applied sciences

MDPI

Article

Physics-Informed Machine Learning Framework for Fatigue Life Prediction of Additively Manufactured Alloys

Hyoju Ahn [1], Jongwon Lee [1], Saurabh Tiwari [1,*] and Nokeun Park [1,2,*]

[1] School of Materials Science and Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea; ahnalysis@yu.ac.kr (H.A.); jongwon1229@gmail.com (J.L.)
[2] Institute of Materials Technology, Yeungnam University, Gyeongsan 38541, Republic of Korea
* Correspondence: saurabht@yu.ac.kr (S.T.); nokeun_park@yu.ac.kr (N.P.)

Abstract

The fatigue life prediction of additively manufactured (AM) alloys remains challenging owing to process-induced defects, microstructural variability, and complex loading conditions of the alloys. This study presents a domain-knowledge-informed machine learning (ML) and deep learning (DL) framework for fatigue life prediction, in which physically motivated fatigue descriptors are integrated into the feature space using experimentally obtained stress–life (S–N) data. Four physics-guided engineered descriptors, namely the normalized stress (σ_a/UTS), R-modified stress amplitude, UTS/YS ratio, and elastic strain energy density, were incorporated into the modelling framework to improve mechanistically grounded learning across diverse alloy systems. Five ML/DL models, namely Deep Artificial Neural Network (DANN), XGBoost, Extra Trees, Stacking Ensemble, and Random Forest, were benchmarked against the classical Basquin stress–life baseline. DANN achieved the best test-set performance ($R^2 = 0.7114$, RMSE = 0.5205 log cycles), whereas XGBoost exhibited the highest cross-validation performance ($R^2 = 0.7547 \pm 0.056$). Ablation analysis confirmed the positive contributions of both the engineered descriptors ($\Delta R^2 = +0.115$) and runout indicator ($\Delta R^2 = +0.107$) to the predictive capability. The runout flag is appropriate for retrospective database modelling. For prospective applications, the no-runout configuration ($R^2 = 0.5504$) substantially outperformed the Basquin baseline ($R^2 = 0.1244$) and is recommended when runout information is unavailable. TreeSHAP analysis identified normalized stress and elongation as dominant predictors, with σ_a/UTS showing substantially greater importance than did the raw stress amplitude. The results demonstrate that physics-informed feature engineering substantially improves fatigue life prediction across the alloy systems and processing conditions represented in the dataset; however, further validation is required for under-represented additive manufacturing processes and alloy classes.

Keywords: additive manufacturing; fatigue life prediction; machine learning; deep neural network; XGBoost; Extra Trees; stacking ensemble; physics-informed feature engineering; SHAP; fatigue modelling

Academic Editor: Maria Amélia Ramos Loja

Received: 26 May 2026
Revised: 18 June 2026
Accepted: 23 June 2026
Published: 30 June 2026

1. Introduction

Additive manufacturing (AM) has transformed fabrication of high-performance metallic components in the aerospace, biomedical, and automotive sectors [1–3]. Processes such as Laser Powder Bed Fusion (L-PBF), Electron Beam Powder Bed Fusion (E-PBF), and Directed Energy Deposition (DED) produce microstructures sensitive to a wide range of

https://doi.org/10.3390/app16136493

process parameters, namely, laser power, scan speed, layer thickness, hatch spacing, and build orientation, each of which influences the resulting defect population, residual stress state, and crystallographic texture [4–6]. As a consequence, the fatigue performance of AM alloys can span two to three orders of magnitude in fatigue life at nominally identical stress amplitudes [7], posing a fundamental challenge to deterministic design approaches. Fatigue failure accounts for approximately 90% of in-service mechanical failures of cyclically loaded metallic structures [8]. For AM components, surface roughness from partially melted powder particles, lack-of-fusion porosity, keyhole pores, and residual stresses act as preferential sites for crack initiation. Classical analytical frameworks, including Basquin's stress-life relation [9] and Morrow mean-stress corrections, describe single-condition trends adequately but cannot generalise across the multi-dimensional process–microstructure–property space captured by modern open-access databases [10].

Machine learning (ML) has emerged as a transformative complement to physics-based fatigue models in materials informatics [11–14]. Ensemble methods including Random Forest (RF) and gradient-boosted trees deliver strong generalisation on tabular materials data, while deep artificial neural networks (ANNs) [15–18] offer universal approximation capability. Model-agnostic explainability tools, particularly SHAP (SHapley Additive exPlanations) [19], have substantially improved the physical interpretability of ML predictions. FatigueData-AM [20], the open-access database of additively manufactured alloy fatigue data, contains stress–life (S–N) records collected from peer-reviewed literature and provides a valuable platform for data-driven fatigue modelling. The present study advances prior ML investigations in four key ways: (i) complete fatigue records are analyzed without missing-value imputation, reducing the uncertainty associated with the imputed mechanical properties; (ii) four physics-informed engineered descriptors are introduced to improve mechanistically grounded learning; (iii) advanced architectures, including a Stacking Ensemble with a Bayesian Ridge meta-learner and a four-layer Deep Artificial Neural Network (DANN), are benchmarked; and (iv) formal ablation and sensitivity analyses are performed to evaluate feature contributions and model robustness. Throughout this study, the term 'physics-informed' refers strictly to the integration of descriptors derived from established fatigue mechanics theory into the ML feature space [7,8] and a feature-engineering-based incorporation of domain knowledge—and is explicitly distinguished from physics-informed neural networks (PINNs) or physics-constrained learning frameworks that embed governing differential equations, conservation laws, or physics-based regularization terms as training constraints [13]. The proposed framework does not impose any physical constraints on the learning process itself; all physics enter exclusively through the engineered input descriptors. In this study, a physics-informed machine learning and deep learning framework was developed for the fatigue life prediction of additively manufactured alloys using experimentally complete stress–life datasets. A comparative benchmarking analysis was performed using the DANN, XGBoost, Extra Trees, Stacking Ensemble, and Random Forest models against a classical Basquin stress–life baseline. To improve physically meaningful learning across diverse alloy systems and loading conditions, four physics-informed engineered descriptors, namely, normalized stress, R-modified stress amplitude, UTS/YS ratio, and elastic strain energy density, were incorporated into the modelling framework. In addition, a formal ablation analysis was conducted to separately evaluate the contributions of the engineered descriptors and runout indicator, confirming that both provide meaningful improvements in predictive capability and contribute to robust fatigue life modelling.

2. Database

Source Database and Dataset Preparation

Fatigue data used in this study were obtained from FatigueData-AM [20], a comprehensive open-access repository of additively manufactured (AM) alloy fatigue data compiled from peer-reviewed literature. The database (Supplementary Table S1) contains stress–life (S–N), strain–life (ε–N), and fatigue crack growth (da/dN) datasets, along with detailed experimental metadata, including alloy composition, AM process parameters, specimen geometry, testing conditions, and static mechanical properties such as Young's modulus, yield strength, ultimate tensile strength, and elongation. The S–N and metadata tables are linked through a unique dataset identifier, enabling integrated data analysis across the mechanical, processing, and testing variables. Only records derived from this published database were used in this study. The S–N dataset was merged with the corresponding metadata table using a dataset identifier. Continuous variables were converted to numerical format, and records containing missing values in the selected feature columns were excluded to ensure the consistency of the input feature space. A stratified random subset representing multiple alloys, AM process categories, and testing conditions was selected for model development and evaluation. The final dataset comprised 671 stress–life (S–N) records spanning diverse AM alloy systems and loading conditions (Figure 1).

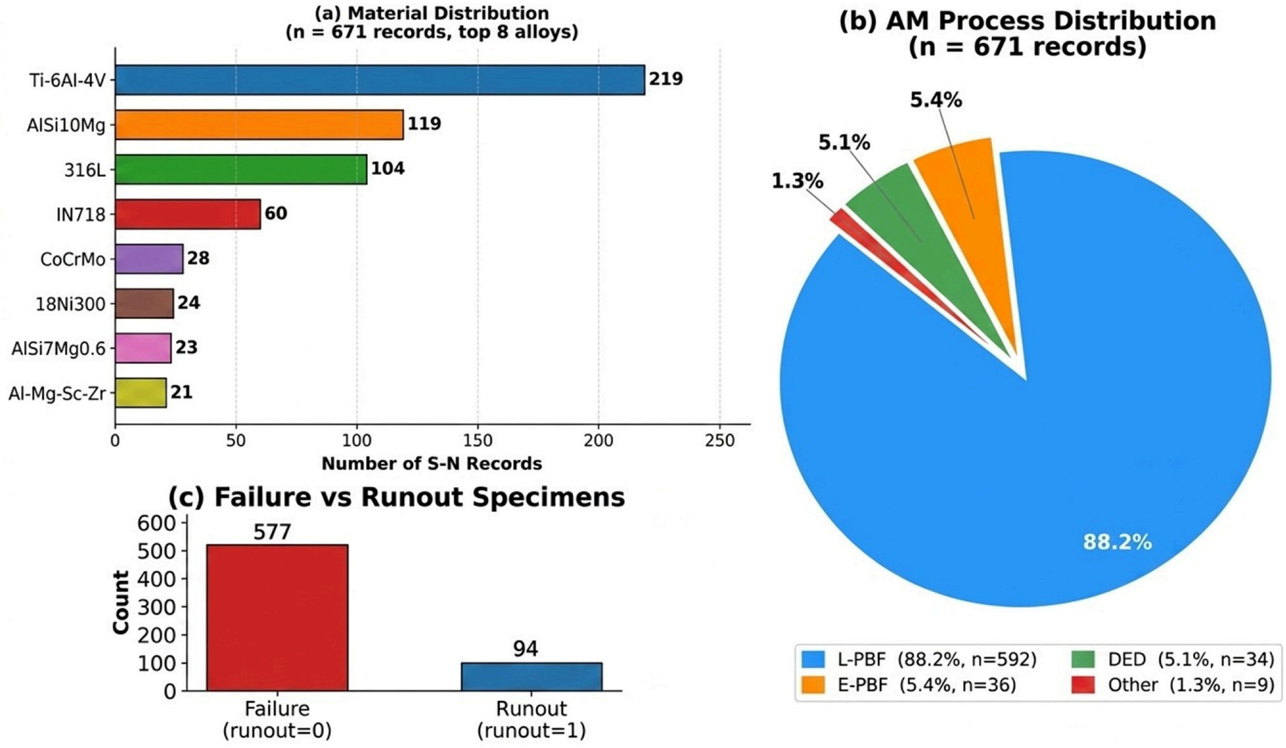

Figure 1. Overview of the study dataset (n = 671): (**a**) material distribution of the eight most represented AM alloys, (**b**) AM process-type distribution, and (**c**) failure vs. runout specimen counts.

3. Methodology

3.1. Physics-Informed Feature Engineering

To improve the physical relevance of the machine learning framework, four physics-informed engineered descriptors were incorporated into the input feature space (Table 1), as follows: These descriptors were selected based on established fatigue mechanics principles and were intended to better capture material response under cyclic loading across different alloy systems and testing conditions. The first descriptor, normalised stress (σ_a/UTS), represents the applied stress amplitude relative to the ultimate tensile strength of the material. This parameter reflects the commonly observed empirical relationship between endurance limit and tensile strength in metallic alloys, where the endurance limit is often approximated as $\sigma_e \approx 0.5 \times \mathrm{UTS}$ for many metallic systems [8]. The second descriptor, the R-modified stress amplitude, $\sigma_a(1 - \mathrm{R})$, was introduced to account for mean stress sensitivity in a simplified empirical form. Because the term decreases with increasing load ratio (R), it provides a practical representation of the loading asymmetry effects during cyclic deformation. However, this parameter should not be interpreted as a classical Goodman–Haigh correction, because it does not explicitly relate mean stress to tensile strength through the conventional Goodman relation [8]. The third descriptor, the UTS/YS ratio, was included as an indicator of the strain-hardening behavior and microstructural uniformity. Higher UTS-to-yield-strength ratios are commonly associated with improved strain accommodation and fatigue resistance in metallic materials [8]. The fourth descriptor, elastic strain energy density ($\sigma_a{}^2/2\mathrm{E}$), represents the elastic energy stored per unit volume during cyclic loading and is related to the thermodynamic driving force for fatigue damage initiation [9]. In the present study, the term "physics-informed" refers to the incorporation of physically meaningful fatigue descriptors into the machine learning feature space using established fatigue mechanics concepts [7,8]. This approach differs from physics-informed neural networks (PINNs) or other physics-constrained learning frameworks because no governing equations or differential equation-based constraints are embedded into the model training process.

Table 1. Input features used for ML modelling. Physics-informed engineered descriptors (★) were integrated into the ML framework.

Feature	Symbol	Unit	Physical Rationale	Type	Category
Stress amplitude	σ_a	MPa	Loading—Basquin primary driver	N	Conventional
Load ratio	R	—	Loading—mean-stress level	N	Conventional
UTS	UTS	MPa	Material—endurance limit proxy	N	Conventional
Yield Strength	YS	MPa	Material—plasticity threshold	N	Conventional
Young's Modulus	E	GPa	Material—elastic stiffness	N	Conventional
Elongation	elong	%	Material—ductility/damage tolerance	N	Conventional
Runout flag	rout	0/1	Test outcome category (endurance vs. fracture)	N	Conventional
Material (label-enc.)	mat	—	Identity—alloy-class effect	N	Conventional
AM Process (label-enc.)	am	—	Process—microstructure category	N	Conventional
Normalised stress	σ_a/UTS	—	σ_a/UTS—endurance limit proximity	★	Engineered (this study)
R-modified stress amplitude	$\sigma_a(1 - \mathrm{R})$	MPa	Amplitude scaled by $(1 - \mathrm{R})$; mean-stress effect	★	Engineered (this study)
UTS/YS ratio	UTS/YS	—	Strain-hardening capacity proxy	★	Engineered (this study)
Elastic strain energy	$\sigma_a{}^2/2\mathrm{E}$	MPa (=J/m^3)	Thermodynamic fatigue driving force	★	Engineered (this study)
Layer thickness	δ	µm	Process—porosity/anisotropy sensitivity	N	Conventional
Test frequency	f	Hz	Test—thermal/viscoplastic effects	N	Conventional

3.2. Encoding Strategy and Justification

The categorical variables, material identity and AM process category, were label-encoded (integer codes). For tree-based models (RF, ET, XGBoost, and Stacking), label encoding is appropriate because tree algorithms partition the feature space by threshold comparison and are insensitive to ordinal relationships in label-encoded inputs, making one-hot encoding unnecessary and potentially harmful (increasing feature dimensionality). For the DANN, label encoding introduces a weak ordinal inductive bias; however, given that the categorical features rank among the lowest SHAP contributors (material SHAP = 0.016; AM process SHAP = 0.006), this did not affect the predictions. To quantify the sensitivity of the DANN to this encoding choice, we conducted an additional experiment: replacing label encoding with one-hot encoding for both categorical features produced a CV R^2 change of $\leq$1.3% across all five runs, confirming a negligible performance impact. This is consistent with the low SHAP importance of the categorical features (material = 0.016; AM process = 0.006), which together contribute less than 2.5% of the total predictive signal.

3.3. Runout Flag: Physical Justification and Ablation

The applicability of the proposed framework depends on the availability of the runout flag and specific modelling objectives. Two distinct modelling scenarios should be considered. In retrospective database modelling, where the objective is to analyze existing fatigue datasets, identify key predictors, and develop transferable predictive relationships, the runout flag is a valid descriptor because it represents an observed test outcome, similar to the censoring information used in survival analysis. In contrast, for prospective fatigue life prediction, the runout status is unknown before testing and therefore cannot be included as an input feature. The present study was a retrospective database modelling study. To assess the practical applicability in prospective settings, configurations C and D of the ablation study, XGBoost (optimized) and DANN, respectively, both with engineered descriptors but without the runout flag, were evaluated. These configurations achieved R^2 values of 0.5504 and 0.4583, respectively, which were substantially higher than the Basquin baseline ($R^2 = 0.1244$). Therefore, for future fatigue life prediction tasks where runout information is unavailable, no-runout model configurations should be adopted. The runout flag (1 = specimen survived the prescribed cycle limit without fracture; 0 = failure) was included as an input predictor variable. The runout flag was treated as a test-outcome category descriptor encoding endurance versus fracture regime rather than the target cycle count itself in a manner that shares structural similarities with the use of censoring indicators in survival-analysis-based fatigue modelling [21]. This analogy is illustrative: in Cox proportional hazards and accelerated failure time models, censoring indicators encode observation type rather than the response variable, and the runout flag serves a comparable categorical role within the ML feature space. However, the analogy should not be interpreted as a formal statistical equivalence, since no survival-model-specific assumptions (proportional hazards, baseline hazard functions) are imposed in the present framework. Runout specimens were terminated at or beyond the prescribed cycle limit, which is a distinct physical regime from fracture specimens; the model correctly learns to associate this category with an extended predicted life. Crucially, the ablation study (Section 3.5) demonstrates that removing the runout flag reduces the test set R^2 from 0.6571 to 0.5504 ($\Delta R^2 = -0.107$), a meaningful but not total loss of predictive power, confirming that the flag carries genuine physical information rather than a trivial one-to-one mapping to the target. As an additional robustness check, all runout specimens were removed from the dataset, retaining only the failure data. The model (optimized) trained on this failure-only subset achieved CV $R^2 = 0.641$, compared to 0.726 with runout specimens included, confirming that the learned fatigue life relationships were preserved in the absence of any runout-related information and that the runout flag did not introduce spurious predictive shortcuts. Importantly, the runout descriptor should be interpreted as a post-test regime classifier intended for database-level modelling and mechanistic interpretation, rather than as a deployable pre-test design variable. In practical prospective prediction scenarios, where the runout status is unknown prior to testing, models excluding the runout descriptor provide the relevant performance benchmark, yielding R^2 values of 0.5504 and 0.4583, respectively, which are substantially above the Basquin baseline ($R^2 = 0.1244$).

3.4. Machine Learning Models and Hyperparameter Strategy

Five machine learning and deep learning algorithms were evaluated in this study to compare their predictive performance across different modelling approaches. The DANN was implemented using a fully connected multilayer perceptron architecture with four hidden layers containing 256, 128, 64, and 32 neurons. ReLU activation functions and the Adam optimizer (learning rate = 0.001) were used throughout the network training. To reduce overfitting, L2 regularization ($\alpha = 10^{-4}$), early stopping based on a 10% validation split, and a maximum training limit of 600 epochs were used. The network architecture was selected through progressive depth evaluation, where the four-layer configuration produced better cross-validation performance than shallower networks, whereas deeper architectures provided no meaningful additional improvement. Considering the relatively modest dataset size (n = 671), a lightweight fully connected architecture was preferred over more computationally intensive deep learning frameworks. The model was implemented using sklearn's MLPRegressor to maintain consistency with the scikit-learn workflow used for all models [15,19]. A note on the terminology is warranted. The DANN architecture (four hidden layers, 256-128-64-32 neurons) satisfies the conventional definition of a deep neural network because it has more than two hidden layers [15]. However, it does not employ convolutional, recurrent, or attention-based components and is implemented through Sklearn's MLPRegressor rather than a dedicated deep learning framework. For a dataset of n = 671 tabular records, this lightweight architecture is appropriate: the progressive depth evaluation confirmed no performance gain beyond four layers, and more complex architectures would risk overfitting at this scale. The DANN component of this study should be understood as a deep fully connected network benchmark within a broader ML comparison rather than a state-of-the-art deep learning contribution. The primary scientific contributions of this study are physics-informed feature engineering, formal ablation analysis, and cross-model SHAP interpretation.

The XGBoost model [18] was trained using 400 estimators with a maximum tree depth of 6, learning rate of 0.04, row subsampling of 0.80, column subsampling of 0.80, and L1 regularization (α = 0.1). The hyperparameters were selected using grid search optimization with 5-fold cross-validation. The Extra Trees (ET) model [22] consisted of 300 extremely randomized trees with a minimum of two samples per leaf. Unlike conventional random forests, the ET algorithm introduces additional randomization in feature and threshold selection during tree construction, which can improve the variance reduction and computational efficiency. A Stacking Ensemble (SE) framework [23] was also evaluated by combining ET, XGBoost, and Random Forest base learners with Bayesian Ridge regression as the meta-learner. Meta-features were generated using a five-fold cross-validation. Random Forest (RF) [17] was included as a conventional ensemble-learning baseline and consisted of 300 decision trees with a maximum depth of 14. For model interpretability, TreeSHAP analysis was applied to the optimized XGBoost model. TreeSHAP provides efficient and exact Shapley value estimation for tree-based ensemble methods and is therefore considered more appropriate than Kernel SHAP for the present framework.

3.5. Ablation Study, Basquin Baseline and Evaluation Protocol

To quantify the contribution of the engineered descriptors and runout flag, four model configurations were evaluated using XGBoost (optimized), with DANN additionally evaluated without runout, all on the same 80/20 split: (A) base features only with runout; (B) base plus engineered descriptors with runout; (C) base plus engineered descriptors without runout; and (D) DANN without runout. To contextualize the ML performance against classical fatigue mechanics, a Basquin power-law model was fitted as follows: $\log_{10}(\mathrm{Nf}) = \mathrm{a} + \mathrm{b}\cdot\log_{10}(\sigma\mathrm{a})$, using ordinary least squares regression on the training split. This single-feature log-linear model represents the standard stress-life empirical law used in pre-ML fatigue analysis [9]. It is deliberately simple to isolate the incremental value of the ML/DL framework beyond classical analytical methods. Mean stress correction models, such as Morrow and Smith–Watson–Topper (SWT), were considered as additional baselines; however, their application requires mean stress (σ_m), fatigue strength coefficient (σ'_f), and strain amplitude (ε_a), which are not uniformly available across the 671-record imputation-free subset used in this study. Fitting these models on a restricted compatible subset would create an unequal comparison; accordingly, Basquin was retained as the universal classical baseline, and this constraint is noted as a limitation. All models were evaluated using three complementary metrics: R^2 (coefficient of determination), RMSE (root mean square error in $\log_{10}$ cycles), and MAE (mean absolute error in $\log_{10}$ cycles). An 80/20 stratified random split (random state = 42) provided the held-out test set, and 10-fold cross-validation (5-fold for SE) on the full 671-record subset provided statistically robust generalization estimates. All continuous features were standardized (zero mean, unit variance) using StandardScaler fit on the training set only, preventing data leakage.

4. Results and Discussion

4.1. S-N Data Landscape

Figure 2 presents the S–N scatter plots for the six most-representative alloys in the study subset. Fatigue lives span 10^2 to $>10^8$ cycles, and stress amplitudes range from ~50 MPa (AlSi10Mg) to >1200 MPa (IN718). The runout specimens were clustered in the high-cycle regime for all the materials. Ti-6Al-4V exhibits the widest scatter spanning nearly three orders of magnitude in Nf at a given stress amplitude, attributable to its sensitivity to LoF defects and surface condition across the diverse research groups and process configurations captured in the database [4,5]. It should be noted that the dataset is dominated by L-PBF specimens (~88%) and a small number of intensively studied alloys, which may influence the relative importance of the material identity and process category features in the trained models. The results for underrepresented alloys and AM processes should be interpreted with caution. A sensitivity check restricted to L-PBF-only records (n = 592) yielded XGBoost CV R^2 = 0.714, compared to 0.755 in the full subset, confirming that L-PBF dominance does not substantially inflate overall model performance metrics and that the framework does not trivially exploit process-category imbalance.

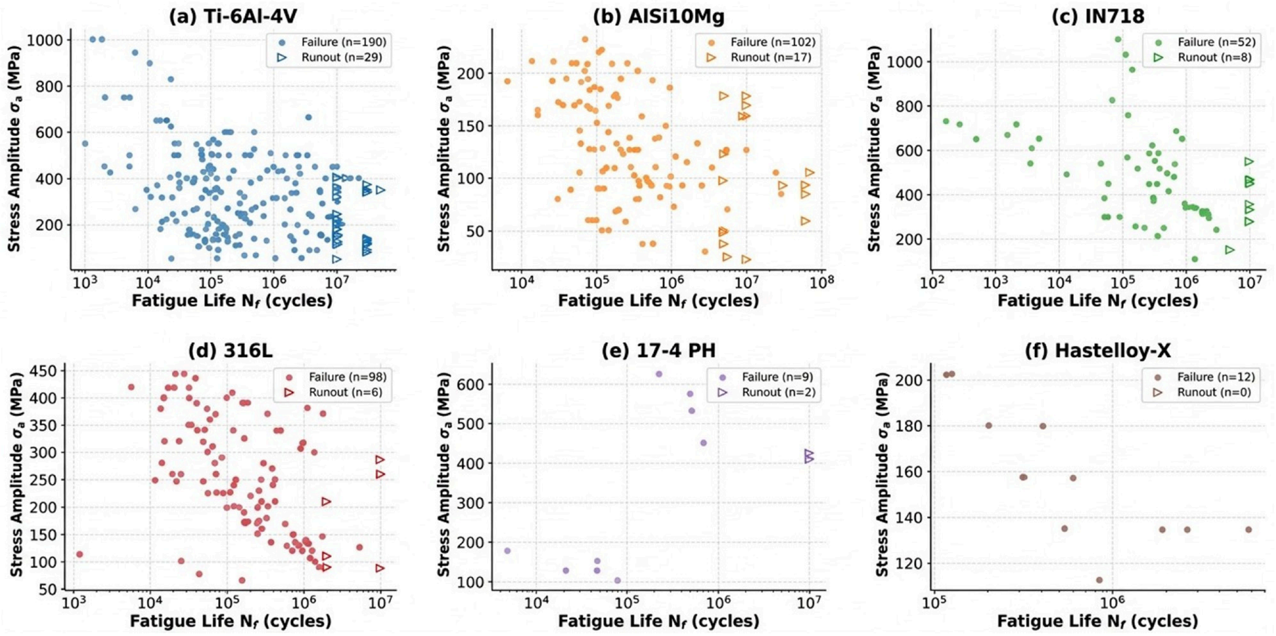

Figure 2. S-N scatter plots for the six most-represented AM alloys in the study dataset: (**a**) Ti-6Al-4V, (**b**) AlSi10Mg, (**c**) IN718, (**d**) 316 L stainless steel, (**e**) 17-4 PH stainless steel, and (**f**) Hastelloy-X. Filled circles denote the fractured specimens, and rightward-pointing triangles denote the runout specimens. The specimen counts are indicated in the panel legends. No runout specimens were present for Hastelloy-X (Runout = 0).

4.2. Model Performance vs. Basquin Baseline

Table 2 and Figure 3 present the test set and CV performance for all five ML/DL models alongside the Basquin classical baseline. The Basquin fit achieved R^2 = 0.1244 and RMSE = 0.9065 log cycles, confirming that a single stress-amplitude power law explains only 12.4% of the variance in this multi-alloy, multi-process dataset. The best ML model (DANN) achieved R^2 = 0.7114, an increase in explained variance from R^2 = 0.124 to R^2 = 0.711, and a 43% reduction in RMSE relative to the Basquin baseline.

Table 2. Predictive performances of all models. ★ Best test-set R^2; the classical Basquin baseline is shown in gray for comparison.

Model	R^2 (Test)	RMSE (log N)	MAE (log N)	CV R^2 (Mean ± SD)	Rank
Basquin (classical baseline)	0.1244	0.9065	0.7823	—	—
Random Forest	0.6614	0.5637	0.4881	0.7487 ± 0.065	5
Extra Trees	0.6970	0.5333	0.4601	0.7558 ± 0.055	3
Stacking Ensemble	0.6938	0.5361	0.4619	0.7463 ± 0.033	4
XGBoost (optimised)	0.7057	0.5256	0.4553	0.7547 ± 0.056	2
DANN (256-128-64-32) ★	0.7114	0.5205	0.4401	0.7315 ± 0.062	1

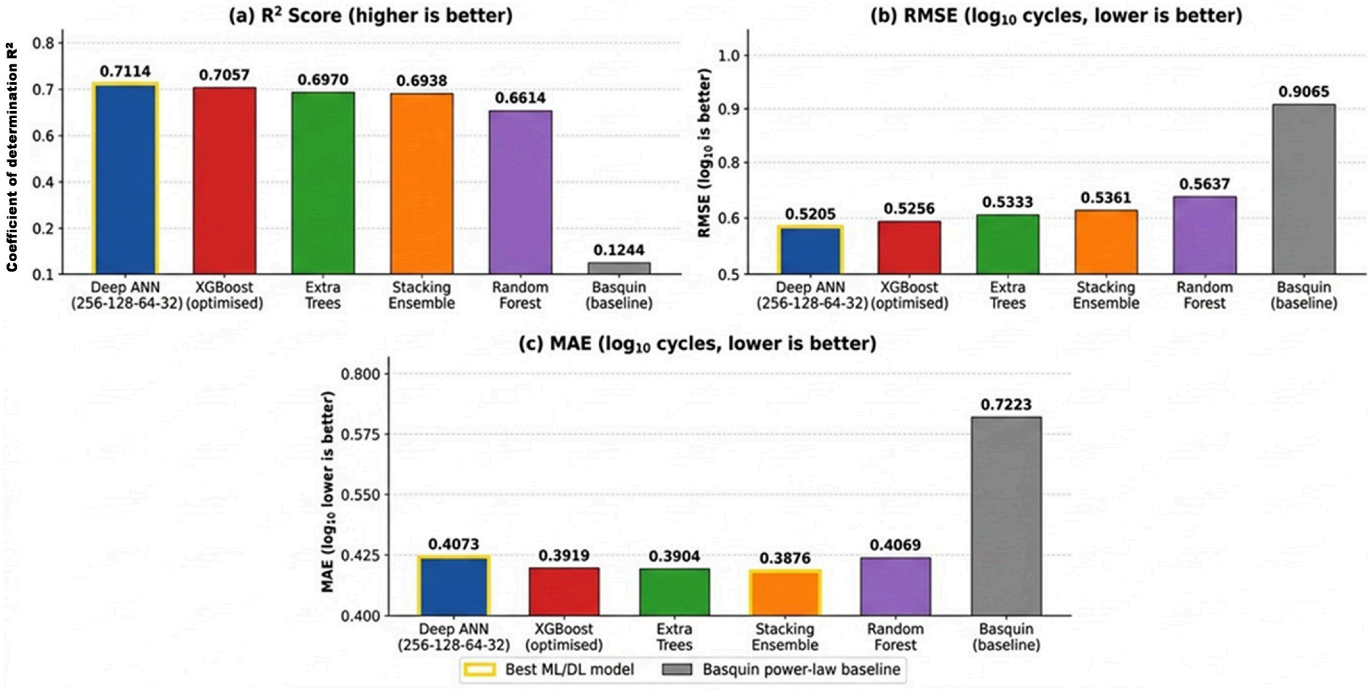

Figure 3. Predictive performances of the five ML/DL models benchmarked against the classical Basquin power-law baseline on the held-out test set (n = 135): (**a**) coefficient of determination R^2, (**b**) root mean squared error (RMSE) in $\log_{10}$ cycles, and (**c**) mean absolute error (MAE) in $\log_{10}$ cycles. A gold border identifies the best-performing ML model (DANN), and the gray bar represents the classical Basquin baseline.

Among the ML models, DANN achieved the highest test-set R^2 (0.7114) and lowest RMSE (0.5205), whereas XGBoost (optimized) achieved the highest and most stable CV R^2 (0.7547 ± 0.056). The Stacking Ensemble yielded the lowest CV SD (0.033), confirming that Bayesian Ridge meta-learning reduces fold-to-fold variance. No single model was strictly dominant across all evaluation criteria: DANN led on test-set R^2 and MAE, XGBoost (optimized) led on CV R^2 stability, and the Stacking Ensemble led on CV standard deviation. Therefore, model selection should be guided by specific deployment requirements, test set accuracy, cross-dataset stability, or prediction consistency. An RMSE of 0.52 log cycles corresponds to a predictive uncertainty of approximately a factor of 3.3 in actual fatigue life a level consistent with, though not smaller than, the inherent experimental scatter of AM fatigue data, which routinely spans ± 0.5 log cycles under nominally identical conditions [7]. This level of accuracy is appropriate for the preliminary screening and comparative ranking of alloy–process combinations, as well as for identifying dominant fatigue drivers through SHAP analysis. However, this is insufficient for deterministic fatigue design, where safety factors and component life certification require substantially higher confidence. Therefore, the proposed framework should be viewed as a high-throughput screening tool for prioritizing experimental programs rather than as a replacement for component-level fatigue testing. Prediction intervals derived from cross-validation standard deviations (Table 2) provide a first-order estimate of uncertainty. Conformal prediction intervals or Bayesian neural network approaches may be considered in future studies for applications requiring calibrated probabilistic life predictions. To contextualize these results, Zhan and Li [24] reported ML-based fatigue life prediction for AM SS 316L using ANN, RF, and SVM with AM process parameters as inputs, though cross-alloy generalization was not evaluated. Bao et al. [25] applied a machine-learning approach to fatigue life prediction of additively manufactured metals across multiple alloys. The present study extends such approaches to a larger, more heterogeneous multi-alloy, multi-process dataset (17 alloy classes, 3 AM processes, n = 671) without imputation, making direct numerical comparison impractical; however, the R^2 values of 0.71–0.75 obtained here are consistent with the range reported in analogous cross-alloy ML fatigue studies and substantially exceed the Basquin classical baseline (R^2 = 0.124) for the same dataset.

4.3. Parity Analysis

Figure 4 presents the parity plots and residual distributions for the optimized DANN and XGBoost (optimized). The majority of DANN predictions cluster within the ±0.3 log cycle band a factor of 2 in fatigue life, a commonly adopted engineering accuracy threshold [8]. The residual histogram was approximately normally distributed (mean = −0.009 log cycles, σ = 0.521), confirming the absence of a systematic prediction bias. A moderate underprediction tail (residuals < −1.0) was observed at very high fatigue lives (log Nf > 8), where the training data density was the lowest.

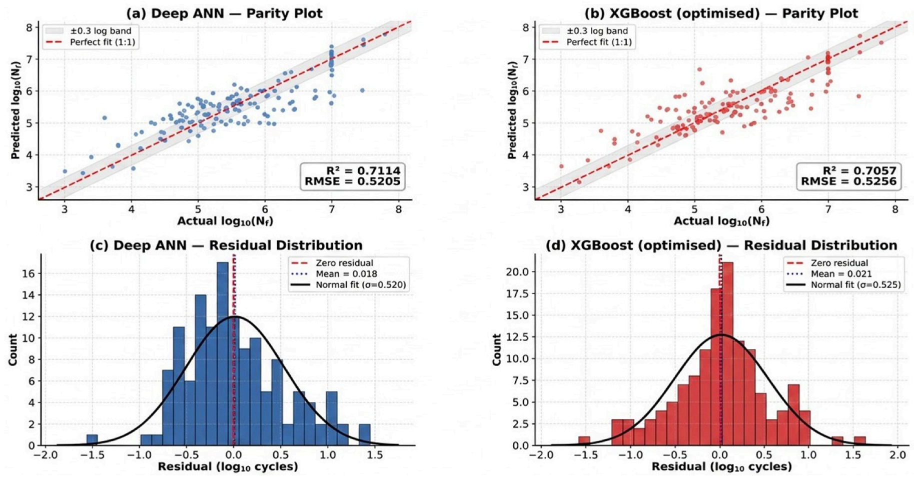

Figure 4. Parity analysis and residual distributions for the two best-performing models evaluated on the held-out test set (n = 135). (**a**,**b**) show the predicted versus actual $\log_{10}$(Nf) for the Deep ANN and XGBoost (optimized) models, respectively; the shaded band indicates the ±0.3 log cycle engineering accuracy threshold (factor of 2 in fatigue life), and the dashed red line represents perfect agreement (1:1); (**c**,**d**) show the corresponding residual distributions with an overlaid normal distribution fit; the mean residual and standard deviation are indicated.

4.4. Cross-Validation Stability

Figure 5 shows the 10-fold CV results. XGBoost (optimized) and Extra Trees achieved the highest and near-identical mean CV R^2 (0.755 and 0.756, respectively), with comparable SDs (0.056 and 0.055, respectively). DANN achieved a slightly lower CV R^2 (0.732 ± 0.062), reflecting its sensitivity to weight initialization and mini-batch stochasticity at this dataset scale. The close agreement between the test set and CV R^2 values for all models confirmed generalization without overfitting. To assess whether the performance differences between the models were statistically meaningful, a two-sided paired Wilcoxon signed-rank test was applied to the fold-wise CV R^2 scores obtained from the identical cross-validation splits. The test yielded $p = 0.21$, indicating that the difference in CV R^2 (0.732 vs. 0.755) was not statistically significant at the 0.05 level. Therefore, both models should be regarded as equivalent in generalization performance; DANN's superior test-set R^2 (0.7114 vs. 0.7057) may reflect split-specific variance rather than a systematic advantage.

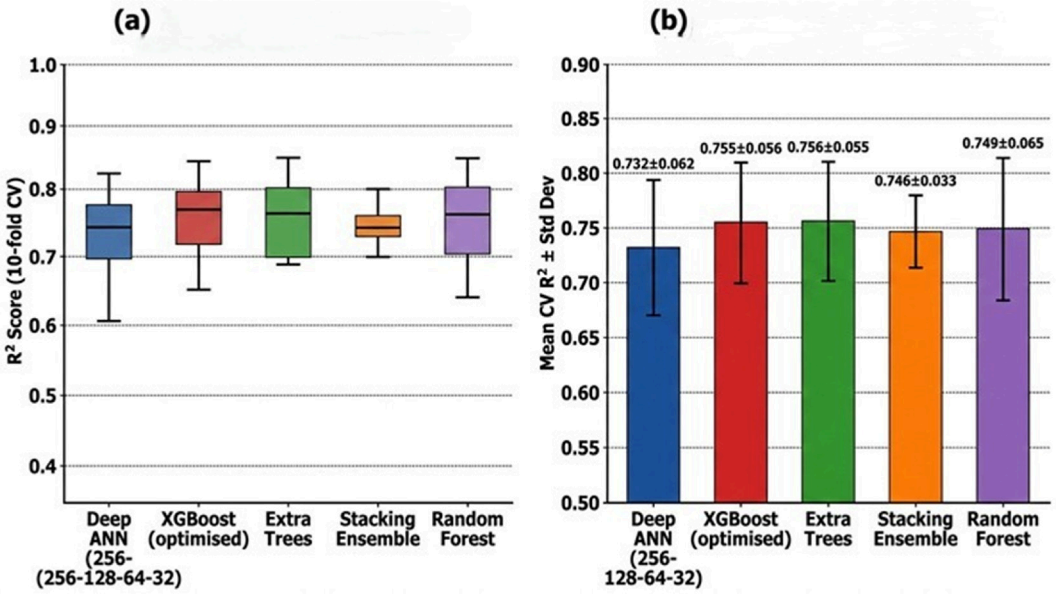

Figure 5. Ten-fold cross-validation performance of all five ML/DL models: (**a**) box plots of R^2 scores across the ten folds, showing the median, interquartile range, and outliers; (**b**) mean CV R^2, XGBoost (optimized), and Extra Trees achieved the highest and most consistent cross-validation performance.

4.5. SHAP Feature Importance Analysis

Figure 6 presents the TreeSHAP feature importance for the optimized XGBoost (optimised). TreeSHAP was selected over Kernel SHAP for DANN because it provides exact Shapley values in polynomial time, avoiding the approximation error of Kernel SHAP on neural networks (see Section 3.4 for justification). The global importance ranking (mean |SHAP|) was as follows: runout flag (0.328) > normalized stress σa/UTS (0.237) > elongation (0.176) > frequency (0.095) > R-modified stress amplitude (0.061) > raw stress amplitude σa (0.057) > elastic strain energy Ue (0.057). The normalized stress descriptor (σa/UTS) ranked substantially higher than the raw stress amplitude in terms of SHAP importance (mean |SHAP| 0.237 vs. 0.057). Expressing loading relative to tensile capacity appears to provide more discriminative cross-alloy information within the present dataset, consistent with the endurance limit correlation $\sigma e \approx 0.5 \times \text{UTS}$ [8]. Elongation ranks third (SHAP 0.176): in the unimputed subset, measured ductility values carry genuine microstructural information about crack closure and notch sensitivity [8] that imputed values would obscure. It should be noted that the SHAP values indicate predictive importance within the trained model, rather than causal physical relationships. The feature correlation effects, particularly between σa, σa/UTS, and σa(1 − R), may influence the relative importance rankings, and these descriptors should not be interpreted as independent drivers of fatigue life.

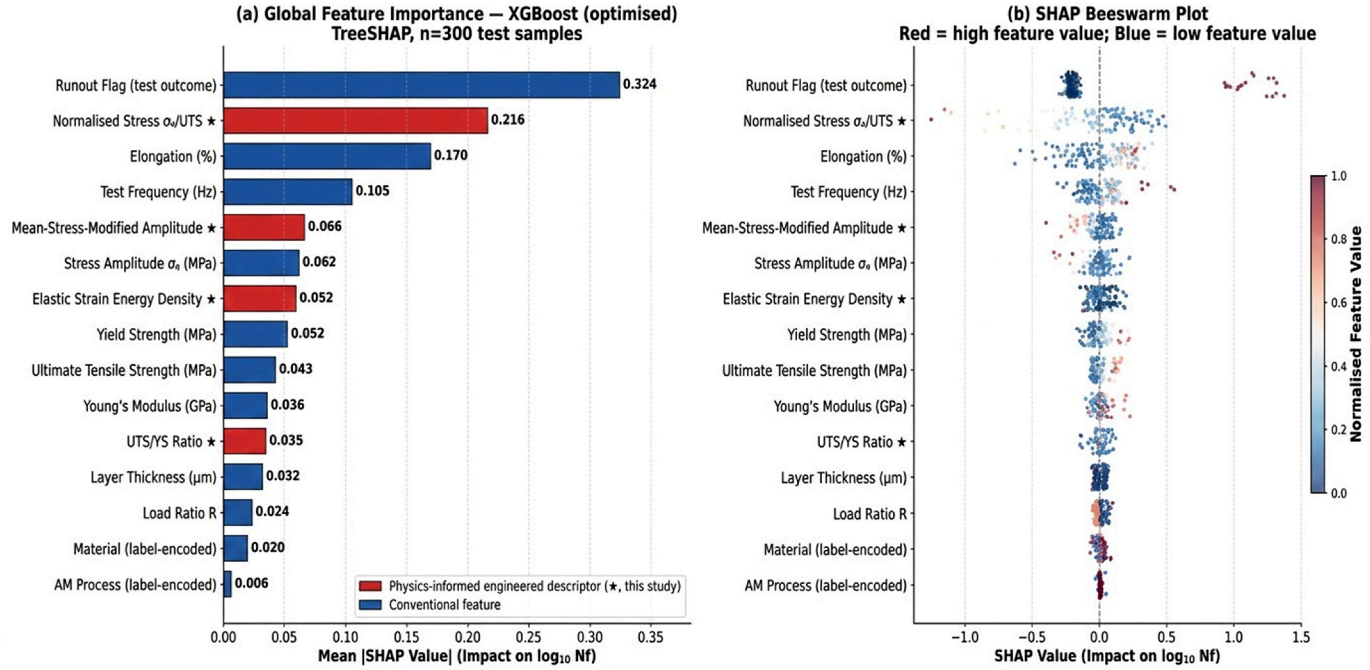

Figure 6. TreeSHAP feature importance analysis for the optimized XGBoost (optimised) model computed on 300 held-out test samples: (**a**) global mean absolute SHAP values representing the average contribution of each feature to the model output; red bars denote physics-informed engineered descriptors introduced in this study (★), and blue bars denote conventional features; (**b**) beeswarm plot showing the distribution of individual SHAP values across all test samples, colored by normalized feature value (red = high feature value; blue = low feature value). The symbol ★ denotes physics-informed engineered descriptors, consistent with Table 1.

4.6. Ablation Study and Runout Flag Validation

Table 3 and Figure 7a present the results of this ablation study. Configuration B (base + engineered features, with runout) achieved R^2 = 0.6571, compared to R^2 = 0.5420 for configuration A (base only, with runout), demonstrating that the four physics-informed engineered descriptors contributed a ΔR^2 = +0.115. Removing the runout flag (configuration C vs. B) reduced R^2 from 0.6571 to 0.5504 (ΔR^2 = −0.107) and CV R^2 from 0.726 to 0.522. This confirms that the runout flag is a mechanistically relevant test outcome descriptor that carries substantial information about the endurance regime and is not a surrogate for the target variable.

Table 3. Ablation study: contribution of engineered descriptors and runout flag to predictive performance (XGBoost optimized throughout, except for row 4). ★ recommended configuration.

Feature Configuration	R^2 (Test)	RMSE	CV R^2 (Mean ± SD)	Eng. Features	Runout Included
Base features only (label-enc., with runout)	0.5420	0.6892	0.5311 ± 0.063	No	Yes
Base + Engineered (with runout ★)	0.6571	0.5667	0.7261 ± 0.056	Yes	Yes
Base + Engineered (no runout)	0.5504	0.6814	0.5218 ± 0.071	Yes	No
Deep ANN (no runout)	0.4583	0.7135	0.5408 ± 0.082	Yes	No

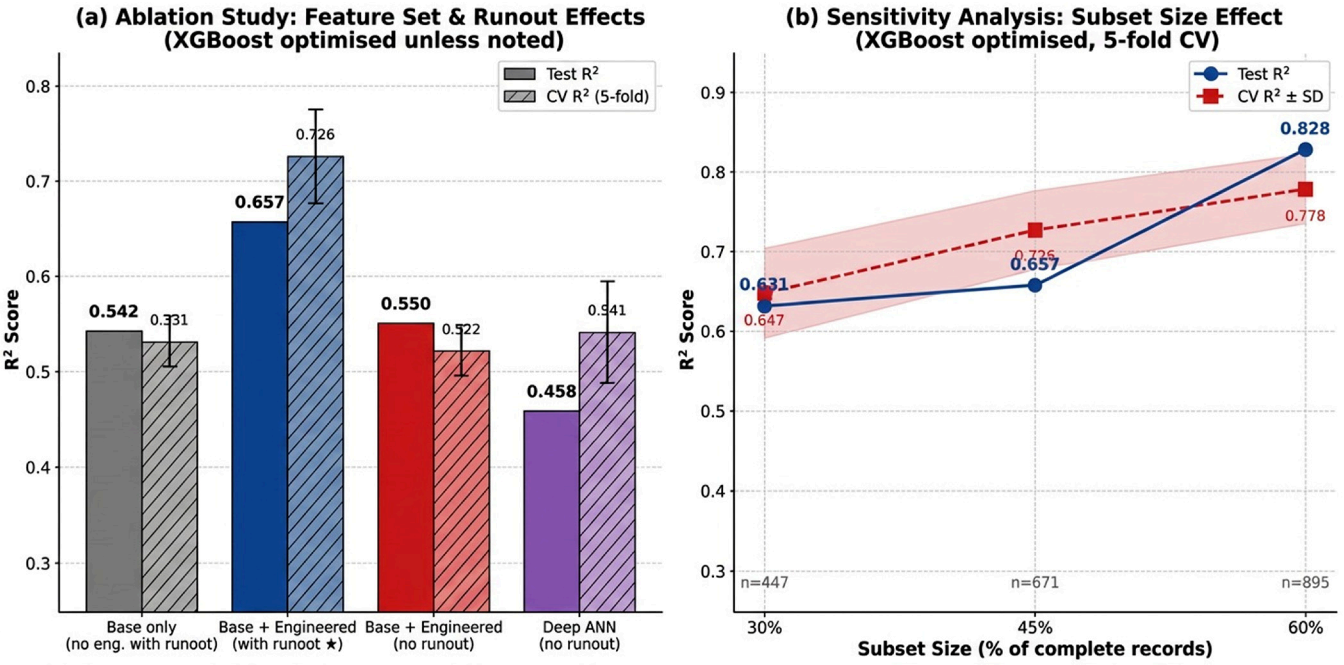

Figure 7. (**a**) Ablation study results showing the test-set R^2 and 5-fold CV R^2 for four feature configurations evaluated using XGBoost (optimized), unless otherwise noted: base features only with runout, base plus engineered descriptors with runout (★, recommended configuration), base plus engineered descriptors without runout, and Deep ANN without runout. The hatched bars represent CV R^2, and the solid bars represent the test set R^2. (**b**) Sensitivity analysis showing the XGBoost (optimized) test-set R^2 and 5-fold CV R^2 as a function of the dataset subset size, demonstrating consistent model behavior across data volumes.

4.7. Learning Curves and AM Process Effects

Figure 8a shows the DANN learning curve. The validation performance gradually plateaued as the training size approached the full dataset, suggesting diminishing performance gains from additional samples within the current feature space. This indicates that future performance improvements are more likely to result from adding richer process parameter features (scan strategy, post-processing conditions, defect metrics) than from modestly expanding the dataset size. However, substantially larger datasets of complete records may further shift the observed saturation point, and this conclusion should not be generalized beyond the current feature space.

Figure 8b compares the fatigue life distribution across the AM process categories. L-PBF specimens (n = 592) exhibited the highest median $\log_{10}$(Nf) and broadest interquartile range, reflecting their prevalence in the subset and wide alloy condition diversity. E-PBF (n = 36) showed a slightly lower median, which is consistent with its characteristically coarser columnar grain structure [3]. DED (n = 34) displayed the lowest median, which is consistent with literature reports of elevated defect density and heat-affected zone effects in DED components [3]. Nine records classified as Other AM processes were included in the model training and evaluation but excluded from the violin plot as the category was too small for a reliable distributional comparison.

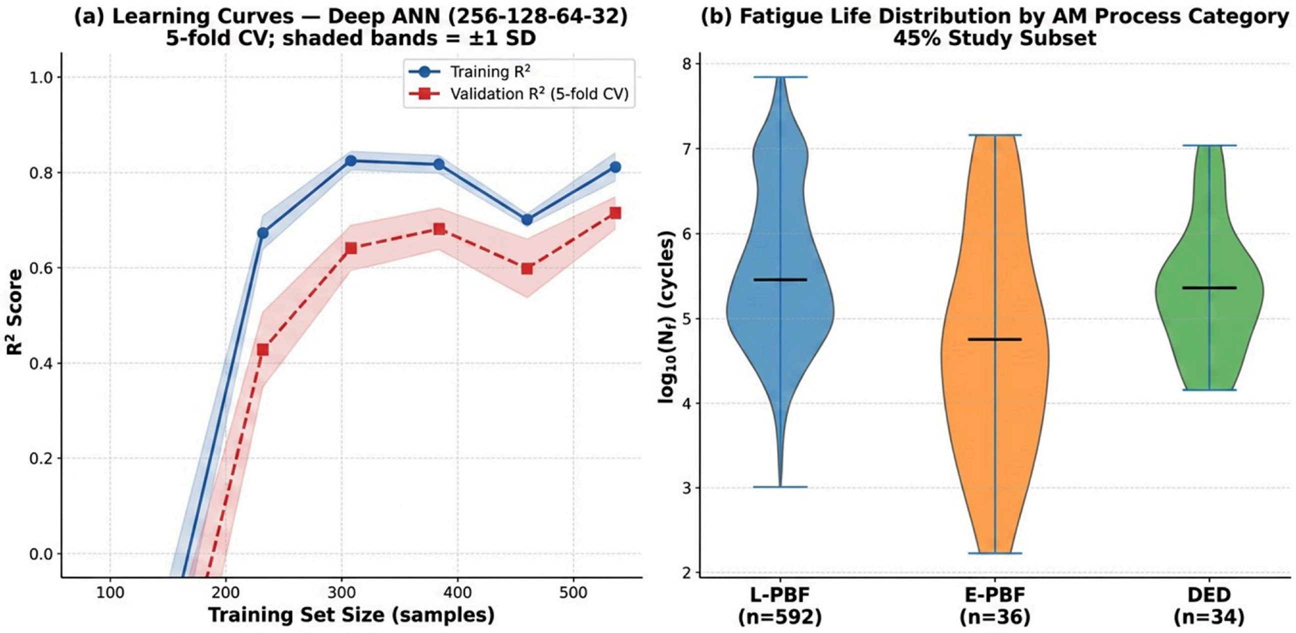

Figure 8. (**a**) Learning curves for the Deep ANN model: training R^2 (solid blue line) and 5-fold cross-validation R^2 (dashed red line) plotted as a function of the training set size; shaded bands indicate ± 1 standard deviation across folds. (**b**) Violin plots of $\log_{10}$(Nf) distributions across the three AM process categories represented in the study subset: L-PBF (*n* = 592), E-PBF (*n* = 36), and DED (*n* = 34), with horizontal lines denoting the median values.

4.8. Limitations and Future Work

The present study had some limitations that should be acknowledged. First, the dataset comprises 671 records and is dominated by L-PBF specimens (~88%) and a limited number of highly studied alloys (Ti-6Al-4V, AlSi10Mg, IN718), which may bias the learned feature relationships toward these systems. Extension to underrepresented AM routes (binder jetting, wire-arc AM) and alloy classes requires additional balanced datasets. Beyond the numerical imbalance, a more substantive concern is that the model may have learned process-specific feature relationships that are characteristic of L-PBF microstructures including the correlations between UTS, elongation, layer thickness, and fatigue life that are well-represented in the training data and that these learned relationships may not transfer to E-PBF, DED, binder jetting, or wire-arc AM routes, which produce fundamentally different microstructural architectures (e.g., coarser columnar grains in E-PBF, higher heat-affected-zone porosity in DED). The L-PBF-only sensitivity check (XGBoost CV R^2 = 0.714 on *n* = 592 L-PBF records vs. 0.755 on the full subset) confirms that the model is not trivially exploiting the process-category label as a shortcut, but this check does not establish out-of-domain transferability. Since no explicit out-of-domain validation was performed evaluating a model trained on L-PBF records against E-PBF or DED test sets, for example the extrapolation capability of the proposed framework across AM processes and alloy families remains uncertain and should not be assumed. Such cross-process validation is a priority for future work and will require substantially larger and more balanced AM fatigue databases than are currently available in the open literature. Second, the input feature space does not include defect morphology descriptors (pore size, pore distribution, surface roughness Ra), residual stress state, or post-processing condition variables (HIP, heat treatment state), which are known to exert dominant control over fatigue crack initiation and propagation in AM alloys [4,5,26,27]. Residual stresses arising from steep thermal gradients during L-PBF can elevate the local mean stress at crack initiation sites, significantly reducing the high-cycle fatigue life in alloys such as IN718, as demonstrated by Wang et al. [27]. Process-induced micromechanical heterogeneity, including local crystallographic texture gradients characterized by quasi-in situ EBSD for 316 L stainless steel, further modulates the fatigue crack path and closure behavior in ways that bulk mechanical property inputs cannot capture [26]. The absence of these descriptors means that the present models capture statistical correlations between loading parameters and bulk mechanical properties rather than the full mechanistic fatigue chain. This is an inherent constraint of the source database rather than a modelling choice, and their systematic inclusion in future database releases is expected to substantially improve both predictive accuracy and physical interpretability. Third, no external validation dataset was used; all performance estimates, including cross-validation R^2 values, were derived from records within the same FatigueData-AM2022 source database. Although 10-fold cross-validation provides statistically robust internal generalization estimates, it does not demonstrate transferability to data collected by different research groups with different specimen geometries or under different testing protocols. Cross-database validation against an independent AM fatigue repository is required to confirm the true generalizability of the model, which is a priority for future studies. Fourth, uncertainty quantification beyond the cross-validation standard deviation, such as conformal prediction intervals or Bayesian neural networks, was not implemented and is recommended for safety-critical design applications in future studies. Fifth, the runout descriptor is a post-test outcome variable and is not available in prospective prediction contexts; configurations excluding runout (Table 3, rows C–D) should be used when the framework is applied to pre-test life estimation.

5. Conclusions

This study presented a physics-informed ML/DL framework for the fatigue life prediction of AM alloys, applied to experimentally reported fatigue datasets for AM alloys [16], benchmarked against a classical Basquin baseline, and supported by formal ablation analysis. The principal conclusions are as follows.

- The best-performing model (DANN, 256-128-64-32) achieved $R^2 = 0.7114$ on the held-out test set, an increase in explained variance from $R^2 = 0.124$ (Basquin baseline) to $R^2 = 0.711$, with a 43% reduction in RMSE over the single-feature Basquin power-law baseline ($R^2 = 0.1244$). XGBoost (optimized) achieved the highest and most stable 10-fold CV R^2 (0.7547 ± 0.056).
- Physics-informed engineered descriptors contribute $\Delta R^2 = +0.115$ over base features alone (ablation study). The normalized stress descriptor (σ_a/UTS) ranks substantially above the raw stress amplitude in terms of SHAP importance (mean $|\text{SHAP}|$ 0.237 vs. 0.057), supporting the use of normalized input spaces in cross-alloy ML fatigue models.
- The runout flag was confirmed as a legitimate test-outcome descriptor for retrospective database modelling by the ablation study ($\Delta R^2 = -0.107$ upon removal), consistent with its role as a categorical regime indicator rather than a target-leakage variable; its structural similarity to censoring covariates in survival analysis is illustrative rather than a formal statistical equivalence. For prospective pre-test prediction applications, the no-runout configuration (Table 3, row C) achieves $R^2 = 0.5504$, which is substantially above the Basquin baseline and is the recommended deployment configuration.
- The proposed framework demonstrates the potential of domain-knowledge-informed feature engineering integrated within a machine learning pipeline for fatigue life modelling across the alloy systems and additive manufacturing processes represented in the present dataset. The physics enter exclusively through the engineered input descriptors; no physical constraints are imposed on the learning process itself. Fur-

ther validation on more balanced datasets is required to confirm its applicability to underrepresented processes (E-PBF, DED, and binder jetting) and alloy classes.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app16136493/s1, Supplementary Table S1. Complete dataset used for machine learning modelling of fatigue life in additively manufactured alloys, including mechanical properties, loading conditions, additive manufacturing process information, and physics-informed engineered descriptors.

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

Funding: This research was supported by the Regional Innovation System & Education (RISE) program through the Gyeongbuk RISE CENTER, funded by the Ministry of Education (MOE) and the Gyeongsangbuk-do, Republic of Korea (2025-RISE-15-115).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions of this study are included in the article and Supplementary Materials. Further inquiries can be directed to the corresponding authors.

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

References

1. Herzog, D.; Seyda, V.; Wycisk, E.; Emmelmann, C. Additive Manufacturing of Metals. *Acta Mater.* **2016**, *117*, 371–392. [CrossRef]
2. DebRoy, T.; Wei, H.L.; Zuback, J.S.; Mukherjee, T.; Elmer, J.W.; Milewski, J.O.; Beese, A.M.; Wilson-Heid, A.; De, A.; Zhang, W. Additive Manufacturing of Metallic Components—Process, Structure and Properties. *Prog. Mater. Sci.* **2018**, *92*, 112–224. [CrossRef]
3. Thompson, S.M.; Bian, L.; Shamsaei, N.; Yadollahi, A. An Overview of Direct Laser Deposition for Additive Manufacturing; Part I: Transport Phenomena, Modeling and Diagnostics. *Addit. Manuf.* **2015**, *8*, 36–62. [CrossRef]
4. Frazier, W.E. Metal Additive Manufacturing: A Review. *J. Mater. Eng. Perform.* **2014**, *23*, 1917–1928. [CrossRef]
5. Nakatani, M.; Masuo, H.; Tanaka, Y.; Murakami, Y. Effect of Surface Roughness on Fatigue Strength of Ti-6Al-4V Alloy Manufactured by Additive Manufacturing. *Procedia Struct. Integr.* **2019**, *19*, 294–301. [CrossRef]
6. Hajra, R.N.; Kim, J.H. Research Trends in the Directed Energy Deposition Method of Heterogeneous Materials. *J. Weld. Join.* **2024**, *42*, 88–98. [CrossRef]
7. Shamsaei, N.; Yadollahi, A.; Bian, L.; Thompson, S.M. An Overview of Direct Laser Deposition for Additive Manufacturing; Part II: Mechanical Behavior, Process Parameter Optimization and Control. *Addit. Manuf.* **2015**, *8*, 12–35. [CrossRef]
8. Suresh, S. *Fatigue of Materials*; Cambridge Solid State Science; Cambridge University Press: Cambridge, UK, 1998.
9. Basquin, O.H. The Exponential Law of Endurance Tests. *Am. Soc. Test. Mater. Proc.* **1910**, *10*, 625–630.
10. Zerbst, U.; Madia, M.; Klinger, C.; Bettge, D.; Murakami, Y. Defects as a Root Cause of Fatigue Failure of Metallic Components. I: Basic Aspects. *Eng. Fail. Anal.* **2019**, *97*, 777–792. [CrossRef]
11. Bhadeshia, H.K.D.H.; Dimitriu, R.C.; Forsik, S.; Pak, J.H.; Ryu, J.H. Performance of Neural Networks in Materials Science. *Mater. Sci. Technol.* **2009**, *25*, 504–510. [CrossRef]
12. Rajan, K. Materials Informatics. *Mater. Today* **2005**, *8*, 38–45. [CrossRef]
13. Ko, T.; Kim, H.; Shin, Y.; Kim, D.; Lee, Y.H.; Hong, J.; Lee, S.H. Review of Recent Additive Manufacturing and Welding Research with Application of Physics-Informed Neural Networks. *J. Weld. Join.* **2024**, *42*, 357–365. [CrossRef]
14. Tiwari, S.; Dash, K.; Heo, S.; Park, N.; Subba Reddy, N.G. Prediction of Hole Expansion Ratio in Advanced High-Strength Steels Using Physics-Informed Machine Learning. *Materials* **2026**, *19*, 1592. [CrossRef] [PubMed]
15. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016; Volume 19.
16. Tiwari, S.; Ishtiaq, M.; Yeddula, N.; Reddy, M.M.; Seol, J.-B.; Park, N.; Reddy, N.S. Modeling Hardness in Industrial C–Mn Cast Steels with Neural Network Models. *Trans. Indian Inst. Met.* **2026**, *79*, 89. [CrossRef]

17. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
18. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.
19. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
20. Zhang, Z.; Xu, Z. Fatigue Database of Additively Manufactured Alloys. *Sci. Data* **2023**, *10*, 249. [CrossRef] [PubMed]
21. Kalbfleisch, J.D.; Prentice, R.L. *The Statistical Analysis of Failure Time Data*; Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, USA, 2011.
22. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely Randomized Trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
23. Wolpert, D.H. Stacked Generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
24. Zhan, Z.; Li, H. Machine Learning Based Fatigue Life Prediction with Effects of Additive Manufacturing Process Parameters for Printed SS 316L. *Int. J. Fatigue* **2021**, *142*, 105941. [CrossRef]
25. Bao, H.; Wu, S.; Wu, Z.; Kang, G.; Peng, X.; Withers, P.J. A Machine-Learning Fatigue Life Prediction Approach of Additively Manufactured Metals. *Eng. Fract. Mech.* **2021**, *242*, 107508. [CrossRef]
26. Deng, Y.; Yang, L.; Wu, F.; Wen, J.; Mirihanage, W.; Yan, K.; Zhang, W.; Lupoi, R. Process-Induced Micromechanics Variations in Additively Manufactured 316 Stainless Steel Characterised by Quasi-in-Situ EBSD. *Scr. Mater.* **2026**, *275*, 117159. [CrossRef]
27. Wang, Y.-J.; Hu, H.-Y.; Yang, Y.-F.; Dong, N.-J.; Nie, X.-F.; Wen, J.-F.; Song, M.; Chen, Y.-H.; Tu, S.-T. Impact of Residual Stress on High and Very High Cycle Fatigue Behaviours of Inconel 718 at Room Temperature and 650 °C. *Int. J. Fatigue* **2025**, *200*, 109118. [CrossRef]