Aerospace Science and Technology 148 (2024) 109118

Contents lists available at ScienceDirect

Aerospace Science and Technology

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

A ML strategy for the identification of optimal LPT design region and related blade shape

Daniele Petronio *, Pietro Paliotta, Davide Lengani, Daniele Simoni

University of Genova, Department of Mechanical, Energy, Management and Transport Engineering (DIME), Genova, Italy

ARTICLE INFO

Communicated by Christian Circi

Keywords:
Machine-learning
POD
LPT
Gaussian-process

ABSTRACT

This work presents a machine-learning (ML) strategy for the identification of the design region that guarantees minimum losses for Low Pressure Turbine (LPT) blades, allowing the definition of the optimal blade shape. The data-driven procedure is twofold. Firstly, an advanced loss-correlation model (M1) that describes the LPT efficiency as a function of the main flow and geometrical parameters, also accounting for unsteady effects, has been trained from a numerical database. Then, a second model (M2) has been tuned to interpret the corresponding blade geometries. Proper Orthogonal Decomposition (POD) has been applied to formally decompose the blade shape into modes and coefficients. The modes provide basis functions, while the coefficients give the weights that, depending on the combination of the design parameters, define the blade shape. Gaussian Process (GP) and Cross-Validation techniques have been used for tuning both M1 and M2. Once properly tuned, the overall procedure provides the loss-correlation model (M1) for the identification of the design region that is expected to minimize losses, and the geometrical model (M2) for a quick definition of the corresponding optimal blade shape. The procedure can be extended to other engineering applications where own efficiency and geometrical data are available.

1. Introduction

Machine-Learning (ML) techniques offer new potential opportunities for the analysis of complex systems, providing the unique advantage to learn the system response directly from data/experiments. The field of ML covers several relevant applications, from regression and dimensionality-reduction, which are of primary interest for the present work, to classification and clustering [1].

For what concerns regression, i.e., the problem of learning from data an input-output relationship in a continuous space, a large variety of ML tools are available today. Linear models are suitable for low-complexity applications, where it is easy to define a priori the structure of the model (e.g., linear, polynomial, exponential, etc.). Regularization methods, like Lasso [2], Ridge [3] and Elastic Net [4], can be considered for more complex systems. Here, sparsity promotion and automatic predictor selection help the identification of the right compromise between model accuracy and generalizability. Lengani et al. [5] used Elastic Net to tune the Reynolds stresses into a Laminar Kinetic Energy model, with the aim of improving closure schemes for RANS simulations. Brunton [6] instead proposed sparsity methods for the description of generic dynamical systems, discovering the governing equations directly from data.

However, non-linear models of various shapes and sizes should be considered when the complexity of the problem increases. Gaussian Processes [7], Artificial Neural Networks (ANN) [8], Genetic Programming algorithms [9] are advanced tools able to learn complex functions, also accomplishing with big databases and with high-dimensional spaces. An example is provided in Dominique et al. [10] where a network for the prediction of the aeroacustic noise in turbulent boundary layers was tuned. The main challenge here is to balance model complexity versus available data. Several criteria have been proposed for this purpose. While Bayesian approaches are often used to this end [11,12], the Cross-Validation (CV) methods are probably the most popular strategies for model selection [8]. They are based on the idea of splitting the available cloud of data in training and validation sets, providing different possibilities to directly estimate accuracy and generalizability and to balance them. Picard and Cook [13] and Dreyfus [8] report a detailed description of CV methods. Dellacasagrande et al. [14] recently proposed a model selection strategy that integrates the Leave-One-Out

* Corresponding author.
E-mail address: daniele.petronio@edu.unige.it (D. Petronio).

https://doi.org/10.1016/j.ast.2024.109118
Received 29 November 2023; Received in revised form 23 February 2024; Accepted 4 April 2024
Available online 10 April 2024

Cross-Validation (LOO-CV) criteria and the data variance for promoting generalizable models.

Dimensionality-reduction constitutes another important field of application of ML algorithms. Here, the aim is to find compressed data representation, allowing the identification of the main features embedded in the data and defining low order spaces where data can be conveniently observed. Among different techniques, the Proper Orthogonal Decomposition (POD), or Principal Component Analysis (PCA), is probably the most widely used data-driven tool for dimensionality reduction. It is usually applied in the fluid dynamic community thanks to its capability to separate spatial from temporal information, providing the main flow structures as POD modes and the related dynamics by the POD eigenvectors (or coefficients). Many works related to POD can be found in the fluid dynamic community (e.g., [15,16]), and also in other applications (e.g., [17,18]). Indeed, independently from the specific case, POD can be generally applied to high-dimensional data, of whatever nature they are, to optimally identify low order representations of the cloud of data, as compact as possible with respect to the variance of the system. For example, POD can be used to decompose geometrical surfaces, as shown in [19], where orthogonal basis functions describing a series of supercritical airfoils have been identified. Toal et al. [20] instead used POD to filter out badly performing geometries from an optimization process. Indeed, as stated by Li et al. [21], POD can be effectively used both for modal parameterization and filtering of surfaces. Duan et al. [22] used the POD-based approach for optimizing a compressor rotor blade (the NASA Rotor 37), while Luo et al. [23] proposed a Gappy-POD method for improving the performance of both turbine and compressor blades. The aim is to reduce the dimensionality of the design space with respect to the conventional geometric methods (e.g., Bezier and B-spline [24] curves, Free-Form Deformation [25]), also excluding abnormal solutions. In this context, a lot of works adopted the UIUC (University of Illinois at Urbana-Champaign) open database (e.g., [26]).

In the turbomachinery field, modern ML routines offer great possibilities. For example, new accurate efficiency correlations can be learned from experimental and numerical data, and they can drive the design of turbomachinery components towards optimized solutions. The advantages are twofold. On one hand, the capability to handle large databases allows the definition of new correlations over wider design spaces, learning the effects of several parameters at the same time. For example, the unsteady effects due to the aerodynamic interaction between stationary and rotating blade rows may be accurately described (see [27] for example), improving previous correlations (e.g., [28–30]). On the other hand, the dynamic nature of the ML tools provides the possibility to update and re-calibrate the models every time new data are available.

In the present work, a ML strategy has been developed to identify the optimal LPT design region and the corresponding blade shape. The procedure, applied to a numerical database (see Acknowledgment) that provides a great amount of LPT cascade efficiency data, consists of two fundamental steps. Initially, an advanced loss-correlation model (M1) has been tuned by means of GP following the learning strategy proposed in [14]. The new loss-correlation describes the LPT cascade efficiency as a function of two sets of parameters: a set of flow parameters defining the unsteady operating conditions of the cascade (i.e., the Reynolds number, the flow coefficient, the reduced frequency and the axial gap between stator and rotor), and a set of geometrical parameters which essentially define the cascade geometries (i.e., the Zweifel number, the velocity ratio, the diffusion rate and the maximum thickness with respect to the chord). Subsequently, a POD-based model (M2) has been tuned to interpret the blade geometry (camber and thickness) as a function of the aforementioned parameters. The blade geometries have been decomposed into POD modes (numerically derived basis shape functions) and POD coefficients, also offering the opportunity to reduce data-dimensionality. POD coefficients have then been related to the set of geometrical parameters entering in the loss-correlation, adopting again GP. Once properly tuned, the overall procedure involves the loss-correlation model (M1) for the identification of the design region that is expected to minimize losses, and the POD-based model (M2) for a quick definition of the corresponding optimal blade shape. The procedure can be generally extended to other engineering applications, where own data and geometries are available.

2. Training database

A numerical database constitutes the foundation for applying the ML tools for the dual purpose of this work: the definition of an advanced loss-correlation model (M1), also accounting for the unsteady effects, aimed at identifying the optimal LPT cascade design region, and the definition of a POD-based model (M2) for identifying the corresponding optimal blade shape. The database consists of an ensemble of $N = 500$ two-dimensional URANS simulations, computed on $M = 25$ different LPT cascades for different inflow conditions. Each cascade has been simulated for meanly 20 different combinations of the main influencing parameters like the Reynolds number, the reduced frequency and the flow coefficient. The simulations indeed reproduce the aerodynamic interaction between the stator and the rotor by means of moving bars passing in front of the tested cascade, as for the sketch of Fig. 1. This configuration is usually adopted in the experiments for simulating the unsteady flow characterizing the real engine operation (see [31,32] for instance). To obtain an accurate resolution of the flow field, and considering the global computational time, unstructured meshes constituted of about 100k polyhedral cells have been used. For all the calculations, prismatic refined layers have been adopted in the boundary layer and in the wake regions to guarantee an y^+ value smaller than one. The direction of the refinement in the wake region has been automatically determined by a procedure that estimates the flow exit angle from literature available correlations. The automatic generation of the mesh has been implemented in STAR-CCM+ by a dedicated Java script. URANS equations are solved adopting a second-order scheme for both the convective and the diffusive terms. The SST k-ω turbulence model coupled with the γ-Re_θ transition model has been used to capture the transition process into the blade boundary layers. The (profile) losses, calculated as for Eq. (1), have been considered as the objective function describing the cascade aerodynamic performances.

$$\omega = \frac{\overline{p_{t1}} - \overline{p_{t2}}}{\overline{p_{t1}} - p_2} \tag{1}$$

The performances of each cascade have been evaluated for different inflow conditions, varying the inlet flow velocity, the bar pitch, the peripheral velocity and the axial distance of the bars with respect to the cascade leading edge plane. The different cascades instead differ for their geometry. The main parameters changing among the cascades are the blade inlet and outlet geometrical angles, the chord, the pitch, the maximum thickness and the curvature of the blade. The main variables are indicated in Fig. 1. A SOBOL sequence has been adopted to efficiently sample the design space around the main characteristics defining the T106A cascade, that is a typical LPT cascade available in the literature (see [31,33]). Fig. 2 reports, on the left, the main T106A parameters and shows, on the right, the distribution of the different samples of the present database into the subspace identified by the inlet angle α_1, the outlet angle α_2 and the cascade solidity $\sigma_x = C_x/S$. It is worth noticing that the axes in this plot, as for all the plots in the following figures, are centered and normalized, for confidentiality reason. Fig. 3 instead presents four different normalized blades (B1, B2, B3, B4, subplots a-d), that will be used as reference for the discussion of the procedure proposed in this work. The camber and thickness distributions are reported in the subplots (e) and (f). The B1 is representative of a standard-loaded LPT blade. The B2 is instead a thick blade designed for supporting high mechanical stresses. The B3 and B4 are highly-deflected profiles, characterized by different loading distributions, with the B3 more aft-loaded with respect to the B4 (see [34]

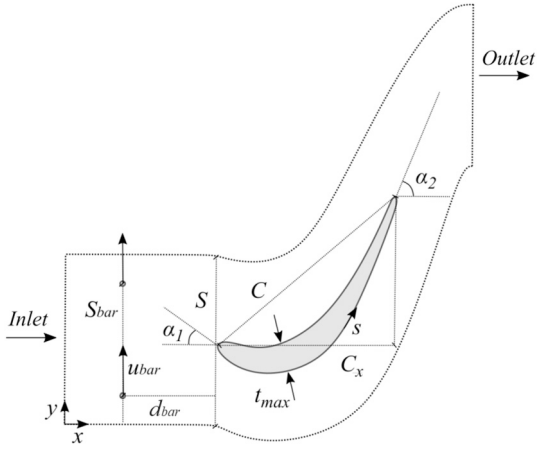

Fig. 1. Sketch of the numerical setup and main variables of the LPT cascade.

for details on the effect due to the loading distribution on the profile losses).

3. Data analysis

Fig. 4 presents the workflow here proposed. The loss-correlation model (M1) is educated by means of two sets of non-dimensional parameters: a set of flow parameters defining the unsteady operating conditions (highlighted in blue) and a set of geometrical parameters describing the cascade geometries (highlighted in red). The flow parameters are those that characterize the unsteady operating conditions of the cascade: the Reynolds number Re, the reduced frequency f^+, the flow coefficient φ and the axial gap AG. The geometrical ones (in red) instead condense the variations induced on the blade loading by the different geometries characterizing the tested cascades. Specifically, the velocity ratio VR, the Zweifel number Zw, the diffusion rate Ds' and the maximum thickness with respect to the chord t_{max}/C have been used. The definition of the parameters is reported in Table 1. As highlighted in the sketch of Fig. 4, they define the main features of the blade loading distributions, being the VR and Zw number related to the loading (dashed area on the plot), and the Ds' strictly linked to the adverse pressure gradient on the blade suction side. Both these parameters are well known to affect the boundary layer evolution, thus the cascade performance (see [35] for instance). The t_{max}/C has instead been chosen to condense different blade thicknesses. On the other hand, the POD-based model (M2) for blade shape reconstruction takes as inputs the subset of geometrical parameters only, and learns the relation between them and the POD coefficients. They, once combined with the modes, provide the blade geometry. For the sake of clarity, the inlet and outlet flow angles (α_1 and α_2) have been here used in place of the Zweifel number and the Velocity Ratio (Zw and VR) in order to separate the effect due to the pitch to chord ratio from the definition of the blade shape, thus increasing the M2 accuracy. It is worth noticing that the library of influencing parameters can be modified and/or augmented by others, depending on availability of data.

3.1. M1: loss-correlation model

M1 is aimed at defining the loss-correlation model, thus a generic function $\omega = f(\underline{x})$ that relates the loss coefficient to the flow and geometrical parameters:

$$\omega_{M1} = f(\underline{x}) = f\left(Re,\ f^+,\ \varphi,\ AG,\ Zw,\ Ds',\ VR,\ t_{max}/C\right) \tag{2}$$

Different regression methods (Least-Square Method LSM, Lasso, Ridge and Elastic Net regressions, Gaussian Process, Neural Networks) may be considered to build the model. For example, the Lasso regression [2]

Table 1
Definition of the flow and geometrical parameters.

Parameter	Definition
Reynolds number	$Re = \frac{c_2 C_x}{\nu}$
Reduced frequency	$f^+ = \left(\frac{u_{bar}}{S_{bar}}\right)\left(\frac{C}{c_2}\right)$
Flow coefficient	$\varphi = \frac{c_x}{u_{bar}}$
Velocity ratio	$VR = \frac{M_2}{M_1}$
Zweifel number	$Zw = \frac{2\cos^2(\alpha_2)}{\sigma_x}$
Diffusion rate	$Ds' = \left(\frac{M_{max} - M_2}{M_2}\right)/\left(\frac{s_{decel}}{s_{tot}}\right)$
Axial gap	$AG = \frac{d_{bar}}{C_x}$
Max thickness vs. chord	t_{max}/C

provides a linear relation between some predictors $\underline{\phi}(\underline{x})$ and some coefficients $\underline{\beta}$, where the coefficients are selected solving the minimization problem of the residuals constrained by the ℓ_1-norm of the coefficients (Eq. (3)).

$$f(\underline{x}) = \Phi(\underline{x}) \cdot \underline{\beta}\,, \quad \underline{\beta} = \min_{\underline{\beta}} \left(||\underline{y} - \Phi(X)\underline{\beta}|| + \lambda_1 |\underline{\beta}| \right) \tag{3}$$

For clarity, $\Phi(X)$ is the matrix collecting the predictors evaluated for the input parameters X, while $\underline{y}$ is the calculated profile loss coefficient for available data. Instead, when a Gaussian Process is used, $f(\underline{x})$ takes the structural form of Eq. (4), where the key element is the kernel function $k(x, x')$ expressing the non-linear relation between the inputs, and σ_n^2 is a free parameter (noise variance).

$$f(\underline{x}) = [k(\underline{x}, X)]\left[k(X, X) + \sigma_n^2 I\right]^{-1} [\underline{y}] \tag{4}$$

In this work, for what concerns M1, a GP with an ard-squared exponential (SE) kernel and two additive physical-informed predictors able to mimic the Reynolds number effects for both the laminar ($Re^{-0.5}$) and the turbulent ($Re^{-0.2}$) boundary layer growth (see also [27] for instance) has been considered. Here, the form of $f(\underline{x})$ is like Eq. (5), while the SE kernel is reported in Eq. (6). The model has different free (hyper)-parameters: the length scales $\underline{l}$, the signal variance σ_f^2, the noise variance σ_n^2 and the coefficients $\underline{\beta}$. The combination of these parameters leads the tuning of the model toward the optimal compromise between accuracy and generalizability.

$$f(\underline{x}) = [k(\underline{x}, X)]\left[k(X, X) + \sigma_n^2 I\right]^{-1} [\underline{\omega}] + [\underline{\beta}]\left[Re^{-0.2},\ Re^{-0.5}\right] \tag{5}$$

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{1}{2}\left[(]\underline{x} - \underline{x}'\right]^T [\underline{l}]\left[\underline{x} - \underline{x}'\right] \right) \tag{6}$$

The learning strategy described in [14] has been used to obtain the best relationship $f(\underline{x})$, among different regression methods and among different definition of the (hyper) parameters, by optimizing three indicators of the model performance:

- The leave-one-out mean error E_{LOO};
- The uniformity of the leverage distribution μ;
- The variance ratio between extrapolation and interpolation ϵ.

The leave-one-out mean error E_{LOO} is defined as the mean of the (absolute) errors computed in predicting the k^{th} observation when the k^{th} observation does not participate at the training of the model (Eq. (7)). It is representative of the expected accuracy of the model in making predictions. Lower values are explicative of higher model performance.

The indicator μ is instead indicative of the uniformity of the leverages h_{kk}, which describe the influence of each k^{th} observation on the model, being defined as the orthogonal projection of the Jacobian matrix Z of the model (see [8,14] for details). Values of μ tending to 1 should be preferred being representative of a more homogeneous influence of the different observations.

The variance ratio ϵ is defined as the mean variance predicted in a set of Q test points outside of the training set domain and the noise-

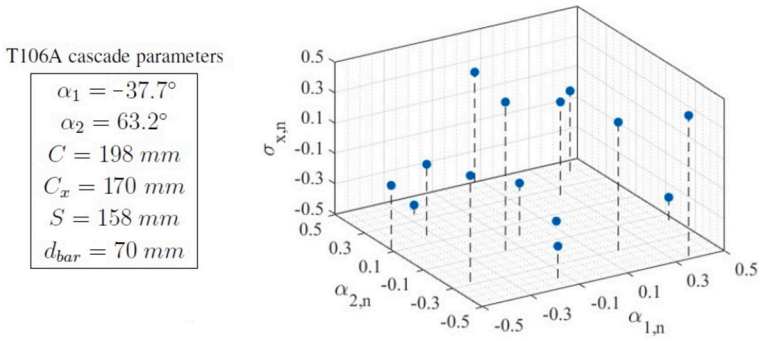

Fig. 2. Main characteristics of the T106A cascade, on the left, and distribution of the different samples of the present database in the subspace $\{\alpha_{1,n}, \alpha_{2,n}, \sigma_{x,n}\}$, on the right.

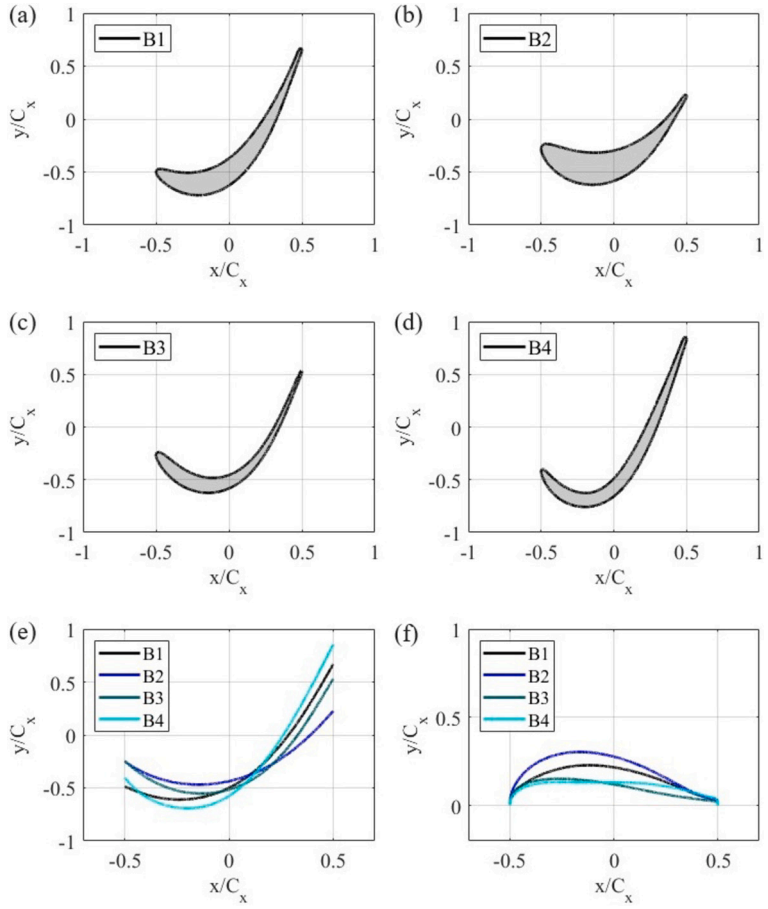

Fig. 3. Four blades from the database (upper subplots: *a-d*) and their decomposition into camber and thickness distributions (bottom subplots: *e-f*).

variance, practically giving a measure of the ratio between the variance (i.e., uncertainty) in extrapolation (outside of the dataset domain) and the variance in interpolation (inside the dataset domain). High values of ϵ indicate models that are prone to overfitting, thus contained values further enhance the selection of parsimonious models. For clarity, in this work the Q test points are selected varying each input parameter from the center to the 20% outside of the hypercube comprising the dataset matrix.

$$E_{LOO} = \frac{1}{N}\sum_{k=1}^{N}\left|\frac{y_k - f^{-k}(\underline{x}_k)}{y_k}\right| \tag{7}$$

$$\mu = \frac{1}{N}\sum_{k=1}^{N}\sqrt{\frac{N}{P}h_{kk}}\ , \quad h_{kk} = diag\left(Z(Z^TZ)^{-1}Z^T\right) \tag{8}$$

$$\epsilon = \frac{1}{Q}\sum_{q=1}^{Q}\frac{\sigma^2_{f(\underline{x}_q)}}{\sigma^2_n} \tag{9}$$

3.2. M2: POD-based model for blade geometry

For the definition of M2 the different blade geometries have been parametrized with their camber and thickness distributions, as usual in the blade design practice. Then, two snapshot matrices, $[C]$ for cambers (Eq. (10)) and $[T]$ for thicknesses (Eq. (11)), have been created. Each column vector collects the y/C_x values identifying both the camber and the thickness lines of the B^{th} blade. Distributions are sampled with n points along the x/C_x axis (thus, for the sake of clarity, n here indicates the number of points discretizing the blade camber and thickness).

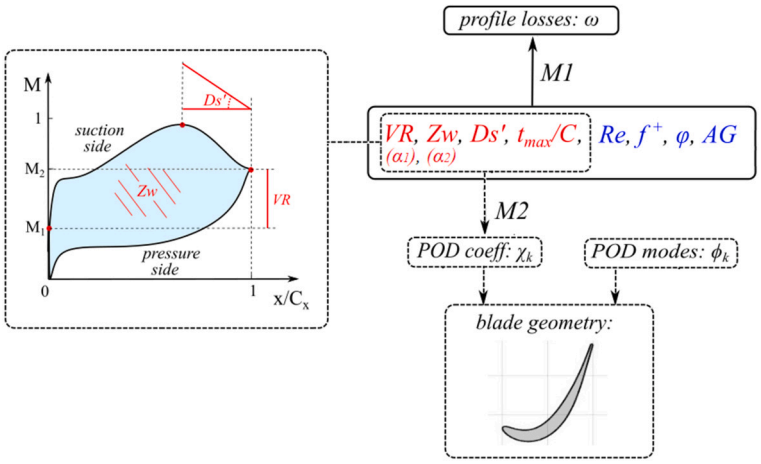

Fig. 4. Sketch of the workflow with the loss-correlation (M1) and the POD-based geometry model (M2). (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

$$[C] = \begin{bmatrix} \frac{y}{C_x}\left(\left(\frac{x}{C_x}\right)_1, B_1\right) & \frac{y}{C_x}\left(\left(\frac{x}{C_x}\right)_1, B_2\right) & \dots & \frac{y}{C_x}\left(\left(\frac{x}{C_x}\right)_1, B_M\right) \\ \frac{y}{C_x}\left(\left(\frac{x}{C_x}\right)_2, B_1\right) & \frac{y}{C_x}\left(\left(\frac{x}{C_x}\right)_2, B_2\right) & \dots & \frac{y}{C_x}\left(\left(\frac{x}{C_x}\right)_2, B_M\right) \\ \vdots & \vdots & \dots & \vdots \\ \frac{y}{C_x}\left(\left(\frac{x}{C_x}\right)_n, B_1\right) & \frac{y}{C_x}\left(\left(\frac{x}{C_x}\right)_n, B_2\right) & \dots & \frac{y}{C_x}\left(\left(\frac{x}{C_x}\right)_n, B_M\right) \end{bmatrix} \tag{10}$$

$$[T] = \begin{bmatrix} \frac{t}{C_x}\left(\left(\frac{x}{C_x}\right)_1, B_1\right) & \frac{t}{C_x}\left(\left(\frac{x}{C_x}\right)_1, B_2\right) & \dots & \frac{t}{C_x}\left(\left(\frac{x}{C_x}\right)_1, B_M\right) \\ \frac{t}{C_x}\left(\left(\frac{x}{C_x}\right)_2, B_1\right) & \frac{t}{C_x}\left(\left(\frac{x}{C_x}\right)_2, B_2\right) & \dots & \frac{t}{C_x}\left(\left(\frac{x}{C_x}\right)_2, B_M\right) \\ \vdots & \vdots & \dots & \vdots \\ \frac{t}{C_x}\left(\left(\frac{x}{C_x}\right)_n, B_1\right) & \frac{t}{C_x}\left(\left(\frac{x}{C_x}\right)_n, B_2\right) & \dots & \frac{t}{C_x}\left(\left(\frac{x}{C_x}\right)_n, B_M\right) \end{bmatrix} \tag{11}$$

POD has been applied to decompose both the snapshot matrices into POD modes and POD coefficients. For clarity, the mean (baseline) camber and thickness have been subtracted from the dataset, and the two snapshot matrices have been factorized into related modes φ, eigenvalues λ and coefficients χ, as for Eqs. (12) and (13). The decomposition can be equivalently computed either by economy-sized Singular Value Decomposition [36] or by obtaining λ and χ as eigenvalues and eigenvectors of the cross-correlation matrix of the snapshot matrices and φ by projection (see [37] for further details).

$$[C] - [c_{baseline}] = [\phi]_c [\lambda]_c^{1/2} [\chi]_c^T \tag{12}$$

$$[T] - [t_{baseline}] = [\phi]_t [\lambda]_t^{1/2} [\chi]_t^T \tag{13}$$

The modes $[\phi]_c$ and $[\phi]_t$ are spatial basis function, i.e., numerically derived blade shape functions depending on x/C_x. The eigenvalues λ directly rank the different modes, while the coefficients χ are scaling factors weighting each mode across the design space. According to the POD properties, a data dimensionality reduction can also be obtained truncating the matrices at order $r < min(n, M)$ (see [37] for instance). For each r, the low order representation is the most compact linear representation of the dataset. The reduced order r is chosen separately for cambers and thicknesses (r_c and r_t).

Then, to allow the reconstruction of the optimal blade shape adopting the set of flow and geometrical parameters identified by the loss-correlation model (M1), the POD coefficients have to be related to the set of geometrical parameters. Here, for clarity, the camber lines have been assumed independent of the maximum thickness, while thickness distributions independent of the flow angles (see Eqs. (14) and (15)). The learning strategy [14] has again been used for tuning the functions f_i. f_j:

$$\chi_{i,C} = f_i\left(\alpha_1, \alpha_2, Ds', C/C_x\right), \quad i = 1, 2, \dots, r_c \tag{14}$$

$$\chi_{j,T} = f_j\left(Ds', t_{max}/C_x\right), \quad j = 1, 2, \dots, r_t \tag{15}$$

Different regression methods have been tested again in terms of the triplet (E_{LOO}, μ and ϵ). Here, differently from M1, GPs with dot-product kernel (Eq. (16)) have also been considered.

$$K(x, x') = \underline{\phi}(\underline{x})^T \cdot \underline{\phi}(\underline{x}) \tag{16}$$

Moreover, the functions to be tuned here are more than one (r_c and r_t), and the output is multidimensional (i.e., the camber and the thickness distribution, and finally the blade shape). Thus, the triplet of indicators for model selection has been used as follows:

- The leave-one-out mean error E_{LOO}, as the mean between the errors on the y/C_x coordinates of the blade shape predicted by the model, when the blade itself does not participate to the tuning of the model.
- The uniformity of the leverage distribution μ, as the mean of the μ values of each $\chi_{c,i}$, $\chi_{t,j}$ weighted on the relative eigenvalues.
- The variance ratio between extrapolation and interpolation ϵ, as the mean of the single variance ratios calculated on the different y/C_x coordinates of the blade shape.

4. Results

Results here reported are aimed to show the capabilities of the introduced ML strategy. In the defined workflow: M1 is able to identify the design region that guarantees minimum losses, in terms of best combinations of the design parameters, and successively M2 can be interrogated in this region to generate the optimal blade shape. For what concern M1, the results of the model selection are shown in terms of the triplet of indicators used for tuning to discuss the obtained model performance, while for confidentiality the tuned correlation and the related response surface cannot be shown. Instead, for M2, the capabilities of the POD-based model in providing LPT blade shapes are presented in more details.

4.1. M1: model selection

In order to discuss the performance of the loss-correlation model, Fig. 5 reports the triplet {E_{LOO}, μ and ϵ} for different candidates. The plot in the background presents the E_{LOO}-μ chart. The blue squares represent the performance of linear models tuned by Least-Square Method (LSM), while the orange ones represent models tuned with the Lasso algorithm. The green points are different GPs, the yellow ones GPs with squared-exponential (SE) kernel and different values of data-variance

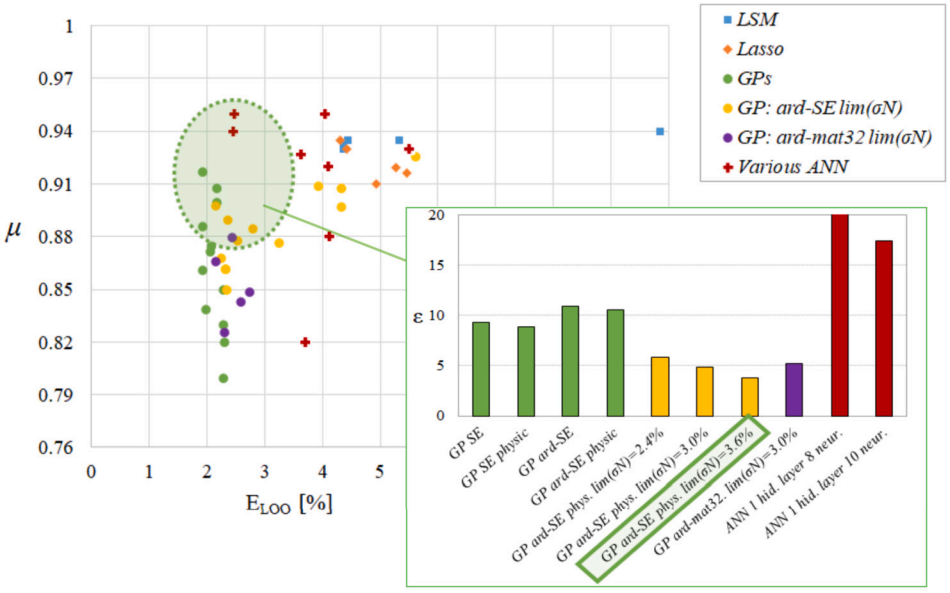

Fig. 5. E_{LOO}-μ chart (background) and ϵ histogram (foreground) for the selection of the loss-correlation model.

(σ_n^2), and the violet points are GPs with matern32 (SE) kernel. A Quasi-Newton gradient-based optimization algorithm has been used to learn the GP length scales and hyperparameters. Different initial values have been tested to verify the convergence of the optimization process to a global minima. In addition, the red crosses represent the performance obtained with various sizes feed-forward neural-networks (ANN). Particularly, networks with 1 hidden layer and different number of neurons (from 5 to 15) have been tested, considering both hyperbolic and radial-basis activation functions. The plot allows the identification of the most-promising candidate models (i.e., lower E_{LOO} and high μ, green area), that are further evaluated by means of the variance ratio ϵ. The box in the foreground reports the ϵ histogram for the models lying inside the green area. The most generalizable model characterized by the minimum variance in extrapolation (i.e., lowest ϵ) is selected at the end of the procedure. Therefore, the present procedure identifies the best loss-correlation model as a GP with ard-SE kernel, physical-informed predictors and noise-variance set to 3.60% with respect to the mean efficiency (see Eq. (5)). It has a leave-one-out mean error of $E_{LOO} = 2.35\%$ (expected error in making predictions), a uniformity coefficient $\mu = 0.89$ and a variance ratio of $\epsilon = 3.75$.

4.2. M2: POD modes for blade camber and thickness

Fig. 6 presents the cumulative sum of the eigenvalues of the snapshot matrices defining the camber (on the left plot) and the thickness (on the right plot) distributions. The plots make evident that a reduced number of modes contains the most of the variance related to both the snapshot matrices. Particularly, 3 modes for the camber and 2 for the thickness account for the 98% of the total variance. Therefore, only 5 modes and related coefficients provide an accurate compressed representation of all the blades in the database (i.e., the blade geometry is defined by 5 degrees of freedom). Thus, in terms of degrees of freedom, the POD low-order representation is comparable to the analytical methods usually used for the preliminary design (e.g., parabolic camber, NACA distributions, Lemniscate functions). It is instead more compact with respect to the parametrization used for CFD-based optimization (e.g., Bezier, B-splines, Free-Form Deformation), since in these cases a larger number of parameters is involved. The capability of the POD low-order representation in reconstructing the blade shapes is shown in Fig. 7, where the low-order representation of the $B1$-$B4$ blades (red lines) is compared with the original profiles (black lines). The legend also reports the mean absolute error in the reconstruction of the blade points y/C_x for the different x/C_x positions. For clarity, a value of $E_{y/C_x} = 0.22\%$ means that for a blade with an axial chord of $C_x = 100$ mm, the localization of the blade points is subjected to a mean error of 0.22 mm in the y direction. The differences between the original and the reconstructed profiles are very low for all the four blades here presented, highlighting an almost perfect low-order reconstruction with 3 modes for the cambers and 2 for the thickness distributions.

The baseline camber and thickness distributions, the 3 leading modes for the camber and the 2 for the thickness are presented in their normalized form in Fig. 8. The subplots (a-d) are related to the camber, while (e-g) are representative of the thickness distributions. The baseline camber distribution, reported in the subplot (a) as the black line, represents the mean of the cambers of all the blades included in the database. It is a typical camber line of LPT profiles with mean inlet and outlet angles in the surround of $-35°$ and $65°$, respectively, and a mean deflection of about $100°$. Indeed, its distribution is very similar to a classical parabolic distribution with the same inlet and outlet angles (shown in red line in the plot), which is often used during turbine blade preliminary design practice (see [38,39] for instance). The first POD mode (b) is instead close to a straight line passing from zero. Thus, when added to the baseline, it acts modifying the baseline camber line in the fore and the aft part of the blade profile only (and only marginally in the middle), mostly influencing the blade inlet and outlet angles. Once properly weighted with the corresponding POD coefficient (depending on the geometrical parameters), it acts providing the proper stagger, inlet and outlet angles. The second mode (c) is a convex curve, almost symmetric with an almost constant curvature as well. It is expected to affect the blade deflection along its entire length. The third mode (d) finally captures an higher-order oscillation, that is able to differentiate the curvature of the camber between the front and the aft parts of the profile. On the other hand, the baseline thickness distribution (e) shows the mean thickness of all the blades of the database (black line). It is compared with two NACA distributions with 13% and 15% thick with respect to the chord (red and blue lines, respectively). As it can be seen from the plot, the mean thickness distribution characterizing the present database is more shifted toward the central part of the profile with respect to the NACA distributions. The first thickness POD mode (f) is instead a curve of almost constant curvature, except for the rear part of the blade where it approaches to zero close to the trailing edge. When added to the baseline, it increase/decrease the overall blade thickness with respect to the mean thickness distribution, except for the trailing edge region, being all the blades designed

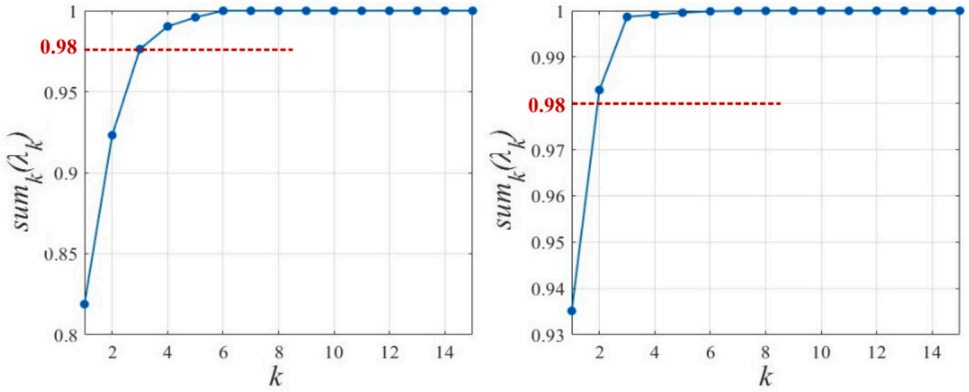

Fig. 6. Cumulative sum of the eigenvalues for the camber (on the left) and thickness (on the right) snapshot matrices.

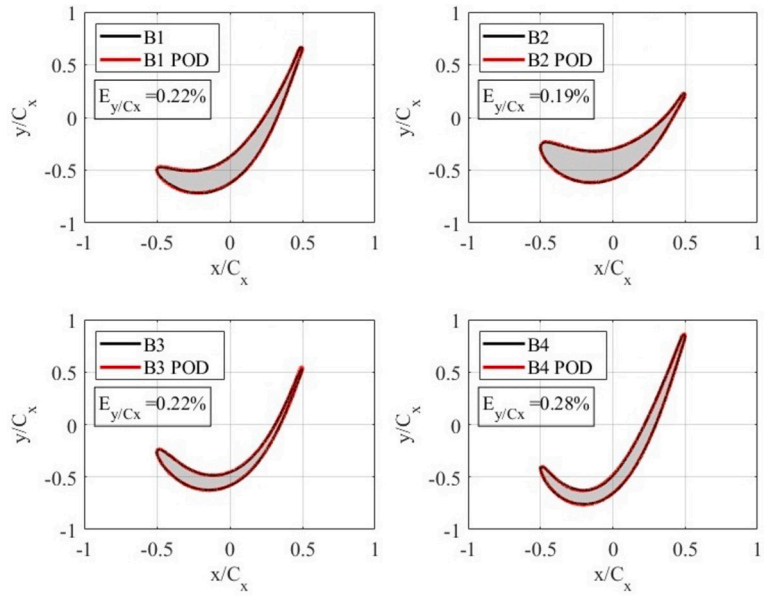

Fig. 7. POD reconstruction of the four blades $B1$-$B4$.

with a similar trailing edge. The second mode (g) provides an oscillation that redistribute the thickness between the front and the aft portion of the blade.

4.3. POD coefficients and relation with the geometrical parameters

With the aim of tuning Eqs. (14) and (15), Fig. 9 presents the variation of the POD coefficients with the geometrical parameters, reporting as examples the first (camber) POD coefficient χ_{1C} with respect to α_2 (on the left) and the first (thickness) POD coefficient χ_{1T} with respect to t_{max}/C_x (on the right). For what concerns the plot on the left, the POD coefficient χ_{1C} acts increasing/decreasing the outlet angle when it multiplies the mode if it is positive/negative, respectively (see also Fig. 8b). The POD coefficient χ_{1T} is instead responsible for thickening the profile when it is negative, and for thinning it when positive (see also Fig. 8f). It is worth pointing out for the sake of clarity that, although an overall trend can be observed (and highlighted in red in the plots), some scatter is present, and some points are with the same abscissa, because the blades differ for more than one parameter at a time.

For a complete description of the blade geometry, the 5 POD coefficients weighting the leading modes should be tuned according to Eqs. (14) and (15). The learning strategy proposed in [14] is again employed. Fig. 10 provides the E_{LOO}-μ chart obtained in this case. Different order LSM models (blue points) and different kernel GPs with SE (yellow points), matern32 (violet), and dot kernels (brown points) have been tested. As before, the most promising models (low E_{LOO} and high μ) are further inspected in terms of the variance ratio ϵ (showed in the box in the foreground). Although linear LSM models provide a good solution, probably because of the limited complexity of the geometrical database, slightly better performance can be reached using dot-kernel of the 3^{rd} order. It is characterized by a mean error $E_{LOO} = 1.24\%$, during the prediction of the blades when they are withdrawn from the model tuning, a leverage uniformity coefficient $\mu = 0.95\%$, and a variance ratio $\epsilon = 1.27$.

Fig. 11 shows the capability of the proposed POD-based geometry model in predicting the blade shapes. Here, the four blades B1-B4 are predicted when they do not participate to the model education (i.e., when the model coefficients are tuned using only the other blades of the database). For the sake of clarity, a leave-one-out cycle has been implemented (e.g., B1 is predicted when B1 is withdrawn from the model education, etc.). Predicted blade shapes are reported as the red profiles, while the original blades are the black ones. The legends report the mean percentage error E_{y/C_x}. The accuracy of the predictions is high for all the four blades here considered, as evidenced by the low values of the errors. The B1 shows an almost perfect superposition in the central part of the blade, with a slightly more curved nose as main difference with respect to the original profile. The B2 is well predicted, highlighting the capability of the model in reproducing blades of sig-

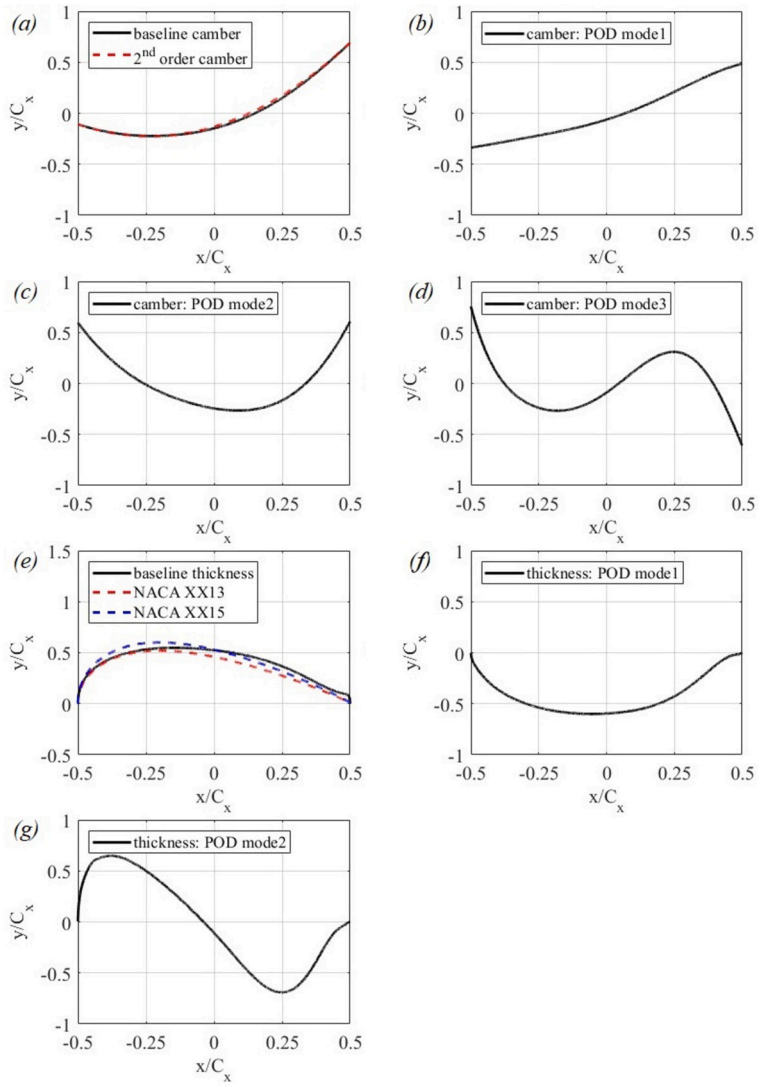

Fig. 8. Baseline and POD modes for camber (subplots *a-d*) and for thickness (subplots *e-g*) distributions.

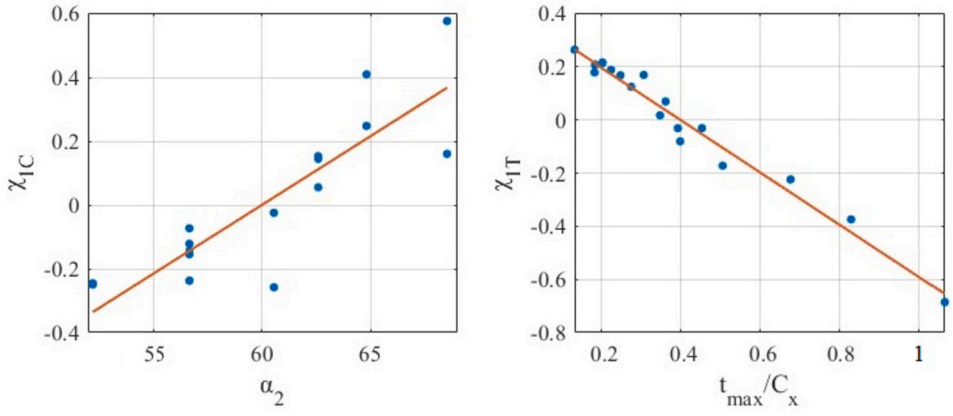

Fig. 9. First camber POD coefficient vs. outlet flow angle (on the left) and first thickness POD coefficient vs. maximum thickness with respect to the chord (on the right).

nificant different thickness distribution. The B3 and B4 are also well reproduced, although the errors are slightly higher (however limited to $< 1.63\%$). This reveals high accuracy also in the prediction of blade shapes that are characterized by different loading distributions (the B3 is more aft-loaded than the B4).

The workflow reported in Fig. 12 provides a direct view of the proposed procedure. Given a target flow condition identifying the LPT operation $\{Re, f^+, \varphi, AG\}_{tg}$, a request deflection VR_{tg} and a thickness constrain $(t_{max}/C)_{tg}$, M1 can be interrogated to provide the optimal values of the blade loading parameters $\{Zw, Ds'\}_{opt}$ minimizing losses.

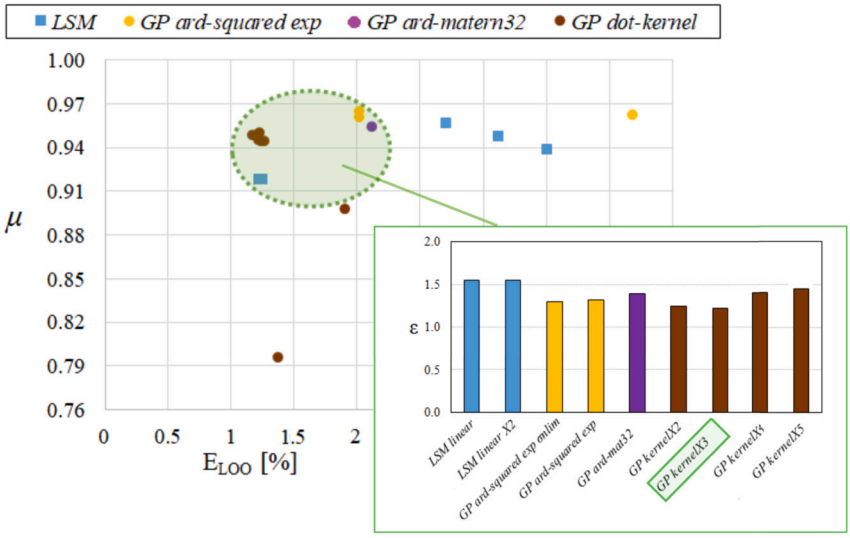

Fig. 10. E_{LOO}-μ chart (background) and ϵ histogram (foreground) for the selection of the POD-based geometry model.

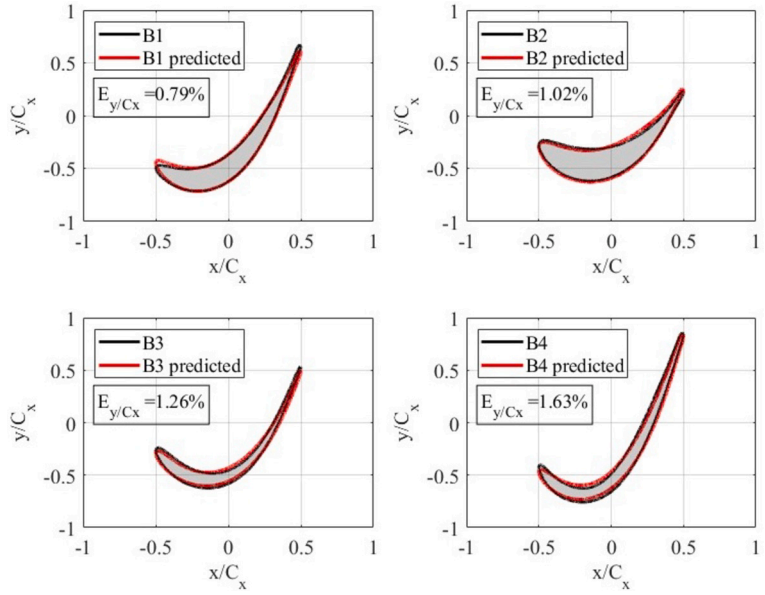

Fig. 11. Prediction of the four blades $B1$-$B4$ from the POD-based geometry model, when they are withdrawn from the model tuning.

Then, M2 can be used for the prediction of the coefficients that, once combined with the modes extracted from the database, provide the optimal blade shape.

In order to directly quantify the improvement in the aerodynamic performance that can be obtained by means of the proposed procedure, Fig. 13 shows the (normalized) loss coefficient characterizing the unsteady operation of the T106A cascade (baseline) for different Reynolds numbers, compared with the optimal cascade suggested by the procedure, when it is interrogated for a (normalized) targeting Reynolds number $Re_{tg} = 0.8$. The lapse rates predicted by the M1 model are reported with continuous lines. Dots instead represent new simulations carried out for three Reynolds numbers and with the other parameters frozen to the same values for both cascades. A 6% performance improvement has been obtained at the targeting Reynolds number. This is due to the improved values $\{Zw, Ds'\}_{opt}$ that optimize the flow diffusion over the blade suction side for that specific flow condition. The price to be paid is a lower performance at low Reynolds number due to higher sensitivity to flow separation. This is also confirmed by the simulations carried out at the lowest Reynolds number, where the baseline T106A cascade shows lower losses than the optimized one. Importantly, the reduced performance of the optimized cascade observed at the smallest Reynolds number here inspected is also properly predicted by the model M1. Thus, the proposed workflow will be able to define a different blade shape for each targeting Reynolds number.

5. Conclusions

This work presented a ML strategy for the identification of optimal LPT design region and related blade shape. The workflow involves a loss-correlation model that can be used to identify the combination of the design parameters minimizing the expected losses, and a POD-based geometry model able to generate the corresponding blade shape. The loss-correlation model has been shown to be able to predict losses with an expected error of 2.35% (leave-one-out mean error) with respect to the mean value numerically computed. The uncertainty in interpolation was found about the 3.60% of the loss coefficient, which grows to 3.75 times the one in interpolation when moving to the 20% outside of the data domain. The POD-based geometry model has instead been defined by 5 shape functions. Indeed, POD has been shown able to provide a compressed representation of the blade geometry, with five modes able

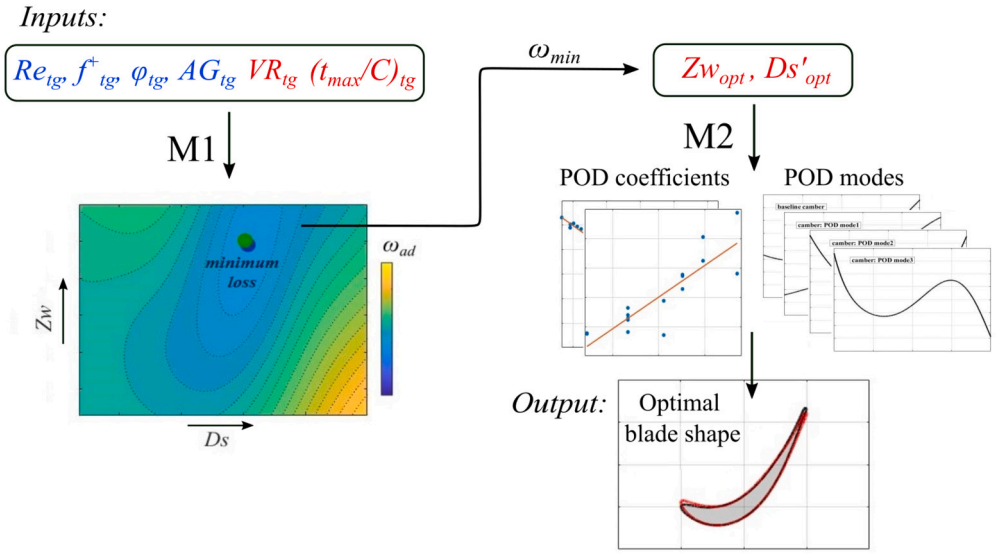

Fig. 12. Proposed workflow for the identification of the optimal design region and the related blade shape.

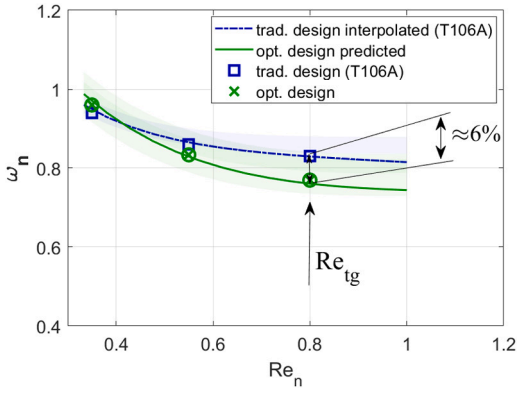

Fig. 13. Performance comparison between baseline (T106A cascade) and ML optimal cascade for different Reynolds numbers.

to describe all the blades in the database. The relation between the POD coefficients and the geometrical parameters allows the accurate generation of different blade profiles, reproducing their main geometrical characteristics. The workflow is self-adaptive, i.e., the models can be updated and re-calibrated whenever new data are available. Moreover, the set of parameters used to condense the cascade performance can be extended when larger databases are considered, to further improve the accuracy of the procedure. The strategy here proposed may generally be of inspiration for other engineering applications where efficiency and geometrical data are available and/or can be easily produced.

Nomenclature

α_1	Flow inlet angle
α_2	Flow outlet angle
χ	POD coefficients
ϵ	Variance ratio
λ	POD eigenvalues
μ	Uniformity of the leverage distribution
ω	Profile loss coefficient
ϕ	POD modes
φ	Flow coefficient
AG	Axial Gap
C	Blade chord, Camber
C_x	Blade axial chord
Ds'	Diffusion rate
E_{LOO}	Leave-one-out mean absolute error
f^+	Reduced frequency
$M1$	Loss-correlation model
$M2$	POD-based geometry model
M	Mach number
Re_x	Reynolds number
s	Coordinate on the suction side
S_{bar}	Pitch between the bars
s_{decel}	Decelerating part of the suction side
s_{tot}	Suction side length
t_{max}	Maximum thickness
u_{bar}	Velocity of the bars
VR	Velocity Ratio
Zw	Zweifel number

CRediT authorship contribution statement

Daniele Petronio: Conceptualization, Data curation, Methodology, Visualization, Writing – original draft. **Pietro Paliotta:** Writing – original draft, Data curation, Methodology, Visualization. **Davide Lengani:** Supervision, Conceptualization. **Daniele Simoni:** Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgement

The authors wish to sincerely thanks MorfoDesign for providing the CFD numerical database, on which the routines and the methods discussed in this work have been developed.

References

[1] M.A. Mendez, J. Dominique, M. Fiore, F. Pino, P. Sperotto, J.V.d. Berghe, Challenges and opportunities for machine learning in fluid mechanics, https://doi.org/10.48550/ARXIV.2202.12577, 2022.

[2] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. B 58 (1) (1996) 267–288, https://www.jstor.org/stable/2346178.
[3] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12 (1) (1970) 55–67, https://doi.org/10.1080/00401706.1970.10488634.
[4] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc., Ser. B, Stat. Methodol. 67 (2) (2005) 301–320, https://www.jstor.org/stable/3647580.
[5] D. Simoni, D. Barsi, M. Dellacasagrande, D. Lengani, Y. Vianney, Modified formulation of laminar kinetic energy transition models by means of elastic-net of a big experimental database of separated flows, Flow Turbul. Combust. 105 (2020), https://doi.org/10.1007/s10494-020-00124-2.
[6] S.L. Brunton, J.L. Proctor, J.N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proc. Natl. Acad. Sci. 113 (15) (2016) 3932–3937, https://doi.org/10.1073/pnas.1517384113.
[7] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, The MIT Press, 2005.
[8] G. Dreyfus, Neural Networks - Methodology and Applications, 2005.
[9] M. Schmelzer, R. Dwight, P. Cinnella, Discovery of algebraic Reynolds-stress models using sparse symbolic regression, Flow Turbul. Combust. (2020), https://doi.org/10.1007/s10494-019-00089-x.
[10] J. Dominique, J.V. den Berghe, C. Schram, M.A. Mendez, Artificial neural networks modeling of wall pressure spectra beneath turbulent boundary layers, Phys. Fluids 34 (3) (2022) 035119, https://doi.org/10.1063/5.0083241.
[11] H. Akaike, Information Theory and an Extension of the Maximum Likelihood Principle, Springer, New York, New York, NY, 1998, pp. 199–213.
[12] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (2) (1978) 461–464, https://doi.org/10.1214/aos/1176344136.
[13] R.R. Picard, R.D. Cook, Cross-validation of regression models, J. Am. Stat. Assoc. 79 (387) (1984) 575–583, https://doi.org/10.2307/2288403.
[14] M. Dellacasagrande, D. Lengani, P. Paliotta, D. Petronio, D. Simoni, F. Bertini, Evaluation of different regression models tuned with experimental turbine cascade data Volume 10B: Turbomachinery — Axial Flow Turbine Aerodynamics, https://doi.org/10.1115/GT2022-81357, 2022.
[15] G. Berkooz, P. Holmes, J.L. Lumley, The proper orthogonal decomposition in the analysis of turbulent flows, Annu. Rev. Fluid Mech. 25 (1) (1993) 539–575, https://doi.org/10.1146/annurev.fl.25.010193.002543.
[16] L. Shi, H. Ma, X. Yu, Pod analysis of the unsteady behavior of blade wake under the influence of laminar separation vortex shedding in a compressor cascade, Aerosp. Sci. Technol. 105 (2020) 106056, https://doi.org/10.1016/j.ast.2020.106056.
[17] N. Peters, A. Wissink, J. Ekaterinaris, Machine learning-based surrogate modeling approaches for fixed-wing store separation, Aerosp. Sci. Technol. 133 (2023) 108150, https://doi.org/10.1016/j.ast.2023.108150.
[18] M. Guo, J.S. Hesthaven, Reduced order modeling for nonlinear structural analysis using Gaussian process regression, Comput. Methods Appl. Mech. Eng. 341 (2018) 807–826, https://doi.org/10.1016/j.cma.2018.07.017.
[19] G.M. Robinson, A.J. Keane, Concise orthogonal representation of supercritical airfoils, J. Aircr. 38 (3) (2001) 580–583, https://doi.org/10.2514/2.2803.
[20] D.J.J. Toal, N.W. Bressloff, A.J. Keane, C.M.E. Holden, Geometric filtration using proper orthogonal decomposition for aerodynamic design optimization, AIAA J. 48 (5) (2010) 916–928, https://doi.org/10.2514/1.41420.
[21] J. Li, X. Du, J.R. Martins, Machine learning in aerodynamic shape optimization, Prog. Aerosp. Sci. 134 (2022) 100849, https://doi.org/10.1016/j.paerosci.2022.100849.
[22] Y. Duan, W. Wu, P. Zhang, F. Tong, Z. Fan, G. Zhou, J. Luo, Performance improvement of optimization solutions by pod-based data mining, Chin. J. Aeronaut. 32 (4) (2019) 826–838, https://doi.org/10.1016/j.cja.2019.01.014.
[23] J. Luo, Y. Zhu, X. Tang, F. Liu, Flow reconstructions and aerodynamic shape optimization of turbomachinery blades by pod-based hybrid models, Sci. China, Technol. Sci. 60 (2017), https://doi.org/10.1007/s11431-016-9093-y.
[24] D. Rajnarayan, A. Ning, J.A. Mehr, Universal airfoil parametrization using B-splines, https://doi.org/10.2514/6.2018-3949, 2018.
[25] T.W. Sederberg, S.R. Parry, Free-form deformation of solid geometric models 20 (1986) 151–160, https://doi.org/10.1145/15922.15903.
[26] D.A. Masters, N.J. Taylor, T.C.S. Rendall, C.B. Allen, D.J. Poole, Geometric comparison of aerofoil shape parameterization methods, AIAA J. 55 (5) (2017) 1575–1589, https://doi.org/10.2514/1.J054943.
[27] D. Simoni, D. Lengani, D. Petronio, F. Bertini, A Bayesian approach for the identification of cascade loss model strategy Volume 2B: turbomachinery, https://doi.org/10.1115/GT2020-14625, 2020.
[28] H.R.M. Craig, H.J.A. Cox, Performance estimation of axial flow turbines, Proc. Inst. Mech. Eng. 185 (1) (1970) 407–424, https://doi.org/10.1243/PIME_PROC_1970_185_048_02.
[29] S.C. Kacker, U. Okapuu, A mean line prediction method for axial flow turbine efficiency, J. Eng. Power 104 (1) (1982) 111–119, https://doi.org/10.1115/1.3227240.
[30] J.D. Coull, H.P. Hodson, Predicting the profile loss of high-lift low pressure turbines, (2), J. Turbomach. 134 (2) (2012), https://doi.org/10.1115/1.4002961.
[31] R. Stieger, D. Hollis, H. Hodson, Unsteady surface pressures due to wake-induced transition in a laminar separation bubble on a low-pressure cascade, J. Turbomach. 126 (2003), https://doi.org/10.1115/1.1773851.
[32] D. Lengani, D. Simoni, M. Ubaldi, P. Zunino, F. Bertini, V. Michelassi, Accurate estimation of profile losses and analysis of loss generation mechanisms in a turbine cascade, J. Turbomach. 139 (12) (2017), https://doi.org/10.1115/1.4037858.
[33] R. Stieger, H. Hodson, The transition mechanism of highly loaded low-pressure turbine blades, J. Turbomach. 126 (2004), https://doi.org/10.1115/1.1773850.
[34] F. Satta, D. Simoni, M. Ubaldi, P. Zunino, F. Bertini, Loading distribution effects on separated flow transition of ultra-high-lift turbine blades, J. Propuls. Power 30 (2014) 845–856, https://doi.org/10.2514/1.B34968.
[35] H. Schlichting, Boundary-Layer Theory, Mc Graw Hill, 1979.
[36] P.C. Hansen, The truncated svd as a method for regularization, Tech. Rep., Stanford, CA, USA, 1986, https://api.semanticscholar.org/CorpusID:37591557.
[37] S. Volkwein, Proper orthogonal decomposition: Theory and reduced-order modelling, Lecture Notes, University of Konstanz, 2013.
[38] A.R. Howell, Axial flow turbines. J. H. Horlock. Butterworths, London. 1966. 275 pp. Diagrams. 97s. 6d., Aeronaut. J. 71 (673) (1966) 57–58, https://doi.org/10.1017/S0001924000056591.
[39] J.D. Coull, H.P. Hodson, Blade loading and its application in the mean-line design of low pressure turbines, J. Turbomach. 135 (2) (2012) 021032, https://doi.org/10.1115/1.4006588.