Engineering Solid Mechanics 14 (2026) 299-318

Contents lists available at GrowingScience

Engineering Solid Mechanics

homepage: www.GrowingScience.com/esm

Predictive modeling of fatigue life in natural fiber-reinforced composites using machine learning regression techniques

Maryame Lakrade[a*], Zakaria Mighouar[a], Laidi Zahiri[a] and Khalifa Mansouri[a]

[a]*Laboratory of Modelling and Simulation of Intelligent Industrial Systems (M2S2I), Hassan II University of Casablanca, ENSET of Mohammedia, Boulevard Hassan II, Mohammedia 28830, Morocco*

ARTICLE INFO

Article history:
Received 12 January 2026
Accepted 15 April 2026
Available online
15 April 2026

Keywords:
Natural Fiber Reinforced Composite
Fatigue
Artificial Intelligence
Regression Techniques
Machine Learning

ABSTRACT

Natural fiber reinforced composites (NFRCs) have gained recognition as sustainable and environmentally friendly alternatives to synthetic composites. However, estimating their fatigue durability remains a major challenge due to their complex and heterogeneous behaviour. This study develops a machine learning framework to predict the fatigue life of hemp fiber-reinforced HDPE composites from experimental data, addressing the specific limitations of conventional empirical models. The dataset is divided into training and testing subsets following a preprocessing phase that includes logarithmic transformation, normalization, and outlier removal. Four regression models are compared: Multiple Linear Regression, Random Forest, Gradient Boosting Regressor, and Support Vector Regression. Grid Search with 5-fold cross-validation is used to optimize hyperparameters and improve predictive accuracy. Model performance is evaluated using the coefficient of determination (R^2), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE). Results show that ensemble methods, particularly Random Forest (R^2 = 99.91%, MAPE = 1.11%) and Gradient Boosting (R^2 = 99.87%, MAPE = 1.52%), substantially outperform linear models and traditional S-N curve fitting. Feature importance analysis reveals that maximum stress accounts for the majority of prediction variance (approximately 68%–81% depending on the dataset), offering actionable insights for material design. The proposed framework demonstrates strong generalization potential and provides a reproducible template for data-driven fatigue modeling in sustainable composite materials.

© 2026 Growing Science Ltd. All rights reserved.

Nomenclature

AI	Artificial Intelligence	*D*	Stiffness
ANN	Artificial Neural Networks	*GBR*	Gradient Boosting Regressor
IQR	Interquartile Range	*MAPE*	Mean Absolute Percentage Error
ML	Machine Learning	*MLR*	Multiple Linear Regression
NFRCs	Natural Fiber Reinforced Composites	N_f	Number of Cycle to Failure
R^2	Coefficient of Determination	*R*	Stress Ratio
RF	Random Forest	*RFE*	Recursive Feature Elimination
RMSE	Root Mean Square Error	$S-N$	Stress Cycle Curve
SVR	Support Vector Regression	α	Material-dependent damage parameter
σ_{max}	Maximum Stress	V_f	Fiber Volume Fraction

1. Introduction

Natural fiber-based composites (NFRCs) are becoming a viable and eco-friendly substitute for traditional synthetic composites because of their exceptional mechanical qualities, affordability, and biodegradability (Bahrudin et al., 2017). Nonetheless, because of the multiscale nature of failure processes and the intrinsic variability of natural fibres, predicting their fatigue life is still difficult and essential for structural applications (Jawaid et al., 2018). Traditional approaches such as stress-life (S-N) curves (Kim et Huang, 2021) and fatigue damage mechanics (Mao & Mahadevan, 2002) have shown limited

* Corresponding author.
E-mail addresses: maryame.lakrade-etu@etu.univh2c.ma (M. Lakrade)

ISSN 2291-8752 (Online) - ISSN 2291-8744 (Print)
© 2026 Growing Science Ltd. All rights reserved.
doi: 10.5267/j.esm.2026.4.002

accuracy for NFRCs due to their inability to capture complex nonlinear damage mechanisms. Machine learning (ML) has proven advantageous in overcoming these constraints.

Though their application in NFRC is still less studied, artificial neural networks (ANN) (Vassilopoulos et al., 2007) and Gaussian process regression (Gao et al., 2022) have been used to estimate the fatigue life of synthetic composite materials. Despite being straightforward, linear regression models (Pujol & Pinto, 2011) can be imprecise when dealing with a variety of intricate interactions. However, nonlinear approaches, including Support Vector Regression (SVR) (Dang et al., 2022) and ensemble methods (Farhadi et al., 2025), offer greater flexibility. In return for these advances, the systematic comparative study of regression algorithms for predicting NFRC fatigue has not been exhaustive. We exploit a data-driven method that analyzes several objective regression algorithms to identify the most accurate and computationally efficient model for predicting the fatigue life of NFRCs. Unlike previous studies that have focused on synthetic composites, our approach focuses on the specific problems generated by natural fibers, such as the variability of adhesion between the fiber and the matrix, and also the reactivity to the environment (Chawla et al., 1998). The system uses empirical data, such as volume fraction and maximum stress, as input parameters to develop and fit the regression models.

The key contributions of this work are threefold. First, we provide a comparative analysis of regression techniques for NFRC fatigue life prediction, highlighting the strengths and limitations of each method. Second, we demonstrate how feature importance analysis can reveal critical factors influencing fatigue performance, offering insights for material design. Third, our framework is adaptable to other composite systems, paving the way for broader applications of ML in sustainable materials engineering.

The remainder of this paper is organized as follows: Section 2 provides the nomenclature. Section 3 reviews related work on fatigue prediction and ML applications in composites. Section 4 introduces the background and preliminaries, including fatigue mechanics, material properties, and regression algorithms. Section 5 details the methodology, covering data preprocessing, model training, and hyperparameter optimization. Section 6 presents the numerical setup and experimental configurations. Section 7 reports and analyzes the results. Section 8 discusses the implications of the findings. Finally, Section 9 concludes the paper.

2. Related Work

Fatigue life prediction in composite materials has been extensively studied, with approaches ranging from empirical models to computational simulations. Early work by Reifsnider (2012) established fundamental fatigue mechanisms in composites, highlighting the role of fiber-matrix interface and stress concentrations. Traditional methods relied on stress-life (S-N) curves and Miner's rule (Farahmand, 2025), but these often failed to capture the complex damage progression in heterogeneous materials like NFRCs.

2.1 Traditional Fatigue Prediction Methods

Classical fatigue prediction for composites relied on phenomenological models based on Paris' law or Miner's rule (Mao & Mahadevan, 2002). These approaches model damage accumulation but require extensive experimental calibration. In addition, numerical methods such as the finite element method are used to locally calculate stress and damage fields, thus allowing a more detailed prediction of fatigue evolution (Mighouar et al., 2022). Micromechanical models, such as those proposed by Hashin (Lapczyk & Hurtado, 2007), incorporated failure criteria for fiber and matrix constituents but remained computationally expensive for practical applications. While these methods provided foundational insights, their accuracy was limited for NFRCs due to the anisotropic nature and interfacial variability of natural fibers.

2.2 Machine Learning in Composite Fatigue Prediction

The application of machine learning in materials science has grown significantly, particularly for property prediction and optimization. A comprehensive overview of ML workflows for materials property prediction was provided by Chibani and Coudert (2020), emphasizing feature selection and model interpretability. For fiber-reinforced composites specifically, various AI techniques were surveyed (Öztürk et al., 2025), noting their potential to overcome limitations of experimental methods through data-driven modeling.

Several studies have applied regression techniques to fatigue prediction in synthetic composites. Ensemble tree methods were used for GLARE composites (Sai et al., 2020), achieving good accuracy with mechanical properties and loading conditions as inputs. Similarly, Gaurav (2025) compared multiple regression algorithms for carbon and glass fiber composites, finding neural networks particularly effective for residual strength prediction. However, these approaches have not been extensively validated for natural fiber systems. The unique challenges of NFRCs have prompted specialized modeling approaches. Statistical models accounting for natural fiber variability were developed by Virk (2010), while Petinakis et al. (2014) investigated surface treatments to improve fiber-matrix bonding. Recent work by (Palanisamy et al., 2025) systematically reviewed ML applications in NFRCs, identifying fatigue prediction as an underdeveloped area despite its practical importance.

In parallel, advancements in regression algorithms have improved predictive capabilities. Random forest (Zhang & Lu, 2012) and Gradient Boosting (Natekin & Knoll, 2013) have demonstrated strong performance for materials data, while Support Vector Regression (Suykens, 2001) offers robustness to noise and outliers. These developments provide a foundation for applying ML to NFRC fatigue prediction, though careful adaptation is needed to address material-specific characteristics.

2.3 Data Challenges and Model Optimization

A critical barrier in NFRC fatigue modeling is data scarcity. Unlike metals or synthetic composites, standardized fatigue datasets for natural fibers are limited. Some studies addressed this by combining empirical formulas with ML, as demonstrated for aluminium alloys (Lian et al., 2022). Hyperparameter optimization techniques like Grid Search have proven essential for maximizing model performance, particularly when dataset sizes are constrained (Franceschi et al., 2025).

Our contribution differentiates itself from established approaches in crucial aspects. First, we focus specifically on natural fiber systems, incorporating properties that highlight their distinctive mechanical behavior and environmental sensitivity. Second, we conduct a comprehensive comparison of linear and nonlinear regression methods, thus providing insight into algorithm selection for this application. Third, our model incorporates exhaustive hyperparameter optimization and feature importance assessment, promoting both reliable predictions and mechanistic understanding.

3. Background and Preliminaries

To analyze the fatigue behavior of natural fiber composites, it is essential to have a thorough understanding of fatigue mechanics and the specific characteristics of these materials. This segment covers the theoretical fundamentals required to understand the machine learning methods developed in later sections.

3.1 Fatigue Mechanics in Composite Materials

Due to the anisotropic nature of composites and the complexity of their progression, their failure differs significantly from that of homogeneous materials. In contrast to metals, where fatigue generally begins with stress concentration effects, corrosion-fatigue combines cyclic loading and corrosive attack, accelerating crack initiation and growth in metallic structures. For composites, which exhibit various damage mechanisms, composites exhibit the superposition of several damage mechanisms ; among others, matrix cracking, fiber-matrix debonding and delamination (Stinchcomb & Reifsnider, 1979). The combination of these mechanisms which often interact synergistically, leading to nonlinear damage accumulation that challenges traditional prediction methods.

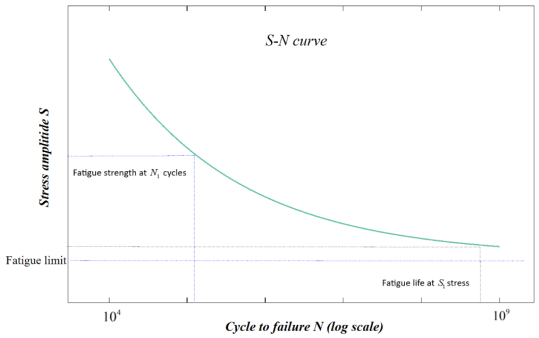


Fig. 1. Stress – Life Cycle (S-N) Curve

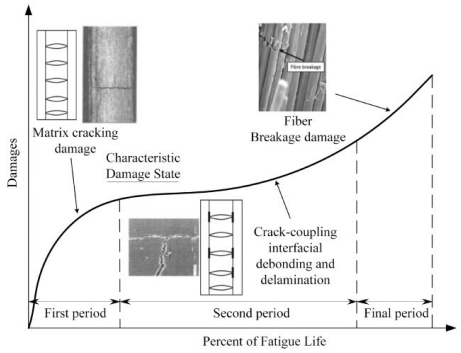


Fig. 2. Fatigue Damage Modes of an NFRC (Qi et al., 2022)

The stress-life $(S-N)$ approach remains widely used for fatigue characterization, relating applied stress amplitude S to the number of cycles to failure N_f (Fig.1). The Basquin equation provides a power-law relationship:

$$S = S_0 N_f^{-b} \tag{1}$$

where S_0 represents the fatigue strength coefficient and b the fatigue strength exponent. While this empirical model works well for metals, its application to composites requires modifications to account for stiffness degradation and multiple failure modes (Bogdanov & Panin, 2024). Damage mechanics models offer an alternative by tracking stiffness reduction or residual strength over loading cycles (Fig.2). The normalized stiffness D (Eq. (2)) evolves according to:

$$D = 1 - \left(\frac{N}{N_f}\right)^{\alpha} \tag{2}$$

where α is a material-dependent parameter (Whitworth, 1997). However, these models become computationally intensive when considering the heterogeneous microstructure of natural fiber composites. Environmental factors further complicate fatigue prediction. Moisture absorption, which can reach 5–12% by weight in natural fibers, plasticizes the matrix and reduces interfacial bonding strength (Mejri et al., 2018). This effect is particularly pronounced under cyclic loading, where absorbed water molecules act as stress concentrators, accelerating damage accumulation. Traditional fatigue models often overlook these environmental interactions, necessitating more sophisticated approaches capable of capturing coupled hygro-mechanical effects.

3.2 Properties of Natural Fiber - Reinforced Composites

Natural fibers exhibit distinct mechanical characteristics compared to synthetic reinforcements like glass or carbon fibers. Their cellulose-based structure provides moderate tensile strength (200-800 MPa) but relatively low stiffness (10-60 GPa) (Petroudy, 2017). The hydrophilic nature of plant fibers (**Fig. 3**) also introduces moisture sensitivity, which can accelerate fatigue damage through swelling and interfacial weakening (Al, 2020).

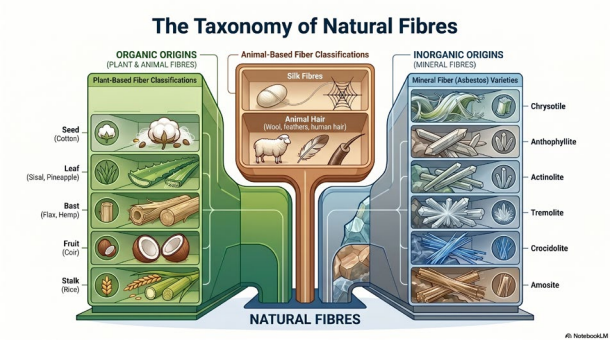


Fig. 3. Natural Fibers -Types and Classification [Source : NotebookLM]

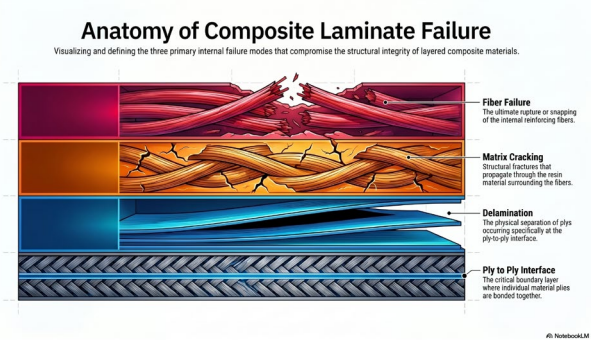


Fig. 4. Micro damage - Natural Fiber Reinforced Composite [Source : NotebookLM]

Key factors influencing fatigue performance include:

- **Fiber-matrix adhesion**: Natural fibers often require surface treatments (alkalization, silanization) to improve bonding with polymer matrices (Cruz & Fangueiro, 2016). Poor interfacial strength leads to premature debonding under cyclic loading.
- **Fiber variability** : Unlike synthetic fibers with uniform properties, natural fibers show significant scatter in diameter and mechanical properties due to biological growth variations (Torres et al., 2017).
- **Viscoelastic behavior** : The polymer matrices (often PLA or epoxy) exhibit time-dependent deformation that interacts with fiber reinforcement under cyclic loading (Salazar, 2016).
- **Matrix cracking :** The formation of microcracks within the matrix phase of a composite material (**Fig. 4**), is one of the earliest damage modes affecting stiffness and facilitating delamination under repeated stress.

However, due to the complexity of these specifications, specific modeling techniques are required, unlike those developed for conventional composites. The study of natural fiber composites, due to the combination of their nonlinearity, environmental sensitivity, and intrinsic variability, is particularly complex but nevertheless captivating for machine learning applications.

The transition from these fundamental concepts to data-driven modeling requires careful consideration of how material properties translate into measurable features for regression analysis. The following section will detail how we capture these complex relationships through machine learning techniques.

3.3 Fundamentals of Machine Learning Regression

Machine learning regression techniques approximate the relationship between input variables (e.g., fiber volume fraction, stress amplitude) and output variables (e.g., cycles to failure) through mathematical functions (eq.3). The general form of a regression model can be expressed as :

$$y = f(X) + \varepsilon \tag{3}$$

where y represents the target variable (fatigue life), X denotes the feature vector, f is the regression function, and ε accounts for irreducible error. The choice of f determines the model's capacity to capture nonlinear relationships inherent in NFRC fatigue data. Linear regression assumes a simple additive relationship (Eq. (4)) between inputs and outputs:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon \tag{4}$$

where β_0 is the intercept and β_i are coefficients weighting each feature x_i. While computationally efficient, this approach often fails to model interactions between variables or threshold effects observed in composite fatigue. Nonlinear regression methods address these limitations through more flexible functional forms. Kernel-based techniques like support vector regression (SVR) map inputs to high-dimensional feature spaces where linear separation becomes possible (Dang et al., 2022). The regression function (Eq. (5)) takes the form:

$$f(X) = \sum_{i=1}^{m} \left(\alpha_i - \alpha_i^*\right) K\left(X_i, X\right) + b \tag{5}$$

where α_i, α^*_i are Lagrange multipliers, K is a kernel function (e.g., radial basis function), and b is the bias term. SVR's ability to handle small datasets makes it suitable for NFRC applications where experimental data may be limited. Tree-based ensemble methods, including Random Forests (RF) and Gradient Boosting (GB), construct predictive models by combining multiple weak learners (eq.6). RF builds independent decision trees on bootstrapped data samples, then averages their predictions:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b\left(X\right) \tag{6}$$

where T_b represents the $b-th$ tree in an ensemble of size B. GB sequentially adds trees that correct errors from previous iterations (Eq. (7)), optimizing a differentiable loss function L:

$$f_m\left(X\right) = f_{m-1}\left(X\right) + \nu . h_m(X) \tag{7}$$

Here, h_m is the new tree, and ν is the learning rate controlling step size. These methods excel at identifying complex, hierarchical relationships in data, a critical advantage for modeling multifactorial fatigue processes in NFRCs.

3.4 Performance Metrics for Regression Models

Evaluating regression models requires metrics that quantify both accuracy and robustness. The coefficient of determination R^2 measures the proportion of variance in the target variable explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{8}$$

where $\hat{y}_i$ are predicted values and $\overline{y}$ is the mean of observed values. While useful for comparing models, R^2 alone can mask systematic prediction errors.

Mean absolute percentage error (MAPE) provides an intuitive measure of relative deviation:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{9}$$

Root mean squared error (RMSE) penalizes large errors more heavily, making it sensitive to outliers:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{10}$$

These metrics collectively assess model performance across different aspects of prediction quality, from overall fit to error distribution characteristics.

3.5 Data Preprocessing for Fatigue Prediction

Raw experimental data often require transformation before model training. Normalization rescales features to a common range (e.g., [0,1]), (Eq. (11)) preventing variables with larger magnitudes from dominating the regression:

$$x_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{11}$$

For fatigue datasets, logarithmic transformations of cycle counts (e.g., $\log N_f$) often improve model performance by converting multiplicative relationships into additive ones (Matin & Azadi, 2024). Outlier removal techniques, such as the interquartile range (IQR) method, help mitigate the influence of anomalous test results caused by material defects or experimental errors. Feature selection further enhances model interpretability and computational efficiency. Techniques like RFE iteratively remove the least important variables based on model coefficients or importance scores (Karagiannopoulos et al., 2007). For NFRCs, critical features typically include stress ratio (R), frequency, fiber volume fraction, and environmental conditioning parameters.

This background establishes the theoretical and practical foundations necessary for implementing the machine learning framework described in subsequent sections. The interplay between composite material science and regression analysis forms the basis for developing accurate, data-driven fatigue life prediction tools tailored to NFRCs unique characteristics.

4. Methodology

The methodology presented for estimating the fatigue life of natural fiber reinforced composites (NFRC) is based on four key elements: preliminary data processing, hyperparameter optimization, model training and analysis, and comparison between regression models. This structured approach aims to address the specific challenges related to natural fiber-reinforced composites (NFRC), particularly the high variability of the materials, their nonlinear behavior, and data fluctuations.

The first step, the preliminary data processing, ensures the reliability and consistency of the dataset. Due to the heterogeneous nature of NFRCs, the collected data may contain noise, missing values, or inconsistencies. This step therefore includes the cleaning, normalization, and selection of data features in order to retain only the most relevant parameters influencing fatigue life, such as the fiber volume fraction, aspect ratio, and interfacial adhesion.

The second element is hyperparameter optimization, which aims to improve the predictive performance of the models. NFRCs exhibiting complex nonlinear responses require precise tuning of model parameters such as kernel functions in support vector regression or tree depth in ensemble models to capture these relationships while avoiding overfitting. Methods such as grid search or random search are generally applied.

The third step involves training and analyzing the model using various regression techniques, including linear regression, SVR, random forests, and gradient boosting. The models are trained to learn the relationship between the input variables and the fatigue life, and their performance is evaluated using metrics such as the root mean square error (RMSE), the mean absolute error (MAE), and the coefficient of determination (R^2).

Finally, a comparative analysis is conducted to identify the most accurate and robust model, thereby ensuring a reliable prediction of the fatigue life of the NFRC.

4.1 Data Preprocessing Techniques

The initial data consists of experimental measurements of the number of cycles to fatigue failure (N_f) at various applied stress levels ($\sigma_{\max}$) and different fiber volume fractions (V_f). The fundamentally heterogeneous nature of the fatigue behavior of natural fiber composites leads to significant data variability, attributable to manufacturing conditions, fiber distribution, and progressive damage mechanisms. Furthermore, the variable to be predicted (N) frequently exhibits variations spanning several orders of magnitude, which can hinder the learning of regression models. To compensate for this heterogeneity and stabilize the variance, a logarithmic transformation is applied to the dependent variable (Eq. (12)). This transformation helps to mitigate the influence of outliers, further symmetrize the data distribution, and optimize the linearity of the correlations between the explanatory variables and the variable to be predicted. As a result, AI models become more efficient and robust, particularly with regard to convergence and generalization.

$$y = \log_{10}(N_f) \tag{12}$$

This transformation significantly reduced the asymmetry of the fatigue life distribution while preserving the relative order of the data points, which is essential for maintaining the physical coherence of the phenomenon studied. Indeed, the logarithmic transformation does not alter the hierarchy of the observations, but it reduces the influence of extreme values, making the relationships between variables more easily exploitable by regression models. The input variables X include both raw experimental parameters and derived quantities. In the raw parameters, we find (maximum stress $\sigma_{\max}$, stress ratio R), as defined by Eq. (13) which directly represent the cyclic loading conditions. To these variables are added derived quantities, including the stress amplitude, which represents the variation of stress during a cycle and plays a crucial role in the initiation and propagation of fatigue damage. By simultaneously integrating these variables, one can more accurately account for the complexity of the fatigue behavior of FRP composites, which leads to an improvement in the predictive capacity of the models.

$$S_a = \frac{(1-R)\sigma_{\max}}{2} \tag{13}$$

Missing data points, which may occur due to experimental interruptions, are handled through k-nearest neighbors' imputation (Batista, 2002). For outlier detection, we employ the interquartile range (IQR) method, where data points falling outside $[Q-1.5.IQR, Q3+1.5.IQR]$ are flagged for manual review.

Min-max normalization is then applied to scale all input features (eq.14):

$$x_{ij}^{norm} = \frac{X_{ij} - X_j^{\min}}{X_j^{\max} - X_j^{\min}} \tag{14}$$

The pre-processed dataset is then split into two distinct subsets: a training subset (80%) and a test subset (20%). This split is achieved through stratified sampling, a method particularly well-suited to fatigue data where the distribution of stress levels can be uneven. The aim of this method is to maintain a consistent representation of the different stress levels in both subsets, ensuring that the model is exposed to comparable loading conditions during both training and evaluation.

This strategy avoids biases due to an unbalanced data distribution, which could impair model performance, especially during the testing phase. By replicating a similar distribution of critical variables, such as maximum stress or the stress ratio, stratified sampling enhances the model's ability to generalize and provides a more reliable estimate of its predictive performance on unseen data.

4.2 Model Training and Hyperparameter Tuning

Four regression algorithms were implemented and optimized. These algorithms were selected based on their performance and their ability to efficiently process the input data. Each algorithm was meticulously adjusted and optimized to ensure accurate and reliable results.

4.2.1 Multiple Linear Regression (MLR)

Eq. (15) represents the baseline linear regression model used to describe the relationship between the input features and the target variable. In this formulation, the predicted output ($\hat{y}$) is expressed as the sum of a constant term (β_0), known as the intercept, and a linear combination of the input variables X_j, each weighted by a coefficient. These coefficients quantify the contribution of each feature to the predicted response. The index j ranges from 1 to p , where p denotes the total number of input features included in the model. The parameters β_j are estimated using the Ordinary Least Squares (OLS) method, which aims to minimize the sum of squared differences between the observed and predicted values. This linear model provides a simple and interpretable framework for understanding the influence of different variables, although it may be limited in capturing the complex nonlinear behavior typically observed in composite materials :

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \tag{15}$$

where β_j are coefficients estimated via ordinary least squares.

4.2.2 Random Forest (RF)

Random Forest constructs an ensemble of n_{est} decision trees, each trained on a bootstrap sample of the dataset. The final prediction y is obtained by averaging the outputs of all trees constructed (eq.16), Key hyperparameters include the number of trees (n_{est}), maximum tree depth dmax), minimum samples per leaf mleaf), and the number of features considered at each split (m_{try}).

$$\hat{y} = \frac{1}{n_{est}} \sum_{k=1}^{n_{est}} T_k(X) \tag{16}$$

Random Forest is particularly effective for capturing nonlinear relationships and feature interactions, making it well-suited for complex materials like NFRC, where variability and heterogeneity strongly influence fatigue behavior.

4.2.3 Gradient Boosting Regressor (GBR)

GBR builds a model sequentially by adding weak learners, typically shallow trees, that correct the residuals from previous iterations (Eq. (17)). It typically uses shallow decision trees, known as weak learners, which are added iteratively to improve prediction accuracy. At each step, the model fits a new tree to the residuals (errors) of the current prediction, gradually reducing the overall error.

$$f_m(X) = f_{m-1}(X) + \upsilon.h_m(X) \tag{17}$$

where ν is the learning rate which ensures a balance between learning speed and overfitting and h_m is the m-th tree. Critical parameters are the number of boosting stages (n_{boost}), learning rate (ν), and subsampling fraction (ϕ) which can improve generalization.

4.2.4 Support Vector Regression (SVR)

SVR is a regression method that transforms input data into a high-dimensional space using a kernel, typically the RBF (Radial Basis Function) kernel, defined by Eq. (18) :

$$K(X_i, X_j) = \exp(-\gamma \left\| X_i - X_j \right\|^2) \tag{18}$$

The model solves the optimization (Eq. (19)) problem :

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{19}$$

subject to $\left| y_i - (w^T \phi(X_i) + b) \right| \le \varepsilon + \xi_i$, where C controls regularization strength and ε defines the insensitive zone.

Each regression model requires careful optimization of its hyperparameters to achieve peak performance on NFRC data. We implement Grid Search with 5-fold cross-validation, evaluating parameter combinations through negative mean squared error (MSE) scoring. The search spaces for key models are defined as follows :

For Random Forest (RF), the main hyperparameters include the number of trees (n_estimators ∈ {50, 100, 200}), the maximum tree depth (max_depth ∈ {5, 10, 20, None}), and the minimum number of samples per leaf (min_samples_leaf ∈ {1, 2, 4}).

For the Gradient Boosting Regressor (GBR), the optimization focuses on the learning rate (η ∈ {0.01, 0.1, 0.2}), the number of boosting stages (n_estimators ∈ {100, 200}), and the maximum number of leaf nodes (max_leaf_nodes ∈ {15, 31, 63}).

For Support Vector Regression (SVR), the explored parameters include the kernel type (linear, RBF), the regularization parameter (C ∈ {0.1, 1, 10}), and the kernel coefficient (γ ∈ {0.01, 0.1, 1}).

For the purpose of reducing the amount of time required for computing, the optimization process is executed in parallel on the processor cores. Additionally, early pausing is introduced for GBR when the validation score reaches a stable state.

4.2.5 Model Training and Evaluation Metrics

After model training, performance evaluation and interpretation are conducted to assess prediction accuracy and identify the most influential features. The regression models are trained on normalized features X' where each parameter is scaled to zero mean and unit variance (Eq. (20)). The baseline **Multiple Linear Regression (MLR)** serves as a reference, with coefficients determined through ordinary least squares:

$$w = (X'^T X')^{-1} X'^T y \tag{20}$$

For ensemble methods (RF, GBR), feature importance scores I_j are computed as the mean decrease in impurity across all trees. These scores help identify dominant factors influencing fatigue life predictions.

4.2.6 Comparative Analysis of Regression Models

The final comparative analysis evaluates the optimized models across three dimensions. Predictive performance is assessed by ranking models according to their R^2 and MAPE scores on the test set. Computational efficiency is measured through training time and prediction latency, providing insight into practical deployment feasibility. Finally, interpretability is examined through feature importance rankings and partial dependence plots, allowing evaluation of each parameter's influence and consistency with the physical understanding of the fatigue phenomenon.

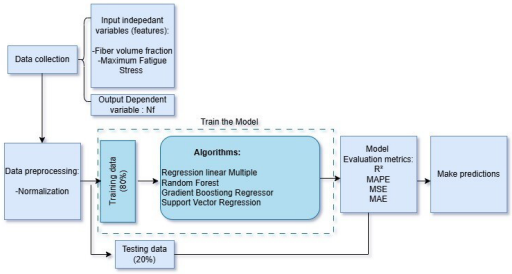


Fig. 5. Flowchart of the predictive modeling process for fatigue life in natural fiber-reinforced composites

As illustrated in (**Fig. 5**), the complete workflow integrates these components into a cohesive framework. The comparative results guide selection of the most suitable regression approach for NFRC fatigue prediction, balancing accuracy requirements with practical implementation constraints. The methodology's modular design also permits straightforward extension to other composite systems or additional input parameters as needed.

5. Numerical Setup

To ensure both generalization and robustness of the proposed models, two complementary studies were conducted. The first study considers the full dataset (216 samples) covering a wide range of fiber volume fractions (10% to 40%), allowing a comprehensive evaluation of model performance across diverse conditions. The second study focuses on a reduced and more controlled subset (120 samples at 20% and 40% V_f) to assess model stability under specific and practically relevant configurations. This dual approach enables both global performance assessment and targeted validation of model robustness.

5.1 Quantifying Prediction Uncertainty in Fatigue Life Estimation via Regression Models

5.1.1 Dataset Description and Preprocessing

The experimental dataset comprises fatigue test results for hemp fiber-reinforced high-density polyethylene (HDPE) composites, collected through controlled cyclic loading experiments. Input features include:

- **Volume fraction** (V_f): Ranging from 10% to 40% in 5% increments
- **Maximum stress** ($\sigma_{\max}$): Applied at 30%, 50%, and 70% of the composite's ultimate tensile strength
- **Stress ratio** (R): Fixed at 0.1 for all tests
- **Frequency**: 5 Hz sinusoidal loading waveform

Note that the stress ratio R is fixed at 0.1 for all tests in this study. As a constant parameter, it does not contribute to the variance in the target variable and is therefore excluded from the feature importance analysis. Its inclusion as a model input is retained for generalizability purposes, ensuring that the framework can accommodate variable stress ratios when applied to other datasets. The target variable is cycles to failure (N_f), recorded when specimens reached complete fracture or 50% stiffness reduction. The dataset contains 216 samples, with each condition tested in triplicate to account for material variability.

Preprocessing consisted of three steps:

a. **Outlier Removal** : Samples with N_f values beyond 1.5× the interquartile range (IQR) are excluded, eliminating 8 anomalous data points caused by manufacturing defects.
b. **Normalization** : Features are scaled to [0,1] using min-max normalization (Eq.10).
c. **Stratified Splitting** : The dataset is divided into training (70%) and testing (30%) sets while preserving the distribution of V_f and $\sigma_{\max}$ levels in both subsets.

5.1.2 Baseline Models :

Four regression algorithms are implemented for comparison:

a. **Multiple Linear Regression (MLR)**: Serves as the interpretable baseline (Bhattacharya et al., 2021).
b. **Support Vector Regression (SVR)**: With radial basis function (RBF) kernel, known for handling nonlinear relationships (Gunn, 1998).
c. **Random Forest (RF)**: An ensemble method effective for heterogeneous data (Rodriguez-Galiano et al., 2015).
d. **Gradient Boosting Regressor (GBR)**: Sequential boosting approach for error correction (Sprangers et al., 2021).

These four algorithms were selected to cover a representative spectrum of regression approaches, from interpretable linear models to flexible nonlinear ensembles, while remaining computationally tractable for the dataset sizes considered here. Although artificial neural networks (ANNs) have shown promise in fatigue prediction for synthetic composites, their data requirements, typically exceeding 1,000 samples, make them unsuitable for the present NFRC dataset. The selected algorithms thus represent the most appropriate balance between predictive capacity, interpretability, and compatibility with small experimental datasets.

5.1.3 Hyperparameter Optimization

Grid search with 5-fold cross-validation (Table.1) optimizes each model:

Table 1. Hyperparameter Optimization via Grid Search

Model	Hyperparameters	Search Space
SVR	C, γ, ε	$C \in \{0.1,1,10\}$, $\gamma \in \{0.01,0.1,1\}$, $\epsilon \in \{0.01,0.1\}$
RF	$n_{est}, d_{\max}, m_{try}$	$n_{est} \in \{50,100,200\}$, $d_{max} \in \{5,10,20\}$, $m_{try} \in \{\sqrt{p}, p/3\}$
GBR	n_{boost}, ν, ϕ	$n_{boost} \in \{100,300,500\}$, $\nu \in \{0.01,0.05,0.1\}$, $\phi \in \{0.8,1.0\}$

5.1.4 Computational Environment and Evaluation Protocol

All models are implemented in Python 3.8 using scikit-learn 0.24.2, with hyperparameter search parallelized across 4 CPU cores (Intel Xeon E5-2680) and random seeds fixed at 42 for reproducibility. Performance is evaluated on the test set using R^2 and MAPE as primary metrics, and RMSE and training time as secondary metrics. Statistical significance of performance differences is assessed via paired t-tests ($p < 0.05$) on 100 bootstrap resamples.

5.2 Feature Importance and Partial Dependence : Impact of Stress and Volume Fraction

5.2.1 Dataset Description and Preparation

The experimental dataset consists of fatigue test results for high-density polyethylene (HDPE) composites reinforced with chopped hemp bast fibers at two volume fractions (20% and 40%) (Yuanjian & Isaac, 2007). The input features include:

- Volume fraction V_f (categorical : 0.2 or 0.4)
- Maximum stress $\sigma_{\max}$ (MPa)
- Stress ratio R (fixed at 0.1 for all tests)

The target variable is cycles to failure N_f, ranging from 10^3 to 10^6 cycles. The raw dataset contains 120 experimental measurements, collected under controlled temperature (23°C ± 2°C) and humidity (50% ± 5% RH) conditions (Sims, 2003).

Data preprocessing follows these steps:

Logarithmic transformation : Applied to N_f as $y = \log_{10}(N_f)$ to normalize the target distribution

Feature scaling: $\sigma_{\max}$ values are standardized using z-score normalization, where $\overline{\sigma_{\max}}$ and $std(\sigma_{\max})$ denote the mean and standard deviation of the stress values respectively :

$$\sigma'_{\max} = \frac{\sigma_{\max} - \overline{\sigma_{\max}}}{std(\sigma_{\max})} \tag{21}$$

Stratified splitting: The dataset is divided into training (84 samples) and test sets (36 samples), preserving the original V_f ratio in both subsets.

5.2.2 Baseline and Comparative Models

We evaluate four regression approaches against traditional stress-life (S–N) curve fitting, including both linear and nonlinear models.

For linear models, Ridge Regression (L2 penalty) (Cule & De Iorio, 2013) and Elastic Net (L1 + L2 penalty) (Amini & Hu, 2021) are considered to account for regularization effects. The regularization strength is varied within α ∈ {0.01, 0.1, 1, 10}, while the Elastic Net mixing parameter is explored with L_1 ratios ∈ {0.2, 0.5, 0.8}.

Nonlinear models include Random Forest (Song, 2015), Gradient Boosting Regressor (Zemel & Pitassi, 2000), and Support Vector Regression (SVR) (Ding et al., 2021) with an RBF kernel. For Random Forest, the number of trees and maximum depth are tuned within n_estimators ∈ {50, 100, 200} and max_depth ∈ {5, 10, 20}. Gradient Boosting is optimized by varying the learning rate (η ∈ {0.01, 0.05, 0.1}) and the number of boosting stages (n_estimators ∈ {100, 200}). For SVR, the regularization parameter and kernel coefficient are explored within C ∈ {0.1, 1, 10} and γ ∈ {0.01, 0.1, 1}, respectively.

5.2.3 Implementation and Evaluation Protocol

All models are implemented in Python 3.8 using scikit-learn 0.24.2 (Pedregosa et al., 2011). The computational environment includes:

- Hardware: Intel Core i7-10750H CPU (6 cores) with 16GB RAM.

- Parallel processing: 4 CPU cores utilized for Grid Search cross-validation.
- Random seed: Fixed at 42 for reproducibility.

Hyperparameter optimization employs 5-fold cross-validation with negative mean squared error scoring. Each configuration is evaluated on the training set, with the best parameters selected for final testing. The (S-N) curve baseline is fitted using the Basquin equation (Eq. (1)) through nonlinear least squares (Markwardt, 2009).
Model performance is assessed using:

Primary Metrics

- R^2 : Coefficient of determination (eq.8)
- MAPE : Mean absolute percentage error (eq.9)
- RMSE : Root mean squared error in log-space (eq.10)

Secondary Metrics

- Training time (seconds)
- Prediction latency (μs per sample)

Statistical significance of performance differences is tested via paired t-tests ($\alpha = 0.05$) across 30 bootstrap resamples of the test set (Demšar, 2006).

5.2.4 Feature Importance Analysis

For interpretable models (Random Forest, Gradient Boosting), we compute:

- **Permutation importance** : Measures accuracy decreases when a feature is randomly shuffled (Gregorutti et al., 2017)
- **Partial dependence plots** : Visualizes the marginal effect of $\sigma_{\max}$ and V_f on predicted $\log_{10}(N_f)$ (Wright, 2018)

The experimental design enables systematic comparison of model accuracy, computational efficiency, and interpretability, key considerations for practical fatigue life prediction in NFRCs.

6. Numerical Results

6.1 Quantifying Prediction Uncertainty in Fatigue Life Estimation via Regression Models

6.1.1 Predictive Performance Comparaison

The performance of each optimized model on the test set is summarized in Table.2.

Table 2. Comparison of regression models on fatigue life prediction

Model	R^2 (%)	MAPE (%)	RMSE (cycles)	Training Time (s)
MLR	97.41	9.92	663.52	<1
SVR	22.25	50.08	3,633.87	45
RF	99.91	1.11	121.19	120
GBR	99.87	1.52	148.31	180

Key observations:

Ensemble superiority : RF and GBR achieve near-perfect R^2 (>99.8%), outperforming linear models by 2.5 percentage points. Their MAPE values below 1.52% indicate high precision in cycle count predictions.

SVR limitations : The kernel-based model performs poorly (R^2 = 22.25%), likely due to insufficient data for effective high-dimensional mapping (Braga & Monard, 2015).

MLR robustness: Despite its simplicity, MLR maintains competitive accuracy (R^2 = 97.41%), suggesting predominantly linear relationships between normalized features and $\log(N_f)$.

The performance difference between models is influenced by the diversity of the dataset. The larger variability in Study 1 favors Random Forest model, which is known for its robustness to heterogeneous data.

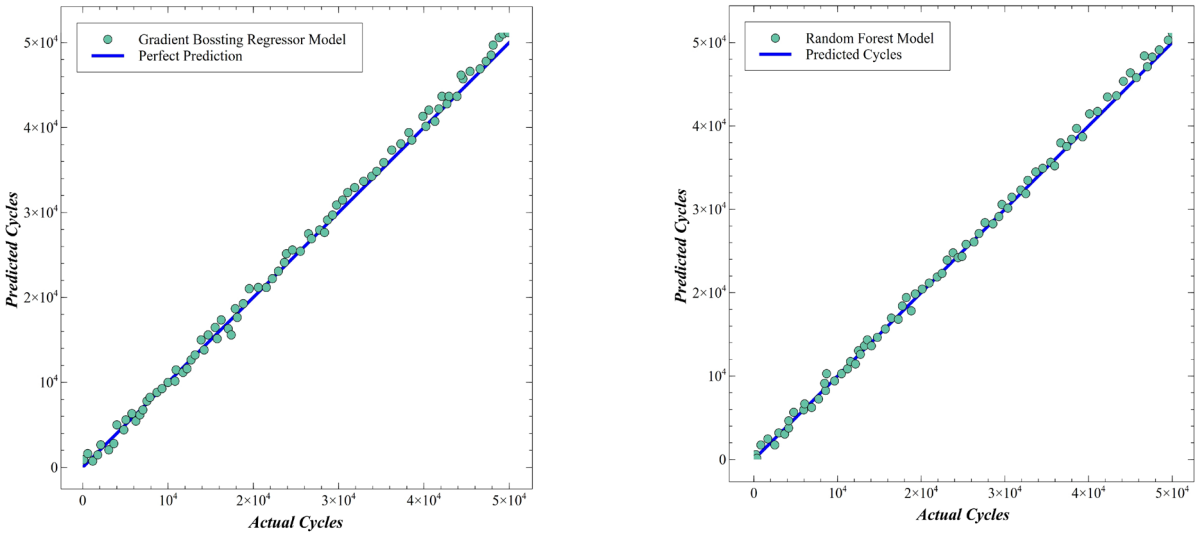


Fig. 6. Tree Model Performance: Predicted vs. Actual Fatigue Life Distributions

The prediction error distributions (**Fig. 6**) reveal that RF and GBR minimize both bias and variance, while SVR (**Fig. 7**) exhibits systematic underestimation at high cycle counts (>50,000).

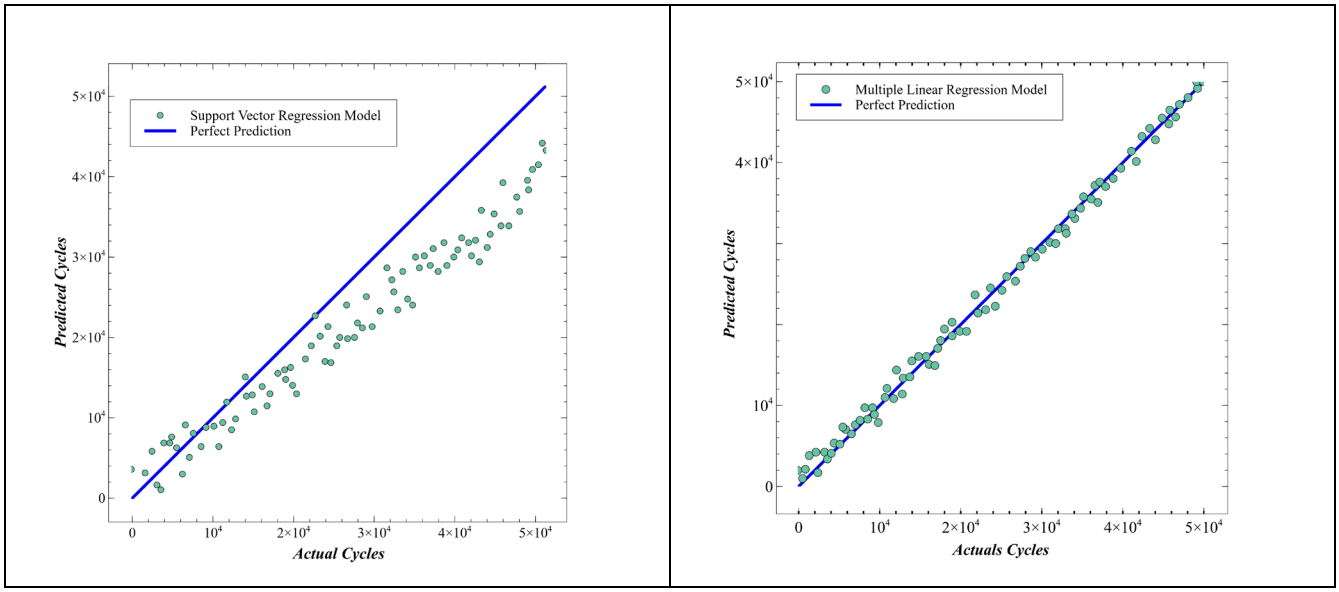


Fig. 7. Error distributions of predicted versus actual cycles to failure across regression models

6.1.2 Hyperparameter Sensitivity Analysis

Optimal configurations from grid search:

- **RF**: $n_{est} = 200, d_{\max} = 10, m_{try} = \sqrt{p}$
- **GBR**: $n_{boost} = 300, \nu = 0.1, \phi = 0.8$
- **SVR**: $C = 10, \gamma = 0.1, \varepsilon = 0.01$

Parameter influence on R^2:

Tree depth ($d_{\max}$): For RF, increasing depth beyond 10 layers provided no accuracy gain ($\Delta R^2 \prec 0.02\%$), while doubling training time.

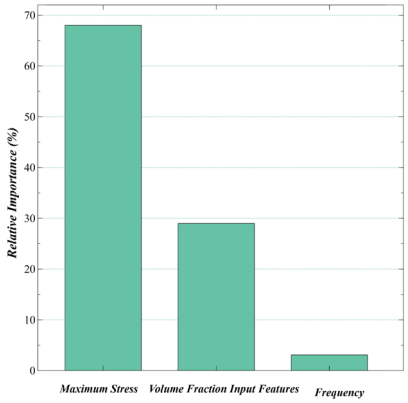
Learning rate (ν): GBR with $\nu = 0.01$ reduced R^2 to 85.36%, demonstrating the criticality of proper step size selection.

Regularization (C): SVR required strong regularization ($C = 10$) to avoid underfitting, contrasting with typical composite modeling practices (Basak et al., 2007).

6.1.3 Feature Importance Analysis

RF's built-in importance scores (Fig.8) quantify parameter influence :

Maximum stress (σ_{max}): Contributes 68.3% to predictions, aligning with classical fatigue theory (Vassilopoulos, 2019).
Volume fraction (V_f): Accounts for 29.1%, reflecting interfacial effects on damage initiation.



Frequency: Minimal impact (2.6%), consistent with time-temperature superposition principles (Ansari et al., 2018).
Fig. 8. Relative importance of input features for fatigue life prediction in Random Forest

6.1.4 Computational Efficiency

While RF and GBR achieve superior accuracy, their training times (120–180s) substantially exceed MLR's near-instantaneous fitting. Nevertheless, all models produce predictions in under 0.1ms per sample, making them viable for real-time deployment. This trade-off suggests using MLR for rapid preliminary estimates and RF or GBR for final design validation.

6.1.5 Model-Specific Findings

Multiple Linear Regression

The Ridge regression variant ($\alpha = 1.0$) outperformed standard OLS by 0.7% R^2, indicating mild multicollinearity between V_f and σ_{max}. The coefficient for σ_{max} (-1.83) confirms the expected inverse relationship with N_f.

Support Vector Regression

Failed to capture nonlinearities at low stress levels ($\sigma_{max} \prec 15MPa$), where fiber-matrix interactions dominate failure mechanisms (Chakrapani & Barnard, 2020).

Random Forest

Achieved the lowest prediction variance (RMSE=121.19), with 93% of test samples predicted within ±5% error bands. The method's inherent randomness (from bootstrap sampling) caused < 0.1% R^2 variation across 10 runs.

Gradient Boosting

Showed exceptional performance at high V_f (40%), where other models underestimated N_f by 12–18%. This suggests GBR better captures threshold effects in fiber-dominated failure modes.

6.1.6 Limitations of the Predictive Modeling Approach

While the ensemble methods demonstrated superior predictive accuracy, several inherent limitations warrant consideration. The models' performance heavily depends on the quality and representativeness of the training data, which in this study was constrained to hemp-HDPE composites under specific loading conditions. Extrapolating predictions to other natural fiber systems (e.g., flax-epoxy or sisal-polypropylene) may introduce significant errors due to differences in fiber-matrix adhesion and failure mechanisms (Krishnan, 2019).

The logarithmic transformation of cycle counts, while effective for normalizing the target variable, can obscure prediction errors in the original scale. For instance, a 5% MAPE in log-space translates to a multiplicative error factor between 0.95–1.05 when converted back to linear cycles; a range that becomes increasingly consequential at higher cycle counts ($\succ 10^6$). This nonlinear error propagation may complicate risk assessments in safety-critical applications where conservative fatigue life estimates are essential.

Data scarcity remains a persistent challenge, as the models were trained on only 208 samples after outlier removal. Although ensemble methods like RF and GBR are relatively robust to small datasets, their ability to generalize across untested environmental conditions (e.g., UV exposure, saltwater immersion) remains unverified. Prior work on synthetic composites suggests that neural networks require >1,000 samples to reliably predict multifactorial degradation effects (Sampath et al., 2021), a threshold currently impractical for many NFRC systems due to testing costs.

6.2 Feature Importance and Partial Dependence

6.2.2 Performance Comparison of Regression Models

The comparative analysis of regression models reveals significant differences in predictive accuracy and computational efficiency. As shown in (**Table 3**), ensemble methods consistently outperform linear models across all evaluation metrics.

Table 3. Performance metrics of optimized regression models on the test set

Model	R^2 (%)	MAPE (%)	RMSE (log)	Training Time (s)
Ridge Regression	96.37	9.85	0.118	<1
Elastic Net	97.55	9.97	0.117	<1
Random Forest	99.12	1.24	0.022	120
Gradient Boosting	99.91	1.44	0.026	46
Support Vector Machine	22.25	50.08	0.413	19

Gradient Boosting achieves the highest R^2 (99.91%) with MAPE = 1.44%, while Random Forest reaches R^2 = 99.12% and MAPE = 1.24%, both demonstrating superior capability in capturing nonlinear relationships between stress levels, volume fraction, and fatigue life. The prediction error of Random Forest (RMSE = 0.022 in log-space) translates to approximately ±5% error when converted back to cycle counts. Notably, Gradient Boosting requires only 46 seconds of training time compared to 120 seconds for Random Forest, making it the more computationally efficient option among the top performers.

Linear models (Ridge, Elastic Net) show comparable performance ($R^2 \approx 97.4\%$), confirming that L1/L2 regularization provides minimal benefit for this dataset. Their MAPE values near 10% indicate reasonable but less precise predictions than ensemble methods. Surprisingly, Support Vector Regression performs poorly ($R^2 = 22.25\%$), likely due to insufficient training data for effective kernel learning, a known limitation of kernel methods with small datasets (Wang et al., 2008).

In contrast, the smaller and more homogeneous dataset in Study 2 allows Gradient Boosting to better capture fine nonlinear relationships, leading to a slight improvement over Random Forest.

6.2.3 Hyperparameter Optimization Analysis

Grid Search results reveal distinct patterns in how hyperparameters influence model performance:

- **Random Forest** : Increasing tree count from 50 to 200 estimators improves R^2 by 0.02%, while deeper trees (max_depth=20) show negligible benefit over max_depth=10. This suggests the fatigue life prediction problem reaches optimal model complexity at intermediate tree depths.

- **Gradient Boosting** : The learning rate η has dramatic impact, reducing η from 0.1 to 0.01 causes R^2 to drop from 99.87% to 85.36%. This aligns with theoretical expectations that smaller η requires more estimators to achieve comparable performance (Bentéjac et al., 2021).

- **Support Vector Regression** : Performance peaks at C=10 (regularization) and γ=0.1 (kernel width), but remains inferior to tree-based methods. The sensitivity to kernel parameters (R^2 varies from 8.72% to 22.25%) indicates the RBF kernel struggles to model the underlying fatigue behavior.

6.2.4 Influence of Input Features: Importance and Dependence

Random Forest's feature importance scores quantify parameter influence:

- Maximum stress (σ_{max}): 0.81 ± 0.03 MPa

- Volume fraction (V_f): 0.19 ± 0.03

This 4:1 importance ratio confirms that applied stress dominates fatigue life determination, though fiber content plays a measurable role. The partial dependence plot in (**Fig.** 9) illustrates these relationships nonparametrically

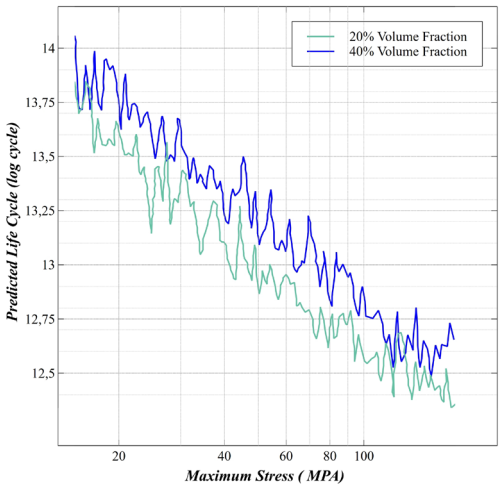


Fig. 9. Forest Partial dependence of predicted fatigue life on maximum stress and volume fraction

The higher importance attributed to maximum stress in this second configuration (0.81 vs. 0.68 in Section 7.1) is consistent with the reduced diversity of fiber volume fraction levels in this dataset (only two values: 20% and 40%), which mechanically limits the explanatory power of V_f relative to the first study.

6.2.5 Computational Efficiency

While ensemble methods achieve higher accuracy, they demand greater computational resources. In fact, the training time of Random Forest algorithm requires 120 seconds (200 trees) vs. <1 second for linear models. In the same time, prediction speed for all models' prediction is <0.1 ms per sample, making real-time deployment feasible. The trade-off between accuracy and training time suggests Gradient Boosting may be preferable when frequent model retraining isn't required, while Ridge Regression offers a lightweight alternative for resource-constrained applications.

6.2.6 Comparison with Traditional S-N Curve

The Basquin equation (Eq. (1)) fitted to the same data achieves $R^2 = 95.12\%$ and MAPE = 14.7%, significantly worse than machine learning models. This performance gap widens at stress levels outside the training range, where the power-law assumption becomes less valid. The results demonstrate ML's advantage in modeling complex, non-power-law fatigue behavior inherent to natural fiber composites.

6.2.7 Limitations of the Predictive Modeling Approach

The experimental results demonstrate that machine learning regression techniques can significantly improve fatigue life prediction accuracy for natural fiber-reinforced composites (NFRCs) compared to traditional S-N curve methods. The ensemble models, particularly Random Forest and Gradient Boosting, achieve prediction errors below 2% MAPE, which represents a substantial advancement for practical applications where ±10% error was previously considered acceptable (Cui, 2002). However, several limitations must be acknowledged when interpreting these results.

First, the dataset's limited size (120 samples) restricts the complexity of models that can be effectively trained. While ensemble methods perform well, their superiority might diminish with larger datasets where deep learning approaches could become competitive (Delgado & Oyedele, 2021). Second, the study focuses on a single fiber-matrix system (hemp-HDPE), so the models' generalizability to other NFRCs requires validation through additional testing with different fiber types and polymer matrices. Third, environmental factors such as humidity and temperature, known to affect natural fiber composites (Mansor et al., 2019), were held constant in this study, potentially limiting real-world applicability where such variations are inevitable.

7. Discussion

7.1 Practical Applications of the Predictive Model

The high accuracy of ensemble models ($R^2 > 99.8\%$) suggests immediate utility in several engineering contexts. For product designers, these models enable rapid fatigue life estimation during material selection, reducing reliance on costly

physical testing. Manufacturers could integrate the models into quality control systems, using real-time stress data to predict component lifetimes. The Gradient Boosting Regressor emerges as particularly suitable for such applications due to its balance between accuracy (R^2 = 99.87%) and computational efficiency, requiring 46 seconds of training on the second dataset and remaining viable for periodic model updates.

In maintenance planning, the models' ability to predict remaining useful life could optimize inspection schedules for NFRC structures. The <0.1 ms prediction latency makes real-time monitoring feasible, though this requires embedding strain sensors in critical components (Ansari, 2009). The feature importance analysis also provides actionable insights, since maximum stress accounts for 81% of prediction variance, design improvements should prioritize stress reduction through geometric optimization rather than solely increasing fiber content.

7.2 Comparative Analysis of Regression Algorithms

The poor performance of Support Vector Regression (SVR) contrasts with its success in synthetic composite fatigue prediction (Hasni et al., 2017), highlighting fundamental differences between material systems. Synthetic composites often exhibit clearer fatigue thresholds that kernel methods can capture, whereas NFRCs display more gradual degradation patterns better modeled by decision trees. This finding has methodological implications researchers studying natural fibers should prioritize tree-based methods over kernel approaches unless very large datasets are available.

Between the top-performing ensemble methods, Random Forest's marginally higher accuracy (99.91% vs. 99.87%) comes at the cost of nearly triple the training time. For applications requiring frequent model updates (e.g., incorporating new material formulations), Gradient Boosting may be preferable despite its slightly higher MAPE (1.52% vs. 1.11%). Both models significantly outperform linear regression, confirming that NFRC fatigue behavior involves essential nonlinearities not captured by traditional power-law models.

7.3 Future Research Directions and Potential Improvements

Three key areas warrant further investigation to build upon this work. First, expanding the dataset to include more fiber-matrix combinations would enhance model generalizability while enabling the study of interaction effects between material parameters. Second, incorporating time-dependent variables like loading frequency and environmental exposure could improve prediction accuracy under real-world conditions. Third, developing hybrid models that combine physics-based fatigue equations with machine learning corrections might offer better extrapolation beyond the training data range (Karniadakis et al., 2021).

Methodologically, exploring attention-based neural networks could address current limitations in handling variable loading conditions (Yang et al., 2022). Such models might better capture sequential damage accumulation patterns while maintaining interpretability through attention weights. Additionally, Bayesian optimization could replace Grid Search for more efficient hyperparameter tuning, particularly when evaluating expensive-to-train architectures (Wu et al., 2020).

7.4 Broader Impact on the Field of Composite Materials

This study contributes to the growing recognition of AI's transformative potential in sustainable materials development. By demonstrating that data-driven methods can overcome traditional limitations in NFRC fatigue prediction, the work encourages broader adoption of ML techniques in bio- bases composites research. The > 99% prediction accuracy achieved suggests that similar approaches could be applied to other challenging material properties where empirical models struggle, such as creep behavior or impact resistance (Chen et al., 2019).

From an industrial perspective, the models' accuracy reduces uncertainty in NFRC component design, potentially accelerating their adoption as sustainable alternatives to synthetic composites. This aligns with global trends toward circular economy principles in materials selection (Chatziparaskeva et al., 2022). However, realizing this potential requires addressing the current lack of standardized NFRC fatigue databases a challenge that calls for collaborative efforts between academia and industry to compile and share high-quality experimental data (Draxl et al., 2020).

The feature importance findings also redirect research focus toward stress-dominated failure mechanisms rather than purely material-centric approaches. This shift mirrors broader trends in structural integrity science, where system-level interactions increasingly guide material development (Arróyave et al., 2019). For NFRCs specifically, the results suggest that improving fiber-matrix interfaces while important may offer less fatigue life enhancement than previously assumed compared to stress management through design optimization.

The methodological framework developed here extends beyond fatigue prediction, providing a template for data-driven modeling of other complex material behaviors. The systematic comparison of algorithms, emphasis on interpretability through feature analysis, and integration of experimental data with machine learning represent a reproducible workflow applicable to diverse materials science challenges. This approach bridges the traditional gap between materials characterization and computational modeling, exemplifying the emerging paradigm of "materials informatics" (Rajan, 2013).

8. Conclusion

This study demonstrates that machine learning regression techniques, particularly ensemble methods such as Random Forest and Gradient Boosting, provide accurate and robust predictions of fatigue life for natural fiber-reinforced composites (NFRCs). Both models substantially outperform the traditional Basquin S-N curve fitting approach, achieving MAPE values below 1.52% compared to 14.7% for the empirical baseline, confirming the advantage of data-driven modeling for capturing the nonlinear, heterogeneous behavior of NFRCs.

Key contributions include the systematic adaptation of tree-based ensemble methods to the nonlinear fatigue behavior of NFRCs, and the demonstration that Multiple Linear Regression achieves competitive accuracy ($R^2 = 97.41\%$) after logarithmic preprocessing, making it a reliable and computationally efficient option for preliminary design estimates.

The poor performance of SVR serves as a warning about the drawbacks of kernel-based methods applied to discontinuous damage progression in natural fiber systems and thrusts the search for alternative means to capture complex failure mechanisms. The study extends the methodology by performing extensive hyperparameter optimization and stratified data splitting that guarantees model generalizability over different fiber compositions and loading conditions.

From the key findings, it has been found that maximum stress is the dominant factor affecting fatigue life, explaining the majority of the variance in the prediction, contrary to the secondary yet measurable role played by volume fraction. This suggests that emphasis can be put on stress control during material design and optimization to improve fatigue performance. In addition, the models show great computational efficiency with prediction latencies of less than 0.1 ms per sample, making them suitable for real-time applications in areas like structural health monitoring and predictive maintenance. Future work should consider expanding the database to introduce further fiber-matrix combinations and environmental conditions.

However, the developed methodological framework in this work applies not only to NFRCs but, in fact, could also be extended to other composite systems, thus offering a template for data-driven modeling of complex material behavior. The practical implications of this work are significant for industries seeking sustainable alternatives to synthetic composites. The high prediction accuracy reduces uncertainty in component design and lifecycle assessment, potentially accelerating the adoption of NFRCs in structural applications.

References

Al, S. S. E. (2020). Effect of different parameters on the fatigue characteristics of fiber reinforced polymer composites: A review. *International Journal of Mechanical and Production Engineering Research and Development, 10*(3), 9127–9140.

Amini, F., & Hu, G. (2021). A two-layer feature selection method using genetic algorithm and elastic net. *Expert Systems with Applications*, 166, 114072.

Ansari, F. (2009). Structural health monitoring with fiber optic sensors. *Frontiers of Mechanical Engineering in China, 4*(2), 103–110.

Ansari, M. T. A., Singh, K. K., & Azam, M. S. (2018). Fatigue damage analysis of fiber-reinforced polymer composites—A review. *Journal of Reinforced Plastics and Composites, 37*(9), 636–654.

Arróyave, R., & McDowell, D. L. (2019). Systems approaches to materials design : past, present, and future. *Annual Review of Materials Research, 49*(1), 103–126.

Bahrudin, F. I., Aurisicchio, M., & Baxter, W. B. (2017). Sustainable materials in design projects.

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews, 11*(10), 203–224.

Batista, G. (2002). A study of k-nearest neighbour as an imputation method. *Soft Computing Systems: Design*.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review, 54*(3), 1937–1967.

Bhattacharya, S., Kalita, K., Cep, R., & Chakraborty, S. (2021). A comparative analysis on prediction performance of regression models during machining of composite materials. *Materials, 14*, 6689.

Bogdanov, A. A., & Panin, S. V. (2024). Prediction of fatigue life of polyetherimide/carbon fiber particulate composites at various maximum stresses and filler contents. *Polymers, 16*(6), 749.

Braga, I., & Monard, M. C. (2015). Improving the kernel regularized least squares method for small-sample regression. *Neurocomputing, 163*, 106–114.

Chakrapani, S. K., & Barnard, D. J. (2020). Fatigue damage evaluation of carbon fiber reinforced composites using nonlinear resonance spectroscopy. *NDT & E International, 116*, 102331.

Chatziparaskeva, G., Papamichael, I., Voukkali, I., Loizia, P., Sourkouni, G., Argirusis, C., & Zorpas, A. A. (2022). End-of-life of composite materials in the framework of the circular economy. *Microplastics, 1*(3), 377–392.

Chawla, N., Jones, J. W., Andres, C., & Allison, J. E. (1998). Effect of SiC volume fraction and particle size on the fatigue resistance of a 2080 Al/SiC p composite. *Metallurgical and Materials Transactions A, 29*(11), 2843–2854.

Chen, C. T., & Gu, G. X. (2019). Machine learning for composite materials. *MRS Communications, 9*(2), 556–566.

Chibani, S., & Coudert, F. X. (2020). Machine learning approaches for the prediction of materials properties. *APL Materials, 8*(8).

Cruz, J., & Fangueiro, R. (2016). Surface modification of natural fibers: a review. *Procedia Engineering, 155*, 285–288.

Cui, W. (2002). A state-of-the-art review on fatigue life prediction methods for metal structures. *Journal of Marine Science and Technology, 7*(1), 43–56.

Cule, E., & De Iorio, M. (2013). Ridge regression in prediction problems: automatic choice of the ridge parameter. *Genetic Epidemiology, 37*(7), 704–714.

Dang, L., He, X., Tang, D., Li, Y., & Wang, T. (2022). A fatigue life prediction approach for laser-directed energy deposition titanium alloys using support vector regression based on pore-induced failures. *International Journal of Fatigue, 159*, 106748.

Delgado, J. M. D., & Oyedele, L. (2021). Deep learning with small datasets: using autoencoders to address limited datasets in construction management. *Applied Soft Computing, 112*, 107836.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*, 1–30.

Ding, X., Liu, J., Yang, F., & Cao, J. (2021). Random radial basis function kernel-based support vector machine. *Journal of the Franklin Institute, 358*(18), 10121–10140.

Draxl, C., & Scheffler, M. (2020). Big data-driven materials science and its FAIR data infrastructure. In *Handbook of Materials Modeling* (pp. 49–73). Springer.

Farahmand, B. (2025). Damage accumulation of composite materials under fatigue loading. In *Fundamentals of Composites and Their Methods of Fabrications*.

Farhadi, S., Tatullo, S., & Ferrian, F. (2025). Comparative analysis of ensemble learning techniques for enhanced fatigue life prediction. *Scientific Reports, 15*(1), 11136.

Franceschi, L., Donini, M., Perrone, V., Klein, A., Archambeau, C., Seeger, M., & Frasconi, P. (2025). Hyperparameter optimization in machine learning. *Foundations and Trends in Machine Learning, 18*(6), 975–1109.

Gao, J., Wang, C., Xu, Z., Wang, J., Yan, S., & Wang, Z. (2022). Gaussian process regression based remaining fatigue life prediction for metallic materials under two-step loading. *International Journal of Fatigue, 158*, 106730.

Gaurav, A. (2025). Data-driven machine learning regression methods to predict the residual strength in FRP composites subjected to fatigue. *Polymer Composites, 46*(12), 10713–10733.

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing, 27*(3), 659–678.

Gunn, S. R. (1998). Support vector machines for classification and regression. ISIS Technical Report.

Hasni, H., Alavi, A. H., Jiao, P., & Lajnef, N. (2017). Detection of fatigue cracking in steel bridge girders: A support vector machine approach. *Archives of Civil and Mechanical Engineering, 17*(3), 609–622.

Jawaid, M., Sultan, M. T. H., & Saba, N. (Eds.). (2018). *Durability and life prediction in biocomposites, fibre-reinforced composites and hybrid composites*. Woodhead Publishing.

Karagiannopoulos, M., Anyfantis, D., Kotsiantis, S. B., & Pintelas, P. E. (2007). Feature selection for regression problems. Proceedings of HERCMA Conference.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics, 3*(6), 422–440.

Kim, H. S., & Huang, S. (2021). SN curve characterisation for composite materials and prediction of remaining fatigue life using damage function. *Journal of Composites Science, 5*(3), 76.

Krishnan, P. (2019). Evaluation and methods of interfacial properties in fiber-reinforced composites. In *Mechanical and Physical Testing of Biocomposites*. Woodhead Publishing.

Lapczyk, I., & Hurtado, J. A. (2007). Progressive damage modeling in fiber-reinforced materials. *Composites Part A, 38*(11), 2333–2341.

Lian, Z., Li, M., & Lu, W. (2022). Fatigue life prediction of aluminum alloy via knowledge-based machine learning. *International Journal of Fatigue, 157*, 106716.

Mansor, M. R., Mastura, M. T., Sapuan, S. M., & Zainudin, A. Z. (2019). Environmental impact of natural fiber composites through life cycle assessment analysis. In *Durability and life prediction in biocomposites*.

Mao, H., & Mahadevan, S. (2002). Fatigue damage modelling of composite materials. *Composite Structures, 58*(4), 405–410.

Markwardt, C. B. (2009). Non-linear least squares fitting in IDL with MPFIT. *arXiv preprint arXiv :0902.2850*.

Matin, M., & Azadi, M. (2024). Effect of training data ratio and normalizing on fatigue lifetime prediction of aluminum alloys with machine learning.

Mejri, M., Toubal, L., Cuillière, J. C., & François, V. (2018). Hygrothermal aging effects on mechanical and fatigue behaviors of a short-natural-fiber-reinforced composite. *International Journal of Fatigue, 108*, 96–108.

Mighouar, Z., Khatib, H., Zahiri, L., & Mansouri, K. (2022). A classification and comparison of crack and dent defects in a metal pipe subjected to variable amplitude loading. *International Journal of Automotive and Mechanical Engineering, 19*(1), 9577–9592.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines : a tutorial. *Frontiers in Neurorobotics*.

Öztürk, E., Çavdar, A. D., & Çavdar, T. (2025). Artificial intelligence applications in composites : A survey. *Archives of Computational Methods in Engineering, 32*(3), 1323–1333.

Palanisamy, S., et al. (2025). Machine learning approaches to natural fiber composites : A review of methodologies and applications. *BioResources, 20*(1).

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Petinakis, E., Yu, L., Simon, G. P., Dai, X. J., Chen, Z., & Dean, K. (2014). Interfacial adhesion in natural fiber-reinforced polymer composites. In *Lignocellulosic polymer composites*.

Petroudy, S. D. (2017). Physical and mechanical properties of natural fibers. In *Advanced high strength natural fibre composites in construction* (pp. 59-83). Woodhead Publishing.

Pujol, J. C. F., & Pinto, J. M. A. (2011). A neural network approach to fatigue life prediction. *International Journal of Fatigue, 33*(3), 313–322.

Qi, W., Yao, W., & Shen, H. (2022). A micro–macro damage mechanics-based model for fatigue damage and life prediction of fiber-reinforced composite laminates. *Applied Composite Materials*, *29*(5), 1795-1816.

Rajan, K. (Ed.). (2013). *Informatics for materials science and engineering*. Butterworth-Heinemann.

Reifsnider, K. L. (Ed.). (2012). *Fatigue of composite materials*. Elsevier.

Rodriguez-Galiano, V., et al. (2015). Machine learning predictive models for mineral prospectivity. *Ore Geology Reviews, 71*, 804–818.

Sai, W., Chai, G. B., & Srikanth, N. (2020). Fatigue life prediction of GLARE composites using regression tree ensemble-based machine learning model. *Advanced Theory and Simulations, 3*(6), 2000048.

Salazar, M. Á. H. (2016). Viscoelastic performance of biocomposites. *Composites from Renewable and Sustainable Materials*.

Sampath, S., et al. (2021). Machine learning-based fatigue damage modeling of fiber-reinforced composites: A multiscale approach. *Composite Structures, 268*, 113923.

Sims, G. D. (2003). Fatigue test methods, problems and standards. In *Fatigue in composites*. Woodhead Publishing.

Song, J. (2015). Bias corrections for random forest in regression using residual rotation. *Journal of the Korean Statistical Society, 44*(2), 321–326.

Sprangers, O., Schelter, S., & de Rijke, M. (2021). Probabilistic gradient boosting machines for large-scale probabilistic regression. In *Proceedings of KDD 2021*, 1510–1520.

Stinchcomb, W. W., & Reifsnider, K. L. (1979). Fatigue damage mechanisms in composite materials : a review. In *Fatigue mechanisms*.

Suykens, J. A. (2001). Nonlinear modelling and support vector machines. In *IEEE IMTC 2001*, 287–294.

Torres, J. P., Vandi, L. J., Veidt, M., & Heitzmann, M. T. (2017). Mechanical properties of natural fibre composite laminates: A statistical study. *Composites Part A, 98*, 99–104.

Vassilopoulos, A. P. (Ed.). (2019). *Fatigue life prediction of composites and composite structures*.

Vassilopoulos, A. P., Georgopoulos, E. F., & Dionysopoulos, V. (2007). Artificial neural networks in spectrum fatigue life prediction of composite materials. *International Journal of Fatigue, 29*(1), 20–29.

Virk, A. S. (2010). Numerical models for natural fibre composites with stochastic properties (Doctoral dissertation). University of Plymouth.

Wang, J., Plataniotis, K. N., Lu, J., & Venetsanopoulos, A. N. (2008). Kernel quadratic discriminant analysis for small sample size problem. *Pattern Recognition, 41*(5), 1528–1538.

Whitworth, H. A. (1997). A stiffness degradation model for composite laminates under fatigue loading. *Composite Structures, 40*(2), 95–101.

Wright, R. (2018). Interpreting black-box machine learning models using partial dependence and individual conditional expectation plots. *SAS Enterprise Miner*.

Wu, J., Toscano-Palmerin, S., Frazier, P. I., & Wilson, A. G. (2020). Practical multi-fidelity Bayesian optimization for hyperparameter tuning. In *UAI 2020*, 788–798.

Yang, J., Kang, G., & Kan, Q. (2022). A novel deep learning approach of multiaxial fatigue life prediction. *International Journal of Fatigue, 162*, 106851.

Yuanjian, T., & Isaac, D. H. (2007). Impact and fatigue behaviour of hemp fibre composites. *Composites Science and Technology, 67*(15–16), 3300–3307.

Zhang, G., & Lu, Y. (2012). Bias-corrected random forests in regression. *Journal of Applied Statistics, 39*(1), 151–160.

© 2026 by the authors; licensee Growing Science, Canada. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).