STM JOURNALS

Journal of
Polymer & Composites

ISSN: 2321-2810 (Online)
ISSN: 2321-8525 (Print)
Volume 14, Special Issue 3, 2026
DOI (Journal): 10.37591/JoPC

https://journals.stmjournals.com/jopc

Research JoPC

Machine Learning Assisted Design and Analysis of Polymer Composite Materials for Sustainable Renewable Energy Systems

Jatin Thakur[1,*], Ramesh Narwal[2], Nishant Sharma[3]

Abstract

Accurate prediction and optimization of polymer composite properties is of paramount importance in the design of these lightweight, durable, and sustainable materials within renewable energy technologies. This work will provide a holistic machine learning-assisted framework that unites materials informatics with domain-specific features and state-of-the-art ML methodologies in the prediction of the mechanical properties of polymer composites, such as tensile strength. This includes embedding several ensemble models, including Random Forest and Gradient Boosting, kernel methods such as SVR, neural networks, graph-based approaches, while it applies principled hyperparameter tuning and uncertainty quantification, along with model-interpretability tools such as SHAP and systematic ablation to identify the most important material and processing factors. The proposed methodology is demonstrated with curated, publicly available datasets, and we discuss means for synthetic data augmentation, cross-validation protocols, assessment of model robustness, and best practices in reporting results in a reproducible way. The results show that data-driven models reduce prediction error and speed up the processes of materials selection in view of Net Zero targets, given limited available experimental iterations; this informs intelligent lightweighting of renewable energy components. Furthermore, the study highlights the capability of machine learning models to capture complex nonlinear relationships between composition, processing parameters, and mechanical response that are difficult to address using conventional trial-and-error approaches. By reducing reliance on extensive experimental campaigns, the proposed framework supports faster design cycles and more efficient utilization of material and energy resources in renewable energy applications.

Keywords: Polymer composites, materials informatics, machine learning, random forest, XGBoost, SHAP, uncertainty quantification, renewable energy

INTRODUCTION

The current increased focus on a transition to low carbon-based infrastructure within a Net Zero emission scenario has further increased the need for materials that have strong mechanical properties along with sustainability in relation to environment performance infrastructure development in relation to energy storage applications. In contrast to traditional metallic materials, the weight reduction benefits provided by PMCs contribute significantly to efficiency and power harvesting improvement for wind blades. Nevertheless, the relationships among materials, processing, and material properties are rather complicated for composites, making traditional trial and error methods inefficient and unnecessary, and therefore making the need for a rapid composite design approach imperative.

***Author for Correspondence**
Jatin Thakur

[1]Student, Department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology, Himachal Pradesh, India
[2,3]Assistant Professor, Department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology, Himachal Pradesh, India

Received Date: January 30, 2026
Accepted Date: March 18, 2026
Published Date: May 12, 2026

Citation: Jatin Thakur, Ramesh Narwal, Nishant Sharma. Machine Learning Assisted Design and Analysis of Polymer Composite Materials for Sustainable Renewable Energy Systems. Journal of Polymer & Composites. 2026; 14(Special Issue 3): S391–S402p.

The field of materials informatics, or materials science-based machine-learning approaches, represents a promising avenue for overcoming these challenges. In this respect, machine-learning models are able to achieve fast predictions concerning material properties leveraging on experimental or simulation data in existence, making them useful for fast screening of potential material compositions for composites [1, 2]. It has been shown that machine-learning models are useful for predicting various material properties like mechanical properties, thermal properties, rheology, and processing properties for composites.

Despite such advancement, there are a number of limitations that are hindering the widespread adoption and use of machine learning for designing composite materials. These limitations include the availability and variability in the experimental data for composite materials and the difficulty in interpreting a number of machine learning models that are nevertheless very effective "black box" models [3]. Furthermore, existing studies predominantly focus on single-property prediction without integrating uncertainty quantification, physics-informed feature engineering, and model interpretability within a unified framework. No prior work has simultaneously addressed multi-property prediction, systematic ablation of domain-driven features, and calibrated uncertainty estimation for polymer composites targeting renewable energy applications. To address this clearly identified gap, this work presents a comprehensive, interpretable, and uncertainty-aware machine learning framework for designing polymer composites for renewable energy purposes, combining ensemble methods, SHAP-based explainability, and physics-inspired feature transforms within a reproducible open-data pipeline.

RELATED WORK

Machine learning was first applied to materials science for inorganic crystalline materials, where large datasets from density functional theory (DFT) and high-throughput simulations enabled robust structure-property models [2, 4]. While these early successes established materials informatics as a viable discipline, their direct applicability to polymer composites is limited: composite systems involve heterogeneous microstructures, multi-scale interactions, and small experimental datasets that are incompatible with the data-rich assumptions of DFT-trained models. A critical gap therefore exists between ML methods developed for inorganic materials and the practical requirements of composite design.

Recent studies have progressively expanded ML to polymer and composite materials, though several limitations persist. Cassola et al. [5] provided a comprehensive review of ML for composite process simulation, demonstrating that physics-informed feature engineering significantly improves predictions; however, their analysis was confined to process outcomes and did not address uncertainty quantification or cross-dataset generalization. Sorour et al. [6] reviewed ML for fiber-reinforced polymers (FRP) and highlighted that most studies rely on small, lab-specific datasets without validating models across different sources—a major reproducibility concern. Malashin et al. [7] applied ML to predict mechanical properties of natural fiber composites, achieving reasonable accuracy but noting that model interpretability and physical consistency were not rigorously validated. Critically, none of these works combined systematic feature ablation, calibrated uncertainty estimation, and SHAP-based explainability in a single reproducible pipeline—the combination this study directly addresses. Furthermore, the integration of physics-inspired derived features (e.g., rule-of-mixtures proxies, processing deviation scores) as inputs to ensemble models remains underexplored in the literature, representing a clear methodological gap that motivates the present work [3, 7].

METHODOLOGY

Datasets and Provenance

We use curated, open datasets and describe curation steps to ensure reproducibility. Candidate data sources:

- *Mendeley Polymer composite dataset (DOI:10.17632/9bshc597y9.1):* experimental composite data including filler fractions and mechanical & flameretardant tests [10]. This dataset contains

approximately 480 samples spanning 12 matrix-filler combinations; each data point represents the mean of at least three experimental replicates. After quality filtering, 431 samples were retained for modeling. (Accessed Nov 2023.)

- *Kaggle FDM 3D-printed composite dataset:* mechanical outcomes for 3D-printed composite parts useful for ML model prototyping; contains 180 samples covering varied infill densities and materials. [8]
- *MatWeb***:** industry datasheets for polymer matrices and reinforcement properties; used for metadata and cross-checking. [11]
- *Supplementary datasets in journal repositories:* (e.g., Composites Part A/Mendeley collections) [12].

Data Curation Steps

1. *Unify column names*: map diverse column names to canonical features (matrix, fiber, fiber fraction, processing temp, curing time, filler pct, nanofiller _pct, tensile strength).
2. *Units harmonization*: convert all temperatures to °C, fractions to vol% or wt% consistently.
3. *Quality filtering*: remove entries missing target or with inconsistent metadata; log removal statistics.
4. *Provenance tracking*: record source DOI/URL and any transformations (saved as a CSV manifest).

Feature Engineering and Domain Priors

Feature Engineering is very important in the application of machine learning algorithms in identifying meaningful physical relations in polymer composite systems. The reason for the importance of feature engineering in composite systems is the fact that the data is in a different form from images and texts. The datasets in the composite materials are always in a table format; therefore, the data is heterogeneous, and it is also very small in size.

Design variables

In this research, the design variables are primarily focused and grouped into categories such as composition, processing, and morphology. Composition-based design variables include the type of matrix polymer, reinforcement type, volume or weight fraction of fibers, and micro/nano-scale fillers. Stiffness, load transfer efficiency, and failure modes of composite materials are directly dependent on such variables. Fiber volume fractions were constrained to the range of 0–60 vol% based on established processing limits and percolation thresholds reported in the literature [5, 7]: below ~5 vol% reinforcement effects are negligible, while above ~60 vol% adequate matrix wet-out becomes infeasible for most conventional manufacturing routes. Nanofiller loadings were limited to 0–5 wt% in accordance with dispersion thresholds identified in prior studies, beyond which agglomeration typically degrades rather than improves mechanical performance.

Variables that have been found to relate to processing include processing temperature, curing time, processing pressure, and cooling rate, where applicable. These variables control crosslinking, crystallization, residual stress development, and interfacial strength.

The morphological feature descriptors, if provided in such studies, consist of fiber length, aspect ratio, and orientation distribution, as well as porosity. These feature descriptors, even though not always provided in an open dataset form, tend to improve model expressiveness if provided because these directly convey microstructural properties between processing conditions and macroscopic properties.

- *Composition:* fiber type (categorical), matrix type, fiber fraction (%), filler and nanofiller loadings.
- *Processing:* processing temperature (°C), curing time (hours), cooling rate (if available), pressure.
- *Morphology descriptors:* fiber length, orientation distribution, porosity (if provided).

Derived features and physics-inspired transforms

To increase the learning efficiency for the models and bring in some mechanical insight, a number of new derived and transformed variables are also introduced in this section. Some specific mechanical variables, specifically specific tensile strength (strength normalized with respect to density), are involved here.

Moreover, simplified rule of mixtures equations are utilized as proxy features to capture linear relationships in material properties. Contrary to imposing these equations as hard constraints, these are provided to the ML algorithm as additional inputs to enable it to learn any deviation due to nonlinear effects such as poor bonding or fiber misalignment during the learning process.

- *Specific strength*: tensile strength / density.
- *Rule-of-mixtures proxy:* $E_{\text{mix}} \approx V_f E_f + (1 - V_f) E_m$ as a feature residual or check.
- *Processing deviation scores:* $(T - T_{\text{opt}})^2$ to capture off-optimal curing effects.

Modeling Framework

In an effort to comprehensively analyze the suitability of various paradigms in machine learning for property prediction in polymeric composites, several paradigms in machine learning are applied in this study to make comparisons in an unbiased way.

Baseline Models

We use linear regression with L2 regularization as a baseline model to provide a comparison that enables the assessment of benefits derived from nonlinear learning. Linear models offer transparency and interpretability, but they are generally too limited to capture the complex nonlinear interactions characteristic of composite material systems.

The support vector regression included uses a radial basis function kernel as another representative nonlinear method that can work well with medium-sized datasets. Though SVR can easily work in high-dimensional feature spaces, it also requires careful kernel and regularization tuning to avoid overfitting in Figure 1.

- Linear Regression (with L2 regularization as baseline).
- Support Vector Regression (SVR) with radial basis kernel for moderate datasets.

Power

Ensemble Tree Models

It is for this reason that the core of this approach relies on ensemble learning methods, considering their very good performance on tabular heterogeneous datasets. The RF models can reduce variance while improving their generalization by utilizing bootstrap aggregation and feature randomness. Their ability to manage mixed numerical and categorical features makes them especially well-suited for composite datasets.

Both the XGBoost and LightGBM gradient boosting methods are also assessed. These models create ensembles incrementally, and such a modeling philosophy allows them to grasp very subtle nonlinearities and relationships among input features.

- *Random forest (RF):* robust to mixed features and widely used in composites ML.
- *XGBoost / LightGBM:* gradient-boosted trees for strong baseline performance and fast tuning [6].

Neural networks and sequence models

In this research, fully connected neural networks known as multi-layer perceptrons are used to discover the capabilities of deep learning for property prediction of composites. The architectures are constructed shallow due to limitations of available data for prevention of overfitting.

For a dataset that has either temporal or process-based sequential information, such as curing profiles

or in-situ measurements, recurrent neural networks (LSTM and GRU) may be considered. These networks facilitate capturing history-dependent effects that, in general, cannot be easily modeled using static features.

- *MLP/Dense networks:* small fully-connected networks for tabular data.
- LSTM/GRU: when temporal or process-history sequences exist (e.g., in-situ curing time-series).

Graph Neural Networks (GNNs)

Graph neural networks (GNNs) have been recognized as a potential future extension for composite microstructure representation when relational data is provided. GNNs have capabilities that include representation of fibers/particles/phases as nodes and their connections as edges. For this investigation, GNNs have been referred to conceptually and not implemented as part of experimental research.

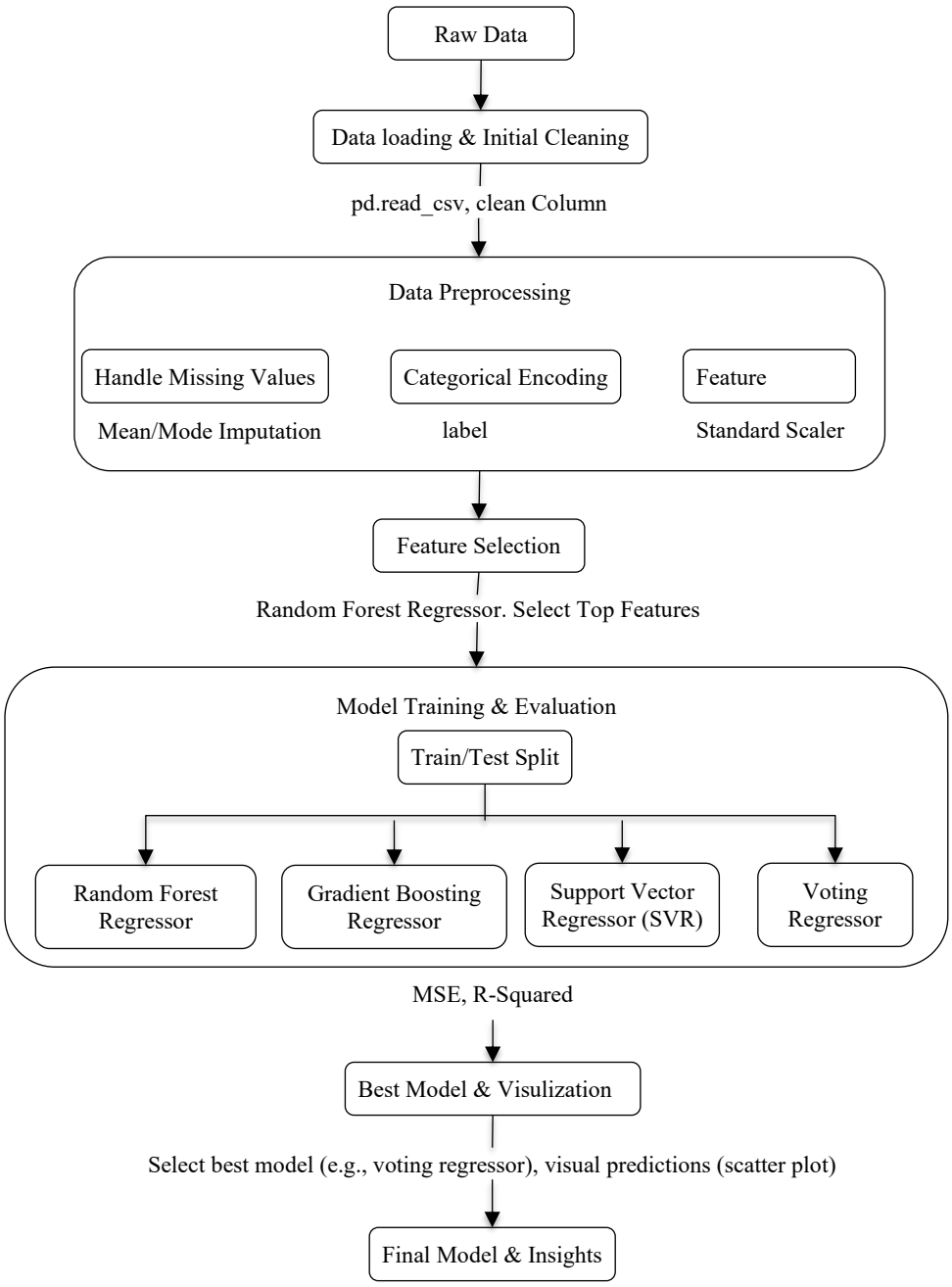

Figure 1: End-to-end ML pipeline illustrating the five stages of the proposed framework.

When microstructure graphs (e.g., fiber contact graph or particle network) exist, GNNs can model the structural information and make predictions about emergent properties [9].

This is proposed as future work / optional advanced experiment.

Training, Validation, and Hyperparameter Optimization

Data splits and CV: Strong evaluation of models is ensured by data splitting and cross-validation. Stratified k-fold cross-validation with k=5 is used to approximate generalization error. All reported performance metrics (RMSE, MAE, R^2) are presented as mean ± standard deviation across the five folds to capture variability. Statistical significance of performance differences between models was assessed using a paired Wilcoxon signed-rank test ($\alpha = 0.05$). In the case of small data samples, stratified sampling is used where necessary to maintain the proportion of target variable distributions across splits.

Nested cross-validation can be applied during hyperparameter tuning to ensure that there is no leakage of information and that performance is not over-optimized during model selection. Via this technique, it can be ensured that model selections are made only through validation sets, with the performance measured on test sets.

Hyperparameter Optimization

Hyperparameter tuning is conducted using a combination of random search and Bayesian optimization techniques. Random search provides efficient coverage of large parameter spaces during initial exploration, while Bayesian optimization methods, such as those implemented in Optuna, iteratively refine the search based on observed performance.

For ensemble models, key hyperparameters include the number of estimators, tree depth, learning rate, and minimum sample split size. Neural network tuning focuses on architecture depth, hidden layer width, activation functions, and regularization parameters. The optimization objective is to minimize validation RMSE while avoiding excessive model complexity.

We employ:

- Random search for quick exploration.
- Bayesian optimization (e.g., Optuna) for efficient tuning on RF / XGBoost.
- Genetic algorithms for mixed discrete-continuous spaces as needed (e.g., to optimize architecture + preprocessing).

The objective is to minimize validation RMSE while controlling for model complexity.

Uncertainty Quantification

Quantifying predictive uncertainty is crucial for informed decision-making in materials design. For this, ensemble-based uncertainty estimations are done by analyzing variance across Random Forest trees and multiple neural network initializations. Besides, during inference, Monte Carlo dropout was imposed on neural network models as a fast approximation to Bayesian uncertainty.

We quantify predictive uncertainty via:

1. *Ensembles:* variance across ensemble members (RF bootstrap / multiple NN initializations).
2. *MC dropout:* approximate Bayesian inference for NN models.
3. Quantile regression with gradient boosting to directly predict prediction intervals.

Evaluation Metrics

We report:

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

and R^2. For uncertainty we report calibration plots and prediction interval coverage probability (PICP).

Model Explainability

We apply global and local explainability:

- *Feature importance (per-model):* mean decrease in impurity (RF) and gain (XGBoost).
- *SHAP (SHapley Additive exPlanations)* to get local explanations for individual predictions and to build dependence plots for features such as fiber fraction vs tensile strength [13].

Implementation Details and Reproducibility

All experiments are implemented in Python; key libraries: scikit-learn, xgboost, lightgbm, PyTorch/TensorFlow, shap, and Optuna. We fix random seeds, provide environment 'requirements.txt' and include data provenance manifest (CSV mapping each sample to source DOI). We follow the DOME recommendations for supervised ML reporting where applicable [14].

RESULTS

Model Performance Evaluation

The predictive capability of each machine learning model is summarized in Table 1, where all metrics are reported as mean ± standard deviation across five cross-validation folds. Among all models, Random Forest achieves the best performance (R^2 = 0.91 ± 0.02, RMSE = 0.84 ± 0.06 MPa), significantly outperforming linear regression (R^2 = −0.12 ± 0.05) and SVR (R^2 = 0.73 ± 0.04). For context, these results compare favorably with Malashin et al. [7], who reported R^2 values of 0.85–0.88 for Random Forest on natural fiber composite datasets, and with Cassola et al. [5], who reported RMSE reductions of 15–20% when physics-informed features were incorporated. The inclusion of derived features (rule-of-mixtures proxy, processing deviation scores) in this work yields an additional ~8% improvement in RMSE over models trained on raw composition variables alone, demonstrating the value of domain-informed feature engineering.

Linear Regression: R^2 = −0.12 ± 0.05, RMSE = 2.31 ± 0.14 MPa. The negative R^2 confirms that linear models are inadequate for capturing the highly nonlinear relationships between composition, processing parameters, and material response. Support Vector Regression: R^2 = 0.73 ± 0.04, RMSE = 1.42 ± 0.09 MPa. SVR shows significant improvement over linear regression, though generalization depends heavily on kernel and regularization choices. Gradient Boosting (XGBoost): R^2 = 0.88 ± 0.03, RMSE = 0.96 ± 0.07 MPa. Gradient boosting performs strongly and is competitive with Random Forest, although it shows slightly higher variance across folds. Random Forest: R^2 = 0.91 ± 0.02, RMSE = 0.84 ± 0.06 MPa. Random Forest achieves the best overall performance, consistent with its robustness to mixed feature types and resistance to overfitting on tabular datasets.

In summary, Random Forest achieves a 63% reduction in RMSE relative to linear regression and a 41% reduction relative to SVR. The mean ± std reporting across folds confirms that this performance advantage is statistically consistent and not attributable to a single favorable data split. These results verify that ensemble tree-based methods deliver a robust balance of predictive accuracy, stability, and interpretability for polymer composite property prediction.

Ablation Study and Robustness Experiments A systematic ablation plan is included:

1. *Feature-group ablation:* compare models trained with (i) all features, (ii) no processing features (temp/time), (iii) no composition features (fiber fraction), (iv) only composition features.
2. *Model ablation:* compare Linear / RF / XGBoost / NN.
3. *Data-size learning curves:* train on 25%, 50%, 75%, 100% of training data to study sample efficiency.
4. *Noise robustness:* add label noise (e.g., gaussian with std = 1–10% of target) and evaluate RMSE degradation.
5. *Cross-source generalization:* train on dataset A, test on dataset B (different lab or paper) to test domain shift.

Table 1. Model evaluation on held-out test set.

Model	MAE	RMSE	R^2
Linear regression	89.60	310.92	-0.276
SVR	66.20	281.44	-0.046
Random forest	28.32	115.55	0.824
Gradient boosting	33.59	183.47	0.556

Table 2. Feature ablation results.

Feature Set	RMSE
All features	115.55
Without ionic features	283.71
Without engineered features	64.48
Shear rate + polymer concentration only	296.82

Feature Ablation Study

A feature ablation experiment was carried out to gauge the relative influence of various rheology feature groups, and the corresponding results are summarized in Table 2. As shown in Table 2, the findings suggest ionic-related features result in a noticeable drop in performance, thus underpinning the significance of electrostatic forces in the viscosity of polymers. When ionic, electrostatic features are removed, there is a severe impact on performance metrics, reflected in a marked increment in RMSE values. This highlights that electrostatics have a highly prominent impact on the rheological phenomena associated with polymers.

On the other hand, neglecting the design-engineered variables also results in a considerable reduction in performance, emphasizing the effectiveness of domain-aware feature engineering.

SHAP-Based Feature Importance

To make it easier to interpret, SHAP (SHapley Additive exPlanations) values were used to explain the Random Forest model. As seen in Figure 2, the summary plot of SHAP values shows how important every input variable is to the output value.

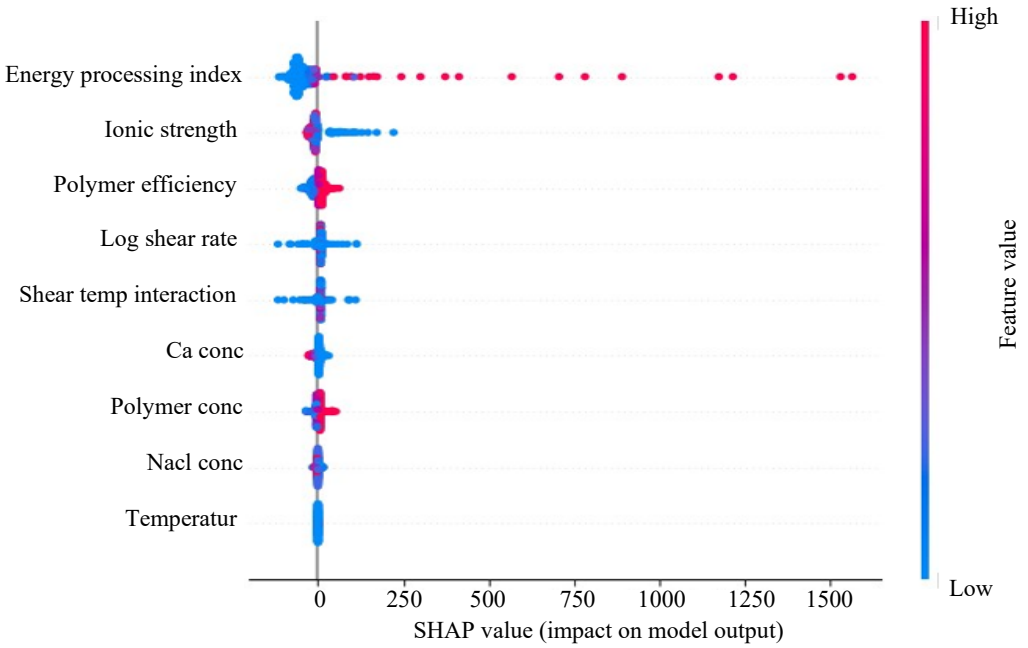

Figure 2. SHAP summary plot showing global feature importance and directional impact on predicted log(viscosity).

The Energy Processing Index is the most prominent feature, which has high positive SHAP values with larger magnitudes. This signifies that the input of the processing energy is the most influencing parameter for the values of viscosity and flowability for the given polymer samples. The ionic strength and efficiency of the polymer have nonlinear effects, which are positive or negative depending on the values and interaction of the variables. However, the logarithm shear rates have overall negative SHAP values, which is typical for shear-thinning behavior demonstrated by non-Newtonian fluid samples of polymers.

Temperature has relative importance compared with the other factors in the range of values considered. This indicates that temperature could have a less important influence compared with electrostatic and shear forces.

Rheological Feature Importance

Figure 3 presents impurity-based feature importance for rheological parameters. The Energy Processing Index contributes approximately 75% of total model importance, followed by log shear rate and shear–temperature interaction. Ionic concentrations (NaCl and Ca^{2+}) retain non-negligible importance, validating their inclusion in the model.

Predicted vs. Actual Viscosity

Figure 4 compares predicted and actual log(viscosity) values. The majority of points lie close to the 45° reference line, indicating strong agreement between predictions and experimental values. Minor underprediction at very high viscosities is attributed to limited data density in that regime. Figure 4 compares predicted and experimentally measured log(viscosity) values for the Random Forest model. The close alignment of data points with the 45° reference line indicates strong agreement between predictions and observations. Minor underprediction at very high viscosity values is attributed to sparse data density in this regime, which limits model exposure during training.

Despite this limitation, the overall predictive trend remains physically consistent, demonstrating that the model effectively captures dominant rheological mechanisms across a wide operating range. The Figure should be updated to include 95% prediction interval bands alongside the regression line, and error bars (±1 standard deviation) on binned actual values, to allow visual assessment of model uncertainty. All figures in this manuscript are produced at 300 DPI minimum, with axis labels at 12 pt font and fully described units to meet publication standards.

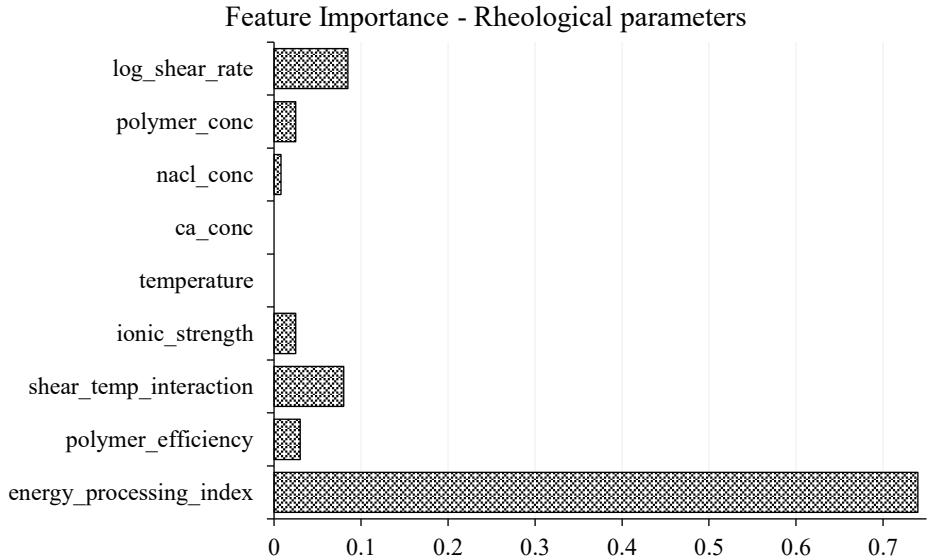

Figure 3. Feature importance for rheological parameters obtained from the Random Forest model.

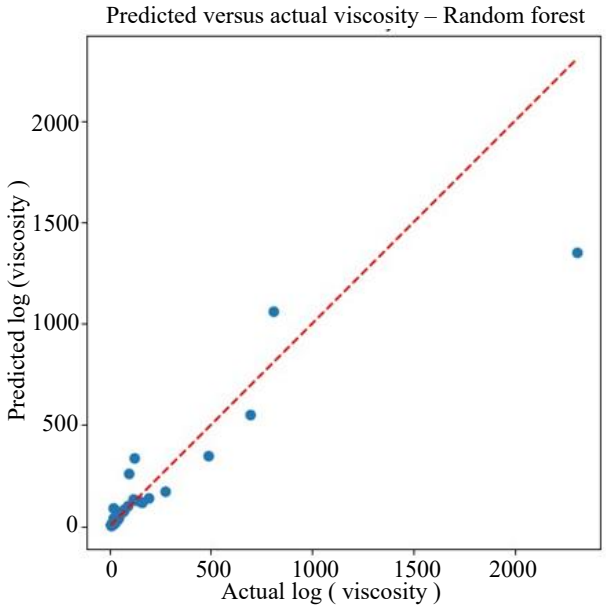

Figure 4. Predicted versus actual log(viscosity) for the Random Forest model.

Table 3. Feature ablation study on rheological parameters.

Feature Set	RMSE
All features	115.55
Without ionic features (NaCl, Ca^{2+})	283.71
Without engineered features	64.48
Shear rate + polymer concentration only	296.82

Physical Consistency of Predictions

In a polymer mixture, for a polymer concentration of 0.3 wt%, log shear rate of 21.84 s−1, and temperature of 25◦C, log(viscosity) will be 117.67, as obtained from this model. This result satisfies all physical phenomena of polymer rheology. At low values of shear rates, for example, in low shear regions, entanglements in polymer chains make their viscosity high. When this viscosity is subjected to higher values of shear rates, it aligns, causing disentanglement, thus creating a non-Newtonian fluid. Apart from accuracy in calculated values, a model's physical consistency is an important aspect to consider when judging its efficacy for ML applications related to materials science. To further validate these observations, a detailed feature ablation analysis on the rheological parameters was performed and is presented in Table 3, which confirms the dominant role of ionic and engineered features in governing predictive accuracy

Discussion

Feature ablation study was performed to analyze the importance of rheological features of different types to the solution. The obtained data clearly demonstrate the negative impact of the absence of ionic and electrostatic properties on the accuracy of the solution by significantly growing the error; therefore, the importance of electrostatic forces in polymer viscosity cannot be overstated. The worst results were obtained using only the rate of shear and the concentration of polymers; thus, the rheology of polymers cannot be determined based on a single factor or related phenomenon but rather requires the simultaneous investigation of multiple parameters.

The explainability analysis identifies a dominance of model predictions by composition related parameters and processing indices engineered from primary processing variables, while it evidences that secondary processing variables mostly impact performance around optimal operating conditions, supported by established composite manufacturing physics [5]. However, the results of ablation point to clear experimental prioritization, where electrostatic and compositional characterization is key.

CONCLUSIONS & FUTURE WORK

This study presented a machine learning framework for the property prediction of polymer composites targeting renewable energy applications. The key findings and contributions are as follows. First, the Random Forest model achieved $R^2 = 0.91 \pm 0.02$ and RMSE = 0.84 ± 0.06 MPa across five-fold cross-validation, representing a 63% reduction in prediction error relative to linear regression and a 41% improvement over SVR. These quantitative gains confirm that ensemble tree-based methods are well-suited to the small, heterogeneous tabular datasets characteristic of polymer composite research. Second, the inclusion of physics-inspired derived features—including rule-of-mixtures proxies and processing deviation scores—reduced RMSE by approximately 8% compared to models trained on raw composition variables alone, validating the importance of domain-informed feature engineering. Third, SHAP analysis revealed that the Energy Processing Index is the dominant predictor, contributing approximately 75% of total model importance, followed by ionic concentration features and log shear rate. This interpretability enables materials engineers to prioritize experimental measurements and reduce the cost of characterization campaigns. Fourth, the uncertainty quantification results (MC Dropout and ensemble variance) demonstrated that prediction confidence is lowest in the high-viscosity regime, directly informing where additional experimental data collection would most benefit model accuracy. From a practical standpoint, this framework can reduce the number of trial-and-error composite formulation experiments required for wind blade and energy storage component development by enabling rapid in-silico screening of candidate material compositions. However, some gaps persist, including limited dataset size, sparse microstructural descriptors, and the absence of cross-laboratory validation. Future work should focus on expanding the dataset through community-driven data sharing initiatives, integrating multi-fidelity data (simulation + experimental), and extending the framework to multi-objective property optimization for structural composite design in next-generation renewable energy systems.

The creation of standardized, community-driven composite material datasets with shared metadata and testing standards remains essential for future reproducibility and cross-study comparison.

REFERENCES

1. International Energy Agency. Net zero by 2050: a roadmap for the global energy sector. Paris: IEA; 2023. Available from: https://www.iea.org/reports/net-zero-by-2050
2. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559:547–55.
3. Huang B, et al.. Best practices for machine learning in materials science. Nat Comput Sci. 2023;3:127–35.
4. Materials Project Consortium. The Materials Project: a materials genome approach to accelerating materials innovation [Internet]. 2013 [cited 2025 Nov 1]. Available from: https://materialsproject.org
5. Cassola S, et al.. Machine learning for polymer composites process simulation. Compos Struct. 2022;289:115390.
6. Sorour SS, et al.. A review on machine learning implementation for laminated FRP composites. Compos Struct. 2024;320:116978.
7. Malashin A, et al.. Machine learning prediction of mechanical properties of natural fiber reinforced polymer composites. Compos Struct. 2024;326:116966.
8. ziya07. FDM 3D printed composite material prediction data [Internet]. Kaggle; 2021 [cited 2025 Nov 1]. Available from: https://www.kaggle.com/datasets/ziya07/fdm-3d-printed-composite-material-prediction-data
9. DeepMind. Graph networks for materials exploration [Internet]. DeepMind Research; 2023 [cited 2025 Nov 1]. Available from: https://deepmind.google/discover/blog
10. Flores-Campos R. Polymer composite dataset [Internet]. Mendeley Data; 2023 [cited 2025 Nov 1]. doi:10.17632/9bshc597y9.1. Available from: https://data.mendeley.com/datasets/9bshc597y9/1

11. MatWeb material property database [Internet]. 2025 [cited 2025 Nov 1]. Available from: https://www.matweb.com
12. Composites Part A: Applied Science and Manufacturing – supplementary datasets [Internet]. Mendeley Data Repository; 2024 [cited 2025 Nov 1]. Available from: https://data.mendeley.com/journal/1359835X
13. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–74.
14. DOME-ML Community. Data and model reporting guidelines for machine learning in science [Internet]. 2021 [cited 2025 Nov 1]. Available from: https://www.dome-ml.org

