Korean J. Chem. Eng., 40(9), 2091-2101 (2023)
DOI: 10.1007/s11814-023-1502-3

pISSN: 0256-1115
eISSN: 1975-7220

An LSTM model with optimal feature selection for predictions of tensile behavior and tensile failure of polymer matrix composites

Jaewook Lee, Nagyeong Lee, Jinkyung Son, and Dongil Shin†

Department of Chemical Engineering, Myongji University, Yongin, Gyeonggi-do 17058, Korea
(*Received 24 January 2023 • Revised 24 January 2023 • Accepted 24 May 2023*)

Abstract−Mechanical properties such as tensile strength, ductility, and tensile modulus are essential criteria in polymer matrix composites (PMC) design and are determined through the stress-strain curve obtained from the tensile test. Material designers can examine the stress-strain curve trends based on the combination and composition, but it is difficult to predict using numerical analysis software due to the complex correlation based on chemical properties. To address these limitations in PMC design, this study uses feature engineering methods such as principal component analysis (PCA) and recursive feature elimination with cross validation (RFECV) to find the minimal and optimal set of features necessary for predicting the tensile behavior of PMC. The Long Short-Term Memory (LSTM) and feedforward neural network (FNN) models are trained using the optimal feature set and 1,270 PMC's tensile test data to predict the tensile stress-strain curve. The predictive model developed in this study provides stress-strain curves of tensile tests, including tensile failure of PMC, which can be challenging due to the high nonlinearity of PMC. Material designers can reduce the time and labor costs of PMC design through this tensile behavior prediction model that has an accuracy of R^2=92% and requires fewer features. In addition, the model can be used as a high-throughput screening model for PMC inverse design systems.

Keywords: Artificial Intelligence, Machine Learning, Mechanical Property, Polymer Matrix Composites

INTRODUCTION

As the demand for sustainable energy in the international community increases, there is a growing need to reduce the carbon emissions of internal combustion engine vehicles and improve the fuel efficiency of electric vehicles. One way to achieve this is by reducing the weight of vehicles, as the weight of a vehicle is closely related to its fuel efficiency and carbon emissions. Current automotive materials range from heavy steel to lightweight aluminum alloys, plastics, and ceramics. Among these materials, plastic has the advantages of being lightweight, easy to mold, and low cost; it is used in about 10% of each vehicle. However, compared to metal, plastics have lower strength and stiffness. To overcome the limitations of plastics, there is a need to develop new PMCs that combine the advantages of plastics with improved mechanical properties and heat resistance [1,2]. The mechanical properties of these materials are greatly influenced by the work condition and are closely related to safety and performance, making them essential considerations when designing PMC. In general, two variables can affect the mechanical behavior of PMC:

① Combination variables, such as the type and composition of polymers and reinforcements.
② Manufacturing process variables, such as the injection speed or cooling temperature.

Designing PMC with exceptional mechanical properties is difficult only by optimizing the manufacturing process variables [3]. It is essential to consider the combination and composition of the polymers and reinforcements. Designing PMC by considering multiple variables simultaneously is time-consuming and costly for material designers.

While first-principles modeling has been used to predict the mechanical properties of PMC [4-8], it is a highly theoretical approach that heavily relies on accurate knowledge of the material structure and interaction between atoms. This level of detail is often not feasible for practical applications due to its computational intensity and complexity. Moreover, the molecular structure of the polymer, which is the constituent material of PMC, is inherently disordered, resulting in high nonlinearity and uncertainty in its mechanical properties. Additionally, when reinforcement is added to form a composite material, it becomes challenging to characterize the molecular structure, making it extremely difficult to obtain accurate predictive results [3,6]. In recent years, there have been significant advances in first-principles modeling for predicting the mechanical properties of PMC. For instance [7,8], used density functional theory (DFT) to model the mechanical properties of PMCs, while [4,5] applied molecular dynamics simulations (MD). However, these methods still demand significant computational resources and often oversimplify due to the complexity of the materials, leading to inaccuracies in prediction [9].

Research on data-driven models using machine learning, which is well-suited for nonlinear modeling, is actively being conducted to address the limitations of first-principles models. Also, the hybrid modeling approach, which integrates first-principles models with machine learning, has garnered considerable attention in the realm of chemical process systems due to its potential to enhance inter-

†To whom correspondence should be addressed.
E-mail: dongil@mju.ac.kr

pretability and extrapolation [10]. Among these, physics-informed machine learning has demonstrated remarkable results in various fields related to dynamic modeling, such as material manufacturing processes, physics, and robotics [11-13]. Through these precedents of applying hybrid modeling, there is potential to improve the accuracy and interpretability of the predictive model we aim to develop. But there are considerable challenges [10]. One of the primary issues is the computational cost. Hybrid models, especially those combining first-principles models with machine learning, can be computationally expensive and time-consuming. Additionally, first-principles models require detailed atomic or molecular information, which cannot be readily accessible for all materials and challenging to obtain with high accuracy. These limitations could potentially impact the model's applicability and versatility. Given these challenges, this study focuses on developing a purely data-driven model, recognizing its potential to offer an effective and efficient way to predict the mechanical behavior of PMCs.

Previous studies with data-driven models have faced challenges in predicting the mechanical properties of PMC based on their chemical composition and other characteristics [14-25]. In this study, we addressed four specific issues. First, previous studies on predicting the mechanical properties of PMC have focused on predicting properties such as tensile strength, modulus, and elongation [14-16]. Still, this approach has the drawback of not considering measurement standards and method variations. Therefore, it is crucial to predict the entire stress-strain curve of the PMC.

The second challenge is that there is a lack of experimental data on the mechanical properties of PMC based on their combination and composition [3]. This is a common issue in data-based modeling. Furthermore, the high nonlinearity of PMC leads to high uncertainty in the data, and even PMC with the same combination and composition of constituent materials can exhibit different tensile behavior. To overcome this, it is essential to improve the reliability of the data. This study addressed this problem by obtaining 1,270 highly reliable stress-strain curve data by conducting repeated experiments five to ten times for each PMC, as discussed in more detail in section 2.

The third challenge is selecting the optimal set of features for predicting the properties of PMC. Selecting the appropriate number of features is one of the most important issues in machine learning [17-19]. If the number of input features is too low, the model's explanatory power decreases, leading to lower accuracy in predictions. However, using too many features can also cause the model to perform poorly, as it needs to interpret all of them correctly. Previous studies developed the models using the feature set, which is suitable for mechanical property prediction [20-23]. There has been no prior research on the optimal feature set for predicting tensile stress-strain curves. Therefore, we sought to find the optimal feature set that can sufficiently explain the tensile behavior of the PMC. This is explained in more detail in section 3.

Finally, because of the PMC's high uncertainty, it is challenging to predict the failure point of the stress-strain curve. Previous studies did not consider the failure point's variance [24,25]. The user is required to specify the output range manually, or the model only outputs without the failure point, which leads to a lack of reliability in the model and its predictions. We proposed a tensile stress-strain curve prediction model that includes the tensile failure point using the post-padding method, which is commonly used in natural language processing and time series processing.

DATA COLLECTION, PREPROCESSING AND SPLIT FOR MODEL TRAINING AND VALIDATION

In this study, we collected PMCs' tensile stress-strain curve data for model training. The PMCs used for the tensile tests are chosen as they are commonly used as lightweight materials. Polypropylene (PP), Polycarbonate (PC), Polyamide6 (PA6), and Polyamide6,6 (PA66) are used as the polymer matrix, and Al_2O_3, Al_2O_5Si, Boron Nitride (BN), and Si_3N_4 are used as the reinforcements. The PMC specimens are composed of a 90-40 wt% polymer matrix based on the type and combination of PMC. The PMC specimens have a thickness of 4 cm and a width of 10.13 cm and are manufactured by ASTM regulations. To eliminate variables other than the type and combination of PMC, Tensile tests are conducted at a fixed speed of 50 mm/min at room temperature (25 °C). The tensile tests are performed using a Universal Testing Machine (UTM), which is depicted in Fig. 1(a).

As mentioned, the uncertainty and the nonlinear characteristic of the tensile test data of PMC can be seen in Fig. 1(b) and Fig. 1(c). Fig. 1(b) shows the tensile stress-strain curve of PMC made of 40 wt% PP, 60 wt% Al_2SiO_5, and Fig. 1(c) shows the curve of PMC made of 40 wt% PA66, 60 wt% BN. The tensile test data of PMC with the same combination and composition show a large variance (uncertainty) of 7% to 15%. To consider this uncertainty when designing PMC, we used the average value of the stress-strain curves through repeated tensile tests as train data. Therefore, repeated tensile tests were conducted 5-10 times for one type of PMC. As a result, 1,270 tensile test data of 85 types of PMC were obtained.

1. Data Preprocessing for Improved Model Training

The obtained data is time-series data expressing the stress as a function of the strain from the tensile test, and it has 128,000 based on the index. Data reduction is performed to prevent excessive time consumption for model training and hyperparameter tuning due to the large data size. Even though the index of the tensile test data is reduced from the default 0.01-second interval to a 0.1-second interval, the accuracy of the data is not compromised. Therefore, the average of the 0.01-second data is replaced with the 0.1-second interval data, and the data is reduced. As a result, the size of the entire data is decreased to 10% of the original data, improving the efficiency of model training.

2. Data Split Using k-fold Cross-validation

Since the most important factor in the quantitative structure-property relationship (QSPR) problem, which predicts the properties of chemical substances from their structures, is the extrapolation capability, the data are partitioned using the k-fold cross-validation method as shown in Fig. 2. Specifically, materials composed of the same polymer matrix and reinforcement are divided into validation sets. The remaining data are used for model training. Finally, for a total of 85 PMC, training and verification are performed by excluding specific combinations one by one, and the average of the performance obtained in the 85 verification processes is used as the final performance.

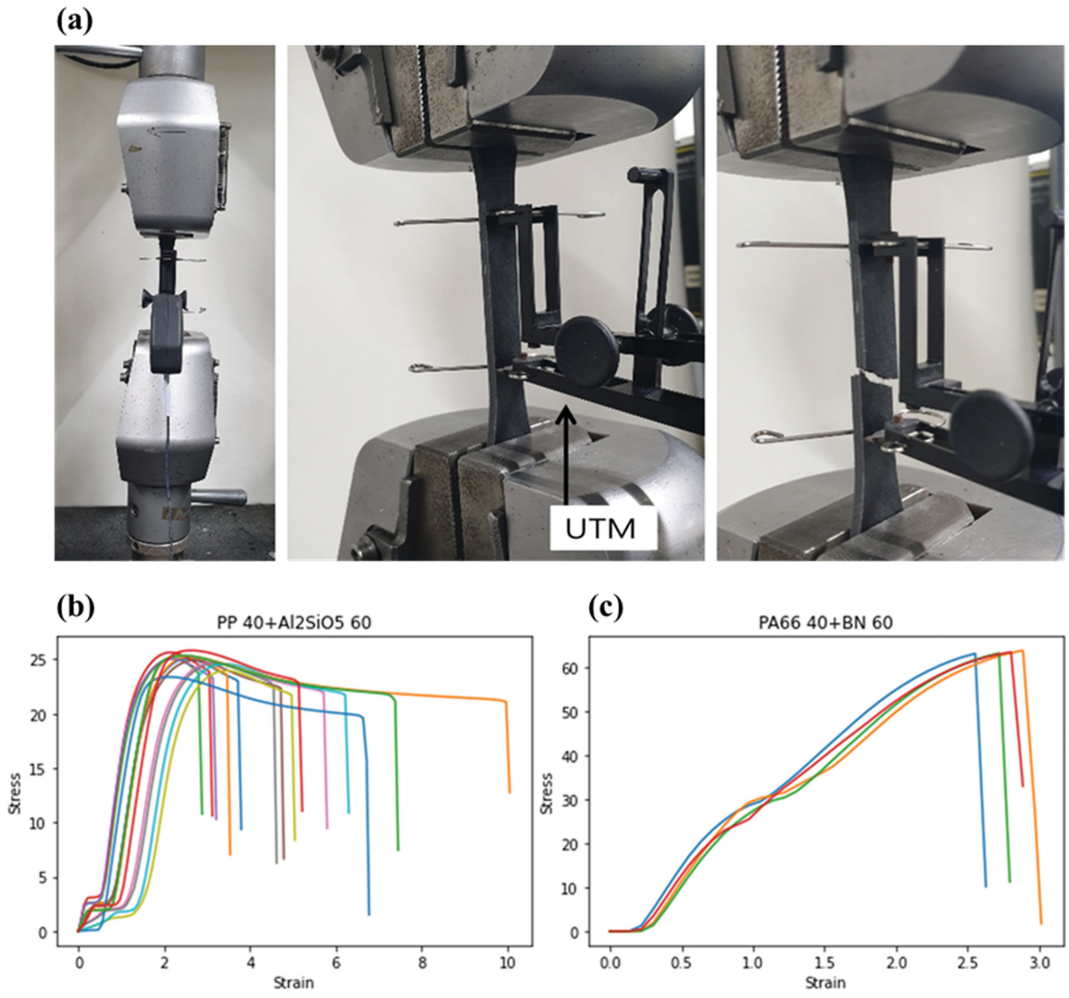

Fig. 1. Tensile test and stress-strain curve data of PMC.

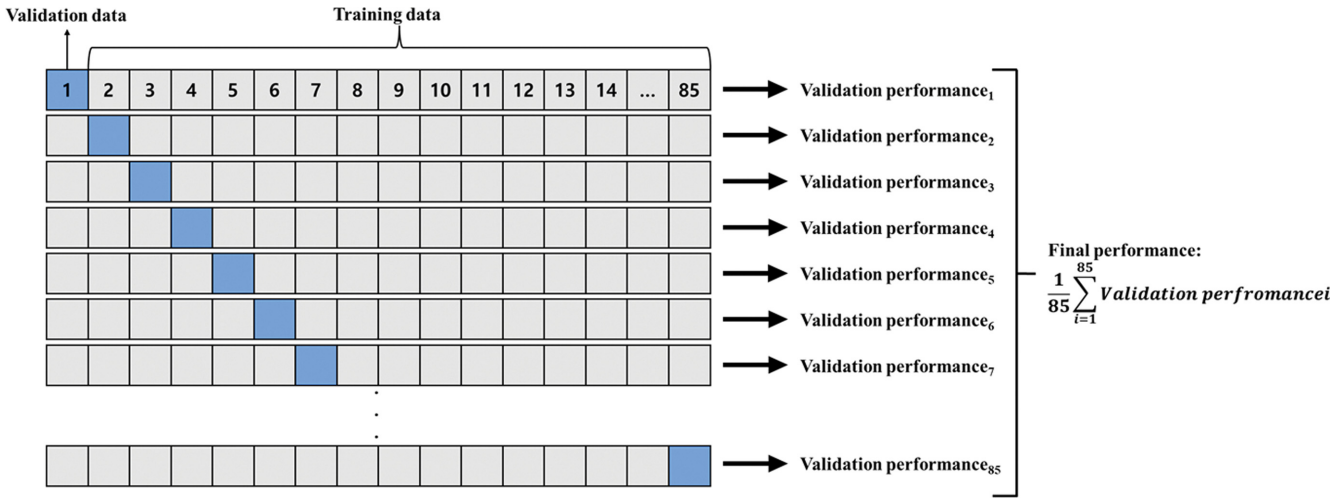

Fig. 2. Dividing the verification data through the k-fold method.

IDENTIFICATION OF OPTIMAL FEATURE SET THROUGH FEATURE ENGINEERING

For prediction through a machine learning-based model, it is crucial to identify a feature set that can adequately reflect the characteristics of polymers and reinforcing materials, which are the constituent materials of PMC. Features are collected, as shown in Fig. 3, and the feature's dimension is reduced using the PCA method

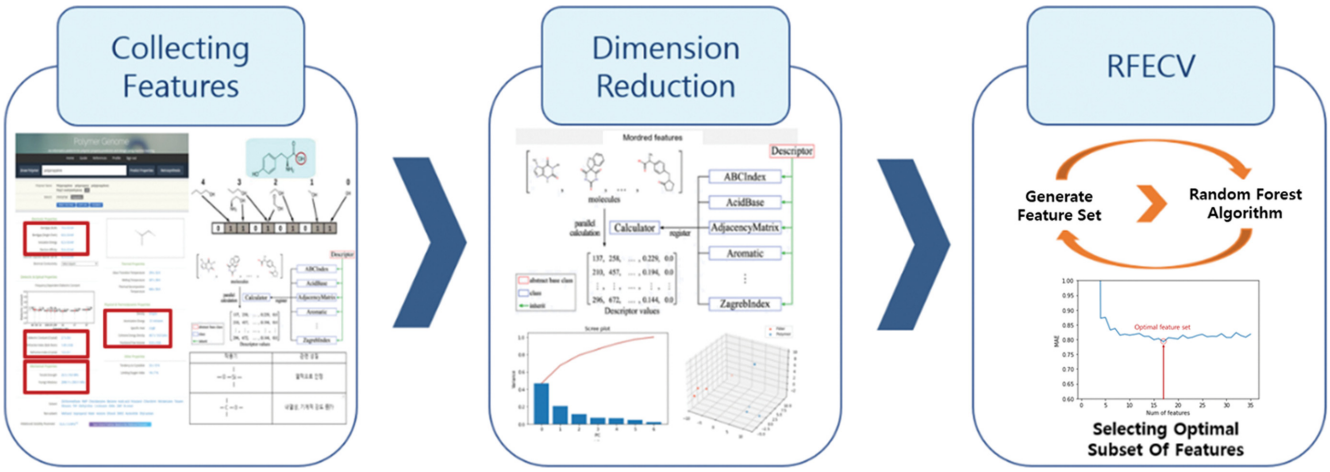

Fig. 3. Feature engineering.

[26]. And the RFE, a feature selection method, is employed to secure a feature set essential for modeling by eliminating unnecessary parts among the collected and preprocessed features.

1. Feature Collection for PMC Modeling

We propose a model for predicting the tensile behavior of PMC by utilizing four types of features, including chemical, mechanical, thermodynamic properties and chemical descriptors (Mordred descriptors). The features are collected from PubChem and PolyInfo, and a total of 21 properties were used for modeling [27,28]. The properties selected as features include composition, molecular weight (matrix & reinforcement), bandgap (matrix), ionization energy (matrix), electron affinity (matrix), density (matrix & reinforcement), atomization energy (matrix), cohesive energy density (matrix), fractional free volume (matrix), dielectric constant (matrix), refractive index (matrix), specific heat (matrix), glass transition temperature (matrix), Young's modulus (matrix & reinforcement), Poisson's ratio (matrix & reinforcement), and tensile strength (matrix & reinforcement).

Mordred is an open-source library developed as a molecular descriptor calculator to solve the QSPR problem [29]. It provides a vast amount of feature sets, up to 1825, for a single material, including molecular weight, types of constituent atoms, presence of functional groups, and molecular structure represented in an adjacency matrix. In this study, it is used to reflect the molecular characteristics of the polymer and reinforcement for predicting the properties of PMC. To comprehensively interpret the mechanical behavior of PMC, a large amount of information provided by Mordred is utilized as a feature to reflect the molecular characteristics of the polymer and reinforcement, and 1825 Mordred descriptors for each polymer and reinforcement are obtained for training the models using the python package RDKit [30].

2. Feature Extraction for Dimension Reduction

Mordred, which provides ample information about the material with 1825 descriptors for each polymer and reinforcement, is utilized as the main feature for predicting the mechanical properties of PMC. However, high-dimension features can often increase the complexity of many algorithms, negatively impacting performance. This problem, referred to as the curse of dimensionality, arises as data dimension increases and causes unexpected behavior in commonly used Euclidean distances. To address this issue, it is necessary to transform the vast number of features into more compressed information. One solution is to use dimension reduction methods, such as PCA, to reduce the number of features used in the model. PCA, a widely used dimension reduction method, is utilized to compress the Mordred descriptors and reduce the dimension of the data.

PCA is a multivariate nonparametric method that reduces correlated multivariate data to low-dimensional data with as little loss of information as possible. Principal components are obtained by compositing by giving weights to each variable. Weights are applied using the original variable's information as much as possible to maximize individual variance. If the original variable has a plurality of principal components, each principal component is determined so that there is no correlation with each other; that is, the covariance between different principal components is zero. To utilize PCA, the number of dimensions to be compressed is determined through cumulative explained variance and individual explained variances

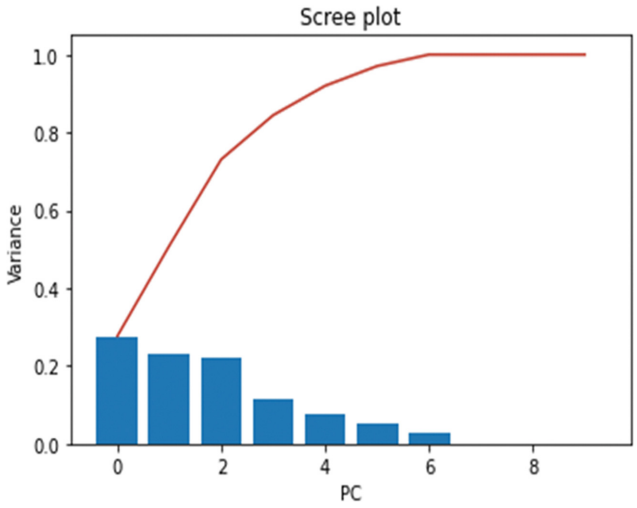

Fig. 4. Result of dimension reduction using PCA.

of each principal component. The result of dividing the Mordred descriptors of the PMC using PCA is shown in Fig. 4. The bar chart represents individual variance explained by different principal components, and the plot describes the total variance explained by different principal components. As a result of the analysis, the existing data is best explained when there are seven principal components. So the 1825 Mordred descriptors of polymer and reinforcement are reduced to 7 features, and a total of 14 Mordred descriptors (polymer & reinforcement) are added to the feature set.

3. Optimal Feature Set Selection

When a model requires too many features as input, it can be burdensome and potentially detrimental to performance due to the risk of overfitting. To simplify the model and improve generalization, it is crucial to extract and use only the most predictive features. One common approach is recursive feature elimination (RFE), which removes the least important features one by one until a specified number of features remains. However, RFE can be computationally expensive, especially when dealing with a large number of features, and requires a predetermined number of features to select. An alternative is recursive feature elimination with cross validation (RFECV), a feature selection method that combines the power of RFE and the robustness of cross-validation. RFECV begins by training a model on the initial feature set and assigning a score to each feature based on its importance. The least important feature is then removed and the process is repeated. At each step, the model's performance is assessed using k-fold cross-validation, providing a more reliable estimate of its generalization ability. The RFECV method automatically identifies the optimal number of features, which corresponds to the highest cross-validated performance.

In this study, we employed the RFECV method with a random forest model, an ensemble algorithm based on decision trees, to select the optimal feature set from our original 35 features. These features consisted of 14 Mordred descriptors and 21 previously collected material property data. The model's performance was evaluated using mean absolute error (MAE) as the metric. As illustrated in Fig. 5, a similar MAE value is achieved when the number of features is eight or more, with the feature set comprising 17 features showing the least error.

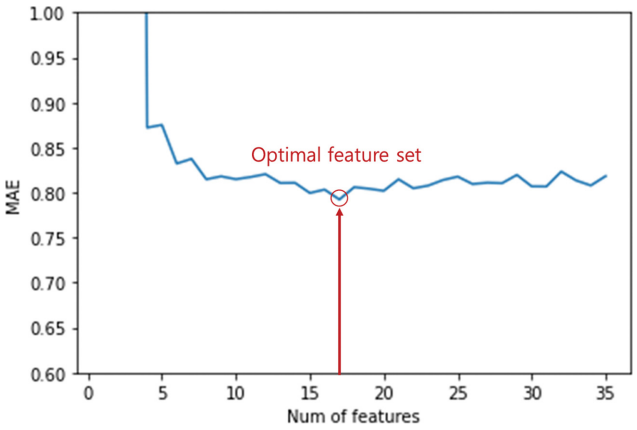

Fig. 5. Optimal feature set result found through RFECV.

In this study, we chose an optimal feature set of 17 features, even though reducing the number of features below 17 might have still provided a reasonable MAE. Our decision was based on several considerations. First, while decreasing the number of features below 17 could potentially maintain a comparable MAE, it would also result in a marginal increase in error. Given that a stress-strain curve can be seen as a kind of raw data for verifying various mechanical properties, even a slight increase in error can cause a change in the measured mechanical property values, which can have significant implications for practical applications. Second, we also considered the interpretability and robustness of our model. The chosen set of 17 features enabled a more comprehensive representation of the inherent complexity within our dataset. As evidenced by previous studies [3,31,32], each of these 17 features provides unique and valuable information that represents the mechanical properties, chemical properties, and structural information of the constituents of the polymer composite resin. This information enhances the interpretability of the fundamental patterns of the stress-strain curve. In con-

Table 1. The 17 selected features in the optimal set identified using RFECV method

Index	Feature name
1	Composition of the polymer matrix composite
2	Density of the polymer
3	Density of the reinforcement
4	Tensile Strength of the polymer
5	Tensile Strength of the reinforcement
6	Molecular weight of the polymer
7	Molecular weight of the reinforcement
8	Glass transition temperature of the polymer
9	Poisson's ratio of the reinforcement
10	Eight Mordred descriptors, which are computational descriptors calculated from the molecular structure of the material capturing various properties such as size, shape, and electronic structure
…	
17	

clusion, we elected to use the optimal feature set of 17 features because it provided the best balance between prediction accuracy and interpretability.

The selected optimal feature set, as detailed in Table 1, comprised 17 key features that played significant roles in predicting the target variable. Each of these features contributed unique and valuable information, justifying their inclusion in the final model. Utilizing this condensed yet potent feature set, we proceeded with the modeling process. This approach substantially simplified the model, reducing potential issues of overfitting and computational complexity, while maintaining robust predictive power. Our strategy underscores the importance of prudent feature selection in model optimization, striking a balance between simplicity and predictive accuracy.

DESIGN OF PREDICTION MODELS

1. Post-padding for Prediction of Tensile Failure

The obtained stress-strain curve data of PMC is time-series data measuring stress according to strain, and the length of the data is different depending on the material's ductility. When we train the machine learning models, we need to set the length of the model's output, but the length of the tensile test data is different depending on the PMC's elongation. Therefore, a padding process is required to match the data length. The average length of the stress-strain curve data of the PMC is 12 cm, and the maximum length is 17.5 cm. Therefore, post-padding is performed to add data to the back of the data to unify all data to a standard length of 17.5 cm.

Zero padding is performed to match the data length with a value of 0 but, in this study, it is most important to predict the failure point, which is the point at which the stress-strain curve ends. So, to force the model to learn about the failure point, we process the post-padding with a physically impossible value of −10, as seen in Fig. 6. The same length stress-strain curve, including the failure point, is processed as training data.

2. Evaluation of the Optimal Feature Set Using an FNN Model

FNN is a type of artificial neural network in which data is transmitted in a forward direction [33]. Fig. 7 is an FNN structure with four hidden layers. The input layer, hidden layer, and output layer exist, and respective weights connect the neurons of each layer. Input data ($X=[X_1, X_2, \ldots, X_n]$) is given as an input signal; the FNN goes through the following calculation process. Obtained through Eq. (1) is received as an input of an active function, the output value of neurons in each layer is derived, and the calculation formula is shown in Eq. (2). In the case of regression analysis, the activation function of the output layer uses an identity function that returns the input value as the output value. Hence, the value output from the output layer's identity function is the final output value of the forward neural network. The calculation formula is as shown in Eq. (3). The cost function between $\hat{Y}_m$, which is the output value of Eq. (3), and the actual data ($Y=[Y_1, Y_2, \ldots, Y_m]$) mainly uses mean squared error (MSE), as shown in Eq. (4) can indicate The training method of forward neural networks is to find weights and biases so that the loss function value is minimized.

$$a_m^{(k+1)} = \Sigma W_{n,m}^{(k)} \cdot x_n^{(k)} + b^{(k)} \text{ for } k=1, 2, 3 \quad (1)$$

$$X_m^{(k+1)} = h(a_m^{(k+1)}) \text{ for } k=1, 2, 3 \quad (2)$$

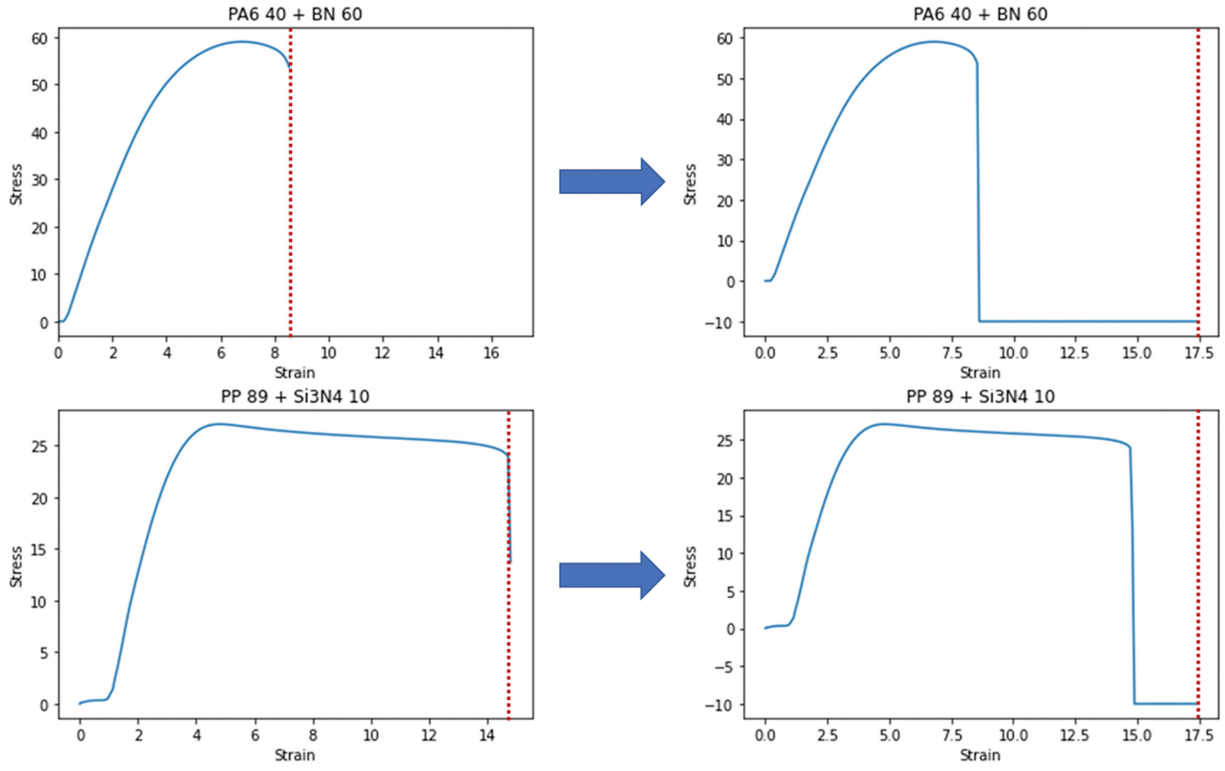

Fig. 6. Post-padding of PMC's tensile stress-strain curve.

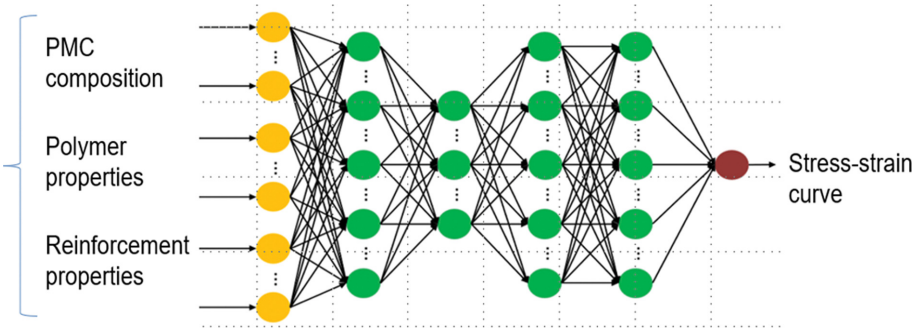

Fig. 7. Deep learning model for predicting tensile behavior of PMC.

Table 2. Hyperparameters of the developed models after tuning

Hyperparameter	FNN-1	FNN-2	FNN-3	LSTM
Layer number	6	4	4	3
Units of layers	64-256-128-64-32-16	256-128-64-32	128-128-64-16	32-128-32
Optimizer	Adam	Adam	Adam	Adam
Activation function	ReLU	ReLU	ReLU	ReLU
Learning rate	0.089	0.032	0.074	0.0061
Loss function	Mean Squared Error	Mean Squared Error	Mean Squared Logarithmic Error	Mean Squared Logarithmic Error

$$\hat{Y}_m = X_m^4 = h'(a_m^4) \text{ for } m = 1 \ldots N \quad (3)$$

$$L(W) = \frac{1}{N}\sum_{m=1}^{N}(Y_m - \hat{Y}_m)^2 \quad (4)$$

We trained three FNN models, which were optimized using Bayesian optimization for their hyperparameters [34]. The structures of the three optimized models can be observed in Table 2. The first model was trained with the entire feature set of 35 features, and the second model was trained with the optimal feature set of 17 features to confirm the impact of the optimal feature set. Additionally, the third model was developed to predict the entire stress-strain curve, including the failure point, which was encoded using post-padding. The results are described in Section 5.

3. LSTM Model to Predict the Stress-strain Curve with Tensile Failure

Unlike FNN, recurrent neural networks have a hidden layer consisting of repetitive cells that are affected by both past state and current input. LSTM consists of an input gate, output gate, forget gate, and a memory cell in a proposed method to alleviate the vanishing gradient problem of the recurrent neural network [35]. Input

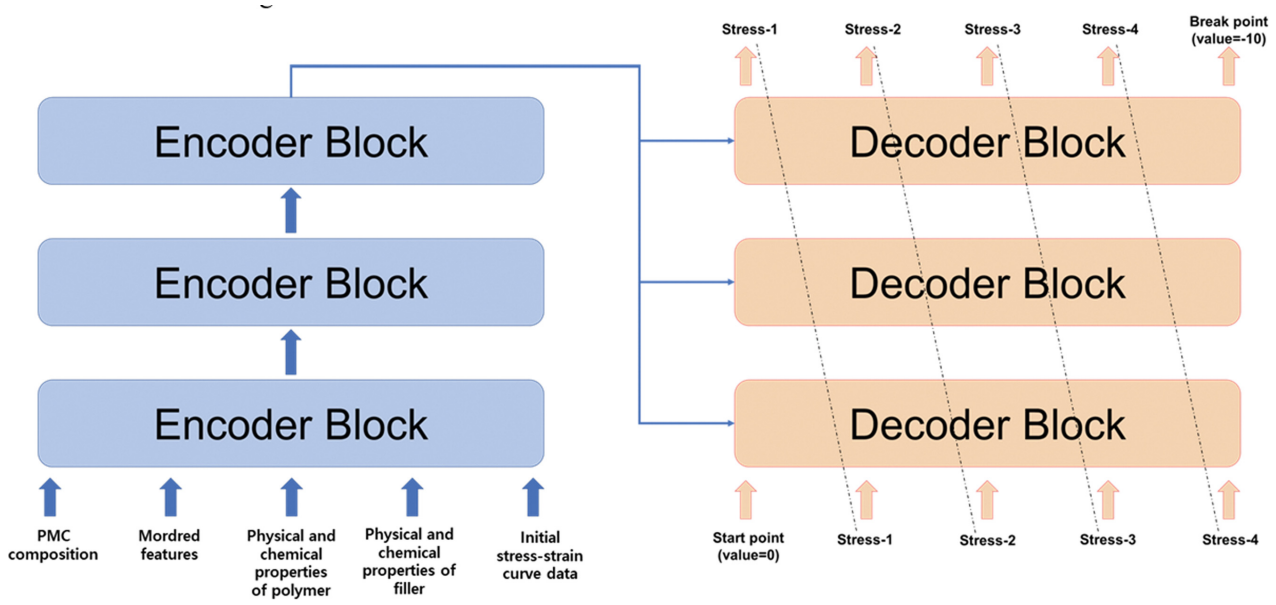

Fig. 8. LSTM model structure.

gate (i_t) is a process of storing current data and can be expressed as Eq. (5). Output gate (Q_t) serves to determine the output value and can be expressed as Eq. (7). Forget gate (f_t) is a step of determining whether to remember the past information (h_{t-1}). It can be expressed as Eq. (9). Memory cell (C_t) is a process of updating the current state to the past state (C_{t-1}), as shown in Eq. (10).

$$i_t = \sigma(W_i h_{t-1} + W_i x_t + b_i) \quad (5)$$

$$\tilde{C}_t = \tanh(W_c h_{t-1} + W_c x_t + b_c) \quad (6)$$

$$O_t = \sigma(W_0 h_{t-1} + W_0 x_t + b_0) \quad (7)$$

$$Y_t = O_t \times \tanh(C_t) \quad (8)$$

$$f_t = \sigma(W_f h_{t-1} + W_f x_t + b_f) \quad (9)$$

$$C_t = (f_t \times C_{t-1}) + (i_t \times \tilde{C}_t) \quad (10)$$

An LSTM model with such a learning algorithm is used for efficient prediction by storing information coming from the input gate into a short-term state or long-term state. The LSTM model needs an initial input value. We developed the model to predict the entire stress-strain curve, including the failure point, using the initial data of the stress-strain curve as an additional input. The developed model structure is shown in Fig. 8 and is optimized using Bayesian optimization for its hyperparameters. The structure of the optimized LSTM model can be observed in Table 2 and the modeling results are described in Section 5.

RESULTS AND DISCUSSION

1. FNN Model-based Verification of the Performance of the Optimal Feature Set

A comparison test was conducted using FNN models to prove the optimal feature set's performance. The R^2, the coefficient of determination, was used as a performance evaluation index, and R^2 is a value that indicates how well the independent variable explains the dependent variable. The closer to 1, the higher the explanatory power, and the closer to 0, the lower the model's explanatory power. When training the FNN model with 35 feature sets without feature selection, R^2 is 84% on average. The result of understanding with 17 optimal feature sets obtained through RFECV can be seen in Fig. 9. A performance improvement of about 5% is confirmed.

Unlike other descriptors that do not provide enough features for PMC, the Mordred descriptors verified in this study can be used to develop models for predicting the mechanical behavior and physical properties of a wide range of PMCs.

2. Performance of the FNN and LSTM Models for Predicting the Entire Stress-strain Curve Including the Tensile Failure

The behavior before the failure point is well predicted when training data does not include the failure point with the FNN model. Still, only limited information is provided to the material designer because the point where the breaking occurs is yet to be predicted. Therefore, when the entire stress-strain curve prediction model is developed using the post-padding method to predict the failure point, the basic FNN model learns the failure point encoded

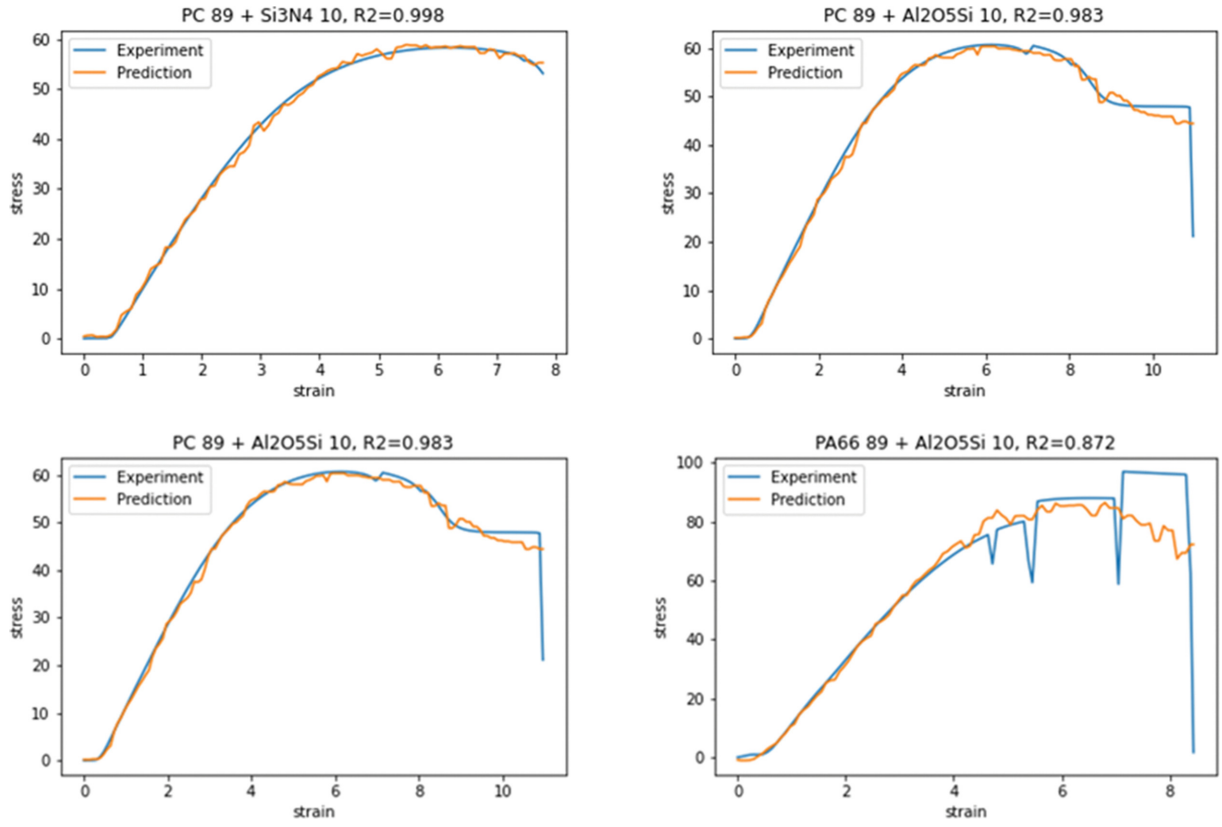

Fig. 9. Prediction result of FNN model using optimal feature set.

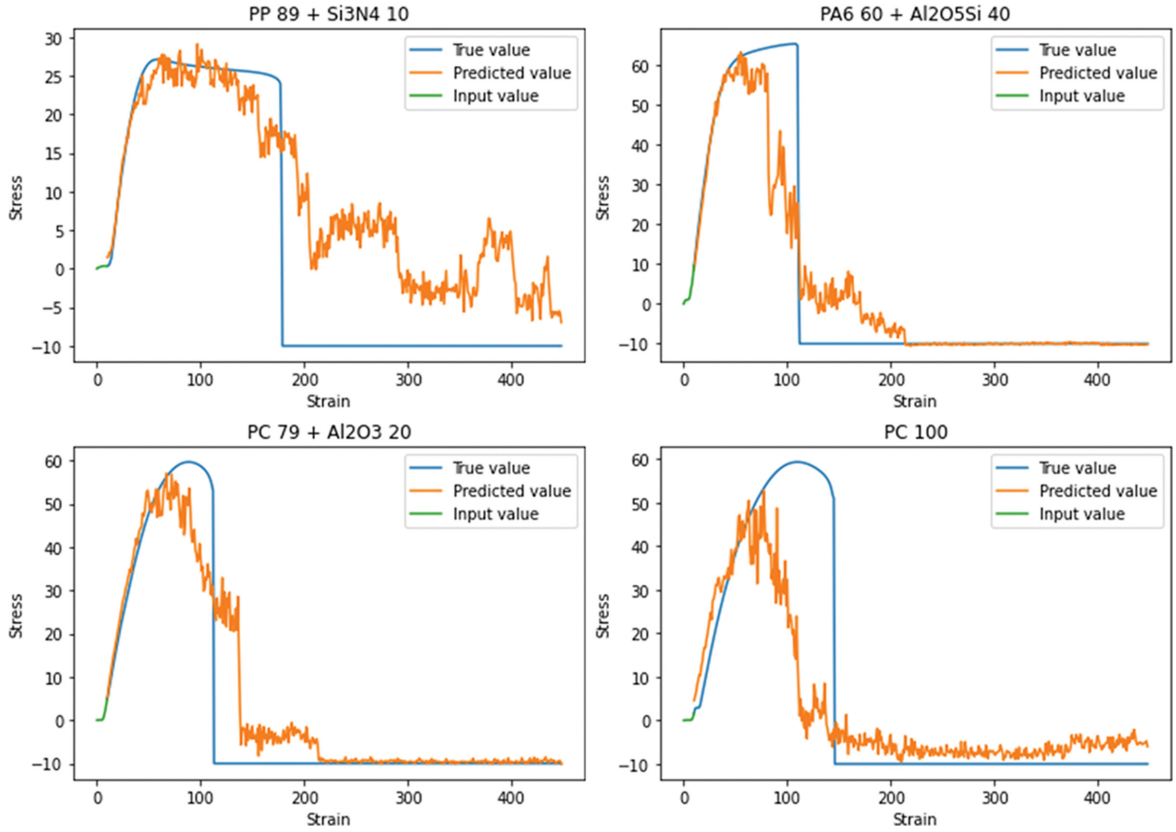

Fig. 10. Prediction result of FNN model considering failure point.

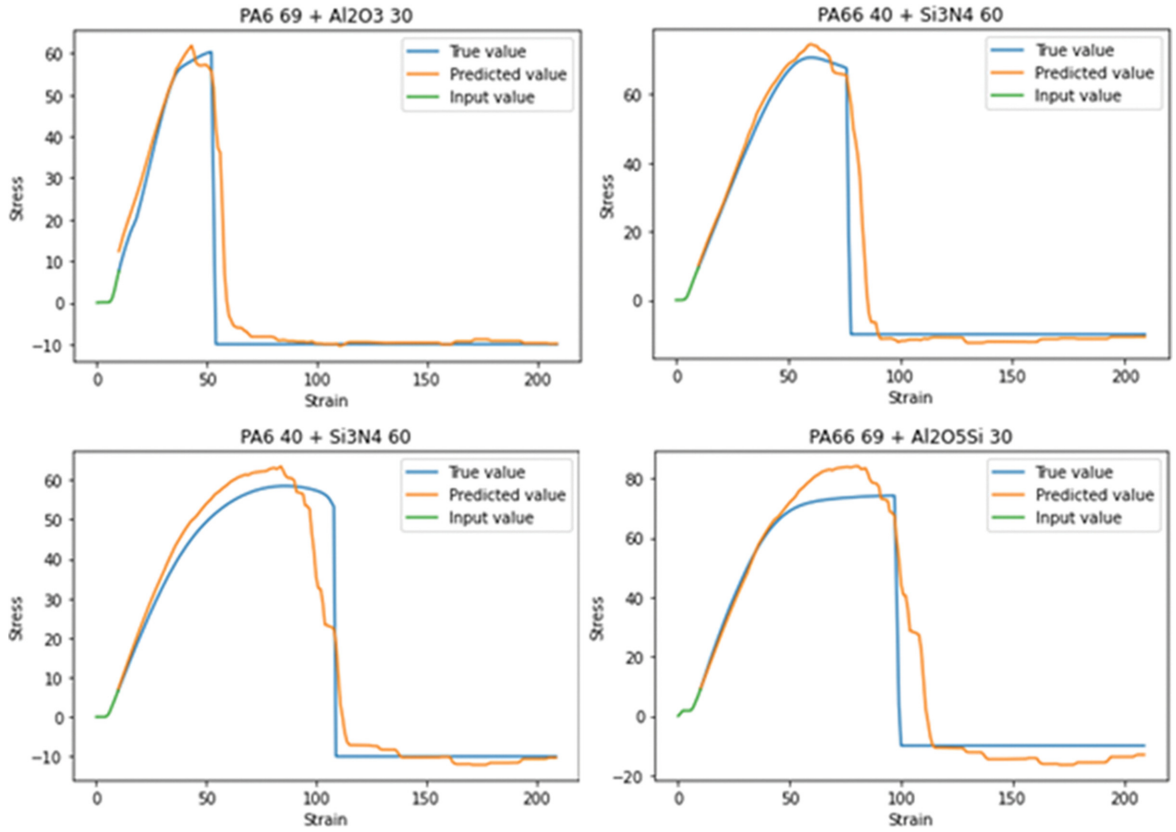

Fig. 11. Prediction result of LSTM model considering failure point.

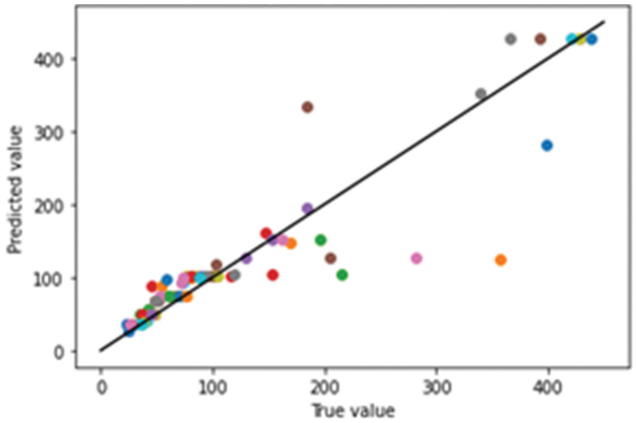

Fig. 12. Prediction result of failure point.

as an outlier, indicating that the model's performance is deficient and R^2 is 53%; this can be seen in Fig. 10.

On the other hand, the LSTM model accurately interprets the failure point as a failure point of the stress-strain curve, confirming the results shown in Fig. 11 and achieving an average R^2 of 92% of the entire stress-strain curve. Additionally, we assessed the accuracy of failure point prediction. As the predictions typically indicate a more gradual fracture occurrence than the actual values, we defined the failure point as the instance when the predicted stress-strain curve first reached a negative value, and this was compared with the actual point of failure. The results, as presented in Fig. 12, indicate an R^2 value of 83%, and a small error of approximately 4.79 based on the MAE, affirming the reliability of failure point prediction.

3. Performance Evaluation of the LSTM Model Using Limited Initial Input Data

The LSTM model requires initial input values of the tensile stress-strain curve data. Therefore, we verified how the model's performance changes depending on the data size used as the initial input. The initial curve data used as input was created by generating the model with 1% to 20% of the data with a 1% interval. The verification was carried out using the validation data. As a result, as seen in Fig. 13, it was confirmed that there is no significant change up to 20% of the data corresponding to the elastic section, and it was confirmed that 92% accuracy could be obtained with only the initial input value of 6%, including the failure point.

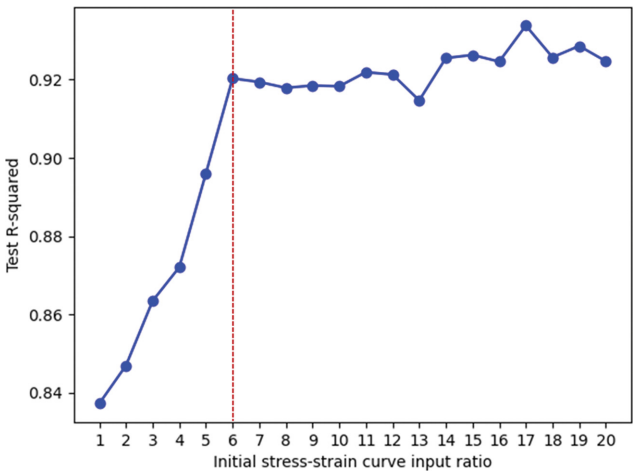

Fig. 13. Test R^2 of LSTM according to the initial stress-strain curve input ratio.

CONCLUSIONS

The stress-strain curve of a PMC, an essential indicator for directly observing mechanical properties required for material design, is difficult to model with only general nonlinear regression because of its strong nonlinearity even with general nonlinear regression methods. Simulations and other programs can only interpret results within the elastic range. In this study, we compared the performance of the FNN model with the LSTM model, which is specialized for time-series data processing, using 1270 tensile test data from 85 types of PMC made with matrix materials PA6, PA66, PC, and PP and reinforcements BN, Al_2O_3, Al_2O_5Si, Si_3N_4 at a composition ratio of 90% to 40%. Mordred descriptors, which reflect characteristics such as molecular-level phase information, chemical properties of polymers and reinforcements such as molecular weight and ionization energy, mechanical properties such as tensile modulus and strength, and thermodynamic properties such as melting point and glass transition temperature, are used as input features to propose a model that predicts the failure point and the entire stress-strain curve, which have not been previously studied.

The training data is constructed by incorporating a failure point into the existing tensile test data using end-encoding, a natural language processing method. It is found that predictions are possible with this data. Although the error tended to increase for the predicted failure point at high tensile elongation, it is very accurate for elongation of 200% or less, which is common among most materials. The overall prediction accuracy is 83%.

The LSTM model has the disadvantage of requiring an initial input value. However, as a result of the model test, it is found that it is possible to predict a stable stress-strain curve using only about 6% of the initial data. It can be applied when strain under higher stresses needs to be calculated. Additionally, unlike the prediction models for mechanical properties of PMC developed in previous studies, which relied on the developer's domain knowledge to select a large number of features, this study used feature engineering methods to choose the minimal and optimal set of 17 features that best explain the tensile behavior of PMC. By examining these elements, the efficiency and performance of the model are improved.

ACKNOWLEDGEMENTS

This work was supported by Korea Institute for Advancement of Technology (KIAT) through the Virtual Engineering Platform of Virtual Test, Data, and AI for Chemical Materials project (P0022334) and the Smart Digital Engineering Education and Training for Lead Engineer project (P0008475) funded by the Ministry of Trade, Industry and Energy (MOTIE).

REFERENCES

1. KPMG, Light weighting of materials: a shift in the center of the

automobile light weighting trend, *Issue Monitor*, **96** (2018).
2. M.-Y. Lyu and T. G. Choi, *Int. J. Precision Eng. Manuf.*, **16**, 1 (2015).
3. A. Sharma, T. Mukhopadhyay, S. M. Rangappa, S. Siengchin and V. Kushvaha, *Arch. Computat. Methods Eng.*, **29**, 3341 (2022).
4. U. F. Röhrig and I. Frank, *J. Chem. Phys.*, **115**(18), 8670 (2001).
5. J. Koyanagi, N. Takase, K. Mori and T. Sakai, *Compos. Part C: Open Access*, **2**, 100041 (2020).
6. H. J. Kreuzer and M. Grunze, *Europhys. Lett.*, **55**(5), 640 (2001).
7. B. B. Yin, J. S. Huang, W. M. Ji and K. M. Liew, *Carbon*, **200**, 10 (2022).
8. N. Keshmiri, P. Najmi, B. Ramezanzadeh and G. Bahlakeh, *J. Mol. Liq.*, **331**, 115800 (2021).
9. J. T. Orasugh and S. S. Ray, *Polymers*, **14**(4), 704 (2022).
10. W. Bradley, J. Kim, Z. Kilwein, L. Blakely, M. Eydenberg, J. Jalvin, C. Laird and F. Boukouvala, *Comput. Chem. Eng.*, **166**, 107898 (2022).
11. T. Wu and J. Movellan, Semi-parametric Gaussian process for robot system identification, *IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE (2012).
12. J. Lu, K. Yao and F. Gao, *AIChE J.*, **55**(9), 2318 (2009).
13. S. Yang, S. W. K. Wong and S. C. Kou, *Proc. National Acad. Sci.*, **118**(15), e2020397118 (2021).
14. B. A. Shuvho, M. A. Chowdhury and U. K. Debnath, *Mater. Perform. Charact.*, **8**, 288 (2019).
15. M. A. S. Matos, S. T. Pinho and V. L. Tagarielli, *Carbon*, **146**, 265 (2019).
16. I. Argatov, *Front. Mech. Eng.*, **5**, 30 (2019).
17. D. Koller and M. Sahami, *Toward optimal feature selection*, Stanford InfoLab Technical Report (1996).
18. J. Cai, J. Luo, S. Wang and S. Yang, *Neurocomputing*, **300**, 70 (2018).
19. C. Lee and G. G. Lee, *Inf. Process. Manage.*, **42**, 155 (2006).
20. M. V. Pathan, S. A. Ponnusami, J. Pathan, R. Pitisongsawat, B. Erice, N. Petrinic and V. L. Tagarielli, *Sci. Rep.*, **9**, 1 (2019).
21. Z. Jiang, Z. Zhang and K. Friedrich, *Compos. Sci. Technol.*, **67**, 168 (2007).
22. D. W. Abueidda, M. Almasri, R. Ammourah, U. Ravaioli, I. M. Jasiuk and N. A. Sobh, *Compos. Struct.*, **227**, 111264 (2019).
23. R. Haddad and M. Haddad, *Struct. Concr.*, **22**, 1 (2021).
24. M. S. Nashed, J. Renno and M. S. Mohamed, *Fatigue Fract. Eng. Mater. Struct.*, **45**, 9 (2022).
25. H. Byun and J. J. Song, *Tunnel Underground Space*, **28**(3), 277 (2018).
26. H. Abdi and L. J. Williams, *Wiley Interdisciplinary Rev.: Comput. Statistics*, **2**(4), 433 (2010).
27. Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang and S. H. Bryant, *Nucleic Acids Res.*, **37**(2), 623 (2009).
28. S. Otsuka, I. Kuwajima, J. Hosoya, Y. Xu and M. Yamazaki, PoLyInfo: Polymer database for polymeric materials design, *2011 International Conference on Emerging Intelligent Data and Web Technologies*, IEEE, 22 (2011).
29. H. Moriwaki, Y. S. Tian, N. Kawashita and T. Takagi, *J. Cheminformatics*, **10**, 1 (2018).
30. G. Landrum, *Rdkit documentation*, Release 2019.09.1 (2019).
31. T. S. M. Kumar, K. Senthilkumar, M. Chandrasekar, S. Subramaniam, S. M. Rangappa, S. Siengchin and N. Rajini, *Biofibers and Biopolymers for Biocomposites: Synthesis, Characterization and Properties*, 111 (2020).
32. P. Mareri, S. Bastide, N. Binda and A. Crespy, *Compos. Sci. Technol.*, **58**(5), 747 (1998).
33. B. Yegnanarayana, *Artificial neural networks*, PHI Learning Pvt. Ltd. (2009).
34. B. Shahriari, K. Swersky, Z. Wang, R. P. Adams and N. De Freitas, *Proc. IEEE*, **104**(1), 148 (2015).
35. Y. Yu, X. Si, C. Hu and J. Zhang, *Neural Comput.*, **31**, 7 (2019).