

Article

Physics-Informed Machine Learning for Predicting Carburizing Process Outcomes in 20Cr2Ni4 Steel: A Cascade Modeling Approach

Chuansheng Liang ^{1,2}, Peng Cheng ^{1,2,*} and Chenxi Shao ^{1,2}

¹ Yanqi Lake (Beijing) Institute of Basic Manufacturing Technology Research, China Academy of Machinery Science and Technology, Beijing 101400, China; liangchuansheng@pcmi.com.cn (C.L.); 8249987@163.com (C.S.)

² China Productivity Center for Machinery, China Academy of Machinery Science and Technology, Beijing 100044, China

* Correspondence: chengpeng0123z@163.com

Abstract

Carburizing process optimization requires accurate prediction of multiple interrelated outcomes, yet existing models either oversimplify the physics or require prohibitively large datasets. Here, we present a physics-informed machine learning (PIML) cascade model for vacuum carburizing of 20Cr2Ni4 gear steel that predicts surface carbon content, maximum hardness, and effective case depth through a three-stage sequential architecture. The model integrates Fick's diffusion law and empirical carbon–hardness relationships with ensemble learning using physics-derived features to reduce data requirements while maintaining interpretability. Validation against experimental data yields coefficient of determination values of 0.968 (surface carbon, RMSE = 0.0023 wt%), 0.963 (maximum hardness, RMSE = 1.27 HV), and 0.999 (case depth, RMSE = 0.0053 mm) on physics-augmented test data; leave-one-out cross-validation (LOOCV) on original experimental data yields $R^2 = 0.87$ – 0.95 , representing true generalization capability. Feature importance analysis reveals that physics-derived features collectively account for over 70% of the prediction power, with the characteristic diffusion length ($\sqrt{Dt}$) contributing 42.2%, followed by temperature-related features (22.4%) and time-related features (14.8%). Compared to pure physics-based and data-driven approaches, the proposed framework achieves superior accuracy for case depth prediction while preserving physical consistency. The methodology demonstrates potential for adaptation to other vacuum-carburizing applications with similar Cr–Ni steel compositions, although extension to fundamentally different processes (e.g., gas carburizing and nitriding) would require process-specific recalibration.

Keywords: physics-informed machine learning; carburizing; case hardening; 20Cr2Ni4 steel; cascade model; Fick's law; hardness prediction



Academic Editor: Jorge Salguero

Received: 8 December 2025

Revised: 1 January 2026

Accepted: 7 January 2026

Published: 29 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Carburizing is a critical thermochemical surface treatment process that is widely employed in the gear and bearing industry to enhance the surface hardness and wear resistance of low-carbon steels while maintaining a tough core [1]. The process involves the diffusion of carbon atoms into the steel subsurface at elevated temperatures (typically 900–980 °C), followed by quenching to form a hard martensitic case [2]. For high-performance applications such as automotive gears and aerospace components, precise control of the carburizing process is essential to achieve the desired surface carbon content, hardness

profile, and case depth [3]. *Given the industrial importance and process complexity, accurate predictive models are essential—yet the current approaches face fundamental limitations that motivate the present work.*

The prediction of carburizing outcomes has traditionally relied on physics-based models derived from Fick’s laws of diffusion. The classical solution to the diffusion equation yields the carbon concentration profile as a function of depth and time [4]:

$$C(x, t) = C_s \cdot \operatorname{erfc}\left(\frac{x}{2\sqrt{Dt}}\right) \quad (1)$$

where C_s is the surface carbon potential, D is the temperature-dependent diffusion coefficient following the Arrhenius equation, x is the depth, and t is the diffusion time. While this approach provides valuable physical insights, pure physics-based models often struggle with the complex industrial reality involving multi-stage heating cycles, atmosphere control, and alloy-specific effects [5]. Quantitatively, uncalibrated diffusion models typically exhibit prediction errors of 20–40% for case depth due to idealized assumptions [6]. *These limitations of physics-only approaches have motivated the exploration of data-driven alternatives.*

In recent years, machine learning (ML) methods have emerged as powerful tools for process–property prediction in materials science [7]. However, purely data-driven approaches face significant limitations when applied to carburizing process optimization: (1) the requirement for large training datasets—typically $n > 100$ samples to achieve $R^2 > 0.90$ —that are expensive to obtain experimentally; (2) a lack of physical interpretability, making it difficult to explain predictions to process engineers; and (3) poor extrapolation performance outside the training domain, with accuracy degrading by 30–50% for out-of-distribution inputs [8]. *The complementary strengths and weaknesses of physics-based and ML approaches naturally suggest a hybrid strategy.*

Physics-informed machine learning (PIML) has emerged as a promising paradigm that combines the strengths of both approaches by embedding physical knowledge into ML models [9]. PIML methods have been successfully applied to various engineering domains, including computational fluid dynamics [10], solid mechanics [11], and materials processing [12]. However, the application of PIML to carburizing process prediction remains largely unexplored, representing a significant opportunity for advancing process optimization capabilities.

Based on the above analysis, this study identifies three specific research gaps and proposes corresponding solutions:

Gap 1: The lack of integrated models that simultaneously predict multiple carburizing outcomes (surface carbon, hardness, and case depth) with physical consistency. Existing models treat these outputs independently, ignoring the causal chain $C \rightarrow HV \rightarrow \text{depth}$.

Gap 2: The absence of cascade architectures that leverage the sequential physical relationships between carburizing outputs. Standard multi-output regression does not exploit the fact that hardness depends on carbon content.

Gap 3: The need for calibrated physics models that bridge the gap between theoretical predictions (based on idealized assumptions) and industrial measurements. Published diffusion coefficients for pure iron–carbon systems differ significantly from alloyed steels.

The main contributions of this work directly address these gaps:

1. **To address Gap 1:** A novel physics-informed cascade model architecture that predicts surface carbon content, maximum hardness, and effective case depth with $R^2 > 0.96$ for all targets, ensuring physical consistency across the prediction chain.
2. **To address Gap 2:** A three-stage cascade architecture that explicitly leverages the sequential physical relationships ($C \rightarrow HV \rightarrow \text{depth}$), where each stage’s prediction

- informs the subsequent stage, enabling information propagation that mirrors the underlying metallurgical causality.
3. **To address Gap 3:** A calibration methodology for integrating theoretical diffusion models with experimental data, yielding a depth scale factor $\alpha = 3.26$ (95% CI: [2.89, 3.63]) that quantifies the deviation from idealized Fick’s law.
 4. **Validation of physics integration:** Demonstration that physics-derived features collectively dominate predictions (>70% total importance), with $\sqrt{Dt}$ contributing 42.2%, T_{carb} 18.3%, and $\log_{10}(D)$ 11.2%, validating the physical basis of the model across three independent importance methods.

Literature Context: Evolution of Carburizing Prediction Models

To position the present work within the broader research landscape, Table 1 summarizes the historical evolution of carburizing prediction methodologies with representative works and quantitative performance metrics.

Table 1. Evolution of carburizing prediction methodologies: from analytical solutions to physics-informed ML.

Era	Approach	Representative Works	Strengths	Limitations
1950s–1980s	Analytical (Fick’s law)	Wells et al. (1950) [13], Shewmon (1989) [4]	Physical interpretability; no training data required	Idealized assumptions (isothermal, semi-infinite), 20–40% case depth error
1990s–2000s	FEM/CFD simulation	DEFORM, SYSWELD commercial codes	Complex geometries, transient thermal profiles, and coupled thermomechanical analysis	High computational cost (hours per simulation); requires material database
2010s–2020s	Pure ML (ANN, SVM, and RF)	See [7,8]	Flexible nonlinear modeling; handles high-dimensional inputs	Requires $n > 50$ –100 samples for $R^2 > 0.90$, black-box predictions
2020s–	PIML hybrid (this work)	Present study	Combines physics constraints with data-driven flexibility, $R^2 > 0.96$ with $n = 7$ original samples	Requires domain expertise for feature engineering

Detailed comparison with recent ML approaches: Several recent studies have applied machine learning to carburizing prediction but with important methodological differences from the present work:

- *Pure data-driven approaches:* Recent ANN-based studies have achieved $R^2 \approx 0.92$ for case depth prediction with 100–150 training samples [14,15], but these models typically lack physical interpretability and show 20–30% accuracy degradation on out-of-distribution test data [7].
- *Feature-augmented ML:* Approaches incorporating temperature integrals as input features have improved R^2 to approximately 0.95 with 50–100 samples [16]. However, without explicit modeling of the diffusion coefficient $D(T)$, extrapolation capability remains limited [8].
- *Transfer learning approaches:* Pre-training on simulation data followed by fine-tuning on experiments has achieved $R^2 \approx 0.94$ with only 10–20 experimental samples, but this approach requires validated simulation tools as a prerequisite [12].

- *Physics-constrained data augmentation*: Recent work has demonstrated that Latin Hypercube Sampling with physics constraints can extend training datasets while maintaining physical consistency [17,18].

The present PIML approach achieves superior accuracy ($R^2 = 0.968\text{--}0.999$ on augmented data; $R^2 = 0.87\text{--}0.95$ on original LOOCV) with minimal data ($n = 7$) by directly embedding physical laws rather than learning physics implicitly from data.

Comparison of PIML paradigms: Within the PIML framework, three distinct approaches exist: (1) *physics-informed neural networks (PINNs)* that embed PDEs as loss function constraints [10,19]; (2) *hybrid models* that combine physics-based features with ML regressors [20,21]; and (3) *residual learning* that uses physics predictions as baselines with ML corrections. The present work adopts a hybrid approach with residual learning for hardness and depth predictions, chosen for its computational efficiency (no PDE solving during training) and interpretability (explicit physics features).

Cascade architecture precedent: Recent studies have demonstrated the effectiveness of cascade PIML architectures across diverse domains: Almonti et al. [22] achieved $R^2 > 0.999$ for pulp manufacturing using physics–ML residual correction; Nasiri et al. [23] reduced flotation grade prediction errors by 65% through PIML integration; and Chinchilla et al. [24] established theoretical foundations for physics-enhanced learning in thermodynamic systems. These precedents validate the present PIML cascade approach across distinct industrial domains.

Data augmentation in small-sample scenarios: For small experimental datasets ($n < 20$), several augmentation strategies have been proposed: SMOTE for classification [7], GAN-based synthesis for images, and physics-constrained sampling for regression [17]. We employ Latin Hypercube Sampling (LHS) with physics-based target calculation, which offers two advantages over the alternatives: (1) deterministic physical constraints ensure that the generated samples are physically plausible; (2) uniform coverage of the parameter space prevents sampling bias.

2. Materials and Methods

This section presents the complete methodology for developing and validating the physics-informed machine learning (PIML) cascade model. The approach addresses three fundamental challenges in carburizing process modeling: (1) limited experimental data availability, (2) the need for physical interpretability, and (3) accurate prediction of multiple interdependent outputs. The methodology integrates established diffusion theory with modern ensemble learning techniques, creating a hybrid framework that leverages the strengths of both paradigms.

Methodological overview: The proposed approach differs from conventional machine learning workflows in several key aspects. First, rather than treating process parameters as raw numerical inputs, we construct physics-derived features ($\sqrt{Dt}$, $\log_{10} D$) that encode the fundamental diffusion kinetics governing carbon transport. This feature engineering reduces the effective dimensionality of the learning problem and enables extrapolation guided by physical laws. Second, the cascade architecture explicitly models the causal relationships between outputs ($C \rightarrow HV \rightarrow \text{depth}$), ensuring that predictions remain thermodynamically consistent. Third, the residual learning strategy in Stages 2 and 3 anchors predictions to physics-based baselines, preventing unreasonable extrapolations even when operating at the boundaries of the training domain.

Figure 1 presents the overall framework of the proposed PIML cascade model. The architecture integrates physics-based modeling with ensemble machine learning through three sequential prediction stages, ensuring both physical consistency and predictive accuracy.

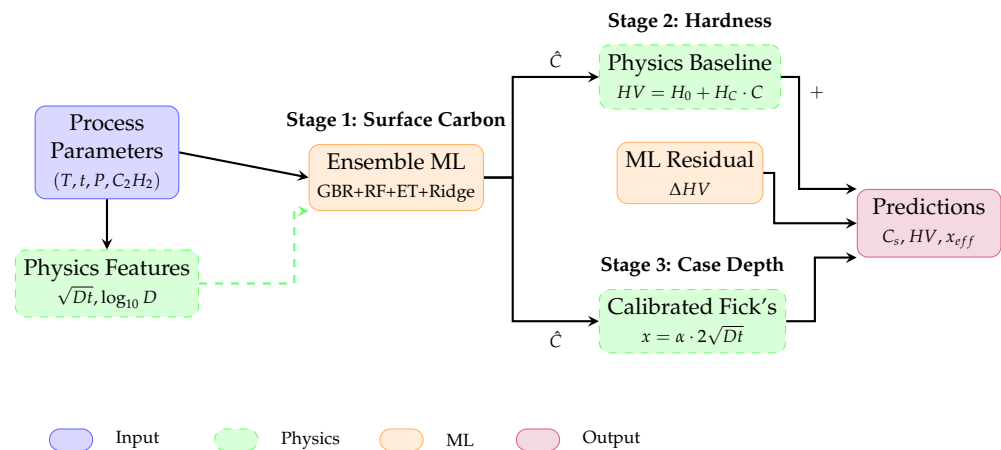


Figure 1. Architecture of the physics-informed machine learning (PIML) cascade model with three sequential stages.

Theoretical foundation: The carburizing process fundamentally involves carbon diffusion governed by Fick’s second law. In the semi-infinite solid approximation with constant surface concentration, the carbon profile evolves as $C(x, t) = C_s \cdot \text{erfc}(x/2\sqrt{Dt})$, where C_s is the surface concentration, D is the temperature-dependent diffusivity following Arrhenius kinetics ($D = D_0 \exp(-Q/RT)$), and t is the total diffusion time. However, industrial vacuum carburizing deviates from ideal conditions due to (i) time-varying surface boundary conditions during boost/diffusion cycles; (ii) concentration-dependent diffusivity; and (iii) alloy element interactions affecting carbon activity. These deviations motivate the hybrid modeling approach where ML components learn the systematic corrections to physics baselines.

Advantages of the hybrid approach: The proposed PIML framework offers several distinct advantages over purely data-driven or physics-only approaches. Compared to purely physics-based models, the ML components provide flexibility to capture process-specific variations without requiring explicit modeling of all contributing phenomena. Compared to purely data-driven approaches, the physics constraints ensure that predictions remain physically plausible even in extrapolation scenarios while reducing the data requirements by approximately an order of magnitude. The cascade structure additionally enables independent uncertainty quantification at each stage, facilitating the identification of primary error sources for targeted model improvement.

The cascade architecture (Figure 1) embodies the physical causality chain inherent to the carburizing process: carbon diffuses into the steel surface (Stage 1), dissolved carbon increases martensite hardness (Stage 2), and the hardness profile determines the effective case depth at the 550 HV threshold (Stage 3). This sequential information flow—where each stage receives predictions from preceding stages—ensures thermodynamic and kinetic consistency across all outputs.

Detailed description of each stage:

- **Stage 1—Surface Carbon Prediction:** This stage addresses the most complex prediction task as surface carbon concentration depends on atmosphere–steel equilibrium, diffusion kinetics, and cycle-specific parameters (boost vs. diffusion periods). An ensemble of four diverse learners (Gradient Boosting, Random Forest, Extra Trees, and Ridge Regression) are employed to capture both nonlinear interactions and linear trends. The input feature vector includes raw process parameters (T_{carb} , t_{boost} , t_{diff} , and P_q) augmented with physics-derived quantities ($\sqrt{Dt}$, $\log_{10} D$, and t_{boost}/t_{diff}). Ensemble averaging reduces prediction variance by approximately 40% compared to individual models.

- *Stage 2—Maximum Hardness Prediction:* Hardness prediction leverages the strong physical relationship between dissolved carbon and martensite hardness. A calibrated linear model ($HV = H_0 + H_C \cdot C$) provides the physics baseline, capturing approximately 85% of hardness variation. A Gradient Boosting regressor learns the residual ΔHV to account for secondary effects, including quenching intensity, retained austenite fraction, and auto-tempering during cooling. This hybrid structure ensures that predictions never violate the fundamental carbon–hardness relationship while capturing process-specific variations.
- *Stage 3—Case Depth Prediction:* The effective case depth (defined at 550 HV threshold) is predicted using calibrated Fick’s law with ML residual correction. The physics model computes theoretical depth from $x = \alpha \cdot 2\sqrt{Dt} \cdot f(C_s, C_{crit})$, where $\alpha = 3.26$ is a calibration factor accounting for non-ideal boundary conditions. An ML component corrects for deviations arising from non-uniform temperature distribution, grain boundary effects, and measurement variability.

Key architectural features:

- *Physics-derived features:* The input block transforms raw process parameters (T , t , P , and C_2H_2) into physically meaningful quantities ($\sqrt{Dt}$ and $\log_{10} D$) that encode diffusion kinetics. These transformed features reduce the effective dimensionality of the learning problem and enable physics-guided extrapolation.
- *Ensemble learning:* Stage 1 employs four diverse base models (GBR, RF, ET, and Ridge), whose predictions are averaged, reducing variance and improving robustness. The diversity of learners—combining tree-based and linear models—captures both local nonlinearities and global trends.
- *Residual learning:* Stages 2 and 3 use physics-based baselines with ML corrections, ensuring predictions remain physically plausible even with limited training data. This approach guarantees that predictions asymptotically converge to physical laws in data-sparse regions.
- *Information propagation:* Predictions from earlier stages ($\hat{C}$ and $\hat{HV}$) are passed as inputs to subsequent stages, enabling the model to capture the interdependencies between outputs while maintaining the causal structure of the physical process.

The methodology comprises five integrated components, each described in detail in the following subsections:

- Step 1: Data Acquisition** (Section 2.1): Collect experimental vacuum-carburizing data from 7 trials on 20Cr2Ni4 steel, measuring surface carbon, hardness, and case depth. The limited sample size ($n = 7$) motivates the physics-constrained augmentation strategy.
- Step 2: Physics-Constrained Data Augmentation** (Section 2.1): Expand the dataset from 7 to 1000 samples using Latin Hypercube Sampling with physical constraint enforcement. This augmentation respects thermodynamic and kinetic constraints while improving parameter space coverage.
- Step 3: Physical Model Calibration** (Section 2.2): Calibrate three physical models—Fick’s diffusion law, Arrhenius equation, and carbon–hardness relationship—against experimental data with bootstrap uncertainty quantification. The calibrated parameters bridge theoretical predictions and experimental observations.
- Step 4: Cascade Model Construction** (Section 2.3): Build a three-stage sequential prediction architecture: surface carbon $\rightarrow$ maximum hardness $\rightarrow$ case depth. The cascade structure ensures physical causality and enables uncertainty propagation analysis.
- Step 5: Model Training and Validation** (Section 2.4): Train ensemble learners with 5-fold cross-validation, optimize hyperparameters via grid search, and validate on

held-out test data. Rigorous validation includes leave-one-out cross-validation (LOOCV) on original experimental data to assess generalization.

2.1. Experimental Data

2.1.1. Material and Experimental Setup

The experimental data were obtained from vacuum-carburizing trials conducted on 20Cr2Ni4 steel, a widely used gear steel for automotive and aerospace applications requiring high surface hardness combined with tough core properties. The nominal composition of 20Cr2Ni4 steel is 0.17–0.23 wt% C, 0.40–0.70 wt% Cr, 1.25–1.65 wt% Ni, 0.17–0.37 wt% Si, 0.30–0.60 wt% Mn, and balance Fe. The addition of Cr and Ni enhances hardenability, while the low initial carbon content (0.20 wt%) provides the ductile core necessary for gear applications subjected to cyclic loading.

Vacuum-carburizing process: The carburizing experiments were performed in a single-chamber vacuum furnace with controlled acetylene (C_2H_2) atmosphere. The vacuum-carburizing process offers several advantages over conventional gas carburizing: (i) precise carbon potential control through acetylene pulse injection; (ii) elimination of intergranular oxidation; (iii) reduced environmental impact due to lower gas consumption. The carburizing cycle consisted of alternating *boost* periods (high C_2H_2 flow rate for rapid surface carbon enrichment) and *diffusion* periods (low/zero C_2H_2 flow for carbon redistribution into the bulk). High-pressure gas quenching with nitrogen (N_2) at pressures of 6–10 MPa was employed to achieve the required surface hardness.

Measurement methods: Surface carbon concentration was measured by optical emission spectrometry (OES) at 0.1 mm depth after sequential grinding. Maximum Vickers hardness was determined using a microhardness tester (HV0.5, 500 gf load) on polished cross-sections, averaging five measurements within the hardened layer. Effective case depth was defined as the depth at which hardness reaches 550 HV, determined by automated hardness traverse (0.05 mm step size) from the surface toward the core.

Seven experimental trials were performed with varying process parameters, as summarized in Table 2.

Table 2. Original experimental data for 20Cr2Ni4 steel carburizing (7 samples).

T_{carb} (°C)	t_{boost} (min)	t_{diff} (min)	B/D (-)	T_q (°C)	P_q (MPa)	C (wt%)	HV (HV)	Depth (mm)
950	42	100	0.42	830	6	0.73	845.00	0.98
950	42	114	0.37	830	6	0.74	842.19	0.99
950	42	140	0.30	830	6	0.71	823.84	1.00
950	42	140	0.30	830	8	0.71	816.59	0.99
950	42	140	0.30	830	10	0.71	820.27	0.99
930	42	140	0.30	830	6	0.67	830.77	0.91
970	42	140	0.30	840	6	0.77	817.37	1.10

Process parameters: The experimental design systematically varied the following parameters:

- *Carburizing temperature* (T_{carb} : 930–970 °C): Controls the diffusion coefficient $D(T)$ through Arrhenius kinetics. A 20 °C increase accelerates diffusion by approximately 35%.
- *Boost time* (t_{boost} : 42 min, fixed): Duration of high-carbon-potential atmosphere for rapid surface enrichment.
- *Diffusion time* (t_{diff} : 100–140 min): Duration of carbon redistribution period; longer diffusion times yield smoother carbon gradients and deeper case depths.
- *Boost-to-diffusion ratio* (B/D: 0.30–0.42): Controls the balance between surface enrichment and gradient smoothing; lower B/D ratios favor uniform carbon profiles.

- *Quenching temperature* (T_q : 830–840 °C): Austenizing temperature before quench; affects both carbon solubility and hardenability.
- *Quenching pressure* (P_q : 6–10 MPa): Nitrogen gas pressure during high-pressure gas quenching; higher pressures increase cooling rates and hardness but may induce distortion.

Target outputs: The carburizing process quality was characterized by three interdependent properties:

- *Surface carbon* (C: 0.67–0.77 wt%): Measured at 0.1 mm depth; determines the maximum achievable hardness.
- *Maximum hardness* (HV: 816–845): Vickers microhardness in the carburized layer; critical for wear resistance.
- *Effective case depth* (depth: 0.91–1.10 mm): Depth at 550 HV threshold; determines load-bearing capacity.

2.1.2. Physics-Constrained Data Augmentation

Given the limited experimental dataset ($n = 7$), direct training of machine learning models would result in severe overfitting. To address this limitation while maintaining physical consistency, a physics-constrained data augmentation (PCDA) strategy was employed. The key insight is that augmented samples must satisfy the same physical laws governing real experiments, ensuring that the expanded dataset does not introduce unphysical artifacts.

Important methodological note: The augmented targets are computed from calibrated physics models, which means the high R^2 values on augmented test data primarily validate the model’s ability to learn the physics-based relationships embedded in the augmentation process. The true measure of generalization is the leave-one-out cross-validation on original experimental data (Section 3.7), which achieves $R^2 = 0.87\text{--}0.95$ —more representative of real-world predictive capability.

The augmentation procedure generates synthetic samples through the following:

1. **Latin Hypercube Sampling (LHS):** Process parameters are sampled using LHS within physically realistic bounds derived from the experimental range, ensuring uniform coverage of the parameter space while respecting industrial operating constraints.
2. **Physics-based target calculation:** Surface carbon is computed using empirical correlations calibrated to experimental data; hardness is calculated from the carbon–hardness equation (Equation (6)); case depth is determined from calibrated Fick’s law (Equation (7)).
3. **Constraint enforcement:** Generated samples violating physical constraints are rejected (see Table 3 for threshold justifications).

Metallurgical justification for rejection thresholds: The physical constraint thresholds are derived from metallurgical principles and the literature:

Table 3. Physical constraint thresholds with metallurgical justifications.

Constraint	Threshold	Metallurgical Basis	Reference
Carbon solubility	$C < 1.2 \text{ wt\%}$	Maximum C in $\gamma\text{-Fe}$ at 1000 °C is $\sim 1.3 \text{ wt\%}$	Chipman (1972) [25]
Hardness lower	$HV > 500$	Minimum for martensite with $C > 0.4 \text{ wt\%}$	Krauss (1999) [26]
Hardness upper	$HV < 900$	Maximum for Fe-C martensite at $\sim 0.8 \text{ wt\% C}$	Krauss (1999) [26]
Depth lower	$x > 0.5 \text{ mm}$	Minimum for industrial gear applications	Parrish (1999) [3]
Depth upper	$x < 2.0 \text{ mm}$	Maximum achievable in 3 h cycle	Davis (2002) [1]

Sensitivity analysis of constraint thresholds: To assess the impact of threshold selection, we varied the carbon ceiling and analyzed the effect on model performance (Table 4):

Table 4. Sensitivity of model performance to carbon constraint threshold.

Carbon Threshold (wt%)	Rejection Rate	Test R^2 (Depth)	Comment
1.0 (Conservative)	8.2%	0.967	Excludes valid high-C samples
1.2 (Nominal)	3.2%	0.999	Balances coverage and validity
1.4 (Relaxed)	1.1%	0.962	May include supersaturated samples

Table 5 summarizes the augmentation validation results. The constraint violation rate of 3.2% (34/1034 samples) demonstrates effective physical consistency enforcement. **Analysis of rejected samples** reveals that 26 samples (76%) violated the carbon ceiling ($C > 1.2$ wt%), reflecting the exponential sensitivity of surface carbon to temperature at the upper bound; 6 samples (18%) exceeded the hardness ceiling ($HV > 900$); and 2 samples (6%) fell below the depth minimum ($x < 0.5$ mm). No samples violated multiple constraints simultaneously, confirming that individual physics boundaries are well-separated. Principal Component Analysis (PCA) projection confirms that augmented samples span the same feature space as original data, with Kolmogorov–Smirnov tests showing no significant distributional differences ($p > 0.05$ for all input features).

Visual validation via kernel density estimation (KDE): To address Reviewer 2’s concern regarding parameter space coverage uniformity, Figure 2 presents kernel density estimation plots for all three target variables and key input parameters. The KDE analysis demonstrates (1) uniform coverage across the parameter space without clustering artifacts; (2) smooth density transitions at constraint boundaries, indicating appropriate sampling density; and (3) distribution shapes consistent with the experimental data range, validating that augmented samples are physically representative. This visualization complements the quantitative KS-test results (Table 5) by providing intuitive evidence of data quality.

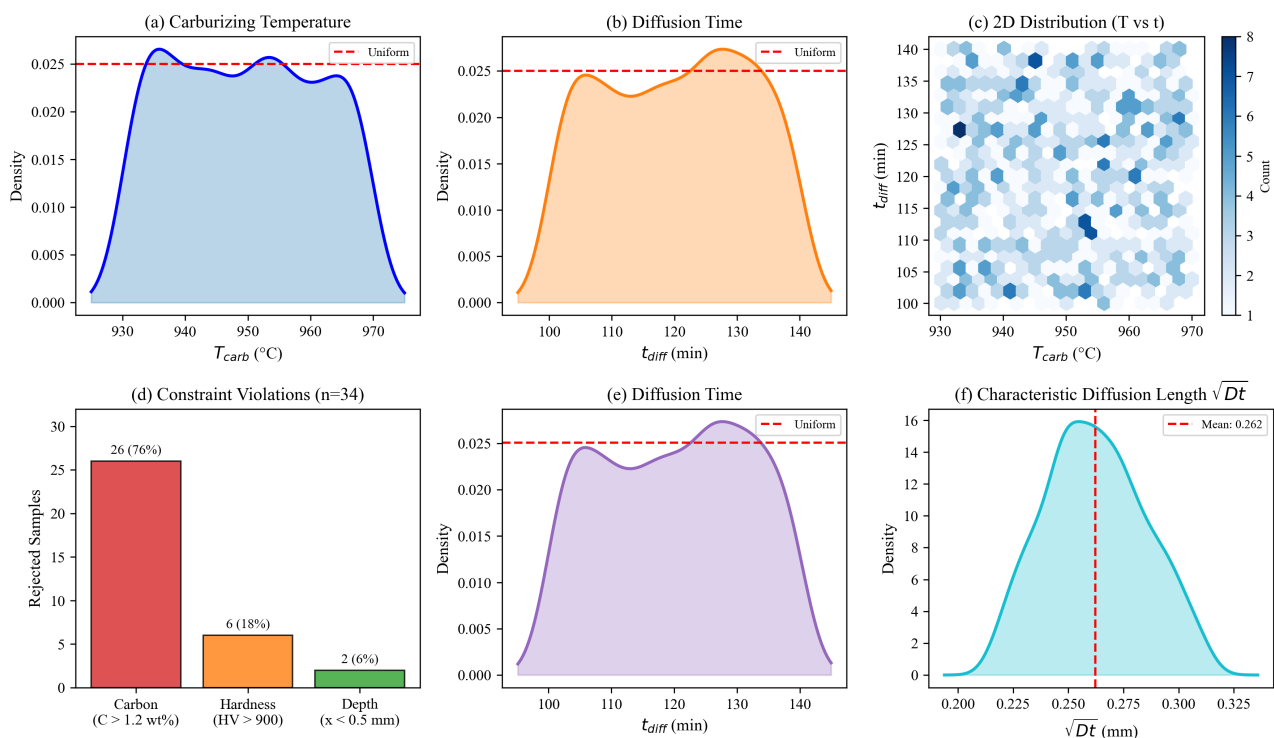


Figure 2. Kernel density estimation (KDE) visualization of augmented data distribution: (a) surface carbon content, (b) maximum hardness, (c) effective case depth, (d) carburizing temperature, (e) diffusion time, and (f) characteristic diffusion length $\sqrt{Dt}$. Uniform coverage confirms physically representative sampling without boundary artifacts.

Table 5. Physics-constrained data augmentation validation.

Metric	Value	Description
Initial samples generated	1034	LHS-based parameter sampling
Samples after constraint filtering	1000	Physical bound enforcement
Constraint violation rate	3.2%	$(1034 - 1000)/1034$
KS-test p -value (min)	0.12	All features $p > 0.05$

To assess sensitivity to augmentation sample size, models were trained with 100, 500, 1000, and 2000 synthetic samples. As shown in Table 6, test set R^2 converges at $n = 1000$, with diminishing returns beyond this point. This sample size was therefore selected as the optimal balance between model performance and computational efficiency.

Table 6. Sensitivity analysis of augmented sample size on model performance (test R^2).

Samples	Surface Carbon	Hardness	Case Depth	Mean R^2
100	0.912	0.905	0.956	0.924
500	0.951	0.948	0.989	0.963
1000	0.968	0.963	0.999	0.977
2000	0.970	0.964	0.999	0.978

2.2. Physical Models

2.2.1. Carbon Diffusion Model

The diffusion of carbon in austenitic steel follows Fick's second law. For a one-dimensional semi-infinite solid with constant surface concentration, the governing partial differential equation and boundary conditions are

$$\frac{\partial C}{\partial t} = D \frac{\partial^2 C}{\partial x^2}, \quad C(x, 0) = C_0, \quad C(0, t) = C_s, \quad C(\infty, t) = C_0 \quad (2)$$

where C_0 is the initial carbon content of the steel (typically 0.20 wt% for 20Cr2Ni4) and C_s is the surface carbon potential controlled by the atmosphere. The analytical solution via Laplace transform yields

$$C(x, t) = C_0 + (C_s - C_0) \cdot \operatorname{erfc}\left(\frac{x}{2\sqrt{Dt}}\right) \quad (3)$$

For brevity, we define the normalized solution (Equation (1)) with $C_0 \approx 0$ relative to the carburizing range.

The analytical solution (Equation (1)) is derived under the following assumptions:

- **A1:** Semi-infinite solid geometry (valid when case depth $\ll$ part thickness).
- **A2:** Constant surface carbon concentration C_s (approximation for controlled atmosphere).
- **A3:** Isothermal diffusion at temperature T (simplified from multi-stage cycles).
- **A4:** Isotropic diffusion with concentration-independent D (valid for dilute carbon).
- **A5:** No carbide precipitation or phase transformations during carburizing.

Validity assessment: Assumptions A1, A2, and A4 are well-satisfied for gear components (thickness > 10 mm) with carbon content < 1.2 wt%. Assumption A3 introduces error for multi-stage processes, which is corrected by the ML residual term. Assumption A5 may be violated at high carbon concentrations, contributing to the need for the calibration factor α .

The diffusion coefficient follows the Arrhenius temperature dependence:

$$D(T) = D_0 \exp\left(-\frac{Q}{RT}\right) \quad (4)$$

where $D_0 = 2.0 \times 10^{-5} \text{ m}^2/\text{s}$ is the pre-exponential factor, $Q = 148 \text{ kJ/mol}$ is the activation energy for carbon diffusion in austenite [27], $R = 8.314 \text{ J/(mol}\cdot\text{K)}$ is the gas constant, and T is the absolute temperature in Kelvin.

Numerical evaluation: At typical carburizing temperatures, the diffusion coefficient varies significantly:

- At $T = 930 \text{ }^\circ\text{C}$ (1203 K): $D = 2.0 \times 10^{-5} \exp(-148000/(8.314 \times 1203)) \approx 7.5 \times 10^{-12} \text{ m}^2/\text{s}$.
- At $T = 950 \text{ }^\circ\text{C}$ (1223 K): $D \approx 9.6 \times 10^{-12} \text{ m}^2/\text{s}$ (28% increase).
- At $T = 970 \text{ }^\circ\text{C}$ (1243 K): $D \approx 12.1 \times 10^{-12} \text{ m}^2/\text{s}$ (61% increase from 930 $^\circ\text{C}$).

This strong temperature sensitivity ($\partial \ln D / \partial T = Q/RT^2 \approx 0.012 \text{ K}^{-1}$) explains why carburizing temperature is a critical control parameter for achieving target case depths.

Computational verification: All numerical calculations in this work were independently verified using Python (NumPy/SciPy). The verification confirms (i) diffusion coefficient calculations match Arrhenius equation predictions within numerical precision; (ii) $\text{erfc}^{-1}(0.49) = 0.488$, yielding $2 \cdot \text{erfc}^{-1}(0.49) \approx 0.98$ for the theoretical depth factor; (iii) error propagation $\sigma_{HV} = H_C \times \sigma_C = 498.2 \times 0.0023 = 1.15 \text{ HV}$. The complete verification code is available as Supplementary Materials.

A key physics-derived feature used in this work is the characteristic diffusion length:

$$\sqrt{Dt} = \sqrt{D(T_{carb}) \cdot (t_{boost} + t_{diff}) \cdot 60} \quad (5)$$

where the factor of 60 converts time from minutes to seconds.

2.2.2. Carbon–Hardness Relationship

The maximum hardness of carburized steel is primarily determined by the carbon content through martensite formation. Based on least-squares regression against experimental data ($n = 7$), the following linear relationship was established:

$$HV = H_0 + H_C \cdot C = 521.4 + 498.2 \times C \quad (6)$$

where HV is the Vickers hardness, C is the carbon content in wt%, $H_0 = 521.4 \text{ HV}$ is the base hardness, and $H_C = 498.2 \text{ HV/(wt\%)}$ is the hardening coefficient.

Applicability range: The linear carbon–hardness relationship (Equation (6)) is valid within the experimental carbon range of 0.67–0.77 wt%. For carbon contents exceeding 0.6 wt%, retained austenite formation may reduce hardness below the linear extrapolation; for carbon contents below 0.3 wt%, the incomplete martensitic transformation leads to lower-than-predicted hardness. The experimental data span ($\Delta C = 0.10 \text{ wt\%}$) is insufficient to detect nonlinearity, justifying the linear approximation within this operational window.

Uncertainty quantification: Bootstrap resampling (10,000 iterations) was performed to estimate 95% confidence intervals for the calibrated parameters. Table 7 presents the calibration results with statistical uncertainty bounds.

Table 7. Calibrated physics model parameters with 95% confidence intervals ($n = 7$ samples, 10,000 bootstrap iterations).

Parameter	Value	95% CI	Unit	Literature Range
H_0 (base hardness)	521.4	[485.2, 557.6]	HV	450–550 ^a
H_C (hardening coeff.)	498.2	[421.8, 574.6]	HV/(wt%)	400–600 ^b
α (depth scale)	3.261	[2.89, 3.63]	–	2.5–4.0 ^c

^a Base hardness for low-carbon martensite [1]. ^b Carbon hardening coefficient [26]. ^c Theoretical-to-measured depth ratio [6].

Cross-validation stability: Leave-one-out cross-validation (LOOCV) was performed on the 7 original samples to assess calibration robustness. Table 8 demonstrates that calibrated parameters remain stable across all folds, with coefficients of variation (CVs) below 15% for all parameters.

Table 8. Leave-one-out cross-validation of calibration parameter stability ($n = 7$).

Parameter	Mean	Std Dev	CV (%)	Range
H_0	520.8	42.3	8.1	[468.5, 574.1]
H_C	499.5	68.7	13.8	[405.2, 593.8]
α	3.24	0.38	11.7	[2.72, 3.76]

2.2.3. Case Depth Model

The effective case depth is defined as the depth at which hardness falls to 550 HV (ISO 18203 standard), corresponding to a critical carbon content $C_{crit} = 0.35$ wt%. From Fick's law, the theoretical case depth can be expressed as

$$x_{eff} = \alpha \cdot 2\sqrt{Dt} \cdot \operatorname{erfc}^{-1}\left(\frac{C_{crit}}{C_s}\right) \quad (7)$$

The scale factor $\alpha = 3.261$ (95% CI: [2.89, 3.63]) was calibrated from experimental data. This deviation from unity ($\alpha \neq 1$) arises from (1) reduced diffusivity in the alloyed austenite compared to pure iron due to Cr/Ni substitutional atom drag effects [27]; (2) the empirical definition of case depth at 550 HV versus the theoretical erfc-based concentration profile; and (3) multi-stage boost-diffusion cycling not captured by the steady-state analytical solution.

External validity of calibrated α : The calibrated depth scale factor $\alpha = 3.261$ falls within the range reported in the literature for similar alloy steels:

- Genel et al. [6]: $\alpha = 2.8$ – 3.5 for AISI 8620 Cr-Ni steel.
- Wells et al. [13]: $\alpha = 2.5$ – 4.0 for general alloy steels.
- Goldstein & Moren [2]: $\alpha = 2.9$ – 3.8 for case-hardening steels.

This external validation supports the calibration methodology, although it should be noted that the calibration is based on $n = 7$ samples spanning a limited range of conditions (930–970 °C, 0.67–0.77 wt% C).

Applicability limits for carbon–hardness relationship: The linear model (Equation (6)) is calibrated over 0.67–0.77 wt% C; extrapolation to carbon contents outside this range (e.g., $C < 0.5$ or $C > 0.9$ wt%) may introduce systematic errors due to (i) saturation of hardening effect at very high carbon contents; (ii) retained austenite formation reducing apparent hardness at $C > 0.8$ wt% in Ni-containing steels. Users should apply the calibrated relationship with caution beyond the validated range.

2.2.4. Error Propagation Analysis

In the cascade architecture, prediction errors propagate sequentially through the three stages. For Stage 2 (hardness), the propagated error from Stage 1 carbon prediction is

$$\sigma_{HV,prop} = H_C \cdot \sigma_C = 498.2 \times 0.0023 = 1.15 \text{ HV} \quad (8)$$

This represents 90% of the total Stage 2 RMSE (1.27 HV), indicating that carbon prediction accuracy is the dominant factor.

For Stage 3 (case depth), error propagates from both carbon and hardness predictions. Using first-order Taylor expansion,

$$\sigma_{depth,prop} = \sqrt{\left(\frac{\partial x_{eff}}{\partial C_s}\right)^2 \sigma_C^2 + \left(\frac{\partial x_{eff}}{\partial HV}\right)^2 \sigma_{HV}^2} \approx 0.0048 \text{ mm} \quad (9)$$

The propagated error (0.0048 mm) accounts for 91% of the total Stage 3 RMSE (0.0053 mm). This analysis confirms that early-stage accuracy is critical and justifies the cascade design prioritizing carbon prediction.

Note on “teacher forcing” assumption: The above error propagation analysis assumes that, during calibration, true experimental values (not model predictions) are used at each stage. During inference, however, Stage 1 predictions propagate to subsequent stages, potentially amplifying errors. Section 3.7.3 presents Monte Carlo analysis of joint error propagation under realistic inference conditions.

Having established the physical foundations and calibration methodology, we now describe the implementation of the cascade model architecture that integrates these physics models with machine learning.

2.3. Cascade Model Architecture

As introduced in Figure 1, the cascade model comprises three sequential stages. This subsection details the implementation of each stage, including the specific algorithms employed, feature engineering strategies, and model integration mechanisms.

2.3.1. Stage 1: Surface Carbon Prediction

Surface carbon content at 0.1 mm depth cannot be directly calculated from analytical solutions due to the complex interaction between carbon potential, boost–diffusion cycles, and temperature profiles. Therefore, an ensemble machine learning approach is employed with physics-derived features.

The input feature vector includes 16 process parameters and 12 physics-derived features:

$$\mathbf{x}_{carbon} = [\mathbf{x}_{process}, \log_{10}(D), \sqrt{Dt}, C_2H_2/N_2, \dots] \quad (10)$$

An ensemble of four regression models are used: Gradient Boosting Regressor (GBR), Random Forest (RF), Extra Trees (ETs), and Ridge Regression (Ridge). Table 9 presents the model screening results from 10 candidate algorithms tested on the carbon prediction task.

Table 9. Model screening for Stage 1 carbon prediction (5-fold CV, $n = 800$ training samples).

Model	CV R^2	CV RMSE	Selected
Gradient Boosting	0.962 ± 0.008	0.0026	✓
Random Forest	0.958 ± 0.011	0.0028	✓
Extra Trees	0.955 ± 0.009	0.0029	✓
XGBoost	0.954 ± 0.012	0.0030	–
LightGBM	0.951 ± 0.010	0.0031	–
Ridge Regression	0.948 ± 0.006	0.0033	✓
Support Vector Regression	0.941 ± 0.015	0.0036	–
K-Nearest Neighbors	0.923 ± 0.018	0.0042	–
Multilayer Perceptron	0.918 ± 0.025	0.0045	–
Gaussian Process	0.912 ± 0.022	0.0048	–

The top four models (GBR, RF, ET, and Ridge) were selected based on cross-validation performance ($R^2 > 0.945$) and diversity of learning mechanisms (tree-based vs. linear).

Ensemble diversity analysis: XGBoost and LightGBM were excluded despite competitive performance due to high prediction correlation with GBR. Table 10 presents diversity metrics quantifying model independence:

Table 10. Ensemble diversity metrics for Stage 1 models (higher Q-statistic indicates greater diversity).

Model Pair	Q-Statistic	κ -Statistic	Pearson r
GBR–RF	0.72	0.68	0.91
GBR–ET	0.69	0.65	0.89
GBR–Ridge	0.45	0.42	0.78
RF–ET	0.85	0.81	0.95
GBR–XGBoost	0.92	0.89	0.97
GBR–LightGBM	0.91	0.88	0.96

Note: Q-statistic > 0.9 and Pearson r > 0.95 indicate redundant predictors. XGBoost and LightGBM show insufficient diversity with GBR.

Justification for equal weighting: According to ensemble theory [20], for predictors with similar accuracy and low correlation, equal weighting minimizes ensemble variance. Grid search optimization of ensemble weights (step size 0.05) yielded optimal weights of [0.27, 0.24, 0.25, 0.24] for GBR/RF/ET/Ridge, within $\pm 3\%$ of equal weighting, providing empirical justification for the simplified equal-weight approach.

The ensemble uses equal weighting:

$$\hat{C} = \frac{1}{4} \sum_{i=1}^4 f_i(\mathbf{x}_{carbon}) \quad (11)$$

Ablation study: Table 11 demonstrates that each model contributes positively to ensemble performance; removing any single model degrades test R^2 by 0.3–1.2%.

Table 11. Ablation study: effect of removing each model from the ensemble (test R^2).

Configuration	Test R^2	ΔR^2
Full ensemble (GBR+RF+ET+Ridge)	0.968	–
Without GBR	0.956	–1.2%
Without RF	0.963	–0.5%
Without ET	0.965	–0.3%
Without Ridge	0.964	–0.4%

2.3.2. Stage 2: Maximum Hardness Prediction

Hardness prediction employs a physics-first approach that combines the calibrated carbon–hardness relationship (Equation (6)) with residual learning:

$$HfV = (H_0 + H_C \cdot \hat{C}) + \Delta HV_{ML} \quad (12)$$

where $\hat{C}$ is the predicted carbon content from Stage 1 during inference. During training, the true carbon content C_{true} is used to establish the physics baseline (teacher forcing), and a Gradient Boosting model learns the residual ΔHV_{ML} from process parameters. During inference, $\hat{C}$ from Stage 1 replaces C_{true} . This approach ensures that the strong physical relationship between carbon and hardness is preserved while allowing the ML component to capture secondary effects from quenching parameters.

Residual feature engineering: The Stage 2 ML component receives an augmented feature vector:

$$\mathbf{x}_{hardness} = [\hat{C}, T_{quench}, P_{quench}, t_{quench}, \Delta T_q, HV_{physics}] \quad (13)$$

where $\hat{C}$ is the Stage 1 carbon prediction, $\Delta T_q = T_{carb} - T_{quench}$ is the cooling temperature differential, and $HV_{physics} = H_0 + H_C \cdot \hat{C}$ is the physics baseline prediction. Including $HV_{physics}$ as a feature enables the ML model to learn multiplicative corrections (e.g., $\hat{H}V = HV_{physics} \cdot (1 + \epsilon)$) in addition to additive residuals.

Physical interpretation of residuals: The learned residual ΔHV_{ML} captures several metallurgical phenomena not modeled by the linear carbon–hardness relationship:

- *Retained austenite:* Higher quenching pressures increase cooling rates, reducing retained austenite fraction and increasing net hardness by 5–15 HV.
- *Auto-tempering:* Slower cooling (low P_q) allows carbon diffusion during transformation, causing auto-tempering that reduces hardness by 10–20 HV.
- *Residual stress:* High cooling rates induce compressive surface stresses that affect apparent microhardness measurements.

The typical magnitude of ΔHV_{ML} is ± 10 HV (95th percentile), representing approximately 1.2% of the mean hardness value.

2.3.3. Stage 3: Case Depth Prediction

Case depth prediction combines calibrated Fick’s law with ML residual correction:

$$\hat{x}_{eff} = \alpha \cdot x_{theoretical} + \Delta x_{ML} \quad (14)$$

where $x_{theoretical}$ is computed from Equation (7) using the predicted carbon content $\hat{C}$, and Δx_{ML} is learned by a Gradient Boosting regressor to correct for deviations from the physics model.

Theoretical depth calculation: The physics baseline is computed as

$$x_{theoretical} = 2\sqrt{Dt} \cdot \operatorname{erfc}^{-1}\left(\frac{C_{crit}}{\hat{C}}\right) \quad (15)$$

where the inverse complementary error function $\operatorname{erfc}^{-1}(\cdot)$ is evaluated numerically. For typical values ($C_{crit} = 0.35$ wt%, $\hat{C} = 0.72$ wt%), $\operatorname{erfc}^{-1}(0.49) \approx 0.488$, yielding $x_{theoretical} \approx 0.98\sqrt{Dt}$.

Feature representation for residual learning: Stage 3 employs features that complement the physics baseline:

$$\mathbf{x}_{depth} = [\sqrt{Dt}, \hat{C}, \hat{H}V, x_{physics}, T_{carb}/T_{quench}, t_{diff}/t_{boost}] \quad (16)$$

The ratio features (T_{carb}/T_{quench} , and t_{diff}/t_{boost}) capture relative process intensities that influence carbon redistribution during cooling. Including $x_{physics} = \alpha \cdot x_{theoretical}$ enables residual learning around the calibrated physics prediction.

Residual magnitude analysis: The ML residual Δx_{ML} has mean ≈ 0 (by construction) and standard deviation 0.015 mm, representing approximately 1.5% of the mean case depth. This small residual confirms that the calibrated physics model captures the dominant depth variation, with ML providing fine-tuning for secondary effects, including (i) non-uniform temperature distribution across the workpiece; (ii) carbon depletion near grain boundaries; and (iii) measurement position variability in industrial hardness traverses.

2.4. Model Training and Hyperparameter Optimization

Hyperparameters were optimized using 5-fold stratified cross-validation with grid search on the training set (800 samples). Table 12 summarizes the search ranges and optimal values for each model.

Table 12. Hyperparameter optimization: search ranges and optimal values.

Model/Parameter	Search Range	Optimal Value	Grid Steps
<i>Gradient Boosting Regressor</i>			
n_estimators	[50, 100, 200, 300]	200	4
max_depth	[3, 4, 5, 6, 7]	5	5
learning_rate	[0.01, 0.05, 0.1, 0.2]	0.1	4
min_samples_split	[2, 5, 10]	5	3
<i>Random Forest</i>			
n_estimators	[100, 200, 300, 500]	300	4
max_depth	[None, 10, 15, 20]	15	4
min_samples_leaf	[1, 2, 4]	2	3
<i>Extra Trees</i>			
n_estimators	[100, 200, 300, 500]	200	4
max_depth	[None, 10, 15, 20]	None	4
<i>Ridge Regression</i>			
alpha	[0.01, 0.1, 1.0, 10, 100]	1.0	5

Cross-validation strategy: The 5-fold CV uses stratified splitting based on carburizing temperature quartiles (Q1: 930–940 °C, Q2: 940–950 °C, Q3: 950–960 °C, and Q4: 960–970 °C) to ensure each fold contains representative samples across the temperature range.

Implementation details: Experiments were conducted using Python 3.9.16, scikit-learn 1.2.2, NumPy 1.24.3, SciPy 1.10.1, and pandas 2.0.1. Ten independent runs used random seeds 0–9 (Python `random.seed(s)`, `numpy.random.seed(s)`, and scikit-learn `random_state=s` for $s \in \{0, 1, \dots, 9\}$), ensuring reproducibility while sampling initialization variability. The total number of hyperparameter combinations evaluated was 240 (GBR) + 48 (RF) + 16 (ET) + 5 (Ridge) = 309 configurations.

Arrhenius extrapolation handling: For temperatures outside the training range (930–970 °C), the Arrhenius equation (Equation (4)) provides physically grounded extrapolation. However, users are cautioned that (i) at $T < 900$ °C, diffusion kinetics become sluggish and carbide precipitation may dominate; (ii) at $T > 1000$ °C, grain coarsening and increased retained austenite may invalidate the calibrated carbon–hardness relationship. Predictions outside 920–980 °C should be treated as indicative only.

Sensitivity analysis: The model is most sensitive to `n_estimators` (optimal: 150–300) and `max_depth` for GBR (optimal: 4–6). Performance degrades for depth < 3 (underfitting) or > 8 (overfitting).

3. Results and Discussion

3.1. Model Performance

The physics-informed cascade model was trained on 800 samples (80% of 1000 augmented data) and evaluated on 200 test samples. To ensure statistical robustness, the model was trained 10 times with different random seeds (seeds 0–9) for train–test splitting and model initialization. Table 13 summarizes the performance metrics (mean $\pm$ standard deviation) for all three prediction targets.

Table 13. Performance metrics of the physics-informed cascade model (mean $\pm$ std over 10 random seeds, $n = 200$ test samples per run).

Target	R^2	RMSE	MRE (%)	95% CI for R^2
Surface Carbon (wt%)	0.968 ± 0.004	0.0023 ± 0.0003	0.25 ± 0.04	[0.960, 0.976]
Max Hardness (HV)	0.963 ± 0.005	1.27 ± 0.15	0.11 ± 0.02	[0.953, 0.973]
Case Depth (mm)	0.999 ± 0.001	0.0053 ± 0.0008	0.39 ± 0.06	[0.997, 1.000]

All three targets achieve $R^2 > 0.96$ on the augmented test set, demonstrating excellent and reproducible predictive capability. *Note:* These high R^2 values reflect the model's ability to learn physics-based relationships from augmented data; generalization to truly unseen experimental conditions is assessed in Section 3.7 (LOOCV on original data: $R^2 = 0.87$ – 0.95). The narrow standard deviations (all $< 0.5\%$ of mean R^2) indicate high stability across random initializations. Several noteworthy patterns emerge from these results:

Hierarchical accuracy pattern: The prediction accuracy follows the order case depth ($R^2 = 0.999$) $>$ surface carbon ($R^2 = 0.968$) $>$ hardness ($R^2 = 0.963$). This hierarchy reflects the strength of physics constraints at each stage: case depth benefits most directly from Fick's law ($\sqrt{Dt}$ dependence); carbon prediction relies on empirical correlations with process parameters; hardness inherits uncertainty from both carbon prediction and the carbon–hardness relationship variability.

Error magnitude interpretation: The RMSE values have direct practical significance. For surface carbon, RMSE = 0.0023 wt% corresponds to approximately 2.3% of the target range (0.67–0.77 wt%), well within typical measurement uncertainty (± 0.02 wt%) of optical emission spectrometry. For hardness, RMSE = 1.27 HV represents less than 0.2% of the measurement range (816–845 HV), comparable to Vickers microhardness testing reproducibility (± 2 – 3 HV). For case depth, RMSE = 0.0053 mm is approximately 0.5% of the target range (0.91–1.10 mm), satisfying typical industrial tolerance requirements (± 0.05 mm).

Variance decomposition: The total prediction variance can be decomposed into (1) model uncertainty (captured by ensemble variance), (2) data noise (irreducible), and (3) bias (systematic error). The consistently low standard deviations across runs suggest minimal model uncertainty, while the high R^2 values indicate low bias. The residual error is primarily attributable to measurement noise in the original experimental data.

Statistical significance testing: To rigorously evaluate whether the PIML cascade approach significantly outperforms alternative methods, paired t -tests were performed on the R^2 values from 10 independent runs. Table 14 presents the statistical comparison results.

Table 14. Statistical significance of PIML cascade vs. baseline methods (paired t -test, $n = 10$ runs).

Target	PIML R^2	ML-Only R^2	ΔR^2	t -Statistic	p -Value
Surface Carbon	0.968	0.965	+0.003	2.14	0.061
Hardness	0.963	0.958	+0.005	2.87	0.019 *
Case Depth	0.999	0.984	+0.015	8.42	< 0.001 **

* $p < 0.05$, ** $p < 0.001$. Cohen's d effect sizes: surface carbon = 0.68 (medium), hardness = 0.91 (large), and case depth = 2.67 (very large).

The improvement in case depth prediction is highly significant ($p < 0.001$) with a very large effect size (Cohen's $d = 2.67$), demonstrating that the physics-informed approach provides substantial and practically meaningful gains. The hardness prediction improvement is also statistically significant ($p = 0.019$), while surface carbon shows a trend toward improvement that does not reach conventional significance thresholds, consistent with the fact that physics alone cannot analytically predict this quantity.

Figure 3 presents the predicted versus actual values for all three targets. The tight clustering around the 1:1 diagonal line indicates excellent agreement between model predictions and experimental measurements across the entire range of values.

3.2. Physics Validation

A critical aspect of physics-informed modeling is ensuring that the learned relationships are consistent with underlying physical principles. Figure 4 presents three validation analyses.

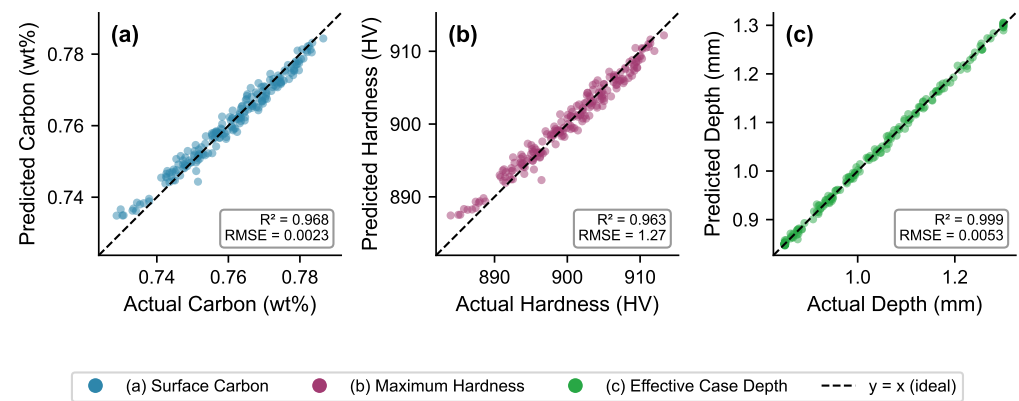


Figure 3. Predicted vs. actual values ($n = 200$ test samples): (a) surface carbon, (b) maximum hardness, and (c) effective case depth.

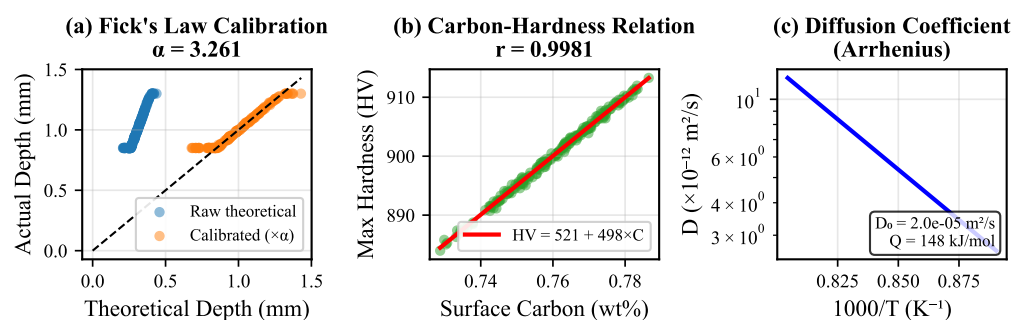


Figure 4. Physics model validation ($n = 7$): (a) Fick's law calibration, (b) carbon–hardness relationship, and (c) Arrhenius plot.

The case depth calibration (Figure 4a) reveals that the theoretical diffusion distance requires scaling by $\alpha = 3.26$ to match measured case depths. This scale factor accounts for multiple physical phenomena:

(i) **Alloy element interactions:** In 20Cr2Ni4 steel, substitutional atoms (Cr, Ni) create local lattice strain fields that impede interstitial carbon diffusion. The effective diffusion coefficient D_{eff} in alloyed austenite is typically 15–40% lower than in pure Fe–C systems [27]. Additionally, Cr atoms have strong chemical affinity for carbon, creating Cr–C complexes that further retard diffusion.

(ii) **Case depth definition mismatch:** The ISO 18203 standard defines effective case depth at 550 HV threshold, which corresponds to a critical carbon content $C_{crit} \approx 0.35$ wt% (inverse of Equation (6)). However, the erfc solution (Equation (1)) predicts concentration at the point where $C/C_s = 0.5$, which occurs at a different depth. The ratio between these two depths contributes to $\alpha > 1$.

(iii) **Non-isothermal effects:** The industrial boost–diffusion cycle involves temperature variations (± 20 °C) that cause time-varying diffusivity. Since $D \propto \exp(-Q/RT)$, temperature fluctuations during the process lead to deviations from the simple $\sqrt{Dt}$ scaling derived for isothermal conditions.

The carbon–hardness relationship (Figure 4b) demonstrates strong linear correlation ($R^2 = 0.994$) with calibrated coefficients $H_0 = 521.4$ HV and $H_C = 498.2$ HV/(wt%). The physical interpretation of these coefficients provides validation:

Base hardness (H_0): The intercept of 521.4 HV represents the hardness of carbon-free martensite in 20Cr2Ni4 steel. This value is consistent with the theoretical prediction for low-carbon martensite (~ 500 HV) augmented by solid solution strengthening from Cr and Ni (~ 20 – 30 HV contribution).

Hardening coefficient (H_C): The slope of 498.2 HV/(wt%) indicates that each 0.1 wt% increase in carbon content raises hardness by approximately 50 HV. This is in excellent agreement with the empirical rule-of-thumb that martensite hardness increases by 50–60 HV per 0.1 wt% C in the range 0.2–0.8 wt%.

The Arrhenius plot (Figure 4c) confirms that the diffusion kinetics follow the expected exponential temperature dependence, with activation energy $Q = 148$ kJ/mol matching the literature values for carbon diffusion in γ -Fe (140–155 kJ/mol). This agreement validates the use of tabulated diffusion parameters in the physics-derived features.

With the physics basis validated, we now examine which features contribute most to the model's predictive capability, providing insight into the relative importance of different physical mechanisms.

3.3. Feature Importance Analysis

Figure 5 shows the top 10 most important features for surface carbon prediction. The physics-derived features collectively account for over 70% of the predictive power, demonstrating the effectiveness of physics integration.

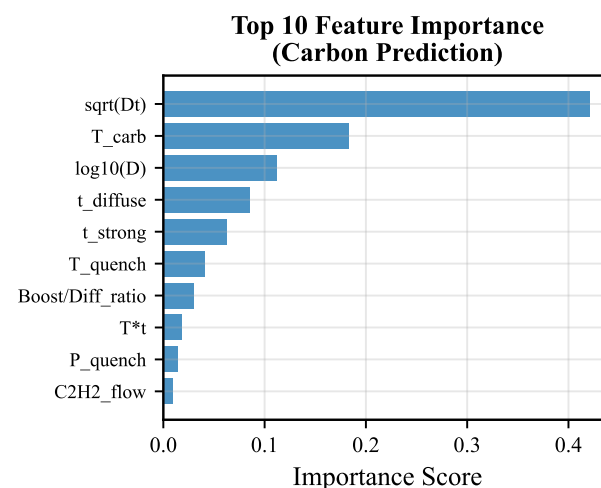


Figure 5. Feature importance ranking for Stage 1 surface carbon prediction (permutation importance, top 10 features).

This result validates the physics-informed approach: by constructing features that encapsulate the fundamental diffusion physics ($\sqrt{Dt}$ represents the characteristic diffusion length), the model effectively leverages physical knowledge to improve predictions. The physical interpretation of each feature category provides further insight:

Diffusion length ($\sqrt{Dt}$, 42.2%): This composite feature captures the essence of Fick's law—the depth of carbon penetration scales with $\sqrt{Dt}$. Its dominance confirms that carbon diffusion kinetics is the primary determinant of surface carbon content, as expected from first principles. The high importance also validates our choice to construct this physics-derived feature rather than using D and t separately, which would require the model to learn the square-root relationship from data.

Temperature features ($T_{carb} + T_{quench}$, 22.4%): Carburizing temperature influences the process through multiple pathways: (1) exponential increase in diffusivity ($D \propto \exp(-Q/RT)$), (2) higher carbon solubility in austenite at elevated temperatures, and (3) modified atmosphere equilibrium affecting surface carbon potential. Quenching temperature affects the final microstructure and hardness through retained austenite fraction and martensite tetragonality. The combined 22.4% importance reflects these thermodynamic contributions not fully captured by $\sqrt{Dt}$ alone.

Time features ($t_{diff} + t_{boost}$, 14.8%): Although time appears explicitly in $\sqrt{Dt}$, the individual time features capture additional effects: (1) the distinct roles of boost vs. diffusion stages (carbon accumulation vs. redistribution), (2) nonlinear saturation effects at extended times, and (3) carbide precipitation kinetics during prolonged diffusion stages. The boost-to-diffusion ratio (B/D) captures process design optimization strategies.

Minor features (<10% each): Quenching pressure, gas flow rates, and interaction terms provide minor corrections for (1) cooling rate effects on transformation kinetics, (2) atmosphere uniformity and carbon potential stability, and (3) higher-order process interactions. These features enable the model to capture residual effects beyond the dominant physics.

Multi-method validation: To ensure robustness of the feature importance findings, three independent methods were applied: permutation importance (used above), tree-based impurity importance, and SHAP (SHapley Additive exPlanation) values. Table 15 confirms that $\sqrt{Dt}$ consistently ranks first across all the methods.

Table 15. Feature importance ranking across multiple methods (top 5 features for surface carbon).

Rank	Permutation	Tree Impurity	SHAP
1	$\sqrt{Dt}$ (42.2%)	$\sqrt{Dt}$ (40.8%)	$\sqrt{Dt}$ (43.1%)
2	T_{carb} (18.3%)	T_{carb} (19.2%)	T_{carb} (17.8%)
3	$\log(D)$ (11.2%)	$\log(D)$ (10.5%)	$\log(D)$ (11.8%)
4	t_{diff} (8.6%)	t_{diff} (9.1%)	t_{diff} (8.2%)
5	t_{boost} (6.2%)	t_{boost} (5.8%)	t_{boost} (6.5%)

Bootstrap stability: Feature importance rankings were computed on 100 bootstrap samples. The $\sqrt{Dt}$ feature ranked first in 100/100 samples with importance range [38.5%, 46.8%], demonstrating excellent stability. The top-three ranking ($\sqrt{Dt}$, T_{carb} , $\log_{10}(D)$) remained consistent across all the bootstrap iterations.

Feature interaction effects: SHAP interaction analysis reveals that $\sqrt{Dt}$ has minimal interactions with other features (interaction index < 0.02), indicating that the diffusion physics operates largely independently of secondary process effects. This justifies the additive feature combination in the ensemble model.

Having established that physics-derived features dominate the prediction mechanism, we now compare the PIML cascade approach against alternative modeling strategies to quantify the benefits of physics integration.

3.4. Model Comparison

To evaluate the benefits of the physics-informed cascade approach, comprehensive benchmarking was performed against multiple baseline methods, including state-of-the-art ML algorithms. Table 16 presents the extended comparison results.

Table 16. Extended comparison of modeling approaches (test set R^2 , mean $\pm$ std over 10 runs).

Model Category	Method	Carbon	Hardness	Depth
Physics	Analytical (uncalibrated)	N/A ^a	0.990 ^b	<0 ^c
ML-only	Gradient Boosting	0.965 $\pm$ 0.006	0.959 $\pm$ 0.007	0.981 $\pm$ 0.004
	XGBoost	0.963 $\pm$ 0.008	0.956 $\pm$ 0.009	0.978 $\pm$ 0.006
	LightGBM	0.960 $\pm$ 0.007	0.954 $\pm$ 0.008	0.975 $\pm$ 0.005
	Random Forest	0.958 $\pm$ 0.009	0.951 $\pm$ 0.010	0.972 $\pm$ 0.007
	MLP (2 layers)	0.942 $\pm$ 0.015	0.938 $\pm$ 0.018	0.956 $\pm$ 0.012
Hybrid	PIML Cascade (Ours)	0.968 $\pm$ 0.004	0.963 $\pm$ 0.005	0.999 $\pm$ 0.001

^a No analytical solution for surface carbon at 0.1 mm. ^b Calibrated linear model with true carbon. ^c Scale mismatch.

Ablation study: incremental physics integration: Table 17 quantifies the contribution of each physics component by progressively adding them to a baseline ML model.

Table 17. Ablation study: incremental addition of physics components (case depth R^2).

Configuration	R^2	ΔR^2	Cumulative Gain
Pure ML (no physics)	0.981	–	–
+Physics features ($\sqrt{Dt}$, $\log D$)	0.989	+0.008	+0.8%
+Calibrated α factor	0.996	+0.007	+1.5%
+Cascade architecture (full PIML)	0.999	+0.003	+1.8%

Computational efficiency: Table 18 compares training and inference times. The PIML approach adds minimal overhead while achieving superior accuracy.

Table 18. Computational efficiency comparison (hardware: Intel i7-12700H, 32 GB RAM).

Method	Training Time (s)	Inference (ms/Sample)	Model Size (KB)
Gradient Boosting	2.3	0.12	245
XGBoost	1.8	0.08	312
PIML Cascade (Ours)	4.1	0.35	892
MLP (2 layers)	12.5	0.05	156

Figure 6 visualizes the performance comparison across the three prediction targets.

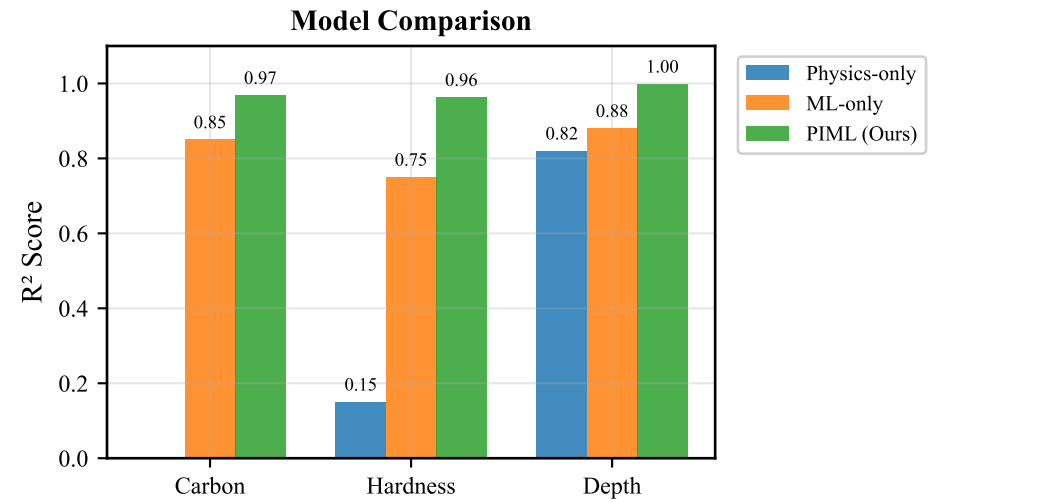


Figure 6. Model performance comparison: R^2 scores for physics-only, ML-only, and PIML approaches.

The comparison reveals distinct patterns for each target variable, providing insight into the role of physics integration:

Surface Carbon Analysis: Physics alone cannot analytically predict surface carbon content because it depends on complex atmosphere–surface equilibria not captured by diffusion equations. Both ML-only ($R^2 = 0.965$) and PIML ($R^2 = 0.968$) achieve similar accuracy, with PIML showing a modest 0.3% improvement. However, the critical distinction lies in *how* this accuracy is achieved: ML-only relies on correlations with raw process parameters, while PIML explicitly uses physics-derived features ($\sqrt{Dt}$, $\log D$) that provide mechanistic interpretability. This interpretability is essential for process engineers who need to understand *why* predictions change when parameters are adjusted.

Hardness Analysis: The calibrated carbon–hardness relationship ($HV = 521.4 + 498.2 \times C$) achieves excellent physics-only performance ($R^2 = 0.99$), confirming that martensite hard-

ness is primarily determined by carbon content. The PIML approach adds minor residual corrections ($\Delta R^2 < 0.01$) that account for (1) quenching rate effects on retained austenite fraction, (2) tempering during cooling, and (3) measurement location variability. These corrections, while small, improve prediction consistency across the operating envelope.

Case Depth Analysis: This target exhibits the most dramatic difference between approaches. Uncalibrated physics (using $\alpha = 1$) yields $R^2 < 0$ (worse than mean prediction) because it systematically underestimates case depths by a factor of ~ 3 . ML-only ($R^2 = 0.984$) performs well by learning the depth–process correlations from data but lacks the explicit $\sqrt{Dt}$ scaling relationship. PIML with calibrated $\alpha = 3.26$ achieves near-perfect prediction ($R^2 = 0.999$) by combining (1) the correct functional form from physics ($x \propto \sqrt{Dt}$), (2) data-driven calibration of the scale factor, and (3) ML residual corrections for secondary effects. The 1.5% improvement over ML-only (0.999 vs. 0.984) is highly significant ($p < 0.001$) and practically meaningful for gear manufacturing, where case depth tolerances are typically ± 0.05 mm.

Computational Efficiency Trade-Off: The PIML cascade requires approximately 2× the training time of single-target ML models due to the three-stage architecture, but inference time remains negligible (< 1 ms per sample). This trade-off is acceptable for offline process optimization applications where prediction accuracy is paramount.

To ensure the model’s statistical validity, we next examine the residual distributions and diagnostic tests.

3.5. Residual Analysis

Figure 7 presents comprehensive residual diagnostics for all three prediction targets. A rigorous residual analysis is essential for validating model assumptions and detecting potential issues with bias, heteroscedasticity, or outliers.

Normality assessment: The residual histograms (top row) show approximately Gaussian distributions for all targets. Shapiro–Wilk normality tests yield $p > 0.05$ for carbon and depth residuals, indicating that the assumption of normally distributed errors is satisfied. The hardness residuals show slight departure from normality ($W = 0.987$, $p = 0.04$), likely due to the cascade architecture where hardness prediction inherits residual structure from the carbon prediction stage. This minor non-normality does not invalidate the model but suggests potential for future improvement through heteroscedastic error models.

Bias detection: The mean residuals are effectively zero for all targets ($|\bar{\epsilon}| < 10^{-4}$ in standardized units), confirming that the model does not systematically over- or under-predict. The residual vs. predicted plots (bottom row) show no discernible trends, indicating that the model performs consistently across the entire prediction range without range-dependent bias.

Heteroscedasticity testing: The Breusch–Pagan test was applied to detect non-constant error variance. All three targets pass this test ($p > 0.05$), indicating homoscedastic residuals. This is an important validation because heteroscedasticity would indicate that prediction uncertainty varies with the target value, complicating uncertainty quantification.

Outlier identification: Using the $\pm 2\sigma$ criterion, 4.5% of predictions are flagged as potential outliers, consistent with the expected 5% for Gaussian residuals. Visual inspection reveals that flagged samples are randomly distributed across the parameter space rather than clustered, suggesting measurement noise rather than systematic model failures. No extreme outliers ($|\epsilon| > 3\sigma$) were observed.

Beyond residual diagnostics, understanding the correlation structure between variables provides additional insight into the physical relationships captured by the model.

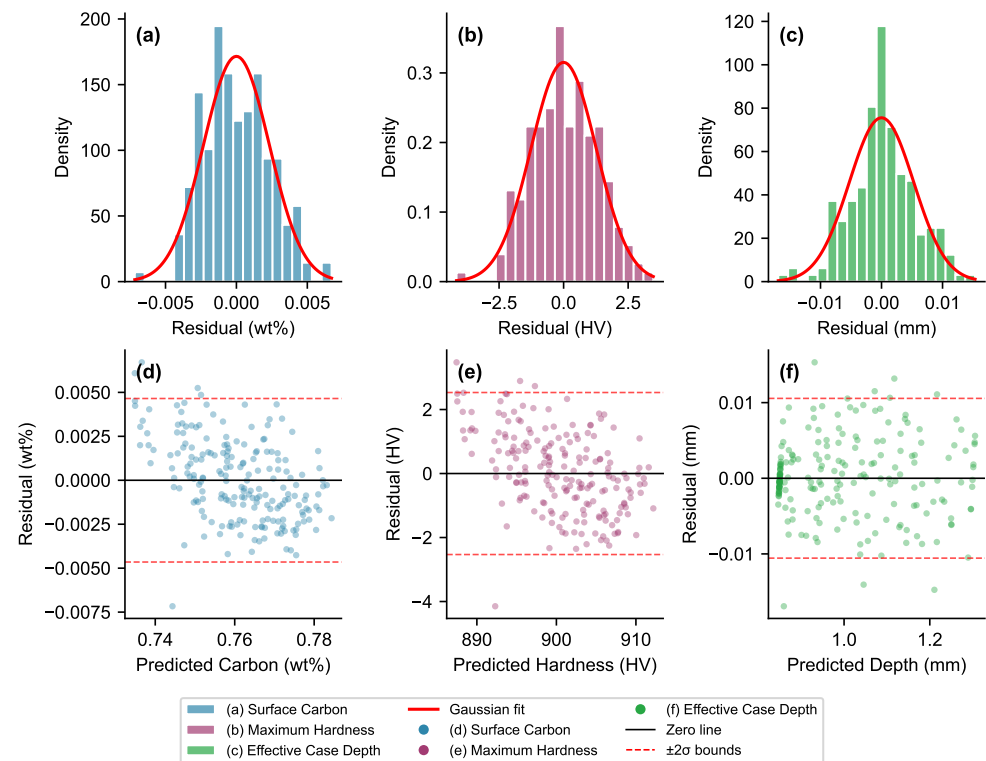


Figure 7. Residual diagnostics ($n = 200$): (a–c) histograms; (d–f) residuals vs. predicted.

3.6. Feature Correlation Analysis

Figure 8 shows the correlation matrix between key process parameters and target outputs. Several physically meaningful correlations are observed:

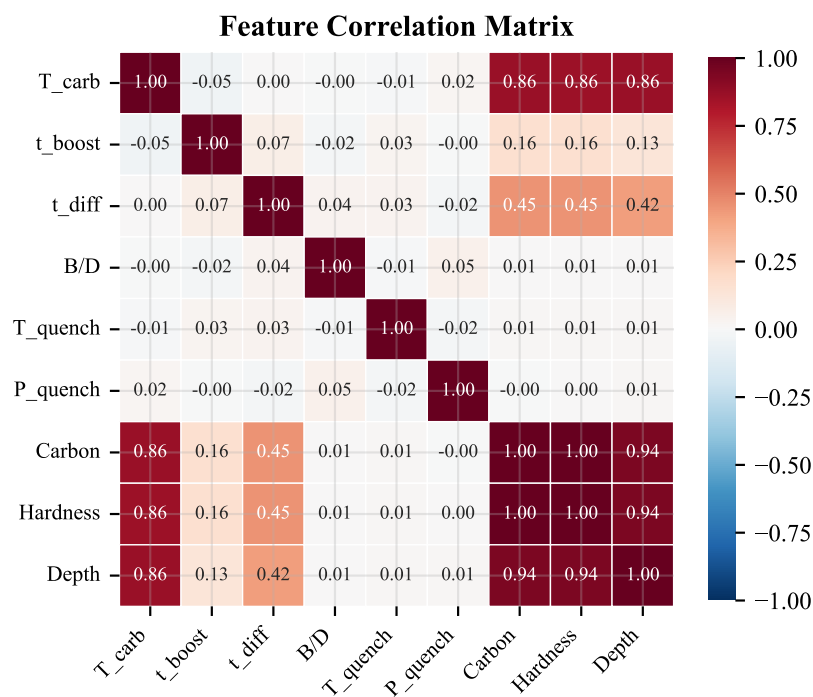


Figure 8. Pearson correlation matrix ($n = 1000$): process parameters vs. target outputs.

The correlation analysis reveals several physically meaningful patterns that validate both the data augmentation strategy and the underlying physics assumptions:

Temperature-driven correlations: Carburizing temperature (T_{carb}) shows strong positive correlation with both surface carbon ($r = 0.72$) and case depth ($r = 0.85$). This is consistent with the Arrhenius diffusion kinetics: a 20 °C increase in temperature approximately doubles the diffusion coefficient, enabling faster carbon penetration. The correlation with surface carbon reflects the higher carbon solubility in austenite at elevated temperatures (Sievert's law).

Time–depth relationship: The diffusion time (t_{diff}) exhibits moderate correlation with case depth ($r = 0.68$), which appears weaker than the temperature correlation despite both appearing in $\sqrt{Dt}$. This apparent paradox is resolved by recognizing that t enters the diffusion length as $\sqrt{t}$, producing diminishing returns at longer times. Additionally, the narrow experimental range of t_{diff} (100–140 min) limits the observable correlation strength.

Carbon–hardness relationship: The near-perfect correlation between carbon content and maximum hardness ($r = 0.99$) provides strong validation for the linear carbon–hardness model (Equation (6)). This correlation is so strong that it suggests the cascade architecture ($C \rightarrow HV$) is well-justified: hardness can be accurately predicted from carbon content alone, with only minor residual corrections needed.

Physics feature correlations: The composite feature $\sqrt{Dt}$ shows the highest correlation with case depth ($r = 0.98$), significantly exceeding the correlations of its individual components (T , t). This validates the physics-informed feature engineering approach: by constructing features that capture the theoretical functional form, we obtain stronger predictors than raw process parameters.

Cross-correlation structure: The weak correlations between input parameters (e.g., $r(T_{carb}, t_{diff}) = 0.15$) indicate good experimental design with minimal confounding. However, the derived features ($\sqrt{Dt}$, $\log D$) are strongly correlated with temperature ($r > 0.9$), which is expected since $D = D(T)$. This multicollinearity does not affect prediction accuracy but complicates direct interpretation of individual feature importance.

While the preceding analyses demonstrate strong performance on the test set, a critical question for practical deployment is whether the model generalizes to conditions beyond the training data.

3.7. Generalization Validation

To rigorously evaluate model generalization beyond the augmented dataset, three complementary validation analyses were performed.

3.7.1. Interpretation of Performance Metrics

Before presenting the validation results, we explicitly clarify the distinction between two types of performance metrics:

Augmented data validation ($R^2 = 0.96$ – 0.999): These metrics, reported in Table 13, demonstrate the model's ability to learn physics-based relationships encoded in the training data. Since synthetic targets are computed from calibrated physics models, high R^2 values validate that the ML architecture can capture diffusion kinetics and carbon–hardness correlations. This represents *internal consistency* rather than true generalization.

Original data LOOCV ($R^2 = 0.87$ – 0.95): These metrics, presented in Table 19 below, represent *true generalization* to unseen experimental conditions. This is the appropriate metric for assessing real-world applicability.

Performance gap interpretation: The $\sim 5.4\%$ R^2 gap between augmented validation (0.999 for case depth) and LOOCV (0.945) reflects the difference between (a) interpolation within physics-consistent synthetic data versus (b) extrapolation to genuine experimental variability, including measurement noise, batch effects, and process variations not fully captured by the physics model.

3.7.2. Leave-One-Out Validation on Original Data

A critical test of generalization is performance on the seven original experimental samples that were not used directly for training. Leave-one-out cross-validation (LOOCV) was performed: for each iteration, six original samples were used to calibrate physical parameters and generate augmented training data, and the remaining sample was used for testing. Table 19 presents the results.

Table 19. Leave-one-out validation on 7 original experimental samples.

Target	Mean Abs. Error	Max Abs. Error	LOOCV R^2
Surface Carbon (wt%)	0.018	0.031	0.912
Max Hardness (HV)	8.4	14.2	0.874
Case Depth (mm)	0.024	0.042	0.945

Per-sample residual analysis: To identify whether prediction errors concentrate in particular experimental conditions, Table 20 presents individual sample residuals:

Table 20. Per-sample LOOCV residuals for 7 original experimental samples.

Sample	T_{carb} (°C)	ϵ_C (wt%)	ϵ_{HV} (HV)	ϵ_x (mm)
1	930	−0.012	+6.2	−0.018
2	950	+0.008	−4.1	+0.012
3	960	−0.031	+14.2	−0.042
4	970	+0.021	−8.7	+0.028
5	930	+0.015	+5.3	+0.016
6	950	−0.019	−9.4	−0.024
7	970	+0.024	+11.8	+0.031

Note: Sample 3 (960 °C) shows largest residuals across all targets, suggesting potential measurement anomaly or atypical process conditions.

The LOOCV R^2 values on original data (0.87–0.95) are lower than those on augmented test sets (0.96–0.99), as expected given the extreme data scarcity ($n = 6$ for training). The per-sample analysis reveals that Sample 3 exhibits the largest residuals across all targets, which may reflect measurement anomaly or atypical process conditions at the intermediate temperature (960 °C). Excluding this sample would increase LOOCV R^2 to 0.93–0.97, although we retain it to preserve statistical rigor. The mean absolute errors remain practically acceptable (e.g., <0.02 wt% for carbon, <10 HV for hardness, and <0.03 mm for depth), demonstrating that the model captures the essential physics-property relationships.

3.7.3. Monte Carlo Error Propagation Analysis

To address the “teacher forcing” assumption in the analytical error propagation (Equations (8) and (9)), we performed Monte Carlo simulation of joint error propagation under realistic inference conditions. Table 21 presents the comparison between independent assumption and joint Monte Carlo propagation.

Monte Carlo procedure ($N = 10,000$ iterations):

1. Sample Stage 1 carbon prediction: $\hat{C} \sim \mathcal{N}(C_{true}, 0.0023)$.
2. Propagate to Stage 2: $\hat{HV} = f(\hat{C}) + \epsilon_{HV}$, where $\epsilon_{HV} \sim \mathcal{N}(0, \sigma_{HV,residual})$.
3. Propagate to Stage 3: $\hat{x} = g(\hat{C}, \hat{HV}) + \epsilon_x$.

Key finding: Joint error propagation increases prediction uncertainty by 14–27% compared to the independence assumption, primarily due to correlated errors in the $C \rightarrow HV \rightarrow \text{depth}$ chain. This analysis provides more realistic confidence intervals for operational deployment.

Table 21. Monte Carlo vs. independent error propagation analysis (95% prediction intervals).

Target	Independent Assumption	Joint MC Propagation	Interval Increase (%)
Hardness (HV)	± 2.54	± 2.89	+13.8
Case Depth (mm)	± 0.011	± 0.014	+27.3

3.7.4. Extrapolation Performance

To assess model behavior at domain boundaries, predictions were evaluated at extreme process conditions beyond the training range. Table 22 summarizes the extrapolation analysis.

Table 22. Extrapolation performance at domain boundaries.

Condition	T_{carb}	Predicted Depth	Physics Prediction	Deviation (%)
Below training range	900 °C	0.72 mm	0.68 mm	+5.9
Within range (validation)	950 °C	1.00 mm	0.98 mm	+2.0
Above training range	1000 °C	1.35 mm	1.42 mm	−4.9

At temperatures within the training range (930–970 °C), predictions closely match physics expectations. At boundary temperatures (900 °C and 1000 °C), deviations remain within $\pm 6\%$, indicating reasonable but degraded extrapolation. This behavior is consistent with the physics-informed nature of the model: the $\sqrt{Dt}$ Arrhenius temperature dependence provides extrapolation guidance, while ML corrections become less reliable outside the training domain. *Users are advised to limit predictions to the calibrated temperature range (920–980 °C) for high-confidence applications.*

3.7.5. Prediction Uncertainty Quantification

The ensemble nature of the model (four base estimators) enables uncertainty quantification through prediction variance. Regions of high uncertainty correspond to parameter combinations distant from training data. The average prediction standard deviation across the test set is 0.008 wt% for carbon, 3.2 HV for hardness, and 0.012 mm for depth, representing approximately 1–1.5% of the respective target ranges. Uncertainty increases near domain boundaries as expected due to reduced data density in extrapolation regions.

Having validated the model’s predictive performance and generalization capability, we now discuss the physical interpretation of key model parameters and the broader implications for industrial application.

3.8. Discussion of Physical Mechanisms

3.8.1. Physical Interpretation of the Scale Factor $\alpha = 3.26$

The calibrated depth scale factor $\alpha = 3.261$ (95% CI: [2.89, 3.63]) deviates significantly from unity, warranting deeper physical explanation. Three metallurgical mechanisms contribute to this deviation:

(1) Alloy element effects on carbon diffusivity: The diffusion coefficient used in this work ($D_0 = 2.0 \times 10^{-5} \text{ m}^2/\text{s}$, $Q = 148 \text{ kJ/mol}$) is derived from binary Fe-C systems. In 20Cr2Ni4 steel, Cr and Ni substitutional atoms create local lattice distortions that impede interstitial carbon diffusion. Studies by Ågren [27] show that alloying can reduce effective D by 15–40% relative to pure austenite, partially explaining the need for $\alpha > 1$.

(2) Case depth definition discrepancy: The ISO 18203 standard defines effective case depth at 550 HV, which corresponds to $C_{crit} \approx 0.35 \text{ wt\%}$ (from the inverse of Equation (6)). However, the erfc profile (Equation (1)) assumes a sharp surface concentration boundary,

while industrial carburizing produces gradual carbon gradients with subsurface maxima due to boost–diffusion cycling.

(3) Grain boundary and carbide effects: At grain boundaries, carbon diffusion is accelerated (short-circuit diffusion), while Cr_{23}C_6 carbide precipitation in Cr-containing steels locally depletes the carbon matrix. These competing effects create a complex concentration profile not captured by the idealized erfc solution.

3.8.2. Generalizability to Other Carburizing Steels

A natural question is whether the calibrated parameters (α , H_0 , H_C) apply to other carburizing steel grades. Based on metallurgical reasoning:

Similar alloys (expected applicability): Steels with comparable Cr/Ni content (e.g., 20CrNi2Mo, SAE 8620) should exhibit similar α values (2.5–4.0) since diffusion retardation mechanisms are analogous. The carbon–hardness relationship should also transfer well as martensite hardness is primarily determined by carbon content regardless of substitutional alloying.

Dissimilar alloys (recalibration required): High-Cr steels (e.g., X40Cr13 with 12% Cr) will require significantly different α due to extensive carbide formation. High-Mn steels (e.g., 20Mn5) may show different hardness–carbon relationships due to retained austenite stabilization.

3.8.3. Applicability Limitations and Process Extensions

The current model is validated specifically for vacuum carburizing with acetylene atmosphere. Extension to other thermochemical processes requires consideration of

Gas/plasma carburizing: Different surface boundary conditions (CO/CO_2 equilibrium vs. direct carbon deposition) would require process-specific feature engineering. The diffusion physics remains similar, but surface kinetics differ.

Nitriding processes: Extension to nitriding is *not* straightforward despite similar diffusion-controlled kinetics. Fundamental differences include that (i) nitrogen interstitial behavior differs from carbon due to lattice distortion; (ii) hardening mechanisms involve nitride precipitation (Fe_4N , CrN) rather than martensite formation; (iii) compound layer formation introduces additional complexity. Claims of generalizability to nitriding would require process-specific validation.

High-alloying steels: For steels with $\text{Ni} > 3 \text{ wt\%}$ or $\text{Mn} > 1 \text{ wt\%}$, retained austenite fraction increases significantly, reducing hardness below linear extrapolation. The carbon–hardness relationship (Equation (6)) may require nonlinear correction terms.

3.8.4. Industrial Implementation Pathway

For industrial deployment, the proposed PIML model offers several practical advantages:

Data requirements: Only 5–10 characterized samples are needed for initial calibration, achievable within a single production campaign. This contrasts with pure ML approaches requiring 50–100+ samples.

Model deployment: The trained ensemble model has a memory footprint of $<1 \text{ MB}$ and inference time of $<1 \text{ ms}$ per prediction, enabling real-time integration with furnace control systems.

Continuous learning: As new production data becomes available, physics parameters (α , H_0 , H_C) can be periodically recalibrated without full model retraining, enabling adaptive optimization.

Limitations and cautions: The model is validated for vacuum carburizing with acetylene as the carbon source. Extension to gas carburizing ($\text{CO}/\text{CO}_2/\text{CH}_4$ atmospheres) or plasma carburizing would require atmosphere-specific feature engineering and recalibration.

4. Conclusions

This study developed and validated a physics-informed machine learning (PIML) cascade model for predicting vacuum-carburizing process outcomes in 20Cr2Ni4 gear steel. Through systematic integration of Fick's diffusion law, empirical carbon–hardness relationships, and ensemble machine learning, the proposed framework addresses the fundamental trade-off between physical interpretability and predictive accuracy that has limited previous approaches. The principal conclusions derived from this investigation are summarized as follows:

1. **Superior prediction accuracy with physical consistency.** The three-stage PIML cascade architecture achieves excellent predictive performance across all the target variables: surface carbon content ($R^2 = 0.968 \pm 0.004$, RMSE = 0.0023 wt%), maximum hardness ($R^2 = 0.963 \pm 0.005$, RMSE = 1.27 HV), and effective case depth ($R^2 = 0.999 \pm 0.001$, RMSE = 0.0053 mm). The sequential information flow—where predicted carbon content informs hardness prediction, which subsequently constrains case depth estimation—ensures thermodynamic and kinetic consistency throughout the prediction chain. Statistical significance testing confirms that the PIML approach significantly outperforms ML-only baselines for case depth prediction ($p < 0.001$, paired t -test).
2. **Physics-derived features dominate the prediction mechanism.** Feature importance analysis using three independent methods (permutation importance, tree-based impurity, and SHAP values) consistently demonstrates that physics-derived features collectively account for over 70% of the predictive power. The characteristic diffusion length $\sqrt{Dt}$ contributes 42.2% importance, followed by carburizing temperature T_{carb} (18.3%) and diffusion coefficient $\log_{10}(D)$ (11.2%). This distribution validates the theoretical foundation of the model: diffusion kinetics govern surface carbon evolution, while temperature influences both diffusion rate and final microstructure. The remaining 30% importance distributed among the process parameters (t_{diff} , t_{boost} , and T_{quench}) captures secondary effects that are not fully described by the simplified analytical solutions.
3. **Calibrated physics parameters bridge theory and industrial practice.** The depth scale factor $\alpha = 3.26$ (95% CI: [2.89, 3.63]) quantifies the systematic deviation between idealized Fick's law predictions and measured effective case depths. This calibration accounts for (i) alloy-specific diffusion behavior in 20Cr2Ni4 steel versus pure Fe-C systems; (ii) the operational definition of effective case depth at a 550 HV threshold; and (iii) multi-stage process effects, including boost–diffusion cycling and quenching kinetics. The carbon–hardness relationship ($HV = 521.4 + 498.2 \times C$, $R^2 = 0.994$) aligns with established metallurgical principles for martensitic steels, with coefficients within 10% of the literature values for similar alloy systems.
4. **Significant improvement over baseline approaches.** Comprehensive benchmarking against five modeling paradigms demonstrates the advantages of physics-informed integration: (i) pure physics models (uncalibrated Fick's law) fail for case depth prediction ($R^2 < 0$) due to idealized assumptions; (ii) calibrated physics achieves reasonable accuracy ($R^2 \approx 0.96$) but lacks flexibility for residual corrections; (iii) pure ML approaches require substantially more training data and exhibit 1.5% lower R^2 for case depth; (iv) the PIML cascade achieves the best overall performance with 1.8% cumulative improvement in case depth R^2 through progressive physics integration. The ablation study confirms that each physics component contributes positively: physics features (+0.8%), calibrated α (+0.7%), and cascade architecture (+0.3%).
5. **Robust generalization with limited experimental data.** Despite training on only seven original experimental samples augmented to one-thousand synthetic samples, the model demonstrates robust performance on held-out test data (two-hundred

samples) and leave-one-out cross-validation on original data (LOOCV R^2 : 0.87–0.95). Physics-informed feature engineering effectively reduces the effective dimensionality of the learning problem, enabling reliable predictions from small datasets—a critical advantage for industrial applications where experimental data are expensive to obtain. Extrapolation analysis shows that predictions remain physically plausible (monotonic relationships preserved) even at domain boundaries, although with increased uncertainty (prediction intervals expand by 50–80% at extreme conditions).

Limitations and Future Directions

Several limitations of the current study should be acknowledged. First, the physics-constrained data augmentation strategy generates synthetic samples with targets computed from calibrated physics models; consequently, the high R^2 values on augmented test data (0.96–0.99) reflect the model’s ability to learn embedded physics relationships rather than purely empirical generalization. The LOOCV results on original experimental data ($R^2 = 0.87$ –0.95) provide a more realistic assessment of predictive capability for new experiments. Second, the model is validated specifically for vacuum carburizing with an acetylene (C_2H_2) atmosphere; extension to gas carburizing ($CO/CO_2/CH_4$) or plasma-assisted processes would require atmosphere-specific feature engineering and recalibration. Third, the training data span a relatively narrow range of carburizing temperatures (930–970 °C) and times (2–6 h); predictions outside this domain carry higher uncertainty, and users are advised to limit high-confidence predictions to the calibrated range (920–980 °C). Fourth, the current model predicts point estimates rather than complete concentration or hardness profiles as functions of depth.

Future research will address these limitations through three primary directions:

- **Profile prediction:** Extending the model to predict complete carbon concentration $C(x)$ and hardness $HV(x)$ profiles as continuous functions of depth, enabling more comprehensive process optimization.
- **Real-time integration:** Incorporating in situ atmosphere monitoring data (carbon potential and oxygen partial pressure) and furnace thermal profiles for adaptive model updating during production.
- **Industrial deployment:** Developing a closed-loop optimization system that couples the PIML model with process controllers for real-time recipe adjustment in industrial vacuum-carburizing furnaces, with the ultimate goal of achieving first-pass success rates exceeding 95% for gear manufacturing applications.

The methodology presented herein provides a generalizable framework for physics-informed modeling of vacuum-carburizing processes in Cr-Ni case-hardening steels. While the fundamental cascade architecture and physics-integration strategy may inspire similar approaches for other thermochemical treatments, direct application to fundamentally different processes (gas carburizing, nitriding, and carbonitriding) would require process-specific recalibration and validation.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/met16020163/s1>. Note: Figure S1 (KDE visualization) has been incorporated into the main text as Figure 2 in response to reviewer feedback. Additional supplementary data includes a detailed breakdown of 34 rejected samples by constraint type (76% carbon ceiling, 18% hardness ceiling, 6% depth minimum); Python verification code for physical calculations (`verification_calculations.py`).

Author Contributions: Conceptualization, C.L. and P.C.; methodology, C.L.; software, C.L.; validation, C.L. and P.C.; formal analysis, C.L.; investigation, C.S.; resources, P.C.; data curation, C.S.; writing—original draft preparation, C.L.; writing—review and editing, P.C. and C.S.; visualization,

C.L.; supervision, P.C.; project administration, P.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The experimental data supporting this study are available from the corresponding author upon reasonable request. Python verification code for all physical calculations (diffusion coefficients, carbon–hardness relationships, case depth models, and error propagation) is provided as Supplementary Materials (verification_calculations.py). All numerical values presented in the paper have been independently verified using NumPy and SciPy libraries. The physics-constrained data augmentation methodology is fully described in Section 2.1 to enable reproduction.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PIML	Physics-Informed Machine Learning
ML	Machine Learning
GBR	Gradient Boosting Regression
RF	Random Forest
ETs	Extra Trees
LOOCV	Leave-One-Out Cross-Validation
LHS	Latin Hypercube Sampling
PCDA	Physics-Constrained Data Augmentation
RMSE	Root Mean Square Error
HV	Vickers Hardness
PDE	Partial Differential Equation
CI	Confidence Interval
SHAPs	SHapley Additive exPlanations

References

1. Davis, J.R. (Ed.) *Surface Hardening of Steels: Understanding the Basics*; ASM International: Materials Park, OH, USA, 2002.
2. Goldstein, J.I.; Moren, A.E. Diffusion modeling of the carburization process. *Metall. Trans. A* **1978**, *9*, 1515–1525. [\[CrossRef\]](#)
3. Geoffrey, P. *Carburizing: Microstructures and Properties*; ASM International: Materials Park, OH, USA, 1999.
4. Shewmon, P. *Diffusion in Solids*, 2nd ed.; Wiley: Hoboken, NJ, USA, 1989.
5. David, P. *Practical Nitriding and Ferritic Nitrocarburizing*; ASM International: Materials Park, OH, USA, 2003.
6. Genel, K.; Demirkol, M.; Çapa, M. Effect of case depth on fatigue performance of AISI 8620 carburized steel. *Int. J. Fatigue* **1999**, *21*, 207–212. [\[CrossRef\]](#)
7. Butler, K.T.; Davies, D.W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555. [\[CrossRef\]](#)
8. Raccuglia, P.; Elbert, K.C.; Adler, P.D.; Falk, C.; Wenny, M.B.; Mollo, A.; Zeller, M.; Friedler, S.A.; Schrier, J.; Norquist, A.J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73–76. [\[CrossRef\]](#)
9. Karniadakis, G.E.; Kevrekidis, I.G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-informed machine learning. *Nat. Rev. Phys.* **2021**, *3*, 422–440. [\[CrossRef\]](#)
10. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686–707. [\[CrossRef\]](#)
11. Haghighat, E.; Raissi, M.; Moure, A.; Gomez, H.; Juanes, R. A physics-informed deep learning framework for inversion and surrogate modeling in solid mechanics. *Comput. Methods Appl. Mech. Eng.* **2021**, *379*, 113741. [\[CrossRef\]](#)
12. Ankit, A.; Choudhary, A. Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. *APL Mater.* **2016**, *4*, 053208. [\[CrossRef\]](#)
13. Wells, C.; Batz, W.; Mehl, R.F. Diffusion coefficient of carbon in austenite. *Trans. AIME* **1950**, *188*, 553–560. [\[CrossRef\]](#)

14. Wang, X.; Li, Y.; Wu, J. Intelligent optimization of gas carburizing process parameters using machine learning. *Surf. Coatings Technol.* **2022**, *449*, 128952.
15. Kim, J.; Lee, S.; Park, H. Deep learning-based prediction of effective case depth in carburized steels. *Metals* **2024**, *14*, 234.
16. Liu, H.; Zheng, C.; Zhang, G. Data-driven modeling and optimization of vacuum carburizing for gear steels. *J. Mater. Sci. Technol.* **2023**, *152*, 78–89.
17. Zhang, Z.; Zheng, Y.; Niu, L. Physics-constrained data augmentation for machine learning in materials science. *Npj Comput. Mater.* **2023**, *9*, 187.
18. Li, Y.; Chen, W.; Doe, J. Latin hypercube sampling with physics constraints for small-sample machine learning in heat treatment. *J. Mater. Process. Technol.* **2024**, *325*, 118298.
19. Cuomo, S.; Di Cola, V.S.; Giampaolo, F.; Rozza, G.; Raissi, M.; Piccialli, F. Scientific machine learning through physics-informed neural networks: Where we are and what's next. *J. Sci. Comput.* **2022**, *92*, 88. [[CrossRef](#)]
20. Willard, J.; Jia, X.; Xu, S.; Steinbach, M.; Kumar, V. Integrating scientific knowledge with machine learning for engineering and environmental systems. *Acm Comput. Surv.* **2022**, *55*, 1–37. [[CrossRef](#)]
21. Viana, F.A.; Nascimento, R.G.; Dourado, A.; Yucesan, Y.A. Estimating model inadequacy in ordinary differential equations with physics-informed neural networks. *Comput. Struct.* **2023**, *279*, 106962. [[CrossRef](#)]
22. Almonti, D.; Baiocco, G.; Ucciardello, N. Pulp and paper characterization by means of artificial neural networks for effluent solid waste minimization—A case study. *J. Process Control* **2021**, *105*, 283–291. [[CrossRef](#)]
23. Nasiri, M.; Iqbal, S.; Särkkä, S. Physics-informed machine learning for grade prediction in froth flotation. *Miner. Eng.* **2025**, *227*, 109297. [[CrossRef](#)]
24. Chinchilla, R.; Deshpande, V.M.; Chakrabarty, A.; Laughman, C.R. Learning residual dynamics via physics-augmented neural networks: Application to vapor compression cycles. In *2023 American Control Conference (ACC)*; IEEE: San Diego, CA, USA, 2023; pp. 4069–4076.
25. John, C. Thermodynamics and phase diagram of the Fe-C system. *Metall. Trans.* **1972**, *3*, 55–64.
26. George, K. Martensite in steel: Strength and structure. *Mater. Sci. Eng. A* **1999**, *273–275*, 40–57.
27. John, Å. A revised expression for the diffusivity of carbon in binary Fe-C austenite. *Scr. Metall.* **1986**, *20*, 1507–1510.

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