XLIII CILAMCE
21–25, November, 2022
FOZ DO IGUAÇU – PR – BRAZIL

Use of Random Forest to predict the accumulation of plastic strain at grain boundaries of a polycrystalline material

Lara Cristina Pereira de Araújo[1], Renato Bichara Vieira[1], Helon Vicente Hultmann Ayala [1]

[1]*Department of Mechanical Engineering, Pontifical Catholic University of Rio de Janeiro*
Rua Marquês de São Vicente, 225, Prédio Cardeal Leme, Sala 101L Gávea, 22453-900, Rio de Janeiro/Brazil
lara_araujo@aluno.puc-rio.br, renatovieira@puc-rio.br, helon@puc-rio.br

Abstract. The main motivation is the study of the accumulation of plastic strain in the grain scale, through the use of machine learning. This alternative can be a significant contribution towards creating models capable of predicting the accumulation of strains. In this way, machine learning becomes a tool capable of helping to understand which physical parameters control damage accumulation. The objective of this study is to predict the accumulation of plastic strains at grain boundaries using the Random Forest model. For all machine learning models, it is necessary to perform effectiveness tests and in this study cross-validation was used. It is a numerical work, based on machine learning, which uses the Random Forest algorithm and cross-validation to authenticate the model. The metric used to measure the performance of the model was the coefficient of determination (R^2). Results for the predictions of the accumulation of plastic strains, when considering the same microstructure, are coherent and of good quality. When comparing the results obtained in this work with the predictions found in the literature, the results obtained are satisfactory. Concluded that the Random Forest model is reliable for predicting the accumulation of plastic strains in grain boundaries of a polycrystalline material.

Keywords: Random Forest, Accumulation of plastic strain, Machine Learning, Grain boundaries

1 Introduction

Science has gone through four periods over the centuries. According to, Agrawal and Choudhary [1] the first period was that of empirical science, in which the development of science was based on experiments and experiences. The second period was that of laws, where science is based on formulations of laws and mathematical theorems. With the advancement of computers came the third period, that of computational science, that based on the theorems of the second period was possible to solve complex problems experienced in everyday life. And the fourth period is data-based science, which with the advancement of computational power and the amount of data generated in the third period together with the two initial periods has become increasingly popular Zhang et al. [2].

Machine learning is an option for solving complex problems, in which empiricism and/or theory are not enough and an ally to numerical solutions Zhang et al. [2], Gan et al. [3]. Zhang et al. [2] mentions that several researchers have been using machine learning to solve problems of crack prediction, life prediction, fracture toughness. Rovinelli et al. [4] shows that the direction of fatigue crack propagation predicted by their model, which combines in situ testing, a crystalline aggregate and crystal plasticity to compose the training data, obtained reliable predictions for polycrystalline material.

Abuzaid et al. [5] in 2012 performed a quantitative analysis of the correlation between slip transfer and the plastic strain accumulation around grain boundaries using Digital Image Correlation (DIC) and Electron Backscatter Diffraction (EBSD). The grain boundaries also be evaluated to determine fatigue crack formation, as they play an important role in blocking or slip transfer grain Cheong and Busso [6], Abuzaid et al. [7]. Carroll et al. [8] using combined in situ and ex situ DIC techniques obtained measurements of strains close to the increasing fatigue crack on the grain scale. The use of experimental techniques of DIC and EBSD generated large volumes of data, making possible the use of machine learning techniques in the detection of plastic strain accumulation in grain boundaries, in this case the Random Forest method was used. Breiman [9] determines that Random Forest is the combination of trees, where each one has values sampled independently and with the same distribution for all trees in the forest.

Badora et al. [10] some machine learning algorithms, including Random Forest, are used to establish the maximum size of fatigue cracks based on a sample from a high pressure nozzle of a gas turbine. To better understand the application of machine learning in predicting fatigue life at medium stresses, Gan et al. [3] used two machine

learning models, Random Forest and External Kernel Learning Machine, to improve the mapping between fatigue life and the monotonic, fatigue and cyclic characteristics of materials. In the study by Vieira and Lambros [11] neural networks were used to predict the accumulation of plastic strain at the grain boundaries of an austenitic stainless steel. Predictions were made with three different databases obtained through an experimental technique. The results were satisfactory for most predictions, the best result being the prediction that used its own microstructure to perform the prediction. And the second-best result was the prediction performed through a combination of the three databases.

The need to seek comparisons between methods, and for this a validation is necessary. Thus extending the work carried out by Vieira and Lambros [11]. This option can be an important contribution to reduce time in the context of structural integrity assessment, which often implies time-consuming and costly surveys. In this way, machine learning becomes an efficient alternative for predicted of plastic strains in the crack initiation region, which can initiate a fatigue failure process Vieira and Lambros [11].

It is a numerical work that uses the machine learning method through the Random Forest model Breiman [9] and cross-validation to authenticate the model. The same database of Vieira and Lambros [11] was used. Estimates of plastic strains at the grain boundaries were obtained using 4 different predictions for each sample, with the best result being the prediction that considered a database combination. An evaluation of the coefficient of determination (R^2) was also carried out for all predictions, with satisfactory values being observed.

2 Machine learning methods

Machine Learning is present everywhere, for example, when you do a Google search, an algorithm is responsible for the search result. But what is Machine Learning? It is the technique of programming computers so that they can learn from data. For a machine to learn, it is necessary to provide the machine with historical data to create an algorithm that minimizes or maximizes some measure (loss function or objective).

2.1 Randon Forest method

Random Forest is the combination of several decision trees that can be used for classification and regression problems, in this study it will be used as regression. The Random Forest algorithm can be trained through bagging Breiman [12] (bootstrap aggregation) or pasting that serve to create varied training subsets. Bagging creates subsets of data by sampling with substitution and pasting creates subsets also by sampling, but without substitution, in this study pasting was used. And in the construction of each tree, a single random subset is analyzed at each node division Géron [13]. In regression analyses, the algorithm indicates the result based on the predictions of decision tree that are generated in parallel. And that result is the average of the outputs of all decision tree. Random Forest has more accurate results than the decision tree and is highly resistant to over-fitting. The structure of Random Forest is shown in Figure 1a.

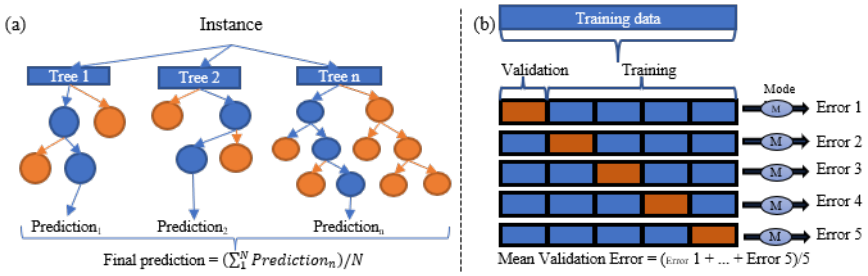

Figure 1. (a) Structure of the Random Forest algorithm. (b) Cross validation model

The hyperparameters used to calibrate the model will be described. The number of trees in the forest, the minimum number of samples needed to split an entire node, the minimum number of samples needed to be in a leaf node. The number of features to consider when looking for the best split, in this case, is equal to the number of samples. And finally, if bootstrap samples are used when building trees, as in this case bootstrap is not used. so the entire dataset is used to build each tree. See Table 1.

Table 1. Values related to the Randon Forest model

Hypermeters					
Description	Value				
Number of trees	10				
Minimum number of split samples	2				
Minimum number of leaf samples	66				
Maximum number of features	None				
Bootstrap	False				
Determination coefficient values R^2					
	Strain	DB 1	DB2	DB3	DB4
Sample 1	ε_{nn}	0.58	0.42	0.32	0.38
	ε_{tn}	0.72	0.41	0.51	0.57
	ε_{tt}	0.68	0.43	0.40	0.50
Sample 2	ε_{nn}	0.58	0.42	0.32	0.38
	ε_{tn}	0.51	0.71	0.52	0.57
	ε_{tt}	0.14	0.61	0.25	0.45
Sample 3	ε_{nn}	0.42	0.40	0.61	0.45
	ε_{tn}	0.39	0.39	0.57	0.46
	ε_{tt}	0.50	0.54	0.68	0.60
Pearson correlation coefficient R					
	Strain	Point 1	Point 2	Point 3	
Sample 1	ε_{nn}	0.81	0.80	0.50	
	ε_{tn}	0.80	0.90	0.67	
Sample 2	ε_{nn}	0.73	0.87	0.72	
	ε_{tn}	0.67	0.13	0.70	
Sample 3	ε_{nn}	0.56	0.44	0.41	
	ε_{tn}	0.75	0.40	0.36	

2.2 Cross validation (*k-folds*)

The cross-validation used in this study was the *k-fold*, which consists of dividing the data set into k sets of the same approximate size, and thus one of the sets is used for testing and the rest for training. This process is repeated until all *k* sets are used at least once as a test set, as shown in Figure 1b. The *k-fold* parameters used were: 10 repetitions for the *k-fold*, the data set was divided by 5 and with a randomness of 4 that controls each repeated instance of the cross-validation.

2.3 Machine learning-based strain prediction

Based on the machine learning technique, a Python code was developed together with the Scikit-Learn package Géron [13] to perform the training and prediction of the random forest regression model using the database validated by Vieira and Lambros [11]. In the code, the data of the three samples used was first loaded and a data frame of each sample was created for a better visualization and location of the data. Subsequently, the data were divided into input and output data (*X,y*). Next, the cross-validation parameters were determined, as described in item 2.2. Finally, the prediction was performed using the hyperparameters described in item 2.1 for each training dataset. After performing the prediction, the R^2 were calculated for each of the outputs. With the results of the predictions, the replacement of the predicted data in the original dataset was made, so that it was possible to save a new data matrix that would allow the generation of visual comparison graphs that were generated through a

MATLAB code provided by Vieira and Lambros [11].

3 Case study description

3.1 Description of the data used

An experimental work presented by Vieira and Lambros [11] was carried out through specimens using the technique of Digital Image Correlation (DIC) and Electron Backscatter Diffraction (EBSD) for the extraction of data around the grains. The material used was alloy 709 which is an austenitic stainless steel. This material was subjected to a heat treatment at 1200ºC for 48 hours so that the grains increased in average size and so the digital image correlation technique could be applied, which was adapted from the work of Carroll et al. [14]. The experimental technique of digital image correlation is widely used to measure displacement and deformation of images, which consists of relating deformed and non-deformed images of a surface by means of markings on the images. To assist in obtaining the grains, an optical microscope was used to take images with high magnifications. After obtaining the images by DIC and EBSD, the datasets were combined to order the strain fields measured with the contained microstructure, in which the alignment was performed with five Vickers markers around the region of interest.

3.2 Description database used

For this study, the same data from Vieira and Lambros [11] were used, which are three different samples, where the first sample is a 199x210 matrix with a total of 41790 data, which in this study is called Database 1 (DB1). The second sample is a 405x419 matrix with a total of 169695 data called Database 2 (DB2). And the third sample is a 409x429 array with 175461 data that has been named Database 3 (DB3). A fourth Database (DB4) was also used, which was composed of 33.33% of each of the previous DBs. Each BD is composed of a dictionary that contains data on plastic s in the normal, tangential, shear, *xx*, *xy* and *yy* directions, geometric angle of the grain boundary α and data that determine where the nucleus and mantle of the grain are located.

For the input data, different parameters can be used, such as displacement rate, force, grain size, temperature, time, geometric angle of the grain boundary, grain core, among others. In this study, the geometric angle of the grain boundary α was used, which was obtained by rotating the strain fields measured in the loading axes (*xy*) to the aligned local coordinate (normal and tangential) in the direction of the local grain boundary.As output data, plastic strains in the three local directions were used (normal ε_{nn}, tangential-normal ε_{tn} and tangential ε_{tt}). In machine learning, DB is usually divided into training and testing, but in this work it was decided to use the entire DB for both training and testing, as the idea was to test the performance of angle α in predicting the accumulation of plastic strains. When creating a fourth DB with 33.33% of all samples, the intention was to verify if the predictor would have a similar performance to the DB with data from a single sample.

4 Results and analysis

In this section, the results of the predictions made for each sample will be presented, as well as the results of R^2. For each sample, predictions were made with four different training databases.First, a comparison of the predictions with the reference image is presented and then a direct comparison with the work of Vieira and Lambros [11]. The hyperparameters used were described in the Table 1.

4.1 Predicted for sample 1

In Figure 2a, the grain boundaries used as a reference are presented, which allows a comparison with the predictions obtained for sample 1. It is observed that in regions with higher concentration of red dots there is a tendency to appear cracks, since in these regions shows the highest level of plastic strain. On the other hand, the dark blue points represent the lowest levels of plastic strain of the material. The grain boundaries generated using 100% of BD1 for training are shown in Figure 2b. When performing the visual comparison with the reference figure, it is possible to affirm that the deformations are coherent with the numerical results presented in Table 1, with the best value, for DB1, of R^2 being 0.72 for the ε_{tn} direction. For DB2 the R^2 value was $\approx$0.42 for the three directions. DB3 reached values between 0.32 and 0.51 for the studied directions. DB4 obtained the highest value of 0.57 for the ε_{tn} direction and the lowest of 0.38 for the ε_{nn} direction.

Sample 1 presents the best result for the prediction obtained by training the dataset of 100% of DB 1, composed by the reference, that is, when generating a prediction using the reference data, there is a better result. DB4 also presented a good prediction of results, similar to DB1, because when using a part of the original data to perform the training, the model is able to make a better prediction.

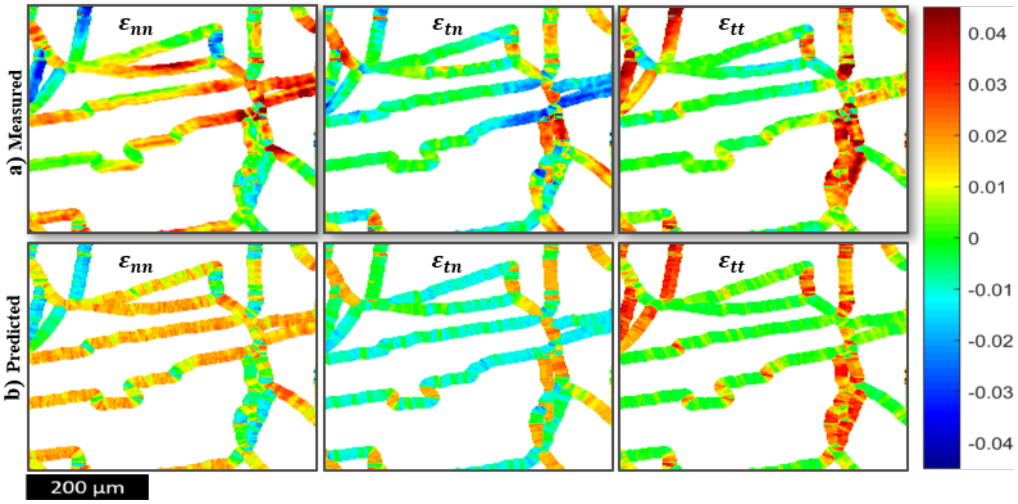

Figure 2. (a) Grain boundary measurements in the mantle regions for sample 1. (b) Equivalent predicted strain field using DB1 for training

4.2 Predicted for sample 2

In Figure 3a, the reference grain boundaries for the sample are shown. Figure 3b shows the prediction values using DB2 to perform the training. Based on the results of R^2 equal to 0.53 (ε_{nn}), 0.71 (ε_{tn}) and 0.61 (ε_{tt}), it can be seen that the levels of plastic strain are similar to the reference. Using DB1 and DB3 for training, similar R^2 values were obtained, as seen in Table 1, with the best result in both cases being in the shear direction (ε_{tn}). The prediction result with BD4 for training is satisfactory, with R^2 values equal to 0.32 (ε_{nn}), 0.57 (ε_{tn}) and 0.45 (ε_{tt}). The sample that presented the best result was DB2, as it is the same data as the reference. The second-best result was the prediction that used the combination of 33.33% of the data from each DB to do the training (see Table 1), showing that the model can improve the prediction when it contains the reference data in the training.

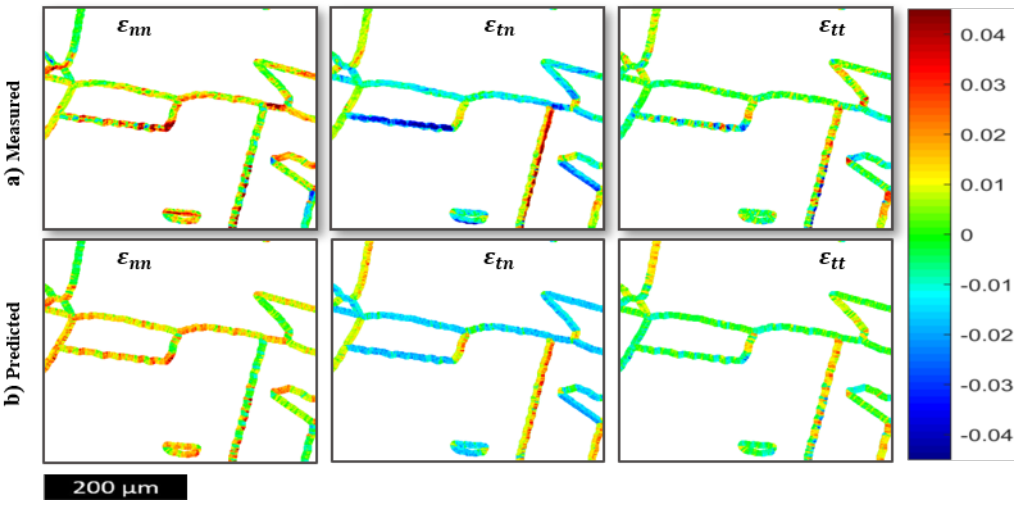

Figure 3. (a) Grain boundary measurements in the mantle regions for sample 2. (b) Equivalent predicted strain field using DB2 for training

4.3 Predicted for sample 3

When using DB1 and DB2 for training the sample, the results are similar, with R^2 values between 0.39 and 0.54 (see Table 1). Using DB3 for training as shown in Figure 4b, the result is ≈19% better than predictions with previous DBs, with an R^2 of 0.68 for the tangential direction. The microstructure generated with BD4 for training is seen in Figure 4c. The result of this prediction had an improvement of ≈6% in relation to DB 1 and 2 and a decrease of ≈11% in relation to DB3, the best value of R^2 for this prediction was 0.60 in the tangential direction.

Same as the previous samples, the best result of sample 3 was constituted by the data set composed by the same of the reference, in this case the BD3. The second-best result was again from DB4 because it contains reference data.

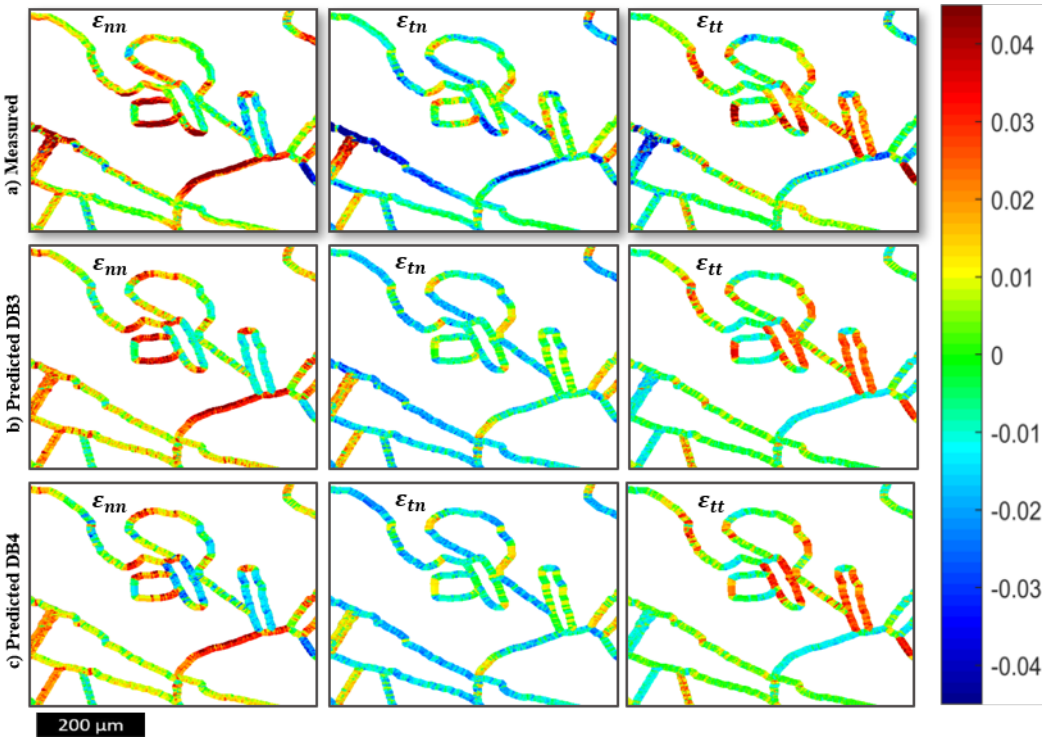

Figure 4. (a) Grain boundary measurements in the mantle regions for sample 3. (b) Equivalent predicted strain field using DB3 for training. (c) Equivalent predicted strain field using DB4 for training.

4.4 Discussion

In the publication by Vieira and Lambros [12] Vieira and Lambros [11], the mean value found for the Pearson correlation coefficient (R) was 0.65 between the samples, with a variation from 0.21 to 0.92. In this work, the average obtained was 0.62 for the samples, with the lowest value being 0.13 and the highest 0.90, as can be seen in Table 1. Observing point 1 of sample 3, in the normal direction, the R value obtained was 0.56, in contrast to the 0.76 of Vieira and Lambros [11]. In the present study, in the shear direction the value obtained was 0.75, however Vieira and Lambros [11] obtained 0.84. Based on the results shown above, it can be verified that the use of the Random Forest method is similar to the Neural Network method in predicting the accumulation of plastic deformation at grain boundaries.

5 Conclusions

In this work, predictions made using the Randon Forest method were presented, considering a reference database for the prediction of plastic strain at the grain boundary of a polycrystalline material. Four different databases were analyzed for the training of the three-sample model. It is concluded that the use of the Random Forest method to predict residual plastic strain in grain boundaries of a polycrystalline material presents similar

results to those obtained by Vieira and Lambros [11]. Observing the coefficient of determination, R^2, it is clear that the values obtained are presented satisfactorily. In view of all the above, the angle α is a good predictor for the relationship between normal and shear strains at grain boundaries, but there are also other parameters that can be used for this same scenario.

The continuation of this research will use a region of the sample as an input parameter of the model and the extraction of the principal components through the algorithm of Principal Component Analysis (PCA) to reduce the dimensionality of the model, then a new prediction will be carried out with the Randon Forest model to evaluate the performance of this new model.

Acknowledgements. The authors would like to acknowledge the support from the Coordination for the Improvement of Higher Education Personnel (CAPES).

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] A. Agrawal and A. Choudhary. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Materials*, vol. 4, n. 5, pp. 53208, 2016.

[2] X. C. Zhang, J. G. Gong, and F. Z. Xuan. A deep learning based life prediction method for components under creep, fatigue and creep-fatigue conditions. *International Journal of Fatigue*, vol. 148, n. March, pp. 106236, 2021.

[3] L. Gan, H. Wu, and Z. Zhong. Fatigue life prediction considering mean stress effect based on random forests and kernel extreme learning machine. *International Journal of Fatigue*, vol. 158, pp. 106761, 2022.

[4] A. Rovinelli, M. D. Sangid, H. Proudhon, and W. Ludwig. Using machine learning and a data-driven approach to identify the small fatigue crack driving force in polycrystalline materials. *npj Computational Materials*, vol. 4, n. 1, pp. 35, 2018.

[5] W. Z. Abuzaid, M. D. Sangid, J. D. Carroll, H. Sehitoglu, and J. Lambros. Slip transfer and plastic strain accumulation across grain boundaries in Hastelloy X. *Journal of the Mechanics and Physics of Solids*, vol. 60, n. 6, pp. 1201–1220, 2012.

[6] K. S. Cheong and E. P. Busso. Effects of lattice misorientations on strain heterogeneities in FCC polycrystals. *Journal of the Mechanics and Physics of Solids*, vol. 54, n. 4, pp. 671–689, 2006.

[7] W. Abuzaid, H. Sehitoglu, and J. Lambros. Plastic strain localization and fatigue micro-crack formation in Hastelloy X. *Materials Science and Engineering A*, vol. 561, pp. 507–519, 2013.

[8] J. D. Carroll, W. Z. Abuzaid, J. Lambros, and H. Sehitoglu. On the interactions between strain accumulation, microstructure, and fatigue crack behavior. *International Journal of Fracture*, vol. 180, n. 2, pp. 223–241, 2013.

[9] L. Breiman. Random Forests. *Machine Learning*, vol. 45, n. 1, pp. 5–32, 2001.

[10] M. Badora, M. Sepe, M. Bielecki, A. Graziano, and T. Szolc. Predicting length of fatigue cracks by means of machine learning algorithms in the small-data regime. *Eksploatacja i Niezawodnosc - Maintenance and Reliability*, vol. 23, n. 3, pp. 575–585, 2021.

[11] R. B. Vieira and J. Lambros. Machine Learning Neural-Network Predictions for Grain-Boundary Strain Accumulation in a Polycrystalline Metal. *Experimental Mechanics*, vol. 61, n. 4, pp. 627–639, 2021.

[12] L. Breiman. Bagging predictors. *Machine Learning*, vol. 24, n. 2, pp. 123–140, 1996.

[13] A. Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly Media, 2017.

[14] J. Carroll, W. Abuzaid, J. Lambros, and H. Sehitoglu. An experimental methodology to relate local strain to microstructural texture. *Review of Scientific Instruments*, vol. 81, n. 8, pp. 083703, 2010.