(t)$ denote the follower-roll radial response of workpiece i over the terminal samples Ω_i retained during dataset construction; $|\Omega_i|$ is the number of samples in that interval. The local mean in this definition is $\bar{g}_i = |\Omega_i|^{-1} \sum_{t\in\Omega_i} g_i(t)$. The fluctuation component is

$$\delta_i(t) = g_i(t) - \bar{g}_i, \qquad t \in \Omega_i, \tag{1}$$

and the corresponding TRF-RMS is

$$r_i = \sqrt{\frac{1}{|\Omega_i|} \sum_{t\in\Omega_i} \delta_i^2(t)}. \tag{2}$$

The resulting workpiece-level values are retained in the modeling database and expressed in millimeters. Subtraction of the local mean removes a constant radial offset, while the root-mean-square aggregation summarizes the fluctuation magnitude. Smaller values correspond to lower terminal radial fluctuation. The scalar target links the terminal response to a planning record but does not reconstruct the ring's circumferential profile.

The quality boundaries of 2.0 and 6.0 mm were adopted from the enterprise quality-control rule and established engineering experience for terminal radial-fluctuation assessment. The three classes are defined as

$$y_i = \begin{cases} 2, & r_i \le 2.0, \\ 1, & 2.0 < r_i \le 6.0, \\ 0, & r_i > 6.0, \end{cases} \tag{3}$$

where 2, 1, and 0 denote good, intermediate, and poor forming quality, respectively. These boundaries are operational criteria for the site's terminal signal. TRF-RMS and the resulting class labels are historical supervision variables only and are excluded from all pre-rolling prediction inputs.

Let

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

denote the resulting dataset, where $\mathbf{x}_i \in \mathbb{R}^P$ contains information available before rolling starts and P is the number of input descriptors ($P = 28$ for FS5). The objective is to estimate the terminal forming-quality grade from $\mathbf{x}_i$:

$$f : \mathbf{x}_i \mapsto \mathbf{p}_i^{\text{final}}, \tag{5}$$

where $\mathbf{p}_i^{\text{final}} = [p_{i,0}^{\text{final}}, p_{i,1}^{\text{final}}, p_{i,2}^{\text{final}}]$ and

$$\hat{y}_i = \arg \max_{c \in \{0,1,2\}} p_{i,c}^{\text{final}}. \tag{6}$$

The poor-quality class score is denoted by

$$s_{\text{poor},i} = p_{i,0}^{\text{final}}. \tag{7}$$

This score is used for relative ranking in the retrospective process-plan review analysis. It is not a calibrated probability of rejection or remelting.

Table 2 summarizes the TRF-RMS-based forming-quality target.

Table 2. Forming-quality target based on the terminal radial-fluctuation root-mean-square (TRF-RMS) indicator.

Item	Definition
Terminal indicator	TRF-RMS r_i, defined from the terminal follower-roll radial response
Physical scope	Process-based terminal forming-quality indicator; not a direct roundness or geometric-tolerance measurement
Threshold source	Enterprise quality-control rule and established engineering experience
Class rule	Good: $r_i \le 2.0$; intermediate: $2.0 < r_i \le 6.0$; poor: $r_i > 6.0$
Role in modeling	Historical supervision only; excluded from pre-rolling model inputs
Prediction timing	All prediction inputs are available before rolling starts

2.3. Pre-Rolling Variables and Physics-Informed Feature Representation

The pre-rolling variables include blank geometry, target geometry, initial temperature, and prescribed rolling time. The time is specified in the plan before rolling and is not the elapsed duration observed after production. In PI-HStack, physics-informed refers to the construction and organization of input descriptors using ring-rolling geometry and planning knowledge. Diameter changes, wall-thickness relationships, geometric-volume proxies, and dimensionless proportions expose prescribed expansion, cross-sectional change, and shape proportions that are implicit in the original dimensions. The purpose is to give the learner an explicit engineering representation and test its value through FS1–FS5. This incorporation occurs at the representation level; the model does not enforce governing equations or conservation-law residuals in its training objective.

The outer diameter, inner diameter, and height are denoted by D, d, and H; subscripts 0 and f denote the blank and target states. The five nested feature sets defined below contain 6, 12, 20, 26, and 28 descriptors, respectively.

FS1: Basic geometry.

$$\mathrm{FS1} = \{D_f, d_f, H_f, D_0, d_0, H_0\}. \tag{8}$$

These six variables specify the absolute blank and target dimensions.

FS2: Blank-to-target geometric change. FS2 adds

$$\Delta D = D_f - D_0, \qquad \Delta d = d_f - d_0, \qquad \Delta H = H_f - H_0, \tag{9}$$

and

$$\rho_D = \frac{D_f}{D_0 + \epsilon_L}, \qquad \rho_d = \frac{d_f}{d_0 + \epsilon_L}, \qquad \rho_H = \frac{H_f}{H_0 + \epsilon_L}, \tag{10}$$

where ϵ_L denotes a denominator safeguard with the same length units as the geometric variables. It is distinct from the dimensionless probability-clipping constants introduced below. The differences and ratios quantify prescribed absolute and relative geometric change from the blank to the target specification.

FS3: Wall-thickness and geometric-volume descriptors. The blank and target wall thicknesses are

$$w_0 = \frac{D_0 - d_0}{2}, \qquad w_f = \frac{D_f - d_f}{2}, \tag{11}$$

with wall-thickness reduction

$$\Delta w = w_0 - w_f. \tag{12}$$

Treating the blank and target as ideal annular cylinders defines the geometric-volume proxies

$$V_0^{\mathrm{geom}} = \frac{\pi H_0 (D_0^2 - d_0^2)}{4}, \qquad V_f^{\mathrm{geom}} = \frac{\pi H_f (D_f^2 - d_f^2)}{4}. \tag{13}$$

FS3 contains blank and target wall thickness, wall-thickness change and ratio, geometric-volume proxies, and their difference and ratio. These descriptors express cross-sectional and idealized geometric-volume relationships in the prescribed plan; they are not measurements of material loss.

FS4: Normalized geometry. FS4 further adds

$$\frac{D_0}{H_0}, \quad \frac{D_f}{H_f}, \quad \frac{D_0}{d_0}, \quad \frac{D_f}{d_f}, \quad \frac{w_0}{H_0}, \quad \frac{w_f}{H_f}, \tag{14}$$

which describe the proportions of the blank and target rings. These dimensionless ratios compactly describe the proportions of the rotationally symmetric annular geometry and facilitate comparisons across different ring sizes.

FS5: Comprehensive pre-rolling representation. FS5 adds the initial temperature T_0 and prescribed rolling time t_{roll}. It contains 28 descriptors and is used in the final PI-HStack model. The complete descriptor definitions and units are listed in Supplementary Table S2. FS2–FS4 are deterministic transformations of the six geometric inputs in FS1, whereas FS5 also introduces the two process variables. The nested comparison therefore evaluates both alternative representations of geometry and the joint addition of temperature and prescribed time.

Table 3 summarizes the five feature sets.

Table 3. Nested pre-rolling feature sets (FS1–FS5) and their engineering interpretation.

Set	No.	Added Descriptors	Engineering Meaning
FS1	6	Target and blank geometry	Absolute blank and target dimensions
FS2	12	Dimensional changes and ratios	Prescribed relative and absolute dimensional changes
FS3	20	Wall-thickness and volume descriptors	Cross-sectional and idealized geometric-volume relationships
FS4	26	Normalized geometric ratios	Dimensionless shape proportions
FS5	28	Initial temperature and prescribed rolling time	Initial thermal condition and planned process duration

2.4. PI-HStack Prediction Model

The two paths in Figure 2 separate historical supervision from prediction of a planned workpiece. PI-HStack is designed to address three requirements of this task. First, alternative tabular learners estimate class relationships from the same moderate-sized planning dataset. Second, OOF meta-learning combines their probability estimates using predictions from models that did not fit the represented workpiece. Third, class-balanced logistic regression and decision offsets selected by Macro-F1 address assignment across the three imbalanced grades. Probability fusion and grade assignment thus have distinct roles, and poor-grade ranking is evaluated separately. The configuration is assessed against its individual learners and alternative fusion/decision settings in Section 3.2; its selection for overall grading is discussed in Section 4.3. Prediction uses the stored fold-specific models and requires no terminal measurement.

2.4.1. Base Learners and Preprocessing

The base learners are the Tabular Prior-data Fitted Network (TabPFN) [21], CatBoost [22], random forest (RF) [23], and extreme gradient boosting (XGBoost) [24]. Prior benchmarks have revisited the performance of deep-learning architectures for tabular data [25]. Other comparisons have documented the competitiveness of tree-based models on typical tabular datasets [26]. The four learners supply alternative class-probability estimates, allowing the meta-learner to learn their relative contributions rather than impose equal averaging or select one learner in advance. Complementarity is a design motivation; its practical value is assessed through the model comparisons.

Numerical predictors are passed through median imputation fitted on the relevant training subset. Although the retained modeling table has no missing entries, the same preprocessing pipeline is used in all evaluations. RF uses balanced-subsample class weighting, CatBoost uses automatic balanced class weights, and XGBoost receives balanced sample weights. No additional external class weighting is supplied to TabPFN. Base-learner parameters are fixed in the evaluation scripts and listed in Supplementary Table S3.

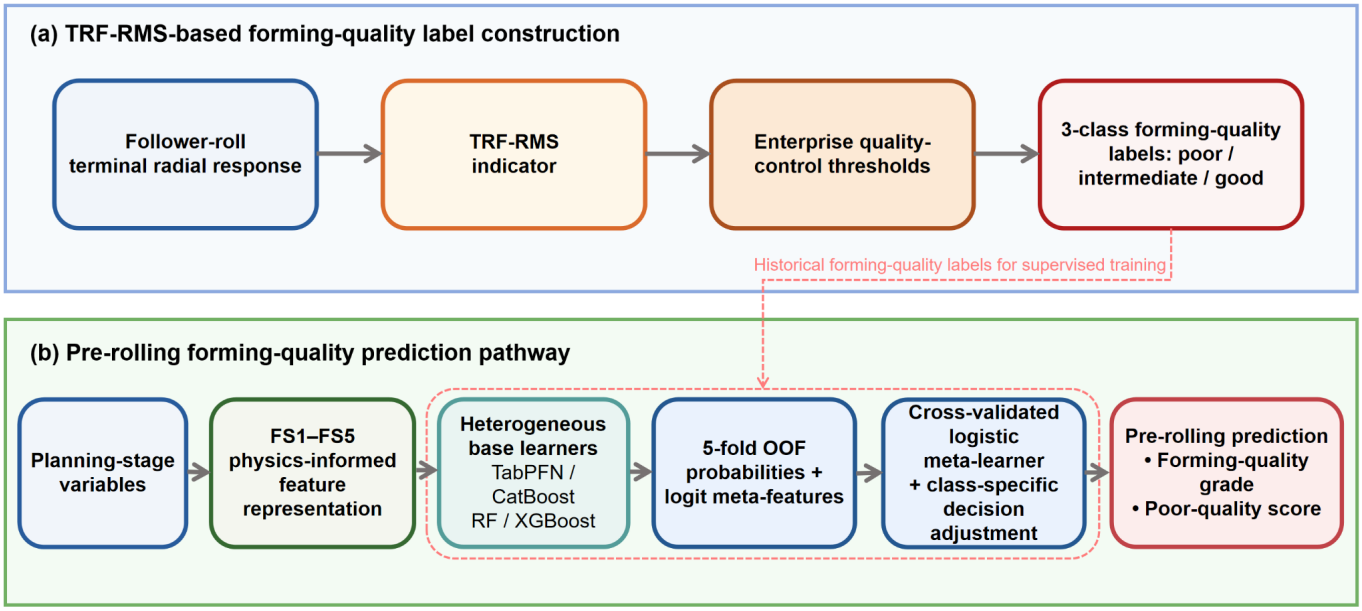


Figure 2. Physics-informed heterogeneous stacking (PI-HStack) for pre-rolling forming-quality prediction. (**a**) Terminal radial-fluctuation root-mean-square (TRF-RMS) label construction using enterprise thresholds and engineering experience. (**b**) Planning-stage variables are organized into five feature sets (FS1–FS5) and processed by a Tabular Prior-data Fitted Network (TabPFN), CatBoost, random forest (RF), and extreme gradient boosting (XGBoost). Five-fold out-of-fold (OOF) probabilities and their logit transforms form the meta-features for logistic-regression fusion and class-specific decision adjustment. Ring-rolling knowledge enters through descriptor construction. The dashed connection applies to training only. Inference uses the fitted models and requires no terminal response or quality label.

For learner $m \in \{1, \ldots, M\}$, with $M = 4$, the probability vector for workpiece i is

$$\mathbf{p}_i^{(m)} = [p_{i,0}^{(m)}, p_{i,1}^{(m)}, p_{i,2}^{(m)}]. \tag{15}$$

The probabilities of the four learners are fused by stacked generalization [27].

2.4.2. OOF Probability Stacking and Meta-Learning

The meta-learner is trained using OOF predictions so that no training sample is represented by a base-model prediction from a model fitted on that sample. Let the training set $\mathcal{D}_{\text{tr}}$ be partitioned into $K = 5$ stratified folds $\{I_1, \ldots, I_K\}$. For learner m and fold k,

$$f_m^{(-k)} = \text{Train}_m(\mathcal{D}_{\text{tr}} \setminus \mathcal{D}_{I_k}), \tag{16}$$

and, for $i \in I_k$,

$$\mathbf{p}_{i,m}^{\text{OOF}} = f_m^{(-k)}(\mathbf{x}_i). \tag{17}$$

The OOF probabilities from all base learners are concatenated as

$$\mathbf{p}_i^{\text{OOF-stack}} = [\mathbf{p}_{i,1}^{\text{OOF}}, \ldots, \mathbf{p}_{i,M}^{\text{OOF}}]. \tag{18}$$

The second-level representation contains the base probabilities and their logit transforms:

$$\mathbf{z}_i = [\mathbf{p}_i^{\text{OOF-stack}}, \text{logit}(\mathbf{p}_i^{\text{OOF-stack}})], \tag{19}$$

where

$$\mathrm{logit}(p) = \log \frac{\mathrm{clip}(p, \epsilon_{\mathrm{logit}}, 1-\epsilon_{\mathrm{logit}})}{1-\mathrm{clip}(p, \epsilon_{\mathrm{logit}}, 1-\epsilon_{\mathrm{logit}})}. \tag{20}$$

The clipping value is $\epsilon_{\mathrm{logit}} = 10^{-6}$. Four learners and three classes yield 12 probability features and 12 logit features, giving 24 meta-features. The logit terms let the linear meta-learner use both probabilities and their log-odds representation. They are a nonlinear re-expression of the same estimates, not additional independent observations; the NoLogit comparison evaluates this representation choice. A class-balanced logistic-regression meta-learner maps $\mathbf{z}_i$ to the pre-adjustment class probabilities $\hat{\mathbf{p}}_i$. Its regularization parameter is selected by three-fold cross-validation using the macro-averaged F1-score (Macro-F1).

For a held-out test workpiece, each fold-specific base model predicts the sample and the K probability vectors are averaged:

$$\mathbf{p}_{j,m}^{\mathrm{test}} = \frac{1}{K}\sum_{k=1}^{K} f_m^{(-k)}(\mathbf{x}_j). \tag{21}$$

The averaged test probabilities are then transformed using the same probability-plus-logit mapping before meta-level prediction; the averaging precedes the logit transformation.

2.4.3. Class-Specific Decision Adjustment

Decision offsets are selected from an additional three-fold split of the training meta-feature matrix. Within each split, a logistic-regression model is fitted to the other two folds, with its regularization strength selected internally by three-fold cross-validation. Predictions for the held-out meta-fold are collected for offset selection. This is a meta-level resampling procedure on the fixed base-level OOF representation, not a fully nested refitting of the entire stack. All such operations are confined to the outer training data. The final meta-learner is then fitted using the complete training meta-feature matrix.

For class c, the selected offsets modify the pre-adjustment probabilities as follows:

$$\tilde{s}_{i,c} = \log(\mathrm{clip}(\hat{p}_{i,c}, \epsilon_{\mathrm{adj}}, 1)) + \beta_c, \tag{22}$$

and

$$\tilde{p}_{i,c} = \frac{\exp(\tilde{s}_{i,c})}{\sum_{\ell=0}^{2}\exp(\tilde{s}_{i,\ell})}. \tag{23}$$

Here, $\epsilon_{\mathrm{adj}} = 10^{-8}$. Adding the same constant to all offsets leaves the normalized probabilities unchanged, so the good-class offset is fixed at zero as a reference. The poor- and intermediate-class offsets are searched over $[-0.8, 0.8]$ with a step of 0.1 and selected to maximize Macro-F1 on training meta-level OOF predictions. The held-out test set is not used for this selection. The final class is

$$\hat{y}_i = \arg\max_{c\in\{0,1,2\}} \tilde{p}_{i,c}, \tag{24}$$

and $\mathbf{p}_i^{\mathrm{final}} = \tilde{\mathbf{p}}_i$ for PI-HStack. Selecting offsets by Macro-F1 aligns grade assignment with equal consideration of the three class-specific F1-scores rather than with poor-class recall alone. This distinction motivates the separate evaluation of overall grading and poor-grade screening. The operation renormalizes the scores; it is not probability calibration and need not preserve the ranking of poor-class scores across workpieces.

2.5. Evaluation Protocol and Metrics

The main evaluation uses five stratified 80/20 hold-out splits with seeds $\{0, 1, 2, 3, 4\}$. Each split contains 1779 training records and 445 test records. Numerical preprocessing,

base-level OOF generation, meta-learner selection, and decision adjustment use only the corresponding outer training data. The fitted model is then evaluated on the held-out test records. Five-fold stratified cross-validation provides an additional evaluation within the same production distribution. The implications of random splitting for recurring specifications are discussed in Section 4.4.

The single-model baselines are RF, XGBoost, CatBoost, TabPFN, and a multilayer perceptron (MLP). The probability-ensemble heterogeneous stacking (PE-HStack) baseline combines TabPFN, CatBoost, and RF using probability-plus-logit meta-features without decision adjustment. PE-HStack+XGBoost adds the fourth learner; PI-HStack-NoLogit retains the four learners and decision adjustment but removes logit meta-features. MultiView PE-HStack combines the three-learner outputs obtained from FS3, FS4, and FS5 without decision adjustment. All variants predict the same three forming-quality grades.

The evaluation settings are summarized in Table 4.

Table 4. Evaluation and model settings: out-of-fold (OOF), feature set (FS), random forest (RF), Tabular Prior-data Fitted Network (TabPFN), multilayer perceptron (MLP), and extreme gradient boosting (XGBoost).

Item	Setting
Main evaluation	Five repeated stratified 80/20 hold-out splits; seeds $\{0, 1, 2, 3, 4\}$
Training/test size	1779/445 records per split
Final input representation	FS5 with 28 pre-rolling descriptors
Compared single learners	RF, XGBoost, CatBoost, TabPFN, and MLP
Final stacking members	TabPFN, CatBoost, RF, and XGBoost
Base-level OOF	Five stratified folds within each outer training split
Meta-learner	Three-fold regularization selection; separate three-fold meta-level OOF predictions for offset selection
Meta representation	12 class-probability features plus 12 logit features (24 total)
Decision adjustment	Poor/intermediate offsets searched over $[-0.8, 0.8]$ with step 0.1; good offset fixed at 0
Robustness analyses	Five-fold cross-validation and TRF-RMS boundary perturbation
Auxiliary screening	Top-q ranking and retrospective threshold-based flagging using the poor-quality score

Classification is assessed by accuracy, Macro-F1, macro-averaged one-vs-rest area under the receiver operating characteristic curve (Macro-AUC), and poor-class recall and F1-score. The F1-score is the harmonic mean of precision and recall. Macro-F1 averages the three class-specific F1-scores equally, whereas Macro-AUC averages the three one-vs-rest areas under the curve. Mean and sample standard deviation are reported over the five splits; model comparisons refer to these descriptive statistics rather than significance tests. Repeated hold-out splits can share workpieces, so the reported variation describes sensitivity to resampling within this dataset, not independent production replications. Confusion matrices and class-specific curves pool prediction instances from the repeated tests, in which individual workpieces may recur. Supplementary Table S1 reproduces the model summaries for reference.

For auxiliary plan-review analysis, test workpieces are ranked by s_{poor}. Let q be the selected fraction, n_{te} the number of test records, and $k = \max[1, \text{round}(qn_{\text{te}})]$. Recall@$q$ is the fraction of poor-quality records among the selected top-k workpieces relative to all

poor-quality records in that test split. Precision@q is the poor-quality fraction in the selected subset. Lift is

$$\text{Lift@}q = \frac{\text{Precision@}q}{\pi_{\text{poor}}}, \tag{25}$$

where π_{poor} is the test-set poor-quality prevalence. This uses the actual selected count k, so rounding can make the selected fraction differ slightly from q.

A retrospective analysis considers score thresholds from 0.20 to 0.50 in steps of 0.05. Records with $s_{\text{poor}} \geq \tau$ are flagged. The flagging rate is

$$\text{FlaggingRate}(\tau) = \frac{1}{n_{\text{te}}} \sum_{i=1}^{n_{\text{te}}} \mathbf{1}(s_{\text{poor},i} \geq \tau). \tag{26}$$

Precision, recall, and F1-score are then calculated with the poor-quality grade as the positive class. The selected fraction represents a hypothetical budget for pre-rolling process-plan review. Thresholds are examined retrospectively, as reported in Section 3.3.

The enterprise grading boundaries remain 2.0 and 6.0 mm. In the separate sensitivity experiment, the routine reconstructs a nearby pair of reference boundaries from the observed TRF-RMS–label correspondence and scales that pair by 0.90, 0.95, 1.00, 1.05, and 1.10. Model evaluation is repeated after relabeling and stratified splitting under each scheme. The reconstructed pair is used only to examine sensitivity to the grading convention; the enterprise rule remains unchanged. The reconstruction routine and its numerical reference values are documented in Supplementary Section S5.

The evaluation addresses reliability at distinct levels: repeated hold-out testing reports split-to-split variation; cross-validation provides an additional resampling check; model-configuration and feature-set comparisons examine the dependence on the evaluated design choices; and boundary perturbation examines local sensitivity to the grading convention. These analyses support interpretation within the observed production distribution. They do not independently validate geometric tolerances or generalization to new specifications. Supplementary Table S3 and Algorithm S1 document the reported settings and evaluation workflow.

3. Results

3.1. Overall Forming-Quality Prediction

PI-HStack achieved mean accuracy of 0.6769, Macro-F1 of 0.6436, and Macro-AUC of 0.8316, the highest reported means for these three metrics in the main comparison (Table 5). Its advantage over XGBoost was modest, particularly for Macro-F1. XGBoost achieved higher poor-class F1, Recall@20%, and Lift@20%, whereas MultiView PE-HStack produced the highest poor-class recall. Thus, the ranking of models depends on whether the objective is overall grade prediction or poor-class screening. The evaluated MLP did not outperform the leading tree-based or stacked models. The corresponding results for each repeated hold-out split are summarized in Supplementary Table S1.

The pooled confusion matrix indicates that most errors occurred between adjacent quality levels, with the intermediate grade accounting for a substantial share of the confusion (Figure 3). Among 430 poor-class prediction instances, 203 were classified as poor, 178 as intermediate, and 49 as good. The matrix contains 2225 prediction instances from five tests of 445 records; a workpiece can occur in more than one test split.

The pooled receiver operating characteristic (ROC) and precision–recall curves provide complementary views of class-specific discrimination (Figure 4). The class-specific area under the ROC curve (AUC) and average precision (AP) summarize these pooled score distributions. They need not equal the mean of the five split-specific metrics in Table 5.

Table 5. Model comparison over five stratified 80/20 hold-out splits. Entries are mean ± sample standard deviation. Metrics are the macro-averaged F1-score (Macro-F1), macro-averaged one-vs-rest area under the receiver operating characteristic curve (Macro-AUC), and poor-class recall within the highest-ranked 20% of records (R@20%). The probability-ensemble heterogeneous stacking (PE-HStack) baselines are defined in Section 2.5.

Model	Accuracy	Macro-F1	Macro-AUC	Poor Recall	Poor F1	R@20%	Lift@20%
RF	0.6521 ± 0.0157	0.6214 ± 0.0169	0.8137 ± 0.0120	0.5302 ± 0.0455	0.5364 ± 0.0291	0.5651 ± 0.0345	2.8256 ± 0.1725
CatBoost	0.6494 ± 0.0053	0.6226 ± 0.0089	0.8266 ± 0.0055	0.5907 ± 0.0312	0.5494 ± 0.0268	0.5837 ± 0.0301	2.9186 ± 0.1505
TabPFN	0.6674 ± 0.0140	0.6227 ± 0.0153	0.8271 ± 0.0099	0.4070 ± 0.0339	0.4989 ± 0.0328	0.5581 ± 0.0368	2.7907 ± 0.1839
MLP	0.6539 ± 0.0185	0.6021 ± 0.0197	0.7932 ± 0.0102	0.3581 ± 0.0529	0.4546 ± 0.0498	0.5000 ± 0.0526	2.5000 ± 0.2632
XGBoost	0.6674 ± 0.0100	0.6427 ± 0.0149	0.8270 ± 0.0075	0.6023 ± 0.0644	**0.5702 ± 0.0441**	**0.5884 ± 0.0517**	**2.9419 ± 0.2587**
PE-HStack	0.6503 ± 0.0146	0.6250 ± 0.0201	0.8295 ± 0.0092	0.6302 ± 0.0595	0.5536 ± 0.0441	0.5744 ± 0.0303	2.8721 ± 0.1516
PE-HStack+XGBoost	0.6539 ± 0.0146	0.6272 ± 0.0186	0.8308 ± 0.0091	0.6116 ± 0.0530	0.5454 ± 0.0404	0.5581 ± 0.0403	2.7907 ± 0.2014
MultiView PE-HStack	0.6526 ± 0.0167	0.6297 ± 0.0206	0.8295 ± 0.0093	**0.6488 ± 0.0482**	0.5680 ± 0.0438	0.5837 ± 0.0380	2.9186 ± 0.1902
PI-HStack	**0.6769±0.0143**	**0.6436±0.0136**	**0.8316±0.0093**	0.4721±0.0585	0.5440±0.0340	0.5628±0.0416	2.8140±0.2080

Bold values indicate column maxima; bold PI-HStack identifies the proposed model.

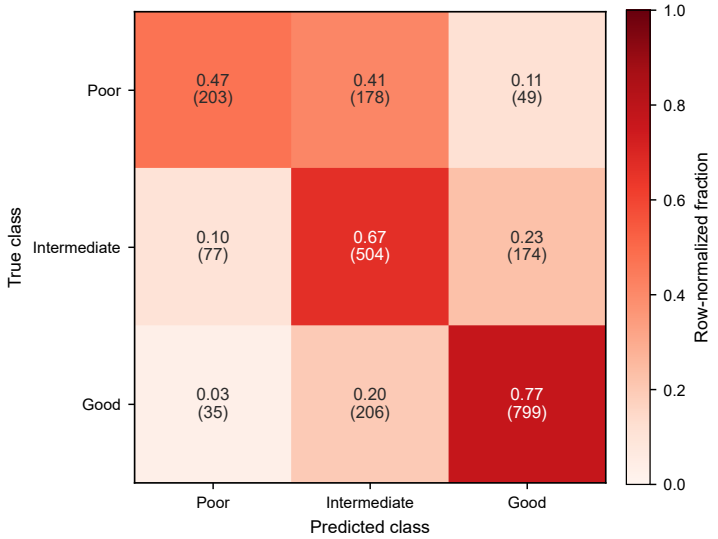


Figure 3. Row-normalized confusion matrix of PI-HStack pooled over the five held-out tests. Parentheses contain prediction-instance counts. The pooled total is 2225; individual workpieces may recur across test splits.

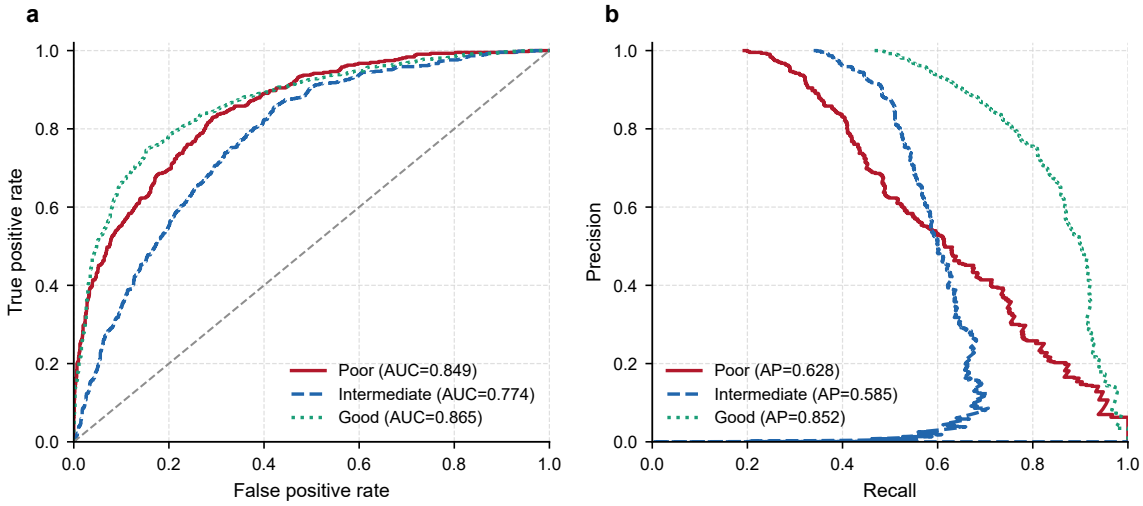


Figure 4. Pooled class-specific discrimination of PI-HStack. (**a**) One-vs-rest receiver operating characteristic (ROC) curves with area under the ROC curve (AUC). (**b**) Precision–recall curves with average precision (AP). The diagonal reference line in panel (**a**) represents the performance of a random classifier. Curves are based on pooled repeated-test prediction instances.

3.2. Ablation and Robustness Analyses

Table 6 summarizes the evaluated learner compositions, logit meta-features, and decision-adjustment settings. Adding XGBoost to the unadjusted three-learner stack changed the mean metrics only slightly. The adjusted four-learner configuration has higher reported accuracy (0.6769 versus 0.6539) and Macro-F1 (0.6436 versus 0.6272) than the unadjusted four-learner configuration, while Macro-AUC differs only slightly (0.8316 versus 0.8308). These are configuration-level comparisons under the reported evaluation protocol; the observed differences are not interpreted as isolated contributions of any individual component. Relative to NoLogit, the full configuration has higher accuracy and Macro-F1 but slightly lower Macro-AUC and Recall@20%. The three-learner PE-HStack retains the highest poor-class F1 and Recall@20% in this comparison.

Table 6. Comparison of learner-fusion and decision configurations; bold metric values indicate the highest reported mean in each column.

Design	TabPFN	CatBoost	RF	XGBoost	Logit	Adj.	Acc.	Macro-F1	Macro-AUC	Poor F1	R@20%
PE-HStack	✓	✓	✓	–	✓	–	0.6503	0.6250	0.8295	**0.5536**	**0.5744**
PE-HStack+XGBoost	✓	✓	✓	✓	✓	–	0.6539	0.6272	0.8308	0.5454	0.5581
PI-HStack-NoLogit	✓	✓	✓	✓	–	✓	0.6710	0.6395	**0.8318**	0.5356	0.5698
PI-HStack	✓	✓	✓	✓	✓	✓	**0.6769**	**0.6436**	0.8316	0.5440	0.5628

Acc. = accuracy; Adj. = class-specific decision adjustment; Logit = logit-transformed probability meta-features. R@20% = poor-class recall within the highest-ranked 20%. The PI-HStack row uses the full-model result reported in Table 5. Differences between configurations are descriptive and should not be read as isolated estimates of individual component effects.

Basic geometry already provided substantial predictive information: FS1 achieved Macro-AUC of 0.8305 and the highest poor-class F1 and Recall@20% among the evaluated feature sets (Table 7). Expanding the geometric representation through FS2–FS4 did not yield consistent gains. Adding initial temperature and prescribed rolling time to FS4 produced the highest mean accuracy, Macro-F1, and Macro-AUC in FS5. Because these two variables were added together, this comparison assesses their joint addition within the expanded representation. The FS1–FS5 contrast does not isolate the benefit of derived geometry at a fixed set of original process inputs. Definitions and units of the pre-rolling descriptors are provided in Supplementary Table S2, and their correlations are shown in Supplementary Figure S1.

Table 7. Feature-set ablation under PI-HStack; bold metric values indicate the highest reported mean in each column.

Feature Set	Basic	Change	Wall/Vol.	Ratio	Temp/Time	Acc.	Macro-F1	Macro-AUC	Poor F1	R@20%
FS1 basic geometry	✓	–	–	–	–	0.6688	0.6370	0.8305	**0.5531**	**0.5953**
FS2 geometry change	✓	✓	–	–	–	0.6665	0.6326	0.8257	0.5305	0.5721
FS3 wall/volume	✓	✓	✓	–	–	0.6656	0.6375	0.8289	0.5474	0.5837
FS4 geometry ratio	✓	✓	✓	✓	–	0.6634	0.6300	0.8278	0.5292	0.5791
FS5 comprehensive	✓	✓	✓	✓	✓	**0.6769**	**0.6436**	**0.8316**	0.5440	0.5628

Basic = blank and target dimensions; Change = dimensional differences and ratios; Wall/vol. = wall-thickness and volume descriptors; Ratio = geometric proportions; Temp/time = initial temperature and prescribed rolling time. Acc. = accuracy; R@20% = poor-class recall within the highest-ranked 20%. The FS5 reference is shared with the full model in Table 5.

Five-fold stratified cross-validation was performed for the representative baseline subset listed in Table 8. PI-HStack and TabPFN both reached accuracy of 0.6677 at the reported precision. PI-HStack had the highest Macro-F1, Macro-AUC, and Recall@20% within this subset, while PE-HStack had the highest poor-class F1. The evaluated subset did not include XGBoost as a stand-alone learner.

Table 8. Five-fold cross-validation robustness analysis; bold metric values indicate the highest reported mean in each column.

Model	Accuracy	Macro-F1	Macro-AUC	Poor F1	R@20%
RF	0.6520	0.6194	0.8132	0.5254	0.5290
CatBoost	0.6506	0.6236	0.8231	0.5431	0.5661
TabPFN	**0.6677**	0.6217	0.8258	0.4960	0.5638
PE-HStack	0.6497	0.6246	0.8267	**0.5456**	0.5660
PI-HStack	**0.6677**	**0.6371**	**0.8295**	0.5407	**0.5846**

Cross-validation and label-boundary sensitivity provide additional within-dataset checks (Figure 5). Accuracy and Macro-AUC changed only slightly over the evaluated boundary multipliers, whereas Macro-F1 showed greater variation. Each boundary scheme changes class assignments and stratified splits, so the results describe local sensitivity to the grading convention. The reconstruction procedure underlying Figure 5b is detailed in Supplementary Section S5.

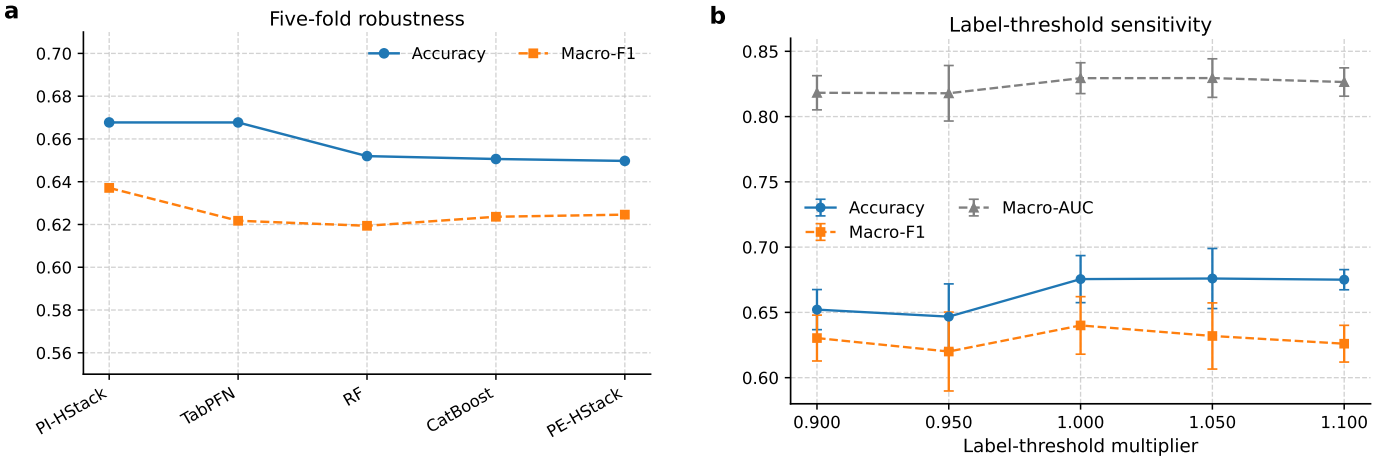


Figure 5. Validation and target-definition sensitivity results. (**a**) Five-fold mean accuracy and macro-averaged F1-score (Macro-F1) for the evaluated subset; lines connect model summaries for readability. (**b**) Label-boundary sensitivity, including the macro-averaged one-vs-rest area under the receiver operating characteristic curve (Macro-AUC). Error bars in panel (**b**) show standard deviation across the five hold-out splits. The multipliers apply to the reconstructed reference boundaries used for sensitivity analysis.

3.3. Auxiliary Pre-Rolling Review Prioritization

The poor-quality score ranks workpieces for possible process-plan review before rolling. In the repeated hold-out evaluation, the highest-ranked 20% of test records contained 56.28% of poor-quality records, giving mean Lift@20% of 2.8140. Selecting 30% contained 71.63% of poor-quality records. These results express the trade-off between review capacity and poor-class coverage: a fixed top-q policy allocates a prescribed fraction of plans to review. They evaluate retrospective prioritization against the TRF-RMS grades, not the effect of subsequent intervention on production quality.

Figure 6a,b illustrates a representative split, in which top-20% recall is 52.3%; Figure 6c,d summarize the repeated tests. Among the evaluated score thresholds, 0.35 produced the highest mean flagging F1, with a flagging rate of 19.87% and poor-class recall of 56.98%. A threshold of 0.30 yielded recall of 65.58% at a flagging rate of 26.07%. Unlike a fixed top-q budget, a score threshold gives a selected fraction that depends on the score distribution. The threshold of 0.35 is a retrospective descriptive result, not a production

recommendation. A threshold for production use should be selected using training or validation data together with the site's review capacity.

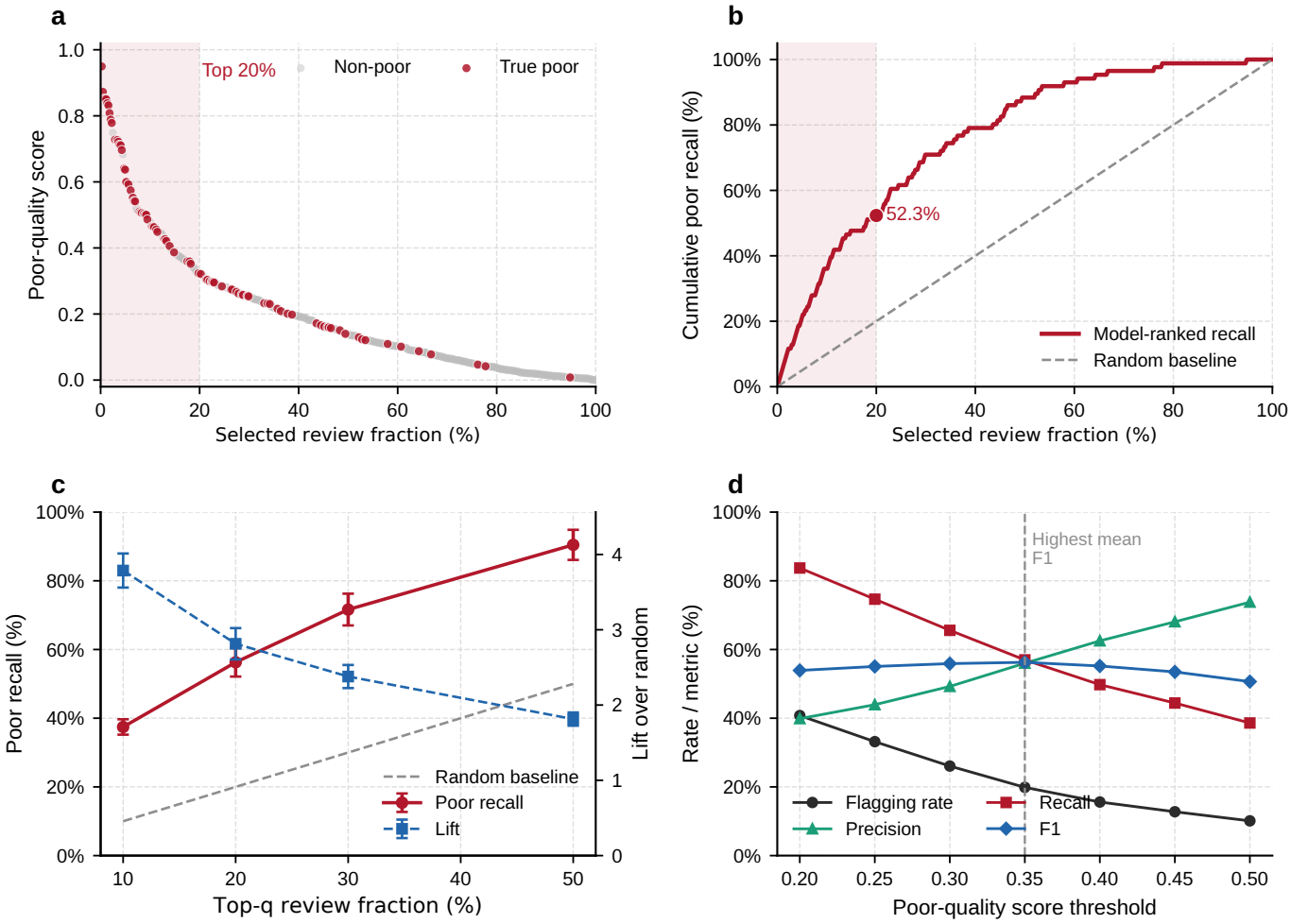


Figure 6. Retrospective prioritization of pre-rolling process-plan review using the poor-quality score. (**a**,**b**) Ranked scores and cumulative recall for one test split. The selected fraction denotes a hypothetical review budget. (**c**) Mean recall and lift across repeated splits for different review fractions; error bars show repeated-split variation. (**d**) Flagging rate, precision, recall, and F1-score versus threshold. The marked threshold has the highest mean F1 among those evaluated in this retrospective analysis. The targets are the TRF-RMS-based quality grades.

3.4. Component-Level Interpretation

Several pre-rolling descriptors share the same underlying dimensions. The correlation matrix in Supplementary Figure S1 summarizes the resulting associations among inputs and the redundancy within the feature representation.

Shapley additive explanations (SHAP) provide feature-attribution values for model predictions [28]. Tree-based SHAP analyses have also been developed to interpret tree-ensemble predictions [29]. In this study, SHAP values were calculated using an auxiliary CatBoost model with the same FS5 variables. CatBoost was fitted separately for each hold-out split, and SHAP values were computed for the corresponding test records. The weighting settings are stated in Supplementary Table S3. This analysis examines a component model family used in PI-HStack, rather than the complete stacked prediction function.

The auxiliary CatBoost analysis identifies target outer and inner diameters, prescribed rolling time, blank dimensions, and geometric ratios among the influential variables (Figure 7). Their dependence patterns are nonlinear, particularly for the target diameters and prescribed rolling time (Figure 8). These attributions concern the auxiliary CatBoost

output, not the final PI-HStack score. They describe fitted associations and do not identify causal effects or parameter settings that would improve production quality.

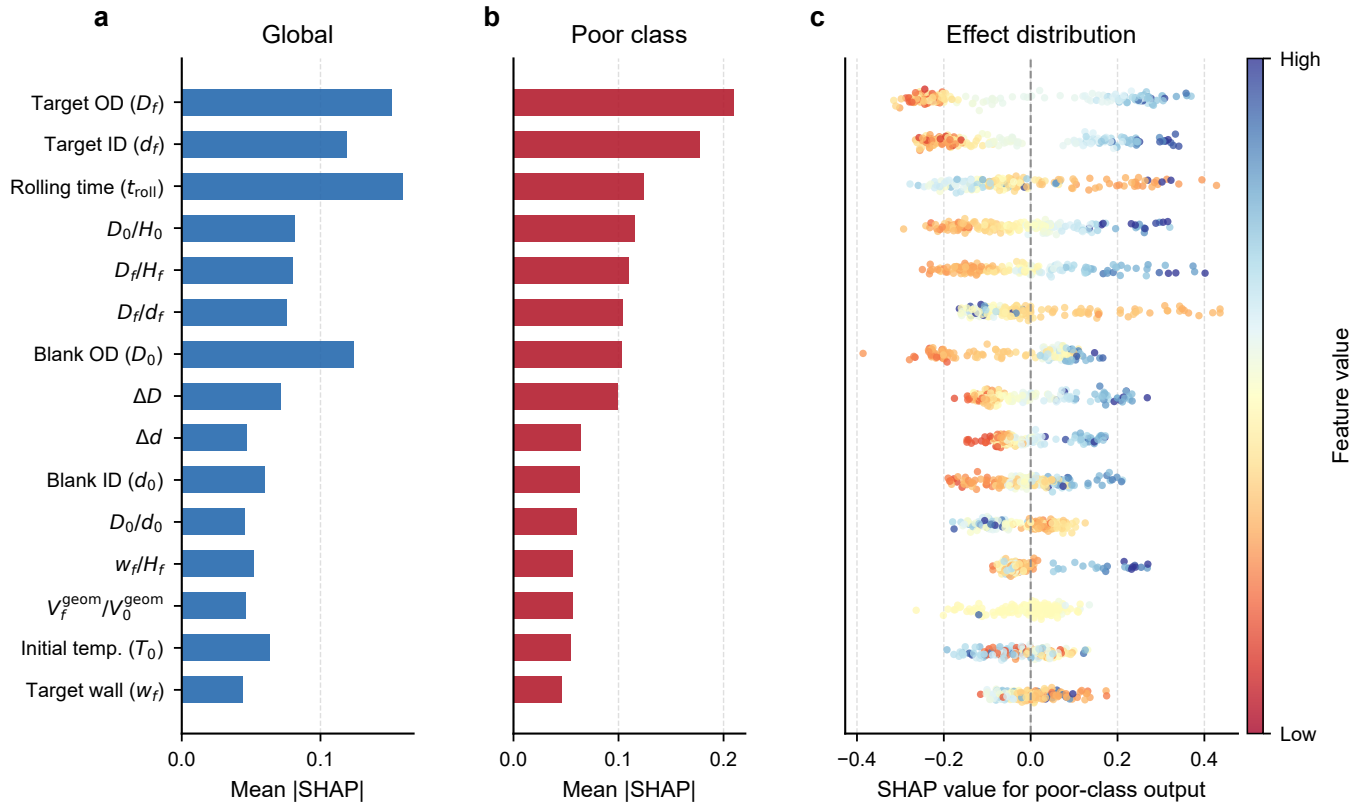


Figure 7. Auxiliary CatBoost interpretation using Shapley additive explanations (SHAP). (**a**) Mean absolute attribution across the three classes. (**b**) Mean absolute attribution for the poor class. (**c**) Poor-class attribution distribution, with color indicating normalized feature values. Outer diameter (OD), inner diameter (ID), and temperature (temp.) are abbreviated in the axis labels. Symbols are defined in Section 2.3 and Supplementary Table S2. Attributions refer to the auxiliary CatBoost fits described in the text.

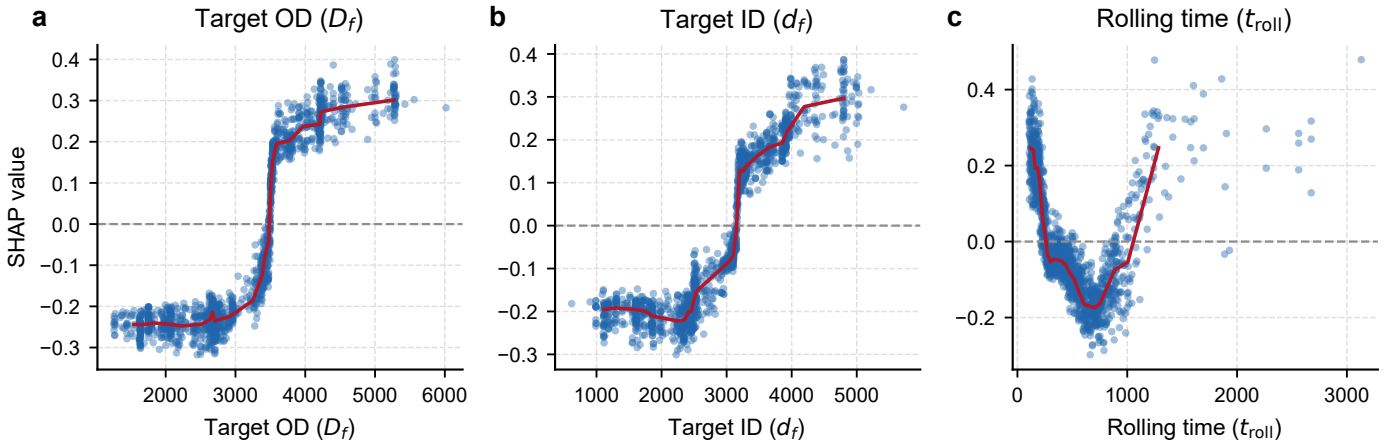


Figure 8. SHAP dependence in the auxiliary CatBoost analysis: (**a**) target outer diameter (OD); (**b**) target inner diameter (ID); (**c**) prescribed rolling time. Red lines show binned median attributions for the poor class. Diameters are expressed in millimeters and prescribed time in seconds. Samples pooled from repeated tests may recur.

4. Discussion

4.1. Forming-Quality Assessment from Industrial Process Records

The scientific contribution concerns the predictability of an operational terminal-quality grade from information fixed before rolling. Linking each historical follower-roll response to its planning record creates supervision at the workpiece level, while excluding the terminal response from new predictions preserves the planning-stage information boundary. The observed performance shows that those planning variables discriminate the TRF-RMS-based grades within this production dataset. This is a different assessment task from quality classification using rolling-process time series and independent outer-shape measurements, as discussed in the Introduction.

The technical contribution is the combination of a process-derived target, an explicit engineering representation, and a prediction-and-ranking evaluation matched to plan review. These elements address target availability, permitted inputs, and the intended decision in one framework. The experiments then qualify that framework: additional engineering descriptors do not provide uniform gains, and overall grade assignment and poor-grade screening favor different configurations. Thus, the findings identify useful information and selection criteria, rather than relying on a single headline accuracy. For rings with rotationally symmetric design geometry, the output is an early indication of an unfavorable terminal radial response, not certification of the full geometric state. A similar formulation could be examined at another site only after defining and validating an appropriate local quality target.

4.2. Engineering Interpretation of Pre-Rolling Descriptors

The feature comparisons distinguish the usefulness of an engineering representation from the amount of independently measured information. Diameter changes describe prescribed expansion, wall-thickness and height changes describe cross-sectional reduction, and geometric ratios express shape proportions. These are relevant aspects of the planned annular transformation, but several are determined by the same dimensions: for example, $\Delta w = (\Delta d - \Delta D)/2$. Geometric-volume proxies likewise follow from the idealized blank and target geometry. Their role is to expose relationships to a finite-data learner, not to add strain histories or measured material loss. This algebraic dependence provides a basis for examining the incremental value of explicit geometric descriptors.

Two observations follow from Table 7. First, FS1 reaches Macro-AUC 0.8305 compared with 0.8316 for FS5, and has higher Recall@20% (0.5953 versus 0.5628). Basic blank and target dimensions therefore provide a meaningful predictive reference, including for limited-budget screening. Second, FS2–FS4 do not improve performance monotonically, whereas jointly adding initial temperature and prescribed time to FS4 yields the highest mean overall metrics in FS5. This supports testing engineering descriptors for the intended objective instead of assuming that more descriptors are always beneficial. Supplementary Figure S1 describes their shared correlations, but does not establish why a metric changes. The comparison neither isolates the individual effects of temperature and time nor compares derived geometry at an identical set of original process inputs.

The auxiliary CatBoost attributions provide a practical focus for review. Target dimensions are product requirements; their importance directs attention to the compatibility of the proposed blank and process conditions with the required product, rather than to changing that product requirement. The nonlinear association with prescribed rolling time motivates examining duration together with geometry rather than assigning a universally favorable direction of change. These associations identify aspects of a plan to inspect using engineering procedures. They do not prescribe a tested corrective setting or establish causal effects of the process variables.

4.3. Model Choice and Pre-Rolling Process-Plan Review

PI-HStack was retained for the overall three-grade task because it provides a common framework for engineering representation, learned fusion, and class-sensitive decisions, and achieves the highest reported mean accuracy, Macro-F1, and Macro-AUC in the main comparison (Table 5). The margins over XGBoost are modest: 0.6769 versus 0.6674, 0.6436 versus 0.6427, and 0.8316 versus 0.8270, respectively. This is a task-specific empirical basis for selecting the configuration, not a statistically established or universal advantage. The design rationale and the comparative results together justify examining PI-HStack for overall grading without making heterogeneous stacking a prerequisite for the problem.

The class-decision objective helps explain why grading and screening must be judged separately. Offsets are selected by Macro-F1 across all three grades, whereas limited-budget review depends on the ordering of poor-grade scores. At a 20% review fraction, XGBoost has higher mean poor-grade coverage than PI-HStack (0.5884 versus 0.5628), despite its lower overall accuracy. A site prioritizing poor-grade screening would therefore have a reason to prefer the simpler XGBoost model under these reported results. Within the feature comparison, FS1 similarly has higher poor-grade coverage than FS5. The practical selection criterion is therefore the performance measure corresponding to the intended review action.

Table 6 shows higher accuracy and Macro-F1 for the adjusted four-learner configuration than for the unadjusted configuration, with a much smaller Macro-AUC difference. That pattern is consistent with the intended role of grade adjustment, but it is a configuration comparison rather than an isolated component-effect estimate. A single learner with matched decision adjustment has not been evaluated, and the logit representation has mixed effects across metrics. These results characterize the tested configurations and support task-level model selection without isolating the contribution of each component. Its multiple stored fold models also require consideration in deployment; inference cost has not been measured.

For engineering use, a high poor-grade score identifies a plan for closer examination, not an automatic rejection. A fixed top-q fraction makes the review workload explicit, while Recall@q measures the historical poor-grade coverage of that workload. PI-HStack's 2.81-fold lift at 20% is relative to random selection. Together with the remaining errors and the XGBoost comparison, it supports prioritizing plans for engineering review rather than assigning acceptance decisions or guaranteed corrective actions.

4.4. Scope and Limitations

The evaluation concerns the observed 2219 aluminum alloy production distribution. The 2224 records include 1029 distinct target-geometry combinations, and random stratified splitting allows recurring or similar specifications in both training and test sets. Neither repeated hold-out testing nor the reported cross-validation establishes generalization to unseen specifications, other machines or alloys, or future production shifts. Grouped, chronological, and external validation address those different questions.

The process-derived target has not been independently validated against complete circumferential profiles. The enterprise boundaries are site-specific operational criteria rather than universal geometric tolerances. Comparison with traceable independent geometric measurements, where available, is needed to quantify correspondence with actual geometric quality. More complete documentation of the historical terminal-window and signal-processing settings would also improve label traceability.

The production route fixes the process stage represented by the dataset: the workpieces enter ring rolling as forged 2219 aluminum alloy ring blanks, while rings accepted for further processing undergo solution treatment and subsequent processing to the T8

temper. Workpiece-specific metallographic observations and hardness measurements were not retained in the historical database and are not included in the predictor set. Consequently, the present analysis does not quantify possible workpiece-to-workpiece variation in pre-rolling microstructure or hardness. Initial temperature is a planning variable and should not be interpreted as a substitute for direct metallographic or hardness characterization. Further analysis of how initial microstructure and material formability relate to the terminal process-based grade would require workpiece-matched metallographic and hardness measurements as well as independently measured formability data, which are not available in the present industrial records.

The review analysis is retrospective. Prospective evaluation is needed to assess operational thresholds, inference requirements, and the outcomes of model-assisted plan review. The poor-class score is not a calibrated probability of rejection or remelting, and this study does not measure gains in tolerance compliance, scrap reduction, or production efficiency. Parameter optimization and quality-feedback control would require intervention evidence beyond the prediction task examined here.

5. Conclusions

This study established a workpiece-level route from retained industrial process signals to pre-rolling forming-quality assessment in 2219 aluminum alloy ring rolling. TRF-RMS and enterprise rules define historical supervision, while a new workpiece is assessed only from its prescribed geometry and process conditions. PI-HStack implements this separation through engineering feature representation, OOF probability fusion, and class-sensitive grade assignment. The contribution is a defined and evaluated planning-stage assessment framework under incomplete independent geometric inspection.

Across five stratified hold-out tests on 2224 records, PI-HStack achieved mean accuracy, Macro-F1, and Macro-AUC of 0.6769, 0.6436, and 0.8316, respectively. The comparisons yield two practical findings. Basic geometry already carries substantial predictive information, and extending the representation does not improve all objectives. Furthermore, the highest reported overall grading metrics and the best poor-grade screening metrics occur in different configurations: PI-HStack leads on the former in the main comparison, while XGBoost leads on poor-class F1 and Recall@20%. Model and feature selection should therefore be tied to the specific quality-assessment decision.

Ranking by the PI-HStack poor-grade score captured 56.28% of poor-grade records within the highest-ranked 20% of test records, quantifying coverage at a limited review budget. The demonstrated outcome is retrospective prediction and ranking of the process-based TRF-RMS grade within the observed production distribution. Independent geometric-tolerance validation, generalization to unseen specifications and future shifts, and production benefits from corrective intervention remain to be established.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/sym18101725/s1: Section S1, Repeated-Hold-Out Summaries (Table S1); Section S2, Pre-Rolling Descriptor Definitions (Table S2); Section S3, Learner and Training Settings (Table S3); Section S4, Evaluation Workflow (Algorithm S1); Section S5, Label-Boundary Sensitivity Procedure; Section S6, Descriptor Correlations (Figure S1); and Supplementary Code S1 (`PI_HStack_All_in_One.py`).

Author Contributions: Conceptualization, J.L. and J.T.; methodology, J.L.; software, J.L.; validation, J.L., R.X., Z.L., X.J., J.H., X.H. and L.Z.; formal analysis, J.L., R.X. and Z.L.; investigation, J.L., X.J. and J.H.; resources, J.T., X.J., L.Z. and L.L.; data curation, J.L., R.X., Z.L., J.H. and L.Z.; writing—original draft preparation, J.L.; writing—review and editing, R.X., Z.L., X.J., J.T. and L.L.; visualization, J.L., Z.L. and X.H.; supervision, J.T. and L.L.; project administration, J.T. and L.L.; funding acquisition, J.L. and L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 12602107, and the Hunan Province Graduate Research Innovation Project, grant number 1053320220390. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Data Availability Statement: The industrial production records contain confidential information on product specifications, process conditions, and enterprise manufacturing practice and are therefore not publicly available. Aggregated results, descriptor definitions, model configurations, and the evaluation workflow are reported in the article and Supplementary Materials. The code is provided as Supplementary Code S1 (`PI_HStack_All_in_One.py`). Access to the underlying confidential production records remains subject to the data owner's permission and applicable confidentiality requirements.

Acknowledgments: During preparation of the manuscript, the authors used ChatGPT (GPT-6; OpenAI) for language editing, manuscript-organization support, and LaTeX-formatting assistance. All scientific content, analyses, interpretations, and conclusions were checked and approved by the authors, who take full responsibility for the manuscript.

Conflicts of Interest: Liang Zhou is employed by Hunan Zhongchuang Aerospace New Materials Co., Ltd., which provided access to the industrial production records and contributed as stated in the Author Contributions section. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

TRF-RMS	Terminal radial-fluctuation root-mean-square
PI-HStack	Physics-informed heterogeneous stacking
PE-HStack	Probability-ensemble heterogeneous stacking
FS	Feature set
OOF	Out-of-fold
TabPFN	Tabular Prior-data Fitted Network
RF	Random forest
XGBoost	Extreme gradient boosting
MLP	Multilayer perceptron
Macro-F1	Macro-averaged F1-score
Macro-AUC	Macro-averaged one-vs-rest area under the receiver operating characteristic curve
ROC	Receiver operating characteristic
AUC	Area under the receiver operating characteristic curve
AP	Average precision
SHAP	Shapley additive explanations
OD/ID	Outer diameter/inner diameter
R@20%	Poor-class recall within the highest-ranked 20% of workpieces

References

1. Jenkouk, V.; Hirt, G.; Franzke, M.; Zhang, T. Finite element analysis of the ring rolling process with integrated closed-loop control. *CIRP Ann.* **2012**, *61*, 267–270. [CrossRef]
2. Allwood, J.M.; Tekkaya, A.E.; Stanistreet, T.F. The development of ring rolling technology. *Steel Res. Int.* **2005**, *76*, 111–120. [CrossRef]
3. Allwood, J.M.; Tekkaya, A.E.; Stanistreet, T.F. The development of ring rolling technology–Part 2: Investigation of process behaviour and production equipment. *Steel Res. Int.* **2005**, *76*, 491–507. [CrossRef]
4. Hua, L.; Zhao, Z.Z. The extremum parameters in ring rolling. *J. Mater. Process. Technol.* **1997**, *69*, 273–276. [CrossRef]
5. Anjami, N.; Basti, A. Investigation of rolls size effects on hot ring rolling process by coupled thermo-mechanical 3D-FEA. *J. Mater. Process. Technol.* **2010**, *210*, 1364–1377. [CrossRef]
6. Zhou, J.; Wang, F.L.; Wang, M.H.; Xu, W.J. Study on forming defects in the rolling process of large aluminum alloy ring via adaptive controlled simulation. *Int. J. Adv. Manuf. Technol.* **2011**, *55*, 95–106. [CrossRef]

7. Hua, L.; Deng, J.D.; Qian, D.S.; Lan, J.; Long, H. Modeling and application of ring stiffness condition for radial-axial ring rolling. *Int. J. Mach. Tools Manuf.* **2016**, *110*, 66–79. [CrossRef]
8. Guo, L.; Yang, H. Towards a steady forming condition for radial-axial ring rolling. *Int. J. Mech. Sci.* **2011**, *53*, 286–299. [CrossRef]
9. Berti, G.A.; Quagliato, L.; Monti, M. Set-up of radial-axial ring-rolling process: Process worksheet and ring geometry expansion prediction. *Int. J. Mech. Sci.* **2015**, *99*, 58–71. [CrossRef]
10. Lafarge, R.; Hütter, S.; Tulke, M.; Halle, T.; Brosius, A. Data based model predictive control for ring rolling. *Prod. Eng.* **2021**, *15*, 821–831. [CrossRef]
11. Zhai, R.; Jiao, Z.; Han, Y.; Zhao, J.; Yu, G. Intelligent Control System of Internal Expansion over Bending and Calibration (IEOBC) Process for Large Pipe Ends. *Symmetry* **2021**, *13*, 1618. [CrossRef]
12. Tao, F.; Qi, Q.; Liu, A.; Kusiak, A. Data-driven smart manufacturing. *J. Manuf. Syst.* **2018**, *48*, 157–169. [CrossRef]
13. Tercan, H.; Meisen, T. Machine learning and deep learning based predictive quality in manufacturing: A systematic review. *J. Intell. Manuf.* **2022**, *33*, 1879–1905. [CrossRef]
14. Cao, J.; Bambach, M.; Merklein, M.; Mozaffar, M.; Xue, T. Artificial intelligence in metal forming. *CIRP Ann.* **2024**, *73*, 561–587. [CrossRef]
15. Fahle, S.; Glaser, T.; Kneißler, A.; Kuhlenkötter, B. Improving quality prediction in radial-axial ring rolling using a semi-supervised approach and generative adversarial networks for synthetic data generation. *Prod. Eng.* **2022**, *16*, 175–185. [CrossRef]
16. Seitz, J.; Moser, T.; Fahle, S.; Prinz, C.; Kuhlenkötter, B. Transfer learning approaches in the domain of radial-axial ring rolling for machine learning applications. In *Proceedings of the 5th Conference on Production Systems and Logistics (CPSL 2023)*; Herberger, D., Hübner, M., Eds.; publish-Ing: Hannover, Germany, 2023; pp. 447–458. [CrossRef] [PubMed]
17. Wang, Q.; Seitz, J.; Lafarge, R.; Kuhlenkötter, B.; Brosius, A. Complementary database generation for machine learning in quality prediction of cold ring rolling. In *Proceedings of the 29. Sächsische Fachtagung Umformtechnik*; Technische Universität Dresden: Dresden, Germany, 2023; Article No. 06. [CrossRef]
18. Xie, J.; Sun, L.; Zhao, Y.F. On the data quality and imbalance in machine learning-based design and manufacturing–A systematic review. *Engineering* **2025**, *45*, 105–131. [CrossRef]
19. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [CrossRef]
20. Wang, X.; Hua, L. Modeling of on-line measurement for rolling the rings with blank size errors in vertical hot ring rolling process. *Int. J. Adv. Manuf. Technol.* **2013**, *68*, 257–262. [CrossRef]
21. Hollmann, N.; Müller, S.; Purucker, L.; Krishnakumar, A.; Körfer, M.; Hoo, S.B.; Schirrmeister, R.T.; Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature* **2025**, *637*, 319–326. [CrossRef] [PubMed]
22. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31, pp. 6638–6648.
23. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
24. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
25. Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; Babenko, A. Revisiting deep learning models for tabular data. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 18932–18943.
26. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 507–520. [CrossRef]
27. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
28. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 4765–4774.
29. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.