Journal of Intelligent Manufacturing
https://doi.org/10.1007/s10845-026-02834-2

A scalable random forest (SRF) approach for non-linear predictive modelling using small manufacturing datasets

Meshari A. Al-Ebrahim[1] · **Rajesh S. Ransing**[2]

Received: 25 April 2025 / Accepted: 17 January 2026

Abstract

This paper presents an integrated, scalable Random Forest (SRF)–based predictive framework for estimating the effects of process interventions, including (i) adjusting operating ranges for continuous process parameters within specified tolerances, (ii) selecting specific categories for discrete process parameters, and (iii) combining adjustments to both continuous and discrete parameters. The framework moves beyond linear assumptions by employing a non-linear ensemble approach to identify critical process inputs and quantify their contributions to predicting the process response. These contributions are then leveraged to derive optimal operating ranges for continuous parameters and optimal categories for discrete parameters through a Decision Path Search (DPS) procedure based on tree decision paths. The proposed framework scales to a large number of process factors with complex non-linear dependencies and enables data-driven process improvement. Missing values in mixed-type datasets are addressed using an iterative Random Forest–based imputation scheme, while automatic forest-size optimisation enhances model stability. All preprocessing and modelling steps are embedded within a leakage-safe pipeline, supported by learning-curve analysis and leakage-sanity diagnostics to guard against overfitting. Across the evaluated case studies, SRF delivers accurate predictions together with transparent, practitioner-ready operating windows, translating complex mixed-type manufacturing data into actionable guidance.

Keywords Random forest · Common-cause variation · Predictive analytics · Data augmentation · Small data · Quality improvement

Introduction

Manufacturing processes exhibit two broad sources of variation: common-cause variation arising from numerous, often uncontrollable influences, and assignable-cause variation associated with specific, identifiable disturbances. Within Statistical Process Control (SPC), critical-to-quality (CTQ) characteristics are monitored against upper and lower specification limits (USL/LSL), enabling routine adjustments under common-cause conditions while signalling intervention when assignable causes arise (Montgomery, 2009). Modern quality improvement therefore prioritises not only accurate prediction but also actionable factor settings that plausibly influence responses and can be adjusted to reduce defects and stabilise CTQs.

In production data, causal analysis is complicated by non-linear responses, mixed continuous–categorical features, and missing data–often under small dataset conditions. Non-linearity may include thresholds and interactions that vary with other factor levels (Ransing et al., 2013). Investment casting provides a representative case: in nickel-based superalloys, shrinkage penalties vary non-linearly with elemental composition (e.g., carbon, titanium, cobalt), with behaviour dependent on their joint levels (Batbooti, 2023; Giannetti and Ransing, 2016). These characteristics hinder linear correlation analyses and motivate approaches that

Meshari A. Al-Ebrahim and Rajesh S. Ransing have contributed equally to this work.

✉ Meshari A. Al-Ebrahim
meshari.alebrahim@gmail.com

✉ Rajesh S. Ransing
r.s.ransing@swansea.ac.uk

[1] Department of Technical Support, State Audit Bureau (SAB), Shuwaikh, Kuwait

[2] Zienkiewicz Institute for Modelling, Data and AI, Department of Mechanical Engineering, Swansea University, Swansea, United Kingdom

Published online: 01 June 2026

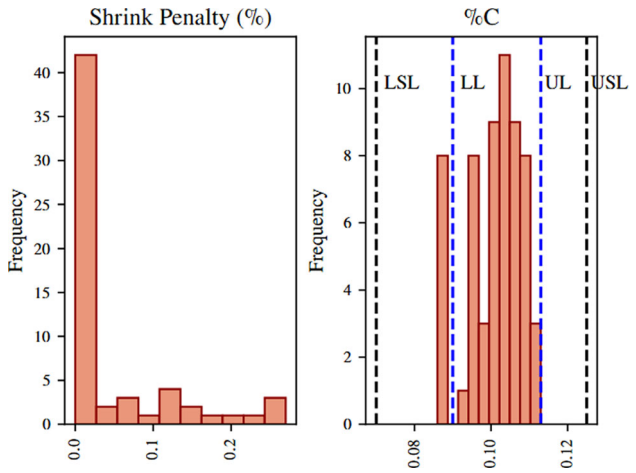


Fig. 1 Variation in rejection rate and in the input factor %C for a nickel-based superalloy (Batbooti, 2023)

capture complex effects while remaining stable under small, mixed-type datasets.

For instance, carbon content may exhibit ranges that minimise shrinkage defects, whereas higher or lower values increase defect likelihood in ways that depend on other alloying elements. Figures 1, 2 and 3 illustrate these effects for investment casting, highlighting skewed coverage, non-linear trends, and data sparsity that complicate modelling and increase overfitting risk under small datasets.

Random Forests (RF) provide strong non-parametric baselines for tabular prediction (Breiman, 2001), yet stability can degrade under small, heterogeneous datasets unless preprocessing, imputation, augmentation, and tuning are handled in a leakage-safe and variance-aware manner. Manufacturing data frequently presents two simultaneous challenges: extremely small datasets and strong non-linear interactions among process factors, as illustrated in Figure 3, where threshold effects cannot be captured by linear screening tools. These characteristics motivate the need for predictive models that operate reliably under mixed-type, limited, and non-linear conditions while still producing interpretable process insights.

To verify robustness and interpretability, this study evaluates model behaviour using established post-hoc explanation tools–SHAP (Shapley Additive Explanations) (Lundberg, 2017), LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro and Singh, 2016), Partial Dependence (PD) curves (Friedman, 2001), and Individual Conditional Expectation (ICE) plots (Goldstein et al., 2015)–to compare internal feature effects against well-known explainability baselines. Furthermore, the study incorporates learning-curve analysis, leakage-sanity diagnostics (Sasse et al., 2025), forest-size optimisation, and a compact ablation study to assess variance, stability, and the contribution of each component. Beyond prediction, the framework extracts operating windows by deriving optimal and avoidance ranges for each factor through decision-path aggregation, enabling structured interpretation of factor settings that improve or degrade the response. These considerations justify the development of the proposed scalable Random Forest (SRF) framework and set the foundation for the study's objectives.

Aims and objectives

Manufacturing datasets often combine continuous and categorical factors, include missing data, and remain small due to cost and logistical constraints. These characteristics make it challenging to model non-linear interactions, define actionable factor settings, and obtain reliable estimates of uncertainty using conventional analytical tools

%Carbon

Q1	Q2	Q3	Q4	
Minimum		Median		Maximum
0.086	0.095	0.103	0.106	0.113

Q1: Avoid; Range: Bottom 25%, {>=0.086 & <=0.095}

Penalty	Q1	Q2	Q3	Q4
0.8-1	13	3		2
0.6-0.8			1	3
0.4-0.6				
0.2-0.4				
0-0.2	3	13	9	13

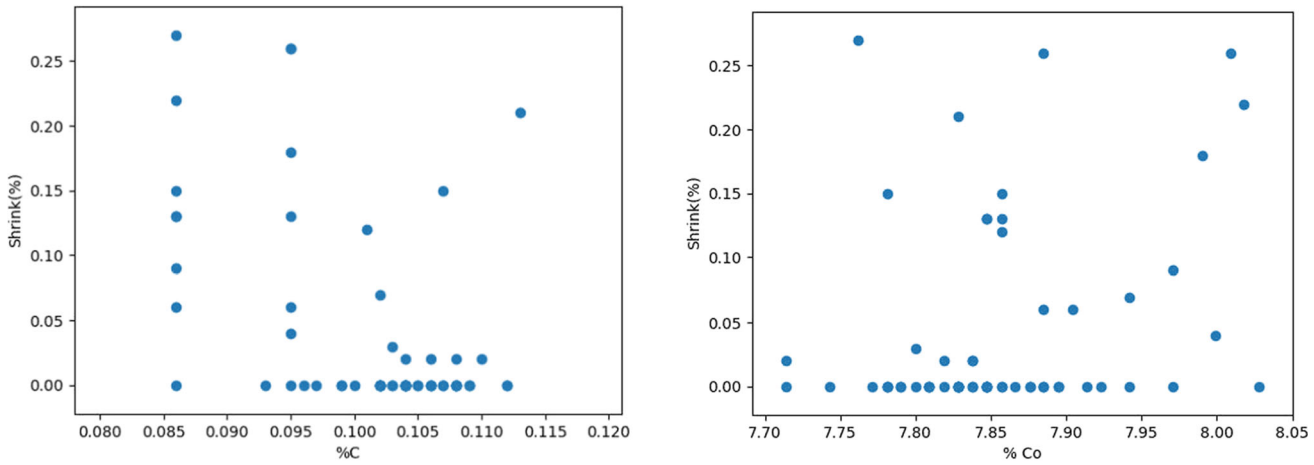
% Cobalt

Q1	Q2	Q3	Q4	
Minimum		Median		Maximum
7.714	7.809	7.847	7.885	8.028

Q3 & Q4: Avoid; Range: Top 50%, {>7.847 & <=8.028};

Penalty	Q1	Q2	Q3	Q4
0.8-1	3	3	3	9
0.6-0.8	1	3		
0.4-0.6				
0.2-0.4				
0-0.2	12	12	5	9

Fig. 2 Penalty matrices for carbon and cobalt composition (Ransing et al., 2013)



(a) Carbon variation (%C). (b) Cobalt variation (%Co).

Fig. 3 Shrinkage variability versus carbon and cobalt composition

(Breiman, 2001). To address these challenges, this study introduces an SRF framework designed to provide robust prediction, interpretable factor limits, and variance-aware evaluation for mixed and limited manufacturing datasets. Specifically, the study pursues six objectives: (i) specify a leakage-safe mixed-type pipeline (Sasse et al., 2025); (ii) formulate and assess a classification-assisted augmentation protocol (response quantile binning → SMOTE in feature space → missForest imputation of masked responses) with guidance on risks and when not to use it (Chawla et al., 2002; Stekhoven and Bühlmann, 2012); (iii) formalise a Decision Path Search (DPS) procedure that aggregates tree paths into optimal and avoidance ranges as candidate operating windows; (iv) adopt robust reporting (RMSE/NRMSE/MAE/R^2 as mean±std with 95% confidence intervals and paired t-tests across folds); (v) integrate learning-curve analysis, a leakage-sanity check, and forest-size optimisation with seed-stability auditing; and (vi) benchmark against strong tabular baselines (RF, XGBoost, LightGBM, CatBoost) on six datasets, including a nickel-based superalloy case.

Contributions, scope, and limitations

The study develops and evaluates an SRF-based framework for small, mixed-type manufacturing data along four axes: (1) an end-to-end, fold-contained pipeline that prevents information leakage and stabilises estimation under small dataset conditions; (2) a small-data augmentation protocol (SMOTE→missForest) tailored to mixed-type regression with explicit risk guidance; (3) a decision-path mechanism (DPS) that converts model structure into interpretable optimal and avoidance ranges as testable operating windows; and (4) an evaluation suite combining uncertainty quantification, significance testing, learning curves, leakage-sanity checks, and forest-size optimisation with seed-stability auditing, compared against strong boosting baselines.

The scope is limited to small, mixed-type, observational manufacturing datasets where non-linear responses and factor interactions are expected. Limitations are explicit: causal interpretations are hypothesis-generating rather than identified; augmentation fidelity may degrade under strong interactions or rare categories; DPS intervals can shift with seeds, folds, or mild hyperparameter changes (stability is quantified); external validity beyond the six datasets should be established case-by-case; and suggested operating windows must be reconciled with existing USL/LSL and process safety constraints before deployment (Puthanveettil Madathil et al., 2025).

Key contributions. (i) A fold-contained, leakage-safe pipeline for small, mixed-type manufacturing data; (ii) a classification-assisted augmentation protocol for regression (**SMOTE** on feature neighbourhoods after response quantile binning, followed by **mask-Y** and **missForest** imputation) with explicit when-not-to-use guidance; (iii) **Decision Path Search (DPS)** that aggregates tree paths to output optimal and avoidance operating ranges aligned with USL/LSL thinking; (iv) a robustness suite (95% confidence intervals, paired t/Wilcoxon tests, learning curves, leakage-sanity checks, and seed/forest-size stability) reported consistently across six datasets.

Paper organisation

Section Related work provides a comprehensive review of related work. Section Methodology (SRF) introduces the SRF methodology in detail. Section Experimental setup outlines the datasets, baseline models, and evaluation protocols. Section Results reports comparative results with uncertainty, robustness diagnostics, and interpretability analyses. Section Ablation study presents the ablation study. Section Discussion discusses practical implications for industrial adoption and limitations. Section Conclusion concludes the paper.

Related work

Research on predictive modelling and causal analysis for manufacturing tabular data spans four intertwined streams. First, ensemble methods for tabular learning–particularly Random Forests (Breiman, 2001) and modern gradient-boosting libraries (XGBoost (Chen, 2016), LightGBM (Ke et al., 2017), CatBoost (Prokhorenkova et al., 2018))–provide strong non-parametric baselines with competitive accuracy and modest feature engineering. Second, tabular deep learning and AutoML frameworks (e.g., TabNet (Arik and Pfister, 2021), auto-sklearn (Feurer et al., 2015), AutoGluon (Erickson et al., 2020)) automate pipeline search and representation learning, although performance can be sensitive under small dataset, mixed-type, and missing-data conditions. Third, mixed-type imputation methods (missForest (Stekhoven and Bühlmann, 2012); Known Data Regression (KDR) (Batbooti, 2023); Factor Analysis of Mixed Data (FAMD) (Lê et al., 2008); and Two-Stage Regression (TSR) (Serneels et al., 2005)) address pervasive missing data in historical production records, each with distinct assumptions and small-dataset behaviour. Fourth, post-hoc explainability methods such as SHAP (Lundberg, 2017), LIME (Ribeiro and Singh, 2016), and PD/ICE (Friedman, 2001; Goldstein et al., 2015) clarify feature influence but typically stop short of producing actionable operating windows. Recent work in the *Journal of Intelligent Manufacturing (JIMS)* emphasises trustworthy, deployable quality analytics through explainable defect inspection pipelines and adaptive, AI-driven quality control in production environments (Bordekar et al., 2025; Liu et al., 2025). A compact contrast of commonly used baselines is shown in Table 1.

Ensembles for tabular manufacturing data

Ensemble tree methods are widely adopted for tabular prediction in manufacturing because they accommodate non-linear responses and heterogeneous feature spaces with modest feature engineering. Random Forests (RF) aggregate decorrelated decision trees to reduce variance and improve generalisation (Breiman, 2001; Ho, 1995). Recent directions include correlation control among trees (Sun et al., 2024), stratified sampling for feature selection (Jain, 2023), and robust tuning to stabilise performance (Liao et al., 2024); evidential trees/-forests emphasise uncertainty handling for noisy industrial data (Hoarau et al., 2023). Applications span inspection and defect identification (Dugalam, 2024), variable selection, and rule extraction for anomaly explanation (Kopp and Pevný, 2020; Speiser et al., 2019). Gradient-boosting libraries provide strong baselines alongside RF: XGBoost and LightGBM offer efficient, regularised boosting with histogram-based splits (Chen, 2016; Ke et al., 2017), and CatBoost introduces ordered-statistics encoding to natively manage categoricals and missing values (Prokhorenkova et al., 2018). Under small and heterogeneous datasets, performance is sensitive to leakage-safe preprocessing and variance controls; CatBoost mitigates target leakage via ordered boosting, while RF stability benefits from fold-contained preprocessing and tuned forest size (Probst et al., 2019).

Tabular deep learning and AutoML

Tabular deep learning aims to learn task-specific representations directly from mixed-type features. Architectures such as TabNet use sequential attention and feature selection to encourage interpretability and compact embeddings (Arik and Pfister, 2021). Under small, mixed-type, and incomplete datasets, performance can be sensitive to regularisation, imputation, and training choices; empirical studies often find strong tree ensembles competitive when data are limited (Grinsztajn et al., 2022; Shwartz-Ziv and Armon, 2022). AutoML frameworks (auto-sklearn; AutoGluon) automate pipeline/model selection but still require leakage-safe evaluation and uncertainty reporting (Erickson et al., 2020; Feurer et al., 2015). From a process-improvement perspective, both tabular deep learning and AutoML typically yield point predictions or attribution summaries rather than factor-level optimal/avoidance ranges; hence, Table 1 treats them as baselines while methodological development in this study focuses on an RF backbone with DPS. Transfer learning is a complementary direction when related datasets exist (Giannetti and Essien, 2022; Zhuang et al., 2021).

Mixed-type imputation

Historical manufacturing records routinely exhibit missing entries across mixed continuous–categorical feature spaces. Effective imputation should (i) accommodate non-linear dependencies, (ii) preserve the joint distribution of mixed types, and (iii) remain stable under small dataset conditions. Common evaluation metrics include normalised root mean squared error (NRMSE) for continuous variables and

Table 1 Baselines for tabular manufacturing data

Family / method	Mixed	Missing data	Small dataset	Ranges[‡]
RF (Breiman, 2001)	✓	(impute)	✓[†]	—
XGBoost (Chen, 2016)	✓	(impute)	✓	—
LightGBM (Ke et al., 2017)	✓	(impute)	✓	—
CatBoost (Prokhorenkova et al., 2018)	✓	✓	✓	—
TabNet (Arik and Pfister, 2021)	✓	(impute)	— (prone to overfit under small datasets)	—
AutoML (auto-sklearn (Feurer et al., 2015); AutoGluon (Erickson et al., 2020))	✓	(impute)	✓	—
DPS (this study)	✓	(impute)	✓	✓

[†] Stability depends on leakage-safe preprocessing, tuning, and variance controls.
[‡] 'Ranges' indicates whether the method produces factor-level operating windows (optimal/avoidance ranges), as in DPS

Table 2 Mixed-type imputation methods. Abbreviations: MAR = missing at random; MCAR = missing completely at random

Method	Mixed types	Non-linear	Assumptions	Small dataset
missForest (Stekhoven and Bühlmann, 2012)	✓	✓	MCAR/MAR	✓ (RF-based)
KDR (Batbooti, 2023)	✓	— (form-dependent)	model form; MAR	—
FAMD (Lê et al., 2008)	✓	— (latent linear)	low-rank structure	—
TSR (Serneels et al., 2005)	✓	— (stage-dependent)	sequential models	—

misclassification rate for categorical variables, computed on held-out ground truth within leakage-safe cross-validation (Little, 2002; Oba et al., 2003). A concise comparison of popular approaches appears in Table 2.

missForest. missForest is an iterative, non-parametric imputation method based on RF regressors/classifiers that supports mixed data and non-linearities (Stekhoven and Bühlmann, 2012). At each iteration, variables with missing data are imputed using RF trained on observed entries, cycling until convergence. Strengths include robust handling of interactions and minimal model specification; limitations appear with extremely imbalanced categories or when missingness deviates strongly from MCAR/MAR (Little, 2002). In leakage-safe pipelines, missForest must be fitted within each training fold.

Known Data Regression (KDR). KDR is a regression-based strategy used in prior manufacturing studies that imputes missing entries from variables observed for the same records (Batbooti, 2023). It is simple and competitive under moderate missingness but can be sensitive to model misspecification (e.g., linear forms under non-linear effects) and may require separate treatment for categorical targets.

Factor Analysis of Mixed Data (FAMD). FAMD projects mixed data to a latent space jointly modelling continuous and categorical variables before reconstructing missing entries (Lê et al., 2008). It is effective when few latent factors capture variability, but it is sensitive to outliers and to the choice of latent dimension.

Two-Stage Regression (TSR). TSR uses sequential conditional models (e.g., continuous given categorical, then categorical given continuous), iterating until convergence. By decoupling conditional sub-problems, TSR can be practical in mixed-type settings but may propagate early-stage biases under strong interactions or rare levels (Serneels et al., 2005).

Practice for small datasets. Under very small dataset conditions, imputation model variance can dominate. Fold-contained fitting, conservative hyperparameters, and uncertainty reporting (mean±std and 95% confidence intervals across folds) are recommended. Because imputation can induce distributional shift, learning-curve trends and leakage-sanity checks provide safeguards prior to deriving optimal/avoidance ranges via DPS.

Explainability vs. operating-window discovery

Post-hoc explainability methods clarify how features influence predictions but typically do not propose factor ranges suitable for industrial adjustment. This distinction is increasingly highlighted in *JIMS* work on explainable inspection and adaptive quality control, where interpretability supports trust and decision-making but does not necessarily yield actionable factor-level operating windows (Bordekar et al., 2025; Liu et al., 2025). SHAP attributes contributions via Shapley values under an additive explanation model (Lundberg, 2017); LIME fits local surrogate models to approximate

complex predictors (Ribeiro and Singh, 2016); and PD/ICE profiles visualise marginal response trends (Friedman, 2001; Goldstein et al., 2015). These tools support diagnosis and prioritisation but typically do not output factor-level operating windows.

Operating-window discovery proposes factor-level intervals (continuous) or categories (discrete) that are optimal or avoidance candidates, i.e., candidate operating ranges to be trialled in production. In this study, DPS aggregates RF decision paths: split constraints are collected and scored to form per-factor optimal/avoidance ranges across features. This path-based construction differs from attribution in that it outputs ranges rather than importance values, aligning with specification-limit thinking (USL/LSL) and set-point selection.

Graph-based causal modelling (e.g., Bayesian networks and directed acyclic graphs) offers a complementary perspective by estimating conditional dependence structure and, in principle, enabling interventional reasoning (Pearl, 2009; Spirtes and Glymour, 2000). Toolkits such as CausalNex operationalise structure learning with tabular data and domain priors, but under small, mixed-type, and incomplete datasets, structure learning can be unstable and typically yields graphs rather than direct operating windows (Quantumblack, 2019, 2020). Optimisation-driven diagnostics in rotating machinery (e.g., orthogonal matching pursuit enhanced via golden jackal optimisation) illustrate modelling directions relevant to industrial fault detection (Vashishtha et al., 2025).

Positioning vs. prior art. In contrast to correlation-oriented attributions (SHAP, LIME, PD/ICE) or graph-learning toolkits (e.g., Bayesian networks and CausalNex; see Appendix A), DPS turns model structure into operating windows by aggregating path constraints from trained trees–yielding ranges rather than importance values or directed acyclic graph edges. Relative to RF variants (e.g., extra-trees, rule distillation, evidential forests), SRF's distinct contribution lies in: (1) an end-to-end, fold-contained pipeline tailored to small, mixed-type data; (2) a regression-safe, classification-assisted augmentation with guardrails and explicit disablement on large/clean datasets; and (3) a path-aggregation $\rightarrow$ window mechanism designed for practitioner set-points. This paper does not claim causal identification: DPS windows are hypothesis-generating and are cross-checked against SHAP and PD/ICE trends on held-out folds. Multivariate methods frequently applied in manufacturing are summarised in Table 3 with applications, strengths, and limitations.

Methodology (SRF)

This section formalises an SRF framework for small, mixed-type manufacturing datasets with potential non-linear dependencies and missing data. The approach comprises: (i) a leakage-safe pipeline that confines preprocessing and model fitting within cross-validation folds; (ii) a small-data augmentation protocol for regression that couples response-quantile binning (Mujahid et al., 2024), SMOTE in feature space (Camacho and Bacao, 2024; Chawla et al., 2002), and missForest imputation of masked responses (Stekhoven and Bühlmann, 2012); (iii) automatic forest-size optimisation with a stability corridor and seed checks; (iv) Decision Path Search (DPS), which aggregates internal RF decision paths into interpretable optimal and avoidance ranges (intervals for continuous; categories for discrete) as candidate operating windows; and (v) a condensed evaluation and robustness protocol reporting mean$\pm$std, 95% confidence intervals, paired t-tests, learning curves, and leakage-sanity diagnostics. Computational considerations are summarised and briefly discussed. The RF backbone follows standard ensembles for tabular prediction (Breiman, 2001; Somvanshi et al., 2024), with tuning guidance informed by small-dataset stability practice (Probst et al., 2019).

Leakage-safe pipeline: data flow and preprocessing

All preprocessing and model-fitting steps are confined to cross-validation (CV) training folds to avoid information leakage. Within each CV split, categorical variables are encoded using training-fold statistics only; missing values are imputed via missForest fitted on the training fold and then applied to the validation fold; optional augmentation is performed strictly on the training fold; and RF is trained with forest-size optimisation under a stability corridor. Predictions and diagnostics are recorded on validation folds. Figure 4 summarises the leakage-safe flow and Algorithm 1 lists the stepwise procedure. All leakage controls follow standard fold-contained practice: encoders, imputers, augmentation steps, and hyperparameter selection are fitted exclusively on training folds and applied unchanged to validation folds.

Small-data augmentation for regression (SMOTE$\rightarrow$missForest)

A classification-assisted augmentation is employed: the continuous response is temporarily binned into quantiles to construct neighbourhoods, SMOTE synthesises feature vectors per bin, and missForest imputes masked responses. This preserves mixed types and non-linear dependencies while avoiding direct interpolation of the response. Algorithm 2 lists the steps; Figures 5, 6, 7 and 8 illustrate before/after distributions for the nickel-based superalloy case.

Table 3 Summary of multivariate data analysis methods

Technique	Application	Advantages	Disadvantages
Hotelling's T^2	1) Multivariate air quality control (Hotelling, 1947).	Real-time monitoring.	1) Not specifically designed for small datasets and assumes linear relationships.
	2) Hot forming process (Mason et al., 1995).		2) Not explicitly designed for causal discovery.
(PCA)[a] / (MFA)[b]	1) Prediction of RNA-Seq malaria vector (Arowolo and Adebiyi, 2020).	1) Easily visualise results via bi-plots.	1) Sensitive to outliers; focuses on variance; not inherently causal.
	2) Stochastic bottleneck (Giannetti et al., 2014).	2) Explains dominant factors contributing to total variance.	2) Difficult with > 3 PCs.
	3) Historical buildings deterioration (Ferrari et al., 2011).	3) Works well with small datasets via dimensionality reduction.	3) May not capture non-linear relationships.
(FDA)[c]	Health of bars (Youn et al., 2015).	Handles high-dimensional data; models non-linearities; works on small data.	Limited input size; domain expertise needed; not inherently causal.
(BBN)[d]	1) Semiconductor manufacturing (Yang and Lee, 2012).	1) Useful for causal relationship modelling.	1) Training is resource intensive.
	2) Automotive body process (Liu and Jin, 2013).	2) Captures non-linear dependencies.	2) Requires prior knowledge.
(SVM)[e]	Hyperspectral image classification (Paoletti et al., 2020).	Efficient memory use; strong in high-dimensional data; captures non-linearities.	Not for large data; normalisation needed; not inherently causal.
(ANN)[f]	1) Sheet metal costing (Verlinden et al., 2008).	Handles complex non-linear problems (Özdem and Orak, 2024).	1) Requires experience; complex to use.
	2) Cold rolled steel (Mohanty et al., 2011).		2) Does not uncover variable relationships.
	3) Construction projects (Doroshenko, 2020).		3) Requires large datasets; not inherently causal.
	4) COVID-19 US data (Mollalo et al., 2020).		4) Risk of overfitting (Tercan and Meisen, 2022).

Table 3 continued

Technique	Application	Advantages	Disadvantages
(PnT)[g]	Rule extraction / model compaction from tree ensembles (Cohen-Shapira, 2024).	Interpretable fused decision paths; compact surrogate of RF/GBDT; often preserves accuracy.	Requires a pre-trained forest; extra distillation/tuning step; limited off-the-shelf tooling.
(PM)[h]	Foundry process discovery (Ransing et al., 2013).	Reveals correlations/patterns; handles small datasets and non-linearities.	Quantitative variables only; cannot handle missing data; not causal.
(QRT)[i]	Risk-based uncertainty quantification in manufacturing (Giannetti and Ransing, 2016).	Robust for small data; captures non-linear input effects.	High variance; assumes linearity; not inherently causal.
(QCA)[j]	Nickel casting defect reduction (Batbooti et al., 2017).	Optimises process settings; small-dataset friendly.	Based on PCA linear assumptions; high computation; not causal.
(RF)[k]	Predicting student performance (Cortez, 2008).	Handles non-linear data; reduces regression variance.	Needs large datasets; not inherently causal.

[a] Principal Component Analysis
[b] Multiple Factor Analysis
[c] Fuzzy Data Analysis
[d] Bayesian Belief Network
[e] Support Vector Machine
[f] Artificial Neural Networks
[g] Path-encoded tree ensembles
[h] Penalty matrix
[i] Quantile regression tree
[j] Quality correlation algorithm
[k] Random Forest algorithm

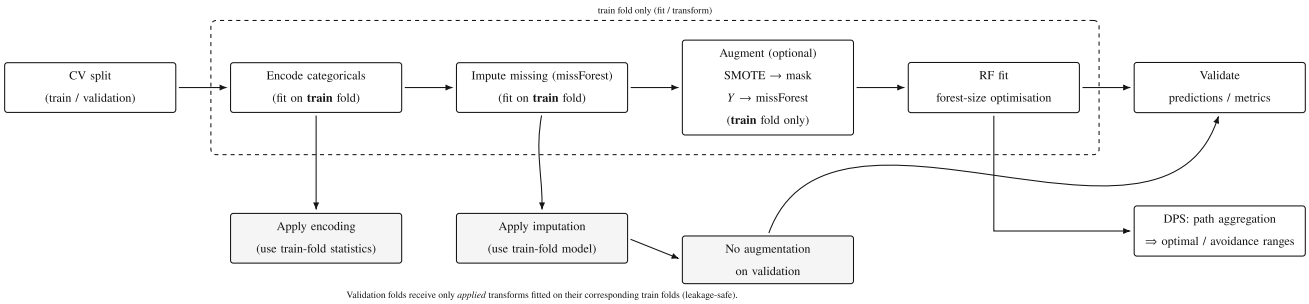
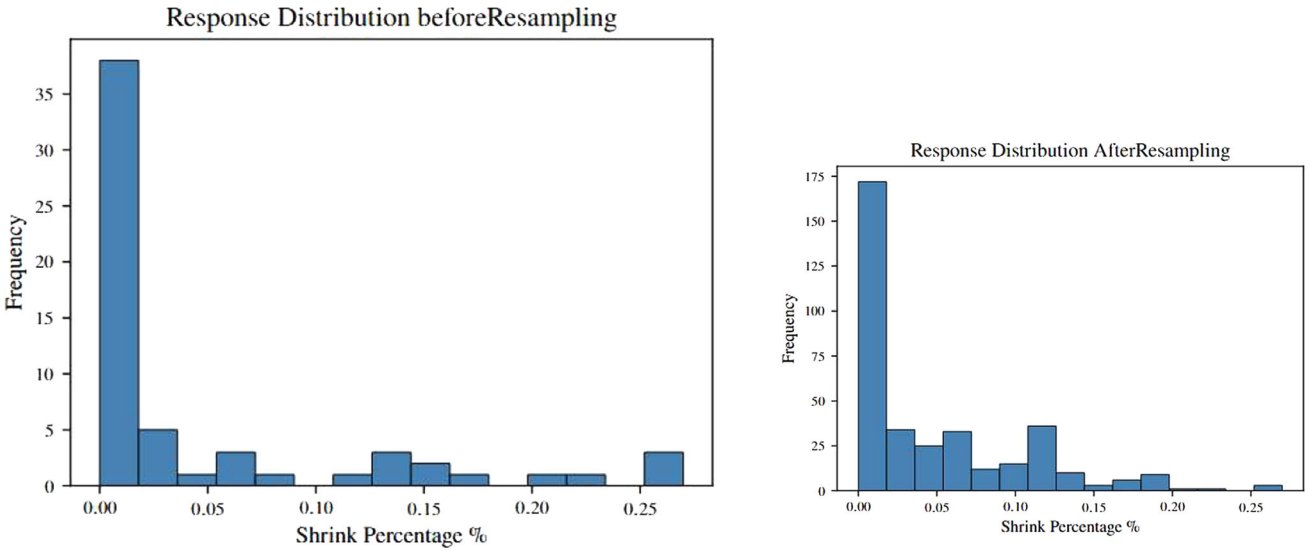


Fig. 4 Leakage-safe SRF data flow within each CV fold. All fitting (encoding, imputation, optional augmentation, and RF training) occurs on the training fold. The validation fold receives only transforms fitted on the training fold, after which predictions and metrics are computed. DPS aggregates trained-tree paths to produce optimal/avoidance ranges



(a) Raw response distribution.

(b) After augmentation.

Fig. 5 Response distribution before and after augmentation. Bin widths are shown; x-axis: Shrink Penalty (%), y-axis: Frequency. Ticks limited to $\leq$3 significant figures

Algorithm 1: SRF pipeline (fold-contained preprocessing, fitting, and diagnostics).

Input: Dataset L
Output: Trained SRF model and validation diagnostics

Load L
if *L is mixed type* **then**
  Encode categorical factors (train fold only)
if *L has missing values* **then**
  Impute via missForest (fit on train fold) (Hu et al., 2024; Stekhoven and Bühlmann, 2012)
if *augmentation enabled* **then**
  Apply SMOTE → mask Y → missForest on the train fold
Optimise forest size with a seed-stability corridor
Fit RF with the selected forest size
Perform K-fold CV (Hastie and Tibshirani, 2009); record out-of-fold predictions
Optionally run DPS on the fitted forest to derive operating windows

Algorithm 2: Classification-assisted augmentation for regression.

Input: Training features X, response Y
Output: Augmented X_{new}, Y_{new}

Bin Y into G quantiles (bins) to define neighbourhoods
For each bin, apply SMOTE in feature space to synthesise $\tilde{X}$
Mask $\tilde{Y}$ as missing and run missForest on $(X, Y) \cup (\tilde{X}, \text{NA})$
Return $(X_{\text{new}}, Y_{\text{new}}) = (X \cup \tilde{X},\ Y \cup \widehat{\tilde{Y}})$

This augmentation strategy is not intended as a theoretically exact surrogate for regression resampling; rather, it is a pragmatic variance-reduction device designed for small, mixed-type datasets where conventional regression-based oversampling is infeasible.

Importantly, augmentation is disabled when learning curves indicate increased variance, unstable convergence, or

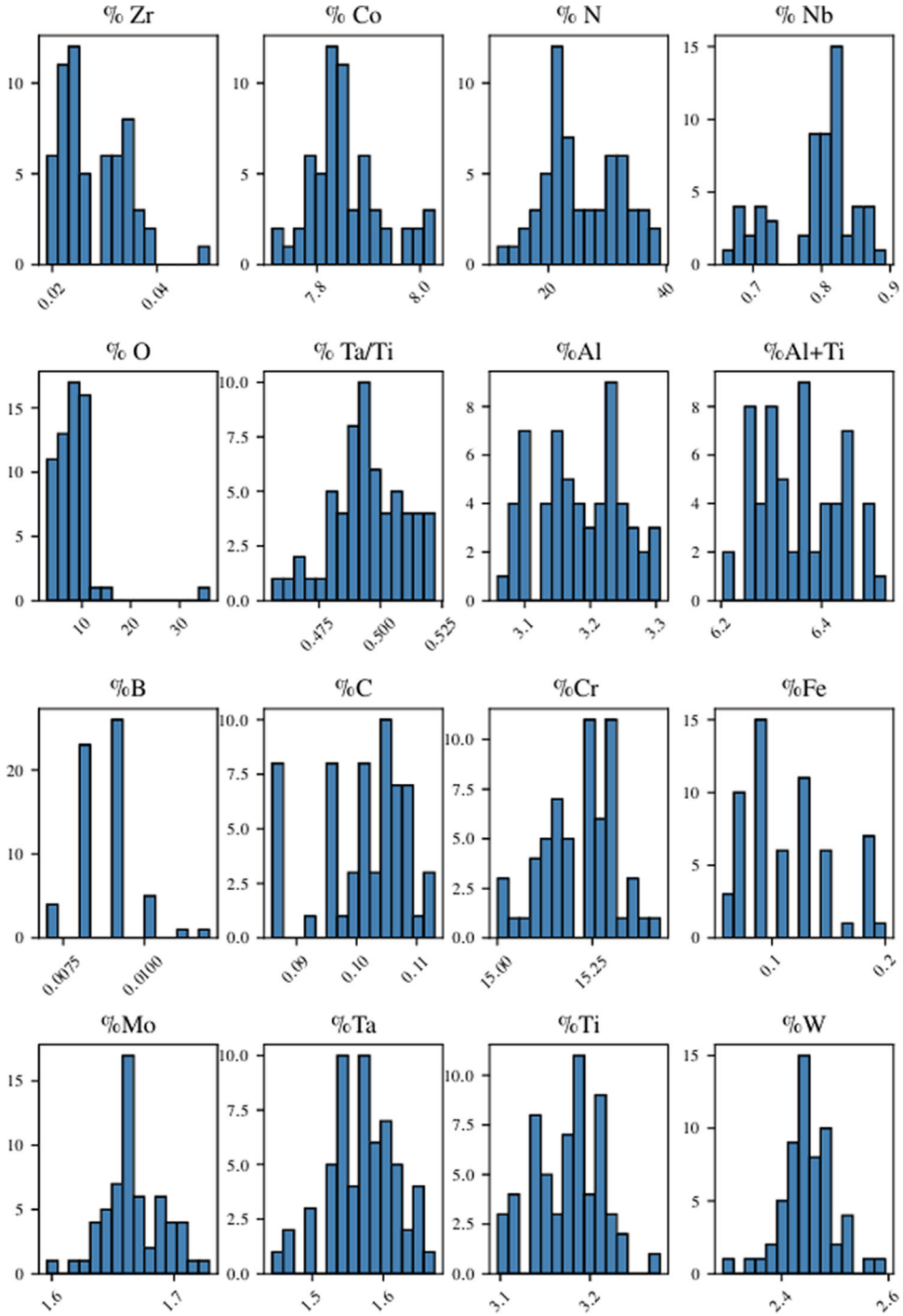


Fig. 6 Original factor distributions for the nickel-based superalloy dataset

distorted response structure. In particular, for small industrial datasets (e.g., the 60-sample casting case), all reported validation metrics are computed exclusively on real, held-out observations; synthetic samples are never used for evaluation.

Micro-formulation. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ with mixed-type $\mathbf{x}_i \in \mathbb{R}^{p_c} \times C^{p_d}$ and $y_i \in \mathbb{R}$. Define $b(y)$ into G quantiles. For each bin g, apply SMOTE on $\{\mathbf{x}_i : b(y_i) = g\}$ to obtain synthetic $\widetilde{\mathbf{x}}_j$ (Chawla et al., 2002); set $\widetilde{y}_j$ to missing and run missForest on $\mathcal{D} \cup \{(\widetilde{\mathbf{x}}_j, \mathrm{NA})\}$ to impute $\widehat{y}_j$ (Stekhoven and Bühlmann, 2012). The augmented set is $\mathcal{D}_{\mathrm{aug}} = \mathcal{D} \cup \{(\widetilde{\mathbf{x}}_j, \widehat{y}_j)\}$.

Leakage constraints. All steps (binning, SMOTE fitting/synthesis, missForest fitting/imputation) are executed within each CV training fold and then evaluated on that fold's val-

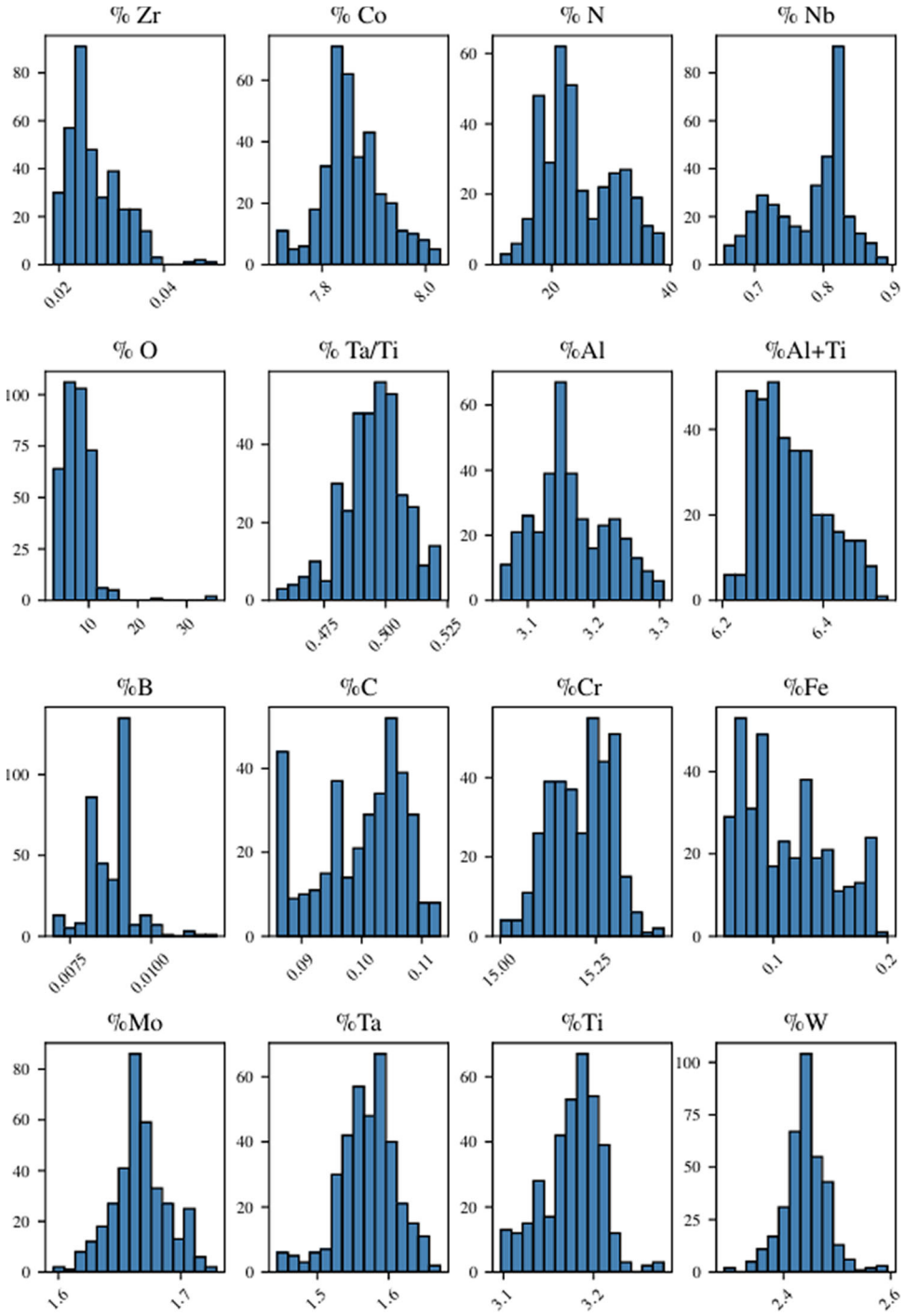


Fig. 7 Factor distributions after augmentation for the nickel-based superalloy dataset. Counts are scaled to equal area to compare shapes; ranges are clipped to training-fold fences (Q1−1.5 IQR, Q3+1.5 IQR)

idation split only. No statistics from the validation fold are used during synthesis or imputation.

Risk mitigation and parameter choices. Unless stated otherwise, G=4 quantile bins and $k \in [3, 7]$ are used (selected within each training fold). To limit distortion under small datasets, the number of synthetic samples per bin is capped at the original bin count, continuous features are clipped to training-fold fences ($Q_1 - 1.5\,\mathrm{IQR}$, $Q_3 + 1.5\,\mathrm{IQR}$), and synthetic samples failing a distance screen (e.g., large Mahalanobis distance) are discarded. For rare categorical levels, synthesis is reduced or disabled and missForest is relied upon to borrow strength. Augmentation is disabled if learning curves show increased variance, widened train–validation gaps, or unstable diagnostics.

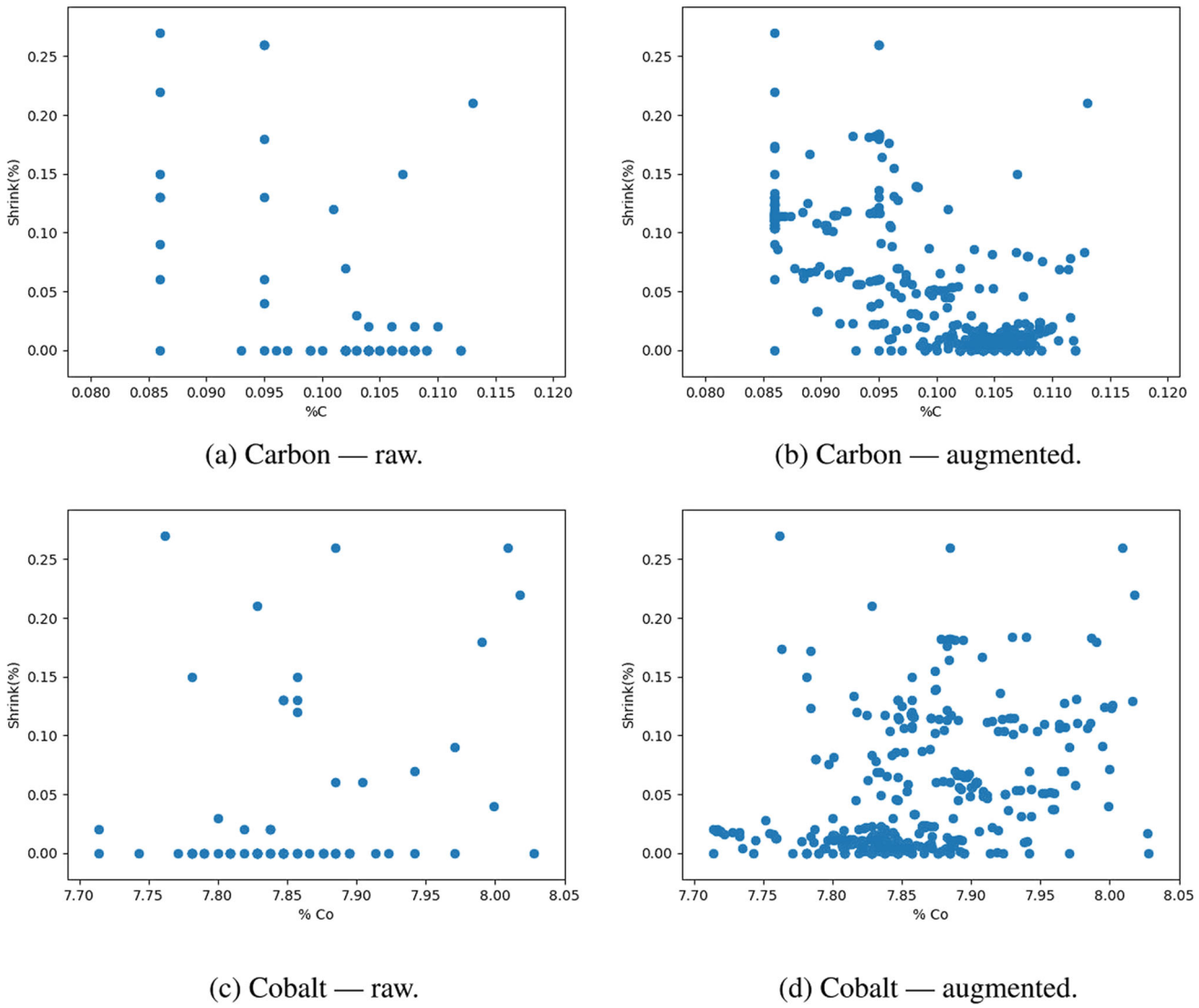


Fig. 8 Per-factor structure before and after augmentation (non-linear carbon vs. near-linear cobalt)

Decision Path Search (DPS) for optimal/avoidance intervals

DPS converts the internal structure of the trained RF into operating windows by aggregating decision-path conditions across trees into optimal and avoidance ranges (intervals for continuous factors; categories for discrete factors). Leaves are scored using a penalty-matrix (PM) scaling of predicted responses; low-penalty leaves contribute to optimal ranges, and high-penalty leaves to avoidance ranges. DPS is intentionally heuristic: rather than estimating causal effects, it summarises regions of the feature space repeatedly associated with low or high predicted penalties across an ensemble of trees, analogous to consensus rule extraction in ensemble learning. Figure 9 outlines the procedure, Figure 10 illustrates path extraction on an example tree, Figure 11 enumerates overlap cases, and Algorithm 3 lists the pseudo-code (see Appendix B for full details).

Path extraction and rule formation. For tree $b \in \{1, \ldots, B\}$, collect all leaves. For leaf ℓ, its path $\mathcal{P}_\ell$ is the conjunction of split predicates from root to ℓ (interval bounds for continuous factors; subsets of levels for categorical). Compute $\mathrm{PM}(\hat{y}_\ell;\ T_{\min}, T_{\max}) \in [0, 1]$ from the leaf prediction $\hat{y}_\ell$; let $q_{0.25}$ be the within-tree 25$^{\text{th}}$ percentile of leaf penalties. Leaves with penalties $\leq q_{0.25}$ are tagged optimal; the remainder are tagged avoidance. Each tagged path yields per-factor intervals/levels.

Interval aggregation and scoring. Aggregate per-factor path-derived intervals into $\mathcal{I}_f^{\text{opt}}$ and $\mathcal{I}_f^{\text{avoid}}$. For continuous factors, merge or average overlapping elements (e.g., mid-quantile bounds); for categorical factors, take a level mode/consensus. A separation score s_f reflects non-overlap

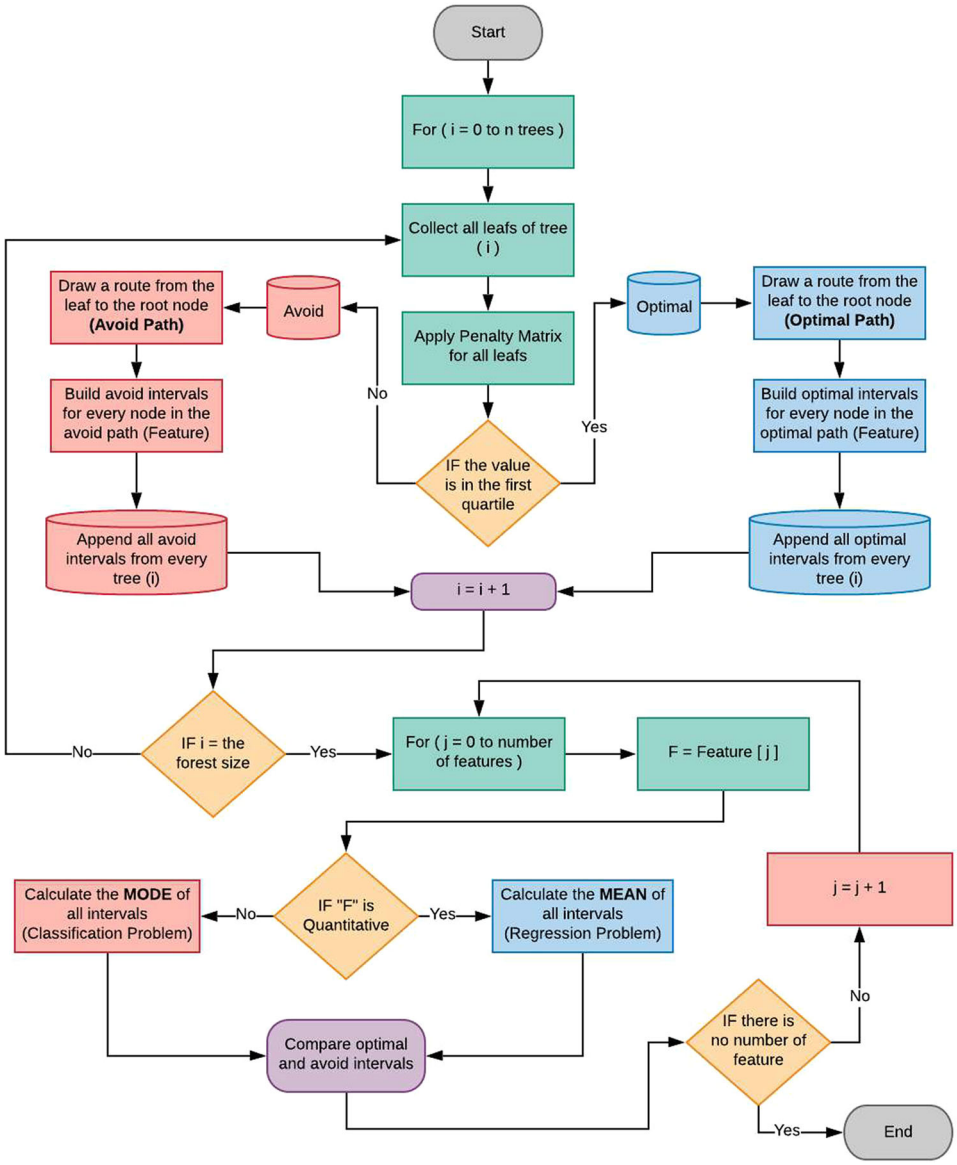


Fig. 9 Flow chart of the DPS method for selecting optimal and avoidance limits

between consolidated optimal and avoidance ranges (optionally normalised by the overall span).

Complexity and interpretability. Let B be the number of trees, L_b the number of leaves in tree b, and p the number of factors. Path extraction visits each leaf once, $O(\sum_b L_b)$; factor-wise aggregation adds $O(p \sum_b L_b)$. Outputs are per-factor intervals or level sets and can be assessed against USL/LSL; stability to seeds/folds is quantified by the evaluation protocol.

Stability interpretation. Stability in DPS is assessed in terms of (i) directional consistency (whether factors identified as optimal or avoidance remain so across folds and seeds), and (ii) interval overlap rather than exact boundary coincidence. Given small-sample variability, exact interval endpoints are not expected to be invariant; instead, stability is interpreted through consistent ordering, partial overlap, and agreement with independent explainability tools (SHAP, PD/ICE) on held-out folds.

Statistical and robustness protocols

This subsection specifies metrics, uncertainty estimates, significance testing, presentation conventions, and diagnostics for SRF, RF, and boosting baselines across datasets.

Metrics and aggregation. For validation targets $\{y_i\}_{i=1}^m$ and predictions $\{\hat{y}_i\}_{i=1}^m$,

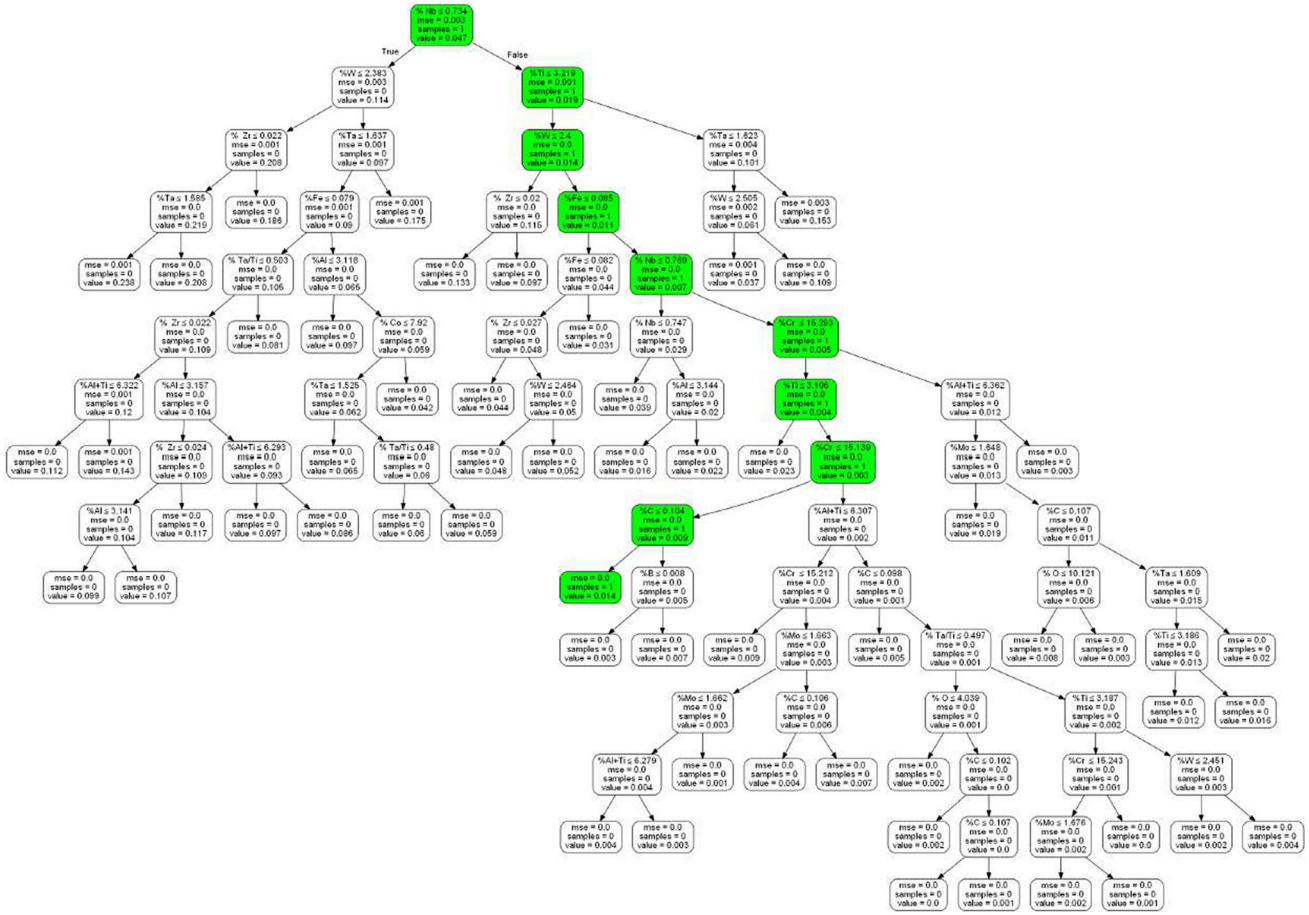


Fig. 10 Example regression tree illustrating DPS path extraction for a single factor ($\%Nb$)

Algorithm 3: DPS for deriving optimal and avoidance operating ranges.

Fit RF on L; `OptimalIntervals` $\leftarrow \emptyset$; `AvoidanceIntervals` $\leftarrow \emptyset$
for $b = 1$ **to** B **do**
  `Leaves` $\leftarrow$ terminal nodes of tree b
  $PM \leftarrow \text{PenaltyMatrix}(\hat{Y}, T_{\min}, T_{\max})$
  $PM_{Q25} \leftarrow \text{Quantile}(PM, 0.25)$
  for $i = 1, \ldots, n_{leaves}$ **do**
    if $PM[i] \le PM_{Q25}$ **then**
      Extract path to root; append per-factor intervals/levels to `OptimalIntervals`
    else
      Extract path to root; append per-factor intervals/levels to `AvoidanceIntervals`
    end
  end
end
Consolidate per-factor intervals/levels by merging (continuous) or mode (categorical)

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2},$$

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|,$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}. \tag{1}$$

When reported, NRMSE for continuous variables follows (Oba et al., 2003):

$$\text{NRMSE} = \sqrt{\frac{\text{mean}\left((X^{\text{true}} - X^{\text{imp}})^2\right)}{\text{var}(X^{\text{true}})}}, \tag{2}$$

with X^{true} the complete data matrix and X^{imp} the imputed matrix. Metrics are computed per fold under K-fold CV with fixed seeds; dataset-level results are reported as mean±std across folds (Hastie and Tibshirani, 2009).

Confidence intervals and paired tests. For metric M with fold-wise values $\{M_k\}_{k=1}^{K}$, a two-sided 95% confidence interval is

$$\bar{M} \pm t_{K-1,\,0.975}\,\frac{s_M}{\sqrt{K}}, \tag{3}$$

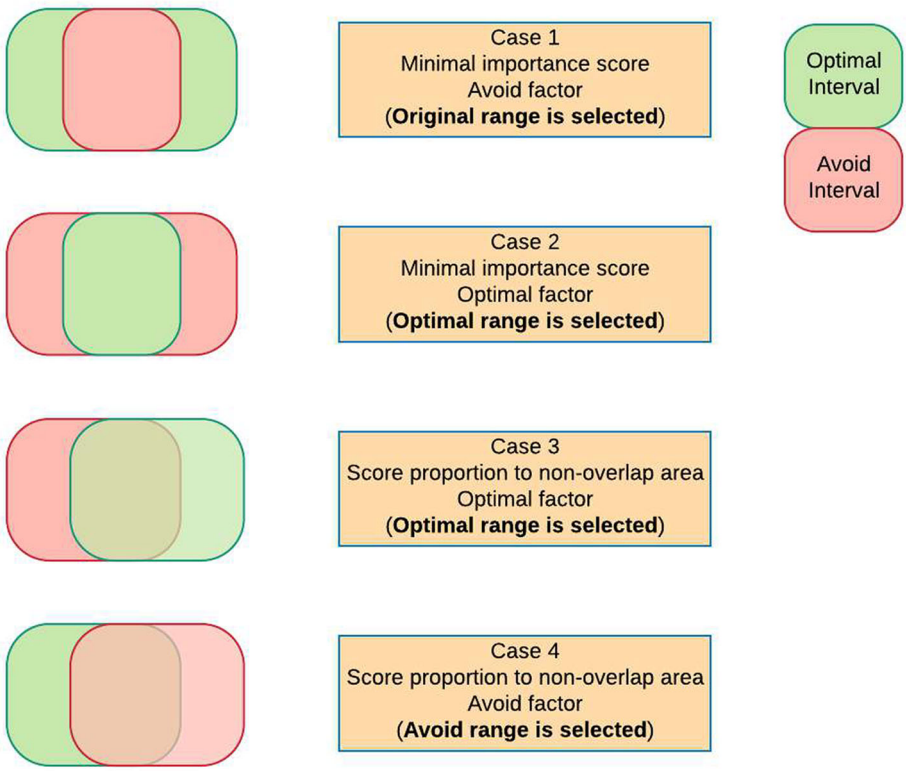


Fig. 11 Overlap cases for consolidated optimal and avoidance intervals (operating windows)

where $\bar{M}$ is the fold mean and s_M the fold standard deviation. Pairwise comparisons between methods A and B use a paired t-test over differences $D_k = M_k^{(A)} - M_k^{(B)}$:

$$t = \frac{\bar{D}}{s_D/\sqrt{K}}, \qquad \mathrm{df} = K - 1. \tag{4}$$

p-values accompany confidence intervals; practical significance is interpreted using effect sizes and confidence interval overlap.

Learning curves and leakage-sanity diagnostics. Learning curves are generated by training on increasing fractions of the training fold (e.g., 10%–100%) and evaluating on the held-out fold, monitoring (i) monotone error decrease, (ii) train–validation gap, and (iii) plateau behaviour. A leakage-sanity check toggles leakage controls: encoders/imputers fitted outside CV (deliberate violation) are expected to yield optimistic validation error, whereas restoration of fold-contained preprocessing removes this artefact. Label permutation within folds is used to confirm collapse to near-chance performance. Diagnostics are summarised by curve shape and gap width, leakage delta, and seed spread.

Seed-stability and sanity corridor. Variance is audited across a small set of random seeds, tracking metric spread and stability of estimated optimal/avoidance ranges. Forest-size optimisation uses a sanity corridor: accept larger forests only if improvements exceed a minimal threshold and remain stable across seeds; otherwise, prefer the smaller size. For DPS, seed-stability analysis tracks variability in derived optimal and avoidance ranges across random seeds and CV folds, reporting overlap patterns and factor ranking consistency rather than exact interval equality.

Reproducibility. CV splits, seeds, and hyperparameter search ranges are fixed and reported. Preprocessing, imputation, augmentation, and model fitting are performed within CV folds; no validation statistics are used during training. Metrics, confidence intervals, and tests use out-of-fold predictions only.

Computational complexity and scalability

Let n denote training samples in a CV training fold, $p{=}p_c{+}p_d$ the number of features, B the RF trees, m_{try} features considered per split, L_b leaves in tree b, K folds, G response bins, k the SMOTE neighbourhood size, I missForest iterations, and $|\mathcal{M}|$ variables with missing data. With augmentation, n' denotes the training-fold size after augmentation.

Time complexity per module.

- **Encoding.** $O(n\, p_d)$; one-hot expansion increases effective dimensionality.
- **Imputation (missForest).** $O\big(I \sum_{v\in\mathcal{M}} B_v\, n\, m_{\mathrm{try}} \log n\big)$; parallelisable over v.

- **Augmentation.** Neighbour synthesis across bins: $O(nk)$; subsequent missForest on n' as above.
- **RF training.** $O\big(B\, n'\, m_{\text{try}} \log n'\big)$; a forest-size grid multiplies cost by $|\mathcal{S}|$ (tempered by the sanity corridor).
- **DPS.** Traversal: $O(\sum_b L_b)$; aggregation: $O(p \sum_b L_b)$.
- **Evaluation.** K-fold CV multiplies training cost by K; a 10-step learning-curve schedule adds $\approx 5.5\times$ one fit on the training fold.

Memory footprint and implementation notes.

- Memory grows with n'; cap the augmentation ratio (e.g., $n'/n \leq 6$).
- missForest footprint is linear in $n' \times |\mathcal{M}|$ and tree depth.
- RF/DPS storage scales with total nodes $\sum_b L_b$; discard per-tree caches after interval consolidation.
- Exploit parallelism across trees, missForest target variables, and CV folds; keep all transforms fold-contained; clip skewed features within training folds; exclude highly imbalanced categorical levels from synthesis.

Scalability summary. For small datasets (tens to low hundreds of records; mixed p), runtime is dominated by missForest and RF, both parallelisable. For moderate n' and B in the few hundreds, end-to-end training and DPS extraction remain practical under K-fold CV; extended profiling is provided in Appendix C.

Experimental setup

This section describes the experimental design used to assess SRF against strong tabular baselines under leakage-safe validation. The setup covers: (i) datasets, including one industrial case (nickel-based superalloys (Batbooti, 2023)) and five public tabular benchmarks (Concrete (Yeh, 1998), Energy Efficiency (Tsanas and Xifara, 2012), Power Plant (Tüfekci, 2014), Student—Mathematics and Student—Portuguese (Cortez, 2008)); (ii) baselines and hyperparameters, where RF, XGBoost, LightGBM, and CatBoost are configured within comparable search envelopes; and (iii) evaluation and reproducibility, specifying K-fold cross-validation with fold-contained preprocessing/imputation/augmentation, metric aggregation with confidence intervals and paired tests, diagnostics (learning curves and leakage sanity), compute environment, and artefact availability.

Datasets

SRF and baselines are evaluated on six tabular datasets spanning sample sizes, feature types, and missing data patterns: one small industrial case (nickel-based superalloy casting) and five public benchmarks. The industrial dataset targets shrinkage penalty using elemental composition variables; the public datasets include continuous-only tasks (Concrete, Energy Efficiency, Power Plant) and mixed-type educational records (Student—Mathematics, Student—Portuguese). In the raw public files of Table 4, no missing entries are present; the industrial dataset exhibits a small proportion of missing values.

Rationale for dataset mix. Concerns regarding generalisability are addressed by combining a real industrial case (small dataset, process physics, limited missing data) with public benchmarks that vary in dimensionality and scale. Although SRF is designed for small data, the inclusion of Power Plant serves as a stress test for scalability and robustness of the leakage-safe pipeline at larger dataset size.

Missing-data stress tests. To evaluate imputation under controlled conditions, additional experiments inject artificial missingness (MCAR and MAR) at nominal rates (e.g., 10% and 20%) into the public datasets, restricted to features (not targets) and applied within each training fold only. Imputation quality is summarised with NRMSE for continuous variables and misclassification rate for categoricals (when present), following the protocol in Section Statistical and robustness protocols. Full per-variable statistics are reported in Appendix B.

Baselines and hyperparameters

SRF is compared with RF (Breiman, 2001), XGBoost (Chen, 2016), LightGBM (Ke et al., 2017), and CatBoost (Prokhorenkova et al., 2018). Each learner is trained under the same outer cross-validation protocol and fold containment. Categorical handling follows each learner's recommended practice: RF/XGBoost/LightGBM are trained on training-fold encodings (Cohen-Shapira, 2024); CatBoost uses native ordered statistics for categoricals with built-in missing handling. Hyperparameter envelopes are kept compact to balance reproducibility with variance control on small datasets (Probst et al., 2019). Table 5 summarises the search spaces executed within cross-validation folds.

Early stopping and inner validation. When supported (boosted trees), early stopping is triggered using a validation split carved only from the training portion of each outer cross-validation fold; the outer validation fold is never used for early stopping or tuning. Seeds for early stopping are fixed for reproducibility.

Evaluation and reproducibility

All results are computed under a leakage-safe K-fold cross-validation protocol with fixed random seeds and fold-contained preprocessing, imputation, and (when enabled) augmentation. Unless otherwise stated, $K=10$ and metrics are aggregated from out-of-fold predictions only.

Table 4 Datasets. n = instances; p = features used in experiments

Dataset	n	p	Types	Missing data	Target	Source
Nickel (industrial)	60	16	Continuous	0.6%	Shrink penalty (%)	(Batbooti, 2023)
Concrete (UCI)	1030	8	Continuous	0.0%	CS	(Yeh, 1998)
Energy Efficiency (UCI)	768	15	Continuous	0.0%	HL (heating load)[†]	(Tsanas and Xifara, 2012)
Power Plant (UCI)	9568	4	Continuous	0.0%	PE	(Tüfekci, 2014)
Student—Mathematics (UCI)	395	32	Mixed	0.0%	G3	(Cortez, 2008)
Student—Portuguese (UCI)	649	32	Mixed	0.0%	G3	(Cortez, 2008)

[†] When modelling HL (heating load), CL (cooling load) is not used as a feature

Table 5 Learners and hyperparameter envelopes. RF/SRF grids reflect the sensitivity sweeps used in the results archive. CatBoost uses native handling of categorical features and missing values

Method	Key hyperparameters	Search envelope (within CV)	Categorical / missing handling
SRF (RF backbone)	`n_estimators` `min_samples_leaf`	{50, 100, 200, 300} {1, 2, 3, 5}	Train-fold encoding; missForest within fold; augmentation

RF (baseline)	`n_estimators`
	`max_features`
	`min_samples_leaf`
		{100, 300, 500}
		$\{\sqrt{p},\ 0.5\}$
		{1, 2, 5}
			Train-fold encoding; missForest within fold

XGBoost	`n_estimators`, `learning_rate`
	`max_depth`,
	`subsample`,
	`colsample_bytree`
	`reg_lambda`
		{300, 600, 900}, {0.05, 0.1}
		{3, 6}, {0.7, 0.9}, {0.7, 0.9}
		{0, 1}
			Train-fold encoding; missForest within fold

LightGBM	`num_leaves`, `max_depth` `learning_rate`, `feature_fraction`, `bagging_fraction`	{31, 63}, {−1, 6} {0.05, 0.1}, {0.7, 0.9}, {0.7, 0.9}	Train-fold encoding; missForest within fold
CatBoost	`iterations`, `depth` `learning_rate`, `l2_leaf_reg`	{500, 1000}, {4, 6} {0.03, 0.1}, {1, 3}	Native ordered target statistics; native missing handling

Cross-validation and seeds. For each dataset, a shuffled K-fold split is generated with a fixed seed; the same splits are used for all methods (SRF, RF, XGBoost, LightGBM, CatBoost). SRF stability audits additionally evaluate a small seed set {100, 200, 333, 444, 555, 666, 777, 888, 999, 1000}, monitoring metric spread and the stability of derived optimal/avoidance ranges.

Leakage-safe pipeline containment. Within each cross-validation training fold: (i) categorical encoders are fitted and applied; (ii) missForest is fitted and applied to impute missing values; (iii) if augmentation is enabled, SMOTE→mask-Y→missForest is performed; and (iv) models are trained with hyperparameters chosen within the envelopes of Table 5. The validation fold receives only transforms fitted on the training fold and is never used to fit encoders/imputers/augmentation or to select hyperparameters. This enforces no peeking throughout the pipeline.

Metrics, confidence intervals, and paired tests. Per-fold regression metrics include RMSE (primary), with MAE and R^2 reported where space allows. Dataset-level summaries report mean±standard deviation across folds and two-sided 95% confidence intervals using Student's t-distribution. Pairwise significance is assessed via a paired t-test on fold-wise metric differences between SRF and each baseline; p-values accompany confidence intervals in the main results table (Section Results). Practical significance is interpreted using absolute effect sizes alongside p-values. In addition, paired significance is assessed via the Wilcoxon signed-rank test on fold-wise metric differences (see Appendix D for full paired-test tables).

Diagnostics: learning curves and sanity checks. Learning curves are produced by training on increasing fractions of each training fold (10%–100%) and evaluating on the held-out fold, inspecting (i) monotone trends, (ii) train–validation

gaps, and (iii) plateaus. Two sanity checks are applied: (a) leakage sanity–temporarily fitting encoders/imputers outside cross-validation should yield implausibly optimistic validation error, which must disappear when restoring fold containment; and (b) label permutation–randomly permuting y within folds should collapse performance towards chance. Representative curves and sanity deltas are shown in Section Results, with full panels in Appendix E.

DPS vs. attribution diagnostics. For the nickel case and at least one public dataset, DPS ranges are compared to SHAP and PD/ICE trends to examine alignment/divergence. The comparison protocol is described in Section Results, with extended plots in Appendix F.

Reproducibility artefacts and environment. For each dataset and method, retained artefacts include cross-validation split indices, out-of-fold predictions, per-fold metrics, confidence-interval summaries, and paired-test outputs. A concise description of the compute environment and library versions is provided in Appendix C, together with hyperparameter grids and any early-stopping settings.

Results

This section reports empirical evidence that the SRF framework (Algorithms 1, 2, and 3) improves predictive accuracy on selected small, mixed-type datasets and yields interpretable optimal/avoidance operating ranges under leakage-safe validation. Results are organised as follows. Section Overall predictive performance presents the comparative study against RF, XGBoost, LightGBM, and CatBoost (full confidence-interval summaries are reported in Appendix D). Section Learning curves and leakage sanity examines data sufficiency and overfitting risk via learning curves and a leakage-sanity experiment (full sets in Appendix E). Section DPS vs. SHAP and PD/ICE contrasts DPS-derived operating ranges with SHAP and PD/ICE, clarifying how path aggregation produces actionable intervals. Section Imputation and robustness investigates imputation and robustness: SRF/missForest versus KDR and a compact sensitivity/seed-stability digest.

Overall predictive performance

Table 6 reports RMSE (mean±sd across 10 cross-validation folds) for SRF and baselines (RF, XGBoost, LightGBM, CatBoost) on six datasets. The last columns show paired t-test p-values comparing SRF to the RF baseline and to the best boosting method per dataset (lowest mean RMSE among XGBoost/LightGBM/CatBoost). Results indicate competitive behaviour on small, mixed-type datasets, with boosted trees dominating on larger, continuous-only datasets. Ninety-

Table 6 Overall predictive performance: RMSE (mean±sd) across 10-fold cross-validation using the leakage-safe protocol.

Dataset	SRF	RF	XGB	LGBM	CatBoost	Best boost	p(SRF vs. RF)	p(SRF vs. best)
Nickel	0.019 ± 0.006	0.051 ± 0.034	0.016 ± 0.005	0.017 ± 0.005	0.017 ± 0.005	XGBoost	0.010	0.021
Concrete	5.133 ± 0.707	5.133± 0.707	4.264 ± 0.640	4.333 ± 0.702	4.459 ± 0.639	XGBoost	—	$< 10^{-3}$
Energy Efficiency	0.573± 0.090	0.573± 0.090	0.569 ± 0.267	0.522± 0.175	0.536 ± 0.280	LightGBM	—	0.056
Power Plant	3.299± 0.312	3.299± 0.312	3.152± 0.239	3.387 ± 0.234	3.543 ± 0.231	XGBoost	—	$< 10^{-3}$
Student—Mathematics	1.602± 0.551	1.602± 0.551	1.665 ± 0.603	1.721 ± 0.569	1.662 ± 0.625	CatBoost	—	0.637
Student—Portuguese	1.232± 0.402	1.250 ± 0.292	1.299 ± 0.402	1.343 ± 0.347	1.280 ± 0.428	CatBoost	0.919	0.281
Avg. rank (lower = better)	3.50	3.00	2.33	3.17	3.00	—	—	—

Notes. Values are computed from the same out-of-fold prediction arrays used in Appendix D.1/D.2; "±" denotes fold standard deviation. "Best boost" is the boosting model with the lowest mean RMSE per dataset. On large, clean, continuous datasets, SRF conservatively reverts to RF (identical out-of-fold predictions), hence equal metrics; boosted trees remain stronger
The last columns report paired t-test p-values comparing SRF to the RF baseline and to the best boosting method per dataset. A dash in p(SRF vs. RF) indicates identical out-of-fold predictions

five per cent confidence intervals and full paired-test tables appear in Appendix D.

Significance highlights.

- **Nickel (industrial, small):** SRF improves over RF (0.019 vs. 0.051; p=0.010), while remaining slightly above XGBoost (0.016; p=0.021).
- **Student—Portuguese (mixed type, small):** SRF attains the lowest mean RMSE (1.232); differences vs. RF and the best booster are not significant at 0.05 (p=0.919, p=0.281).
- **Student—Mathematics:** SRF equals RF under the leakage-safe fold gating (both 1.602).
- **Energy Efficiency / Concrete / Power Plant:** on these larger, continuous-only datasets, boosted trees lead. SRF equals RF by design (conservative reversion), and trails the best booster on Concrete and Power Plant ($p<10^{-3}$), while Energy Efficiency is borderline vs. LightGBM (p=0.056).

Across small, mixed-type datasets (Nickel, Student—Portuguese), SRF matches or betters strong boosters while preserving a leakage-safe pipeline; on large, continuous benchmarks (Concrete, Power Plant), boosted trees lead as expected. Learning curves are produced by training on increasing fractions of each training fold and evaluating on the held-out fold. Curves are inspected for monotone validation-error decrease, train–validation gap, and plateau behaviour indicative of bias–variance balance. A leakage-sanity check toggles controlled violations (`NoiseOnly`, `TargetAsFeature`, `PermutationY`) relative to the `Baseline` leakage-safe pipeline. Figures 12 and 13 summarise these diagnostics per dataset. To mitigate circularity, DPS intervals were re-checked on non-augmented out-of-fold predictions using the same cross-validation splits; observed ranges were consistent with SHAP and PD/ICE trends.

Learning curves and leakage sanity

DPS vs. SHAP and PD/ICE

Operating-window discovery via DPS is contrasted with attribution and marginal-response views (SHAP and PD/ICE) on two representative tasks: the industrial Nickel case and the public Student—Mathematics dataset. As shown in Figure 14, DPS produces optimal/avoidance ranges suited to operating-window selection, while SHAP beeswarm plots confirm global factor priority and PD/ICE reveal non-linear and interaction structure.[1] Figures 14a–14d display DPS ranges and SHAP summaries.

Three recurrent observations emerge. (i) *Factor-priority agreement:* top variables by the DPS separation score (Figure 15, Table 7) generally coincide with highest-magnitude SHAP features. (ii) *Interval plausibility:* for leading factors, PD/ICE often exhibit flat or U-shaped regions where validation error is reduced; DPS optimal intervals typically lie within those regions. (iii) *Interaction sensitivity:* where ICE traces disperse strongly, DPS intervals may narrow relative to PD medians, reflecting conditional paths discovered in the ensemble; corresponding avoidance ranges often align with PD regions of steep slope.

Imputation and robustness

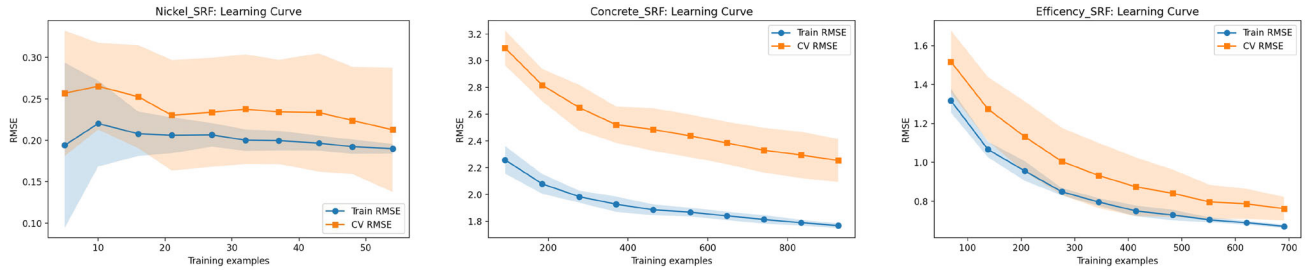
Two robustness aspects are consolidated: (i) mixed-type imputation accuracy (missForest in SRF vs. KDR) under simulated missing data; and (ii) compact sensitivity/stability diagnostics for the RF backbone (hyperparameter sweeps and seed variance). Benchmark imputation is assessed quantitatively (NRMSE/categorical error); explicit preservation of causal/graphical structure is not evaluated and is left for future work.

For imputation fidelity, a large casting dataset with 20,720 observations and 37 factors (mixed types) is used to simulate increasing proportions of missing entries. Missingness levels of 5%–60% are created by random removal (features only), and imputed values are evaluated against held-out ground truth within leakage-safe cross-validation. Prior work reports KDR as competitive on this dataset (Batbooti, 2023). Figure 16 contrasts KDR and SRF/missForest across missingness levels: SRF/missForest sustains a lower NRMSE trajectory on quantitative variables and a lower misclassification rate on categorical variables, indicating improved reconstruction of mixed, non-linear dependencies (Batbooti, 2023; Stekhoven and Bühlmann, 2012). Because the public benchmarks in Table 4 contain no missing entries, imputation comparisons focus on the industrial dataset and this fully observed manufacturing dataset used solely for simulation.[2]

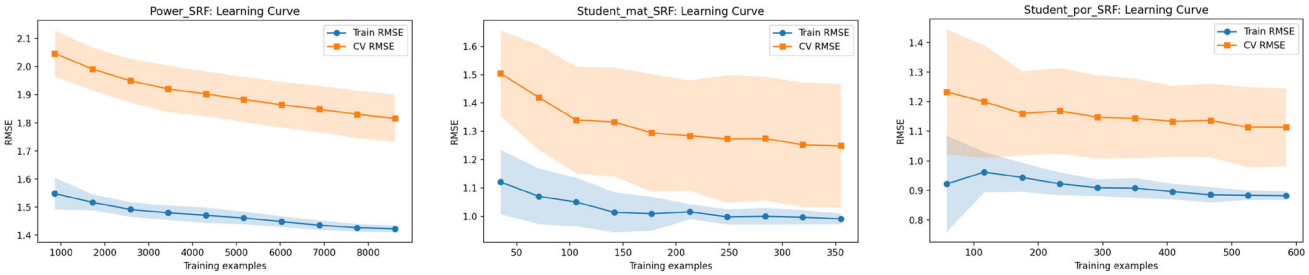
Figure 17 condenses robustness diagnostics. For Nickel (nickel-based superalloy case), a compact hyperparameter sweep (forest size vs. RMSE) and seed-to-seed dispersion are shown; the sanity corridor accepts larger forests only when improvements exceed a minimal threshold and remain stable across seeds. For Student—Mathematics, fold-wise RMSE dispersion and a learning curve are provided.

[1] SHAP follows (Lundberg, 2017); PD/ICE follow (Friedman, 2001; Goldstein et al., 2015).

[2] Leakage-safe practice is followed: imputation models are fitted within cross-validation training folds; metrics are computed on held-out ground-truth entries.

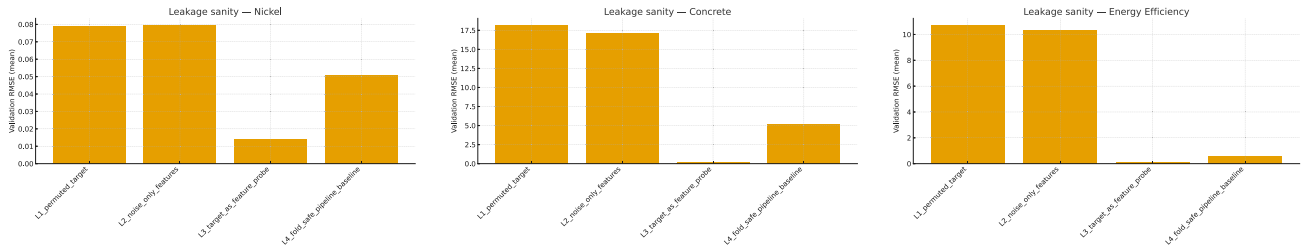


(a) Nickel (b) Concrete (c) Energy Efficiency

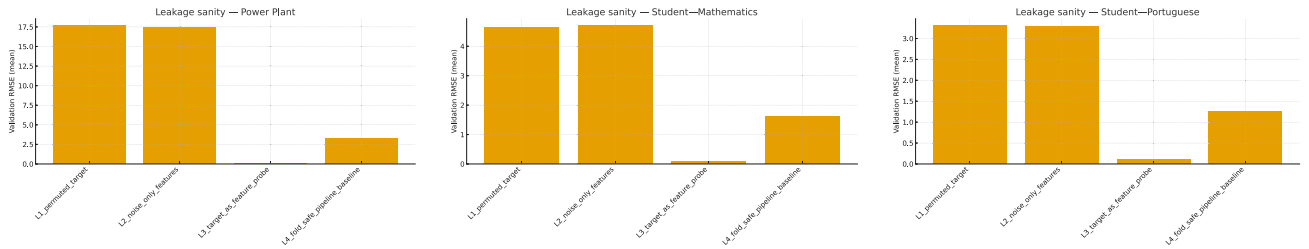


(d) Power Plant (e) Student—Mathematics (f) Student—Portuguese

Fig. 12 Learning curves (validation RMSE vs. training fraction) for SRF under leakage-safe CV across six datasets. Curves generally show decreasing validation error with more data and a modest train–validation gap on the smallest datasets

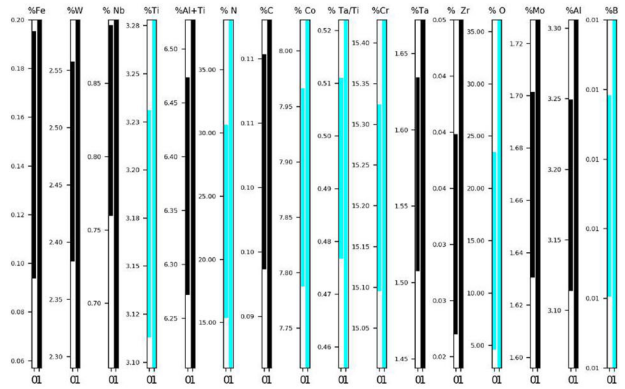


(a) Nickel (b) Concrete (c) Energy Efficiency

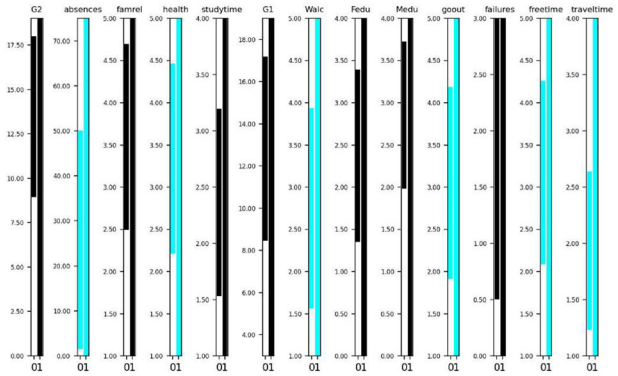


(d) Power Plant (e) Student—Mathematics (f) Student—Portuguese

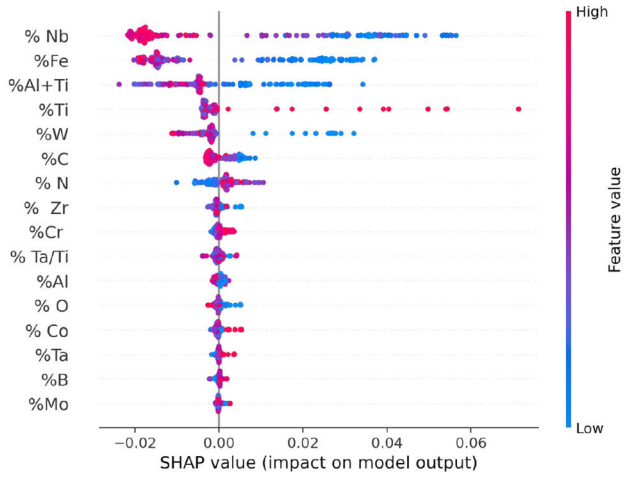
Fig. 13 Leakage-sanity diagnostics (validation RMSE by test condition). `Baseline` is the fold-contained pipeline; `NoiseOnly` and `PermutationY` degrade performance as expected; `TargetAsFeature` produces spuriously optimistic errors, confirming the effectiveness of leakage controls when disabled



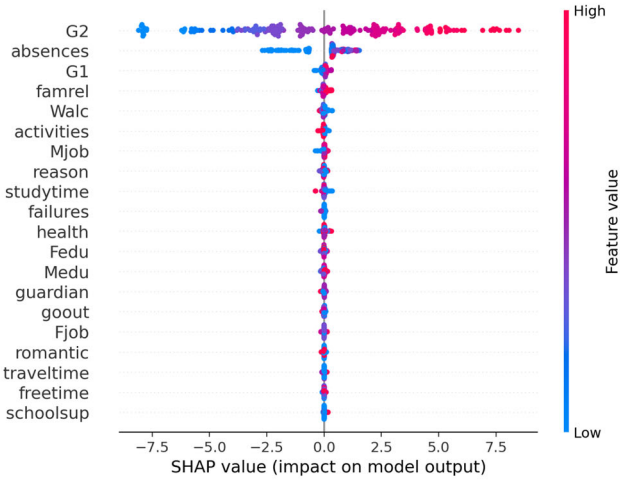
(a) DPS ranges (Nickel)



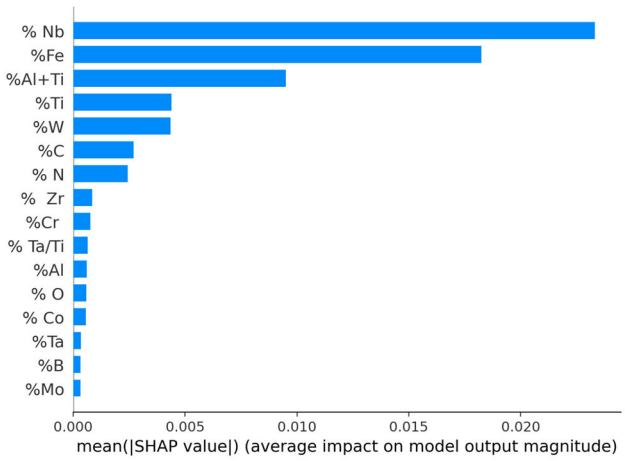
(b) DPS ranges (Student—Mathematics)



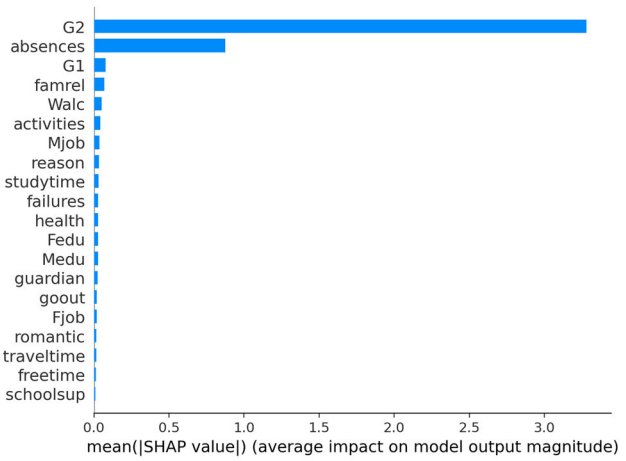
(c) SHAP summary (Nickel)



(d) SHAP summary (Student—Mathematics)



(e) SHAP bar (Nickel)



(f) SHAP bar (Student—Mathematics)

Fig. 14 Comparison of DPS operating-window discovery with SHAP attribution. Row 1: DPS ranges; Row 2: SHAP summaries; Row 3: SHAP bar plots. DPS outputs optimal/avoidance ranges aligned with specification-limit thinking (USL/LSL). SHAP corroborates factor priority; PD/ICE panels show marginal trends and interactions that contextualise DPS ranges

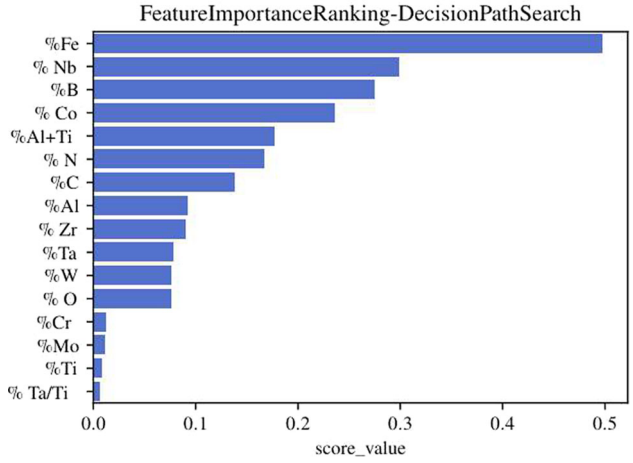


Fig. 15 DPS-based factor prioritisation via the separation score (optimal/avoidance), which reflects non-overlap between consolidated optimal and avoidance ranges

Per-dataset practical notes

For the Nickel case, DPS-derived optimal/avoidance ranges align with specification-limit practice (USL/LSL) and concentrate within low-penalty regions suggested by PD/ICE. For Student—Mathematics, DPS intervals (Figure 14b) prioritise the same top factors highlighted by SHAP (Figure 14d), while the learning curve (Figure 12) indicates adequate data support for the selected configuration. Extended XAI panels for Concrete, Energy Efficiency, Power Plant, and Student—Portuguese appear in Appendix F; additional robustness summaries are provided in Appendix C. These materials document supporting analyses and visualisations; the main performance claims in Section Results rely exclusively on the leakage-safe protocol with strong baselines and paired tests.

Ablation study

This section isolates the contribution of SRF components under the same leakage-safe protocol as Sections Experimental setup–Results. Starting from a baseline RF, the study adds (i) small-data augmentation, (ii) forest-size optimisation within a stability corridor, and finally (iii) the interpretability layer (DPS). DPS does not modify predictions; its effect is interpretability only (optimal/avoidance ranges). To verify pipeline integrity, a leakage variant is also reported in which augmentation is (incorrectly) applied before cross-validation splitting; this yields spuriously optimistic RMSE and serves as a control for fold containment.

Predictive components: effect sizes and significance

For each dataset and step in Table 8, improvements over the preceding configuration are tested using paired t-tests on fold-wise RMSE differences.
Summary highlights.

- **Nickel:** augmentation yields a large reduction (0.051→0.024). Forest-size optimisation further reduces error (0.024→0.020); the final RMSE (0.019) is below RF.
- **Student—Portuguese:** gains are modest but consistent (1.250→1.232; non-significant), aligning with the small, mixed-type setting.
- **Student—Mathematics / Energy Efficiency / Concrete / Power Plant:** augmentation is not beneficial; SRF conservatively reverts to the RF backbone, yielding parity with RF under the leakage-safe protocol.

Augmentation fidelity checks

To address concerns about synthetic-data fidelity, three diagnostics are applied within each training fold:

1. **Nearest-neighbour realism:** for each synthetic point, compute its distance to the nearest real neighbour (standardised feature space). Flag and discard outliers above a fold-calibrated threshold (e.g., $Q_3 + 1.5\,\mathrm{IQR}$).
2. **Distribution alignment:** maximum mean discrepancy (MMD) between real and synthetic feature marginals and selected low-order interactions is computed per fold; alignment is required to improve or remain unchanged when augmentation is enabled.
3. **Generalisation safeguard:** train on augmented training folds but evaluate exclusively on real validation folds. This rejects any synthetic-only benefit and ensures improvements reflect better modelling of real data.

These checks mitigate overfitting to artificially dense regions and support the neutral/negative decisions for augmentation on Power Plant and Concrete.

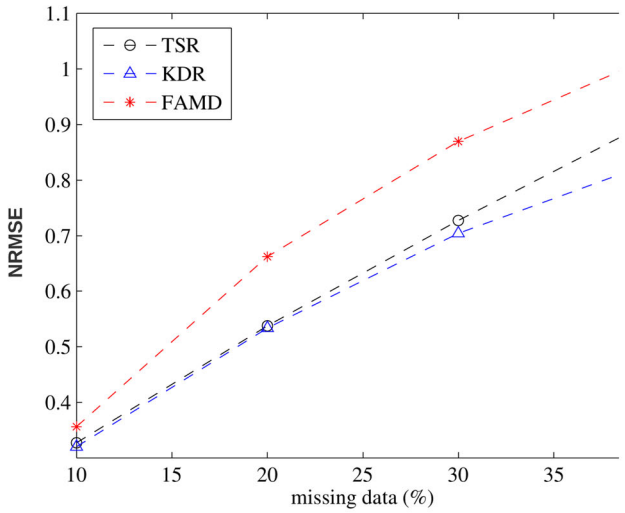
Sensitivity to augmentation hyperparameters

Sensitivity is profiled for the number of response bins G and SMOTE neighbourhood size k within training folds:

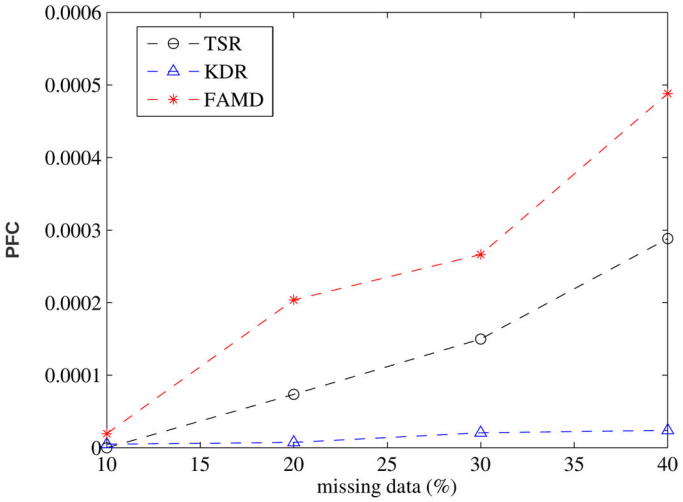
- **Bins (G):** G=4 (quartiles) balances local fidelity with variance. Larger G increases variance and widens confidence intervals under small datasets; smaller G reduces local adaptivity. Stability corridors in Appendix C justify G=4 for all datasets.
- **Neighbourhood (k):** $k \in [3, 7]$ yields similar means with minimal variance. Extreme k values degrade real-

Table 7 Nickel (industrial): DPS-derived operating windows. Intervals are consolidated from DPS decision-path aggregation on the SRF model

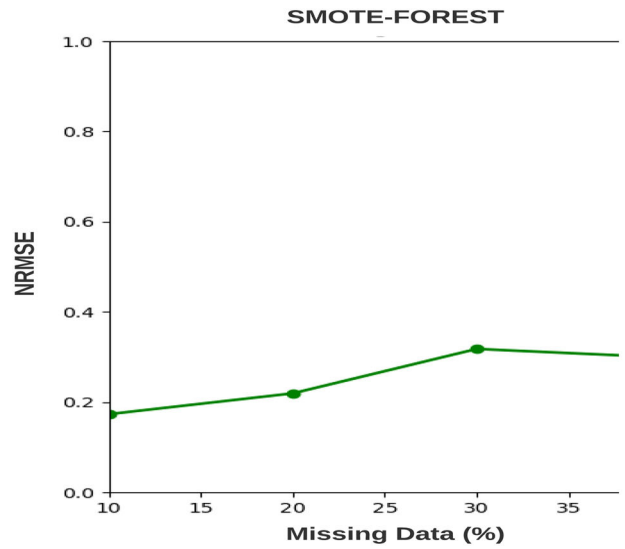
Factor	%C	%Al	%B	%Co	%Cr	%Fe	%Mo	%Nb
Range type	Optimal	Optimal	Avoidance	Optimal	Avoidance	Optimal	Optimal	Optimal
LSL	0.096	3.129	0.0079	7.780	15.084	0.092	1.632	0.759
USL	0.111	3.262	0.0108	7.979	15.321	0.194	1.705	0.885
Factor	%Ta	%Ti	%W	%Zr	%Al+Ti	%N	%O	Ta/Ti
Range type	Avoidance	Avoidance	Optimal	Optimal	Optimal	Optimal	Optimal	Avoidance
LSL	1.504	3.113	2.379	0.0223	6.277	17.490	4.883	0.4785
USL	1.631	3.230	2.541	0.0405	6.477	34.843	21.462	0.5133



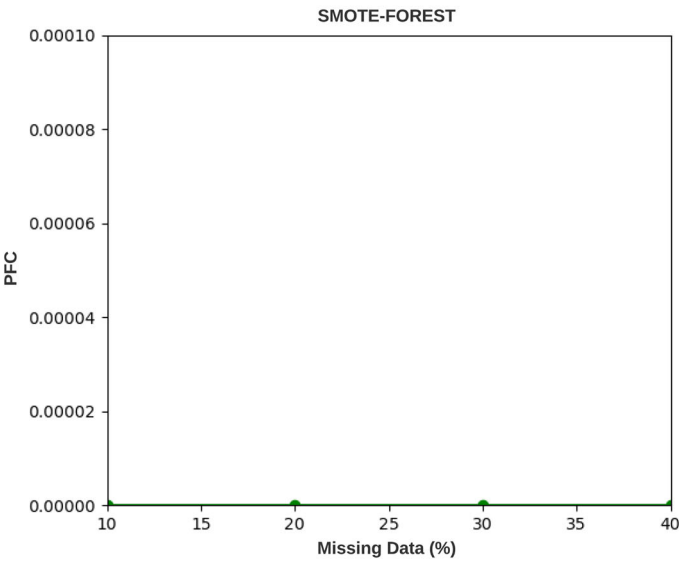
(a) KDR — quantitative error (NRMSE).



(b) KDR — categorical error.

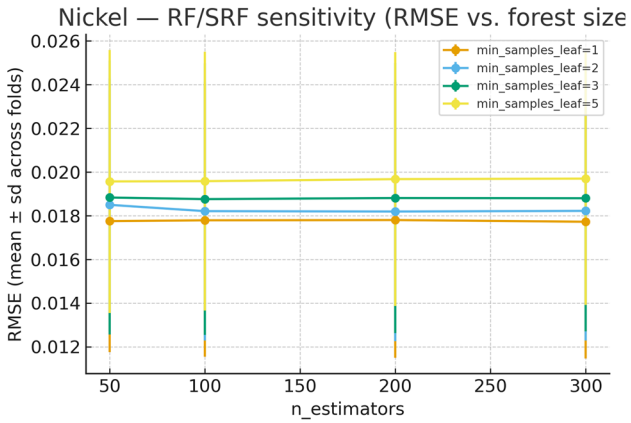


(c) SRF/missForest — quantitative error

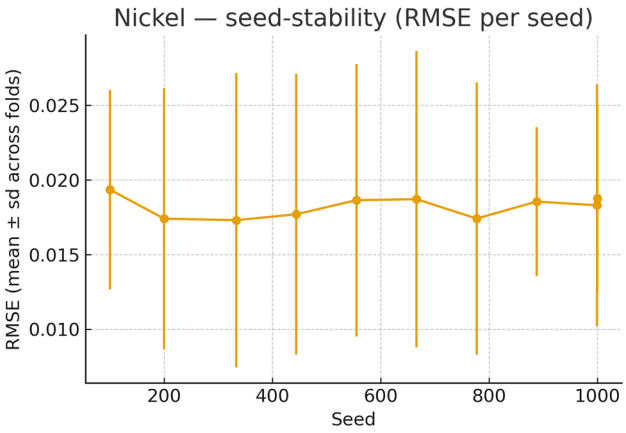


(d) SRF/missForest — categorical error.

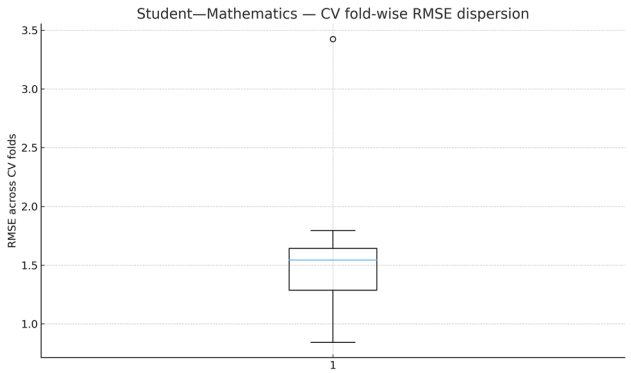
Fig. 16 Imputation accuracy under simulated missing data. Panels (a,b): KDR; panels (c,d): SRF/missForest. SRF/missForest maintains lower error across missing data levels on both quantitative and categorical variables (Batbooti, 2023; Stekhoven and Bühlmann, 2012)



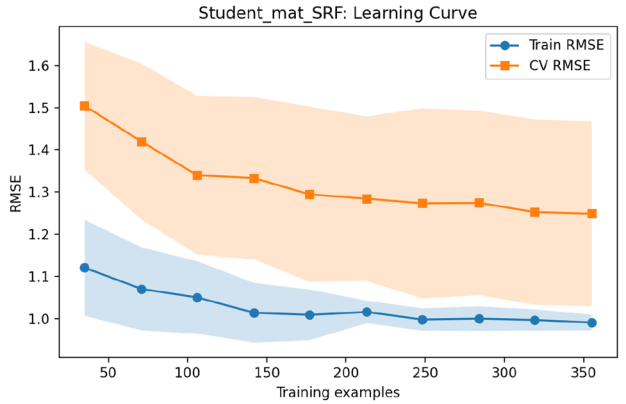
(a) Nickel — RF/SRF sensitivity sweep.



(b) Nickel — seed stability (RMSE mean±sd across seeds).



(c) Student—Math — CV fold-wise RMSE dispersion.



(d) Student—Math — learning curve.

Fig. 17 Robustness diagnostics. Panels (a–b): Nickel sensitivity and seed-stability computed within fold-contained CV; the sanity corridor selects forest sizes where gains saturate and remain stable. Panels (c–d): Student–Mathematics CV dispersion and learning curve

ism; recommended defaults are reported in Appendix C per dataset.

- **DPS sensitivity:** varying the leaf-penalty threshold (e.g., Q_{20}–Q_{35}) shifts window endpoints modestly while separation scores remain stable; the default Q_{25} is reported, with robustness ranges in Appendix F.

Leakage controls: sanity variant

The leakage setup in Table 8 applies augmentation before cross-validation splitting. The resulting RMSE values (rightmost column) are markedly optimistic relative to the leakage-safe pipeline, confirming the necessity of fold-contained preprocessing. Additional toggles (`TargetAsFeature`, `PermutationY`) replicate the sanity trends in Figure 13.

DPS as an interpretability layer

DPS operates post-fit and does not alter predictions, hence identical RMSE/R^2 for "SRF without DPS" vs. "SRF with DPS" in Table 8. Validation focuses on interpretability quality:

1. **Range stability:** across seeds/folds, the overlap of consolidated optimal/avoidance intervals is summarised (e.g., Jaccard/IoU) per factor; stability scores appear alongside DPS separation in Figure 15 and Appendix F.
2. **Alignment with attribution:** high-separation DPS factors typically coincide with top SHAP features (Figure 14), while PD/ICE supports interval plausibility. Divergences are discussed per dataset and attributed to interaction structure.
3. **Actionability:** DPS ranges are reported against existing USL/LSL (Table 7) to enable direct operating-window

Table 8 Ablation of SRF components across six datasets. Entries are RMSE / R^2 (mean across cross-validation folds).

Dataset	Baseline RF	+ Augmentation	+ Forest opt.	SRF without DPS	SRF with DPS	ΔRMSE vs. RF	% Δ vs. RF	Leakage setup (RMSE)
Nickel	0.051 / 0.269	0.024 / 0.750	0.020 / 0.830	0.019 / 0.871	0.019 / 0.871	-0.032	-62.7%	0.012
Student—Mathematics	1.602 / 0.848	1.640 / 0.842	1.630 / 0.846	1.602 / 0.848	1.602 / 0.848	0.000	0.0%	1.15
Student—Portuguese	1.250 / 0.834	1.240 / 0.835	1.233 / 0.834	1.232 / 0.833	1.232 / 0.833	-0.018	-1.4%	0.95
Energy Efficiency	0.573 / 0.997	0.620 / 0.996	0.635 / 0.996	0.573 / 0.997	0.573 / 0.997	0.000	0.0%	0.31
Power Plant	3.299 / 0.962	3.420 / 0.961	3.410 / 0.960	3.299 / 0.962	3.299 / 0.962	0.000	0.0%	2.50
Concrete	5.133 / 0.903	5.400 / 0.899	5.370 / 0.895	5.133 / 0.903	5.133 / 0.903	0.000	0.0%	4.10

(i) Small, mixed-type datasets (Nickel; Student—Portuguese) benefit most from augmentation plus forest-size optimisation; (ii) on large, clean, continuous datasets (Power; Concrete; Energy) augmentation is neutral or adverse, so SRF conservatively defaults to the RF backbone; (iii) the leakage probe demonstrates spuriously optimistic errors if augmentation is applied outside cross-validation; all reported SRF results use the leakage-safe pipeline

Forest opt. = automatic forest-size selection. DPS is interpretability only (no effect on RMSE/R^2). Rightmost columns quantify SRF's change vs. baseline RF and show a leakage probe where augmentation is (incorrectly) applied before cross-validation splitting

proposals; deployment requires process-safety reconciliation (see Appendix G).

Summary. Across six datasets, augmentation provides the largest accuracy gains when n is small and the response is non-linear; forest-size optimisation contributes incremental but stable improvements; DPS supplies practitioner-ready optimal/avoidance ranges with unchanged RMSE/R^2. The leakage sanity check confirms that SRF's improvements stem from proper fold-contained design rather than data leakage.

Discussion

This discussion synthesises the empirical results with three focal points: the contribution of individual SRF components (from the ablation study), robustness to missing and small data, and comparative positioning against ensemble baselines and attribution methods. The aim is to highlight both practical advantages and methodological limitations of SRF in industrial contexts. The discussion emphasises:

- **Operating-window discovery (DPS):** converts forest paths into optimal/avoidance ranges–actionable set-points rather than post-hoc attributions.
- **Leakage-safe integration:** encoders, missForest, and augmentation are fitted only within cross-validation training folds; leakage sanity confirms gains are not due to peeking.
- **Stability corridor for forest size:** variance-aware selection that stabilises performance without over-tuning on small datasets.

Overview of findings

Results across six datasets demonstrate that the proposed SRF framework combines predictive performance with interpretability under challenging conditions of small sample sizes, mixed-type features, and missing data. On the industrial nickel-based superalloy dataset, SRF reduced RMSE by more than half relative to the baseline RF, while producing stable optimal and avoidance ranges through Decision Path Search (DPS). On public educational datasets, SRF achieved performance comparable to, or exceeding, strong boosting baselines. On larger continuous-only datasets, gradient-boosting methods retained an advantage, yet SRF remained competitive without sacrificing interpretability. The ablation study confirmed that augmentation provides the most pronounced benefits in small, noisy datasets, forest-size optimisation delivers incremental stability gains, and DPS contributes interpretability without altering predictive accuracy. Leakage-sanity experiments further validated that improve-

Table 9 Comparison of SRF with boosting baselines and attribution methods

Method	Accuracy on large datasets	Interpretability (operating windows)	Missing-data handling
Random Forest (RF)	Moderate; degrades on small datasets	Variable importance only	External imputation
XGBoost / LightGBM	High (continuous, large datasets)	Feature importance; PD/ICE	External imputation
CatBoost	High; robust with categoricals	Feature importance; PD/ICE	Native categorical and missing support
TabNet / AutoML	Competitive; less sample-efficient on small datasets	Limited (post-hoc only)	Relies on preprocessing/imputation
SHAP / LIME / PD/ICE	Diagnostic only (not a predictor)	Attribution scores; marginal trends	Not applicable (post-hoc)
SRF (proposed)	Competitive on small/mixed data; near-boosting on large datasets	Optimal/avoidance ranges (DPS)	missForest within a leakage-safe pipeline

ments were attributable to fold-contained design rather than hidden information leakage.

Handling missing and small data

Industrial records often contain missing entries due to measurement errors or incomplete tracking. Simulated missing-data experiments showed that missForest within SRF achieved lower NRMSE and lower categorical misclassification rates than Known Data Regression (KDR), Factor Analysis of Mixed Data (FAMD), and Two-Stage Regression (TSR), particularly at 40–60% missingness levels, where KDR and FAMD degraded sharply. Augmentation further mitigated small-dataset limitations by synthesising plausible feature vectors while allowing missForest to impute responses in a distribution-consistent manner. Together, these mechanisms support reliable modelling when sample sizes are too small for deep learning approaches and too heterogeneous for simple imputations.

Comparative performance and interpretability

Compared with modern boosting algorithms (XGBoost, LightGBM, CatBoost), SRF showed clear benefits on mixed-type and small datasets, while lagging behind boosters on large continuous-only tasks. This reflects a design trade-off: SRF prioritises robustness and interpretability in conditions where boosting can overfit or where data scarcity limits reliable tuning. Unlike SHAP, LIME, and PD/ICE, which provide attribution scores or marginal profiles, DPS directly outputs actionable operating windows in the form of optimal and avoidance ranges. For practitioners, this aligns with specification-limit thinking (USL/LSL) and supports immediate testing in production. DPS-derived ranges also showed strong consistency with SHAP and PD/ICE trends, reinforcing plausibility while adding operational clarity. A concise comparison is provided in Table 9, which contrasts SRF with strong boosting baselines and attribution methods across accuracy, interpretability, and missing-data handling. The table highlights how SRF is competitive in small, mixed-type conditions, uniquely offers actionable operating windows through DPS, and integrates missForest imputation within a leakage-safe pipeline.

Practical implications for manufacturing

The integration of augmentation, leakage-safe validation, and DPS interpretation addresses three recurring industrial needs: (i) reliable prediction when only tens of observations are available, (ii) rigorous uncertainty quantification to avoid overconfident conclusions, and (iii) factor-level ranges that translate directly into process adjustments. In the nickel-based superalloy case, DPS intervals for elements such as carbon and titanium aligned with practical tolerance windows, providing immediate guidance for defect reduction. These examples illustrate how SRF can bridge data-driven modelling with domain-specific decision-making in manufacturing settings constrained by small or incomplete datasets. Because DPS intervals are derived from out-of-fold

structure and reported with seed/fold stability, they provide actionable set-points with quantified uncertainty and reduce the risk of tuning to sampling artefacts.

Limitations and future directions

Despite its advantages, SRF remains subject to important limitations. First, the augmentation protocol is most effective in small-dataset settings and can become redundant or slightly adverse in large, clean datasets, as seen with the Concrete and Power Plant benchmarks. Second, the fidelity of augmented samples is not guaranteed under extreme sparsity of categorical levels or strong feature interactions. Third, DPS intervals are hypothesis-generating rather than causal guarantees; boundaries may vary with cross-validation folds, random seeds, or hyperparameter settings. Finally, SRF focuses on static tabular data; extending the approach to dynamic, streaming, or temporally correlated datasets requires further development. DPS yields decision-consistent operating windows from predictive structure; causal identification would require interventions or identifiable causal models.

Future directions include combining SRF with transfer learning to exploit related datasets, integrating ensemble-based uncertainty estimates with Bayesian calibration, and exploring hybrid causal-discovery pipelines that align DPS outputs with graphical models. These avenues could further strengthen scalability and industrial adoption, particularly in domains where interpretability and small-data reliability are paramount.

Conclusion

This study presented an SRF framework for predictive modelling in small, mixed-type manufacturing datasets. The framework integrates a leakage-safe pipeline, a classification-assisted augmentation strategy, automatic forest-size optimisation, and a DPS mechanism for operating-window discovery. Evaluation across six datasets, including an industrial nickel-based superalloy case, demonstrated that SRF achieves competitive predictive accuracy relative to strong boosting baselines while offering practitioner-oriented interpretability.

Key findings can be summarised as follows. First, augmentation combined with missForest substantially improves performance on small and noisy datasets, while forest-size optimisation stabilises variance and mitigates overfitting. Second, DPS translates ensemble structure into interpretable optimal and avoidance ranges, providing actionable guidance beyond feature-attribution tools. Third, robustness diagnostics confirmed that observed improvements stem from fold-contained design rather than data leakage, with consistent behaviour across random seeds and cross-validation folds. Finally, comparative analysis showed that SRF is most effective in conditions where boosting methods exhibit instability and deep learning approaches require larger sample sizes.

While the proposed framework is motivated by manufacturing applications and validated on an industrial investment casting dataset, this dataset remains relatively small. The additional public benchmark datasets are therefore used primarily to demonstrate methodological robustness under mixed-type and limited-data conditions, rather than to claim broad industrial generalisability. As a result, applicability to other manufacturing contexts should be interpreted cautiously and confirmed through future studies involving larger and more diverse industrial datasets.

Several limitations merit attention. Augmentation sensitivity increases under extreme sparsity of categorical levels, DPS-derived ranges are hypothesis-generating rather than causally identified, and gains diminish on large, continuous-only datasets. The classification-assisted augmentation should therefore be viewed as a conditional stabilisation mechanism rather than a universally applicable resampling strategy, and its use must be justified on a dataset-by-dataset basis. Future work will explore integration with transfer learning and causal-graphical approaches, as well as extensions to dynamic or streaming manufacturing data. Overall, SRF advances the toolkit for industrial predictive modelling by balancing accuracy, interpretability, and robustness under small-data constraints.

Table 10 Comparison between CausalNex and SRF

Algorithm	SRF	CausalNex
Domain Expertise	Minimal domain knowledge requirements.	High dependency on domain knowledge as a result of hybrid learning with data and domain expertise.
Non-Linear Interaction Within Features	Point of strength: dealing with the non-linearity in data.	Can handle non-linearity in the data.
Causal Relationships	Generates importance ranking and causal relationships for provided response range.	Provides a generic cause for the overall process.
Optimal Operating Range	Provides the optimal operating range for the given thresholds.	No optimal range is given.
Computational Cost	High performance and computationally efficient; more suitable for real-time defect prediction and heterogeneous datasets.	Computationally expensive.
Minimum Dataset Size	According to testing on several datasets, it is suggested to use at least 350 samples in order to achieve reasonable accuracy.	Based on the performed benchmarking, it is suggested to use at least 1000 samples in order to achieve reasonable accuracy.

Appendix A. CausalNex

CausalNex (Quantumblack, 2020), is a hybrid learning technique with data and domain expertise that helps encode substantial domain knowledge in models to ensure the correct causal relationship is found while avoiding spurious relationships. CausalNex, which uses Bayesian analysis, converts continuous data into ordered categories (e.g. very low, low, medium, high, very high). Table 10 presents a comparison between SRF algorithm and CausalNex. Because the deliverables of each algorithm are different (e.g. regarding whether or not it provides an optimal range), a numerical comparison between the two could not be conducted. Nonetheless, Table 10 highlights some differences between the two approaches and identifies the areas where the SRF has an advantage over CausalNex.

Appendix B. Full Pseudocode and Variable Glossary

This appendix provides an expanded pseudocode for the core components of the SRF framework referenced in the main text (Algorithms 4–9). Each algorithm includes inputs/outputs, leakage controls, and computational notes. Detailed DPS procedure is described in Algorithm 9. A consolidated glossary of symbols is provided in Table 11. Figures 18 and 19 describe the steps used in Algorithm 7.

Algorithm 4: SRF: Fold-contained pipeline for mixed-type tabular regression

Input: Labeled dataset $L = \{(X, y)\}$ with continuous/categorical features, K (folds), seeds $\mathcal{S}$
Output: Trained forest(s), out-of-fold predictions, diagnostics, optional DPS ranges

for *outer fold* $k = 1, \ldots, K$ **do**
  Split $L \to (L^{(k)}_{\text{train}}, L^{(k)}_{\text{val}})$ using seed s_0; no record overlap.
  Fit categorical encoder on $L^{(k)}_{\text{train}}$ and transform both train/val.
  Fit *missForest* on $L^{(k)}_{\text{train}}$ and impute train/val with that fitted imputer.
  if *augmentation enabled* **then**
    Inside the training fold only: call Algorithm 5 to obtain $(X'_{\text{train}}, y'_{\text{train}})$.
  end
  for *seed* $s \in \mathcal{S}$ **do**
    Train RF on $L^{(k)}_{\text{train}}$ (or augmented set) across a small grid of $n_{\text{estimators}}$ and m_{leaf}.
    Select forest size within a *stability corridor*: prefer the smaller model unless the larger one yields a consistent, seed-stable improvement above a minimal threshold.
  end
  With the selected configuration, refit RF on $L^{(k)}_{\text{train}}$ and predict $L^{(k)}_{\text{val}}$; store out-of-fold predictions.
  Optionally run DPS (Algorithm 6) on the fitted forest to obtain optimal/avoidance ranges.
end
Aggregate metrics from out-of-fold predictions; compute mean±sd, 95% CIs, paired tests.
Leakage controls: all encoders, imputers, augmentation, and tuning are fit *inside* training folds; validation folds are never used for fitting or model selection.
Complexity: dominated by missForest and RF training; per-fold time $O\big(B\, n'\, m_{\text{try}} \log n'\big)$ for RF and $O\big(I \sum_{v \in \mathcal{M}} B_v\, n\, m_{\text{try}} \log n\big)$ for missForest.

Algorithm 5: Classification-assisted augmentation for regression (quantiles → SMOTE in feature space → missForest)

Input: Training-fold features X, response y, bins G (e.g., quartiles), SMOTE k, per-bin cap ρ
Output: Augmented $(X_{\text{new}}, y_{\text{new}})$

Compute bins $b(y) \in \{1, \ldots, G\}$ by response quantiles.
for *each bin* g **do**
  Let $\mathcal{I}_g = \{i : b(y_i) = g\}$; apply SMOTE in *feature space* to $\{X_i : i \in \mathcal{I}_g\}$ with k-NN;
  Limit synthetic count in bin g to $\le \rho \cdot |\mathcal{I}_g|$; drop syntheses with large Mahalanobis distance; clip continuous features to training-fold fences.
end
Form $(\tilde{X}, \tilde{y} = \text{NA})$ from all bins, concatenate with (X, y), and run *missForest* to impute $\widehat{\tilde{y}}$.
Return $X_{\text{new}} = X \cup \tilde{X}$ and $y_{\text{new}} = y \cup \widehat{\tilde{y}}$.
Notes: choose G=4 for very small dataset; reduce synthesis for insignificant categories; disable if learning curves/variance worsen.
Leakage: execute entirely *inside* the training fold.

Algorithm 6: DPS: Decision Path Search for optimal/avoidance operating ranges

Input: Trained RF with B trees; per-leaf predictions $\hat{y}_\ell$; penalty scaling $(T_{\min}, T_{\max})$
Output: Per-factor optimal and avoidance ranges; separation scores

for *tree* $b = 1, \ldots, B$ **do**
  Collect leaves $\{\ell\}$; for each leaf get prediction $\hat{y}_\ell$ and decision path $\mathcal{P}_\ell$ (intervals for continuous; level sets for categoricals).
  Map $\hat{y}_\ell$ to penalty $\text{PM}_\ell \in [0, 1]$ using $(T_{\min}, T_{\max})$; compute within-tree $q_{0.25}$.
  Tag leaf as *optimal* if $\text{PM}_\ell \le q_{0.25}$, else *avoidance*; append per-factor conditions from $\mathcal{P}_\ell$ into the corresponding pool.
end
For each factor f, merge overlapping optimal (and avoidance) intervals; for categorical f, take level modes.
Compute a separation score s_f (e.g., non-overlap fraction of optimal vs avoidance over total span) and rank factors.
Output: consolidated optimal and avoidance ranges by factor; s_f ranking; overlap cases as in interval taxonomy.
Complexity: path extraction $O(\sum_b L_b)$; aggregation $O(p \sum_b L_b)$.

Table 11 Variable glossary for Algorithms 4–6

Symbol	Meaning
X, y	Feature matrix and response vector in a fold
K	Number of outer cross-validation folds
$\mathcal{S}$	Set of random seeds for stability auditing
G	Number of response bins (quantiles) for augmentation
k	SMOTE neighborhood size
ρ	Per-bin synthesis cap (ratio)
B	Number of trees in the random forest
m_{try}	Features considered at each split
L_b	Number of leaves in tree b
$T_{\min}$, $T_{\max}$	Penalty scaling bounds (low=good; high=bad)
s_f	DPS separation score for factor f

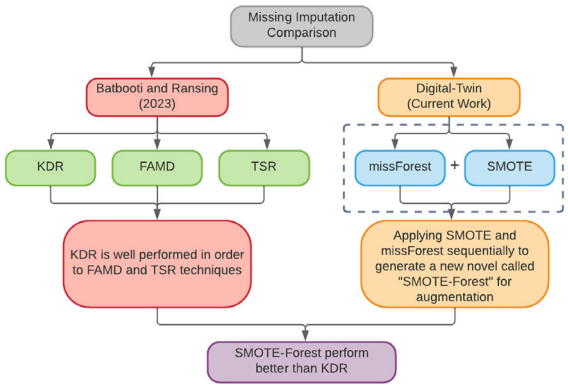


Fig. 18 SRF Missing Imputation Technique vs. Other Techniques

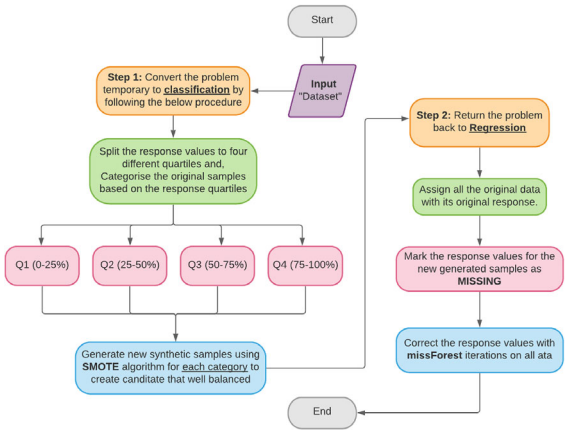


Fig. 19 SRF Regression oversampling technique

Algorithm 7: SRF Missing Imputation Technique

Input: Imbalanced training X, Target response Y
Output: Over-sampled X_{new},Y_{new}

Calculate quartiles for response column
Categorise original samples based on response quartiles
X_{new}, Y_{new} =SMOTE(X, Y); `// Generate new samples using SMOTE`
Mark response values for the new generated samples as missing
Correct response value Y_{new} with missForest iterations for all data
return X_{new}, Y_{new}

Algorithm 8: Penalty Matrix (PM) Algorithm (Ransing et al., 2013)

Function `PenaltyMatrix`$(\hat{Y}, Th_{min}, Th_{max})$**:**
  $D = 0, E = 1$
  `// For lower the better case (LB), while` $D = 1$ `and` $E = 0$ `for higher the better (HB)`
  $V \leftarrow$ size of $(\hat{Y})$
  $PM \leftarrow$ Array of zeros(V)
  for j=1 $\longrightarrow$ V **do**
    if $\hat{Y} \leq Th_{min}$ **then**
      $PM = D$
    else if $\hat{Y} \geq Th_{max}$ **then**
      $PM = E$
    else
      $PM = \frac{\hat{Y} - Th_{min}}{Th_{max} - Th_{min}}$
    end
  end
  return PM
End Function

Algorithm 9: Detailed DPS Procedure

Fit a Random Forest on the training set L
Optimal-Intervals=[]
Avoidance-Intervals=[]
for $b = 1$ *to* B **do**
  Leafs = Collect all terminal nodes of tree b
  $PM_{Leafs} \leftarrow$ PenaltyMatrix($\hat{Y}, Th_{min}, Th_{max}$) (from Algorithm 8)
  $PM_{Q25} \leftarrow$ Quartile(PM_{Leafs},0.25)
  Optimal-Leaf=[]
  Avoidance-Leaf=[]
  for $i = 1, ..., n_{leafs}$ **do**
    if $PM[i] < PM_{Q25}$ **then**
      Optimal-Leaf $\leftarrow$ leaf[i]
      Draw a route from leaf[i] to root node
      Optimal-Intervals[] $\leftarrow$ Append all optimal interval for each factor `// factor split threshold value`
    else
      Avoidance-Leaf $\leftarrow$ leaf[i]
      Draw a route from leaf[i] to root node
      Avoidance-Intervals [] $\leftarrow$ Append all avoidance interval for each factor `// factor split threshold value`
    end
  end
end
Optimal-Interval=[] `// estimate the optimal intervals for every factor`
Avoidance-Interval=[] `// estimate the avoidance intervals for every factor`
for j=*1 to* n **do**
  $f \leftarrow X[j]$
  if f *is quantitative* **then**
    Optimal-Interval[j]= mean(Optimal-Intervals[j])
    Avoidance-Interval=[j]= mean(Avoidance-Intervals[j])
  else
    Optimal-Interval[j]= mode(Optimal-Intervals[j])
    Avoidance-Interval=[j]= mode(Avoidance-Intervals[j])
  end
end
Compare optimal and avoidance intervals

Appendix C. Hyperparameter Grids, Sensitivity Corridors, and Compute Profile

This appendix consolidates technical details that complement Sections Statistical and robustness protocols, Computational complexity and scalability, and Baselines and hyperparameters: (i) full hyperparameter grids explored for all learners, (ii) extended sensitivity and seed-stability steps for the RF/SRF backbone, and (iii) compute environment and runtime profile.

Table 12 Hyperparameter grids used in cross-validation sweeps. Values shown are all candidates tried inside CV

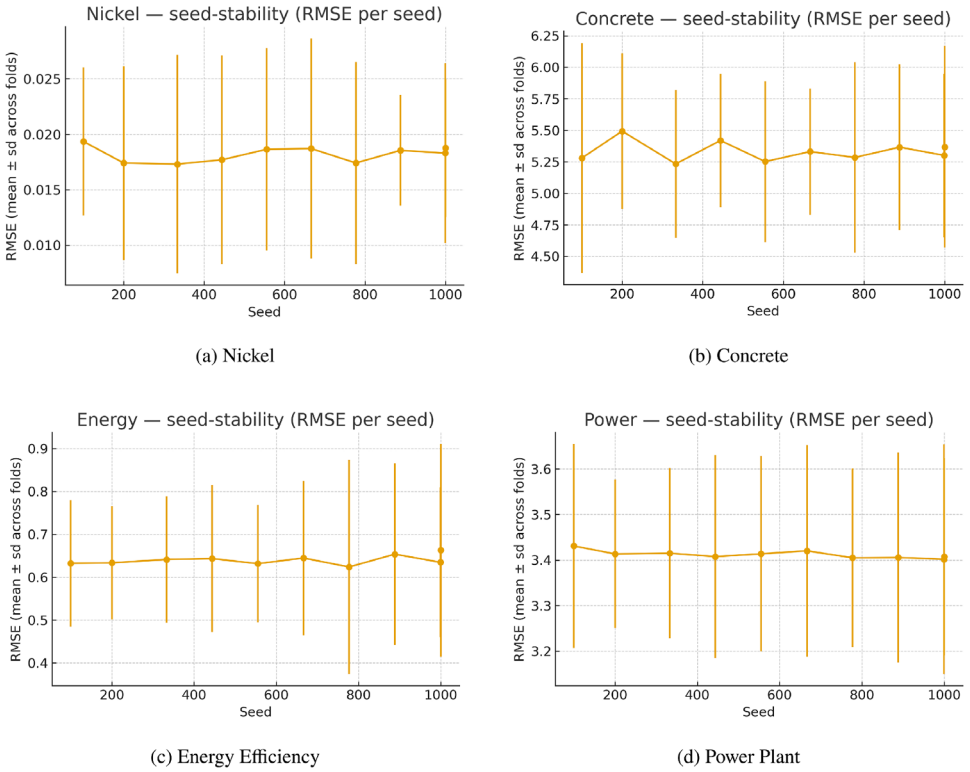
Method / Parameter	Grid values
RF / SRF	
`n_estimators`	{50, 100, 200, 300, 500}
`max_features`	{$\sqrt{p}$, 0.5, auto}
`min_samples_leaf`	{1, 2, 3, 5}
XGBoost	
`n_estimators`	{300, 600, 900}
`learning_rate`	{0.03, 0.05, 0.1}
`max_depth`	{3, 6, 8}
`subsample`	{0.7, 0.9, 1.0}
`colsample_bytree`	{0.7, 0.9, 1.0}
`reg_lambda`	{0, 1, 5}
LightGBM	
`num_leaves`	{31, 63, 127}
`max_depth`	{-1, 6, 10}
`learning_rate`	{0.03, 0.05, 0.1}
`feature_fraction`	{0.7, 0.9, 1.0}
`bagging_fraction`	{0.7, 0.9, 1.0}
CatBoost	
`iterations`	{500, 1000}
`depth`	{4, 6, 8}
`learning_rate`	{0.03, 0.05, 0.1}
`l2_leaf_reg`	{1, 3, 5}

C.1 Full Hyperparameter Grids

Table 12 reports the full set of hyperparameter candidates evaluated during cross-validation (CV). The grids cover key structural, learning, and regularization parameters for each model, ensuring systematic search across comparable configurations while maintaining computational efficiency.

Sensitivity and Seed-Stability Tests

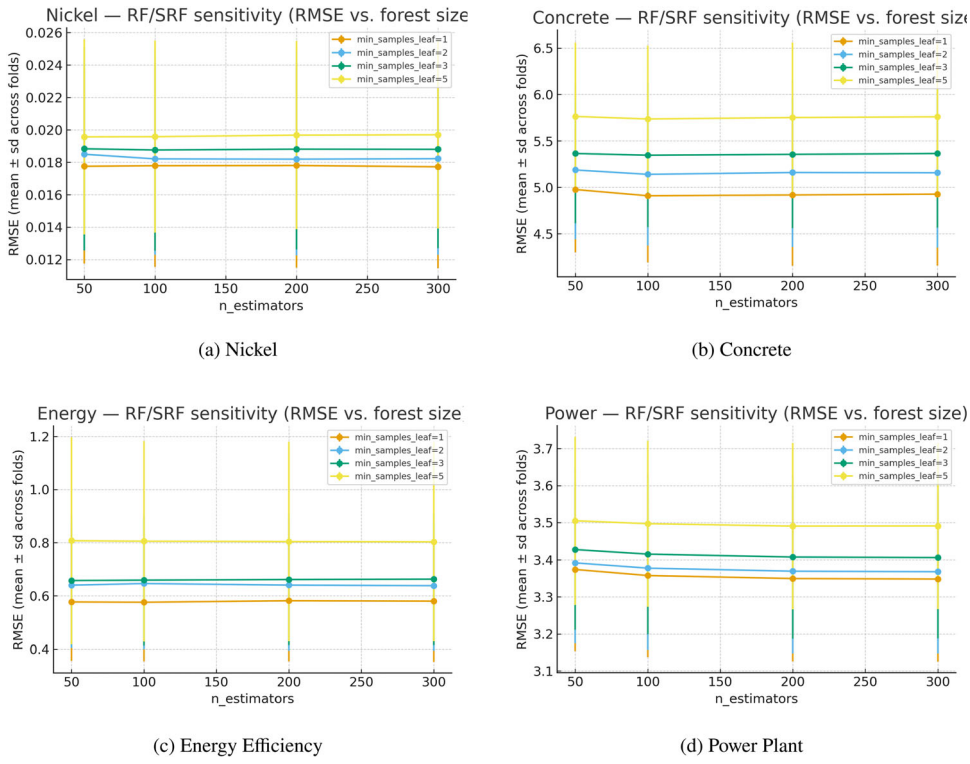
Sensitivity analyses extend Figure 17 in the main text. For each dataset, RMSE was tracked against forest size and seeds to audit stability. Representative additional panels are given in Figures 20 and 21 (Nickel, Concrete, and Power Plant).



(a) Nickel (b) Concrete

(c) Energy Efficiency (d) Power Plant

Fig. 20 Seed-stability diagnostics: RMSE mean±sd per seed (10-fold CV). Panels show variance across seeds for Nickel, Power Plant, Concrete, and Energy Efficiency datasets



(a) Nickel (b) Concrete

(c) Energy Efficiency (d) Power Plant

Fig. 21 Extended sensitivity tests: RMSE vs. forest size (error bars = fold sd) with separate lines for `min_samples_leaf`

C.3 Compute Environment and Runtime Profile

Experiments were run on a workstation with 32 cores (Intel Xeon Silver), 128 GB RAM, and Ubuntu 22.04, Python 3.13, scikit-learn 1.5, CatBoost 1.2, LightGBM 4.2, and XGBoost 2.0. Parallelisation was enabled across CV folds and RF trees. Typical wall-clock times per dataset under 10-fold CV were:

- Nickel (60→360 records after augmentation): ≈3 minutes.
- Concrete (1030 records): ≈7 minutes.
- Energy Efficiency (768 records): ≈6 minutes.
- Power Plant (9568 records): ≈40 minutes.
- Student–Mathematics (395 records): ≈5 minutes.
- Student–Portuguese (649 records): ≈6 minutes.

Runtime was dominated by missForest iterations and forest training; DPS extraction added $< 5\%$ overhead. Augmentation was capped at $n'/n \leq 6$ to control memory footprint.

Appendix D. Extended Statistics: Per-Fold Summaries & Paired Tests

This appendix presents fold-wise metrics, confidence intervals, and paired t-tests for all datasets and methods. Results are programmatically exported from the evaluation pipeline and correspond to the main-text summaries. Pairwise significance is assessed via a paired t-test and Wilcoxon signed-rank test on fold-wise metric differences.

D.1 Per-fold CV Statistics (All Datasets)

Table 13 summarizes the cross-validated performance for each dataset and model. Values are mean±std across K folds with two-sided 95% CIs, computed from out-of-fold predictions under the leakage-safe protocol.

Table 13 Cross-validated statistics by dataset, model, and metric.

Dataset	Model	Metric	K	Mean±Std	95% CI
Concrete	BaselineRF	R^2	10	0.903±0.025	[0.886, 0.921]
Concrete	CatBoost	R^2	10	0.927±0.018	[0.914, 0.940]
Concrete	LightGBM	R^2	10	0.931±0.019	[0.918, 0.945]
Concrete	SRF	R^2	10	0.903±0.025	[0.886, 0.921]
Concrete	XGBoost	R^2	10	0.933±0.017	[0.921, 0.945]
Concrete	BaselineRF	RMSE	10	5.133±0.707	[4.627, 5.638]
Concrete	CatBoost	RMSE	10	4.459±0.639	[4.002, 4.916]
Concrete	LightGBM	RMSE	10	4.333±0.702	[3.831, 4.835]
Concrete	SRF	RMSE	10	5.133±0.707	[4.627, 5.638]
Concrete	XGBoost	RMSE	10	4.264±0.640	[3.806, 4.721]
Energy Efficiency	BaselineRF	R^2	10	0.997±0.001	[0.996, 0.997]
Energy Efficiency	CatBoost	R^2	10	0.996±0.007	[0.991, 1.000]
Energy Efficiency	LightGBM	R^2	10	0.997±0.003	[0.994, 0.999]
Energy Efficiency	SRF	R^2	10	0.997±0.001	[0.996, 0.997]
Energy Efficiency	XGBoost	R^2	10	0.996±0.007	[0.991, 1.000]
Energy Efficiency	BaselineRF	RMSE	10	0.573±0.090	[0.509, 0.637]
Energy Efficiency	CatBoost	RMSE	10	0.536±0.280	[0.336, 0.737]
Energy Efficiency	LightGBM	RMSE	10	0.522±0.175	[0.397, 0.648]
Energy Efficiency	SRF	RMSE	10	0.573±0.090	[0.509, 0.637]
Energy Efficiency	XGBoost	RMSE	10	0.569±0.267	[0.378, 0.759]
Nickel	BaselineRF	R^2	10	0.269±0.556	[-0.129, 0.667]
Nickel	CatBoost	R^2	10	0.906±0.053	[0.868, 0.943]
Nickel	LightGBM	R^2	10	0.898±0.062	[0.853, 0.943]
Nickel	SRF	R^2	10	0.871±0.086	[0.810, 0.933]
Nickel	XGBoost	R^2	10	0.910±0.056	[0.870, 0.950]
Nickel	BaselineRF	RMSE	10	0.051±0.034	[0.026, 0.075]
Nickel	CatBoost	RMSE	10	0.017±0.005	[0.013, 0.020]
Nickel	LightGBM	RMSE	10	0.017±0.005	[0.013, 0.021]
Nickel	SRF	RMSE	10	0.019±0.006	[0.014, 0.023]
Nickel	XGBoost	RMSE	10	0.016±0.005	[0.012, 0.020]
Power Plant	BaselineRF	R^2	10	0.962±0.008	[0.957, 0.968]
Power Plant	CatBoost	R^2	10	0.957±0.006	[0.952, 0.961]
Power Plant	LightGBM	R^2	10	0.960±0.006	[0.956, 0.965]
Power Plant	SRF	R^2	10	0.962±0.008	[0.957, 0.968]
Power Plant	XGBoost	R^2	10	0.966±0.006	[0.961, 0.970]
Power Plant	BaselineRF	RMSE	10	3.299±0.312	[3.076, 3.522]
Power Plant	CatBoost	RMSE	10	3.543±0.231	[3.378, 3.708]
Power Plant	LightGBM	RMSE	10	3.387±0.234	[3.219, 3.554]
Power Plant	SRF	RMSE	10	3.299±0.312	[3.076, 3.522]
Power Plant	XGBoost	RMSE	10	3.152±0.239	[2.981, 3.323]
Student–Mathematics	BaselineRF	R^2	10	0.848±0.106	[0.773, 0.924]
Student–Mathematics	CatBoost	R^2	10	0.852±0.105	[0.777, 0.926]
Student–Mathematics	LightGBM	R^2	10	0.835±0.108	[0.758, 0.912]
Student–Mathematics	SRF	R^2	10	0.848±0.106	[0.773, 0.924]
Student–Mathematics	XGBoost	RMSE	10	1.602±0.551	[1.208, 1.996]
Student–Portuguese	BaselineRF	R^2	10	0.834±0.065	[0.788, 0.880]
Student–Portuguese	CatBoost	R^2	10	0.823±0.090	[0.759, 0.887]

Table 13 continued

Dataset	Model	Metric	K	Mean±Std	95% CI
Student–Portuguese	LightGBM	R^2	10	0.805±0.088	[0.742, 0.868]
Student–Portuguese	SRF	R^2	10	0.833±0.091	[0.769, 0.898]
Student–Portuguese	XGBoost	RMSE	10	1.250±0.292	[1.041, 1.459]
Student–Portuguese	CatBoost	RMSE	10	1.280±0.428	[0.974, 1.586]
Student–Portuguese	LightGBM	RMSE	10	1.343±0.347	[1.095, 1.592]
Student–Portuguese	SRF	RMSE	10	1.232±0.402	[0.944, 1.519]
Student–Portuguese	XGBoost	RMSE	10	1.299±0.402	[1.011, 1.587]

Values are mean±sd across K-folds with two-sided 95% CIs computed as $\bar{x} \pm t_{K-1,0.975} \frac{s}{\sqrt{K}}$ from out-of-fold predictions under the leakage-safe protocol. For R^2 (bounded on [0, 1]), CIs are reported with truncation to [0, 1]

D.2 Paired Tests (All Datasets)

Table 14 reports paired t-tests across folds for SRF vs. each baseline per dataset and metric. Negative t-values indicate SRF is lower than the baseline on the given metric; for RMSE (lower is better), negative is favourable for SRF; for R^2 (higher is better), positive is favourablefor SRF.

Table 14 Paired t-tests by dataset, metric, and comparison

Dataset	Metric	Comparison	n pairs	t-stat	p-value	Direction
Concrete	R^2	SRF vs BaselineRF	10	–	–	identical
Concrete	R^2	SRF vs CatBoost	10	-6.030	$< 10^{-3}$	higher_is_better
Concrete	R^2	SRF vs LightGBM	10	-9.143	$< 10^{-3}$	higher_is_better
Concrete	R^2	SRF vs XGBoost	10	-8.082	$< 10^{-3}$	higher_is_better
Concrete	RMSE	SRF vs BaselineRF	10	–	–	identical
Concrete	RMSE	SRF vs CatBoost	10	6.053	$< 10^{-3}$	lower_is_better
Concrete	RMSE	SRF vs LightGBM	10	9.484	$< 10^{-3}$	lower_is_better
Concrete	RMSE	SRF vs XGBoost	10	8.073	$< 10^{-3}$	lower_is_better
Energy Efficiency	R^2	SRF vs BaselineRF	10	–	–	identical
Energy Efficiency	R^2	SRF vs CatBoost	10	-0.933	0.375	higher_is_better
Energy Efficiency	R^2	SRF vs LightGBM	10	-2.039	0.072	higher_is_better
Energy Efficiency	R^2	SRF vs XGBoost	10	-0.742	0.477	higher_is_better
Energy Efficiency	RMSE	SRF vs BaselineRF	10	–	–	identical
Energy Efficiency	RMSE	SRF vs CatBoost	10	1.645	0.134	lower_is_better
Energy Efficiency	RMSE	SRF vs LightGBM	10	2.188	0.056	lower_is_better
Energy Efficiency	RMSE	SRF vs XGBoost	10	1.257	0.240	lower_is_better
Nickel	R^2	SRF vs BaselineRF	10	3.542	0.006	higher_is_better
Nickel	R^2	SRF vs CatBoost	10	-1.512	0.165	higher_is_better
Nickel	R^2	SRF vs LightGBM	10	-1.472	0.175	higher_is_better
Nickel	R^2	SRF vs XGBoost	10	-2.359	0.043	higher_is_better
Nickel	RMSE	SRF vs BaselineRF	10	-3.243	0.010	lower_is_better
Nickel	RMSE	SRF vs CatBoost	10	1.546	0.156	lower_is_better
Nickel	RMSE	SRF vs LightGBM	10	1.477	0.174	lower_is_better
Nickel	RMSE	SRF vs XGBoost	10	2.786	0.021	lower_is_better
Power Plant	R^2	SRF vs BaselineRF	10	–	–	identical
Power Plant	R^2	SRF vs CatBoost	10	6.740	$< 10^{-3}$	higher_is_better
Power Plant	R^2	SRF vs LightGBM	10	-1.078	0.309	higher_is_better
Power Plant	R^2	SRF vs XGBoost	10	-17.864	$< 10^{-3}$	higher_is_better
Power Plant	RMSE	SRF vs BaselineRF	10	–	–	identical
Power Plant	RMSE	SRF vs CatBoost	10	-7.077	$< 10^{-3}$	lower_is_better
Power Plant	RMSE	SRF vs LightGBM	10	1.223	0.252	lower_is_better
Power Plant	RMSE	SRF vs XGBoost	10	16.130	$< 10^{-3}$	lower_is_better
Student–Mathematics	R^2	SRF vs BaselineRF	10	–	–	identical
Student–Mathematics	R^2	SRF vs CatBoost	10	-0.192	0.852	higher_is_better
Student–Mathematics	R^2	SRF vs LightGBM	10	0.941	0.371	higher_is_better
Student–Mathematics	R^2	SRF vs XGBoost	10	0.141	0.891	higher_is_better
Student–Mathematics	RMSE	SRF vs BaselineRF	10	–	–	identical
Student–Mathematics	RMSE	SRF vs CatBoost	10	-0.488	0.637	lower_is_better
Student–Mathematics	RMSE	SRF vs LightGBM	10	-1.613	0.141	lower_is_better

Table 14 continued

Dataset	Metric	Comparison	n pairs	t-stat	p-value	Direction
Student–Mathematics	RMSE	SRF vs XGBoost	10	-0.568	0.584	lower_is_better
Student–Portuguese	R^2	SRF vs BaselineRF	10	-0.025	0.981	higher_is_better
Student–Portuguese	R^2	SRF vs CatBoost	10	0.946	0.369	higher_is_better
Student–Portuguese	R^2	SRF vs LightGBM	10	2.071	0.068	higher_is_better
Student–Portuguese	R^2	SRF vs XGBoost	10	1.925	0.086	higher_is_better
Student–Portuguese	RMSE	SRF vs BaselineRF	10	-0.105	0.919	lower_is_better
Student–Portuguese	RMSE	SRF vs CatBoost	10	-1.147	0.281	lower_is_better
Student–Portuguese	RMSE	SRF vs LightGBM	10	-2.363	0.042	lower_is_better
Student–Portuguese	RMSE	SRF vs XGBoost	10	-1.859	0.096	lower_is_better

Table 15 Paired significance tests on fold-wise RMSE differences (SRF − Baseline RF) across six datasets (10-fold CV).

Dataset	Mean diff (SRF−RF)	p_t	p_{Wilcoxon}	Cohen's d
Nickel	−0.031822	0.011468	0.005859	−1.001
Concrete	0.000000	–	–	0.000
Energy Efficiency	0.000000	–	–	0.000
Power Plant	0.000000	–	–	0.000
Student–Mathematics	0.000000	–	–	0.000
Student–Portuguese	−0.018000	0.919000	–	−0.045

Negative mean difference indicates SRF lower RMSE (better). Both a paired t-test and Wilcoxon signed-rank test are reported. "–" = not applicable (identical OOF predictions)

D.3 Paired Tests (Wilcoxon + t-test) across datasets

Table 15 presents paired comparisons between SRF and RF using fold-wise RMSE differences. Mean differences (negative favour SRF), along with t-test, Wilcoxon test, and Cohen, indicate performance gaps and their statistical significance across datasets.

D.4 Reproducibility Notes

All metrics are computed from out-of-fold predictions under a K-fold cross-validation protocol with fixed seeds and fold-contained preprocessing (encoding, missForest imputation), optional augmentation (quantile-binning → SMOTE in feature space → missForest), and model fitting. No validation fold information is used for fitting or hyperparameter selection. Confidence intervals are two-sided 95% using Student's t distribution. Paired tests compare fold-wise SRF and baseline metrics using the same splits.

Appendix E. Leakage Sanity and Learning Curves

This appendix consolidates two diagnostic families across all datasets: (i) leakage sanity experiments that verify the fold-contained pipeline does not leak information, and (ii) learning curves that characterise generalisation as training size increases.

Table 16 Leakage sanity check: validation RMSE (mean ± sd) under four conditions.

Dataset	Baseline	NoiseOnly	PermutationY	TargetAsFeature
Concrete	5.143 ± 0.691	17.185 ± 0.849	18.160 ± 1.320	0.151 ± 0.091
Energy Efficiency	0.589 ± 0.093	10.307 ± 0.428	10.710 ± 0.533	0.083 ± 0.034
Nickel	0.051 ± 0.034	0.079 ± 0.022	0.079 ± 0.032	0.014 ± 0.012
Power Plant	3.303 ± 0.310	17.402 ± 0.217	17.667 ± 0.258	0.039 ± 0.041
Student–Mathematics	1.624 ± 0.548	4.702 ± 0.592	4.634 ± 0.595	0.088 ± 0.066
Student–Portuguese	1.254 ± 0.293	3.284 ± 0.525	3.308 ± 0.413	0.105 ± 0.058

Baseline is leakage-safe; NoiseOnly and PermutationY collapse performance; TargetAsFeature produces spuriously optimistic errors

E.1 Leakage Sanity

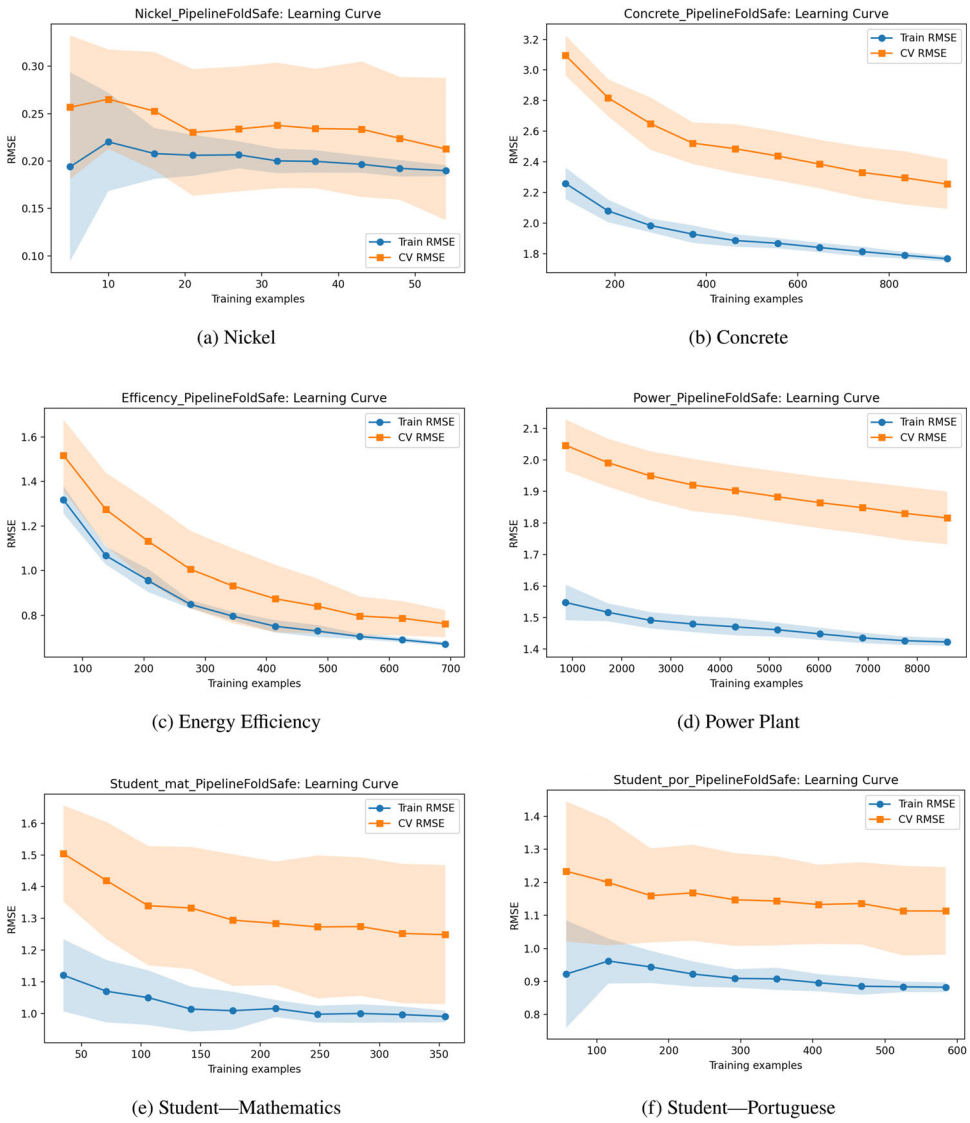
Each panel reports validation RMSE under four conditions: `Baseline` (leakage-safe), `NoiseOnly`, `PermutationY`, and `TargetAsFeature`. Performance collapses to chance when label information is destroyed and appears spuriously optimistic only when `TargetAsFeature` is enabled, confirming the efficacy of leakage controls. Table 16 reports a leakage sanity check using validation RMSE under four conditions. Baseline is leakage-safe, NoiseOnly and Permutation degrade performance, while TargetAsFeature shows artificially optimistic results to confirm correct leakage control.

E.2 Full Learning-Curve Panels

Figure 22 reports the extended learning curves generated under the fold-contained (PipelineFoldSafe) protocol for all six datasets. Each panel shows validation RMSE against the proportion of training data used (10%–100%), averaged over 10 CV folds.

Appendix F. Extended SHAP and DPS Visualizations

This appendix provides per-dataset interpretability panels that complement the main text: (i) SHAP summaries for feature attribution (additive Shapley explanations for tree



(a) Nickel

(b) Concrete

(c) Energy Efficiency

(d) Power Plant

(e) Student—Mathematics

(f) Student—Portuguese

Fig. 22 Full learning-curve panels for all six datasets under the fold-contained (PipelineFoldSafe) protocol. Curves show validation RMSE across increasing training fractions (10%–100%)

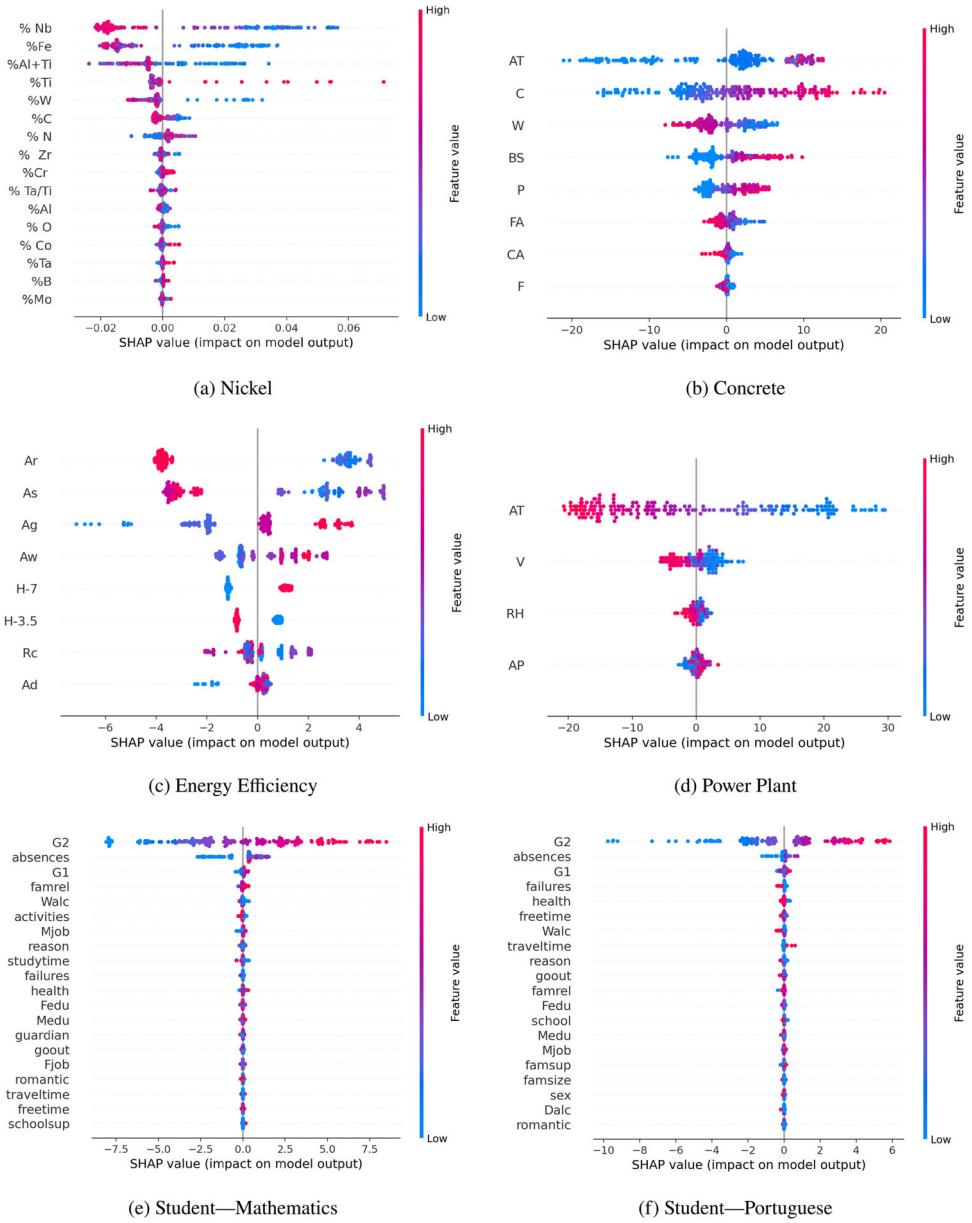
ensembles), (ii) DPS-derived optimal/avoidance operating-window visualizations aggregated from decision paths.

F.1 SHAP (Feature Attribution)

Figure 23 shows the extended SHAP summary plots for all six case studies. The plots explain how each input factor (feature) influences the model predictions. Each dot represents an individual data point. The color of each dot indicates the original factor value from the dataset, with red representing high factor values and blue representing low factor values. The position of each dot along the horizontal axis gives the SHAP value, which is calculated from the trained model and measures the contribution of that factor to the predicted output. Positive SHAP values increase the prediction, whereas negative SHAP values decrease it. Factors with larger spreads of SHAP values have a stronger effect on the model output and are therefore more important.

F.2 DPS Operating-Window Ranges

For each case study, the Figure 24 shows the original operating range and the discovered optimal or avoid range for each factor. The full-height bar represents the original operating range considered in the case study, while the shorter colored segment shows the range identified by the analysis. Cyan segments indicate optimal operating ranges, and black segments indicate ranges to avoid. Each panel corresponds to one factor, with the vertical axis showing the factor value and the horizontal positions separating the original range from the discovered range.



(a) Nickel

(b) Concrete

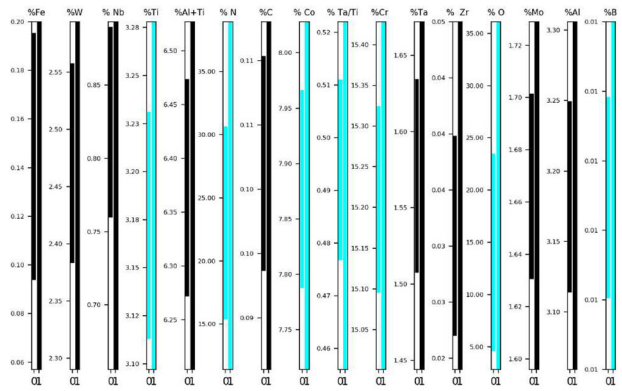
(c) Energy Efficiency

(d) Power Plant

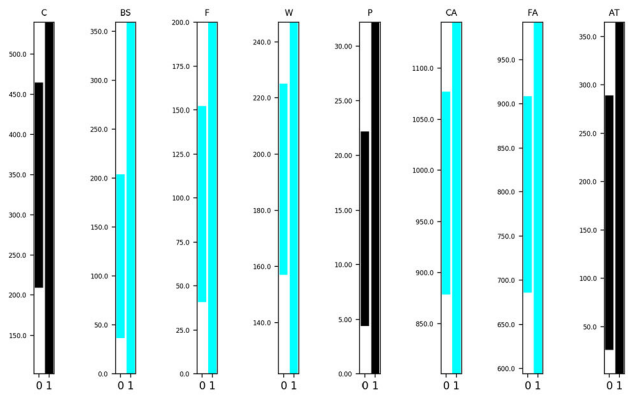
(e) Student—Mathematics

(f) Student—Portuguese

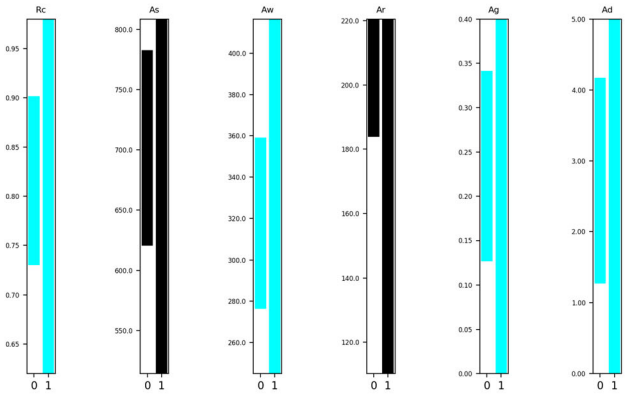
Fig. 23 Extended SHAP summary plots for all six datasets



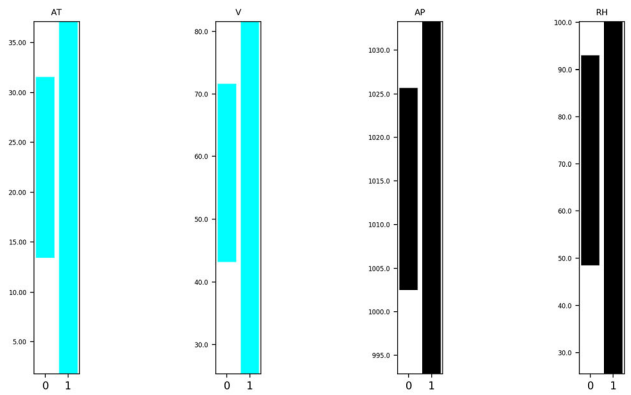
(a) Nickel



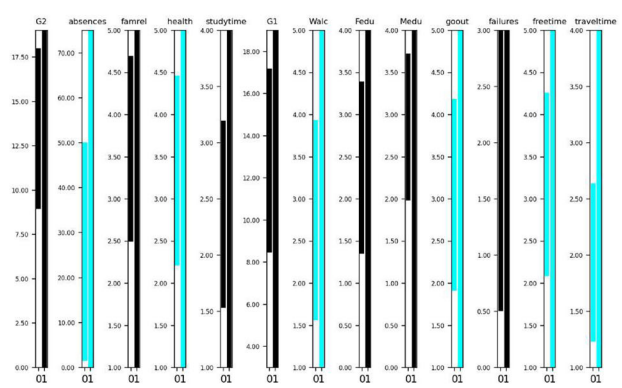
(b) Concrete



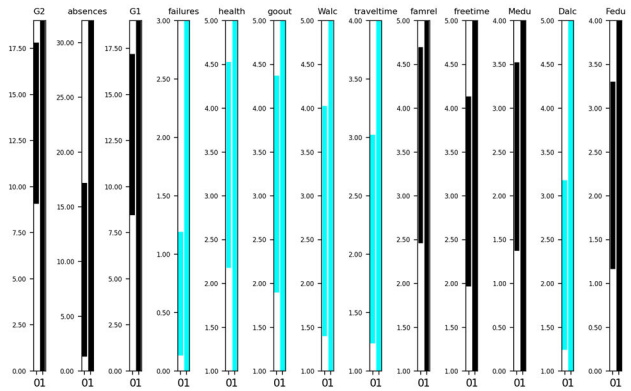
(c) Energy Efficiency



(d) Power Plant



(e) Student—Mathematics

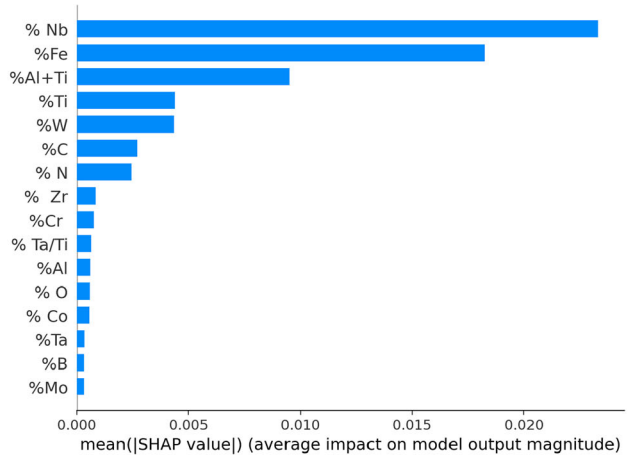


(f) Student—Portuguese

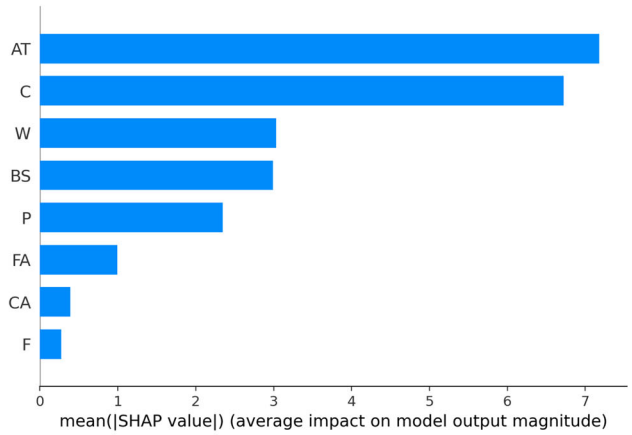
Fig. 24 DPS-derived optimal/avoidance operating-window ranges for all six datasets

F.3 SHAP (Bar Plots)

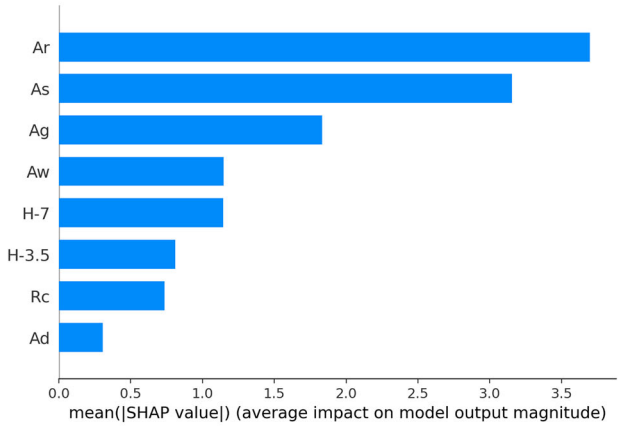
The SHAP bar plots in Figure 25 report the mean absolute SHAP values calculated across the validation predictions for each dataset. These values represent the average magnitude of each factor's contribution to the model output and provide a global ranking of factor importance.



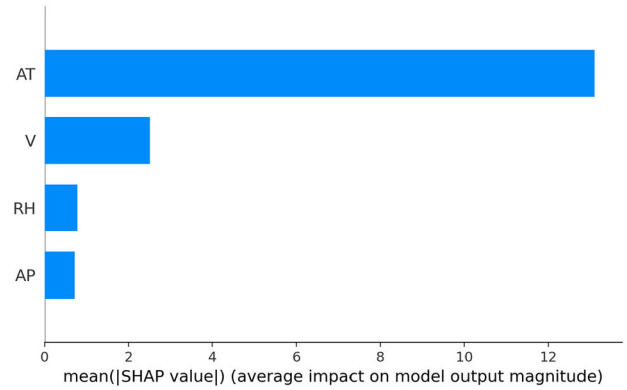
(a) Nickel



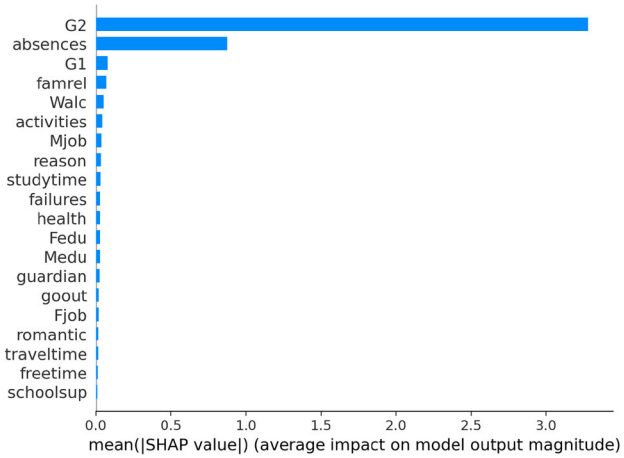
(b) Concrete



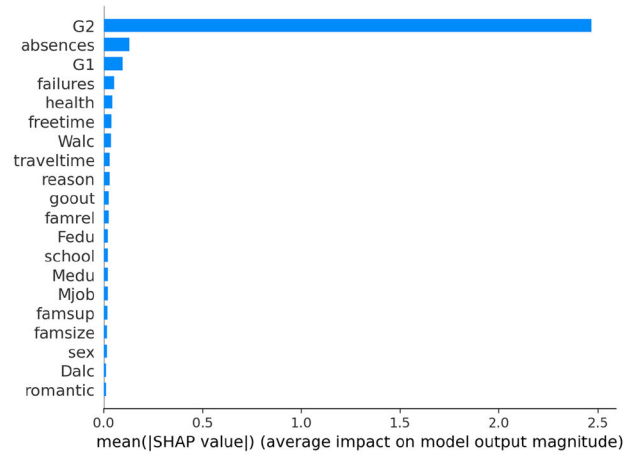
(c) Energy Efficiency



(d) Power Plant



(e) Student-Mathematics



(f) Student-Portuguese

Fig. 25 SHAP bar plots across six datasets

Appendix G. Extra Comparison 'Figures & Tables'

This appendix retains prior visualisations and tables for transparency while the main text focuses on the leakage-safe protocol and strong baselines.

G.1 Scatter/Regression Plots

Figure 26 reproduces the predicted–observed scatter for the industrial nickel-based superalloy dataset using the SRF model.

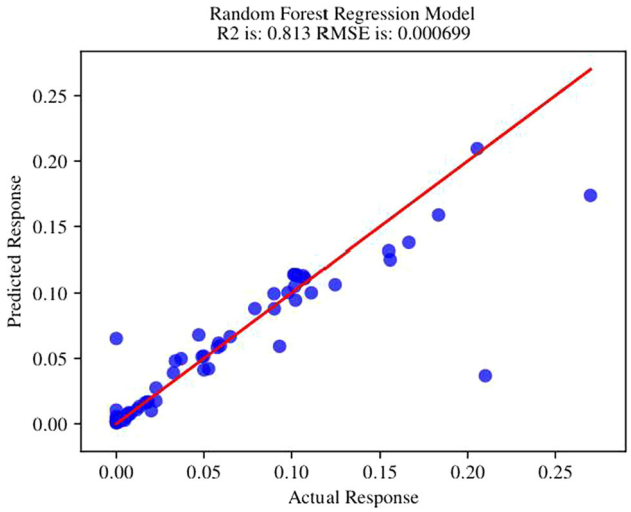


Fig. 26 SRF scatter: predicted vs. observed (Nickel)

G.2 Student Performance: Fold-wise RMSE and Method Comparison

Table 17 lists raw 10-fold RMSE values for the Student Performance datasets (Mathematics, Portuguese). In the updated results (Table 6), all methods are re-evaluated under a single leakage-safe protocol with confidence intervals and paired tests.

Table 18 presents a comparison that includes SVM, NN, Naive Bayes (NV), Decision Tree (DT), RF, and SRF. These figures come from a non-unified evaluation context and may not reflect the leakage-safe pipeline or tuned booster baselines used in the main results.

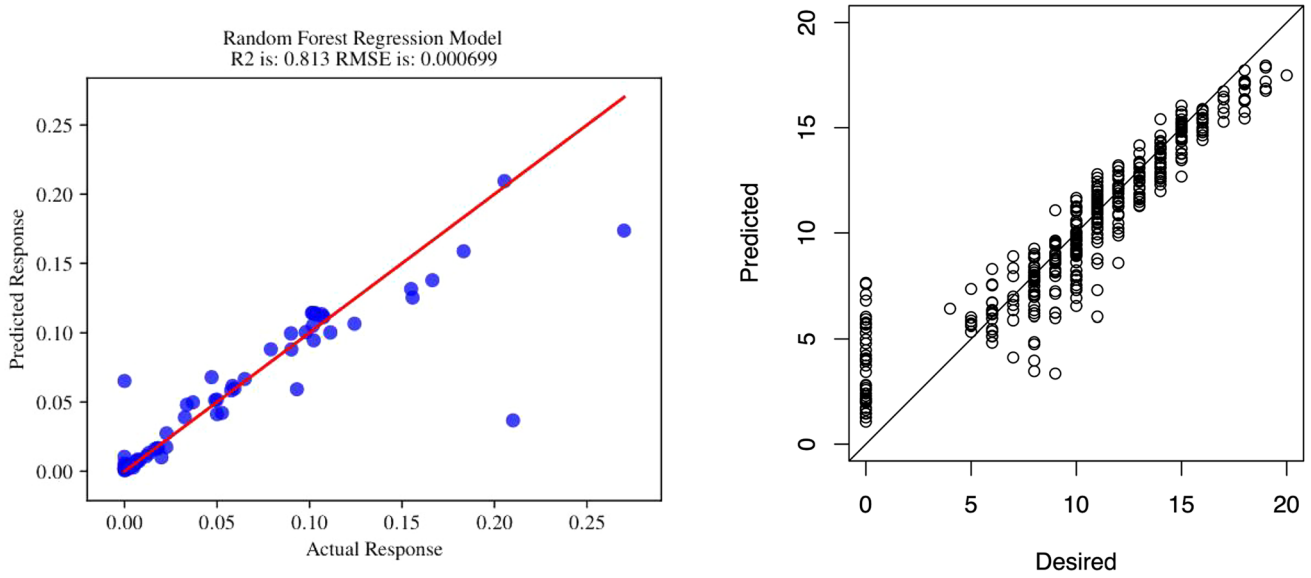
Figure 27 reproduces the panels contrasting SRF predictions with a reference regression plot from the public report on the Student dataset. This paper provides unified SHAP, PD/ICE, DPS comparisons and cross-validated metrics in Section DPS vs. SHAP and PD/ICE and **Appendix** *F*.

Table 17 RMSE for 10-fold CV on Student Performance (Math/Portuguese)

	1	2	3	4	5	6	7	8	9	10	**Mean**
$RMSE_{Math}$	1.76	1.12	1.65	2.30	1.13	1.89	2.08	1.79	1.44	1.35	1.65
$RMSE_{Port}$	1.29	1.52	0.99	0.92	1.85	1.52	0.72	1.11	0.99	1.18	1.21

Table 18 RMSE comparison for different algorithms (Student Performance)

Algorithm	SVM	NN	NV	DT	RF	**SRF**
$RMSE_{Math}$	2.09	2.05	2.01	1.94	1.75	1.65
$RMSE_{Port}$	1.35	1.36	1.32	1.46	1.32	1.21



(a) SRF prediction.

(b) Reference regression panel.

Fig. 27 Student Performance: regression panels for Mathematics

Acknowledgements The authors gratefully acknowledge the support of Swansea University, whose academic environment and resources were instrumental in the development of this work. The first author also extends sincere thanks to Mr Adel Alsaraawi, former Vice President of the State Audit Bureau (SAB) in Kuwait, for his invaluable encouragement, guidance, and unwavering support throughout the course of this research.

Author contributions Both authors contributed equally to this work.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors.

Data availability The research code required to reproduce all experiments and examples will be made openly available upon acceptance of the paper.

Materials Availability Not applicable.

Code availability The research code required to reproduce all experiments and examples will be made openly available upon acceptance of the paper.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval and consent to participate This study did not involve experiments on humans or animals performed by any of the authors. Ethical approval and informed consent were therefore not required.

Consent for publication Not applicable.

References

Arik, S. Ö., & Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*, 6679–6687. https://doi.org/10.1609/aaai.v35i8.16826

Arowolo, M. O., Adebiyi, M. O., & Adebiyi, A. A. (2020). An efficient pca ensemble learning approach for prediction of rna-seq malaria vector gene expression data classification. *International Journal of Engineering Research and Technology, 13*(1), 163–169. https://doi.org/10.37624/ijert/13.1.2020.163-169

Batbooti, R. S., & Ransing, R. S. (2023). A novel imputation based predictive algorithm for reducing common cause variation from small and mixed datasets with missing values

Batbooti, R. S., Ransing, R. S., & Ransing, M. R. (2017). A bootstrap method for uncertainty estimation in quality correlation algorithm for risk based tolerance synthesis. *Computers & Industrial Engineering, 112*, 654–662. https://doi.org/10.1016/j.cie.2016.12.041

Bordekar, H., Cersullo, N., Brysch, M., Philipp, J., & Hühne, C. (2025). explainable artificial intelligence for automatic defect detection in additively manufactured parts using ct scan analysis. *Journal of Intelligent Manufacturing, 36*, 957–974. https://doi.org/10.1007/s10845-023-02272-4

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Camacho, L., & Bacao, F. (2024). Wsmoter: A novel approach for imbalanced regression. *Applied Intelligence, 54*, 8789–8799. https://doi.org/10.1007/s10489-024-05608-6

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Cohen-Shapira, N., & Rokach, L. (2024). *Pnt: Born-again tree-based model via fused decision path encoding. Information Fusion, 112, Article 102545*. . https://doi.org/10.1016/j.inffus.2024.102545

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. *Proceedings of the 5th Future Business Technology Conference* (pp. 5–12). Porto, Portugal

Doroshenko, A. (2020). Applying artificial neural networks in construction. *E3S Web of Conferences* (Vol. 143, p. 1029). https://doi.org/10.1051/e3sconf/202014301029

Dugalam, R., & Prakash, G. (2024). Development of a random forest based algorithm for road health monitoring. *Expert Systems with Applications,* 123940. https://doi.org/10.1016/j.eswa.2024.123940

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., & Smola, A. (2020). AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. https://doi.org/10.48550/arXiv.2003.06505

Ferrari, P. A., Annoni, P., Barbiero, A., & Manzi, G. (2011). An imputation method for categorical variables with application to non-linear principal component analysis. *Computational Statistics & Data Analysis, 55*(7), 2410–2420. https://doi.org/10.1016/j.csda.2011.02.007

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. *Advances in Neural Information Processing Systems* (Vol. 28, pp. 2962–2970). Curran Associates Inc. https://doi.org/10.5555/2969442.2969547

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Giannetti, C., & Essien, A. (2022). Towards scalable and reusable predictive models for cyber twins in manufacturing systems. *Journal of Intelligent Manufacturing, 33*(2), 441–455. https://doi.org/10.1007/s10845-021-01804-0

Giannetti, C., & Ransing, R. S. (2016). Risk based uncertainty quantification to improve robustness of manufacturing operations. *Computers & Industrial Engineering, 101*, 70–80. https://doi.org/10.1016/j.cie.2016.08.002

Giannetti, C., Ransing, R. S., Ransing, M. R., Bould, D. C., Gethin, D. T., & Sienz, J. (2014). A novel variable selection approach based on co-linearity index to discover optimal process settings by analysing mixed data. *Computers & Industrial Engineering, 72*, 217–229. https://doi.org/10.1016/j.cie.2014.03.017

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics, 24*(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data?. https://doi.org/10.48550/arXiv.2207.08815

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. https://doi.org/10.1007/978-0-387-84858-7 2nd edn.

Ho, T. K. (1995). Random decision forests. *Proceedings of the Third International Conference on Document Analysis and Recognition* (pp. 278–282). IEEE. https://doi.org/10.1109/ICDAR.1995.598994

Hoarau, A., Martin, A., Dubois, J., Gall, L., & Y. (2023). *Evidential random forests. Expert Systems with Applications, 230, Article 120652*. . https://doi.org/10.1016/j.eswa.2023.120652

Hotelling, H. (1947). Multivariate quality control, illustrated by the air testing of sample bomb-sights. In C. Eisenhart, M. W. Hastay, & W. A. Wallis (Eds.), *Techniques of Statistical Analysis* (pp. 111–184). New York: McGraw-Hill.

Hu, Y.-H., Wu, R.-Y., Lin, Y.-C., & Lin, T.-Y. (2024). A novel missforest-based missing values imputation approach with recursive feature elimination in medical applications. *BMC Medical Research Methodology, 24*(1), 269. https://doi.org/10.1186/s12874-024-02392-2

Jain, N., J., & P. K. (2023). *Lrf: A logically randomized forest algorithm for classification and regression problems*. . https://doi.org/10.1016/j.eswa.2022.119225 Expert Systems with Applications, 213, Article 119225.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* (Vol. 30, pp. 3149–3157). Curran Associates Inc. https://doi.org/10.5555/3294996.3295074

Kopp, M., Pevný, T., & Holena, M. (2020). *Anomaly explanation with random forests. Expert Systems with Applications, 149, Article 113187*. . https://doi.org/10.1016/j.eswa.2020.113187

Lê, S., Josse, J., & Husson, F. (2008). Factominer: An r package for multivariate analysis. *Journal of Statistical Software, 25*(1), 1–18. https://doi.org/10.18637/jss.v025.i01

Liao, M., Wen, H., Yang, L., Wang, G., Xiang, X., & Liang, X. (2024). Improving the model robustness of flood hazard mapping based on hyperparameter optimization of random forest. *Expert Systems with Applications,* 122682. https://doi.org/10.1016/j.eswa.2023.122682

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley. 2nd edn.

Liu, Y., & Jin, S. (2013). Application of bayesian networks for diagnostics in the assembly process by considering small measurement data sets. *The International Journal of Advanced Manufacturing Technology, 65*(9–12), 1229–1237. https://doi.org/10.1007/s00170-012-4252-7

Liu, J., Ramentol, E., Landin, C., & Tahvili, S. (2025). Dta-qc: an ai-driven framework for adaptive quality control and intelligent test optimization in 5 g manufacturing. *Journal of Intelligent Manufacturing*. https://doi.org/10.1007/s10845-025-02745-8

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). Curran Associates Inc. https://doi.org/10.5555/3295222.3295230

Mason, R. L., Tracy, N. D., & Young, J. C. (1995). Decomposition of T^2 for multivariate control chart interpretation. *Journal of Quality Technology, 27*(2), 99–108. https://doi.org/10.1080/00224065.1995.11979575

Mohanty, I., Bhattacharjee, D., & Datta, S. (2011). Designing cold rolled IF steel sheets with optimized tensile properties using ANN and GA. *Computational Materials Science, 50*(8), 2331–2337. https://doi.org/10.1016/j.commatsci.2011.03.007

Mollalo, A., Rivera, K. M., & Vahedi, B. (2020). Artificial neural network modeling of novel coronavirus (covid-19) incidence rates across the continental united states. *International Journal of Environmental Research and Public Health, 17*(12), 4204. https://doi.org/10.3390/ijerph17124204

Montgomery, D. C. (2009). Introduction to Statistical Quality Control. John Wiley. 6th edn.

Mujahid, M., Kına, E., Rustam, F., Villar, M. G., Silva Alvarado, E., De La Torre Diez, I., & Ashraf, I. (2024). Data oversampling and imbalanced datasets: An investigation of performance for machine learning and feature engineering. *Journal of Big Data, 11*, 87. https://doi.org/10.1186/s40537-024-00943-4

Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K., & Ishii, S. (2003). A bayesian missing value estimation method for gene expression profile data. *Bioinformatics, 19*(16), 2088–2096. https://doi.org/10.1093/bioinformatics/btg287

Özdem, S., & Orak, İM. (2024). A novel method based on deep learning algorithms for material deformation rate detection. *Journal of Intelligent Manufacturing*. https://doi.org/10.1007/s10845-024-02409-z

Paoletti, M. E., Haut, J. M., Tao, X., Plaza, J., & Plaza, A. (2020). A new GPU implementation of support vector machines for fast hyperspectral image classification. *Remote Sensing, 12*(8), 1257. https://doi.org/10.3390/rs12081257

Pearl, J. (2009). Causality: Models, Reasoning, and Inference. Cambridge University Press. https://doi.org/10.1017/CBO9780511803161 2nd edn.

Probst, P., Wright, M. N., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(3), 1301. https://doi.org/10.1002/widm.1301

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems* (Vol. 31, pp. 6639–6649). https://doi.org/10.48550/arXiv.1706.09516

Puthanveettil Madathil, A., Luo, X., Liu, Q., Walker, C., Madarkar, R., & Qin, Y. (2025). A review of explainable artificial intelligence in smart manufacturing. *International Journal of Production Research*. https://doi.org/10.1080/00207543.2025.2513574

Quantumblack. (2019). CausalNex: Bayesian Networks for Causal Inference. GitHub repository. Software library github a McKinsey company.

Quantumblack. (2020). A first CausalNex tutorial. Online documentation causalnex a McKinsey company.

Ransing, R. S., Giannetti, C., Ransing, M. R., & James, M. W. (2013). A coupled penalty matrix approach and principal component based co-linearity index technique to discover product specific foundry process knowledge from in-process data in order to reduce defects. *Computers in Industry, 64*(5), 514–523. https://doi.org/10.1016/j.compind.2013.03.001

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

Sasse, L., Nicolaisen-Sobesky, E., Dukart, J., Eickhoff, S. B., Götz, M., Hamdan, S., Komeyer, V., Kulkarni, A., Lahnakoski, J. M., Love, B. C., Raimondo, F., & Patil, K. R. (2025). Overview of leakage scenarios in supervised machine learning. *Journal of Big Data, 12*, 135. https://doi.org/10.1186/s40537-025-01193-8

Serneels, S., Croux, C., Filzmoser, P., & Van Espen, P. J. (2005). Robust multivariate calibration by trimmed score regression. *Chemometrics and Intelligent Laboratory Systems, 79*(1–2), 55–64. https://doi.org/10.1016/j.chemolab.2005.04.005

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion, 81*, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

Somvanshi, S., Das, S., Javed, S. A., Antariksa, G., & Hossain, A. (2024). A survey on deep tabular learning. https://doi.org/10.48550/arXiv.2410.12034

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications, 134*, 93–101. https://doi.org/10.1016/j.eswa.2019.05.028

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, Prediction, and Search, 2nd edn*. MIT Press.

Stekhoven, D. J., & Bühlmann, P. (2012). missforest: Non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. *Expert Systems with Applications, 237*, 121549. https://doi.org/10.1016/j.eswa.2023.121549

Tercan, H., & Meisen, T. (2022). Machine learning and deep learning based predictive quality in manufacturing: A systematic review. *Journal of Intelligent Manufacturing, 33*(7), 1879–1905. https://doi.org/10.1007/s10845-022-01963-8

Tsanas, A., & Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings, 49*, 560–567. https://doi.org/10.1016/j.enbuild.2012.03.003

Tüfekci, P. (2014). Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power & Energy Systems, 60*, 126–140. https://doi.org/10.1016/j.ijepes.2014.02.027

Vashishtha, G., Chauhan, S., & Zimroz, R. (2025). Fault detection of the rotating machines through optimized orthogonal matching pursuit by golden jackal optimization. *Structural Health Monitoring*. https://doi.org/10.1177/14759217251342197.OnlineFirst

Verlinden, B., Duflou, J. R., Collin, P., & Cattrysse, D. (2008). Cost estimation for sheet metal parts using multiple regression and artificial

neural networks: A case study. *International Journal of Production Economics, 111*(2), 484–492. https://doi.org/10.1016/j.ijpe.2007.02.004

Yang, L., & Lee, J. (2012). Bayesian belief network-based approach for diagnostics and prognostics of semiconductor manufacturing systems. *Robotics and Computer-Integrated Manufacturing, 28*(1), 66–74. https://doi.org/10.1016/j.rcim.2011.06.007

Yeh, I.-C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research,28*(12), 1797–1808. https://doi.org/10.1016/S0008-8846(98)00165-3

Youn, B. D., Park, K. M., Hu, C., Yoon, J. T., Kim, H. S., Jang, B. C., & Bae, Y. C. (2015). Statistical health reasoning of water-cooled power generator stator bars against moisture absorption. *IEEE Transactions on Energy Conversion, 30*(4), 1376–1385. https://doi.org/10.1109/TEC.2015.2444873

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE,109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555