Explainability analysis of a machine learning-based constitutive model for concrete

Saulo S. de Castro[1], Álefe F. Figueiredo[1], Roque L. S. Pitangueira[1]

[1]*Dept. of Structures Engineering, Federal University of Minas Gerais*
Av. Antônio Carlos, 6627, Pampulha, Belo Horizonte, 31270-901, Minas Gerais, Brazil
saullo9@yahoo.com.br, alefefreitas_11@hotmail.com, roque@dees.ufmg.br

Abstract.

Concrete, like other quasi-brittle materials, exhibits intricate mechanical behavior that defies simple mathematical description and remains partially elusive. Limited comprehension of the governing mechanisms hampers the development of a comprehensive theory and broader constitutive models. This limitation suggests that more encompassing models could emerge through advanced techniques, like Machine Learning (ML) algorithms, capable of capturing material behaviors effectively. While various studies endorse ML-based constitutive models and validate the proposed hypothesis, skepticism persists within academia due to concerns about ML models being "black boxes" devoid of interpretable physical consistency. To challenge this perception, this paper introduces the SHAP tool, employed to validate the physical coherence of an ML-based constitutive model focused on concrete representation. The SHAP tool's methodology is outlined, accompanied by illustrative applications showcasing the correlation between input variables and model outputs. Clear demonstrations of physical consistency debunk the notion of ML models as opaque "black boxes." Ultimately, this study debunks skepticism, offering new perspectives on the utilization of ML-based constitutive models, thus fostering broader acceptance and integration in the structural engineering community.

Keywords: Constitutive models, Concrete, Machine learning, Artificial neural network, Explicability.

1 Introduction

Concrete, as well as other quasi-brittle media, presents extremely complex mechanical behavior, which is difficult to mathematically equate and whose complete understanding has not been achieved yet. The development of numerical methods, especially the finite element method (FEM), has enabled a great advance in the way of representing the mechanical behavior of quasi-brittle media.

Many of these constitutive models have been developed from experimental observations, which in turn have underpinned various theories about the mechanical behavior of the observed materials. This dynamic has made the constitutive models conceived by this methodology widely known as phenomenological models. In the specific case of concrete, we highlight the models based on fracture mechanics ([1–8]), those rooted in damage mechanics ([9]) and more recently, the models formulated according to the phase field theory.

Phenomenological models have a good capacity to represent the mechanics of concrete, provided that the limits imposed by the theoretical bases are respected and the model parameters are truly representative of the material. Thus, no model is completely general and the parameters that feed them are mostly difficult to obtain.

Although each phenomenological model has its own limitations, in the final analysis, they all stem from both human limitations and those of classical statistics, in recognizing behavioral patterns of the material and formulating, from these patterns, more representative and broader theoretical bases. This perspective motivated research that sought in Machine Learning (ML) algorithms a way to propose more robust constitutive models. It is pointed out that ML algorithms are currently the gold standard in the art of pattern recognition and function approximation problems.

In the work of [10], the first proposal to use an ML method for concrete constitutive representation was presented. The authors developed a model known as Neural Network-based Constitutive Model (NNCM), by

using a Multilayer Perceptron (MLP), a type of Artificial Neural Network (ANN), to represent the relationships between stresses and strains in concrete plates subjected to biaxial stress states. The model was trained using a set of stress-strain curves obtained from several tests performed on the mentioned plates. In this way, the generated model had the ability to predict the stress increments resulting from strain increments, based on the actual stress and strain states. However, the parameters that characterize the material and the structural system were not included as attributes of the model. After the work of [10], many other researches followed this path, among which are [10–20].

Although the use of ML-based constitutive models has shown promising results, many researchers and engineers express a skeptical view of the practical viability of these models. This skepticism is motivated, in part, by the perception that ML models are "black boxes" whose physical consistency of responses can not be guaranteed in any situation [19]. In fact, some ML models are so robust that it would be impossible to analyze their decision-making mechanisms without the use of auxiliary tools. This occurs due to the large number of variables within the models and the possibility of creating highly complex relationships between input and output variables. Here, it is important to emphasize that this ability to create complex relationships between the variables that govern the problem is responsible for the robustness of ML models.

The need for an explainable ML model adhering to a physical reality is not restricted to materials engineering. The interpretability, explainability and adherence to physical principles are topics of several researches, such as [21–23].

The present work used the SHAP tool to evaluate the physical consistency of a constitutive model formulated based on an MLP and to demonstrate that with proper training, ML models can originate reliable constitutive models. From this methodology it is possible to evaluate how the variables correlates with each other, how each variable influences the responses and to observe whether or not there would be evidence of physical consistency of the predictions made by the model.

2 ML Models Explainability

Explainability is the ability of the model to provide clarity in its results in order to help the user make decisions based on data with confidence and in an auditable way. However, most ML algorithms, such as ANNs, are considered as "black boxes", and therefore, target for much criticism. In the practical word, unexplainability may compromise the reliability of ML models and make its use unfeasible. This occurs because a prediction performance metric, regardless of what it is, is a very shallow description for the vast majority of real-world problems/tasks.

Models of a simple nature, such as linear regressions or decision trees, have the advantage of being explainable from their own structures. In this sense, it is possible to use interpretation methods belonging to the group of Intrinsic methods to accomplish this task. On the other hand, in the case of more complex models, such as ANNs, their explanation requires the application of methods known as Post hoc.

Post hoc methods explore the explainability of a given model by analyzing its predictions. These techniques can be grouped into two categories: Specific and Agnostic. Specific methods are employed to understand the decisions and behavior of a particular learning model, while Agnostic methods are more generic and can be applied to any type of learning model. This differentiation between methods is essential to understand how different degrees of complexity influence the possibilities to explain the predictions of a given model. The proper choice of interpretation method can be crucial for understanding the results obtained and for gaining insights into practical real-world problems.

The central purpose of this work is to offer a contribution to the increase of reliability in the use of constitutive models based on ML techniques. In this sense, this study will present the use of a Post hoc method, whose objective is to understand a NNCM, as well as to demonstrate that it keeps physical consistencies necessary for its practical use. For this purpose, the agnostic method called SHAP (SHapley Additive exPlanations), which was proposed by [24], was adopted.

2.1 SHapley Additive exPlanations

SHAP (SHapley Additive exPlanations) is a local explanation method based on the concepts of Shapley values, which are known to be theoretically optimal. In order to gain a clear understanding of the SHAP method, it is essential to obtain prior knowledge about the workings of the Local Interpretable Model-Agnostic Explanations (LIME) method and to familiarize oneself with the concept of Shapley values, an ideia coming from coalition game theory.

LIME is a technique developed by [25] with the purpose of becoming a tool capable of explaining ML mod-

els. To this end, LIME analyzes the effect of variations in features on the outputs of a given model. To achieve this goal, an interpretable local model that seeks to approximate the predictions for each example individually is developed, thus providing strong explanations at the local level. When considering the perspective of scholars of constitutive modeling, it is found that LIME allows that, from a reliable ML model, an equilibrium path can be discretized and analyzed at "n" points, facilitating the evaluation of the behavior of the model point by point. This characteristic enables the analysis of the physical consistency of the model at each discretization point. Furthermore, the methodology allows a broad and detailed view of how the model recognizes the degradation processes of a given material.

To train the local model, the LIME algorithm introduces small perturbations in the features of the example of interest and observes how the model responds to these pertubations. In other words, the technique modifies some input variables randomly and observes how this impacts the output variables.

Shapley values fairly measure the individual contribution of each player to the overall outcome of a coalition game. In other words, they reflect the influence of each player on the outcome. For the present text, it is sufficient to understand that in a coalition game, two or more players are involved in a strategy that aims to achieve the maximum possible gain. In the context of explainability of ML models, the "game" represents the task of making a prediction for a specific example in the dataset. The "gain" is the difference between the prediction for that example and the average of the predictions for all examples. The "players" are interpreted as the values of the features of the example that contribute to making the prediction. From the associations posed, this concept can be illustrated with the following allegory: the game takes place in a room where each player (representing the value of a given feature) enters randomly and participates in the game. The Shapley value of a feature is the average change in the prediction made by the coalition of that feature with features that were already present in the room.

SHAP combines these two concepts to enable, in addition to individual explanations, the construction of global explanations by combining or averaging all local instances.

3 Methodology

First a MLP was trained with synthetic stress-strain data of the concrete material with the following physical characteristics: characteristic compressive strength of concrete $f_{ck} = 31.0$ MPa, characteristic tensile strength of concrete $f_t = 2.7$ MPa, Young's modulus $E = 25850.0$ MPa, Poission's ratio $\nu = 0.18$, strain corresponding to the concrete compression strength limit $e_c = 2.2 \times 10^{-3}$ and strain corresponding to the concrete tensile strength limit $e_t = 1.925 \times 10^{-4}$. The objective was to develop a NNCM for representing the mechanical behavior of concrete. The synthetic data were generated from numerical simulations of several structural systems under different types of loading and failure modes, and therefore, with different stress states. For this purpose, a conventional constitutive model known as smeared crack model, whose results are known to be consistent, was used. Details of the processes of selection of training data, definition of network architecture and its hyperparameters, training and model validation can be seen in [26].

This paper focuses on the presentation of the explainability analysis of the developed model. The evaluation here presented was performed on the model responses for one of the structural systems that composed the training database: the beam under 4-point bending, presented in Fig. 1. The concrete beam, with a size of $600 \times 150 \times 120$ mm³, was numerically simulated under plane stress state, with a quadratic finite element mesh. The analysis was performed using the cylindrical arc length control method, with an initial increment of the load factor of 0.05 and the convergence was verified in displacements with tolerance of 1×10^{-4}.

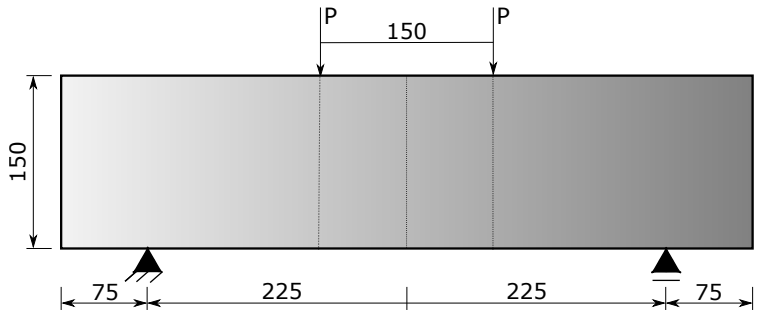

Figure 1. 4-point bending test - Problem setting (measures in mm)

This structure was chosen for the physical consistency assessment because it presents stress states referring to pure bending, which allow to understand more easily how stress increments relate to strain increments, starting

from a given known stress-strain state.

In order to visualize the results of the model predictions, an auxiliary method was developed which, for each iteration, from the input data strain increments, current stresses and strains and historical stresses and strains, calculates the respective stress increments as output data. Then, the method uses such predictions to calculate the current stresses predicted by the MLP, to be used in the next iteration, similarly to Newton-Raphson process. At the end, the method assembles the stress-strain curve with the stresses increments predicted by MLP.

Thus, the predictions made with the ML model were compared with the results obtained by the FEM model, and the consistency of these predictions was analyzed using some of the various tools of the SHAP library. The impact of each input variable on the model's response was assessed locally and globally. The local analysis allows to evaluate the behavior of a specific prediction of the model (a stress increment of a specific gauss point) and the global analysis, the general behavior of all predictions of the model (all stress increments of all gauss points along all equilibrium paths of the analyzed sample). The results of these analyses are presented in section 4.

4 Results

For the local analysis, a point in the lower central region of the beam was chosen, as illustrated by Fig. 2. In this example, only the output variable stress increment at x-direction ($\Delta\sigma_{xx}$) was evaluated.

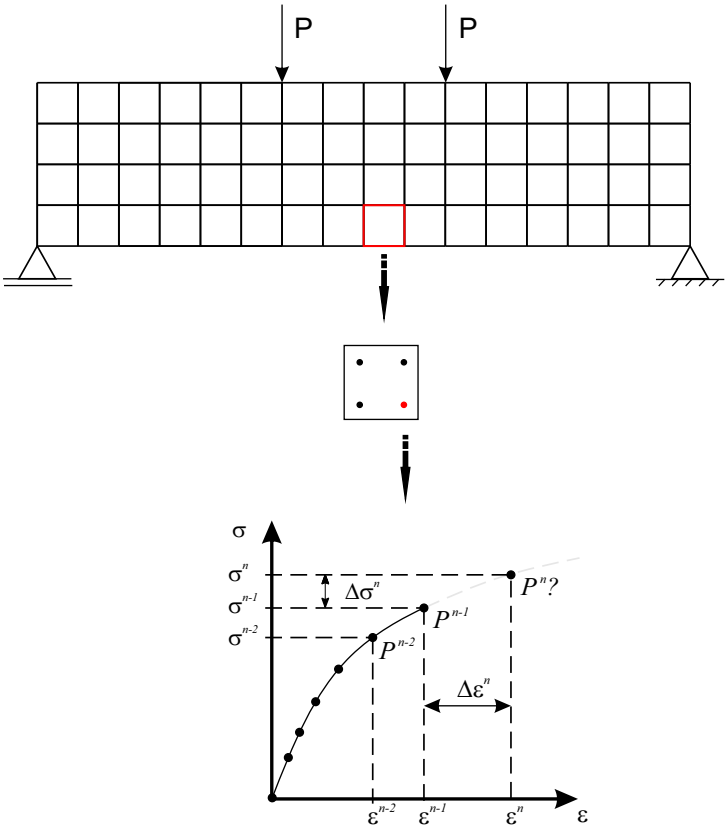

Figure 2. Numerical model used to generate data for the consistency study - Demarcation of the selected point for SHAP analysis.

Fig. 3 and Fig. 4 brings the MLP-predicted stress-strain curve and the waterfall plot for the selected point and component, respectively.

It can be seen that the input variables that have the greatest impact on the output variable analyzed are the stresses, strains and strain increment in the respective direction, for both points evaluated (A and B). As this point of analysis is mostly under normal stresses in the x-direction, while the normal stresses in y-direction and the shear stresses are very low (pure bending state), the variables related to the last two mentioned stress components should not and did not influence the prediction of stress increment in the x-direction, and therefore, the observed results presented themselves coherent and consistent with the reality of the problem.

For the global analysis, the influence of the features on the predictions of all curves present in the region

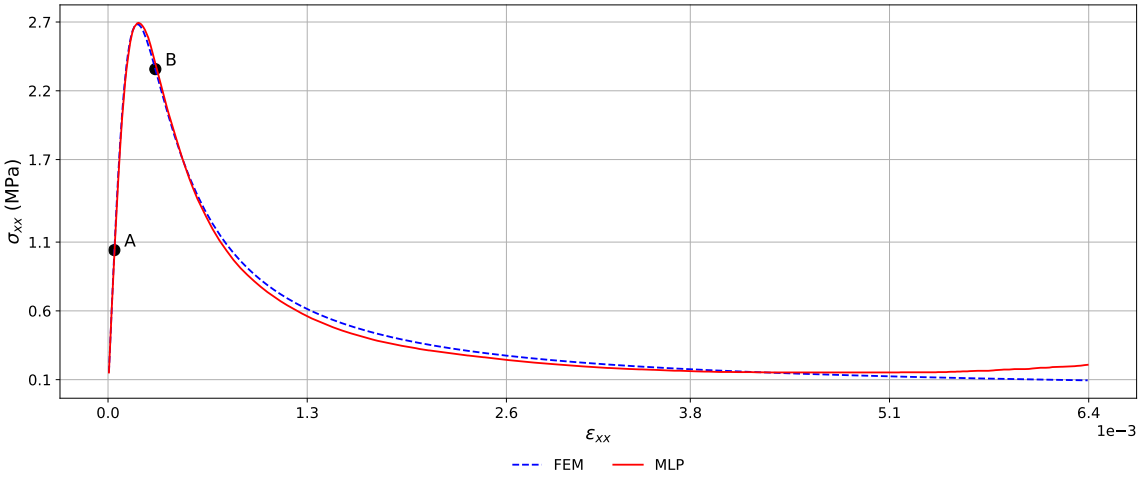

Figure 3. Comparison between FEM and MLP results - $\sigma_{xx} \times \varepsilon_{xx}$

.

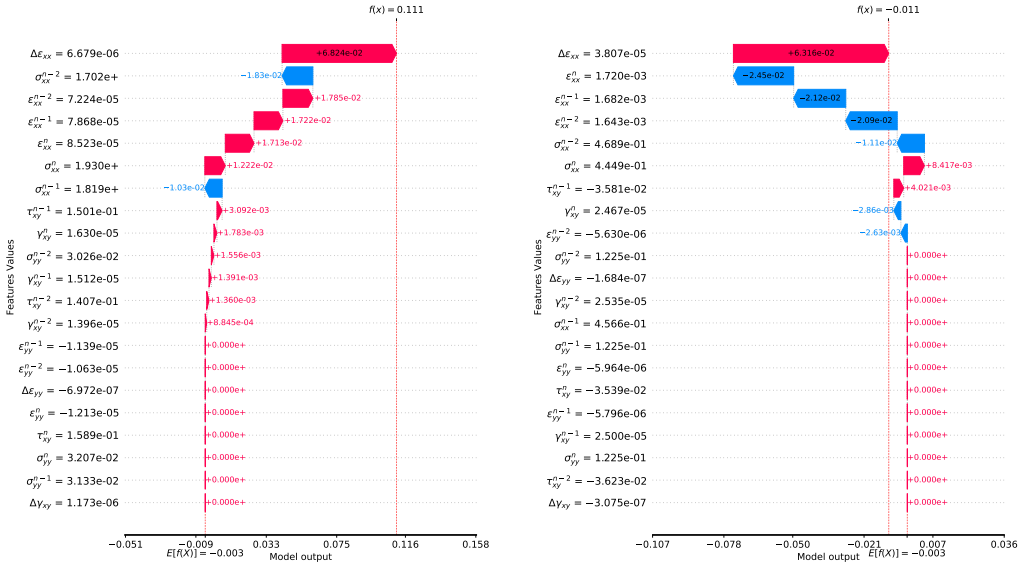

Figure 4. Local explanation for the 4-point bending test in the pure bending region - (a) Feature importance for output $\Delta\sigma_{xx}$ for the ascending branch, (b) Feature importance for output $\Delta\sigma_{xx}$ for the descending branch.

under pure bending was evaluated.

By analyzing Fig. 5, it can be seen that the general behavior of the entire dataset in this region is the same as observed in the individual analysis. For the main stress component of this structure, in the region of interest, the features that most influence the MLP prediction are those that have the same direction as the output variable (x-direction).

5 Conclusions

The analyses performed here attest that the predictions do not represent a function disconnected from physical reality, whose answer coincides with the expected one only as a matter of chance. For the cases analysed, the input variables that have the greatest impact on the output variables are those that the theoretical bases point to as being responsible for the observed response. This verification highlights that the criticism directed at the utilization of ML algorithms is influenced by bias and introduces fresh perspectives on the utility of ML-based constitutive models.

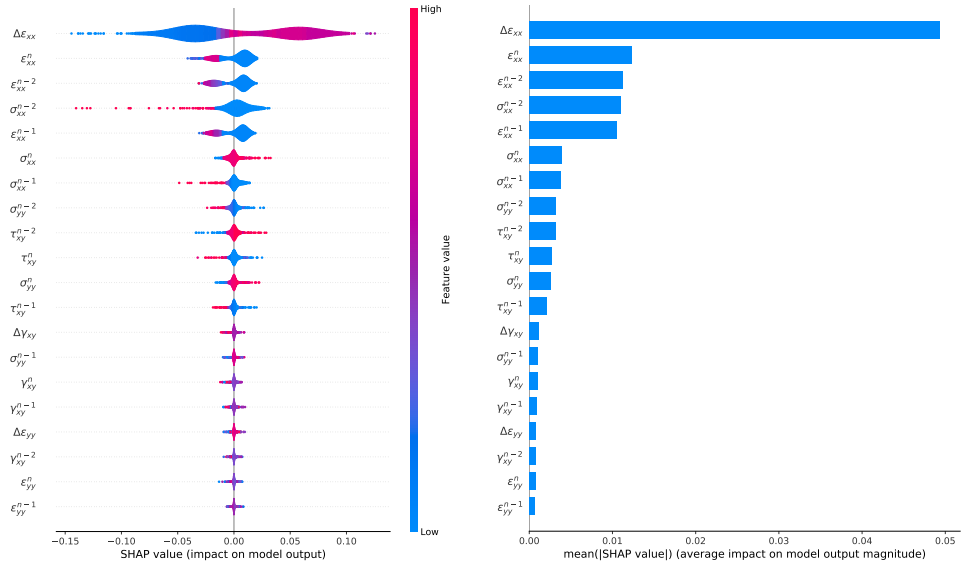

Figure 5. Global explanation for the 4-point bending test in the pure bending region - Feature importance for output $\Delta\sigma_{xx}$ (a) SHAP value, (b) Mean SHAP value.

Acknowledgements. The authors gratefully acknowledge the support from the Brazilian research agency CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), the PROPEEs (Programa de Pós-Graduação em Engenharia de Estruturas) and the UFMG (Universidade Federal de Minas Gerais).

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] D. Dugdale. Yielding of steel sheets containing slits. *Journal of the Mechanics and Physics of Solids*, vol. 8, n. 2, pp. 100 – 104, 1960.
[2] A. Hillerborg, M. Modéer, and P.-E. Petersson. Analysis of crack formation and crack growth in concrete by means of fracture mechanics and finite elements. *Cement and Concrete Research*, vol. 6, n. 6, pp. 773 – 781, 1976.
[3] Z. P. Bažant. Instability, ductility, and size effect in strain-softening concrete. *ASCE J Eng Mech Div*, vol. 102, n. 2, pp. 331–344, 1976.
[4] Z. P. Bažant and L. Cedolin. Blunt crack band propagation in finite element analysis. *ASCE J Eng Mech Div*, vol. 105, n. 2, pp. 297–315, 1979.
[5] Z. P. Bažant and B. H. Oh. Crack band theory for fracture of concrete. *Matériaux et construction*, vol. 16, n. 3, pp. 155–177, 1983.
[6] J. Cervenka, V. Cervenka, and S. Laserna. Modelling softening materials in engineering practice using fem and crack band method. *Mecánica Computacional*, vol. 36, n. 1, pp. 5–5, 2018.
[7] C. Carloni, G. Cusatis, M. Salviato, J.-L. Le, C. G. Hoover, and Z. P. Bažant. Critical comparison of the boundary effect model with cohesive crack model and size effect law. *Engineering Fracture Mechanics*, vol. 215, pp. 193 – 210, 2019.
[8] J. Planas, B. Sanz, and J. M. Sancho. Vectorial stress-separation laws for cohesive cracking: in concrete and other quasibrittle materials. *International Journal of Fracture*, pp. 1–16, 2020.
[9] D. Kochanov. Time of rupture process under creep conditions. *Izvestia Akademii Nauk, USSR*, vol. 8, pp. 26–31, 1958.
[10] J. Ghaboussi, J. Garrett Jr, and X. Wu. Knowledge-based modeling of material behavior with neural networks. *Journal of engineering mechanics*, vol. 117, n. 1, pp. 132–153, 1991.
[11] G. Ellis, C. Yao, R. Zhao, and D. Penumadu. Stress-strain modeling of sands using artificial neural networks. *Journal of geotechnical engineering*, vol. 121, n. 5, pp. 429–435, 1995.
[12] J. Ghaboussi, D. A. Pecknold, M. Zhang, and R. M. Haj-Ali. Autoprogressive training of neural network constitutive models. *International Journal for Numerical Methods in Engineering*, vol. 42, n. 1, pp. 105–126, 1998.

[13] I.-C. Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete research*, vol. 28, n. 12, pp. 1797–1808, 1998.
[14] R. Haj-Ali, D. A. Pecknold, J. Ghaboussi, and G. Z. Voyiadjis. Simulated micromechanical models using artificial neural networks. *Journal of Engineering Mechanics*, vol. 127, n. 7, pp. 730–738, 2001.
[15] M. Lefik and B. A. Schrefler. Artificial neural network as an incremental non-linear constitutive model for a finite element code. *Computer methods in applied mechanics and engineering*, vol. 192, n. 28-30, pp. 3265–3283, 2003.
[16] Y. Hashash, S. Jung, and J. Ghaboussi. Numerical implementation of a neural network based material model in finite element analysis. *International Journal for numerical methods in engineering*, vol. 59, n. 7, pp. 989–1005, 2004.
[17] J. F. Unger and C. Könke. Coupling of scales in a multiscale simulation using neural networks. *Computers & Structures*, vol. 86, n. 21-22, pp. 1994–2003, 2008.
[18] B. Le, J. Yvonnet, and Q.-C. He. Computational homogenization of nonlinear elastic materials using neural networks. *International Journal for Numerical Methods in Engineering*, vol. 104, n. 12, pp. 1061–1084, 2015.
[19] Z. Liu, C. Wu, and M. Koishi. A deep material network for multiscale topology learning and accelerated nonlinear modeling of heterogeneous materials. *Computer Methods in Applied Mechanics and Engineering*, vol. 345, pp. 1138–1168, 2019.
[20] F. Masi, I. Stefanou, P. Vannucci, and V. Maffi-Berthier. Thermodynamics-based artificial neural networks for constitutive modeling. *arXiv preprint arXiv:2005.12183*, 2020.
[21] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
[22] C. Molnar. *Interpretable machine learning*. Lulu. com, 2020.
[23] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, and others. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, vol. 58, pp. 82–115, 2020.
[24] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, vol. 30, 2017.
[25] M. T. Ribeiro, S. Singh, and C. Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.
[26] A. F. Figueiredo. The challenges of constitutive modeling via artificial neural networks. Master's thesis, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil, 2023.