Mechanical Systems and Signal Processing 257 (2026) 114584

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/ymssp

Full length article

Physics-guided neural network architectures for modeling nonlinear behaviors: Application to shape memory alloy wires

Wei Dai [a], Biagio Carboni [b], Giuseppe Quaranta [b,*], Yongjun Pan [a,*], Walter Lacarbonara [b]

[a] College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, China
[b] Department of Structural and Geotechnical Engineering, Sapienza University of Rome, Rome, Italy

ARTICLE INFO

Communicated by J. Baqersad

Keywords:
Differential evolution
Long-short term memory
Machine learning
Physics-enhanced neural network
Parametric identification
Shape memory alloy
Transformer

ABSTRACT

A novel physics-guided machine learning framework is proposed for the numerical simulation of complex nonlinear behaviors and to improve the generalization capability of data-driven models under limited information conditions. It combines an a priori physics-based model with a new neural network (NN) architecture designed to compensate for its inaccuracies. A differential evolution algorithm is adopted for the parametric identification of the physics-based model. On the other hand, an original transformer-based NN, enhanced with a long short-term memory (LSTM) feature fusion layer, is employed to capture the inherent aspects of the nonlinear behavior that are overlooked by the physics-based model so as to reduce overall errors.

This approach is applied to simulate the hysteretic tensile behavior of NiTiNOL shape memory alloy wires subject to different loading rates. A comprehensive experimental comparative assessment is carried out between a physics-based (phenomenological) modeling approach and several NN architectures, including the uni- and bi-directional LSTM models, a transformer model, as well as the proposed architecture merging a transformer component with either uni- or bi-directional LSTM model. These NN architectures are employed both in a fully data-driven setting and within the physics-guided framework herein conceived with the purpose of mitigating the inaccuracies arising from the selected phenomenological model and limited data.

The results indicate that the proposed NN architecture connecting LSTM and transformer components achieves the best overall performance. Furthermore, the physics-guided approach consistently improves both accuracy and reliability of the predictions.

1. Introduction

The development of mathematical models capable of simulating nonlinear behaviors of materials and systems is of utmost importance in engineering. Historically this challenge has been addressed through different modeling paradigms [1], namely: (i) white box modeling, in which the model structure is fully defined based on physical principles; (ii) black box modeling, which relies entirely on data-driven techniques to derive models that reproduce observed behaviors without reference to the underlying physics; (iii) gray box modeling, which combines data-driven approaches with physical insight to enhance the accuracy of the models.

Among these, white box (physics-based) modeling has traditionally been the dominant approach in engineering. Its appeal lies in the ability to interpret the modeled phenomena in purely physical terms, as well as in the control it provides over the assumptions

* Corresponding authors.
E-mail addresses: weidai@cqu.edu.cn (W. Dai), biagio.carboni@uniroma1.it (B. Carboni), giuseppe.quaranta@uniroma1.it (G. Quaranta), yongjun.pan@cqu.edu.cn (Y. Pan), walter.lacarbonara@uniroma1.it (W. Lacarbonara).

https://doi.org/10.1016/j.ymssp.2026.114584
Received 23 October 2025; Received in revised form 5 June 2026; Accepted 12 June 2026

and approximations embedded within the mathematical formulation. Nonetheless, the increasing complexity of modern materials and systems, along with the challenges associated with developing comprehensive physics-based models, have contributed to the widespread use of black and gray box (data-driven) models across research and practical applications [2].

In this context, data-driven modeling approaches based on machine learning (ML) techniques are currently garnering increasing attention. These methods have demonstrated strong potential in capturing complex nonlinear behaviors across a wide range of engineering domains, including solid and fluid mechanics [3–5], material science [6,7], structural mechanics and engineering [8–11]. Although a black box modeling approach dominates much of the research in this field, its potential limitations should not be overlooked. In fact, relying on data limits the robustness of this modeling approach because datasets are often sparse, and the models typically fail to generalize in the presence of unseen scenarios since comprehensive datasets covering all possible conditions are not always available. Moreover, although significant theories [12–14] and useful tools [15–17] have been developed to explain the internal workings of black box models, the lack of a clear physical interpretation remains a significant concern, especially when dealing with very sophisticated ML models. The development of gray box modeling approaches, which are aimed to enhance data-driven ML models by integrating physical knowledge, is a promising path to overcome these issues. Depending on how the underlying physics is integrated, physics-enhanced ML models can be classified according to the following categories [18,19]:

- physics-informed ML models, when they explicitly incorporate physics laws (often expressed as differential equations, conservation laws, or constraints) directly into the learning process;
- physics-guided ML models, when their training is guided by a priori physics principles or existing physical models;
- physics-encoded ML models, in which physical knowledge is embedded directly into the model architecture (often as hardwired components) to eliminate the need for additional constraints during training.

Due to their complex nonlinear behavior and relevance in several real-world applications, the numerical simulation of shape memory alloys (SMAs) is one of the areas in which alternative modeling approaches have been explored in the attempt to continuously improve the accuracy of the final predictions. SMAs are intelligent materials that exhibit the shape memory effect or the pesudoelastic behavior – recovering large inelastic strains upon heating or unloading – where their response is strongly dependent on the loading history [20]. Among the available SMAs, NiTiNOL (a Ni–Ti alloy) is the most widely used due to its superelasticity and excellent biocompatibility, finding applications in aerospace [21,22], biomedical devices [23–25], and structural engineering [26–29]. Initial modeling of SMAs adopted a white box strategy based on physics-driven phenomenological formulations. For instance, Zhang and Zhu [27] extended Wilde's hyperelastic SMA model to better simulate the energy dissipation in the proposed SMA-based hysteretic damper. Carboni et al. [30] incorporated a bell-shaped function into the standard Bouc-Wen model to capture the pinching effect in NiTiNOL steel strands under flexure. Pei et al. [31] developed a bi-directional SMA actuator model based on Brinson's constitutive equations, accounting for stress- and temperature-induced martensite fraction evolution to predict the wire displacements under thermoelectric coupling. More recently, black box, fully data-driven approaches have been investigated following the broader availability of experimental data. In this context, ML tools have been mainly adopted to predict phase transformation and hysteresis behavior [32,33]. For example, Honrao et al. [34] employed over 8000 multi-component Ni-Ti data points to uncover how composition and processing affect transformation temperature and strain. To this end, they trained an ensemble ML algorithm leveraging on several algorithms, such as Gaussian processes (GP), support vector regression, random forest regression, gradient boosting regression, and neural networks (NNs). Zadeh et al. [35] compiled a dataset of 3024 Ni-Ti samples (spanning compositions, deformation modes, and crystallographic orientations) and applied gradient boosting algorithm over decision trees to predict transformation strains with strong generalization across alloy variants. Wang and Wang [36] proposed a NN-based SMA model that successfully captures phase transformation effects, providing a data-driven complementary approach to traditional physical constitutive methods. Lenzen and Altay [37] developed a ML-enhanced dynamic model for superelastic SMA wires, utilizing a feedforward NN (FNN) to learn strain-rate-sensitive thermodynamic parameters from macroscopic constitutive simulation data and achieving high-precision parameter identification. Subsequently, they proposed a dual FNN inverse-identification framework with Latin hypercube sampling and transfer learning [38], effectively addressing the ill-posed nature of this inverse identification problem and improving convergence and consistency across multiple loading scenarios.

In addition to white box and black box modeling approaches, gray box models with physics-based predictions enhanced by ML algorithms are emerging, although their application in the field of SMA modeling remains still very limited. For instance, Lenzen and Altay [39] proposed a physics-informed deep operator network for predicting martensite evolution in superelastic SMAs from cyclic tensile tests. To circumvent the need for direct measurement of martensite evolution, they utilized cyclic stress–strain tests. Thus, the physics-based model of Zhu and Zhang [40] was employed to map the estimated martensite fraction to the corresponding stress response, which was directly used in defining the network's loss function. The study by Joy et al. [41] introduces a physics-guided ML framework that leverages experimentally derived thermomechanical data and recurrent NNs (RNNs) to model strain evolution during the actuation of NiTiHf SMAs. Physics-based constraints (such as partial phase transformation behavior and thermoelastic unloading conditions) are incorporated in a weak form as supplementary training data to guide the learning process, rather than being imposed as strict constitutive equations. Khatamsaz et al. [42] developed a ML framework for the optimum design of Ni-Ti SMAs. In particular, statistical data for the design are supplied by GPs, which construct models of the underlying objective function. Their approach integrates physical understanding of diffusion-controlled precipitation and phase transformation behavior into the computational framework by tailoring the GP kernel to reflect known physical trends. Although Khatamsaz et al. [42] describe their approach as a physics-informed ML framework, it is more accurately characterized as a physics-encoded method, since the a priori physical knowledge is embedded into the ML model (specifically into the kernel function of the GP).

In light of the promising results from integrating a priori physics into ML models and the limited application of such approach to SMAs, this paper proposes a novel physics-guided NN for the numerical simulation of complex nonlinear behaviors, with a particular focus on the stress–strain response of NiTiNOL wires. Compared to previous proposals in this field, e.g., [41], the key contributions of this paper lie in both the method of physics integration and the layout of the ML architecture. In the present work, in fact, the ML model is aimed at compensating the residuals of an a priori physics-based model. To this end, a new transformer-based NN enhanced with a long short-term memory–based feature fusion layer is proposed. An extensive comparative assessment using experimental data demonstrates that the proposed NN architecture delivers superior overall performance in simulating the complex nonlinear response of NiTiNOL wires. Moreover, the results show that the physics-guided approach consistently compensates for the inaccuracies of the underlying physical model under limited data.

The remainder of the paper is organized as follows. Section 2 presents the physics-guided ML framework for modeling nonlinear behavior, including the physical prior model (Section 2.1) and the NN-based residual compensation module (Section 2.2). Section 3 applies the approach to the tensile response of NiTiNOL wires under varying loading rates, covering the experimental campaign (Section 3.1), physical modeling (Section 3.2), ML implementation (Section 3.3), and discussion of results (Sections 3.4 and 3.5). Concluding remarks are given in Section 4.

2. Physics-guided machine learning framework

The considered physics-guided ML framework is illustrated in Fig. 1. It consists of an a priori physics-based model augmented by a special NN architecture that compensates for its inherent inaccuracies.

As shown in Fig. 1, the computational procedure starts from the collection of experimental data (step ①), which will later be used to calibrate the a priori physics-based model and to train the NN architecture tailored to reduce its inaccuracies. Next, a physical prior model is formulated, and its relevant parameters are identified (step ②). Following this step, the residuals of the physics-based model (defined as the differences between its numerical estimates and the experimental data) are computed (step ③). These residuals are then organized and used to train the NN that functions as a corrector of the physics-based model (step ④), ultimately aimed at mitigating its errors (step ⑤).

2.1. Physical prior model

The physical prior model $f_{\text{physics}}(\mathbf{x})$ refers to a model governed by parameters $\mathbf{x}$, which may be formulated based on phenomenological relationships (i.e., phenomenological model) or derived from first mechanical principles (i.e., mechanical model). Irrespective of the formulation approach, the parameters of the physical prior model colleted into $\mathbf{x}$ must be initially identified from the available experimental data. Among the alternative techniques available for this task, the Differential Evolution (DE) algorithm proposed by Storn and Price [43] is here adopted.

The DE algorithm is a population-based global optimization metaheuristic algorithm. It iteratively manipulates a population $\mathcal{P}$ of candidate solutions $\mathbf{x}_i = \{x_{i,1}, \dots, x_{i,k}, \dots, x_{i,D}\}$ with $i = 1, \dots, N$ (being D the number of model parameters and N the population size), in order to find the optimal solution $\mathbf{x}_*$ that minimizes an objective function $\varphi(\mathbf{x})$ within a bounded search space defined by the lower and upper limits $\mathbf{x}_l = \{x_{l,1}, \dots, x_{l,k}, \dots, x_{l,D}\}$ and $\mathbf{x}_u = \{x_{u,1}, \dots, x_{u,k}, \dots, x_{u,D}\}$, respectively. Starting from an initial, random distribution of candidate solutions, the DE algorithm applies two main operators to manipulate the current population at each iteration or generation g (being G the total number of generations). The first operator is the mutation, which generates the so-called provisional offspring $\mathbf{v}_i = \{v_{i,1}, \dots, v_{i,k}, \dots, v_{i,D}\}$ by perturbing a candidate solution through the differences between couples of selected individuals in the population. Moreover, the crossover operator acts to obtain the final offspring (also named trial vector) $\mathbf{u}_i = \{u_{i,1}, \dots, u_{i,k}, \dots, u_{i,D}\}$. It is worth noting that this sequence of the evolutionary operators, where mutation is followed by crossover, is consistent with the original formulation by Storn and Price [43] and aligns with most implementations reported in the existing literature [1]. An updated population of candidate solutions is then generated through the selection operator, and the iterative procedure terminates once a specified criterion is met.

The implementation of the DE algorithm for parametric identification of the physical prior model is outlined in Algorithm 1. The pseudo-code shows that the so-called rand/1 mutation scheme is here adopted at each generation in combination with the binomial crossover. Boundary violations are handled by means of a projection scheme through the operator `clip`$(\cdot)$ while a selection stage ensures that only elites are retained for the next generation. In this study, the initial population is generated via Sobol sequences to ensure the uniform sampling of the bounded search space while the objective function $\varphi(\mathbf{x})$ is defined as the mean squared error (MSE) between the experimental data $\mathbf{y}_{\text{exp}}$ and corresponding numerical predictions carried out from the physical prior model $f_{\text{physics}}(\mathbf{x})$. The control parameters of the DE algorithm are selected following commonly recommended guidelines in the literature [1]. The population size is set to $N = 10D$ to ensure sufficient diversity for effective global exploration. The scale factor of the mutation operator is set to $F = 0.5$, providing a balanced trade-off between exploration and exploitation. The crossover probability is set to $P_c = 0.9$ to promote efficient information exchange between offspring and accelerate convergence. All control parameters are kept constant throughout the optimization process.

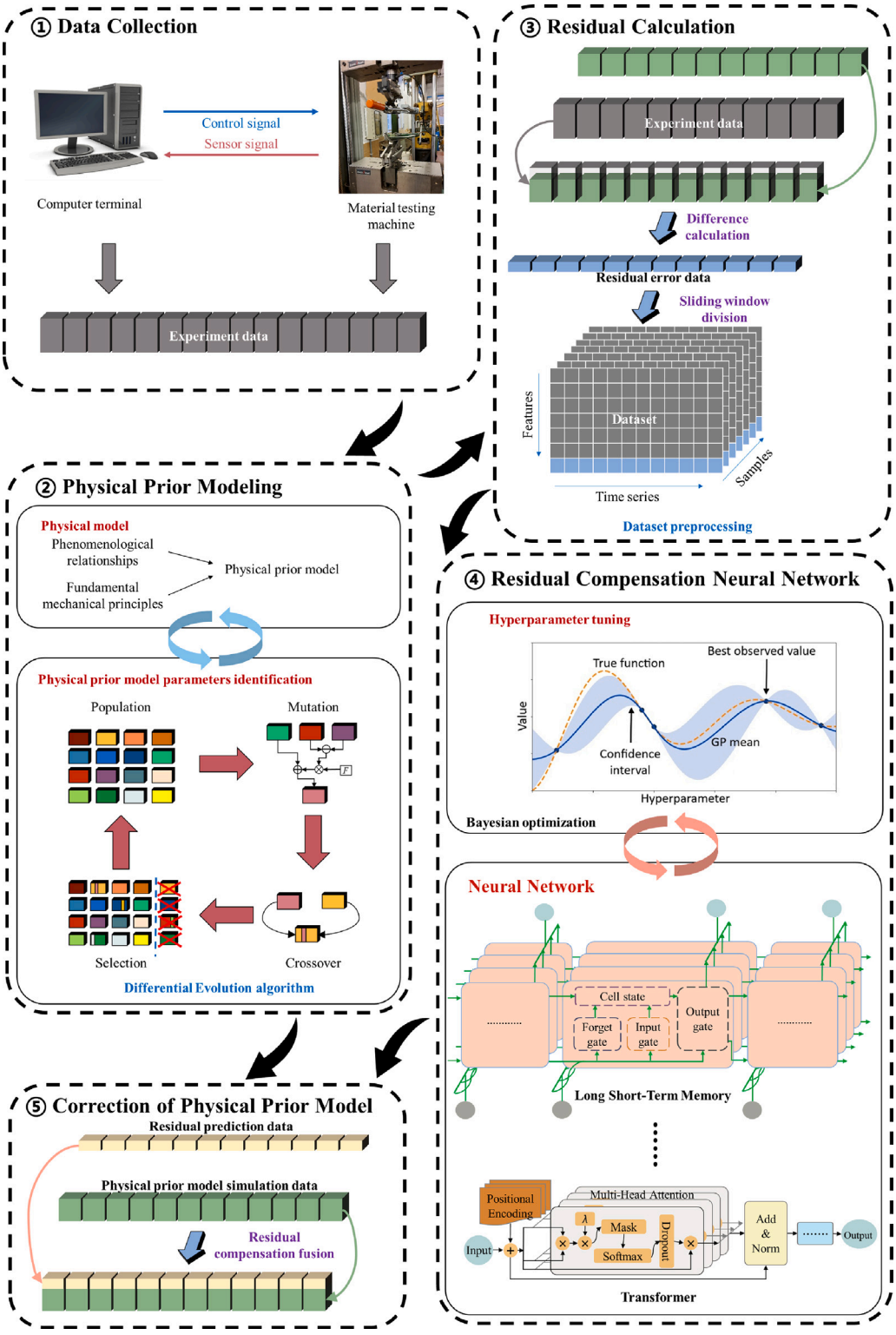


Fig. 1. Flowchart of the physics-guided ML framework.

Algorithm 1 DE algorithm for parametric identification of the physical prior model

Input: Objective function $\varphi(\mathbf{x})$, search space dimensions D, lower bound $\mathbf{x}_l$ and upper bound $\mathbf{x}_u$ of the search space, population size N, max number of generations G, scale factor F, crossover probability P_c

Output: Optimal solution $\mathbf{x}_*$, convergence history

Initialize

a. Generate the initial population via Sobol sequence: $\mathbf{X}_{\text{sobol}} \in [0,1]^D$

b. Map the initial population onto the search space: $\mathbf{x}_i = \mathbf{x}_l + \mathbf{X}_{\text{sobol}}(i) \odot (\mathbf{x}_u - \mathbf{x}_l),\ \forall i = 1, \dots, N$

c. Calculate the initial objective function value (via parallel computing): $\varphi(\mathbf{x}_i),\ \forall i = 1, \dots, N$

for $g = 1$ to G **do**

a. Mutation – generate the provisional offspring $\forall i = 1, \dots, N$

i. Select distinct candidate solutions by generating random indices $r_1, r_2, r_3 \in \{1, \dots, N\}$

ii. Generate provisional offspring: $\mathbf{v}_i = \mathbf{x}_{r_1} + F(\mathbf{x}_{r_2} - \mathbf{x}_{r_3})$

b. Binomial crossover – generate the final offspring $\forall i = 1, \dots, N$

i. Randomly select basis vector $\mathbf{x}_b$

ii. Generate final offspring:

$$u_{i,k} = \begin{cases} v_{i,k} & \text{if } \texttt{rand}() < P_c \ \ \% \ \texttt{rand}() \text{ is a uniform random number generator} \\ x_{b,k} & \text{otherwise} \end{cases}$$

c. Handling boundary violations: $u_{i,k} \leftarrow \texttt{clip}(u_{i,k}, x_{l,k}, x_{u,k}),\quad \forall i = 1, \dots, N$

d. Selection – Retain elites only

e. Update convergence history

f. Termination condition check

end for

Return

Carry out the physical prior model parameters $\mathbf{x}_* = \arg\min \varphi(\mathbf{x}_i; i = 1, \dots, N)$ and convergence history

2.2. Residuals compensation module

Physics-based models, whether phenomenological or mechanical, are inherently limited, as they often overlook or approximate some aspects of actual behavior, thereby limiting their predictive accuracy and generalizability. To address these deficiencies, a residuals compensation module is integrated as the core component of the hybrid framework. This is formally defined as follows:

$$\hat{\mathbf{y}} = f_{\text{physics}}(\mathbf{x}_*) + f_{\text{residual}}(\boldsymbol{\theta}|\mathbf{x}_*), \tag{1}$$

where $\hat{\mathbf{y}}$ denotes the predictions while $f_{\text{residual}}(\boldsymbol{\theta}|\mathbf{x}_*)$ represents a NN-based model parameterized by $\boldsymbol{\theta}$ and trained to compensate for the residuals of the physics-based model $f_{\text{physics}}(\mathbf{x}_*)$ after parametric identification (i.e., residual learning). This formulation seeks to preserve physical consistency while accounting for unmodeled phenomena that lie beyond the predictive capacity of the physical prior model. Crucially, this hybrid paradigm maintains physical interpretability while mitigating catastrophic extrapolation errors endemic to pure data-driven surrogates. The considered NN architectures serving as residual compensation module are presented hereafter.

2.2.1. Uni-directional long short-term memory model

The uni-directional Long Short-Term Memory (LSTM) model [44] addresses the limitation of the standard RNN in modeling long-range dependencies within sequential data [45], a problem commonly referred to as the vanishing gradient phenomenon. Let $\mathbf{z}_1, \dots, \mathbf{z}_t, \dots, \mathbf{z}_T$ be the sequence of inputs. The LSTM processes the sequence in the forward direction from $t = 1$ to $t = T$, considering for each step t the hidden state $\mathbf{h}_{t-1}$ and the cell state $\mathbf{c}_{t-1}$ at step $t-1$ given $\mathbf{z}_t$. In so doing, it computes the new hidden state $\mathbf{h}_t$ as follows:

$$\mathbf{i}_t = \sigma\left(\mathbf{W}_i \mathbf{z}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{b}_i\right) \quad \text{(input gate)}, \tag{2a}$$

$$\mathbf{f}_t = \sigma\left(\mathbf{W}_f \mathbf{z}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f\right) \quad \text{(forget gate)}, \tag{2b}$$

$$\mathbf{o}_t = \sigma\left(\mathbf{W}_o \mathbf{z}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{b}_o\right) \quad \text{(output gate)}, \tag{2c}$$

$$\tilde{\mathbf{c}}_t = \tanh\left(\mathbf{W}_c \mathbf{z}_t + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{b}_c\right) \quad \text{(cell input)}, \tag{2d}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \quad \text{(cell state update)}, \tag{2e}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh\left(\mathbf{c}_t\right) \quad \text{(hidden state update)}, \tag{2f}$$

where $\sigma(\cdot)$ is the sigmoid activation function while $\mathbf{W}_{\blacksquare}$ and $\mathbf{U}_{\blacksquare}$ are learnable weight matrices and $\mathbf{b}_{\blacksquare}$ are learnable biases. Through such mechanisms consisting of input, forget, and output gates, the LSTM model regulates information flow into retention within, and emission from the cell state. In particular, the input gate modulates new information integration into the cell state, the forget gate filters irrelevant historical data, and the output gate controls state updates. This architecture mitigates vanishing gradient problem while preserving critical temporal patterns. Its selective memory management based on three gates enables robust modeling of complex nonlinear phenomena. However, while the computational cell of the LSTM addresses the long-term dependency issue more effectively than that of the RNN, Eq. (2) also underscores that its operation depends on a greater number of hyperparameters, leading to increased computational effort during training. The output of the LSTM is obtained via linear transformation of the updated hidden state $\mathbf{h}_t$ (through a fully connected layer) and provides the error compensation of the physics-based model at each step t.

2.2.2. Bi-directional long short-term memory model

A bi-directional LSTM (BiLSTM) model enhances sequence modeling by processing data in both forward and reverse directions [46]. A BiLSTM consists of two separate LSTM layers, namely a forward LSTM layer that processes the sequence from $t = 1$ to $t = T$ and a backward LSTM layer that processes the sequence from $t = T$ to $t = 1$. Therefore, the forward LSTM layer computes the hidden state $\vec{\mathbf{h}}_t$ processing the sequence from $t = 1$ to $t = T$ as follows:

$$\vec{\mathbf{h}}_t = \texttt{LSTM}(\vec{\mathbf{h}}_{t-1}, \vec{\mathbf{c}}_{t-1} | \mathbf{z}_t), \tag{3}$$

where $\vec{\mathbf{h}}_{t-1}$ and $\vec{\mathbf{c}}_{t-1}$ are the hidden state and the cell state at step $t-1$ proceeding from $t = 1$ to $t = T$, respectively, while $\texttt{LSTM}(\cdot)$ represents the set of operations in Eq. (2). The backward LSTM layer instead computes the hidden state $\overleftarrow{\mathbf{h}}_t$ processing the sequence from $t = T$ to $t = 1$ as follows:

$$\overleftarrow{\mathbf{h}}_t = \texttt{LSTM}(\overleftarrow{\mathbf{h}}_{t+1}, \overleftarrow{\mathbf{c}}_{t+1} | \mathbf{z}_t), \tag{4}$$

which still implies the use of Eq. (2) but proceeding in the reverse direction. At each step t, the updated hidden state is the concatenation of the forward and backward hidden states:

$$\mathbf{h}_t = \mathbf{W}_f \vec{\mathbf{h}}_t + \mathbf{W}_b \overleftarrow{\mathbf{h}}_t + \mathbf{b}_t, \tag{5}$$

where $\mathbf{W}_f$ and $\mathbf{W}_b$ are learnable weights and $\mathbf{b}_t$ is a bias term. The outputs of the BiLSTM are obtained through a linear transformation of the updated hidden state $\mathbf{h}_t$ (through a fully connected layer), yielding the error compensation for the physics-based model at each step t. The bi-directional formulation is especially effective in characterizing the directional asymmetries that may arise in nonlinear behaviors.

2.2.3. Transformer model

Unlike recurrent architectures (such as standard and gated RNNs), a transformer (Trans) employs a self-attention mechanism [47] to model global dependencies without sequential processing constraints [48]. During the encoding stage, for a single attention head, denoted as $\texttt{head}_j$, the input matrix $\mathbf{Z}$ is linearly projected into queries $\mathbf{Q}_j$, keys $\mathbf{K}_j$, and values $\mathbf{V}_j$ via learnable weights:

$$\mathbf{Q}_j = \mathbf{Z}\mathbf{W}_{Q,j}, \quad \mathbf{K}_j = \mathbf{Z}\mathbf{W}_{K,j}, \quad \mathbf{V}_j = \mathbf{Z}\mathbf{W}_{V,j}, \tag{6}$$

where $\mathbf{W}_{Q,j}$, $\mathbf{W}_{K,j}$ and $\mathbf{W}_{V,j}$ are the learned projection matrices specific to $\texttt{head}_j$. Multi-head self-attention splits the representation into H attention heads, computes attention in parallel, concatenates the results, and projects them back to the model dimension:

$$\texttt{MultiHeadSelfAttention}(\mathbf{Z}) = \texttt{Concatenate}(\texttt{head}_1, \ldots, \texttt{head}_j \ldots, \texttt{head}_H)\mathbf{W}_O, \tag{7}$$

where each attention head is computed as:

$$\texttt{head}_j = \texttt{SelfAttention}(\mathbf{Q}_j, \mathbf{K}_j, \mathbf{V}_j) = \texttt{Dropout}\left(\texttt{Softmax}\left(\texttt{Mask}(\lambda \mathbf{Q}_j \mathbf{K}_j^{\top}, \mathbf{M}_j)\right), p\right)\mathbf{V}_j. \tag{8}$$

Here, λ represents the scaling factor and $\mathbf{W}_O$ is the learned output projection matrix. The function $\texttt{Mask}(\cdot)$ applies an attention mask to restrict attention to specific positions. The $\texttt{Softmax}(\cdot)$ function converts attention scores into a probability distribution across the sequence, ensuring the weights sum to 1. $\texttt{Dropout}(\cdot, p)$ is a regularization technique that randomly sets a proportion p of the attention weights to zero during training to prevent overfitting. Finally, $\texttt{Concatenate}(\cdot)$ stacks the head outputs before applying the final projection. After computing the multi-head attention, the result is added to the original input via a residual connection and normalized (Add&Norm). This is followed by a position-wise FNN, another residual connection, and a subsequent normalization step, together forming the core structure of the transformer encoder block. For data with a uni-directional structure, such as that considered in the following experimental application, an encoder-only architecture is commonly used. In the present study, the input sequence is first augmented with positional encoding to preserve order information, and then processed through two multi-head self-attention layers.

xxxxx

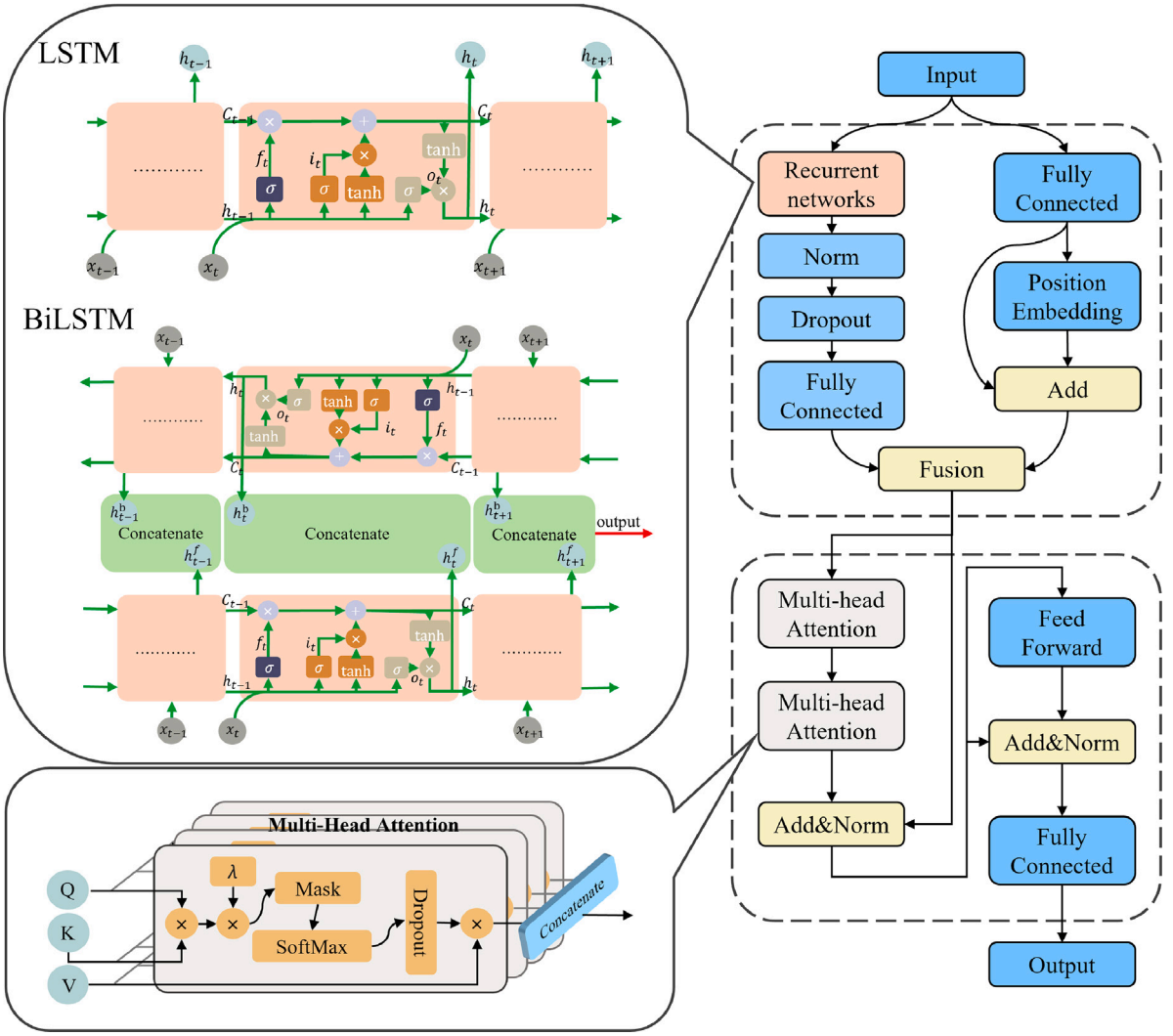


Fig. 2. Proposed dual-branch architecture combining a transformer component with either uni- or bi-directional LSTM model.

2.2.4. Proposed long short-term memory-transformer models

This paper presents an original dual-branch architecture that combine a transformer component with either uni-directional LSTM (LSTM-Trans) or bi-direction LSTM (BiLSTM-Trans). In the input stage of the models, either alternative is employed to encode the data in each window. Compared with the LSTM-Trans, the BiLSTM-Trans further incorporates information from both past and future time steps within the window, enabling a more comprehensive representation of the data sequence. These features are then integrated with position-sensitive global features generated by a positional encoding layer. The fused features are subsequently fed into the subsequent multi-head self-attention layers, which capture long-range global dependencies. This feature fusion structure is not only intuitive but also fully utilizes the advantages of recurrent connections in local temporal modeling and the ability of Trans to capture global information dependencies within a window. This enables subsequent layers to utilize local information while focusing on global dependencies, thereby improving the ability to model complex dynamic systems. Compared to a parallel structure, the feature fusion structure is more compact and does not require two separate full branches.

The framework is shown in Fig. 2. After passing through either the uni-directional or bi-directional LSTM layer, normalization and dropout stages, the input data is mapped to a dimension consistent with the transformer branch through a fully connected layer. Meanwhile, the input data is directly mapped through a fully connected layer and positional encoding to form preliminary global features. The features from the two branches are then fused through addition and fed into the subsequent transformer module (consisting of two layers of multi-head self-attention and a FNN). Finally, the prediction results are carried out through layer normalization and a fully connected mapping.

2.2.5. Bayesian hyperparameters tuning

The NN architectures considered in this study involve numerous hyperparameters, whose appropriate values must be carefully selected. To address this, a global optimization technique well-suited for black-box functions is adopted, namely, a Bayesian optimization. It is particularly effective for hyperparameter tuning in NNs, where evaluations are computationally expensive and gradients are unavailable [49]. In this approach, the objective function, here defined as the root mean squared error (RMSE) of predictions on the validation set for a given hyperparameter set $\boldsymbol{\theta}$, is approximated by a surrogate model. An acquisition function then guides the selection of the next hyperparameter configuration to evaluate, balancing exploration and exploitation. Through this iterative process, Bayesian optimization efficiently searches for optimal hyperparameters while minimizing the number of costly model evaluations.

During the Bayesian optimization process, GP is used to construct the surrogate model. Among the different acquisition function types, the Expected Improvement Plus technique is here employed, which is particularly suitable for handling the randomness of training results across different batches in deep learning hyperparameter optimization.

3. Experimental application

All NN architectures considered in the present work are employed to model the nonlinear tensile behavior of NiTiNOL wires under varying loading rates. These models are implemented using both a fully data-driven approach (under which they are denoted as LSTM, BiLSTM, Trans, LSTM-Trans, and BiLSTM-Trans) as well as following the illustrated physics-guided approach (in such a case, they are labeled as Pg-LSTM, Pg-BiLSTM, Pg-Trans, Pg-LSTM-Trans, and Pg-BiLSTM-Trans, respectively). In this context, the employed NN architectures are intended to compensate for the limitations of a rate-independent uniaxial phenomenological model, which is simple but unable to capture strain rate effects. The entire evaluation pipeline was executed on a system equipped with an Intel i5-14600KF CPU and an NVIDIA GeForce RTX 4080 SUPER GPU.

3.1. Experimental dataset

To comprehensively characterize the mechanical properties of the NiTiNOL wire, random-amplitude tensile cyclic tests were conducted at increasing loading rates (i.e., 0.2, 0.4, 0.6, 0.8, 1, 1.5, 2.0 mm/s) at room temperature. The experiments were carried out using a Zwick/Roell Allround Line Z250 universal testing machine shown in Fig. 3a, which has a maximum load capacity of 250 kN, a crosshead speed range of 0.00005–600 mm/min, and a maximum return speed of 1000 mm/min. The test specimen has an initial length of 100 mm and a diameter of 0.6 mm. The influence of wire diameter was not explicitly investigated in this study. In fact, by formulating the response in terms of stress–strain relationship, geometric scaling effects associated with the cross-sectional area are implicitly accounted for, and the identified behavior primarily reflects intrinsic material properties. The wire specimens were not clamped using double capstan grips. Instead, commercial double-eye grips were employed, with particular attention devoted to minimizing boundary effects. Each wire was wrapped around a pin of the grip and clamped by compression over a length of at least 50 times the wire diameter. This configuration enables a gradual transfer of the applied load through friction along an extended contact length, thereby reducing local stress concentrations near the grips and promoting a more uniform stress distribution across the active gauge length. The effect of temperature variations induced by strain rate is implicitly included in the experimental data used to calibrate the models. Although temperature is not explicitly modeled as a separate variable, its influence on the stress–strain response is reflected in the measurements. The variation of the tensile amplitudes is shown in Fig. 3b.

Tensile tests exhibit the typical pseudoelastic behavior, as shown in Fig. 3c. The stress–strain curve first displays a linear relationship between stress and strain in the initial austenitic elastic deformation stage as the material remains predominantly austenitic with fully recoverable deformation. The curve then enters a prolonged plateau region in the phase transformation plateau stage, where part of the austenite transforms into strain-induced martensite under tension when the stress reaches a critical value, forming a stress plateau with almost constant load but continuous strain increase. With continuous loading, most of the material transforms into martensite, and the curve rises again to enter the martensitic elastic deformation stage, reflecting the elastic deformation of the martensite phase. During the unloading stage, the decrease in stress promotes the reverse transformation of part of the strain-induced martensite into austenite, accompanied by significant hysteresis during deformation recovery, which demonstrates the characteristics of energy dissipation and the shape memory effect.

3.2. Selected phenomenological model and parametric identification

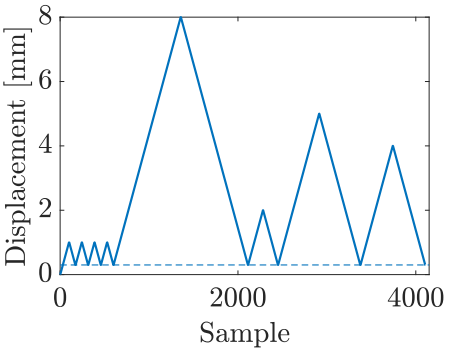
This study employs the rate-independent model proposed by Charalampakis and Tsiatas [50] as the phenomenological model to describe the stress–strain hysteresis of NiTiNOL wires. This constitutive model captures essential features of the material response, establishing a robust foundation for residual compensation. The governing equations are the following:

$$\sigma_w = \int_0^t \dot{\sigma}_w \mathrm{d}t, \tag{9}$$

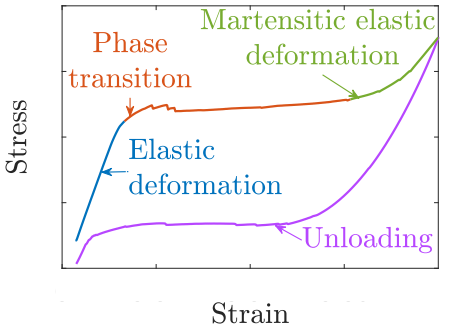
$$\dot{\sigma}_w = \left(1 - s(\varepsilon_w)\right) E \left[\dot{\varepsilon}_w - |\dot{\varepsilon}_w| \operatorname{sgn}(\sigma_w - \beta_w) \cdot \left(\frac{|\sigma_w - \beta_w|}{Y}\right)^n\right] + s(\varepsilon_w) E_m \dot{\varepsilon}_w, \tag{10}$$

$$\beta_w = E\alpha \left[\varepsilon_w - \frac{\sigma_w}{E} + f_T \tanh(a\varepsilon_w)\mathrm{H}(-\varepsilon_w \dot{\varepsilon}_w)\right], \tag{11}$$

(a) Tensile test device.



(b) Applied displacement.



(c) Hysteresis cycle.

Fig. 3. Experimental testing of NiTiNOL wires: (a) testing machine; (b) loading history; (c) general shape of the hysteresis cycle.

Table 1
Search space and parametric identification of the Charalampkis and Tsiatas's model [50].

	E [MPa]	Y [MPa]	n	E_m [MPa]	α	f_T	a	c	ε_t	MSE
$\mathbf{x}_l$	200,000	300	0.5	10,000	0.001	1	100	40	0.01	–
$\mathbf{x}_u$	800,000	600	3	40,000	0.01	5	500	100	0.2	–
$\mathbf{x}_*$	48, 233.241	470.432	1.798	17, 962.050	0.00851	1.630	120.883	64.058	0.0579	3.191%

$$s(\varepsilon_w) = \frac{\tanh\left(c\left(|\varepsilon_w| - \varepsilon_t\right)\right) + 1}{2}, \tag{12}$$

The model described by Eqs. (9)–(12) is defined by nine parameters $\mathbf{x} = \{E, Y, n, E_m, \alpha, f_T, a, c, \varepsilon_t\}$, where ε_w denotes the tensile strain of the NiTiNOL wire, $\dot{\varepsilon}_w$ represents the strain rate of wire deformation, $\dot{\sigma}_w$ is the stress change rate of the NiTiNOL wire at ε_w, and σ_w is obtained by solving Eq. (10) using a suitable numerical solver (such as the Euler method, Runge–Kutta method, or Adams–Bashforth-Moulton method). The parameters E and E_m characterize the material stiffness in the austenite phase and fully martensite phase, respectively, whereas ε_t is the strain value at the transition midpoint between the Graesser-Cozzarelli term (the first term of Eq. (10)) and the martensite term (the second term of Eq. (10)). The parameter Y denotes the critical yield stress of the material, serving as a threshold parameter to determine the deformation stage. The post-yield stiffness of the material is controlled by α while a regulates the smoothness of the material hysteresis curve near the origin. Furthermore, n and c rule the steepness of the transition from the initial elastic stage to the post-yield stage and the transition between the Graesser-Cozzarelli term and martensite term, respectively. Finally, f_T acts as a balance control between twinning hysteresis and superelasticity. Table 1 outlines the parameter search space for $\mathbf{x} = \{E, Y, n, E_m, \alpha, f_T, a, c, \varepsilon_t\}$, regulating the phenomenological model described by Eqs. (9)–(12). The results of the parametric identification consisting of the optimal value of $\mathbf{x}$ are also provided.

This physics-based, phenomenological model proposed by Charalampakis and Tsiatas [50] has been previously adopted to model a variety of devices that employ SMAs [28,51–53]. Nonetheless, as a rate-independent phenomenological model, it predicts the same response regardless of the loading rate. Consequently, the stress response depends exclusively on strain variations, leading to hysteresis loops that remain unchanged for different loading rates. Given that SMAs inherently exhibit rate-dependent behavior, the present work then explores the feasibility of enhancing this versatile rate-independent phenomenological model through an ML-based approach. Indeed, the ultimate objective is to demonstrate the capability of the proposed ML framework to compensate for a simplified physical model. From this perspective, residual learning based on Eq. (1) ensures a physically grounded response, with the data-driven component focused on capturing discrepancies arising from rate-dependent effects. This separation helps alleviate

Table 2
Optimal hyperparameter combinations for all NN-based approaches (including both a fully data-driven approach and the physics-guided approach) and inference time across the test set.

Model	Layout				Training		Complexity		Inference time [s]
	n_{unit}	H	d	r_{drop}	$L_{initial}$	β	Layers	Learnable parameters	
LSTM	197	–	–	0.089628	0.0013079	$2.06911 \cdot 10^{-5}$	6	$158.9 \cdot 10^3$	6.89
BiLSTM	117	–	–	0.14882	0.0042809	$1.1023 \cdot 10^{-4}$	6	$113.9 \cdot 10^3$	6.79
Trans	–	5	31	–	0.00062058	$4.2291 \cdot 10^{-5}$	15	$403.3 \cdot 10^3$	5.94
LSTM-Trans	167	9	12	0.033872	0.0018633	$2.7736 \cdot 10^{-5}$	20	$332.5 \cdot 10^3$	9.88
BiLSTM-Trans	253	12	13	0.014262	0.0003748	$8.3524 \cdot 10^{-5}$	20	$1000 \cdot 10^3$	17.01
Pg-LSTM	411	–	–	0.034342	0.0015149	$1.4602 \cdot 10^{-5}$	6	$683.4 \cdot 10^3$	12.13
Pg-BiLSTM	238	–	–	0.040977	0.0023788	$6.55 \cdot 10^{-5}$	6	$462.1 \cdot 10^3$	10.93
Pg-Trans	–	13	8	–	0.003317	$1.15 \cdot 10^{-8}$	15	$185.7 \cdot 10^3$	6.28
Pg-LSTM-Trans	191	5	20	0.059083	0.00080977	$5.4116 \cdot 10^{-4}$	20	$340.7 \cdot 10^3$	9.69
Pg-BiLSTM-Trans	116	15	21	0.20199	0.00053405	$3.4487 \cdot 10^{-6}$	20	$1800 \cdot 10^3$	13.93

identifiability issues and improves robustness when working with incomplete experimental data. Despite its wide adoption, it should be noted that the phenomenological model proposed by Charalampakis and Tsiatas [50] is not thermodynamically consistent, and the residual compensation provided by the ML module does not enforce this condition. Strict fulfillment of thermodynamic principles would require the incorporation of additional constraints into the learning process. However, the goal is not to develop a thermodynamically rigorous constitutive model, but rather to investigate whether an ML-based residual correction can effectively compensate for the predictive limitations of an existing engineering-oriented physics-based model while preserving its formulation and ease of implementation. In this context, the physics-based component provides a physically meaningful baseline response and guides the learning process toward accurate mechanical behavior. This, in turn, helps reduce the amount of training data required and improves generalization capabilities compared to purely data-driven approaches.

3.3. Training of the neural models

The experimental tensile testing of the NiTiNOL wire is a dynamic process. To fully exploit the information in the time series data and capture the temporal dependencies, a sliding window method with a window step of 1 for data partitioning is employed. The sliding window length is defined based on a fixed 1% strain increment, since the experimental sampling frequency is adjusted according to the applied tensile force rate, resulting in the same number of data points being considered under different loading rates. This strain-based definition ensures consistency in the information collected and enables the networks to learn rate-dependent hysteretic behavior in a physically meaningful way. To ensure that the training, validation, and test sets include experimental data across the full range of speeds and that the sample distribution within each experiment is consistent with the overall data distribution, samples from each experiment are divided at the same ratio. The time-window samples of the last cycle in each experiment (i.e., the cycle exhibiting amplitude variations of 0.3–4–0.3 mm in Fig. 3b) and at different loading rates are used as the test set, since this loading cycle does not appear in the remaining part of the record and thus provides a valuable case for evaluating the model's generalization ability. The remaining samples are randomly shuffled and divided into training and validation sets at a 7:3 ratio. Randomization of windows helps improve model generalization, enhance robustness, reduce training order bias, and prevent the model from overfitting to local time-period data. A unified random seed is set to ensure results reproducibility and control variables to avoid the impact of randomization on algorithm comparison.

To address the inconsistency in sampling rates of cyclic test data at different tensile speeds, in addition to strain and strain rate, the sample input introduces a time interval feature (i.e., adding a "time difference" or "sampling interval" feature to the model input) to enable the model to know the actual interval between each time step. This allows the model to adjust predictions via additional time information even when input data have different sampling rates. Finally, input samples are normalized as follows: strain values to the range $[0, 1]$, strain rate to $[-1, 1]$, and time interval to $[0, 1]$. For output samples, stress is normalized to $[0, 1]$, while the residual stress-output of the NN-based compensation module is standardized to prevent outliers from affecting the results.

The NN architectures considered in this study involve several hyperparameters, including the number of LSTM units (n_{unit}), the number of self-attention heads (H), the dimension of each self-attention head (d), dropout rate (r_{drop}), initial learning rate ($L_{initial}$), and regularization coefficient (β). The hyperparameters were evaluated through 50 iterations of the Bayesian optimization. The optimal feasible hyperparameter combinations ruling the layout and the learning process of the NN architectures are listed in Table 2, together with the inference time across the test set.

3.4. Results

The obtained results are analyzed next to evaluate the accuracy of all NN-based models, implemented using either a fully data-driven or the devised physics-guided approach, with the aim of assessing their absolute performance and isolating the accuracy gain resulting from the integration of physics-based priors.

Fig. 4 compares the physics-based phenomenological model with the different NN architectures, each implemented using either a fully data-driven or a physics-guided approach, across common error metrics on both the training and test sets. Specifically, the

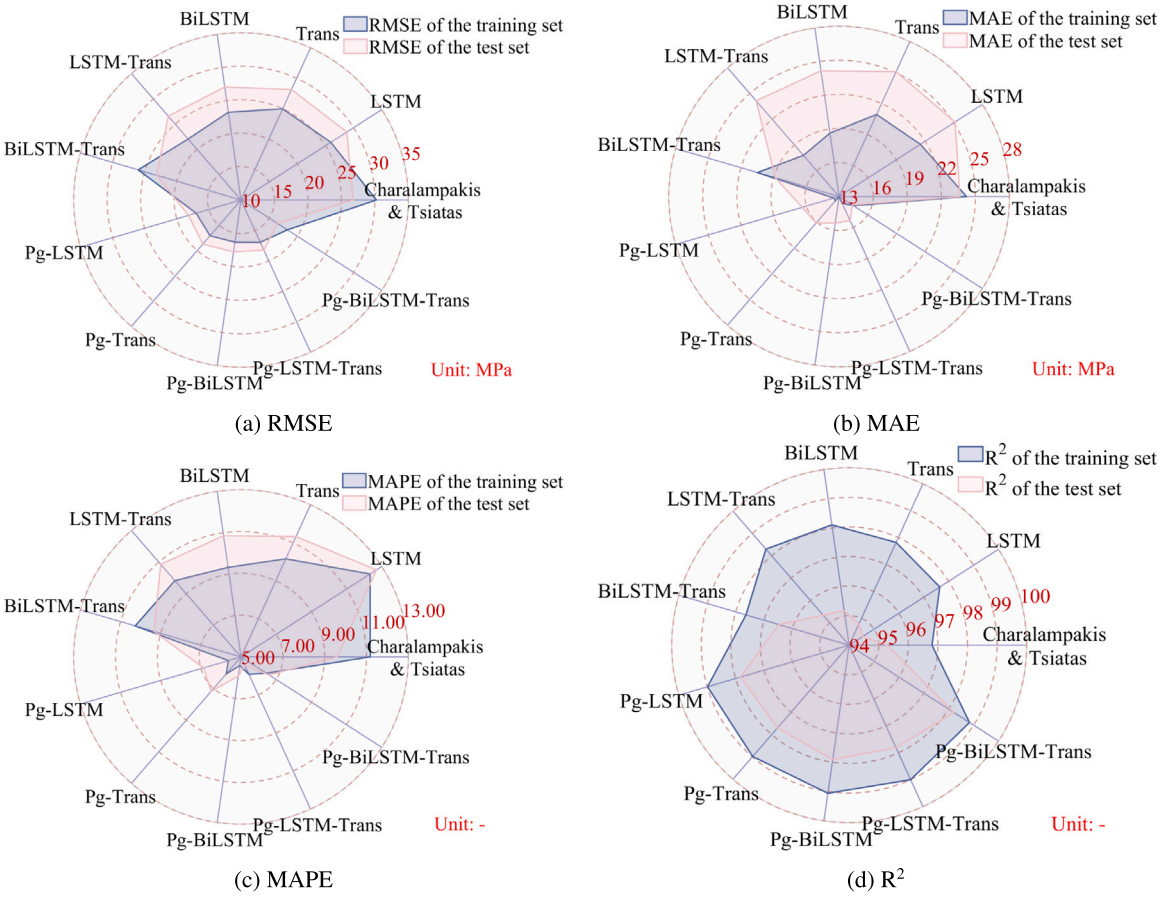


Fig. 4. Error metrics across different approaches and NN architectures: (a) RMSE; (b) MAE; (c) MAPE; (d) R^2.

RMSE assesses the model's overall accuracy by computing the square root of the average squared differences between predictions and observations. The mean absolute error (MAE) provides the average magnitude of these errors without regard to sign, offering a straightforward measure of predictive precision. The mean absolute percentage error (MAPE) expresses the average absolute error as a percentage of the measured values, facilitating comparison across different scales. The coefficient of determination (R^2) measures the proportion of variance in the experimental stress data explained by the model, indicating the strength of correlation between predicted and actual values.

A clear performance gap is observed in Fig. 4 between the phenomenological model and the fully data-driven NN models. Specifically, the phenomenological model proposed by Charalampakis and Tsiatas [50] generally demonstrates lower accuracy on the training set compared to the fully data-driven NN-based architectures, reflecting the limitations of its simpler structure. On the other hand, it outperforms most of the fully data-driven NN-based architectures on the test set across several error metrics. The robustness of the phenomenological model stems from its physics-based foundation, which enables it to approximate the macroscopic hysteretic response under conditions not seen before (i.e., the 4 mm amplitude cycle shown in Fig. 3b). Among the purely data-driven NNs, LSTM ranks lowest, while the proposed BiLSTM-Trans demonstrates markedly superior overall performance on the test set. This comparative assessment restricted to a fully data-driven implementation clearly indicates that the proposed hybrid dual-branch architecture provides a more stable prediction of the force–displacement relationship of NiTiNOL wires, with both forward and backward processing playing a key role toward limiting local overfitting. The performance of the conventional BiLSTM model also deserves consideration, as it is satisfactory despite being the less sophisticated architecture, consisting of only $113.9 \cdot 10^3$ learnable parameters. Ranking the fully data-driven NN models by their test set R^2 scores results in the following order: BiLSTM-Trans, LSTM-Trans, BiLSTM, Trans, and then LSTM.

Fig. 4 also demonstrates that integrating physics priors uniformly boosts the performance of data-driven methods. Across all physics-guided NN models, RMSE and MAE values for the test set drop by over 20%, MAPE decreases by approximately 40%, and R^2 increases from around 95% or lower to nearly 98%. This confirms that residuals correction – anchored by the physics prior – focuses the NN on local nonlinear details, simplifying the learning task and improving precision while preserving global trends.

By leveraging on the physics-guided approach, the proposed Pg-BiLSTM-Trans model exhibits further improvements in both accuracy and stability. On the test set, it achieves the lowest RMSE (yielding a value of 16.50 MPa) and MAE (which equals 14.38 MPa)

among all considered models, together with a high R^2 value (equal to 98.21%), indicating strong generalization capability. It is noted that Pg-BiLSTM-Trans shows lower performance on the training set compared with Pg-LSTM-Trans. This can be attributed to the bi-directional temporal encoding, which imposes stronger regularization. However, while this limits training accuracy, it effectively improves prediction stability and reduces nonphysical fluctuations in the force–displacement response on the test set for unseen data.

Considering both the fully data-driven and physics-guided settings across all error metrics, the presented dual-branch architecture combining Trans component with either LSTM or BiLSTM module demonstrates superior performance in modeling the nonlinear behavior of NiTiNOL wires. However, Table 2 shows that such a new architecture requires larger computational effort for inference than conventional approaches. Among the existing NN architectures, the conventional BiLSTM also yields satisfactory results despite the relatively low number of learnable parameters. It is also confirmed that, when guided by physics, the data-driven approach achieves higher accuracy and better generalization capability.

The analysis of Figs. 5 and 6 provide mechanical insights about the performance of both Pg-LSTM-Trans and Pg-BiLSTM-Trans architectures. Figs. 5 and 6 compare the experimental and numerical results for a loading rate lower and larger than 1 mm/s, respectively. A general overview of Figs. 5 and 6 confirms that the hysteresis loops obtained by means of the phenomenological model by Charalampakis and Tsiatas [50] are nearly indistinguishable across different loading rates. Being a rate-independent phenomenological model, it is insensitive to the tensile rate and thus its stress predictions depend solely on strain variations. Conversely, the shape of the force–displacement cycles produced by the proposed physics-guided models shows a much larger sensitivity with respect to the loading rate.

During the elastic phase of the loading branch, the Charalampakis and Tsiatas's model [50] overestimates the initial stiffness, especially at loading rate larger than 1 mm/s, with an excessively steep initial slope that gradually decreases, resulting in a curved rather than linear response. The turning point also appears overly smooth and lacks definition. In contrast, the Pg-BiLSTM-Trans and Pg-LSTM-Trans models match the experimental slope of the initial stiffness and the inflection point with high precision.

In the phase transformation plateau region, both the phenomenological model by Charalampakis and Tsiatas [50] and the proposed physics-guided NN models exhibit varying degrees of overshoot. The MAE value during the loading branch occurs in the martensitic elastic deformation stage, where the model by Charalampakis and Tsiatas [50] reaches a peak error of approximately 41 MPa at a tensile rate of 1 mm/s. In comparison, the Pg-LSTM–Trans model provides a more accurate prediction, with a MAE in the loading branch of about 28 MPa, corresponding to a relative improvement of 31.7%. The Pg-BiLSTM–Trans model further improves the accuracy, reducing the MAE in the loading branch to approximately 21 MPa, thus resulting into a relative improvement of 48.8%.

During unloading, the overshoot observed in the previous stage leads to an abnormally steep initial unloading slope when using the model by Charalampakis and Tsiatas [50]. This behavior alters the shape of the hysteresis loop from the experimentally observed sharp-tipped form to a flatter profile. This results in a poor agreement in the unloading trajectory, with significant loop opening and distorted energy dissipation characteristics, especially at loading rate larger than 1 mm/s. In contrast, both the Pg-BiLSTM-Trans and Pg-LSTM-Trans models more closely describe the entire experimental unloading path, yielding a well-closed hysteresis loop. This can be properly quantified using the RMSE. Indeed, the maximum RMSE during the unloading branch for the model by Charalampakis and Tsiatas [50] is approximately 37.06 MPa with a tensile rate of 2.0 mm/s. This cumulated deviation is drastically reduced to only 11.77 MPa and 11.22 MPa when using the proposed Pg-LSTM–Trans and Pg-BiLSTM–Trans models, respectively, corresponding to relative improvements of 68.24% and 69.72%.

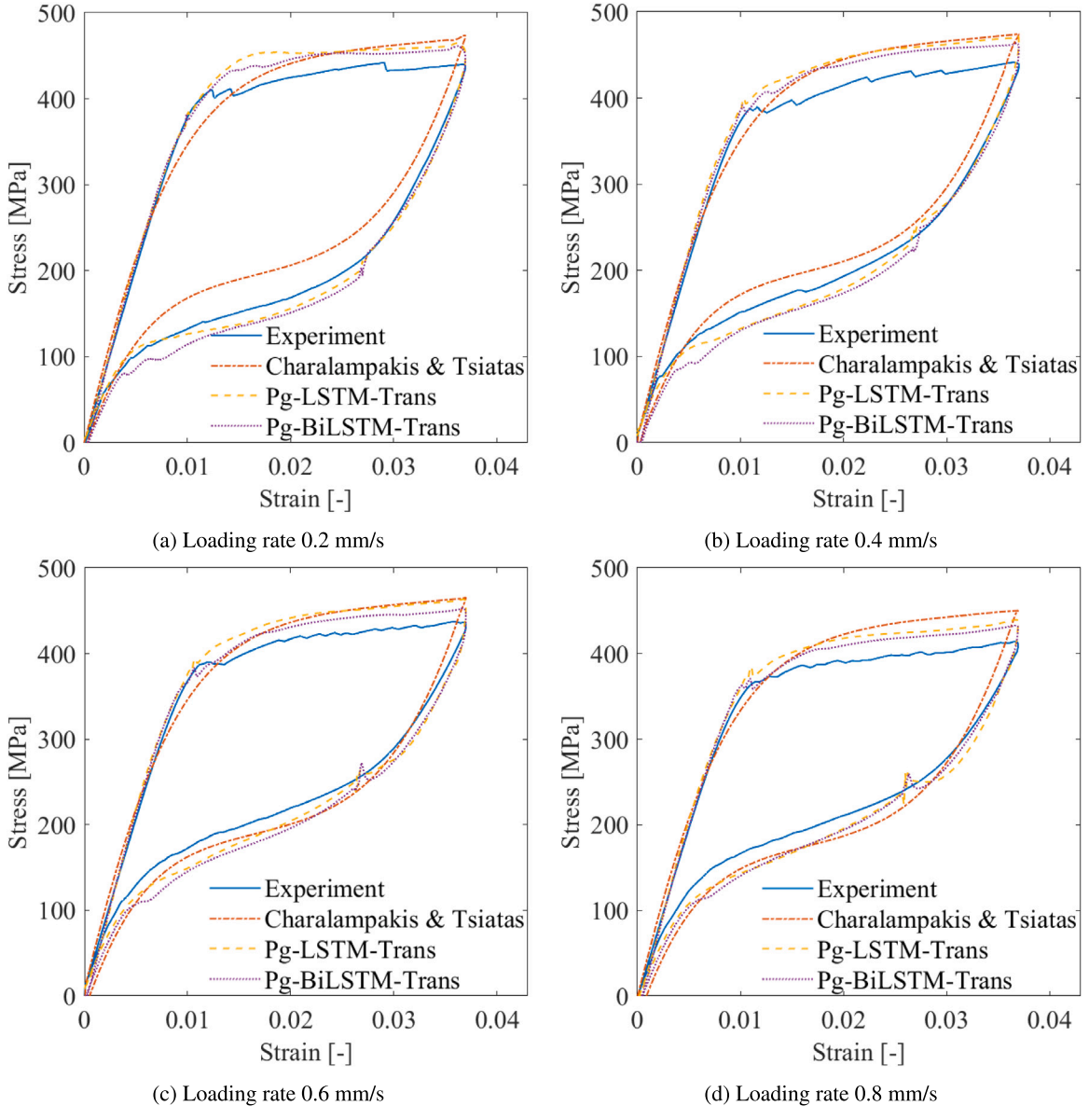
Although the proposed Pg-LSTM-Trans and Pg-BiLSTM-Trans models demonstrate superior performance upon combining a priori physics with an advanced NN architecture, Figs. 5 and 6 also show nonphysical local jitters for all loading rates in the loading branch (at about 1% strain) and, more evident, in the unloading branch (at about 2.5% strain). Such jitters have been observed in all NN architectures, and have been found much larger when a fully data-driven setting is adopted instead of the considered physics-guided approach. In particular, the results in Figs. 5 and 6 show that both the amplitude and spatial extent of these jitters are significantly reduced when using the Pg-BiLSTM–Trans model compared with the Pg-LSTM–Trans model, indicating enhanced stability due to the bi-directional processing of sequential data.

As a final note, it is noted that the proposed dual-branch architecture can be possibly implemented with either a uni-directional Gated Recurrent Unit (GRU) or a bi-directional GRU (BiGRU) in place of LSTM or BiLSTM, respectively. Their reduced number of gates (where the forget and input gates are merged into a single update gate operating alongside a reset gate) results in a simpler computational cell [54]. However, the performance slightly deteriorates. In its physics-guided setting, replacing BiLSTM with BiGRU in the proposed dual-branch architecture including Trans (Pg-BiGRU-Trans) leads to a higher number of units (n_{unit} = 262) after hyperparameter optimization, consistent with previous empirical studies [54]. Moreover, Pg-BiGRU-Trans results in a 10.76% and 12.17% increase in both RMSE and MAE, respectively, compared to Pg-BiLSTM-Trans (with a slight decline in both MAPE and R^2).

3.5. Benchmarking

The performance of the Pg-BiLSTM-Trans model is finally compared with that of a rate-dependent SMA constitutive model. For this purpose, the improved Graesser–Cozzarelli model recently proposed by Liu et al. [55] is adopted. This phenomenological model separates the hysteretic response into loading and unloading stages. During loading, the constitutive law introduces a strain-rate-dependent forward transformation behavior, which is defined as an explicit function of the strain rate as follows:

$$\dot{\sigma}_w = (1 - s(\varepsilon_w))E_a \left[\dot{\varepsilon}_w - |\dot{\varepsilon}_w| \left(\frac{\sigma_w - \beta}{\sigma_{m-s}}\right)^n\right] + s(\varepsilon_w)E_s, \tag{13a}$$



(a) Loading rate 0.2 mm/s

(b) Loading rate 0.4 mm/s

(c) Loading rate 0.6 mm/s

(d) Loading rate 0.8 mm/s

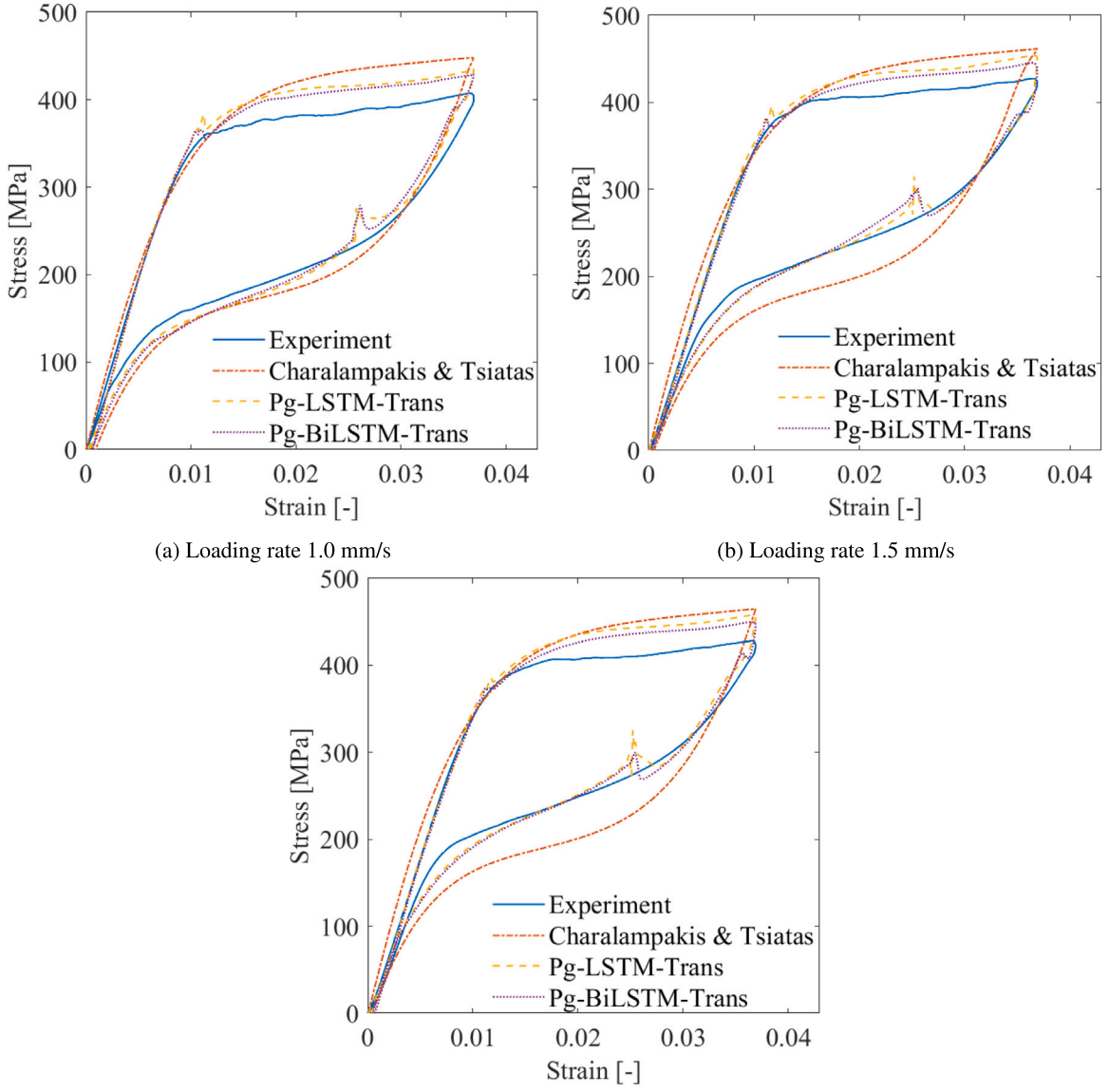
Fig. 5. Comparative analysis of hysteresis cycles in the test set for mid-low loading rate values: loading rate of (a) 0.2 mm/s, (b) 0.4 mm/s, (c) 0.6 mm/s, (d) 0.8 mm/s.

$$\beta = E_a h_a \left(\varepsilon_w - \frac{\sigma_w}{E_a}\right), \quad h_a = \frac{E_{am}(\dot{\varepsilon}_w)}{E_a - E_{am}(\dot{\varepsilon}_w)}, \tag{13b}$$

$$s(\varepsilon_w) = \frac{\tanh(g(|\varepsilon_w| - \varepsilon_{m-f})) + 1}{2}, \tag{13c}$$

where E_a is the deformation modulus of the austenite, β is the back stress, σ_{m-s} is the initiation stress for the forward transformation, n is a material coefficient controlling the corner sharpness of the hysteretic curve, E_s is the deformation modulus of detwinned martensite, h_a is the coefficient adjusting the stiffness of the stress plateau during the loading stage, g controls the sharpness of the martensitic hardening corner, and ε_{m-f} is the completion strain of the forward transformation. $E_{am}(\dot{\varepsilon}_w)$ is loading modulus of the mixed phase, which is the main reason for the differences in material curves under various strain rates. During unloading, both the elastic unloading modulus and the reverse transformation initiation stress are defined as explicit functions of the peak strain as follows:

$$\dot{\sigma}_w = E_m(\varepsilon_{ps}) \left[\dot{\varepsilon}_w - |\dot{\varepsilon}_w| \left(\frac{\sigma_w - \beta}{\sigma_{a-s}(\varepsilon_{ps})}\right)^{n-1} \left(\frac{\sigma_w - \beta}{\sigma_{a-s}(\varepsilon_{ps})}\right)\right], \tag{14a}$$



(a) Loading rate 1.0 mm/s

(b) Loading rate 1.5 mm/s

(c) Loading rate 2.0 mm/s

Fig. 6. Comparative analysis of hysteresis cycles in the test set for high loading rate values: loading rate of (a) 1.0 mm/s, (b) 1.5 mm/s, and (c) 2.0 mm/s.

$$\beta = E_m(\varepsilon_{ps}) h_m \left[\varepsilon_w - \frac{\sigma_w}{E_m(\varepsilon_{ps})} + f_T \tanh(\bar{a}\varepsilon_w) \right], \tag{14b}$$

$$h_m = \frac{E_{ma}}{E_m(\varepsilon_{ps}) - E_{ma}}, \tag{14c}$$

where h_m is the coefficient adjusting the stiffness of the stress plateau during the unloading stage, f_T is the coefficient describing the phase state of SMA at different temperatures, $\bar{a}$ is the adjustment coefficient, and E_{ma} is the unloading modulus of the mixed phase. $E_m(\varepsilon_{ps})$ is elastic unloading modulus and $\sigma_{a-s}(\varepsilon_{ps})$ is initiation stress for the reverse transformation, which depend on peak strain ε_{ps}.

The improved Graesser-Cozzarelli model proposed by Liu et al. [55] exhibits a higher level of complexity, requiring a total of 19 parameters to be determined, i.e., more than twice the number used in the model proposed by Charalampakis and Tsiatas [50]. All parameters are identified via DE, with their specific values listed in Table 3. It is noted that such a large number of parameters increases the risk of ill-posedness in the resulting identification problem, particularly when only limited information is available.

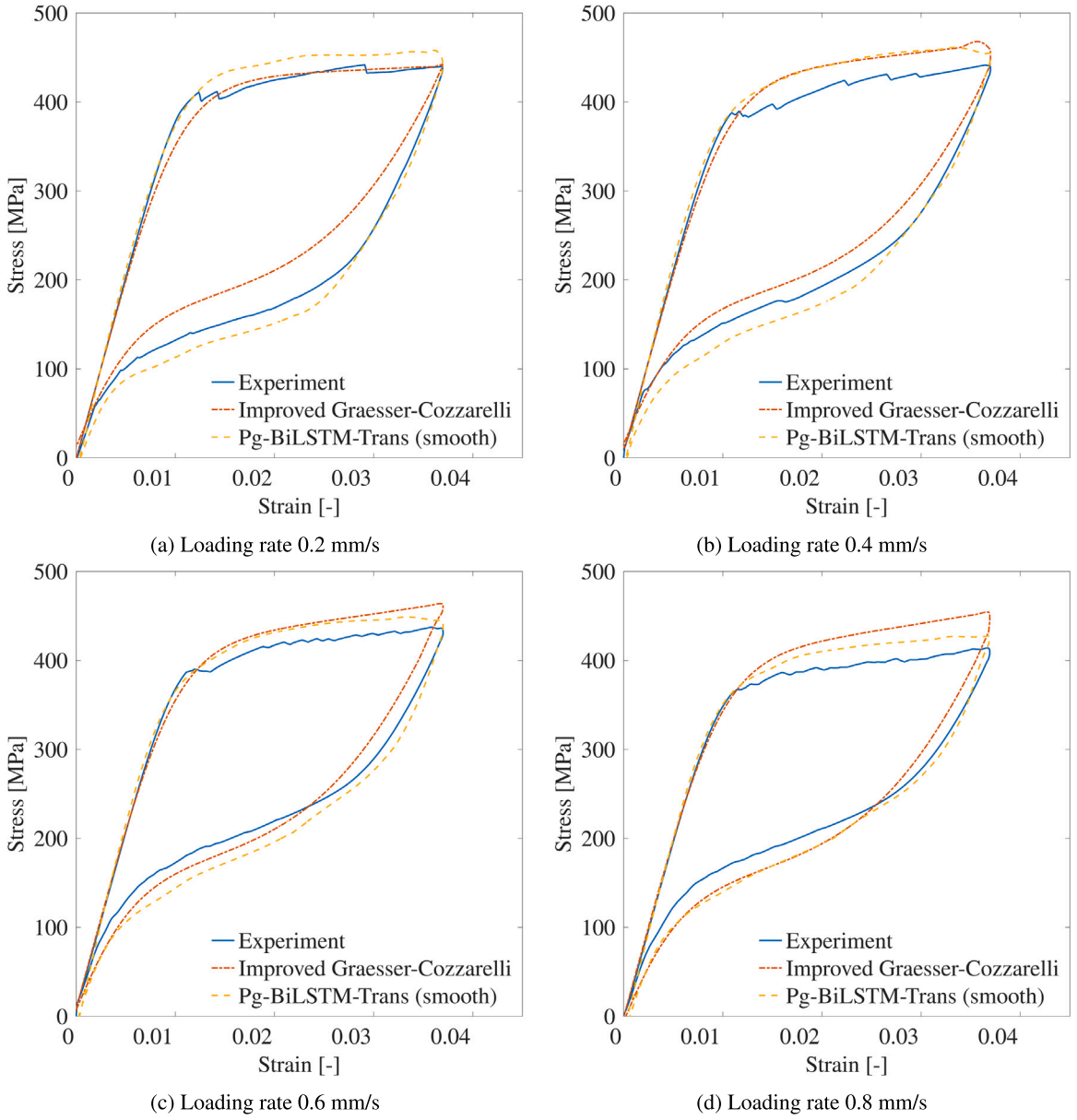


Fig. 7. Comparative analysis between the smoothed response of the proposed Pg-BiLSTM-Trans model and the improved Graesser-Cozzarelli model of hysteresis cycles in the test set for mid-low loading rate values: loading rate of (a) 0.2 mm/s, (b) 0.4 mm/s, (c) 0.6 mm/s, (d) 0.8 mm/s.

More importantly, whereas the physics-guided framework considered here, based on the model by Charalampakis and Tsiatas [50], does not require any prior information on the loading history, the approach proposed by Liu et al. [55] relies on prior knowledge of the peak strain, which may be unavailable in practical applications. The comparison between the hysteresis cycles predicted by the improved Graesser–Cozzarelli model proposed by Liu et al. [55] and those obtained from the Pg-BiLSTM-Trans model proposed in this work is presented in Figs. 7 and 8. In the present comparison, the hysteretic cycles predicted by the Pg-BiLSTM-Trans model were obtained after the application of a second-order Savitzky–Golay filter [56]. This is because, although relatively small and localized, the nonphysical jitters visible in Figs. 5 and 6 in the predictions of the Pg-BiLSTM-Trans model may lead to numerical instabilities when the model is used in structural analyzes. It is therefore useful to mitigate these oscillations either during training through suitable regularization strategies (e.g., by introducing a small smoothness penalty term into the loss function), or in post-processing by smoothing the predicted response.

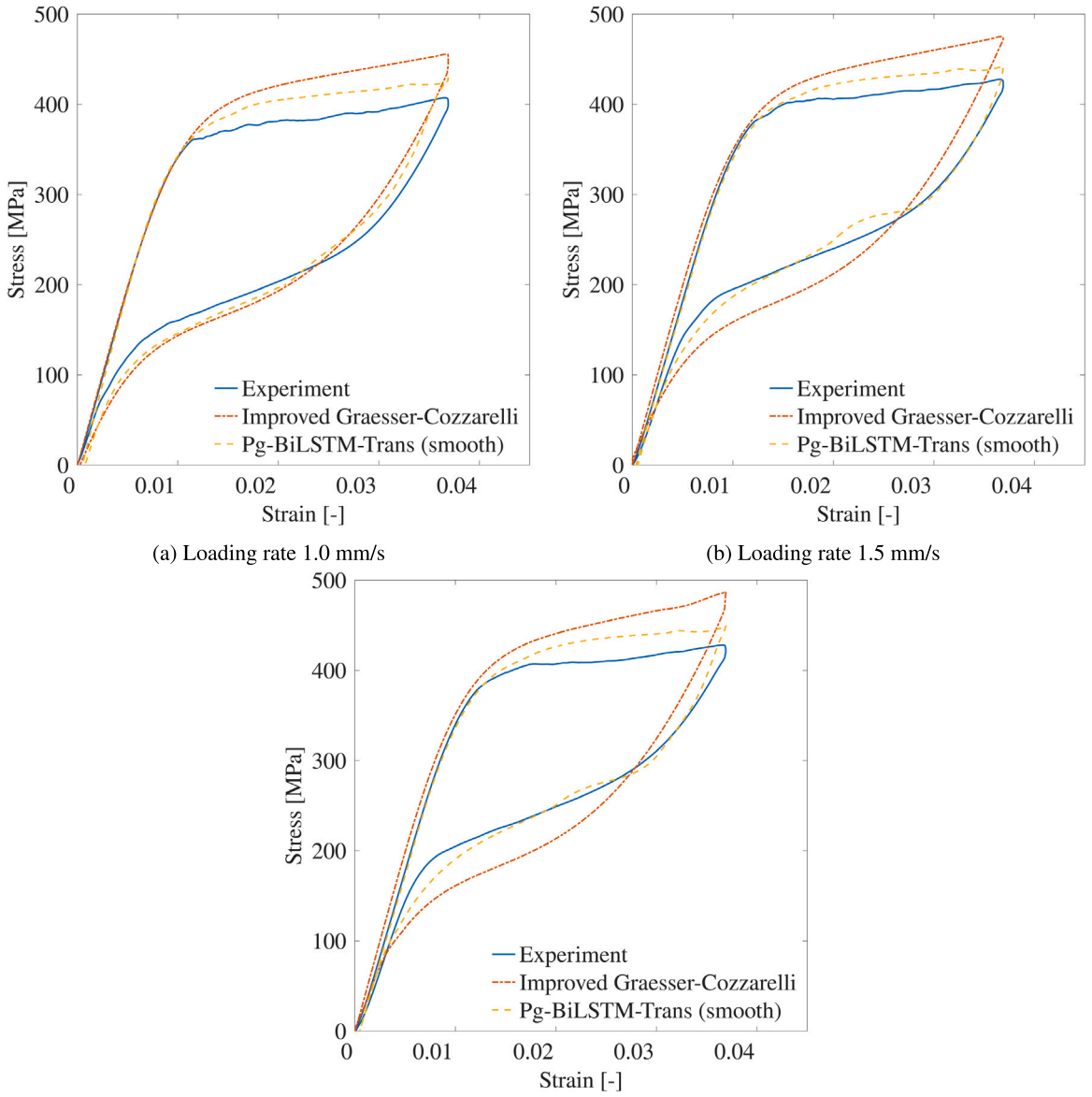
The analysis of Figs. 7 and 8 demonstrates that the Pg-BiLSTM-Trans model not only effectively compensates for the inherent deficiencies of the underlying rate-independent physics-based model but also outperforms a rate-dependent model under limited

Table 3
Parametric identification of the improved Graesser-Cozzarelli model [55].

E_a [MPa]	σ_{m-s} [MPa]	E_s [MPa]	ε_{m-f}	g	E_{ma} [MPa]	f_T	$\bar{a}$	n	MSE
40 668.012	442.645	23 117.431	0.060	85.921	836.752	1.055	142.216	3.670	2.747%

$E_{am} = 4257.614 + 505.043\ln(\dot{\varepsilon}_w - 0.0015)$, $\sigma_{a-s} = 737.537 - 1.0\ln(\varepsilon_{ps} - 1.0)$

$$E_m = 28\,424.802 + \frac{15\,215.286}{1 + \exp\left((\varepsilon_{ps} - 2.315)/0.460\right)}$$



(a) Loading rate 1.0 mm/s

(b) Loading rate 1.5 mm/s

(c) Loading rate 2.0 mm/s

Fig. 8. Comparative analysis between the smoothed response of the proposed Pg-BiLSTM-Trans model and the improved Graesser-Cozzarelli model of hysteresis cycles in the test set for high loading rate values: loading rate of (a) 1.0 mm/s, (b) 1.5 mm/s, and (c) 2.0 mm/s.

information. After smoothing, the minimum R^2 across all strain rates obtained with the Pg-BiLSTM-Trans model is 97.6% at a strain rate of 1.0 mm/s, whereas the corresponding lowest R^2 value for the improved Graesser–Cozzarelli model is 91.7% at a strain rate of 2.0 mm/s. On average, the MAPE associated with the Pg-BiLSTM-Trans model is reduced by 21% compared with that of the improved Graesser–Cozzarelli model. Furthermore, the Pg-BiLSTM-Trans model attains its highest RMSE and MAE values at a

strain rate of 0.4 mm/s, namely 19 N and 17 N, respectively. These values correspond to reductions of 44% and 43%, respectively, compared to the improved Graesser–Cozzarelli model proposed by Liu et al. [55], for which the maximum RMSE and MAE values occur at a strain rate of 2.0 mm/s and are equal to 34 N and 30 N, respectively.

4. Conclusions

This paper has presented a physics-guided approach for the implementation of NN architectures in the numerical simulation of nonlinear, memory-dependent behaviors, with focus on SMA wires. Several NN architectures have been take into account, including the uni- and bi-directional LSTM models, a transformer model, as well as a new architecture merging a transformer component with either uni- or bi-directional LSTM model.

Ablation experiments and comparative evaluations on NiTiNOL wires under tension at varying loading rates indicate that, in a fully data-driven setting, the proposed NN architecture performs very well on the test set. When implemented in a physics-guided manner, the proposed transformer-based NN architecture achieves an even more accurate approximation of the stress–strain hysteresis curves across varying loading rates. The main improvements of the proposed physics-guided NN architectures over the baseline physical model can be attributed to the restored sensitivity to the tensile loading rate and the improved simulation of the unloading branch. Ultimately, the proposed NN architecture combining a transformer component with either an uni- or bi-directional LSTM, also enhanced by physics, shows strong potential for modeling nonlinear, rate-dependent behaviors, particularly in applications involving large-scale data or highly parallelized online simulation and control tasks.

For the sake of completeness, it is pointed out that the present models (employed either in a fully data-driven setting or within the illustrated physics-guided framework) are deterministic and predict a perfectly repeatable transformation course under the same loading conditions. As such, the models do not account for uncertainties associated with the intrinsic microstructural variability of the material.

CRediT authorship contribution statement

Wei Dai: Writing – original draft, Visualization, Software, Formal analysis, Data curation, Methodology. **Biagio Carboni:** Conceptualization, Writing – review & editing, Investigation, Funding acquisition, Supervision. **Giuseppe Quaranta:** Conceptualization, Writing – review & editing, Methodology, Investigation, Funding acquisition, Supervision. **Yongjun Pan:** Conceptualization, Writing – review & editing, Investigation, Funding acquisition, Supervision. **Walter Lacarbonara:** Conceptualization, Writing – review & editing, Investigation, Funding acquisition, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Giuseppe Quaranta and Biagio Carboni report that financial support was provided by the Italian Ministry of University, while Walter Lacarbonara reports financial support from Sapienza University of Rome.

Acknowledgments

Biagio Carboni acknowledges the support received for his work through the project “Engineered basements for vibrations protection of artworks and strategic sensitive equipment” (Grant No. 2022TH5HC2) funded by European Union – Next Generation EU within the PRIN 2022 program, D.D. 104 of the Italian Ministry of University and Research (MUR). Giuseppe Quaranta acknowledges the support received for his work through the project “Artificial Intelligence for ENVIronmental impact minimization of SEismic Retrofitting of Structures (AI-ENVISERS)” (Grant No. P20227KKF5) funded by European Union – Next Generation EU within the PRIN 2022 PNRR program, D.D. 1409 of the MUR. This research reflects only their views and opinions and the MUR cannot be considered responsible for them. Walter Lacarbonara acknowledges partial support through the project “Innovative techniques for vibration control in civil engineering: nonlinear modeling, digital twins, and experimentation” (Grant No. RD12318AAD49FEC0) funded by Sapienza University of Rome and the project Rome Technopole (Grant No. ECS 0000024, CUP B83C22002820006, National Recovery and Resilience Plan Mission 4 Component 2 Investment 1.5) funded by the European Union - NextGenerationEU.

Data availability

The authors do not have permission to share data.

References

[1] G. Quaranta, W. Lacarbonara, S.F. Masri, A review on computational intelligence for identification of nonlinear dynamical systems, Nonlinear Dynam. 99 (2) (2020) 1709–1761.
[2] J.A. Suykens, J. Vandewalle, Nonlinear modeling: advanced black-box techniques, Springer Science & Business Media, 1998.
[3] J.R. Mianroodi, N. H. Siboni, D. Raabe, Teaching solid mechanics to artificial intelligence—a fast solver for heterogeneous materials, Npj Comput. Mater. 7 (1) (2021) 99.
[4] R. Vinuesa, S.L. Brunton, Enhancing computational fluid dynamics with machine learning, Nat. Comput. Sci. 2 (6) (2022) 358–366.
[5] H. Jin, E. Zhang, H.D. Espinosa, Recent advances and applications of machine learning in experimental solid mechanics: A review, Appl. Mech. Rev. 75 (6) (2023) 061001.
[6] Y. Kim, Y. Kim, C. Yang, K. Park, G.X. Gu, S. Ryu, Deep learning framework for material design space exploration using active transfer learning and data augmentation, Npj Comput. Mater. 7 (1) (2021) 140.
[7] A. Farrag, Y. Yang, N. Cao, D. Won, Y. Jin, Physics-informed machine learning for metal additive manufacturing, Prog. Addit. Manuf. 10 (1) (2025) 171–185.
[8] Y. Xie, M. Ebad Sichani, J.E. Padgett, R. DesRoches, The promise of implementing machine learning in earthquake engineering: A state-of-the-art review, Earthq. Spectra 36 (4) (2020) 1769–1801.
[9] B.Z. Cunha, C. Droz, A.-M. Zine, S. Foulard, M. Ichchou, A review of machine learning methods applied to structural dynamics and vibroacoustic, Mech. Syst. Signal Process. 200 (2023) 110535.
[10] A.T.G. Tapeh, M. Naser, Artificial intelligence, machine learning, and deep learning in structural engineering: a scientometrics review of trends and best practices, Arch. Comput. Methods Eng. 30 (1) (2023) 115–159.
[11] M. Haywood-Alexander, W. Liu, K. Bacsa, Z. Lai, E. Chatzi, Discussing the spectrum of physics-enhanced machine learning: a survey on structural mechanics applications, Data-Centric Eng. 5 (2024) e30.
[12] J.-S. Pei, J.P. Wright, A.W. Smyth, Mapping polynomial fitting into feedforward neural networks for modeling nonlinear dynamic systems and beyond, Comput. Methods Appl. Mech. Engrg. 194 (42–44) (2005) 4481–4505.
[13] J.-S. Pei, A.W. Smyth, New approach to designing multilayer feedforward neural network architecture for modeling nonlinear restoring forces. I: Formulation, J. Eng. Mech. 132 (12) (2006) 1290–1300.
[14] C. Molnar, G. Casalicchio, B. Bischl, Interpretable machine learning–a brief history, state-of-the-art and challenges, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2020, pp. 417–431.
[15] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (5) (2018) 1–42.
[16] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, A. Hussain, Interpreting black-box models: a review on explainable artificial intelligence, Cogn. Comput. 16 (1) (2024) 45–74.
[17] E. ŞAHiN, N.N. Arslan, D. Özdemir, Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning, Neural Comput. Appl. 37 (2) (2025) 859–965.
[18] S.A. Faroughi, N. Pawar, C. Fernandes, M. Raissi, S. Das, N.K. Kalantari, S.K. Mahjour, Physics-guided, physics-informed, and physics-encoded neural networks in scientific computing, 2022, arXiv preprint arXiv:2211.07377.
[19] A. Cicirello, Physics-enhanced machine learning: a position paper for dynamical systems investigations, in: Journal of Physics: Conference Series, Vol. 2909, IOP Publishing, 2024, 012034.
[20] G. Olson, M. Cohen, Stress-assisted isothermal martensitic transformation: application to TRIP steels, Metall. Trans. A 13 (11) (1982) 1907–1914.
[21] J.M. Jani, M. Leary, A. Subic, M.A. Gibson, A review of shape memory alloy research, applications and opportunities, Mater. Des. (1980-2015) 56 (2014) 1078–1113.
[22] M.-S. Kim, J.-K. Heo, H. Rodrigue, H.-T. Lee, S. Pané, M.-W. Han, S.-H. Ahn, Shape memory alloy (SMA) actuators: The role of material, form, and scaling effects, Adv. Mater. 35 (33) (2023) 2208517.
[23] N. Eliaz, Corrosion of metallic biomaterials: A review, Materials 12 (3) (2019).
[24] H. Lu, L. Liu, C. Yang, X. Luo, C. Song, Z. Wang, J. Wang, Y. Su, Y. Ding, L. Zhang, et al., Simultaneous enhancement of mechanical and shape memory properties by heat-treatment homogenization of Ti_2Ni precipitates in TiNi shape memory alloy fabricated by selective laser melting, J. Mater. Sci. Technol. 101 (2022) 205–216.
[25] Y.-W. Cui, L. Wang, L.-C. Zhang, Towards load-bearing biomedical titanium-based alloys: From essential requirements to future developments, Prog. Mater. Sci. 144 (2024).
[26] K. Wilde, P. Gardoni, Y. Fujino, Base isolation system with shape memory alloy device for elevated highway bridges, Eng. Struct. 22 (3) (2000) 222–229.
[27] Y. Zhang, S. Zhu, A shape memory alloy-based reusable hysteretic damper for seismic hazard mitigation, Smart Mater. Struct. 16 (5) (2007) 1603–1613.
[28] W. Dai, B. Carboni, G. Quaranta, Y. Pan, W. Lacarbonara, Nonlinear response of a multidirectional negative-stiffness isolation system via semirecursive multibody dynamic approach, Int. J. Mech. Syst. Dyn. 4 (3) (2024) 258–277.
[29] R. DesRoches, B. Smith, Shape memory alloys in seismic resistant design and retrofit: a critical review of their potential and limitations, J. Earthq. Eng. 8 (3) (2004) 415–429.
[30] B. Carboni, W. Lacarbonara, F. Auricchio, Hysteresis of multiconfiguration assemblies of NiTinol and steel strands: Experiments and phenomenological identification, J. Eng. Mech. 141 (3) (2015).
[31] Y.C. Pei, X.Y. Wang, Z.-Y. Yao, B.-H. Wang, Z.-h. Liao, H. Lu, The driving characteristics of bidirectional SMA wire actuators - theoretical modeling and experimental testing, Sensors Actuators A 372 (2024).
[32] G.L.W. Hart, T. Mueller, C. Toher, S. Curtarolo, Machine learning for alloys, Nat. Rev. Mater. 6 (8) (2021) 730–755.
[33] R. Hmede, F. Chapelle, Y. Lapusta, Review of neural network modeling of shape memory alloys, Sensors 22 (15) (2022).
[34] S.J. Honrao, O. Benafan, J.W. Lawson, Data-driven study of shape memory behavior of multi-component Ni-Ti alloys in large compositional and processing space, Shape Mem. Superelastiity 9 (1) (2023) 144–155.
[35] S.H. Zadeh, T.D. Brown, X. Qian, I. Karaman, R. Arroyave, A composition-based predictive model for the transformation strain of NiTi shape memory alloys, Acta Mater. 289 (2025).
[36] F. Wang, L.-X. Wang, A phenomenological constitutive model for one-dimensional shape memory alloys based on artificial neural network, J. Intell. Mater. Syst. Struct. 32 (18–19) (2021) 2338–2348.
[37] N. Lenzen, O. Altay, Machine learning enhanced dynamic response modelling of superelastic shape memory alloy wires, Materials 15 (1) (2022).
[38] N. Lenzen, O. Altay, Feedforward neural network-assisted parameter identification and tuning for uniaxial superelastic shape memory alloy models under dynamic loads, J. Vib. Control 31 (1-2, SI) (2025) 51–69.
[39] N. Lenzen, O. Altay, Physics-informed deep operator network for predicting martensite evolution in superelastic shape memory alloys through cyclic tensile tests, Smart Mater. Struct. 33 (6) (2024) 065039.
[40] S. Zhu, Y. Zhang, A thermomechanical constitutive model for superelastic SMA wire with strain-ratedependence, Smart Mater. Struct. 16 (5) (2007) 1696.

[41] J.K. Joy, B. Haghgouyan, A.A. Karakalas, M. Vasoya, D.C. Lagoudas, A neural network model for shape memory alloy actuation response with physical constraints for partial phase transformation, J. Intell. Mater. Syst. Struct. 1045389X251348269.
[42] D. Khatamsaz, R. Neuberger, A.M. Roy, S.H. Zadeh, R. Otis, R. Arróyave, A physics informed bayesian optimization approach for material design: application to NiTi shape memory alloys, Npj Comput. Mater. 9 (1) (2023) 221.
[43] R. Storn, K. Price, Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces, J. Global Optim. 11 (4) (1997) 341–359.
[44] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[45] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: LSTM cells and network architectures, Neural Comput. 31 (7) (2019) 1235–1270.
[46] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (11) (1997) 2673–2681.
[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[48] S. Islam, H. Elmekki, A. Elsebai, J. Bentahar, N. Drawel, G. Rjoub, W. Pedrycz, A comprehensive survey on applications of transformers for deep learning tasks, Expert Syst. Appl. 241 (2024).
[49] C. Gambella, B. Ghaddar, J. Naoum-Sawaya, Optimization problems for machine learning: A survey, European J. Oper. Res. 290 (3) (2021) 807–828.
[50] A.E. Charalampakis, G.C. Tsiatas, A simple rate-independent uniaxial Shape Memory Alloy (SMA) model, Front. Built Environ. 4 (2018).
[51] A. Salvatore, B. Carboni, W. Lacarbonara, Nonlinear dynamic response of an isolation system with superelastic hysteresis and negative stiffness, Nonlinear Dynam. 107 (2) (2022) 1765–1790.
[52] S. Karmakar, A. Kumar, C. Kolay, Development, characterization, and seismic application of a shape memory alloy-based self-centering damper, Soil Dyn. Earthq. Eng. 183 (2024) 108797.
[53] A. Salvatore, On the shock performance of a hysteretic tri-stable vibration isolation system: Nonlinear phenomena and optimization, Int. J. Non-Linear Mech. 157 (2023) 104557.
[54] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014, arXiv preprint arXiv:1412.3555.
[55] B. Liu, B. Li, S. Wang, Y. Liu, T. Yang, Y. Zhou, A 1D strain-amplitude- and strain-rate-dependent model of super-elastic shape memory alloys for structural vibration control, Structures 25 (2020) 426–435, http://dx.doi.org/10.1016/j.istruc.2020.03.028, URL https://www.sciencedirect.com/science/article/pii/S2352012420301119.
[56] S.J. Orfanidis, Introduction to Signal Processing, Prentice-Hall, Inc, 1995.