A methodology to generate design allowables of composite laminates using machine learning

C. Furtado a,b,*, L.F. Pereira b,c,d, R.P. Tavares b,e, M. Salgado b,c, F. Otero b,f, G. Catalanotti g, A. Arteiro b,c, M.A. Bessa d, P.P. Camanho b,c

a Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, United States
b INEGI, Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Rua Dr. Roberto Dias, 400, 4200-465 Porto, Portugal
c DEMec, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Dias, s/n, 4200-465 Porto, Portugal
d Department of Materials Science and Engineering, Delft University of Technology, 2628 CD Delft, The Netherlands
e Department of Materials, Textiles and Chemical Engineering, Ghent University, Technologiepark-Zwijnaarde 46, B-9502 Zwijnaarde, Belgium
f Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, United States.
g Escola de Ciências e Tecnologia, Universidade de Évora, Colégio Luís António Verney, Rua Romão Ramalho, 59, 7000-671 Évora, Portugal

A R T I C L E   I N F O

Keywords:
Polymer-matrix composites (PMCs)
Machine learning
Fracture mechanics
Design allowables

A B S T R A C T

This work represents the first step towards the application of machine learning techniques in the prediction of statistical design allowables of composite laminates. Building on data generated analytically, four machine algorithms (XGBoost, Random Forests, Gaussian Processes and Artificial Neural Networks) are used to predict the notched strength of composite laminates and their statistical distribution, associated to the uncertainty related to the material properties and geometrical features. This work focuses not only on the so-called Legacy Quad Laminates (0°/90°/±45°), typically used in the design of composite aerostructures, but also on the newer concept of double-double (or double-angle ply) laminates. Very good representations of the design space, translating in low generalization relative errors of around ±10%, and very accurate representations of the distributions of notched strengths around single design points and corresponding B-basis allowables are obtained. All machine learning algorithms, with the exception of the Random Forests, show very good performances, with Gaussian Processes outperforming the others for very small number of data points while Artificial Neural Networks have better performance for larger training sets. This work serves as basis for the prediction of first-ply failure, ultimate strength and failure mode of composite specimens based on non-linear finite element simulations, providing further reduction of the computational time required to virtually obtain the design allowables for composite laminates.

1. Introduction

The generation of design allowables for composite laminates is of utmost importance for the design and certification of the composite structures used in the aerospace industry. The determination of these design allowables, which account for the variability associated with curing/consolidation procedures, geometrical features and defects characteristic of composite structures, usually relies on extensive, expensive and time-consuming experimental test campaigns. With the increase of computational power, and the development of high-fidelity numerical models that accurately represent the response and failure of composite materials, alternatives to generate design allowables based on advanced finite element simulations have also been sought out to reduce the certification costs (Tay et al., 2005; der Meer et al., 2010; Ling et al., 2009; Schuecker and Pettermann, 2006; Camanho et al., 2007a; Vogler et al., 2013; Camanho et al., 2013; Abdi et al., 2016; Zhang et al., 2017; Abumeri et al., 2011; Spendley, 2012). However, these solutions are still computationally expensive, especially if uncertainty is accounted for. The recent advances on machine learning techniques open a new window of possibilities for the faster prediction of the structural response of composite materials and their optimization (Bessa et al., 2017; Bessa and Pellegrino, 2018; Bessa et al., 2019; Bisagni and Lanzi, 2002; Yvonnet and He, 2007), by allowing the definition of surrogate models that continuously and analytically describe the design space.

https://doi.org/10.1016/j.ijsolstr.2021.111095
Received 5 October 2020; Received in revised form 6 May 2021; Accepted 21 May 2021
Available online 26 May 2021
0020-7683/© 2021 Elsevier Ltd. This article is made available under the Elsevier license (http://www.elsevier.com/open-access/userlicense/1.0/).
In recent years, significant effort has been made to use surrogate models to represent the response of composite materials and, consequently, reduce the computation cost of finite element simulations (El Said and Hallett, 2018; Balokas et al., 2018; Yan et al., 2020). In general, the process relies on i) design of experiments, where the descriptors of the design space are defined, ii) data generation, where data to train and test the surrogate models is obtained, and iii) surrogate model definition (or machine learning application) where a model that represents the design space is defined. For instance, El Said and Hallett (2018) and Yan et al. (2020) proposed multiscale modelling approaches based on surrogate models at the mesoscale: the former to predict the elastic response of structures with internal defects (namely wrinkles), and the latter to predict composite structural damage and failure. Both approaches relied on the definition of Representative Volume Elements (RVE) at the mesoscale, which were used to populate the design space and train a surrogate model. The surrogate model was then used to represent the composite response at the macroscale, avoiding the need to run RVE models in parallel with the macroscale simulation, significantly reducing the computational cost of the models. These surrogate models focus on the representation of the material behaviour at the mesoscale level and still rely on numerical simulations at the macroscale. If standard tests for certification of composite materials (plain strength, open-hole strength, bolted strength, among others) can be described parametrically and accurate analytical surrogate models can be built on data from numerical simulations, the virtual certification of composite materials can be greatly simplified. This paper presents the first steps towards that goal.

In this paper, a feasibility study on the application of machine learning techniques for predicting a design allowable, the notched strength of multidirectional composite laminates, is presented, with the main goals of presenting the challenges of applying machine learning techniques for composite laminates, and of evaluating the most appropriate algorithms for the determination of composite design allowables. Even though the data-driven framework is established on data derived from an analytical framework (Furtado et al., 2017; Vallmajo et al., 2019), this work serves as basis and guideline to a more demanding challenge that includes the prediction of first-ply failure strength, ultimate strength and failure mode of composite materials based on nonlinear finite element simulations. The data-driven framework is defined following the procedure schematically shown in Fig. 1:

1. Three ply-level material properties: the longitudinal Young’s modulus, $E_l$, the longitudinal strength, $X_l$, and the critical energy release rate, $G_0$, or longitudinal crack resistance curve, $R$-curve;
2. The stacking sequence of the laminate.
3. Two geometrical features: the width of the specimen, $W$, and the diameter of the hole, $D_h$.

The proposed methodology (Furtado et al., 2017) can be summarized on a combination of three building blocks: a finite fracture mechanics model (Camanho et al., 2012), invariant-based approaches to estimate stiffness and strength (Tsai and Melo, 2014; Tsai and Melo, 2016) and an analytical model based on Fracture Mechanics to estimate the laminate fracture toughness (Camanho and Catalanotti, 2011).

The combination of the models provides an efficient framework to predict the laminate elastic, strength and fracture properties required to predict size effects in notched laminates. The model requires the knowledge of:

1. Three ply-level material properties: the longitudinal Young’s modulus, $E_l$, the longitudinal strength, $X_l$, and the critical energy release rate, $G_0$, or longitudinal crack resistance curve, $R$-curve;
2. The stacking sequence of the laminate.
3. Two geometrical features: the width of the specimen, $W$, and the diameter of the hole, $D_h$.

The proposed methodology (Furtado et al., 2017) can be summarized as follows:

1. Three ply-level material properties: the longitudinal Young’s modulus, $E_l$, the longitudinal strength, $X_l$, and the critical energy release rate, $G_0$, or longitudinal crack resistance curve, $R$-curve;
2. The stacking sequence of the laminate.
3. Two geometrical features: the width of the specimen, $W$, and the diameter of the hole, $D_h$.

The proposed methodology (Furtado et al., 2017) can be summarized as follows:

Fig. 1. Schematic representation of the machine learning framework.
as follows (see also Fig. 2):

1. From the longitudinal Young’s modulus of the ply, \( E_i \), the Master Ply concept is used to determine Tsai’s modulus (Artur et al., 2020), \( T_r \), and estimate the ply elastic properties (Tsai and Melo, 2014; Artur et al., 2019). From Tsai’s modulus, lay-up, and using the universal laminate factors, the elastic properties of the laminate are obtained.

2. From the longitudinal strength of the ply, \( X_i \), and using the universal laminate factors of the laminate determined using Tsai’s modulus, a maximum allowable strain criterion is applied to predict the laminate unnotched strength, \( X^c \).

3. From the critical energy release rate, \( G^0 \), and using the elastic properties of the balanced and 0° sublaminates determined using Tsai’s modulus, an analytical model based on fracture mechanics (Camanho and Catalaniotti, 2011) is used to predict the laminate fracture toughness, \( \sigma^f \).

4. Finally, the laminate unnotched strength, \( X^c \), and the laminate fracture toughness, \( \sigma^f \), are used in the Finite Fracture Mechanics model (Camanho et al., 2012) to predict the notched strength (\( G^f \)) of the laminate for any specimen configuration (\( W \) and \( D \)).

The analytical framework is able to accurately predict the tensile notched strength of composite laminates as shown in references Camano et al. (2012), Erçin et al. (2013), Arteiro et al. (2013), Arteiro et al. (2014), Vallmajör et al. (2019) and Furtado et al. (2017) where its predictions were compared with experimental results. The original analytical model (Camanho et al., 2012) was applicable for predicting the strength of quasi-isotropic laminates, therefore, a recently proposed generalization for highly orthotropic materials was used here (Catalaniotti et al., 2021).

Given its simplicity and efficiency, the analytical framework summarised above is a good candidate to perform a feasibility investigation to appropriately define the input parameters/descriptors, to assess the ability of machine learning algorithms to predict the strength of composite laminates, and to identify the most effective algorithms compatible with structural analysis of composite laminates.

### 2.2. Strategies for laminate definition

Most of the input parameters required by the analytical framework can be treated as continuous and independent variables, appropriate to build training data to feed machine learning algorithms. However, the laminate stacking sequence requires a more detailed discussion. Considering each ply orientation of a stacking sequence an input parameter results in a high dimensional representation, which is inconvenient in machine learning applications since larger datasets are generally required to accurately capture the design space.

In the following section, two strategies for the definition of laminates are presented: one based on laminate parameters (Tsai and Pagano, 1968), and another based on the double-double (or double angle-ply) laminates recently proposed by Tsai et al. (2017) as a practical alternative to the conventional Legacy Quad laminates.

#### 2.2.1. Conventional laminates

The most common way to define a laminate is by its stacking sequence (i.e. assembly of plies with specified fibre orientation angles). This is a simple and convenient representation, but has a major drawback: the number of variables is intrinsically related to the number of plies. This translates in a high-dimensional laminate representation and, more importantly, in a representation with variable dimension.

Lamination parameters, firstly proposed by Tsai et al. in 1968 (Tsai and Pagano, 1968), provide a more compact definition of the laminate: twelve lamination parameters and a thickness variable are sufficient to geometrically define any laminate, independently of the number of plies. However, it is important to note that these twelve lamination parameters are interrelated, i.e. when the values of some parameters are fixed, the other are constrained to a certain feasible region. Significant time and effort have been dedicated to efficiently define the constraints of the 12-dimensional convex (Grenestedt and Gudmundson, 1995) feasible domain of the laminate parameters, both with (Bloomfield and Wagner, 2009) and without (Setoodeh et al., 2006) restrictions on the possible ply orientations. The in-plane, \( \xi^A_{1 \ldots 2 \ldots 3 \ldots 4} \), coupled, \( \xi^B_{1 \ldots 2 \ldots 3 \ldots 4} \), and out-of-plane, \( \xi^D_{1 \ldots 2 \ldots 3 \ldots 4} \), laminate parameters are calculated as:

\[
\xi^A_{1 \ldots 2 \ldots 3 \ldots 4} = 1 \sum_{i=1}^{N} \left\{ \cos(\theta_i) \cos(4\theta_i) \sin(2\theta_i) \right\} (\xi_i - \xi_{i-1})
\]

\[
\xi^B_{1 \ldots 2 \ldots 3 \ldots 4} = 2 \sum_{i=1}^{N} \left\{ \cos(2\theta_i) \cos(4\theta_i) \right\} (\xi_i^2 - \xi_{i-1}^2)
\]

\[
\xi^D_{1 \ldots 2 \ldots 3 \ldots 4} = 4 \sum_{i=1}^{N} \left\{ \cos(2\theta_i) \cos(4\theta_i) \right\} (\xi_i^2 - \xi_{i-1}^2)
\]

where \( h \) is the thickness of the laminate, \( \theta_i \) is the fibre orientation at height \( z \in [z_i \ldots z_{i+1}] \), and \( N \) is the number of plies of a laminate. The twelve lamination parameters fully and uniquely represent a laminate, and, therefore, they can be converted to stacking sequences. However, this is a non-trivial process, whose solution generally requires restrictions to the possible permitted ply orientations and the use of optimization algorithms (Jiesselmuidden et al., 2009; Irissari et al., 2011; Meddaikar et al., 2017; Bloomfield et al., 2010) or other techniques (Todoroki and Sekishiro, 2007; Liu et al., 2019; Viquerat, 2020).

For the problem at hand, only two lamination parameters need to be considered, \( \xi^A_{1 \ldots 2} \), because:

1. The analytical framework previously described deals with in-plane loading only, and considers that the laminate is homogenized (not necessarily symmetric), i.e. the bending-extension coupling matrix \( B \approx 0 \); in other words, it’s only capable of distinguishing lay-ups and not stacking sequences. Since laminate organization is not accounted for, the laminates can be fully described by the in-plane laminate parameters, \( \xi^A_{1 \ldots 2 \ldots 3 \ldots 4} \).

---

1 To compute the notched strength using the finite fracture mechanics model proposed in Ref. Camano et al. (2012), the stress intensity factor for a plate with a central circular hole of radius \( R \) and two symmetric cracks emanating from the hole edge needs to be calculated. However, an analytical expression for the stress intensity factor of this configuration only exists for quasi-isotropic laminates (Newman Jr., 1983). The recent generalization mentioned and used in this work stems from a numerical study to determine the stress intensity factor of cracks emanating from circular and elliptical holes in orthotropic plates. The approach is based on the original work from Suo et al. (1991) and a semi-analytical expression for the correction factor, \( \phi = \frac{1}{\sqrt{2\pi r}} \), was used to take into account the effects of orthotropy.

2 Legacy Quad laminates refers to laminates composed of 0°, 90° and ±45° plies, with a minimum of 10% of each orientation, symmetric at the mid-plane and balanced (same number of 45° and −45° plies). These laminate thumb rules have been used to limit the complexity of composites design.
2. Only balanced laminates can be considered Camanho and Catralanotti, 2011; Furtado et al., 2017, therefore $\zeta_{\{3,4\}} = 0$.

As shown in Fig. 3, this definition based on the lamination parameters $\zeta_{\{1,2\}}$. Results in a non-rectangular (i.e. $\zeta_{\{1,2\}}$ are not fully independent) representation of the normalized stiffness. This will be further commented in Section 3.

2.2.2. Double-double laminates

An alternative to the Legacy Quad laminates, where plies with orientation of $0^\circ$, $90^\circ$ and $\pm 45^\circ$ are used, was recently proposed by Tsai et al. (2017): the Double-Double (DD) laminates. These laminates are composed of plies of two orientations and are defined as $[\pm \phi / \pm \psi]$ or $[+\phi / + \psi / - \phi / - \psi]_n$, where $n$ is the number of repetitions. This bi-angle approach to laminate design results in stronger laminates with higher resistance to micro-cracking and delamination and in other advantages such as faster layup, simpler design and easier tapering through single ply drops (Tsai et al., 2017; Shrivastava et al., 2020).

DD-sublaminates lay-ups can be fully described by two parameters, $\phi$ and $\psi$, which can vary continuously from $0^\circ$ to $90^\circ$. This provides a smaller design space, which, in one hand, reduces the design flexibility attributed to composite materials, but, on the other hand, allows for a more efficient and otherwise nonviable stacking sequence optimization.

As shown in Fig. 4, this definition based on the ply orientations, $\phi$ and $\psi$, results in a continuous and injective representation of the normalized stiffness (Tsai et al., 2017). Furthermore, the normalized stiffness is diagonal symmetric, a characteristic that will be further commented on and taken advantage of in Section 3.

2.3. Definition of dimensionless parameters

As described in Section 2.1, the notched strength of an open-hole specimen, $\sigma$, can be fully described as:

$$\sigma = f_1(E_1, X_T, \zeta_0, D, W/D, \text{lay-up})$$

where $E_1$, $X_T$ and $\zeta_0$ are material descriptors, $D$ and $W/D$ are geometric descriptors and the lay-up can be described by:

$$\text{lay-up} = f_2(\alpha, \beta)$$

where $\alpha = \zeta_{\{1,2\}}$ for a Quad laminate and $\alpha = \phi$ and $\beta = \psi$ for a DD laminate (Section 2.2), comprising a total of seven input parameters. The number of input parameters can be reduced using the Buckingham’s $\Pi$ theorem (Buckingham, 1914) that states that, if there is a physical relation involving $N$ physical variables, then the relation can be rewritten as a relation of $(N-K)$ dimensionless products where $K$ is the number of independent variables.
number of fundamental dimensions required to describe the physical variables. Following the Buckingham II theorem, the number of dimensions can be reduced to four dimensions, using the following dimensionless parameters:

\[
\frac{\sigma}{X_T} = f\left(\frac{E_i}{DX_T}, \frac{W}{D, a, \beta}\right)
\]

This provides a more compact description of the problem at hand, which is convenient in machine learning applications since high dimensional spaces generally require larger datasets to provide accurate surrogate models.

Note that the design of experiments (presented in Section 3) is performed in the original 7D space (Eq. 4), which is then converted to the dimensionless 4D space (Eq. 6), which is in turn used to train and test the machine learning models. Although other strategies could have been defined, this strategy stems from the fact that the relation between the 4D and 7D space is not necessarily injective. For example, it is possible that two specimens with the same geometry (W/D) but made from two different materials have the same 4D parameters and therefore the same notched strength. For this reason, it was more convenient to define the design space in the 7D space, guaranteeing that each variable is well populated in its defined bounds and then convert them to the 4D space before the training step.

3. Design of experiments

The design of experiments was performed using the implementation of Sobol sequence (Sobol, 2001; Saltelli, 2002; Saltelli et al., 2010) in the SALib Python library (Herman and Usher, 2017): a low discrepancy quasi-random sequence designed to explore the space in a more uniform manner than random sampling \(^3\).

Sampling schemes (Sobol, 2001; Saltelli, 2002; Saltelli et al., 2010; McKay et al., 1979) were designed considering continuous variables with fixed bounds (i.e. no variable dependency is allowed). This attribute is compatible with the material property (\(E_i, X_T, \varepsilon_k\)) and geometric (D, W/D) input parameters. However, the laminate descriptors require a more careful analysis: on one hand, when considering conventional laminates, it is clear that the lamination parameters \((\zeta_1, \zeta_2)\) are not independent (represented by the non-rectangular region in Fig. 3) and depending on the hypothesis initially considered (maximum number of plies, ply angles allowed, etc.), yield discrete combinations of values. On the other hand, when considering DD laminates, and even though \(\beta\) and \(\delta\) are independent variables, advantage can be taken of the fact that the stiffness is diagonal symmetric as shown in Fig. 4.

The strategies used to conform the sampling schemes to the laminate descriptors are presented herein.

3.1. Laminate descriptors for conventional laminates

Sampling has to be performed in the lamination parameters space, however, the analytical framework presented in Section 2 requires the lay-up as an input parameter to compute the notched strength. For this reason, there is the need to solve the inverse problem, i.e. obtain the stacking sequence from the lamination parameters. This is a non-trivial problem, that generally involves time-consuming optimization algorithms (IJsselmuiden et al., 2009; Irisarri et al., 2011; Meddaikar et al., 2017; Bloomfield et al., 2010) and can only be performed in an acceptable time frame if restrictions are imposed, such as a fixed number of plies, discrete values of allowed angles, among others. To avoid the in situ computation of the inverse problem, a database that relates all the possible \(\zeta_1/\zeta_2\) combinations to a corresponding lay-up was created. The following laminate restrictions were imposed:

1. Only ply orientations of 0°, +45°, −45° and 90° are considered (Legacy Quad laminates);
2. The number of plies of each orientation range from 0 to 32;
3. All laminates are balanced.

Since laminate organization (i.e. stacking sequence) is not accounted for in the model, and only balanced laminates can be considered (the number of +45°, −45° plies is equal), this yields \(32^2 - 1 = 32767\) laminate permutations (27133 unique orientation-percentage combinations), which configures a database with acceptable size, that can be efficiently accessed to convert lamination parameters to a corresponding lay-up. Note that for less restrictive constraints (more angles considered, larger number of plies, non-balanced laminates, etc.), this will yield a database too large to be accessed efficiently.

Sobol sampling assumes the variables to be continuous and independent from one another. As shown in Fig. 5, where all the 27133 possible pairs of \(\zeta_1/\zeta_2\) points are plotted in grey, this is not verified in the present case. For this reason, the distance of all the Sobol generated points (in red) to the closest allowed point is calculated. The Sobol generated point is either discarded, in case the computed distance is higher than a given threshold value, or approximated to the closest allowed point (in blue). If the defined threshold distance is too high, all points will be approximated to the closest allowed point, leading to a densely populated region at the allowed \(\zeta_1/\zeta_2\) boundary (in this case, the triangle boundary shown in Fig. 5). If the threshold distance is too low, most points will be discarded. A sensitivity analysis was made to analyse the effect of the selected threshold distance on the performance of the trained models and no significant dependency was found (the analysis is not shown here for the sake of conciseness). The authors considered that a threshold distance of 0.1, that would guarantee that all points inside the boundary would be included and most points outside the boundary would be discarded, was appropriate as a criterion for discarding generated points. This is a simple and efficient strategy to select the Sobol sequence sampling points that conform with the constraints of the problem. A more complex sampling method that respects the restrictions imposed to the laminate descriptors could be potentially also be envisioned.

![Fig. 5. Sobol sampling modification for Legacy Quad laminates.](image-url)
3.2. Laminate descriptors for DD laminates

For the DD laminates, sampling is performed in the ply orientation space, which is continuous and composed of independent variables. The two angles are then simply organized to form a laminate: \([\pm \phi/\pm \psi]\). Since laminate organization is not accounted for by the analytical framework, the number of repetitions \(n\) does not need to be considered as the percentage of layers in each direction is independent of the number of repetitions. Moreover, the DD laminates are balanced laminates by definition, and therefore, no further restrictions are required.

However, as shown in Fig. 6, the stiffness is diagonal symmetric with respect to \(\phi\) and \(\psi\) because the angles are interchangeable. Although this is not required, advantage can be taken of the fact that the stiffness is diagonal symmetric to obtain a more densely packed distribution in the region of interest. As shown in Fig. 6, a \(\phi/\psi\) Sobol generated point for which \(\psi > \phi\) is converted into a \(\psi/\phi\) point.

4. Computational analyses

In the previous section, the design of experiments, where the input variables, their bounds and strategies to populate the design space are defined, was described. Here, the analysis, where the output variable (the notched strength) is computed for the defined input data (material, geometric and laminate descriptors), is described. In this work, the output considered is the open-hole strength of a laminate as determined using the analytical framework described in Section 2.

For conventional laminates, for each input data point, \(x_i = [E_1, x_T, E_0, D, W/D, c_1, c_3, \psi_1, \psi_3]\):

1. the laminate parameters, \(c_1\) and \(c_3\), are converted in a lay-up by accessing the database that configures all the possible \(c_1^i/c_3^i \leftrightarrow \) lay-up combinations.
2. The notched strength, \(\sigma\), is computed.
3. The original 7D parameters are converted to the 4D design space: \(x_i^* = \frac{c_1}{c_3}. W/D, c_3^i, \psi_1, \psi_3\).
4. The normalized notched strength, \(\sigma/X_T\), is computed.

For the DD laminates, for each input data point, \(x_i = [E_1, x_T, E_0, D, W/D, \phi, \psi, |\psi|];\)

1. The laminate descriptors, \(\phi\) and \(\psi\), are organized in a lay-up: \([\pm \phi/\pm \psi]\)
2. The notched strength, \(\sigma\), is computed.
3. The original 7D parameters are converted to the 4D design space: \(x_i^* = \frac{c_1}{c_3}. W/D, \phi, \psi, |\psi|\).

Since laminate organization is not accounted for by the analytical framework, the number of repetitions \(n\) does not need to be considered as the percentage of layers in each direction is independent of the number of repetitions. Moreover, the DD laminates are balanced laminates by definition, and therefore, no further restrictions are required.

However, as shown in Fig. 6, the stiffness is diagonal symmetric with respect to \(\phi\) and \(\psi\) because the angles are interchangeable. Although this is not required, advantage can be taken of the fact that the stiffness is diagonal symmetric to obtain a more densely packed distribution in the region of interest. As shown in Fig. 6, a \(\phi/\psi\) Sobol generated point for which \(\psi > \phi\) is converted into a \(\psi/\phi\) point.

4. The normalized notched strength, \(\sigma/X_T\), is computed.

Since the framework is based on an analytical model, the computation of a notched strength is not time-consuming, and therefore, not a limiting factor. For reference, the notched strength of 1000 data points takes 8 s to be computed in a standard laptop. For more computationally expensive problems, e.g. based on finite element simulations, the amount of information available (or that can be generated) is more limited. Therefore, the goal here is to obtain good approximations of the design space with the least amount of data possible and to show how the available number of data points affect performance of the models.

5. Prediction of open-hole tensile strength of composite laminates using machine learning

At this point, a database of well distributed points has been generated and is available to train and test machine learning algorithms. As schematically shown in Fig. 1, in data-rich situations, a given machine learning algorithm, with fixed hyperparameters (parameters that are not optimized during the learning process) is trained using a training set. The trained model is then used to predict the output of the test set points in order to estimate the generalization error of the model. For small number of data points, the randomness associated with the training set used is higher and several repetitions of the evaluation procedure must be performed to obtain a more reliable estimate of the generalization error (or, alternatively, more robust evaluation procedures, such as k-fold cross validation Raschka, 2018a, can be used). It is not trivial to quantify how much data is sufficient, since it depends on several factors, such as the complexity of the problem and the models being fitted to the data (Hastie et al., 2013). Further considerations must be taken into account when several machine learning algorithms and hyperparameter combinations are tested. This will be further commented on in Section 5.2.

Gaussion Processes (Krige, 1951), Artificial Neural Networks (Rosenblatt, 1958) and two tree-ensemble algorithms (Random Forests (Breiman, 2001) and XGBoost (Chen et al., 2016) are considered here. The two tree-ensemble algorithms have a fast learning process, do not require data standardization, have easily tunable hyperparameters and have been reported to be very powerful algorithms when dealing with structured data (Chollet, 2017; Hamidieh, 2018). Furthermore, they have been extensively applied in Kaggle competitions (Hamidieh, 2018; Chollet, 2017), and have been successfully applied in the materials science realm, e.g. for predicting the critical temperature of superconductors (Hamidieh, 2018; Stanev et al., 2018). Artificial Neural Networks are very powerful and extremely scalable, being able to deal with large datasets, but are harder to tune properly, e.g. require adequate hyperparameter tuning to provide good predictions. Gaussian Processes are also very powerful, particularly for regression problems, provided adequate kernel selection (Rasmussen and Williams, 2006; Duvenaud, 2014) is performed, but are poorly scalable. Recently, Gaussian Processes and Artificial Neural Networks have been extensively used in e.g. the design of materials (Bessa et al., 2017; Bessa et al., 2019) and structures (Bessa and Pellegrino, 2018).

The implementations of Random Forests and Gaussian Processes in scikit-learn (Pedregosa et al., 2011), the implementation of XGBoost in xgboost (Chen et al., 2016) and the implementation of Artificial Neural Networks in keras (Chollet et al., 2015) open-source software libraries were used throughout in this work. All the algorithms are summarized hereafter.

5.1. Machine learning algorithms

A Random Forest (Breiman, 2001) is a tree-ensemble technique that combines multiple weak learners (decision trees), each trained with a subset of the training set (this strategy, called bootstrap aggregation or bagging, improves the stability and accuracy of the trained model), and
averages their output in an attempt to produce a strong learner. XGBoost, although also a tree-ensemble technique, is based on gradient boosting concept (Friedman, 2001), i.e. the weak learners are trained sequentially in order to predict the error residuals of previous learners. A relevant feature of these tree-based algorithms is that they provide feature importances, i.e. they quantify the importance of each parameter for the output prediction.

Gaussian Processes are a Bayesian machine learning method (thus, allow the incorporation of prior knowledge in the learning process) that is performed through maximization of the log-marginal-likelihood (Pedregosa et al., 2011). Being a probabilistic method, this approach is able to compute a posterior distribution for the output prediction.

The choice of the best hyperparameters for a given algorithm must be performed without access to the test set (Cawley and Talbot, 2010; Varma and Simon, 2006), i.e. the hyperparameter selection must be viewed as a problem of efficient model selection. After the selection of the best hyperparameters, the model was retrained using the full training set and a generalization error estimate was computed using the test set.

### 5.2. Model selection and assessment

In machine learning, a hyperparameter is a parameter whose value controls the learning process. The hyperparameters must ensure that the machine learning model is flexible enough to adapt to the intricacies of the problem (to avoid underfitting), but not too flexible to over-adapt to the training set (to avoid overfitting). Table 2 shows the possible values established for each hyperparameter (if not mentioned, the defaults of the respective library implementations are used). Due to its simplicity, grid search, a brute-force technique that consists in searching through a manually specified subset of the hyperparameter space of a learning algorithm, is used to optimize the hyperparameters. More automatic approaches, such as Bayesian optimization of the machine learning algorithms ( Snoek et al., 2012), could also be used to tune the hyperparameters of the ML algorithms.

### 5.3. Learning curves

The easiest way to improve the performance of an algorithm is to collect more data (although there is normally a plateau, i.e. after a given number of training points the performance does not improve significantly anymore). Nevertheless, gathering data is usually expensive and, since this work intends to serve as a feasibility study, it is relevant to understand how the size of the dataset affects the performance of the model. In order to study such influence, learning curves for Legacy Quad and DD laminates are presented in Figs. 7 and 8, respectively. The root mean squared error (RMSE) is chosen as error metric.

For each algorithm and number of points presented in the learning curves, the model selection and assessment procedures described in

### Table 1

<table>
<thead>
<tr>
<th>Material Property</th>
<th>Legacy Quad</th>
<th>Double-Double</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_1$ (GPa)</td>
<td>$E_1$ (GPa)</td>
<td></td>
</tr>
<tr>
<td>$X_1$ (MPa)</td>
<td>$X_1$ (MPa)</td>
<td></td>
</tr>
<tr>
<td>$S_0$ (N/mm)</td>
<td>$S_0$ (N/mm)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Geometric</th>
<th>Legacy Quad</th>
<th>Double-Double</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$ (mm)</td>
<td>$D$ (mm)</td>
<td></td>
</tr>
<tr>
<td>$W/D$</td>
<td>$W/D$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layup</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_1^\theta$</td>
<td>$\psi$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Substrate</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\xi$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Hyperparameter</th>
<th>Search values</th>
<th>DD</th>
<th>Quad</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>n_estimators</td>
<td>100, 500, 1000</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td></td>
<td>max_depth</td>
<td>4, 10</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>learning_rate</td>
<td>0.01, 0.1, 0.2</td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>subsample</td>
<td>0.5, 1</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>colsample_bytree</td>
<td>0.5, 1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td>n_estimators</td>
<td>100, 500, 1000, 5000</td>
<td>5000</td>
<td>1000</td>
</tr>
<tr>
<td></td>
<td>max_depth</td>
<td>1, 7, 20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>min_sample_leaf</td>
<td>1, 5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>max_features</td>
<td>auto, sqrt</td>
<td>auto</td>
<td></td>
</tr>
<tr>
<td>Artificial Neural Network</td>
<td>hidden layer 1</td>
<td>8, 16, 64</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>hidden layer 2</td>
<td>8, 16</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hidden layer 3</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Gaussian Processes</td>
<td>Matern kernel (κ)</td>
<td>1/2, 3/2, 5/2</td>
<td>5/2</td>
<td>3/2</td>
</tr>
<tr>
<td>RBF kernel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Section 5.2 were followed. In order to make the results for different algorithms and number of training points comparable, it was assured all the models were trained using the same points and increasingly larger subsets of the training set and the test set was kept constant.

As shown in Figs. 7 and 8, good overall predictions can be obtained with a reduced number of training points. However, the curves have not yet converged: using a higher number of training points will still translate in a reduction of the generalization error. Moreover, Gaussian processes were able to achieve the best performances for very small number of data points, whereas artificial neural networks outperformed all the algorithms for increasing number of data points (Bessa et al., 2017; Neal, 2012; Williams, 1998). Regarding the tree-ensemble methods, XGBoost demonstrated to be a highly competitive algorithm, whereas random forests have not achieved satisfactory performances.

5.4. Representations of the design space

Based on the previous results, 1000 training points are considered to be sufficient to accurately capture the overall trend of the design space. This number of data points is also considered to be reasonable for the application of the present framework to more expensive and accurate methodologies for open-hole strength prediction (or related problems). Therefore, several models were trained for this number of training points following the model selection procedure presented in Section 5.2. Afterwards, about 10000 new points were collected and the predictions of the models compared with the ground truth. As shown in Figs. 9 and 10, the obtained relative errors are highly concentrated between the ±10% range.

In the remainder of this section, representations of the design space are shown to assess the ability of the trained models to capture the intricacies of the open-hole strength function as well as its continuity. Here, a sensitivity analysis, where a single parameter is varied ranging from its allowed minimum and maximum, while the remaining six are kept constant, is performed. The results are shown for Legacy Quad and DD laminates in Fig. 11a and b, respectively.

From Fig. 11, it can be concluded that, in fact, all four algorithms are capable of fitting the analytical model with very good accuracy. However, as expected, and given its discontinuous nature, the tree-based models provide a less smooth response. The Gaussian Processes and Artificial Neural Networks provide the smoother prediction curves due to their continuous nature as well as leading to the lower prediction errors and are, therefore, more appropriate to address the present problem.

6. Generation of B-basis allowables using machine learning

In this section, the generation of design allowables, the B-basis allowables (Handbook, 2002), based on machine learning models is explored. The goal is to verify if the machine learning models are able to capture, not only the overall 7D design space, but also the variability associated with a given design point, consequence of the uncertainty related to the material properties and geometrical descriptors.

By taking the variability of the input parameters (material and geometrical) into account, the uncertainty of the input parameters can be propagated to the notched strength, i.e. a statistical distribution of the notched strength can be obtained, which can then be used to compute the statistical design allowables, namely the B-basis allowable. The B-basis allowable is the standard design allowable used in the aeronautical industry for fail safe structures (Handbook, 2002; Spendley, 2012). It is defined as the 95% lower confidence bound on the tenth percentile of a
specified population of measurements. It is a conservative allowable that ensures with 95% confidence that 90% of the population will have a given property, e.g. strength, higher than the B-basis allowable. Vallmajó et al. (2019) described two methodologies to obtain the B-basis: the CMH-17 approach and a Monte Carlo based approach. The first, is the methodology proposed by the Composite Materials Handbook (Handbook, 2002) and is generally employed for small populations, such as the ones typically obtained experimentally. The second, is a computerized mathematical technique that allows, by repeated, nearly infinite, random sampling of the input parameters, obtaining the distribution of the population of results. Here, given the efficiency of both the baseline analytical model and of the trained machine learning models, the Monte Carlo based approach is used, as described below:

1. 10000 input material ($E_1$, $X_T$, $G_0$) and geometrical ($D$ and $W$) parameters are generated following a given statistical distribution (normal and uniform distributions, respectively). The lay-up was considered to be fixed, i.e. no laminate rotation/misalignment during the cutting procedure was considered. The baseline analytical model allows accounting for lay-up variability, however the machine learning models do not, since any rotation of the laminate (or any single ply) will yield invalid lay-ups.

2. The distribution of the notched strength is obtained by calculating the notched strength (using the analytical model described in Section 2 and the trained machine learning models described in Section 5).

3. The 10th percentile of the distribution of the notched strengths is obtained and approximated to the B-basis allowable (10th percentile $\approx$ B-basis Vallmajó et al., 2019).

In a representative example, which allows the comparison of the distribution of the notched strengths, B-basis values and mean values obtained using the baseline analytical model and the trained machine learning models is shown below. An IM7/8552 [90/0/-45/45]$_s$ quasi-isotropic lay-up and a [45/-45/20/-20] DD lay-up, and specimens with hole diameter-to-width ratios of $3 < W/D < 8$ and hole diameters of 2, 4, 6, 8 and 10 mm were considered (Camanho et al., 2007b; Vallmajó et al., 2019). The material parameters were considered to

| Table 3

<table>
<thead>
<tr>
<th>Properties of the IM7/8552 material system (Camanho et al., 2007b) and variability of the geometric parameters.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_1$ [GPa]</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Standard deviation</td>
</tr>
<tr>
<td>$D$ [mm]</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Tolerance</td>
</tr>
</tbody>
</table>
follow a normal distribution with means and standard deviation shown in Table 3 and the geometrical descriptors were considered to follow an uniform distribution with a tolerance of ±2%, related to tolerances allowed during specimen cutting.

The distributions of the notched strength, B-basis and mean value provided by the analytical model and the trained machine learning models are presented in Figs. 12 and 13 for a specimen with $D = 6 \text{ mm}$ and $W = 36 \text{ mm}$.

As shown in Figs. 12–15, for both the Legacy Quad and DD lay-up, the XGBoost, Artificial Neural Networks and Gaussian Processes models provide very accurate strength distributions compared to those obtained using the analytical model, allowing an accurate determination of design allowables related to material and geometrical variability. It should be noted that for a different number of training samples, some ML algorithms may lead to more accurate distributions than others. Even though no clear distinction between the accuracy of the models on the determination of B-basis allowables was found, the Gaussian Processes model has the advantage of having a continuous nature and a fast training process, in contrast to the XGBoost and Artificial Neural Network models, respectively. The Gaussian Processes models having i) a fast learning process, ii) less hyperparameters to optimize, iii) having provided a continuous accurate representation of the design space, iv) the lowest relative error and v) good B-basis allowable predictions, and vi) being more powerful model for small training sets, were considered the most interesting and convenient Machine Learning models for this.
type of analysis. Particularly the fact that Gaussian Process models provide better performances for smaller training sets can be an extremely important characteristic when training models based on numerical data (e.g. from finite element simulations), whose generation is much more time-consuming.

To further test the Gaussian processes models, a design chart, where the diameter-to-width ratio, $D/W$, is varied between the bounds (1/8 and 1/3) for three hole diameters ($D = 2$, 4, 10 mm) was generated (Figs. 16 and 17). The mean and B-basis values for each geometry are presented. The Gaussian processes models provide very accurate predictions of both the mean values and B-basis and correctly capture the effect of varying the hole diameter as well as the diameter-to-width ratio ($D/W$).

Note that the curves diverge slightly near the boundaries (for $D/W = 1/8$ and $D/W = 1/3$). This can be explained by the fact that when variability is considered, the generated input parameters can be outside of the training space (see bounds used to train the machine learning models in Table 1). This causes some loss in accuracy in the machine learning algorithms, as they are known to be very powerful as interpolation tools, i.e. working inside the bounds in which they were established, and very ineffective when extrapolating outside the bounds of the design space used for training. This drawback can be circumvented by increasing the design space and retraining the models, which would be trivial using the analytical model, but may otherwise be impossible when dealing with more computationally expensive models or experimental data. This further highlights the need to accurately define the design space used to train the machine learning models and the limitations of such techniques.
Details about the text content are not available.
References