metals

MDPI

Article

Comparison of Performances of Machine Learning and Deep Learning Models for Prediction of Creep Rupture Life

Muhammad Bilal Jan, Zengchao Wu and Mengyu Chai *

School of Chemical Engineering and Technology, Xi'an Jiaotong University, Xi'an 710049, China; mbilaljan@stu.xjtu.edu.cn (M.B.J.); wu432812@stu.xjtu.edu.cn (Z.W.)
* Correspondence: chaimy@xjtu.edu.cn

Abstract

Accurate prediction of creep rupture life is essential for ensuring the long-term reliability of high-temperature components in power generation and petrochemical industries. Selecting appropriate data-driven models for limited and heterogeneous creep datasets remains a critical challenge, as conventional accuracy-based comparisons do not fully capture model behavior under varying service conditions. This study presents a unified evaluation framework for systematically comparing multiple machine learning and deep learning models for creep rupture life prediction of 2.25Cr–1Mo steel. The framework integrates predictive accuracy, prediction reliability, regime-specific error analysis, and computational efficiency, enabling a comprehensive assessment beyond global error metrics. The input feature space is reduced from seventeen to eight physically meaningful variables without loss of predictive performance. To further assess model robustness, prediction errors are analyzed across four distinct rupture life regimes, revealing significant variations in model behavior that are not reflected in aggregate metrics. Results indicate that support vector regression (SVR) provides the most consistent overall performance across all regimes and offers a strong balance between accuracy and computational efficiency. Among deep learning models, a Bayesian neural network (BNN) achieves competitive predictive performance while additionally enabling uncertainty estimation. These findings demonstrate that, for small tabular creep datasets, appropriately regularized models outperform complex neural network architectures, highlighting the importance of matching model complexity to dataset characteristics. This study is limited to a single steel grade, moderate dataset size, and extrapolation beyond trained stress and temperature ranges, which are key directions for future work.

Keywords: creep rupture life prediction; machine learning; deep learning; 2.25Cr1Mo steel; high-temperature materials; computational efficiency; feature selection; model comparison

Check for updates

Academic Editor: Alberto Campagnolo

Received: 18 June 2026
Revised: 8 July 2026
Accepted: 11 July 2026
Published: 14 July 2026

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

1. Introduction

Power generation and petrochemical industries are facing creep damage in metallic materials as they are increasing the working temperature and pressure to improve power generation, reduce pollutant emissions, and enhance thermal efficiency [1–4]. Creep damage reduces the mechanical properties of metallic materials, which increases the risk of failure in critical machinery parts [5,6]. To prevent unwanted downtime, avoid economic losses, and ensure structural integrity in these sectors, accurate creep rupture life prediction of industrial machinery is crucial [7–9]. Traditionally, creep rupture life has been predicted using conventional models such as the Larson–Miller parameter (LMP) [10], θ

parameter method [11], Norton equation [12], Monkman–Grant method [13], and Wilshire equation [14]. However, these empirical models are based on curve-fitting and overly simplified assumptions, which limit their ability to use diverse input features and capture complex nonlinear relationships, resulting in poor generalizability and limited predictive capability [15–17]. To overcome these limitations of traditional models, researchers are increasingly employing advanced data-driven approaches.

Machine learning (ML) models and deep learning (DL) architectures are widely used in the area of creep rupture life prediction. From 1997 to 2016, publications were sporadic [18]. However, after 2017, there was a sharp increase in studies applying these models for creep-life prediction, as they can capture complex nonlinear interactions and handle multiple features effectively [18]. This trend reflects the growing research interest in data-driven approaches in this field. For instance, Z. Yang et al. [19] predicted the creep life of 10CrMo910 steel main steam pipelines through ML models such as back propagation (BP), support vector regression (SVR), random forest (RF), and gradient boosting regression (GBR). The particle swarm optimization–support vector regression (PSO-SVR) outperformed other models by capturing the nonlinear relationship between carbide evolution and creep rupture life. T.-X. Yang et al. [20] applied linear regression (LR), support vector regression (SVR), multi-layer perceptron (MLP), kernel ridge regression (KRR), random forest regression (RFR), and eXtreme Gradient Boosting (XGBoost) to predict the creep rupture life of oxide dispersion-strengthened (ODS) steels based on chemical composition, heat treatment parameters, and test conditions. The prediction accuracy of the XGBoost model is better than that of LR, SVR, MLP, KRR, and RFR models. Plus, the calculated values align with the experimental values for the collected alloys. Wang et al. [21] applied ML models such as LR, SVR, RF and GBR to predict the irradiation creep coefficient in nuclear graphite and investigate the relative importance of material properties and environmental factors. Among these models, GBR performed best, followed by RF, which also provided good results. Song et al. [22] predicted the long-term creep life of P91 steel using domain knowledge and a back-propagation artificial neural network (BP-ANN). Their hybrid model, BP-ANN with domain knowledge, outperformed traditional models such as Larson–Miller (LM) and θ projection. Liang et al. [23] predicted the long-term creep rupture life by comparing the traditional method time–temperature parameter (TTP), specifically the Larson–Miller (LM), Manson–Harferd (MH), and Orr–Sherby–Dorn (OSD) method, with the ANN machine learning technique. In creep life prediction, ANN outperformed traditional TTP methods. While these studies demonstrated the potential of ML models, recent research has increasingly explored DL approaches because they can automatically extract hierarchical features, handle larger unified datasets, and model more complex nonlinear relationships, offering advantages over conventional ML methods [24–26]. For instance, Yan et al. [27] developed a method that combines the affinity propagation (AP), generative adversarial networks (GANs), and deep neural networks (DNNs) that predicted the creep fracture life for 7050 aluminum alloy better than ML models. X.C. Zhang et al. [28] combined creep rupture life, fatigue, and creep-fatigue datasets of an austenitic stainless steel and applied conventional ML techniques and deep neural networks (DNNs) to predict lifetime and evaluate model accuracy. DNNs offered higher accuracy than traditional ML techniques.

Despite an excellent contribution to the field of ML/DL-based creep rupture life prediction, limited attention has been given to the model complexity, computational efficiency, parameter-to-data ratio and training and inference time, which are crucial for selecting the most appropriate models that avoid overfitting and provide accurate predictions with less computation with real-life applicability. ML models and DL architectures are two different approaches. Selecting the best model for a dataset needs a comprehensive comparison

based on prediction accuracy, prediction reliability and computational efficiency. Simply selecting an ML model or DL architecture with the highest testing performance does not necessarily guarantee generalizability [29]. High-bias models and simpler models applied to complex data tend to underfit, resulting in poor performance. Meanwhile, high-variance models and complex models with limited data, compared to model complexity, overfit and fail to generalize [30]. This fundamental principle in ML, the "bias–variance" tradeoff, helps when selecting the appropriate model that does not overfit and gives accurate results with less computation. In ML and DL-based studies, the data-to-parameter ratio is one of the most important factors when selecting a model that does not overfit, as most ML studies use very high model parameters compared to available training data in ML-based creep life prediction [18]. Another perspective is that there is a general trend of DL outperforming ML due to its complexity and advancement. However, when it comes to tabular data and small datasets, ML outperforms DL architectures. Moreover, possible tradeoffs must be considered among performance, computational cost, and hyperparameter optimization and model training time, which are important in real-world applications [31]. A further limitation involves error analysis; most of the studies report MAE and RMSE for the entire creep life dataset regimes without examining underfitting or overfitting of models across different data regimes [27,32–34]. This aggregate reporting masks model behavior variations across different operating conditions. To comprehensively assess model behavior—where it performs well and where it performs poorly—creep life needs to be divided into distinct ranges for the purpose of finding model behavior in each rupture time regime.

2.25Cr1Mo steel, a type of low-alloy Cr-Mo heat-resistant steel category, possesses outstanding mechanical properties, superior high-temperature resistance, and weldability. Creep resistance in 2.25Cr1Mo steel is governed by two primary metallurgical mechanisms. First, solid solution strengthening: Cr and Mo atoms substitute into the ferritic matrix, impeding dislocation glide at elevated temperatures. Second, precipitation strengthening: carbides, principally M23C6 and M7C3, precipitate along grain boundaries and within the matrix during service, pinning dislocations and retarding grain boundary sliding [35–37]. The 18 input features in this study were selected to capture both mechanisms. The chemical composition variables (Cr, Mo, Si, C, N, Ni, etc.) directly control solid solution and carbide-forming capacity, while the heat treatment parameters (normalizing and tempering temperatures and times) determine the initial carbide morphology and distribution prior to creep testing. The test conditions (stress and temperature) define the thermomechanical driving force for creep deformation.

Due to these characteristics, 2.25Cr1Mo is widely used to manufacture high-temperature equipment in the power generation and petrochemical industries. Therefore, developing a reliable framework for creep life prediction in low-alloy Cr-Mo steel is essential. The developed model must balance prediction accuracy, computational efficiency, and generalization capability to avoid overfitting. Furthermore, it should be deployable in practical applications, such as in situ creep monitoring systems. This study begins with the acquisition of creep rupture life data from NIMS and published literature. The main objective of this study is to compare various ML models and DL architectures for selecting an appropriate model predicting the creep rupture life of 2.25Cr1Mo steel. Rather than relying solely on prediction accuracy, a unified evaluation framework is employed to systematically assess prediction reliability, range-specific error behavior across different creep life regimes, and model complexity and computational efficiency, including parameter efficiency metric, training time, inference speed, and parameter-to-data ratio. These aspects are essential for avoiding overfitting and underfitting and for identifying appropriate models for a given dataset.

This study is organized as follows. Section 2, materials and methods, focuses on data acquisition and construction, data visualization, data preprocessing, feature screening, model algorithms and architectures with hyperparameter optimization, evaluation methods, and the workflow of this paper. Section 3 presents the results and discussion of feature selection, predictive performances, scatter band analyses, error distribution analysis across four creep life ranges and computational efficiency, including training time, inference speed, and parameters of ML and DL models. The Section 4 concludes this comparative analysis and outlines the directions for future research.

2. Materials and Methods

2.1. Materials

The NIMS (Tsukuba, JAPAN) creep datasheet provides extensive details on the high-temperature creep properties of 2.25Cr1Mo steel, a material widely used in pressure vessels and piping systems for elevated temperature service [38,39]. This study focuses on high-temperature creep tests conducted on round-bar specimens, in accordance with ASTM E139-11 standard [40]. Accordingly, the NIMS creep database for 2.25Cr1Mo steel and literature sources was utilized to analyze its creep performance under high-temperature conditions.

2.2. Dataset Construction

The dataset comprises 464 samples from two sources: 385 samples from the NIMS creep database (Japan) [39] and 79 additional data points collected from published literature [36,41–50] covering diverse experimental conditions and compositions within the 2.25Cr1Mo steel grade. This dual-source construction improves compositional and condition diversity. There are 18 input features and one output feature: creep rupture life. The statistical details of all the features are listed in Table A1. Among the input features, 10 represent chemical composition, 4 relate to the heat treatment parameters and 2 describe physical conditions. The chemical composition includes carbon (C), silicon (Si), manganese (Mn), phosphorus (P), sulfur (S), nickel (Ni), chromium (Cr), molybdenum (Mo), copper (Cu), aluminum (Al), and nitrogen (N). The heat treatment parameters include normalizing temperature (N Temp), normalizing time (N Time), tempering temperature (T Temp), and tempering time (Temp Time). The physical condition includes test stress and test temperature. The temperature range of 450–650 °C reflects both the available NIMS experimental data and the practical service conditions for which 2.25Cr1Mo steel is specified; higher-temperature service applications require different steel grades.

2.3. Data Visualization and Preprocessing

The input features reveal diverse levels of fluctuation in the dataset, as shown in Table A1. To make the input features on the same scale, the standardization technique was employed. The selection of this technique was the result of comparing the coefficient of determination (R^2) obtained after transforming the input data using two methods, Min–Max scaling (which rescales each feature to the [0,1] range) and standardization [51]. The ridge regression model was trained on datasets scaled by each method. The prediction results revealed that standardization-scaled data achieved a higher R^2, indicating better performance with the same model. The calculation of R^2 on both scaled datasets helped in the selection of a more effective transformation method for predicting creep rupture life. The comparison of both techniques is shown in Table 1. The mathematical expression for standardization is shown in Equation (1), where X represents the original feature value, μ is the mean of the feature, σ denotes the standard deviation of the feature, and X_{scaled} is the resulting standardized value.

$$X_{scaled} = \frac{X - \mu}{\sigma} \quad (1)$$

Table 1. Comparison of input features preprocessing methods.

Preprocessing Method	R^2 Score	Model	Performance Summary
Min-Max Scaling	0.817	Ridge Regression	Performs well but slightly less accurate
Standard Scaling	0.896	Ridge Regression	Performs better (Recommended)

The output feature creep rupture life is not normally distributed; its high standard deviation reflects a wide spread of values. The abnormal distribution is verified through both visual inspection (box plot) and statistical testing (D'Agostino K^2 test). The D'Agostino K^2 test evaluates whether a dataset or a feature departs from normality by jointly assessing skewness and kurtosis. The K^2 statistic is calculated as the sum of the squared standardized deviations of skewness and kurtosis from their expected values under a normal distribution, with larger values indicating greater deviation from normality. A *p*-value below the chosen significance level (commonly 0.05) leads to rejecting the null hypothesis, confirming that the data distribution differs significantly from a Gaussian distribution. In brief, a high K^2 statistic along with a *p*-value less than 0.05 indicates deviation from a normal distribution [52]. Accordingly, the D'Agostino K^2 test was applied on the original creep rupture data, which resulted in a higher statistical value (K^2 = 582.44) and an extremely small *p*-value (3.33×10^{-127}); this provides strong evidence against the null hypothesis, which assumes the data is normally distributed. To handle the outliers in the broad distribution of the creep rupture life values and to scale them lower, four transformation techniques were evaluated using the D'Agostino K^2 test. Square root transformation showed the worst result, the power transformation and the Box–Cox showed moderate results, while the log transformation yielded the most favorable results with a lower K^2 statistic (27.36) and higher *p*-value (1.14×10^{-6}). Although the *p*-value of the log transformation is still below the 0.05 threshold, it is higher than those from the other transformations and closer to 0.05, indicating less deviation. The four transformation methods with the corresponding K^2 statistic and *p*-values are presented in Table 2.

Table 2. Comparison of normality test results for various transformations of the target variable (creep rupture life). Classification is based on the K^2 statistic magnitude: lower K^2 indicates reduced deviation from normality. Log transformation yielded the lowest K^2 (27.36) and was selected as the preferred transformation. All transformations remain non-Gaussian ($p < 0.05$), but the log transformation is closest to normality among the four methods tested.

Transformation	K^2 Statistic	*p*-Value	Interpretation
Square Root	109.94	1.34×10^{-24}	Worst
Yeo–Johnson	43.69	3.24×10^{-10}	Moderate
Box–Cox	37.09	8.82×10^{-9}	Moderate
Log	27.36	1.14×10^{-6}	Better

The outliers are detected using the box plot and are satisfactorily handled by applying the logarithmic transformation. In addition, the original data distribution of the target variable, creep life, is illustrated in Figure 1a, and the changes in its overall pattern after logarithmic transformation are depicted in Figure 1b.

Following transformation, for ML models, the data has been split into the training and testing sets, 80% of the data were randomly selected for the model training using the train_test_split function from scikit-learn and 20% for the test set to evaluate the model's generalization performance. The 80/20 split was chosen following standard practice in ML studies with limited datasets [19–21,28,32], balancing sufficient training data with a statistically meaningful test set for generalization assessment. DL architectures are trained

on the same 80% on which the ML models were trained, however DL needs a validation dataset along with testing set, so the remaining 20% of the data was split into 10% testing and 10% validation for early stopping and learning rate scheduling. The division is based on the need to ensure enough data to train the model while preserving unseen data for generalization assessment. The major split of creep rupture life data into 80% and 20% is visualized in Figure 2, where stress and rupture life are plotted, and data points are color-coded based on temperature.

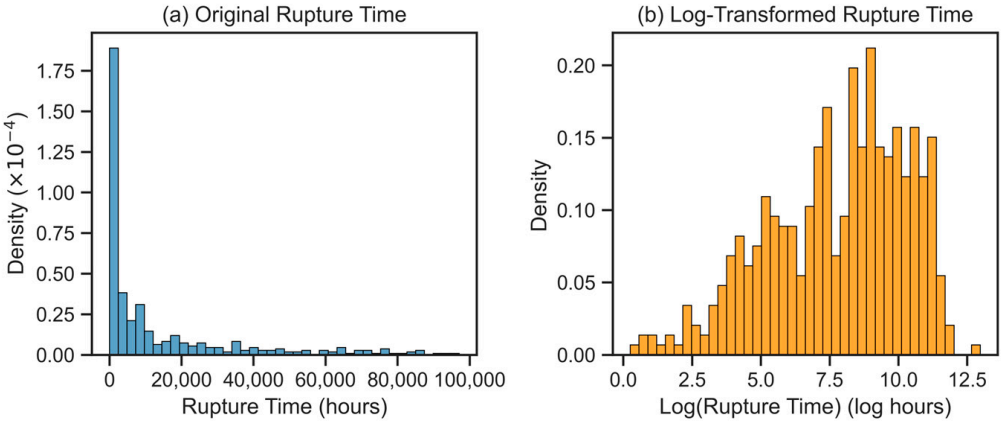


Figure 1. (**a**) Original creep rupture time distribution showing pronounced right skewness with extreme values beyond 400,000 h. Panel (**a**) is capped at 100,000 h; samples above this threshold (out of 464 total) are excluded from the visualization to preserve resolution in the dominant data range, though they remain in all model training and evaluation. (**b**) Log-transformed rupture time displaying a more symmetric distribution. Both panels use probability density on the *y*-axis (area under bars = 1), normalizing for the difference in bin widths between the linear and log scales. The distinct visual shapes reflect the genuine change in distributional symmetry after log transformation, which is the basis for selecting log transformation as the output preprocessing method.

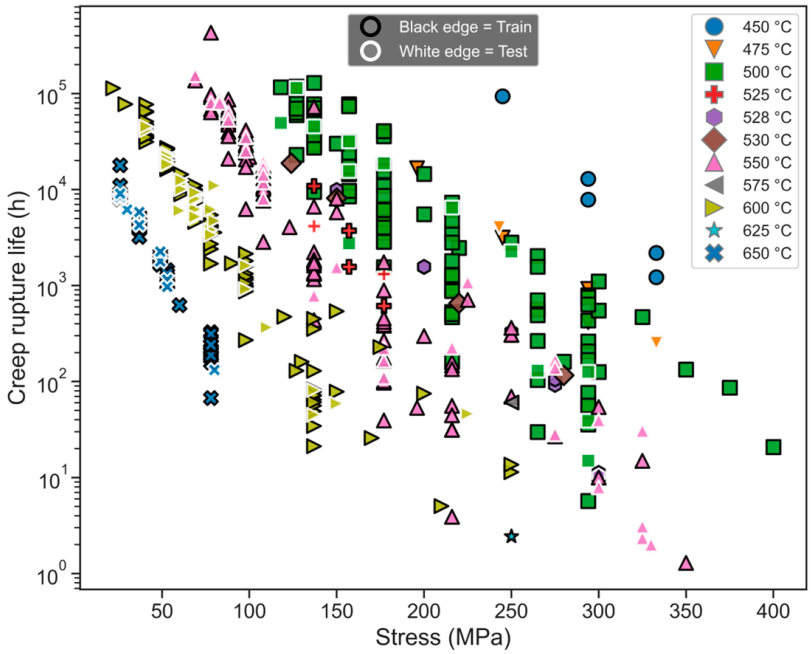


Figure 2. Visualization of creep rupture life dataset split into training and testing sets. Marker shape and color represent the test temperature, while marker edge color distinguishes the training and testing samples.

2.4. Input Feature Screening

Input feature screening is an important step in machine learning-based creep life prediction because it helps identify the features that have a meaningful relationship with creep rupture life while eliminating those that are irrelevant. In ML-based creep life studies, the use of an excessive number of input features can lead to unnecessarily complex models [53]. In our case, the features are screened using Pearson correlation and *p*-values. The Pearson correlation coefficient (r) quantifies the linear association between each input feature "X" and the target variable "y".

The *p*-value is used to screen features based on acceptance or rejection of the null hypothesis. The null hypotheses (H_0) states that there is no relationship between input and output features. A significance threshold of $\alpha = 0.05$ was used for rejecting the null hypothesis. If the *p*-value of a feature is greater than 0.05, it means there is no significant relationship between that feature and creep life. Features were retained only if their *p*-value was less than 0.05, indicating a statistically significant relationship with creep rupture life.

2.5. Machine Learning and Deep Learning Models

To obtain the best predictive, most efficient model, several ML algorithms and DL architectures were applied, including basic models, ensemble models and advanced deep learning models. To the best of our current knowledge, comparison of such models on the basis of feature screening, predictive accuracy, prediction reliability, error distribution characteristics in various regimes, and computational efficiency has not been thoroughly investigated on creep life prediction data. In this section, only the basic general principle of each model is mentioned. The architecture and hyperparameters information are stated in the subsequent section (Section 2.6).

2.5.1. Basic Machine Learning Models

In this study, five ML models, linear regression (LR), support vector regression (SVR), random forest (RF), gradient boosting regression (GBR), and Xtreme gradient boost (XGBoost) were used for training and testing the data. The main principle and respective advantages and limitations of the five ML models are presented in Table 3.

Table 3. Main principles of five machine learning models.

Model	Principle
LR	Linear regression fits a linear relationship between input variables and output [54].
SVR	Kernel-based regression; minimizes error within a margin (ε-insensitive) [55].
RF	Ensemble of trees with bootstrapping and random feature splits [56].
GBR	Boosting technique that combines weak learners to minimize residual errors [57].
XGBoost	Optimized boosting with regularization and parallel computation [58].

2.5.2. Deep Learning Models

In addition to ML models, four DL models, such as deep neural network (DNN), Bayesian neural network (BNN) with Monte Carlo dropout, convolutional neural network (CNN), and long short-term memory (LSTM), were used for training and testing the data. Comparing and evaluating the effectiveness of these DL architectures with ML models through the assessment of predictive accuracy, reliability, error analysis, and computational efficiency on a small dataset with a moderate number of features represents a key novelty

of this study. The main principles and respective advantages and limitations of the four DL architectures are presented in Table 4.

Table 4. Main principles of four deep learning architecture.

Model	Principle
DNN	Fully connected deep network learning nonlinear mappings [59].
BNN	Bayesian neural network introduces distributions over weights to quantify uncertainty in predictions [60].
1D CNN	Convolutional filters extract localized feature patterns from input data [61].
LSTM	Recurrent network capturing long-term sequential dependencies [62].

2.6. Hyperparameter Optimization

All models went through hyperparameter optimization using 5-fold cross-validation on the training set of 371 samples to identify parameters producing accurate predictions without overfitting. ML models used grid search across specified parameter ranges. Each combination was tested to find the best possible values. DL architecture used manual tuning with 4–6 network configurations per model. The selection of architecture options was based on preliminary experiments, with early stopping based on validation loss to prevent overfitting during training.

Cross-validation (CV) R^2 scores were the main metric for optimization, with the final hyperparameter configuration selected as the combination giving the best average cross-validation performance. The optimized models were then retrained on the full training set with the selected optimal parameters and evaluated on independent test sets. All stochastic operations used a fixed random seed (random_state = 42), including the train–test split, k-fold partitioning, and neural network initialization (numpy, Python (3.12), and TensorFlow seeds all set to 42), ensuring full reproducibility. Tables 5 and 6 show the final optimized hyperparameters for all ML models and the DL architecture, respectively.

Table 5. Hyperparameter optimization, search space and the corresponding best CV R^2 of the five ML models.

Model	Optimal Hyperparameters	Search Space Size	Best CV R^2
LR	None	N/A	N/A
SVR	C = 100, γ = 0.1, kernel = RBF	32 combinations	0.96
RF	n_estimators = 500, max_depth = None, min_samples_split = 2, min_samples_leaf = 1, max_features = None	324 combinations	0.88
GBR	n_estimators = 500, learning_rate = 0.1, max_depth = 3, min_samples_split = 5, min_samples_leaf = 2	108 combinations	0.94
XGBoost	n_estimators = 300, learning_rate = 0.1, max_depth = 3, subsample = 0.8, colsample_bytree = 1.0	162 combinations	0.93

Table 6. Architecture and training parameters with corresponding best CV R^2 of the four DL models.

Model	Architecture	Dropout	L2 Reg	Learning Rate	Batch Size	Best CV R^2
DNN	[32, 24, 16]	0.1	0.001	0.001	8	0.92
BNN	[24, 16, 8] filters = [64, 32]	[0.08, 0.04, 0.02]	N/A	0.001	8	0.94

Table 6. *Cont.*

Model	Architecture	Dropout	L2 Reg	Learning Rate	Batch Size	Best CV R^2
CNN	kernels = [3, 3], dense = [32, 16]	0.2	0.001	0.001	8	0.87
LSTM	lstm = [64, 32], dense = [32, 16], recurrent_dropout = 0.1	0.2	0.001	0.001	8	0.89

2.7. Model Evaluation Methods

The evaluation of the model's performance is primarily based on the deviation between actual and predicted values of creep rupture life, quantified using statistical metrics to ensure reliable model validation [29]. This study employed three evaluation methods [63], including the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE), as shown in Equations (2)–(4), where Yi, Ŷi, and Ȳi denote actual, predicted, and mean creep life values, respectively.

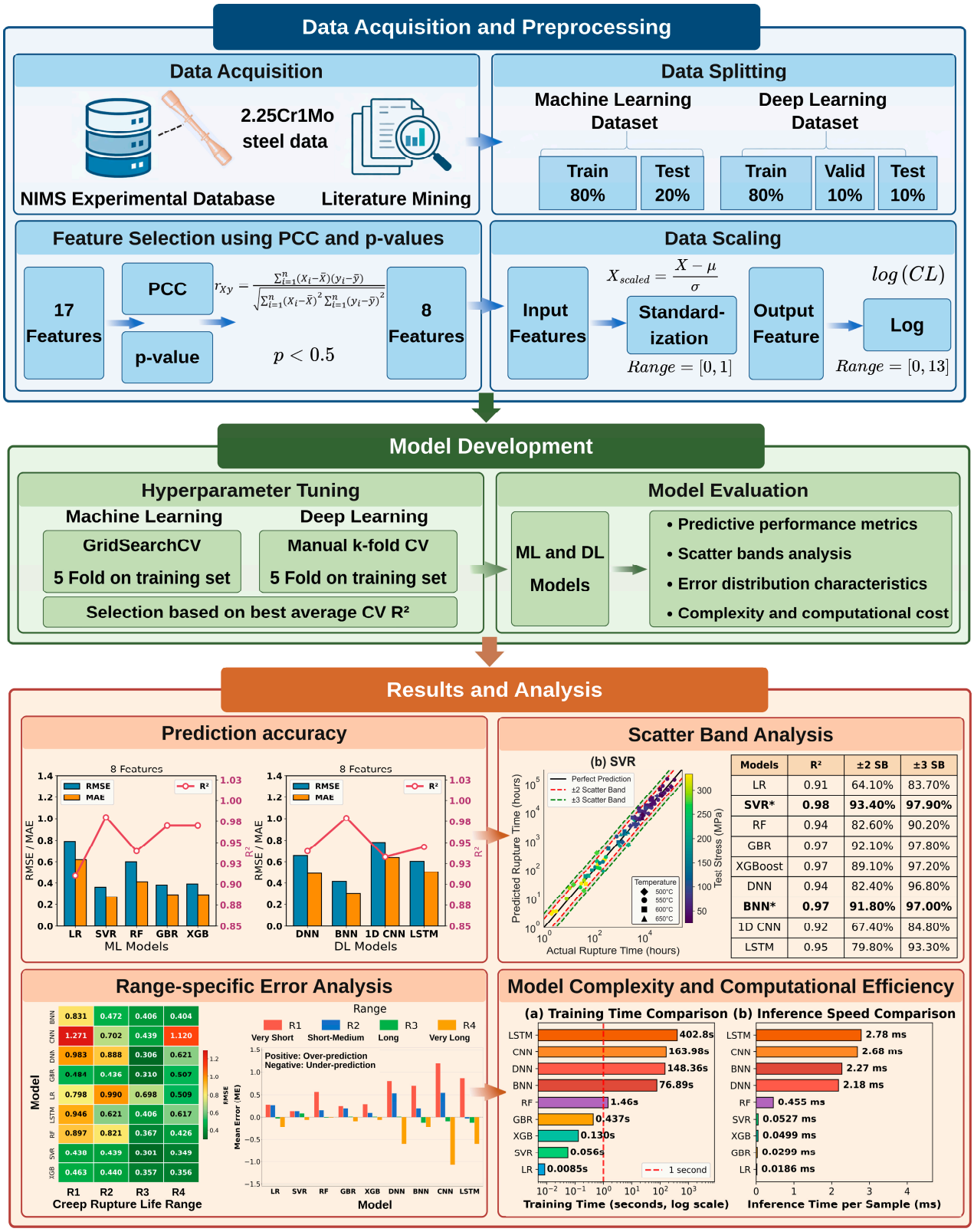
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Yi - \hat{Y}i)^2}{\sum_{i=1}^{n}(Yi - \overline{Y}i)^2} \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Yi - \hat{Y}i)^2} \quad (3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Yi - \hat{Y}i| \quad (4)$$

2.8. Overall Workflow

The workflow of this study, which employs ML models and DL architectures to predict the creep rupture life of 2.25Cr1Mo steel, is illustrated in Figure 3. "Data Acquisition and Preprocessing" section shows that the data were acquired from the NIMS creep database and published literature, and the dataset was established by removing invalid entries and imputing missing values. The data were then divided into training and testing sets; an 80/20 split was used for ML models, whereas DL models used an 80/10/10 split for training, validation, and testing. Feature selection was performed exclusively on the training data using the Pearson correlation coefficient (PCC) with a significance threshold of $p < 0.05$, which reduced the feature set from 17 to 8 variables. The selected features were scaled by applying standardization to the inputs and a logarithmic transformation to the output. The "Model Development" section describes the hyperparameter tuning for all nine models (five ML and four DL architectures), which was carried out using 5-fold cross-validation on the training set. For DL models, the validation set was additionally used for early stopping and learning-rate scheduling during final training. After training, all models were evaluated on the test set using metrics of predictive accuracy (R^2, RMSE, MAE), scatter band analysis, error distribution characteristics across four creep-life regimes, model complexity, and computational efficiency. The final "Results and Analysis" section presents the overall comparison in terms of each evaluation method; the best-performing ML and DL models were identified as the final optimized models.



Models	R²	±2 SB	±3 SB
LR	0.91	64.10%	83.70%
SVR*	**0.98**	**93.40%**	**97.90%**
RF	0.94	82.60%	90.20%
GBR	0.97	92.10%	97.80%
XGBoost	0.97	89.10%	97.20%
DNN	0.94	82.40%	96.80%
BNN*	**0.97**	**91.80%**	**97.00%**
1D CNN	0.92	67.40%	84.80%
LSTM	0.95	79.80%	93.30%

Figure 3. Complete workflow of the unified evaluation framework used in this study, illustrating data acquisition and preprocessing, model development with hyperparameter tuning and evaluation, and the final results and analysis stage comparing all models to identify the best-performing approach. Asterisks (*) indicate the highest percentage in each model category.

3. Results and Discussions

3.1. Feature Selection

3.1.1. Feature Screening

The features for modeling were screened using Pearson correlation and *p*-values. The features with very high significance, moderate significance and no significance are shown by the blue, green and red circles in the heat map (Figure 4). It is emphasized that the objective function in this study is creep rupture life, which is treated as a function of applied stress, test temperature, chemical composition, and heat treatment conditions, all of which are independent input variables. The feature-to-feature correlations visible in Figure 4 are presented as a multicollinearity diagnostic only and do not imply physical dependencies between input variables.

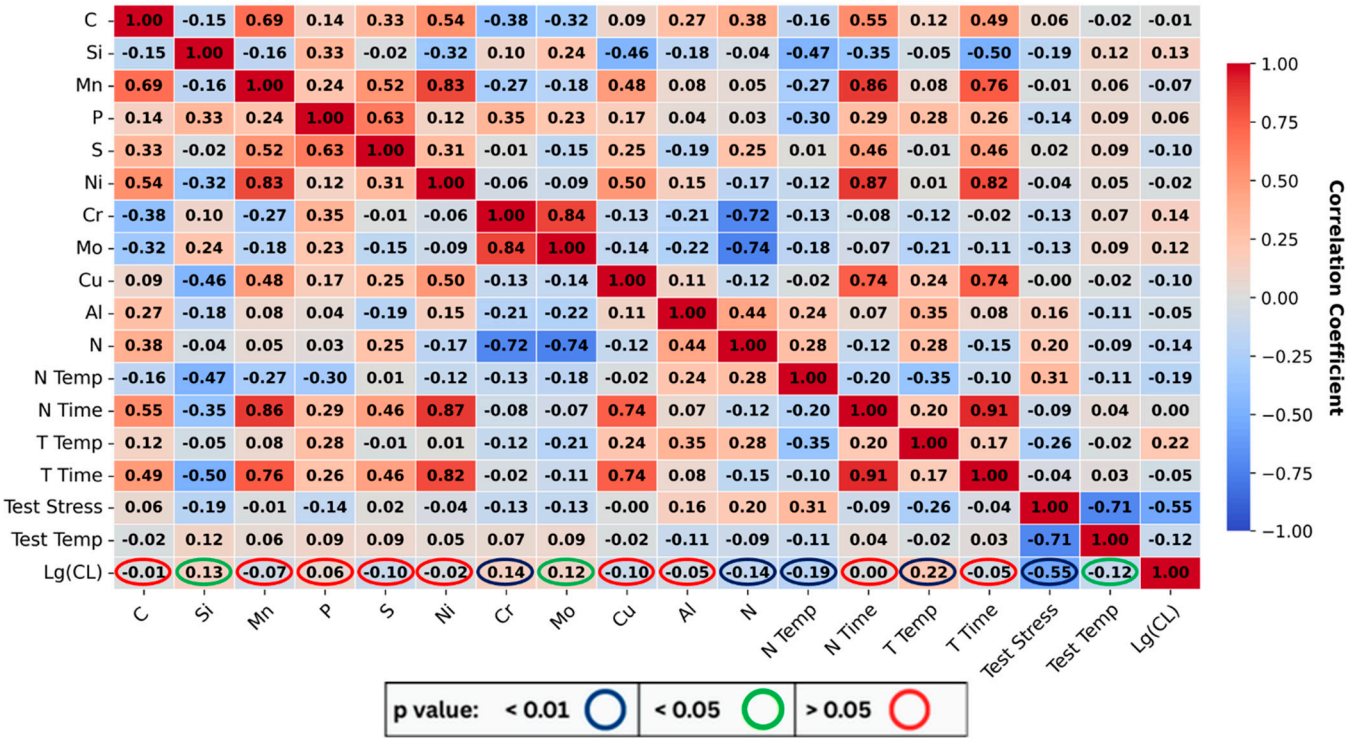


Figure 4. Heat map describing the relationship of features with output feature creep life; the Pearson coefficient, r, values show how much the feature is correlated with creep life, and the *p*-values at the bottom side of the figure are used to complement r values and screen the features based on accepting and rejecting of the null hypothesis.

If the *p*-value is much lower than the threshold 0.05 (less than 0.01), as indicated by the blue circle in Figure 4, these features are the most significant, indicating a strong relation with creep life. Features indicated by the green circle, with *p*-values below 0.05, are also considered important. All these *p*-values are complemented by the r values of the features. To know the extent of impact of these features on creep rupture life, Figure 5 shows the positive and negative influence of the features on creep life, as well as the *p*-value of each feature. The smaller the *p*-value, the greater the significance, with the green color features indicating higher importance and the red color features indicating lower importance. Table 7 provides a brief summary of the screened and excluded features. Based on the combined statistical screening, eight representative features were retained for subsequent modeling.

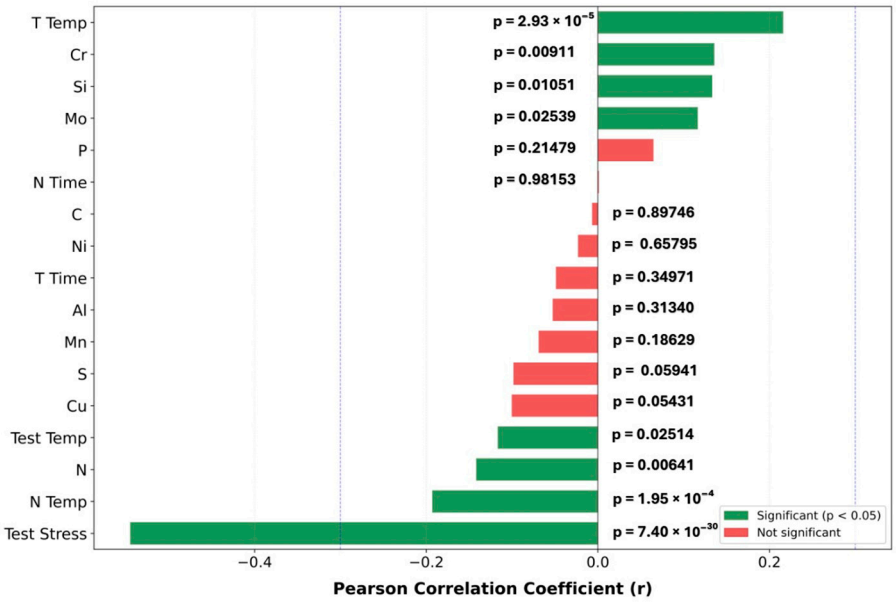


Figure 5. Pearson correlation coefficient (r) values and *p*-values of each feature.

Table 7. Final screened features and eliminated features after input feature screening via Pearson correlation and *p*-values. Significance classification is based on *p*-value thresholds: Very High = $p \leq 0.01$ (strong statistical evidence); High = $0.01 < p \leq 0.05$ (moderate statistical evidence); No significance = $p > 0.05$ (insufficient evidence to reject the null hypothesis of no association).

Class	Significance	Features	*p*-Value
Chemical composition	Very High	Cr, N	≤0.01
	High	Mo, Si	≤0.05
	No significance	C, Ni, Al, Mn, P, S, Cu	≥0.05
Heat treatment parameters	Very High	N Temp, T Temp	≤0.01
	High	N/A	≤0.05
	No significance	N Time, T Time	≥0.05
Physical conditions	Very High	Test stress	≤0.01
	High	Test temperature	≤0.05
	No significance	N/A	N/A

The correlations obtained through r and corresponding *p*-values are strongly consistent with known creep mechanisms in metallic materials. All included features are supported by findings in the literature. For instance, features such as test stress show a strong negative correlation, as higher stress accelerates creep damage by promoting faster dislocation movement [64]. Higher test temperatures accelerate creep deformation, thereby leading to a reduction in creep rupture life [65]. Alloying elements like Cr, Mo positively correlate due to their roles in solid solution and carbide strengthening [66]. Si also positively correlates. This element contributes to carbide strengthening, and it is reported that creep rates are found to decrease with increasing silicon content in alloys [67]. N correlates negatively, possibly due to inclusion or nitride formation. In a recent study, it was found that the increased use of N in alloys decreases the creep life up to 50% [68]. A higher normalizing temperature (N Temp) shows a negative correlation with creep life, possibly due to grain coarsening or over-tempering effects [69]. In contrast, higher tempering temperature (T Temp) correlates positively, likely due to improved carbide precipitation and microstructural stabilization [70]. Regarding excluded impurity elements, such as P, S, Cu, Ni, Al, and Mn, they showed *p*-values ≥ 0.05 in the Pearson correlation test, indicating no statistically significant relationship with log(creep life) in this dataset. For P specifically, while phosphorus embrittlement is a well-documented mechanism in ferritic

steels, the P content in this dataset varies only narrowly (0.001–0.03 wt%, std = 0.005 wt%), with most samples clustered near the lower end of the specification range. This limited variance likely makes P's contribution statistically indistinguishable from noise. The model does not explicitly capture P-induced embrittlement. It is also acknowledged that Pearson correlation detects only linear associations. Non-linear feature importance methods such as SHAP (SHapley Additive exPlanations) may capture additional interaction effects between features like Mo and temperature. However, two results validate the current selection: (1) the r values align with known metallurgical mechanisms (Section 3.1.1), and (2) 8-feature models match 17-feature models in predictive performance (Tables 8 and 9), confirming that discarded features add no predictive value. A SHAP-based analysis is recommended for future work.

Table 8. Performance comparison of machine learning models using 8 selected features versus all 17 original features. Changes in R^2, RMSE, and MAE on test set. Positive ΔR^2 indicates improvement with feature reduction; negative ΔRMSE/ΔMAE indicates improvement. All changes represent minimal variation, confirming effective feature selection. Status classification: $\Delta R^2 < 0.01$ and ΔRMSE < 0.03 = Negligible (less than 5% relative change in RMSE); ΔR^2 of 0.01–0.02 or ΔRMSE of 0.03–0.05 = Minor change. These thresholds reflect practical engineering significance rather than statistical significance.

Model	ΔR^2	ΔRMSE	ΔMAE	Status
LR	+0.003	−0.01	+0.01	Negligible
SVR	+0.005	−0.02	−0.02	Negligible
RF	−0.01	+0.03	+0.03	Minor change
GBR	+0.003	+0.01	−0.01	Negligible
XGBoost	+0.002	−0.02	+0.01	Negligible

Table 9. Performance comparison of DL architectures using 8 selected features versus all 17 original features. Changes in R^2, RMSE, and MAE on test set. Positive ΔR^2 indicates improvement with feature reduction; negative ΔRMSE/ΔMAE indicates improvement. All architectures maintained comparable performance, validating effective feature selection.

Model	ΔR^2	ΔRMSE	ΔMAE	Status
DNN	−0.016	−0.030	−0.005	Minor change
BNN	−0.002	−0.010	+0.001	Negligible
1D CNN	−0.006	−0.002	−0.008	Negligible
LSTM	−0.011	−0.003	+0.001	Minor change

3.1.2. Feature Performance Comparison

To evaluate the effectiveness of the feature selection, ML and DL models were trained and tested using the eight screened features and all the 17 features. Figure 6 demonstrates the ML models' performance on both feature sets, and Figure 7 demonstrates the DL architecture's performance on both sets. The predictive performance of all models trained with the 8 screened features remained consistent with those using all 17 features. The accuracy remained stable during the modeling on 8 selected features and all 17 features, which confirms the effectiveness of feature screening. This stability also indicates that the screened variables sufficiently capture the essential factors governing the creep life of 2.25Cr1Mo steel.

ML models' performance metrics remained stable in both feature sets, as shown in Table 8. R^2 values remained identical for four out of five models, with changes occurring in values only. However, RF showed a marginal decrease from 0.95 to 0.94, representing only a 1% change. The maximum RMSE deviation was only 0.03 in RF, which represents approximately 5% variation. SVR, the best-performing ML model, showed a negligible change in R^2 and RMSE in both cases. These minimal differences show that the 8-feature set has comparable predictive capability.

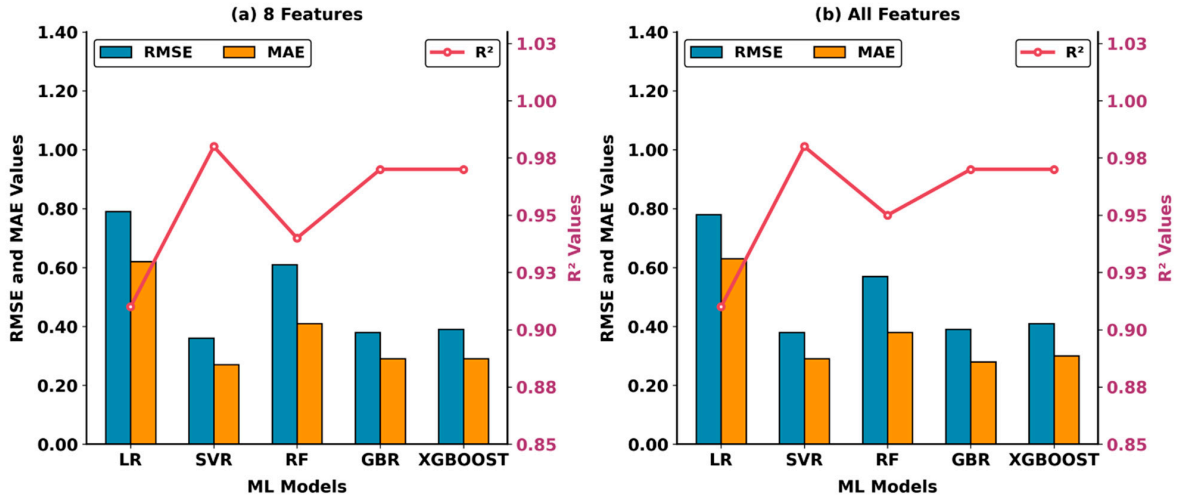


Figure 6. Effect of feature selection on ML model performance. Test set metrics (RMSE, MAE, R^2) for five ML models using (**a**) 8 features selected through Pearson correlation analysis ($p < 0.05$) and (**b**) all 17 original features. Blue bars: RMSE, orange bars: MAE (left axis); red line: R^2 values (right axis).

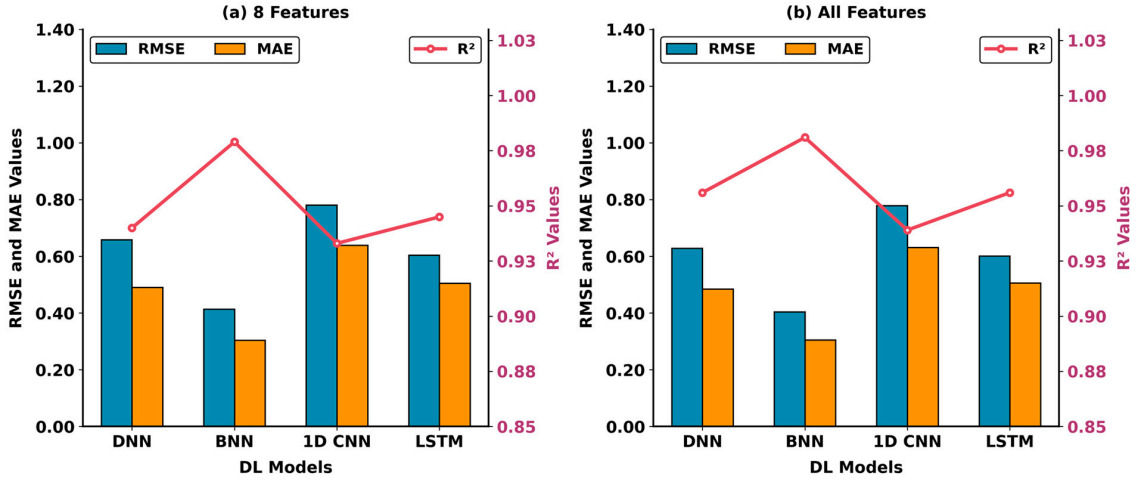


Figure 7. Effect of feature selection on DL architecture performance. Test set metrics (RMSE, MAE, R^2) for four DL architectures using (**a**) 8 features selected through Pearson correlation analysis ($p < 0.05$) and (**b**) all 17 original features. Blue bars: RMSE, orange bars: MAE (left axis); red line: R^2 values (right axis).

Similarly, as shown in Table 9, the results of DL architectures maintained consistent performance across both feature sets. All models achieved comparable R^2 values, showing only minor differences in the values. However, in the case of DNN, the maximum degradation of 0.016 was observed. BNN, the top DL performer, showed stability in all metrics; its R^2 decreased by only 0.002 and RMSE by 0.010. Across all four DL models, the average R^2 change was 0.009, representing a relative variation of less than 2%. RMSE changes ranged from 0.002 to 0.030, all within acceptable tolerance.

3.2. Overall Predictive Performance of Machine Learning and Deep Learning Models

Following feature selection, all nine models (five ML models and four DL architectures) were evaluated on their respective training and testing sets to establish overall baseline predictive performance. This section presents the comprehensive performance metrics and regression quality assessments for both types of models. The details in this section form the foundation for subsequent detailed analyses of error distribution patterns (Section 3.4) and model complexity and computational efficiency (Section 3.5).

Tables 10 and 11 summarize the predictive performance of all ML models and DL architectures across training, validation, cross-validation, and testing datasets, respectively. All metrics (such as R^2, RMSE, and MAE) are reported on the logarithmic scale (log-transformed creep life) to maintain consistency with the modeling approach. Among ML models, SVR achieved the highest predictive accuracy, with a test R^2 of 0.98 and RMSE of 0.36. This is closely followed by GBR, which achieved the test R^2 of 0.97 and RMSE of 0.38. XGBoost achieved an R^2 of 0.97 and an RMSE of 0.39. RF demonstrated moderate performance, achieved an R^2 of 0.94, and an RMSE of 0.60, whereas LR showed the poorest performance among ML models, with an R^2 of 0.91 and an RMSE of 0.79. The poor performance of LR confirms the importance of capturing non-linear relationships in creep behavior modeling.

Table 10. Performance metrics of ML models for creep life prediction. R^2 values for training, test sets, and cross-validation (5-fold) along with RMSE and MAE on the training and test set for LR, RF, SVR, GBR, and XGBoost models. All metrics are calculated on log-transformed creep rupture life values.

		LR	SVR	RF	GBR	XGBoost
R^2	Training	0.85	0.98	0.98	0.99	0.99
	Testing	0.91	0.98	0.94	0.97	0.97
	Cross-Validation	0.84	0.96	0.88	0.94	0.94
RMSE	Training	0.93	0.31	0.28	0.17	0.21
	Testing	0.79	0.36	0.60	0.38	0.39
MAE	Training	0.72	0.18	0.19	0.11	0.15
	Testing	0.62	0.27	0.41	0.29	0.30

Table 11. Performance metrics of DL models for creep life prediction. R^2 values for training, test sets, and cross-validation (5-fold) along with RMSE and MAE on the training and test set for DNN, BNN, 1D CNN, and LSTM models. All metrics are calculated on log-transformed creep rupture life values.

		DNN	BNN	1D CNN	LSTM
R^2	Training	0.96	0.98	0.92	0.94
	Testing	0.94	0.97	0.92	0.95
	Validation	0.95	0.97	0.91	0.94
	Cross-validation	0.91	0.95	0.85	0.88
RMSE	Training	0.58	0.39	0.62	0.56
	Testing	0.65	0.41	0.78	0.60
MAE	Training	0.41	0.29	0.48	0.42
	Testing	0.49	0.30	0.63	0.50

To assess the sensitivity to training set size, the SVR model was retrained using 40–100% of the training data while keeping the hyperparameters and test set fixed. The test R^2 increased from 0.94 using 40% of the training data (N = 147) to 0.97 using 60% (N = 220), after which further increases in training data resulted in only marginal improvements, reaching R^2 of 0.98 using the full training dataset (N = 371). RMSE showed a similar trend, decreasing from 0.44 to 0.36, demonstrating that the model continued to benefit from additional training data despite diminishing returns beyond approximately 220 training samples. Even when trained on 40% of the training data, the SVR model achieved an RMSE (0.61) substantially lower than that of the weakest ML baseline, linear regression (0.79; Table 10), demonstrating reasonable robustness to reduced sample size. The best performance was achieved using the full training dataset (80:20 split), as shown in Table A2 in the Appendix A.

DL models showed greater performance variability. For instance, BNN emerged as the best-performing DL architecture with a test R^2 of 0.97 and RMSE of 0.41, achieving accuracy comparable to top-performer ML models. DNN demonstrated moderate performance, achieved an R^2 of 0.94 and an RMSE of 0.65. LSTM also exhibited almost the same accuracy,

achieved an R^2 of 0.95 and an RMSE of 0.60. Particularly, CNN showed the poorest performance across all models evaluated. CNN achieved an R^2 of 0.92 and an RMSE of 0.78. This indicates the fundamental architectural mismatch with the tabular creep life data structure [71–73].

Cross-validation R^2 values demonstrated strong consistency with test performance (Tables 10 and 11). Among ML models, SVR showed minimal deviation with a cross-validation R^2 of 0.96 and test R^2 of 0.98 (ΔR^2 = 0.02), while GBR and XGBoost maintained a cross-validation R^2 of 0.94 against a test R^2 of 0.97. LR and RF showed larger discrepancies, with LR achieving a cross-validation R^2 of 0.84 and a test R^2 of 0.91, and RF achieving a cross-validation R^2 of 0.88 and a test R^2 of 0.94. For DL models, BNN demonstrated excellent consistency across all evaluation stages: cross-validation R^2 of 0.95, validation R^2 of 0.97, and test R^2 of 0.97 (ΔR^2 = 0.02), confirming robust generalization through Bayesian regularization and MC dropout. DNN showed good alignment with cross-validation R^2 of 0.91 and test R^2 of 0.94, while LSTM and 1D CNN exhibited larger variations (ΔR^2 = 0.07–0.08).

Figure 8 shows the regression plots for all five ML models. It displays predicted versus actual log-transformed creep life for both training and test sets. The closeness of data points to the ideal 45-degree line (perfect prediction) provides visual confirmation of model accuracy. The spread of points in two colors—training (blue markers) and test (orange markers)—indicates prediction variance. The regression plot of SVR demonstrates excellent alignment with the perfect prediction line across the entire range of creep rupture life values, with minimal scatter in both training and test datasets. The model shows consistent predictive capability with no systematic bias. The close fitting of test points around the perfect prediction line confirms the generalization capability of the model. GBR and XGBoost show similar regression characteristics to SVR. However, both models show slightly increased scatter in test predictions. RF shows noticeably greater scatter in test predictions. The regression plot of LR clearly reveals its limitations, showing systematic deviations from the perfect prediction line. This behavior confirms that linear assumptions are insufficient to capture the complex interactions of features in 2.25Cr1Mo steel. Figure 9 shows regression performance for the four DL architectures evaluated. BNN shows the most controlled prediction behavior among the four DL models. Its test predictions show close scatter to the perfect prediction line, indicating effective regularization through Monte Carlo dropout. DNN demonstrates greater prediction variance than BNN. The test predictions show noticeably wider spread around the perfect prediction line. The regression plot of LSTM reveals fundamental challenges in applying recurrent architectures to tabular creep rupture data. While the model captures general trends, the significant deviations of test points appear across the overall creep life range, especially in extreme regions. The sequential processing and treating of the 8 features as a time sequence does not align with the problem structure. These 8 features represent independent material properties and test conditions rather than temporal dependencies. So, this architectural mismatch of LSTM reveals as inconsistent predictions and higher error variance, making it ineffective in regression problems [71,72]. Also, CNN presents the most problematic regression behavior in DL models because it shows extreme scatter and systematic bias in the overall creep life range. Its test predictions deviate substantially from the perfect prediction line, as shown in Figure 9c. Specifically, 1D CNN aggregates information from adjacent input positions, which is a valid inductive bias only when the ordering of input features carries physical meaning, for example, consecutive time steps in a signal, or adjacent wavelengths in a spectrum. In the present dataset, the 8 input features (Cr, Si, Mo, N, N Temp, T Temp, Test Stress, Test Temp) have no inherent order: permuting their column sequence does not change the physical problem, since each feature is an independent material or test-condition variable rather than a position in a sequence. A convolutional kernel,

however, treats column adjacency as meaningful and will produce materially different results depending on arbitrary column ordering. This structural mismatch, not a general limitation of convolutional architectures, explains CNN's poor performance here, and the same architecture performs well on genuinely ordered data such as time series or images. This interpretation makes the model unsuitable for practical application in regression-based creep life prediction despite more layers and the maximum number of parameters [71,73].

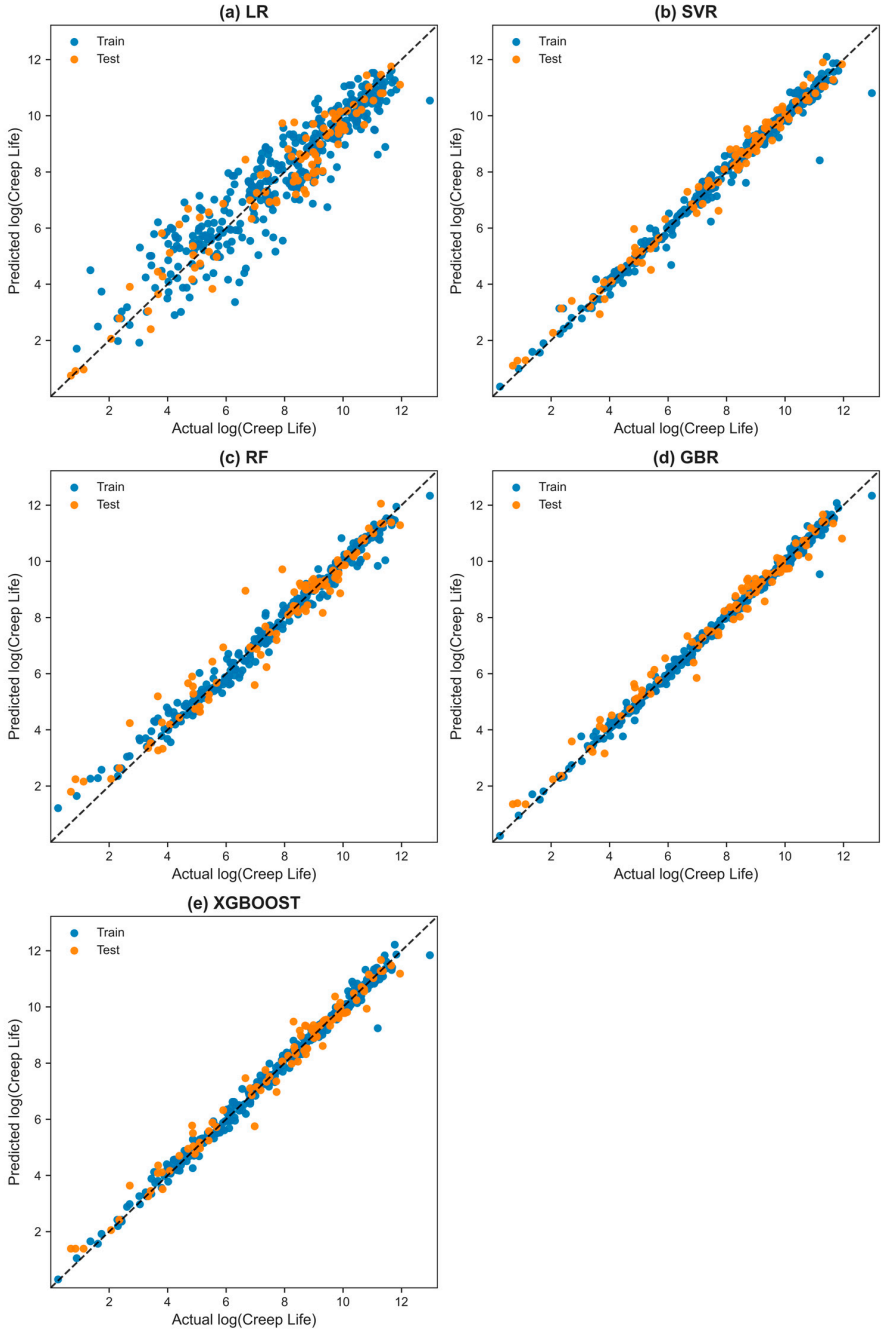


Figure 8. Predicted vs. actual log (creep life) for (**a**) LR, (**b**) SVR, (**c**) RF, (**d**) GBR and (**e**) XGBoost. In all panels, the *x*-axis represents the actual log-transformed creep rupture life (log hours), and the *y*-axis represents the predicted log-transformed creep rupture life (log hours). Blue circles represent training data, orange circles represent test data, and the dashed black line indicates perfect prediction. SVR demonstrates superior predictive performance with minimal scatter from the ideal prediction line.

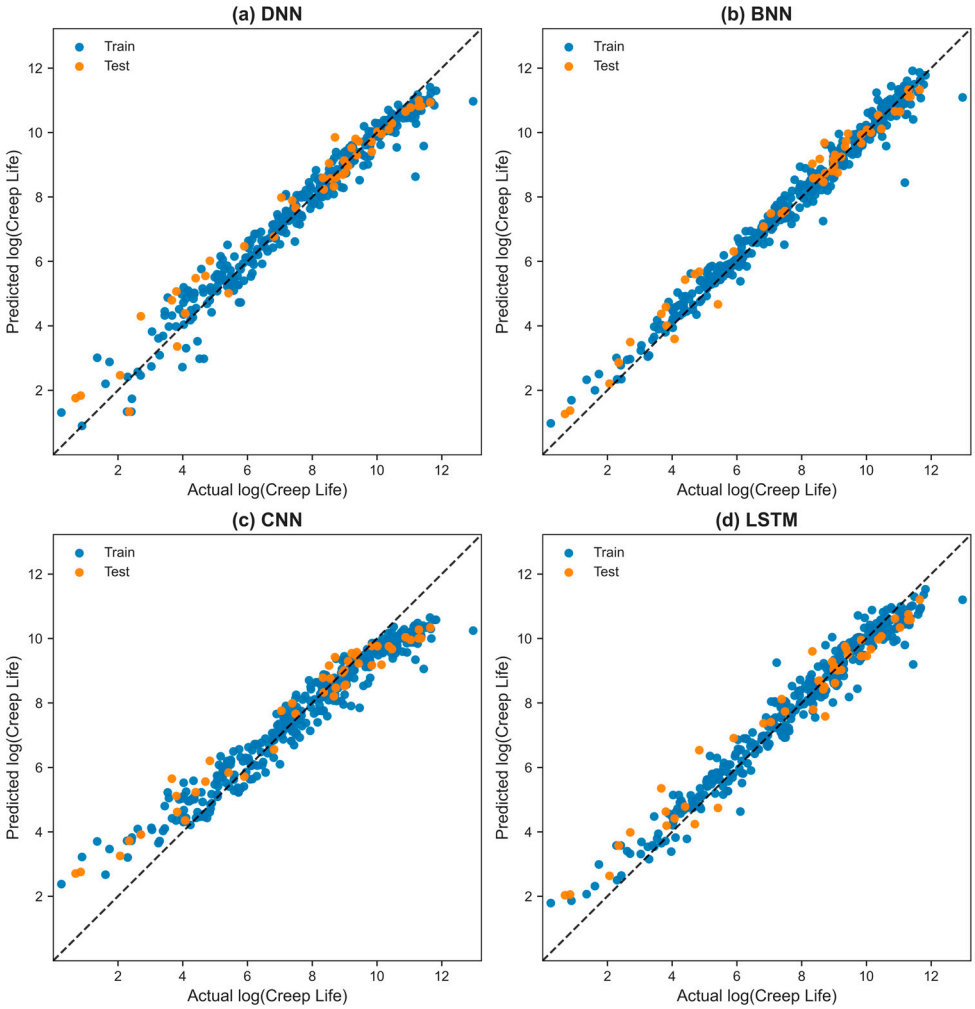


Figure 9. Predicted vs. actual log (creep life) for (**a**) DNN, (**b**) BNN, (**c**) 1D CNN, and (**d**) LSTM network. In all panels, the *x*-axis represents the actual log-transformed creep rupture life (log hours), and the *y*-axis represents the predicted log-transformed creep rupture life (log hours). Blue circles represent training data, orange circles represent test data, and the dashed black line indicates perfect prediction. BNN demonstrates superior predictive performance with minimal scatter from the ideal prediction line.

To validate the physical robustness of the SVR model, test stress and test temperature were each varied by ±5% independently while holding all other inputs constant (Table 12). All four perturbations yielded physically consistent responses: increasing either feature reduced the predicted log(rupture time) and decreasing it extended life, confirming that the model has learned physically meaningful input–output relationships rather than statistical artefacts. The temperature effect (mean $\Delta\log = -1.58$ per +5%) was approximately five times larger than the stress effect (mean $\Delta\log = -0.32$ per +5%), consistent with the exponential Arrhenius-type temperature dependence of thermally activated creep deformation [65]. Sensitivity of predicted rupture life to compositional features (Cr, Mo, Si, N) is qualitatively reflected in Figure 5 through their Pearson r values; quantitative composition-level sensitivity analysis is recommended for future work, particularly for steels with intentional compositional variations outside the standard 2.25Cr1Mo specification range. This study focuses on ML and DL model comparison; therefore, direct comparison of the recommended SVR model with conventional parametric methods (Larson–Miller Parameter and Monkman–Grant) on the same 464-sample dataset was not performed, as it has been widely reported in previous studies [22,23,53].

Table 12. SVR sensitivity analysis: effect of ±5% perturbation in test stress and test temperature on predicted log(rupture time). Life Factor = exp(Mean Δlog); values below 1.0 indicate life reduction, above 1.0 indicate life extension. All perturbations are physically consistent: increasing stress or temperature reduces the predicted rupture life, whereas decreasing either variable increases it.

Feature	Perturbation	Mean Δlog	±Std	Range (Min, Max)	Life Factor
Test Stress (MPa)	+5%	−0.32	0.16	−0.78, −0.05	0.73
Test Stress (MPa)	−5%	+0.33	0.16	+0.04, +0.78	1.38
Test Temp (°C)	+5%	−1.58	0.42	−2.34, +0.09	0.21
Test Temp (°C)	−5%	+1.72	0.41	+0.03, +2.61	5.59

In conclusion, comparing five ML and four DL approaches reveals several notable trends. Among the ML models, SVR outperformed the other ML models. On the other hand, among DL models, BNN outperformed other DL models. SVR even surpassed BNN by a small but measurable margin, reflected in a 3.2% reduction in test RMSE. Furthermore, except for LR, four of five ML models achieved an R^2 greater than 0.94, while only one DL model, BNN, exceeded an R^2 of 0.95. This indicates that classical ML models, especially SVR, demonstrate greater reliability for this specific problem class. These baseline performance results set the stage for deeper analysis of error distribution patterns (Section 3.4) and computational efficiency (Section 3.5) to inform final model selection recommendations.

3.3. Scatter Band Analysis

While overall predictive performance (Section 3.2) establishes baseline accuracy, it does not provide insight into prediction reliability, which is critical for engineering applications. To obtain such insights, scatter band analysis, such as ±2σ and ±3σ confidence intervals, is utilized. The standard deviation σ here refers to the standard deviation of the residuals (actual minus predicted log-transformed creep life) calculated on the test set for each model. These intervals provide a different perspective by measuring how many predictions are within acceptable limits. They also identify potentially risky outliers, which is critical for structural integrity. In statistical modeling, the ±2 and ±3 intervals correspond to approximately 95.45% and 99.73% coverage under the assumption of normally distributed prediction errors [74,75]. For engineering design applications, these intervals translate directly to safety margins. ±2σ represents a standard confidence level for preliminary design, while ±3σ provides the conservative bounds required for safety-critical components [76,77]. In this section, all nine models were evaluated through statistical reliability and engineering safety margins. These insights showed distinct patterns that offer valuable guidance for assessing deployment risks. Table 13 presents the sigma band coverage statistics for all models. These sigma band percentages are calculated on their respective test sets. Figure 10 (ML models) and Figure 11 (DL models) provide visual confirmation through predicted vs. actual plots with ±2σ (red dashed) and ±3σ (green dashed) bands. The data points in the figure are color-coded by test stress level to reveal potential stress-dependent reliability patterns.

Table 13. Scatter band analysis for model prediction accuracy. Percentage of test set predictions falling within ±2 and ±3 sigma bands for ML models (LR, RF, SVR, GBR, XGBoost) and DL models (DNN, BNN, 1D CNN, LSTM). Asterisks (*) indicate the highest percentage in each model category. The model with the best performance is highlighted in bold.

Models	R^2	±2 Sigma Band	±3 Sigma Band
LR	0.91	64.10%	83.70%
SVR *	**0.98**	**93.40%**	**97.90%**
RF	0.94	82.60%	90.20%

Table 13. *Cont.*

Models	R^2	±2 Sigma Band	±3 Sigma Band
GBR	0.97	92.10%	97.80%
XGBoost	0.97	89.10%	97.20%
DNN	0.94	82.40%	96.80%
BNN *	**0.97**	**91.80%**	**97.00%**
1D CNN	0.92	67.40%	84.80%
LSTM	0.95	79.80%	93.30%

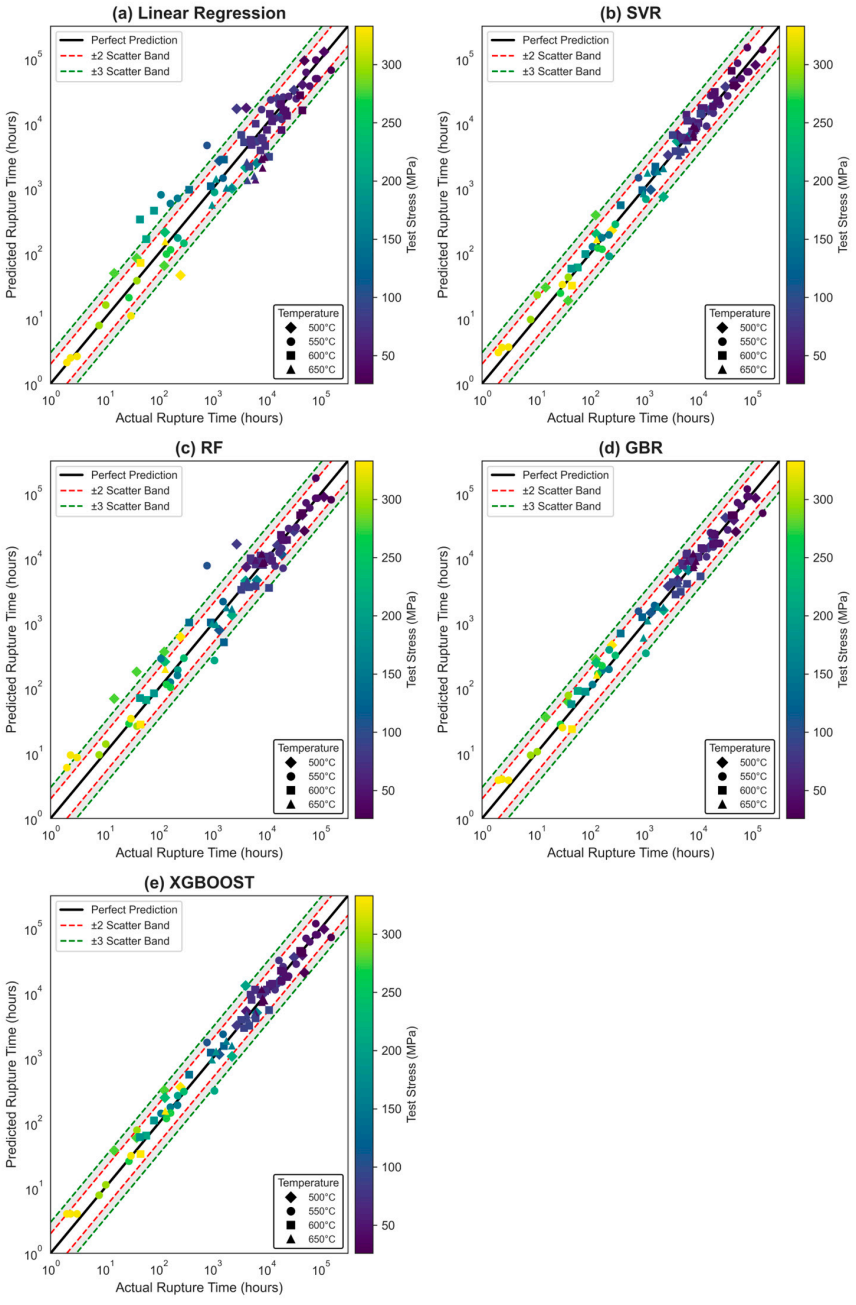


Figure 10. Creep rupture time predictions with scatter band analysis. Performance evaluation of (**a**) LR, (**b**) SVR, (**c**) RF, and (**d**) GBR (**e**) XGBoost models on test data. Black solid line: perfect prediction; red dashed lines: ±2 factor scatter band; green dashed lines: ±3 factor scatter band. Markers indicate test temperature (◆ 500 °C, ● 550 °C, ■ 600 °C, ▲ 650 °C) and stress level, colored by applied stress (yellowish color indicates higher stress, while bluish color indicates lower stress). SVR demonstrates superior accuracy with most predictions (93.40%) within the ±2 scatter band across all temperature–stress combinations.

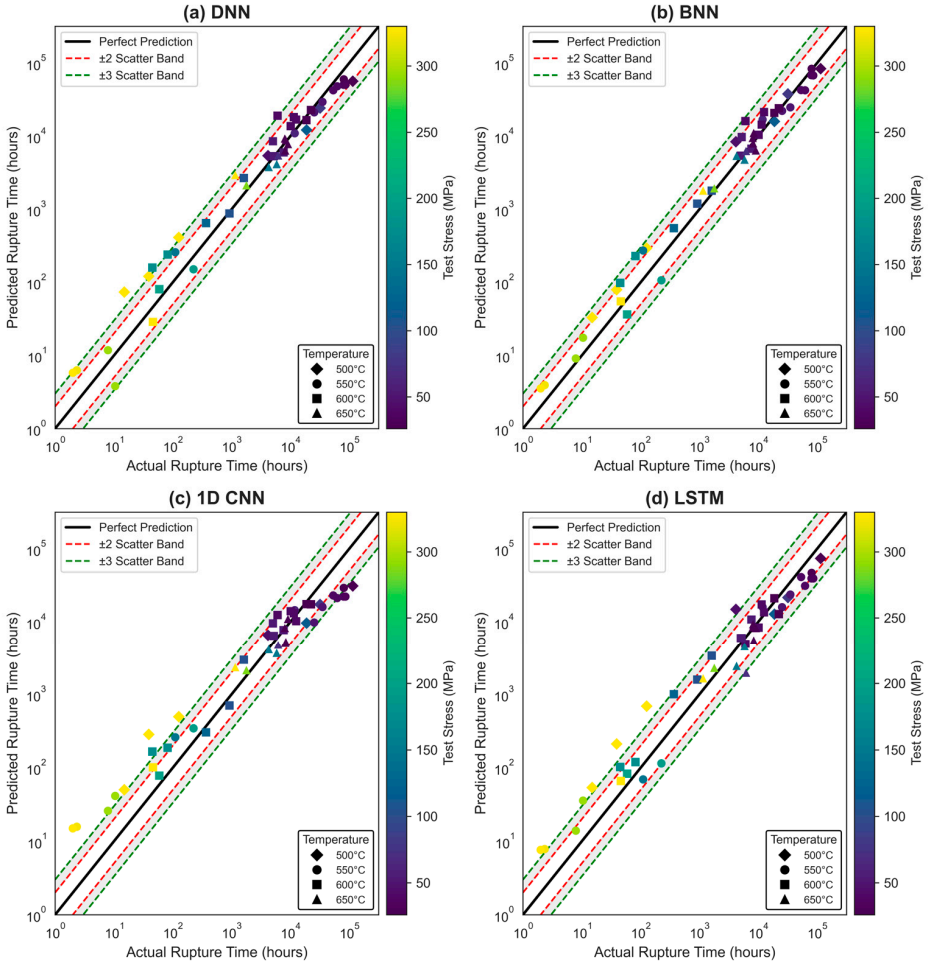


Figure 11. Creep rupture time predictions with scatter band analysis. Performance evaluation of (**a**) DNN, (**b**) BNN, (**c**) 1D CNN, and (**d**) LSTM models on test data. Black solid line: perfect prediction; red dashed lines: ±2 factor scatter band; green dashed lines: ±3 factor scatter band. Markers indicate test temperature (◆ 500 °C, ● 550 °C, ■ 600 °C, ▲ 650 °C) and stress level, colored by applied stress (yellowish color indicates higher stress, while bluish color indicates lower stress). BNN demonstrates superior accuracy with most predictions (91.80%) within the ±2 scatter band across all temperature–stress combinations.

In ML models, SVR showed excellent prediction reliability with 93.40% of test predictions falling within ±2σ and 97.90% within the ±3σ band. It approached the theoretical expectations for normally distributed errors such as 95.45% and 99.73%. The scatter band plot (Figure 10b) reveals tight clustering of predictions around the perfect prediction line across the entire rupture life range. Only three points are out of the ±2σ line, and no points surpass the ±3σ line; however, two points are on the ±3σ bound. SVR showed no catastrophic outliers. GBR achieved comparable reliability, such as 92.10% within ±2σ and 97.80% within ±3σ. Its scatter band coverage was nearly identical to SVR despite a slightly higher aggregate RMSE. XGBoost showed a marginally reduced ±2σ coverage of 89.10% but maintained a strong ±3σ reliability of 97.20%. This indicates the presence of moderate but not extreme outliers. RF presented the first concerning reliability degradation among ML models. It achieved 82.60% coverage within ±2σ and 90.20% within ±3σ, which is substantially below theoretical expectations. LR showed the lowest prediction reliability with only 64.10% within ±2σ and 83.70% within ±3σ. This is the worst performance across all models. The scatter band plot (Figure 10a) clearly illustrates the source, its systematic S-shaped deviation from the ideal line. This reliability failure makes LR unsuitable for practical application, which requires non-linear modeling approaches.

In DL models, BNN achieved the best reliability among DL architectures. It reached 91.80% coverage within ±2σ and 97.00% within ±3σ. BNN approached the performance of top ML models, especially SVR and GBR, in terms of scatter band percentage value. The scatter band plot (Figure 11b) shows remarkably consistent prediction behavior. The data points were distributed uniformly within bands across the stress range. When compared to other DL models, the Bayesian regularization approach, employing Monte Carlo dropout during training, successfully constrains prediction uncertainty to manageable levels. From another perspective, BNN exhibited minimal stress-dependent reliability variation. All four visual stress levels, color-coded markers, show equivalent scatter patterns. This indicates robust generalization across operating conditions. DNN showed moderate predictive reliability, achieving 82.40% within ±2σ and 96.80% within ±3σ scatter bounds. LSTM showed concerning reliability characteristics, achieving 79.80% coverage within ±2σ and 93.30% within ±3σ. The scatter band plot of LSTM (Figure 11d) reveals wide, heterogeneous scatter across the overall rupture life range. Most of the points are outside ±2σ, and several points fall outside ±3σ. CNN showed the lowest reliability among all DL models, achieving only 67.40% coverage within ±2σ and 84.80% within ±3σ. Its unreliability is comparable to that of LR. The scatter band plot of CNN (Figure 11c) further illustrates its failure behavior, showing systematic errors characterized by extreme over-prediction in the short-term regime. The disparity between CNN's moderate R^2 (0.92) and its low ±2σ coverage (67.40%) illustrates why aggregate R^2 alone is an insufficient measure of prediction reliability: a model with systematic directional bias can achieve moderate R^2 while producing practically unacceptable predictions across much of the life range. This further justifies the multi-metric evaluation framework employed in this study.

In summary, comparing ML and DL architectures reveals systematic reliability differences. In terms of scatter band analysis, the top two ML models (SVR, GBR) achieve ±3σ coverage of 97.90–97.80%, while the best DL model, BNN, reaches 97.00%, providing essentially equivalent performance to SVR. At the ±2σ level, the top ML models, such as SVR, GBR and XGBoost, average 91.5% coverage compared to BNN's 91.8%. However, the average of other DL models is very low, i.e., 67.4–82.4%. Apart from BNN, DL models show inherently higher variance in their typical predictions. This is likely because their larger parameter space presents greater optimization challenges with limited data.

3.4. Error Distribution Analysis

The aggregate performance metrics presented in Section 3.2, such as R^2, RMSE, and MAE, provide an overall assessment of ML and DL models' predictive accuracy. However, this overall assessment masks the important variations in predictive performance across different creep rupture life ranges. The four-range division is motivated by the observation that models achieving similar aggregate R^2 can show very different systematic biases across different service-life regimes. Since equipment in power and petrochemical plants operates across a broad range of conditions, from high-stress short-life scenarios to low-stress long-life service, characterizing model accuracy within each regime is more relevant to engineering applications than reporting a single aggregate metric. The acquired creep life data used for modeling in this study have a wide range of service conditions from approximately 1.8 h to 440,000 h. To gain deeper insights into ML and DL models' behavior across different creep life ranges, a comprehensive error distribution analysis was conducted. The creep rupture data was divided into four distinct rupture life ranges based on log-transformed rupture time values. Each range needs a systematic evaluation model to capture different error patterns.

The partition of the data is described in Table 14. This division shows the natural distribution observed in the dataset. Most of the data points fall within the medium range (R2)

and medium-to-long term range (R3). This pattern makes sense because these two ranges (R2 and R3) are the most common conditions in creep rupture life engineering practice.

Table 14. Distribution of creep rupture life data across different time ranges. Creep life ranges shown in logarithmic scale and actual hours, with corresponding sample sizes for ML (80/20 split) and DL (80/10/10 split) datasets.

No.	Range (log)	Range (h)	Description	ML Sample Size (Train/Test)	DL Sample Size (Train/Test)
R1	0.25–4	1.29–55.8	Very short-term	25/12	25/8
R2	4.0–8.0	55.8–3159.5	Short-to-medium	152/26	152/10
R3	8.0–11.0	3159.5–60,359.1	Medium-to-long-term	158/48	158/23
R4	11.0–13.0	60,359.1–430,054	Very long-term	33/6	33/5

The error analysis showed different performance patterns for ML and DL models across the four different rupture life ranges. Among ML models, SVR achieved the most consistent performance, showing the lowest RMSE in three out of four ranges: R1 (RMSE: 0.438), R3 (RMSE: 0.301), and R4 (RMSE: 0.349). GBR displayed comparable accuracy in R2 (RMSE: 0.436), as presented in Table 14, Figure 12a. RF, GBR, and XGB generally showed comparable accuracy in the medium-to-long term ranges (R2 and R3) but demonstrated increased errors in the extreme ranges (R1 and R4). LR consistently underperformed across all ranges, with RMSE values ranging from 0.509 to 0.990. More critically, LR showed systematic bias, over-predicting in short-term ranges such as R1 and R2 (ME: +0.27 to +0.29) and under-predicting in long-term ranges such as R2 and R3 (ME: −0.30 to −0.35). These overpredictions and underpredictions of LR show an inability to capture the non-linear relationships inherent in creep behavior. RF also showed severe bias and overprediction in very short creep life ranges such as R1 (ME: +0.56). This behavior of showing bias in extreme regions can be attributed to RF's tendency to avoid extreme predictions outside its training range, resulting in conservative estimates that deviate from actual values at the lower extreme of the dataset. The details of overprediction and underprediction of all models are shown in Figure 12b.

DL models showed greater variability in performance across ranges compared to ML models. In DL models, BNN emerged as the best-performing DL model, achieving the lowest RMSE in three ranges: Range 1 (RMSE: 0.831), Range 2 (RMSE: 0.472), and Range 4 (RMSE: 0.404), as shown in Table 15, Figure 12a. The Bayesian approach's inherent uncertainty quantification and regularization likely contributed to its robustness across diverse rupture life regimes. DNN showed exceptional performance in R3 (RMSE: 0.306), rivaling the best ML model (SVR: 0.301). It shows that when the data region is well represented and abundant, deep architectures can effectively capture complex patterns. CNN showed catastrophic performance degradation in extreme ranges, with RMSE values exceeding 1.0 in both R1 (1.271) and R4 (1.120). Moreover, CNN showed severe systematic bias, tremendously over-predicting in R1 (ME: +1.19) and intensely under-predicting in R4 (ME: −1.10). This behavior indicates that the convolutional architecture, which excels at spatial pattern recognition, may not be well-suited for capturing the temporal dependencies and extrapolation requirements inherent in tabular data [78]. LSTM performance was intermediate, with moderate errors across most ranges but notable under-prediction in extreme regimes (ME: −0.61 in Range 4), as shown in Figure 12b. Figure 12c shows the comparison between the best ML model (SVR) and best DL model (BNN) across all ranges. The overall mean RMSE across all ranges was 0.519 for ML models and 0.690 for DL models, as shown in Figure 12d, representing a 24% performance advantage for ML approaches. This difference is particularly pronounced in data-sparse extreme ranges (R1 and R4), where DL models' tendency toward overfitting becomes more apparent.

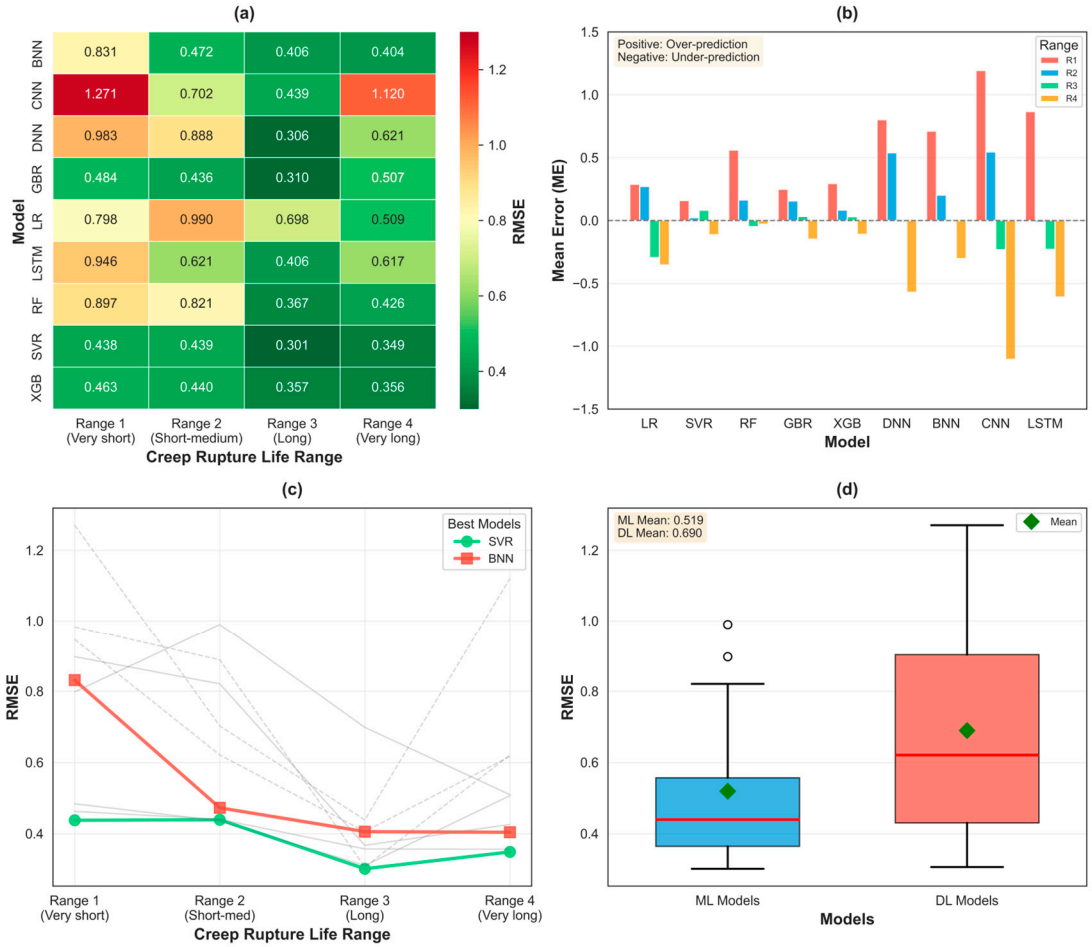


Figure 12. Error distribution analysis across various creep rupture life regimes (Range 1–4). (**a**) RMSE heatmap across models and creep life regimes. (**b**) Systematic bias across models and ranges. (**c**) RMSE trends across rupture life ranges. (**d**) RMSE distribution: ML vs. DL models.

Table 15. Range-specific performance analysis of best-performing models. RMSE and MAE values for the best ML models and DL architectures across different creep life ranges, along with average error metrics (RMSE and MAE) and identification of models exhibiting high bias (mean error > 0.5). Performance metrics calculated on test set predictions in logarithmic scale.

Range	Best ML Model RMSE/MAE	Best DL Model RMSE/MAE	Average Error RMSE/MAE	Models With Highest Bias (ME > 0.5)
R1	SVR RMSE: 0.438 MAE: 0.363	BNN RMSE: 0.831 MAE: 0.711	ML: 0.616/0.475 DL: 1.008/0.913	CNN (ME: 1.19) LSTM (ME: 0.86) DNN (ME: 0.80) RF (ME: 0.56)
R2	GBR RMSE: 0.436 MAE: 0.338	BNN RMSE: 0.472 MAE: 0.405	ML: 0.625/0.472 DL: 0.671/0.552	CNN (ME: 0.54) DNN (ME: 0.53)
R3	SVR RMSE: 0.301 MAE: 0.230	DNN RMSE: 0.306 MAE: 0.258	ML: 0.407/0.315 DL: 0.389/0.319	All models mean error is above 0.5
R4	SVR RMSE: 0.349 MAE: 0.314	BNN RMSE: 0.404 MAE: 0.392	ML: 0.429/0.325 DL: 0.690/0.669	CNN (ME: −1.10) LSTM (ME: −0.60) DNN (ME: −0.57)

To know more about systematic bias and prediction reliability, Figure 12b shows the mean error of all the models, which provides critical insights into systematic predictions. SVR among the ML models and BNN among the DL models showed good performance.

However, when comparing all models, SVR demonstrated balanced performance across all creep life ranges with a mean error close to zero (−0.11 to +0.16). CNN showed bad performances in both directions, with over-prediction in the short term (+1.19 in R1) and severe under-prediction in long-term regimes (−1.10 in R4). Also, LR showed consistent over-predictions and under-predictions, over-predicting short-term and under-predicting long-term behavior. BNN performed very well in R3 and R2, but its performance is poor in R1. Similarly, DNN also showed poor performance in R1 and R4.

The performance in data-sparse regions, the extreme ranges (R1 and R4), presented the greatest modeling challenges due to limited training data representation. In these regions, all models showed performance degradation, but to varying degrees. Table 15 shows that, in R1, the average RMSE for ML models is 0.616 and 1.008 for DL models (CNN and LSTM increased this error). In R4, the average RMSE is 0.429 for ML models and 0.690 for DL models. Box plot analysis, as shown in Figure 12d, demonstrated that ML models not only achieved lower mean RMSE but also demonstrated more consistent performance with tighter error distributions. The interquartile range for ML models (0.30–0.60) is narrower than that for DL models (0.35–0.85), indicating greater prediction stability. The increment of the interquartile range of DL is due to the worst performance of CNN, which emerged as a clear outlier with the highest variability and worst overall performance, making it unsuitable for this regression task.

The relatively normal distribution of residuals for top-performing models (SVR, BNN) in the medium-to-long term ranges (Ranges 2 and 3) suggests that these models have successfully captured the underlying relationships without significant error variance. However, increased residual variance in extreme ranges indicates areas for potential model improvement. In a nutshell, overall, the superior performance of ML models, particularly SVR, in these sparse data regions can be attributed to their simpler architectures and stronger generalization capabilities. DL models, despite their capacity to learn complex patterns, suffer from overfitting when data is limited, particularly at distribution extremes. These findings suggest that, for practical creep rupture life prediction with a limited dataset, SVR provides the best balance of accuracy, consistency, and reliability across the entire service-life spectrum. While DL models show promise in well-represented data regions, where data is well structured and abundant, their increased complexity does not translate to superior performance given the dataset size and characteristics of this application.

3.5. Model Complexity and Computational Efficiency

In model selection, predictive accuracy is one of the primary criteria in engineering applications. Nevertheless, model complexity and computational efficiency are equally important for practical deployment [79,80]. In this section, model complexity is quantified using the number of trainable parameters, as summarized in Table 16. Computational efficiency is then evaluated by analyzing training time, inference speed, and how these metrics relate to predictive performance. All variables used to assess model complexity and computational efficiency are also described in Table 16. In ML models, LR, the simplest model, which has only 9 parameters (8 features plus intercept), while RF has 222,678 parameters across 500 decision trees. Among the DL models, the parameter counts range from 761 in the BNN to 31,489 in the LSTM, reflecting the substantial architectural differences between these approaches.

The relationship between model complexity and performance reveals a famous characteristic bias–variance trade-off pattern, as shown in Figure 13. LR has a smaller number of parameters, not a complex model, but these fewer parameters are unable to capture the non-linear interactions governing creep behavior, which is why it suffers from high bias (underfitting) and has poor performance. On the other hand, complex models like LSTM,

RF, and CNN tend to overfit, learning details specific to the training data instead of general patterns in creep life data. SVR demonstrates less complexity with 263 parameters; it takes place in the best position, between the simplicity of LR and the complexity of ensemble methods such as RF and XGBoost, as shown in Figure 13. This intermediate position, neither simple nor complex, proves optimal for the dataset size and feature dimensionality, as evidenced by its superior performance metrics (R^2 = 0.98, RMSE = 0.36). Similarly, BNN, in the appropriate position, shows competitive performance (R^2 = 0.97) with a smaller number of parameters (761 parameters). SVR occupies the optimal region of this trade-off space, achieving the lowest combined bias and variance through its kernel-based formulation and built-in regularization. The RBF kernel implicitly maps the 8-dimensional feature space to a higher-dimensional space where linear separation becomes possible, while the regularization parameter C controls model complexity to prevent overfitting. This balance proves ideal for the dataset size (464 samples) and problem structure. The competitive performance of BNN is due to Monte Carlo dropout-based regularization and relatively modest architecture. CNN lies in the less efficient region, has high complexity, less predictive performance and high training and inference time. The position of CNN indicates the fundamental architectural mismatch with the problem structure. Basically, convolutional operations are designed to extract spatial hierarchies in images and sequences. It does not provide an advantage for tabular data like creep rupture data with independent features [71]. The enforced local connectivity unnecessarily restricts the model's expressiveness for this application; such local connectivity and weight sharing are highly advantageous for image data [73].

Table 16. Computational efficiency and model complexity analysis of ML and DL models. Number of model parameters, training time (seconds), inference time per sample (milliseconds), test set R^2 performance, and logarithmic parameter count for nine models. Metrics provide a comprehensive assessment of the trade-off between model complexity, computational cost, and predictive accuracy. Values in the last column represent $\log_{10}$ of the parameter count in column 2.

Model	No. of Parameters	Training Time (s)	Inference Time (ms)	R^2	Value of Parameter
LR	9	0.0085	0.0186	0.91	0.95
SVR	263	0.0556	0.0527	0.98	2.41
RF	222,678	1.4609	0.4547	0.94	5.34
GBR	6500	0.4374	0.0299	0.97	3.81
XGBoost	3890	0.1302	0.0499	0.97	3.58
DNN	1785	148.36	2.18	0.95	3.25
BNN	761	76.89	2.27	0.97	2.88
1D CNN	9633	163.98	2.68	0.93	3.98
LSTM	31,489	402.80	2.78	0.95	4.49

Model complexity was evaluated using the parameter efficiency metric ($R^2/\log_{10}$(parameters)). Based on this metric, SVR (0.40) and BNN (0.34) demonstrated the highest efficiency among the evaluated models. This indicates that these models achieved strong predictive performance while maintaining relatively low parameter counts. Another important aspect of model complexity, particularly for assessing performance and mitigating overfitting, is the parameter-to-data ratio [18]. Models with parameter counts compatible with the available data demonstrate more stable performance and better generalization. In our case, SVR, the most efficient model, has a parameter-to-data ratio of 0.56 (263/464), BNN has 1.64 (761/464), while the most complex model, LSTM, has a ratio of 68.17 (31,489/464). In this study, complex DL models such as LSTM, CNN, and DNN indicate severe parameter overabundance relative to available data. This overparameterization demands aggressive regularization (dropout, early stopping, weight decay) that ultimately constrains the model's effective capacity. Also, the Vapnik-Chervonenkis

(VC) dimension of a model bounds the minimum samples required for generalization. LSTM and 1D-CNN, with parameter-to-data ratios of 68.2 and 20.8, respectively (Table 16), exceed safe limits for this 464-sample dataset. Their inclusion in this study is deliberate: to empirically demonstrate, not assume, their limitations in small-data tabular settings, consistent with [31].

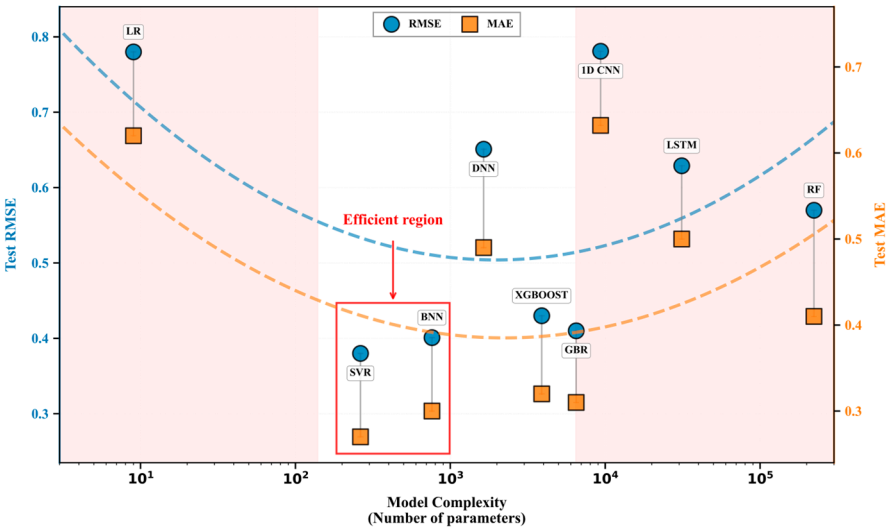


Figure 13. Trade-off between model complexity and predictive performance for various ML and DL models. The shaded light red regions qualitatively indicate underfitting and overfitting zones, while the central white area represents a balanced complexity range. The boundaries are not strict and depend on the ratio of model parameters to available training data. The red-highlighted region marks efficient models (SVR and BNN) that achieve low error with relatively few parameters.

Computational efficiency in terms of training time and inference speed shows clear differences between ML and DL approaches. Figure 14 shows the (a) training time and (b) inference time of all models. ML models require less training time. For instance, LR takes only 0.0085 s to train, while RF, a more complex model with a higher parameter count, takes 1.46 s. The best ML model, SVR, completed training in 0.0557 s. In contrast, DL models require longer training time. For example, the best DL model, BNN, completed training in 76.89 s, while DNN, CNN and LSTM took 148.37 s, 163.98 s, and 402.80 s. Other DL models, the worst performers, consume the longest periods to train as compared to the best DL model, BNN. However, there is a gap between SVR and BNN training times, representing a 1380-fold difference. This dramatic gap arises from fundamental algorithmic differences. For example, ML models like SVR solve convex optimization problems with closed-form or iterative solutions that converge rapidly. On the other hand, DL models require iterative gradient descent through high-dimensional parameter spaces with early stopping mechanisms to prevent overfitting. In the case of LSTM, with its recurrent architecture and 31,489 parameters, it required the longest training time (402.80 s). This highest training time represents a 7230-fold increase over SVR despite achieving inferior predictive performance. This computational overhead becomes significant when considering hyperparameter optimization. In this study, hyperparameter optimization was performed using 5-fold cross-validation for all models. So, the time easily extends to hours, while for ML models, especially SVR, it completes within minutes. This time consumption matters less in offline environments such as laboratory work, but for industrial settings where models must be frequently retrained with new data, this efficiency advantage becomes substantial. However, there are some areas in which simple ML models, such as SVR, cannot perform well. For instance, BNN provides native uncertainty quantification

through Monte Carlo dropout sampling. This offers probabilistic predictions valuable for reliability assessment, while traditional ML models are unable to provide it [81].

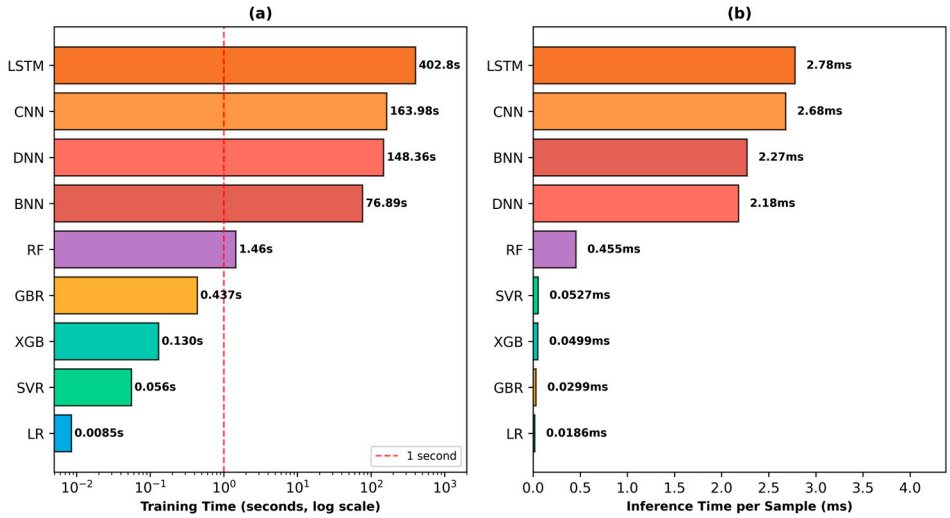


Figure 14. Computational efficiency comparison of ML and DL models for creep life prediction. (**a**) Training time on logarithmic scale ranging from 0.01 s to 402.80 s, and (**b**) inference time per sample ranging from 0.019 ms to 2.789 ms. DL models (LSTM, 1D CNN, DNN, BNN) require 50–4000× longer training times than ML models (RF, GBR, XGBoost, SVR, LR), while inference speeds remain consistent (all < 3 ms per prediction). Red dashed line in (**a**) indicates 1 s reference threshold. Color scheme differentiates DL architectures (red/orange) from ML algorithms (green/purple/blue).

Furthermore, inference time per sample provides critical insight into real-time prediction capabilities. ML models demonstrate good inference efficiency. For example, the inference times range from 0.019 ms (LR) to 0.455 ms (RF). SVR achieves inference in 0.053 ms per sample. This inference of SVR enables rapid predictions suitable for real-time prediction or online monitoring systems, as suggested by Jan and Chai [18] in their recent review. On the contrary, DL models exhibit uniformly slower inference, with per-sample prediction times ranging from 2.19 ms (DNN) to 2.79 ms (LSTM). The 43-fold inference speed advantage of SVR over BNN translates to practical differences in batch prediction scenarios. For example, predicting creep life for 10,000 design variants would require approximately 0.53 s with SVR compared to 22.79 s with BNN. In the future development, the edge computing in online monitoring of creep, as suggested by Jan and Chai [18], for that, SVR's 263 parameters require minimal memory, approximately 2 KB for double-precision storage, enabling deployment on edge devices or embedded systems. On the other hand, LSTM's 31,489 parameters, approximately 250 KB, present greater memory constraints and energy consumption in edge scenarios.

In summary, for structured tabular data with moderate feature dimensionality (8 features) and limited sample size (464 observations), model complexity must align with data availability to achieve optimal performance. SVR and BNN emerged as the most effective models among ML and DL architectures, respectively. This relationship is quantified through parameter efficiency ($R^2/\log_{10}(\text{parameters})$), where SVR (0.40) and BNN (0.34) ranked highest, demonstrating that appropriate model sizing, not architectural complexity, determines performance in limited-data scenarios. Beyond predictive accuracy, these models also exhibited superior computational efficiency. SVR required only 0.056 s training time and 0.053 ms inference latency, while BNN, despite being the best DL model, demanded longer training. These findings confirm that increasing model complexity does not improve accuracy when data are scarce; rather, appropriately sized models like SVR offer optimal

cost–benefit profiles for practical deployment in resource-constrained applications, such as in situ creep monitoring systems.

4. Conclusions

This study compared five ML models and four DL models for creep rupture life prediction using a unified evaluation framework. Feature selection, predictive accuracy, prediction reliability, range-specific error distribution, model complexity, and computational efficiency were systematically evaluated using a dataset of 464 experimental samples.

- Feature selection identified eight optimal input features (Cr, Si, Mo, N, Test Stress, Test Temp, T Temp, and N Temp), which maintained predictive performance while reducing model complexity.
- Among the ML models, SVR achieved the best overall performance (R^2 = 0.98, RMSE = 0.36). Among the DL models, BNN performed best (R^2 = 0.97, RMSE = 0.41), demonstrating that Bayesian regularization can effectively mitigate overfitting in limited-data problems.
- Range-specific error and reliability analyses showed that SVR consistently provided the lowest prediction errors and the most stable performance across the creep life range, particularly in data-sparse extreme regions. BNN was the most reliable DL model, although ML models achieved, on average, 24% lower RMSE than DL models.
- SVR also demonstrated the best balance between predictive accuracy, model complexity, and computational efficiency, whereas DL models required substantially greater computational cost without corresponding improvements in predictive performance.
- Overall, for moderate-sized tabular datasets, conventional ML models provide a better cost–benefit balance than more complex DL architectures. SVR is recommended for creep rupture life prediction owing to its high accuracy, robustness, and computational efficiency, while BNN represents the best-performing deep learning architecture evaluated in this study and provides a framework that can support predictive uncertainty estimation.

Future work can extend this framework to larger multi-steel datasets, incorporate creep rate features, include uncertainty quantification across broader operating conditions, and conduct external validation and extrapolation testing beyond the current training range.

Author Contributions: M.B.J.: Conceptualization, Investigation, Methodology, Visualization, Writing—original draft, Software; Z.W.: Validation, Formal analysis, Data curation; M.C.: Conceptualization, Writing—review & editing, Resources, Supervision, Project administration, Funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Fundamental Research Funds for the Central Universities, grant number [xxj032025061] and National Natural Science Foundation of China, grant number [22208258].

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

Abbreviations

ML	machine learning
DL	deep learning
MC	Monte Carlo
SVR	support vector regression
RF	random forest

XGBoost	extreme gradient boosting
GBR	gradient boosting regression
LR	linear regression
BNN	Bayesian neural network
DNN	deep neural network
LSTM	long short-term memory
CNN	convolutional neural network
RMSE	root mean squared error
RFR	random forest regression
CV	cross-validation
MAE	mean absolute error
MLP	multi-layer perceptron
AP	affinity propagation
GAN	generative adversarial network
TTP	time–temperature parameter
KRR	kernel ridge regression
LMP	Larson–Miller parameter
ODS	oxide dispersion-strengthened
ANN	artificial neural network
MH	Manson–Haferd
OSD	Orr–Sherby–Dorn
Cr-Mo	chromium–molybdenum
N Temp	normalizing temperature
N Time	normalizing time
T Temp	tempering temperature
T Time	tempering time
NIMS	National Institute for Material Science

Appendix A

Table A1. Summary of statistical descriptions of each parameter in the dataset. Statistical descriptors (min, max, mean, std) were computed from the full dataset of 464 samples prior to any train–test partitioning, using numpy descriptive statistics. Source: compiled from the NIMS creep database [39] and literature sources [36,41–50], as described in Section 2.2.

S. No.	Parameter	Unit	Min	Max	Mean	Std
1	C	wt %	0.0600	0.1600	0.1212	0.0173
2	Si	wt %	0.1000	0.4000	0.2710	0.0654
3	Mn	wt %	0.4000	0.6300	0.5007	0.0575
4	P	wt %	0.0010	0.0300	0.0146	0.0048
5	S	wt %	0.0020	0.0230	0.0118	0.0050
6	Ni	wt %	0.0250	0.3200	0.0974	0.0835
7	Cr	wt %	0.5900	2.4600	2.2275	0.2798
8	Mo	wt %	0.4300	1.0700	0.9583	0.0908
9	Cu	wt %	0.0200	0.2200	0.0774	0.0506
10	Al	wt %	0.0040	0.0190	0.0092	0.0046
11	N	wt %	0.0080	0.0142	0.0091	0.0011
12	N Temp	°C	910.00	990.00	926.13	14.769
13	N Time	Min	12.000	360.00	89.374	98.672
14	T Temp	°C	650.00	750.00	725.80	21.662
15	T Time	Min	22.000	480.00	159.74	127.46
16	Stress	Mpa	22.000	400.00	142.37	84.792
17	Temp	°C	450.00	650.00	560.59	52.003
18	Creep Life	h	1.2900	430,054	16,210	31,897
19	lg (CL)	Log h	0.2546	12.972	7.8011	2.5159

Table A2. Effect of reduced training set size on SVR test performance. Random subsamples containing 40%, 50%, 60%, 70%, and 80% of the 371-sample training pool were drawn and used to refit the SVR pipeline, holding the previously selected hyperparameters fixed (C = 100, γ = 0.1, RBF kernel; Table 5). All rows were evaluated on the same fixed test set used throughout this study, so that differences in performance reflect training data volume alone, not changes in the test set. The 100% row uses the complete training pool with no subsampling, corresponding to the standard 80:20 split reported in Table 10, and is included as the reference point against which the reduced-data rows are compared. The training fraction shown in bold represents the 80:20 split, as this data partition yielded the best performance.

Training Fraction	N Training Samples	Test R^2	Test RMSE	Test MAE
40%	147	0.94	0.61	0.45
50%	184	0.94	0.64	0.46
60%	220	0.97	0.44	0.32
70%	257	0.97	0.45	0.31
80%	294	0.97	0.41	0.29
100%	**371**	**0.98**	**0.36**	**0.27**

References

1. Fan, H.; Zhang, Z.; Dong, J.; Xu, W. China's R&D of advanced ultra-supercritical coal-fired power generation for addressing climate change. *Therm. Sci. Eng. Prog.* **2018**, *5*, 364–371. [CrossRef]
2. Rocha, D.H.; Siqueira, D.S.; Silva, R.J. Exergoenvironmental analysis for evaluating coal-fired power plants technologies. *Energy* **2021**, *233*, 121169. [CrossRef]
3. Surywanshi, G.D.; Pillai, B.B.K.; Patnaikuni, V.S.; Vooradi, R.; Anne, S.B. 4-E analyses of chemical looping combustion based subcritical, supercritical and ultra-supercritical coal-fired power plants. *Energy Convers. Manag.* **2019**, *200*, 112050. [CrossRef]
4. Pandey, C.; Mahapatra, M.; Kumar, P.; Vidyrathy, R.; Srivastava, A. Microstructure-based assessment of creep rupture behaviour of cast-forged P91 steel. *Mater. Sci. Eng. A* **2017**, *695*, 291–301. [CrossRef]
5. Holdsworth, S. Creep-fatigue failure diagnosis. *Materials* **2015**, *8*, 7757–7769. [CrossRef] [PubMed]
6. Ashby, M.; Dyson, B. Creep damage mechanics and micromechanisms. In *Fracture*; Elsevier: Amsterdam, The Netherlands, 1984; Volume 84, pp. 3–30. [CrossRef]
7. Wang, K.; Wang, X.; Wen, J.; Zhang, X.; Gong, J.; Tu, S. Creep rupture: From physical failure mechanisms to lifetime prediction of structures. *J. Mech. Eng.* **2021**, *57*, 132–152. [CrossRef]
8. Tan, Y.; Wang, X.; Kang, Z.; Ye, F.; Chen, Y.; Zhou, D.; Zhang, X.; Gong, J. Creep lifetime prediction of 9% Cr martensitic heat-resistant steel based on ensemble learning method. *J. Mater. Res. Technol.* **2022**, *21*, 4745–4760. [CrossRef]
9. Roy, P.K. Creep rupture analysis and life estimation of 1.25 Cr-0.5 Mo, 2.25 Cr-1Mo and modified 9Cr-1Mo steel: A comparative study. *J. Inst. Eng. Ser. D.* **2018**, *99*, 33–38. [CrossRef]
10. Larson, F.R.; Miller, J. A time-temperature relationship for rupture and creep stresses. *Trans. Am. Soc. Mech. Eng.* **1952**, *74*, 765–771. [CrossRef]
11. Evans, R.; Parker, J.; Wilshire, B. The θ projection concept—A model-based approach to design and life extension of engineering plant. *Int. J. Press. Vessel. Pip.* **1992**, *50*, 147–160. [CrossRef]
12. Yao, H.; Zhao, Y.; Song, X.; Jia, J.; Xiang, Z. On the applicability of boundary condition based tensile creep model in predicting long-term creep strengths and lifetimes of engineering alloys. *Eur. J. Mech.-A/Solids* **2019**, *73*, 57–66. [CrossRef]
13. Monkman, F.C. An empirical relationship between rupture life and minimum creep rate in creep-rupture tests. *Proc. ASTM* **1956**, *56*, 593–620.
14. Evans, M.; Williams, T. Assessing the capability of the Wilshire equations in predicting uniaxial creep curves: An application to Waspaloy. *Int. J. Press. Vessel. Pip.* **2019**, *172*, 153–165. [CrossRef]
15. Dyson, B. Use of CDM in materials modeling and component creep life prediction. *J. Press. Vessel Technol.* **2000**, *122*, 281–296. [CrossRef]
16. Ghosh, R. Creep life predictions of engineering components: Problems & prospects. *Procedia Eng.* **2013**, *55*, 599–606. [CrossRef]
17. Sattar, M.; Othman, A.; Kamaruddin, S.; Akhtar, M.; Khan, R. Limitations on the computational analysis of creep failure models: A review. *Eng. Fail. Anal.* **2022**, *134*, 105968. [CrossRef]
18. Jan, M.B.; Chai, M. Machine Learning Approaches for Creep Rupture Life Prediction of Metallic Materials: A Comprehensive Review. *Int. J. Press. Vessel. Pip.* **2025**, *219*, 105690. [CrossRef]
19. Yang, Z.; Dai, T.; Yang, B.; Wu, Y.; Bai, Y.; Zhao, J. Creep life prediction of 10CrMo910 steel main steam pipes based on carbide evolution and machine learning. *Mater. Today Commun.* **2025**, *45*, 112378. [CrossRef]

20. Yang, T.-X.; Dou, P. Prediction of creep rupture life of ODS steels based on machine learning. *Mater. Today Commun.* **2024**, *38*, 108117. [CrossRef]
21. Wang, Y.; Tsang, D.; Zhang, Y.; Zhang, Q.; Zhu, F.; Song, L.; Ma, X. Insights into irradiation creep coefficient in nuclear graphite from machine learning. *Nucl. Eng. Technol.* **2025**, *57*, 103559. [CrossRef]
22. Song, C.; Liu, X.; Zhu, L.; Fan, P.; Ren, S.; Zhang, K.; Chen, J.; Chen, H. Long-term creep life prediction of P91 steel using domain knowledge and back propagation artificial neural network. *Mater. High Temp.* **2025**, *42*, 14–24. [CrossRef]
23. Liang, T.; Liu, X.; Fan, P.; Zhu, L.; Bi, Y.; Zhang, Y. Prediction of long-term creep life of 9Cr–1Mo–V–Nb steel using artificial neural network. *Int. J. Press. Vessel. Pip.* **2020**, *179*, 104014. [CrossRef]
24. Xiang, S.; Chen, X.; Fan, Z.; Chen, T.; Lian, X. A deep learning-aided prediction approach for creep rupture time of Fe–Cr–Ni heat-resistant alloys by integrating textual and visual features. *J. Mater. Res. Technol.* **2022**, *18*, 268–281. [CrossRef]
25. Agrawal, A.; Choudhary, A. Deep materials informatics: Applications of deep learning in materials science. *MRS Commun.* **2019**, *9*, 779–792. [CrossRef]
26. Saoudi, A.; Fellah, M.; Hezil, N.; Lerari, D.; Khamouli, F.; Atoui, L.H.; Bachari, K.; Morozova, J.; Obrosov, A.; Samad, M.A. Prediction of mechanical properties of welded steel X70 pipeline using neural network modelling. *Int. J. Press. Vessel. Pip.* **2020**, *186*, 104153. [CrossRef]
27. Yan, J.; Zhou, J.; Zhang, J.; Zhao, P.; Zhang, Z.; Wang, W.; Xuan, F. AP-GAN-DNN based creep fracture life prediction for 7050 aluminum alloy. *Eng. Fract. Mech.* **2024**, *303*, 110096. [CrossRef]
28. Zhang, X.-C.; Gong, J.-G.; Xuan, F.-Z. A deep learning based life prediction method for components under creep, fatigue and creep-fatigue conditions. *Int. J. Fatigue* **2021**, *148*, 106236. [CrossRef]
29. Guan, X.; Burton, H. Bias-variance tradeoff in machine learning: Theoretical formulation and implications to structural engineering applications. *Structures* **2022**, *46*, 17–30. [CrossRef]
30. Ranglani, H. Empirical Analysis of the Bias-Variance Tradeoff Across Machine Learning Models. *Mach. Learn. Appl. Int. J.* **2024**, *11*, 1–12. [CrossRef]
31. Shwartz-Ziv, R.; Armon, A. Tabular data: Deep learning is not all you need. *Inf. Fusion* **2022**, *81*, 84–90. [CrossRef]
32. Zhang, Z.; Wang, X.; Li, Z.; Xia, X.; Chen, Y.; Zhang, T.; Zhang, H.; Yang, Z.; Zhang, X.; Gong, J. Machine learning-assisted probabilistic creep life assessment for high-temperature superheater outlet header considering material uncertainty. *Int. J. Press. Vessel. Pip.* **2024**, *209*, 105211. [CrossRef]
33. Zare, A.; Hosseini, R.K. A breakthrough in creep lifetime prediction: Leveraging machine learning and service data. *Scr. Mater.* **2024**, *245*, 116037. [CrossRef]
34. Liu, Y.; Wu, J.; Wang, Z.; Lu, X.-G.; Avdeev, M.; Shi, S.; Wang, C.; Yu, T. Predicting creep rupture life of Ni-based single crystal superalloys using divide-and-conquer approach based machine learning. *Acta Mater.* **2020**, *195*, 454–467. [CrossRef]
35. Chu, L.; Chen, X.; Fan, Z.; Zhou, Y.; Wu, Z.; Cui, Q. Characterization of heterogeneous creep deformation in vanadium-modified 2.25 Cr1Mo steel weldments by digital image correlation. *Mater. Sci. Eng. A* **2021**, *816*, 141350. [CrossRef]
36. Chai, M.; Li, H.; Tang, Z.; Lai, C.; Song, Y.; Zhang, Z.; Duan, Q. In situ monitoring of high-temperature creep damage in CrMoV high-strength structural steel using acoustic emission. *Constr. Build. Mater.* **2024**, *446*, 137996. [CrossRef]
37. Li, Y.; Huang, S.; Hui, H.; Han, Y.; Huang, J.; Chen, Z. Revealing influence of warm bending on creep performance of heat-resistant 2.25 Cr-1Mo-0.25 V steel. *J. Mater. Sci.* **2023**, *58*, 3802–3824. [CrossRef]
38. Irmak, F.; Gordon, A.P. A framework for life prediction of 2.25 Cr-1Mo under creep and thermomechanical fatigue. In *Turbo Expo: Power for Land, Sea, and Air*; American Society of Mechanical Engineers: New York, NY, USA, 2018; Volume 51128, p. V006T24A020. [CrossRef]
39. Sawada, K.; Kimura, K.; Abe, F.; Taniuchi, Y.; Sekido, K.; Nojima, T.; Ohba, T.; Kushima, H.; Miyazaki, H.; Hongo, H. Catalog of NIMS creep data sheets. *Sci. Technol. Adv. Mater.* **2019**, *20*, 1131–1149. [CrossRef] [PubMed]
40. *ASTM E139-11*; Standard Test Methods for Conducting Creep, Creep-Rupture, and Stress-Rupture Tests of Metallic Materials. ASTM: West Conshohocken, PA, USA, 2018.
41. Chen, K.; Xu, Y.; Song, S. Effect of impurity antimony on the creep behavior of 2.25 Cr-1Mo heat-resistant steel. *Results Phys.* **2019**, *13*, 102208. [CrossRef]
42. Song, S.-H.; Wu, J.; Wei, X.-J.; Kumar, D.; Liu, S.-J.; Weng, L.-Q. Creep property evaluation of a 2.25 Cr–1Mo low alloy steel. *Mater. Sci. Eng. A* **2010**, *527*, 2398–2403. [CrossRef]
43. Chen, J.; Liu, H.; Pan, Z.; Shi, K.; Zhang, H.; Li, J. Carbide evolution and service life of simulated post weld heat treated 2.25 Cr–1Mo steel. *Mater. Sci. Eng. A* **2015**, *622*, 153–159. [CrossRef]
44. Chaudhuri, S.; Ghosh, R. Creep behavior of 2.25 Cr1Mo steel—Effects of thermal ageing and pre-strain. *Mater. Sci. Eng. A* **2009**, *510*, 136–141. [CrossRef]
45. Kvapilová, M.; Ohanková, M.; Král, P.; Dvořák, J.; Kuchařová, K.; Čmakal, J.; Sklenička, V. Characterization of creep properties and the microstructure of a service-exposed low alloy CrMoV steel steam pipe. *Mater. Sci. Eng. A* **2022**, *853*, 143684. [CrossRef]

46. Goyal, S.; Laha, K.; Panneer Selvi, S.; Mathew, M. Mechanistic approach for prediction of creep deformation, damage and rupture life of different Cr–Mo ferritic steels. *Mater. High Temp.* **2014**, *31*, 211–220. [CrossRef]
47. Song, Y.; Chai, M.; Han, Z.; Liu, P. High-temperature tensile and creep behavior in a CrMoV steel and weld metal. *Materials* **2022**, *15*, 109. [CrossRef] [PubMed]
48. Song, Y.; Chai, M.; Lv, J.; Han, Z.; Liu, P.; Yan, H.; Sha, Z. Creep rupture behavior of 2.25 Cr1Mo0. 25V steel and weld for hydrogenation reactors under different stress levels. *Rev. Adv. Mater. Sci.* **2022**, *61*, 334–349. [CrossRef]
49. Zhang, Z.; Liu, H.; Yang, L.; Kong, L.; Zhang, H.; Li, J. Exploring carbide evolution in bainitic 2.25 Cr1Mo steel subjected to creep rupture tests. *Metall. Mater. Trans. A* **2024**, *55*, 4175–4189. [CrossRef]
50. Chu, D.-J.; Kim, H.-Y.; Lee, J.; Jung, W.-S. Investigation of precipitation sequence during creep in 2.25 Cr-1Mo steel. *Mater. Charact.* **2020**, *164*, 110328. [CrossRef]
51. Sujon, K.M.; Hassan, R.B.; Towshi, Z.T.; Othman, M.A.; Samad, M.A.; Choi, K. When to use standardization and normalization: Empirical evidence from machine learning models and XAI. *IEEE Access* **2024**, *12*, 135300–135314. [CrossRef]
52. D'Agostino, R.B. Transformation to normality of the null distribution of g 1. *Biometrika* **1970**, *57*, 679–681. [CrossRef]
53. Chai, M.; He, Y.; Wang, J.; Wu, Z.; Lei, B. Predicting creep life of CrMo pressure vessel steel using machine learning models with optimal feature subset selection. *Int. J. Press. Vessel. Pip.* **2024**, *212*, 105349. [CrossRef]
54. Hope, T.M. Linear Regression. In *Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 67–81. [CrossRef]
55. Iqbal, M.; Salami, B.A.; Khan, M.A.; Jalal, F.E.; Jamal, A.; Bardhan, A. Computational approach towards shear strength prediction of squat RC walls implementing ensemble and hybrid SVR paradigms. *Mater. Today Commun.* **2024**, *40*, 109921. [CrossRef]
56. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
57. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
58. 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, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
59. Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C.J.; Müller, K.-R. Explaining deep neural networks and beyond: A review of methods and applications. *Proc. IEEE* **2021**, *109*, 247–278. [CrossRef]
60. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*; Proceedings of Machine Learning Research: New York, NY, USA, 2016; Available online: https://proceedings.mlr.press/v48/gal16.html (accessed on 2 February 2026).
61. Kiranyaz, S.; Avci, O.; Abdeljaber, O.; Ince, T.; Gabbouj, M.; Inman, D.J. 1D convolutional neural networks and applications: A survey. *Mech. Syst. Signal Process.* **2021**, *151*, 107398. [CrossRef]
62. Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. D. Nonlinear Phenom.* **2020**, *404*, 132306. [CrossRef]
63. Jaradat, Y.M.; Alia, M.A.; Masoud, M.Z.; Manasrah, A.A.; Jannoud, I.A.; Alheyasat, O. Beyond One-Size-Fits-All: Comparing and Selecting Regression Metrics for Robust Model Assessment. In *Proceedings of the 2025 12th International Conference on Information Technology (ICIT)*; IEEE: Piscataway, NJ, USA, 2025; pp. 416–422. [CrossRef]
64. Wen, J.-F.; Tu, S.-T.; Xuan, F.-Z.; Zhang, X.-W.; Gao, X.-L. Effects of stress level and stress state on creep ductility: Evaluation of different models. *J. Mater. Sci. Technol.* **2016**, *32*, 695–704. [CrossRef]
65. Dorn, J.E. Some fundamental experiments on high temperature creep. *J. Mech. Phys. Solids* **1955**, *3*, 85–116. [CrossRef]
66. Speich, G.; Dabkowski, D.; Porter, L. Strength and toughness of Fe-10Ni alloys containing C, Cr, Mo, and Co. *Metall. Trans.* **1973**, *4*, 303–315. [CrossRef]
67. Dargusch, M.; Dunlop, G.; Bowles, A.; Pettersen, K.; Bakke, P. The effect of silicon content on the microstructure and creep behavior in die-cast magnesium AS alloys. *Metall. Mater. Trans. A* **2004**, *35*, 1905–1909. [CrossRef]
68. Wang, C.; Sheng, N.; Sun, S.; Wang, H.; Fan, S.; Sang, Z.; Shi, K.; Hou, G.; Li, J.; Zhou, Y. Addition of Nitrogen in Ni-based Cast Superalloys: Microstructures and Properties. *Mater. Sci. Eng. A* **2025**, *944*, 148907. [CrossRef]
69. Abe, F. Effect of quenching, tempering, and cold rolling on creep deformation behavior of a tempered martensitic 9Cr-1W steel. *Metall. Mater. Trans. A* **2003**, *34*, 913–925. [CrossRef]
70. Dudova, N.; Mishnev, R.; Fedoseeva, A.; Kaibyshev, R. On the effect of tempering temperature on the long-term creep behavior of a 10% Cr steel with low nitrogen and high boron contents. *Mater. Sci. Eng. A* **2024**, *890*, 145912. [CrossRef]
71. Chen, J.; Liu, X.; Zhu, L.; Fan, P.; Chen, H.; Xie, Y.; Yue, L. A study of creep rupture life prediction for P91 steel with machine learning method: Model selection and sensitivity analysis. *Int. J. Press. Vessel. Pip.* **2025**, *216*, 105494. [CrossRef]
72. Al-Selwi, S.M.; Hassan, M.F.; Abdulkadir, S.J.; Muneer, A. LSTM inefficiency in long-term dependencies regression problems. *J. Adv. Res. Appl. Sci. Eng. Technol.* **2023**, *30*, 16–31. [CrossRef]
73. Li, X.; Yu, C.; Su, F.; Quan, T.; Yang, X. Novel training algorithms for long short-term memory neural network. *IET Signal Process.* **2019**, *13*, 304–308. [CrossRef]
74. Wang, X.; Meng, X.; Huang, H.; Lin, Y.; Yuan, Z.; Wan, X.; Zheng, X.; Hu, P.; Zheng, F.; Li, L. Interpretable machine learning models for predicting corrosion behaviour in Titanium alloys. *Corros. Eng. Sci. Technol.* **2025**, *61*, 25–37. [CrossRef]

75. Kölzow, F. Application of Probabilistic Methods for Lifetime Prediction of High Temperature Components Under Creep-Fatigue Loading. Doctoral Dissertation, Technische Universität Darmstadt, Darmstadt, Germany, 2021. [CrossRef]
76. Mehta, S.; Cruse, T.; Mahadevan, S. Confidence bounds on structural reliability. In Proceedings of the 34th Structures, Structural Dynamics and Materials Conference, La Jolla, CA, USA, 19–22 April 1993; p. 1377. [CrossRef]
77. Rasky, D.; Kolodziej, P.; Newfield, M.; Laub, B.; Chen, Y. Assessing factors of safety, margins of safety, and reliability of thermal protection systems. In Proceedings of the 36th AIAA Thermophysics Conference, Orlando, FL, USA, 23–26 June 2003; p. 4043. [CrossRef]
78. Katzir, L.; Elidan, G.; El-Yaniv, R. Net-dnf: Effective deep modeling of tabular data. In Proceedings of the International Conference on Learning Representations, Virtual, 26 April–1 May 2020; Available online: https://openreview.net/forum?id=73WTGs96kho (accessed on 5 February 2026).
79. Fanariotis, A.; Orphanoudakis, T.; Kotrotsios, K.; Fotopoulos, V.; Keramidas, G.; Karkazis, P. Power efficient machine learning models deployment on edge IoT devices. *Sensors* **2023**, *23*, 1595. [CrossRef] [PubMed]
80. Hu, X.; Chu, L.; Pei, J.; Liu, W.; Bian, J. Model complexity of deep learning: A survey. *Knowl. Inf. Syst.* **2021**, *63*, 2585–2619. [CrossRef]
81. Li, L.; Chang, J.; Vakanski, A.; Wang, Y.; Yao, T.; Xian, M. Uncertainty quantification in multivariable regression for material property prediction with Bayesian neural networks. *Sci. Rep.* **2024**, *14*, 10543. [CrossRef] [PubMed]

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