Energy Conversion and Management 243 (2021) 114339

Contents lists available at ScienceDirect

Energy Conversion and Management

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

Framework of airfoil max lift-to-drag ratio prediction using hybrid feature mining and Gaussian process regression

Yaoran Chen [a], Zhikun Dong [a], Jie Su [a], Yan Wang [a], Zhaolong Han [a,b,c,d,*], Dai Zhou [a,b,c,d,*], Yongsheng Zhao [a,b], Yan Bao [a,b,c,d]

[a] School of Naval Architecture, Ocean & Civil Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[b] State Key Laboratory of Ocean Engineering, Shanghai Jiao Tong University, Shanghai 200240, PR China
[c] Key Laboratory of Hydrodynamics of Ministry of Education, Shanghai Jiao Tong University, Shanghai 200240, PR China
[d] Shanghai Key Laboratory for Digital Maintenance of Buildings and Infrastructure, Shanghai Jiao Tong University, Shanghai 200240, PR China

ARTICLE INFO

Keywords:
Airfoil
Max lift-to-drag ratio
Gaussian process regression
Feature pool
Feature selection

ABSTRACT

The maximum lift-to-drag coefficient of an airfoil directly affects the aerodynamic performance of wind turbine. Machine learning methods are known for being really effective in helping to predict this parameter in a faster and more accurate way. So far, the majority of related studies have focused on the use of artificial neural networks to make this prediction, but this model has issues with its poor interpretation and the confidence level of its results was unclear. In this paper, a novel framework is proposed, involving the Gaussian process regression and a hybrid feature mining process. The aim is to use the new framework to evaluate the maximum lift-to-drag ratio of given airfoils under a turbulent flow condition, where the Reynolds number is around 100,000. The feature mining process here designed contains a hybrid feature pool that comprises various geometric characters, and a hybrid feature selector that can assist the prediction performance and make it better. Based on the airfoil dataset of the University of Illinois at Urbana-Champaign that contains a total of 1432 profiles, a comparative analysis was conducted. The results showed that the current framework can provide a more accurate estimate than parallel models in both single-point and interval aspects of view. Noticeably, the model reached an overall precision of 95.2% and 94.1% on training and testing sets, respectively. Moreover, the simplicity and the confidence reference from the model output were further illustrated with a case study, which also verified that how it can serve real engineering application.

1. Introduction

Due to growing environmental issues, many countries are turning to renewable sources of energy. Among these, wind power is the most prevalent [1]. This energy source relies on turbines to convert wind to electricity and has attracted the attention of many scholars recently [1].

Researches have shown positive correlation between the designed tip speed ratio of wind turbine's peak power coefficient and the lift-to-drag ratio of the blades, and the rotor will achieve its desirable working stage as the blades operate around the maximum ratio [2]. Therefore, being able to predict this ratio is of great importance during the early design stage of the engine.

Traditionally, the analysis of the maximum lift-to-drag ratio depends on experimental procedures (e.g. wind-tunnel test [3], field test [4]) and numerical measurements (e.g. XFOIL, StarCCM + and Fluent) that are based on aerodynamic theories [5] using different turbulence models [6]. These methods are time-consuming since they require specific equipment settings and measuring and post-processing stages [7]. However, owing to the development of data science and artificial intelligence, machine learning (ML) based methods are boosting as an effective alternative to evaluate airfoil performance. Such methods have greatly accelerated the predicting speed and are well known for also guaranteeing its precision [8].

So far, researchers of this field mainly used artificial neural networks (ANN) as their ML models. For instance, Sahuck (2020) [9] used a two-hidden-layer ANN for maximum lift-to-drag ratio prediction and compared the results with the response surface method. In his study, different distributions of control points were investigated as the input feature and the structure of neural network was systematically tuned. Similarly, Wen et al. (2019) [10] adopted a back-propagation ANN to make the same prediction, with hyper-parameters that were

* Corresponding authors at: School of Naval Architecture, Ocean & Civil Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.
E-mail addresses: han.arkey@sjtu.edu.cn (Z. Han), zhoudai@sjtu.edu.cn (D. Zhou).

https://doi.org/10.1016/j.enconman.2021.114339
Received 2 February 2021; Accepted 24 May 2021
Available online 11 June 2021

Nomenclature

Abbreviations

ANN	Artificial Neural Network
CART	Classification and Regression Trees
EVS	Explained Variance Score
IWSS	Incremental Wrapper-based Subset Selection
PARSEC	Parametric Sections
RF	Random Forest
CST	Class-Shape Transformation
DT	Decision Tree
GPR	Gaussian process regression
ML	Machine Learning
RBF	Radial Basis Function
SFS	Sequential Forward Selection

Symbols

C_l	Lift coefficient
C_d	Drag coefficient
C_m	Moment coefficient
C_l/C_d	Lift-to-drag ratio (i.e. Sliding ratio)
T_i	Thickness of i^{th} node of airfoil
C_i	Camber of i^{th} node of airfoil
K_i	Curvature of i^{th} node of airfoil
y_i^{upper}	y-coordinate of i^{th} node at airfoil upper surface
y_i^{lower}	y-coordinate of i^{th} node at airfoil lower surface
d	Euclidean distance between airfoil nodes
$S_{i-1,i,i+1}$	Triangle area formed i^{th} and adjacent nodes
X	x-coordinate on normalized airfoil
T_{max}	Max thickness
$X_{T_{max}}$	Max thickness location
C_{max}	Max camber
$X_{C_{max}}$	Max camber location
$\overline{T}$	Total mean thickness
$\overline{C}$	Total mean camber
$\overline{T_{lead}}$	Mean thickness at leading part
$\boldsymbol{\omega}$	Current feature set
$Q(\boldsymbol{\omega})$	Interval prediction estimator
$\widehat{y}$	Normal distribution from GPR prediction
σ	Standard deviation of GPR prediction
$Var(\hat{A}\cdot)$	Variance
α_0	Noise level of GPR
η	Constant kernel value of GPR
λ_{σ_0}	Value of λ_σ in current model
$N_{shuffle}$	Times of re-shuffle to compute the feature importance in RF
$\overline{T_{mid}}$	Mean thickness at middle part
$\overline{T_{trail}}$	Mean thickness at trailing part
K_{top}	Top-most surface curvature
K_{bot}	Bot-most surface curvature
$T_{c/4}$	Thickness at 1/4 chord
$T_{c/2}$	Thickness at 1/2 chord
$T_{3c/4}$	Thickness at 3/4 chord
$C_{c/4}$	Camber at 1/4 chord
$C_{c/2}$	Camber at 1/2 chord
$C_{3c/4}$	Camber at 3/4 chord
D	Training dataset
$\boldsymbol{x_i}$	Input of training dataset
y_i	Output of training dataset
L_t	Loss of t^{th} tree node in RF
j	Index of feature
$I^{(j)}$	Feature importance of j^{th} feature
T	Total number of tree nodes in RF
N_t	Number of trees in RF
N_{trn}	Number of training airfoils
$P(\boldsymbol{\omega})$	Point prediction estimator
λ_σ	Weight item in objective function
$\widehat{y}$	Mean value of GPR prediction
$\boldsymbol{k}(\hat{A}\cdot)$	Kernel function of GPR
$\overline{\sigma_{set}}$	Mean standard deviation of dataset
l_{rbf}	RBF length scale of GPR
N_t	Number of trees in RF
μ	Point prediction of GPR in case study
S_{split}	Minimum number of samples required to split an internal node in RF

heuristically selected. The accuracy of their prediction was reaching 90%. In the same research direction, Viquerat et al. (2020) [11] terminated their model structure with two fully connected ANN layers to predict the drag of an arbitrary 2D airfoil. Because the flow condition was under a laminar stage, their work presented an impressive result with maximum relative error less than 2%. Despite the merit outcomes of the aforementioned studies, there are two main disadvantages of using ANN models for airfoil evaluation. On the one hand, since the ANNs are "black box" models [12], designers cannot comprehend how changing inputs can influence the prediction. On the other hand, the outcomes of ANN models are merely single-point estimates, which are inadequate for designers to properly evaluate the accuracy and validity of their prediction. Therefore, facilitated its understanding, the model should be re-designed in a way that solves the issues with reliability and interpretability.

A promising alternative to fill this gap is using the Bayesian learning methods, which include a typical algorithm: Gaussian process regression (GPR). Different from the ANN models, the results of the GRP are based on probabilities, which promotes one single aero-performance prediction and a confidence interval for users to assess risks [13]. Furthermore, the GPR belongs to the group of "non-parametric methods", which means that the parameters of the model will automatically self-update as the dataset is expanding. This can render its strong ability of generalization from a small-scaled dataset [14]. Some pioneers have employed the GPR on relative engineering investigations. For instance, Ref. [15] focused on establishing correlations between different types of aerodynamic characteristics, and Ref. [16] tried to use GPR in modeling wind energy time-series problems. However, it is important to address that few scholars have shown concerns about the implementation of the GPR on the maximum lift-to-drag prediction of different airfoils under a high Reynolds number.

To conduct the GPR on the current problem, it is first necessary to determine the most meaningful and suitable inputs of parameters that describe the airfoil features. This step is crucial because the GPR is non-sparse [14], meaning that it cannot use embedded algorithms to penalize or to eliminate the unwanted inputs while training, such as the different regulators in ANN [11], to overcome the model overfitting. Hence, for the GPR, it renders an elaborated feature mining process on inputs, which generally comprises two connected procedures: i) the construction of the feature pool that involves the potential features, such as average thickness or maximum camber of the airfoil, and ii) a careful feature selection process that detects the most useful features of the pool to improve the prediction accuracy as much as possible.

Previous researches have proposed various approaches to find potential parameters to describe the airfoil geometry. Generally, they can be categorized into three types, as listed in Table 1. Among them, the

Table 1
Comparison of methods for feature generation in aerodynamic performance prediction of airfoils.

Methods	Publications	Peculiarities	Set-backs
Control point based	[9,10,17]	Easy-implemented and widespread used; adopts extra methods to fit the profile	Poor generalization; challenging to determine the best number and the location of control points
Curve function based	[18,19,2122,27,28]	Some features have direct or indirect physical meanings	Sometimes redundant; Often limited to a specific type or group of airfoils
Figure based	[11,25,26]	End-to-end; deep features are auto-learned through convolutional kernels	Poor interpretability; difficult to conduct and to reproduce in real applications

most straightforward strategy is to use control-points to fit the contour through Bessel splines [9]. However, it is often challenging to allocate such points, let alone there is a Pareto balance between the total dot number and the geometrical similarities [17]. Besides, curve functions are an alternative for airfoil shape-construction. Based on this possibility, many researchers have shown their interests in applications of the well-known Joukowski transformation [18] and its simplified versions that consider the airfoil camber line and its thickness distribution [19]. Others may divide the whole profile into local components such as PARSEC [20] based methods in Ref. [21], or class-shape transformation (CST) based method in Ref. [22]. Compared with the control-points, one appealing merit of this type is its parameters often possess direct physical meanings, and this can allow users alter those features more effectively [23]. However, these parameters can be redundant since they were initially designed for shape-formulation and not for aerodynamic prediction [21]. Also, one curve expression usually does not cover a wide range of shape types, say, profiles not obeying the conformal rules of Joukowski transformation [24]. To solve this issue, more end-to-end methods are necessary, such as figure-based approaches (refer to Table 1. Row 3), so the airfoil contours can be transformed into 2D pixel-array and its deep features can be extracted with the help of convolutional kernels [25]. Nevertheless, such methods are not easy to implement in applications. To reproduce this technology, one should first format the target profile into the graph according to the given resolution and airfoil location [26]. Also, although the deep features can be automatically learned by ML models, their meanings are not fully understood due to the complexity of hyper-parameter tuning for the best convolutional kernels.

Focused on the second step only, there are various feature selection techniques that can be used to enhance the prediction's accuracy. Such process can be regarded as a re-extraction of the potential inputs to better fit the machine learning model. Typical feature selection methods can be classified into wrapper-based and filter-based approaches, according to whether the selection is simultaneously processed with the regressor by turns. However, for wrapper-based methods, despite of a satisfactory outcome, the computational costs involving them are considerably high [12], especially when coupled with an expensive learning model – e.g., the Gaussian process regression. When it comes to filter based methods, e.g., uni-variate feature selection and tree-based selection [29], although the selection speed is fast, the performance of the sub-set may sometimes not be efficient since the selector is often decoupled with the regressor.

By synthesizing the benefits and limits of the above methods, a hybrid feature mining approach is proposed, containing both a hybrid feature pool and a hybrid feature selector. The hybrid pool's contradicting to point-based and functions-based features, makes it highly compatible and extensible. This happens because all potential features are considered global or local geometric attributes of the given airfoil, instead of parameters that should mathematically compute the whole profile in detail. Also, different from figure-based methods, these features are extremely convenient when it comes to computing the airfoil contour, and has explicit geometric meanings to make the model more interpretable. In addition, a hybrid feature selector is also proposed in this work to assist the process of feature selection. Such selector integrates the tree-based selector (that acts as a filter) and the sequential forward selector (that acts as a wrapper) as a way to improve the prediction accuracy with a reduced computational budget. A detailed description of this hybrid approach is presented in the "Methods" section.

In all, this paper proposes an integrated machine learning framework used to predict the maximum lift-to-drag ratio of a given airfoil in a probabilistic, interpretable and accurate way. The process involves a combination of the aforementioned hybrid feature mining approach and the GPR predictor. First, using the coordinates of the given profile, the potential geometric attributes are extracted to form the hybrid feature pool. Then, the maximum lift-to-drag ratio and its confidence level are estimated using the GPR model, which was elaborately trained through a hybrid feature selection to promote its accuracy. Such selection includes two parts in tandem: i). a tree-based selector to initially determine the significance of the geometric attributes of the hybrid feature pool; ii). a wrapper-based selector to iteratively couples with the GPR model to promote a recurrent feature selection, downward along the rank of the feature significance. The original aspects of this study are stated as follows:

1 A novel statistical model based on the GPR was built, with the aim to offer both single-point and interval prediction of the maximum C_l/C_d for the given airfoil under a turbulence flow condition.
2 A novel hybrid feature mining approach was proposed to promote the prediction accuracy of the GPR model, not only to attempt a more precise point estimation but also a more concentrated interval estimation.
3 The framework is totally end-to-end and the input features are completely interpretable, having explicit meanings from the profile geometry.
4 The feasibility of the framework was tested by a large airfoil dataset obtained from University of Illinois at Urbana-Champaign (UIUC) [30], followed by a clear annotation of the case study to enrich the engineering reference from the outcome.

The roadmap of the article is as below. Section 2 will describe the raw dataset and the way of pre-processing. In Section 3, the methods of the framework will be explained, including feature extraction, hybrid feature selection and GPR. Section 4 will demonstrate and discuss the numerical results. Finally, in Section 5, solid conclusions will be summarized and the future work will be presented.

2. Raw dataset and pre-processing

In this section, the source and overview of the airfoil dataset (Section 2.1) as well as the approaches of pre-processing (Section 2.2) are presented.

2.1. Raw data description

The raw airfoil database of this work was obtained from the up-to-date UIUC data site that is publically available [30]. The database, which has been widely used and cited in previous researches [31], comprises two main parts. One is the profile shape that was depicted by bunches of *x-y* coordinates, belonging to a variety of airfoil families, e.g. NACA, Boeing, NASA, etc. The other is the airfoil performance, which was mostly evaluated through UIUC Low-Speed Airfoil Test program

(LSATs) under a wide range of Reynolds numbers (from 60,000 to 500,000) and angle of attacks. The tested aerodynamic parameters include the lift coefficient (C_l), drag coefficient (C_d) and moment coefficient (C_m) etc.

In this study, as a representative parameter of airfoil, the maximum lift-to-drag ratio is considered as the target to be investigated [10]. Meanwhile, to fit the flow condition for wind turbine applications [32], the target Reynolds number is chosen to be 100,000.

2.2. Data pre-treatment

In this section, a careful formalization is conducted on raw profile's co-ordinates to make it more convenient to generate features for the upcoming ML model.

Firstly, the starting points and end points of all airfoils are fixed, located at (0, 0) and (1, 0). If there is a bias (see Fig. 1), a translation or rotation would be conducted. Through this step, a same zero-base reference is constructed for every airfoil, hence it is convenient to find geometric attributes by simply calling the corresponding coordinates. Secondly, the same curve formulation method is applied on all airfoil coordinates. In this work, instead of using high-order Bezier splines, polygons are adopted to formulate the curve, since the raw coordinates list is stated to be the points *along* the upper and lower airfoil surfaces [30]. Depending on different given dots among different airfoils, this step allows the interpolation of the y value for any given x. Finally, the dataset is fully cleansed to wipe out the outliers as far as possible. The aberrations mainly include airfoils with an extremely thin thickness (e.g. E377 ultralight airfoil and its family [30]), airfoils with sharp local concaves (e.g. St. CYR171 airfoil [30]) or airfoils without a unit chord length (e.g. UA 79-SF-187 sailplane airfoil [30]), etc.

As a result, the final number of the legal airfoils is counted to be 1432, out of the total 1635 raw profiles. The geometry information and the max sliding ratio (C_l/C_d) of these airfoils are the final database feeding into the machine learning model.

3. Methods

In this section, the methods of the proposed machine learning framework are presented. As shown in Fig. 2, after pre-processing (Section 1), the workflow mainly includes three parts: the generation of hybrid feature pool, the hybrid feature selection and the maximum lift-to-drag ratio prediction through Gaussian process regression. The algorithm and the connection of these three parts will be introduced in Section 3.1, Section 3.2 and Section 3.3, respectively.

3.1. Hybrid feature pool

In this work, there were 17 potential geometry features considered for maximum sliding ratio prediction. The features were enlightened form the previous investigations, for instance, the PARSEC, CST and Joukowski transformation. The current feature pool comprised three main kinds of geometrical measurements, including thickness, middle line deflection (i.e. camber) and curvature. They could be easily calculated using the pre-processed co-ordinates of the airfoil through expressions below:

$$T_i = y_i^{upper} - y_i^{lower} \tag{1}$$

$$C_i = \left(y_i^{upper} + y_i^{lower}\right)/2 \tag{2}$$

$$K_i = 4S_{i-1,i,i+1}/d_{i-1,i}d_{i,i+1}d_{i-1,i+1}, \tag{3}$$

whereT_i,C_i and K_i denote the thickness, camber and curvature of i^{th} node; y_i^{upper} and y_i^{lower} respectively denote the y-coordinates on upper and lower surface; d is the Euclidean distance between points; $S_{i-1,i,i+1}$ is the triangle area formed by i^{th} and its adjacent points.

As shown in Table 2 and Fig. 3, the current features were categorized into three levels. (i). global parameters, including the average and max thickness (camber) and their locations; (ii). local profile features in vertical or transverse directions, including the average thickness over different parts of the airfoil (from the lead to trail) and the curvatures of the top-most and bottom-most surface; (iii). single-point features representing the thickness (and camber) at quartering sections.

Obviously, this hybrid feature pool could be further expanded by adding other geometric parameters, but temporarily in this work, those 17 features are used as a basic trail.

3.2. Hybrid feature selection

A hybrid feature selection method was built in this work, comprising both filter and wrapper based selectors. To begin with, the random forest (RF) algorithm was used to provide a sorted sequence consisting of the

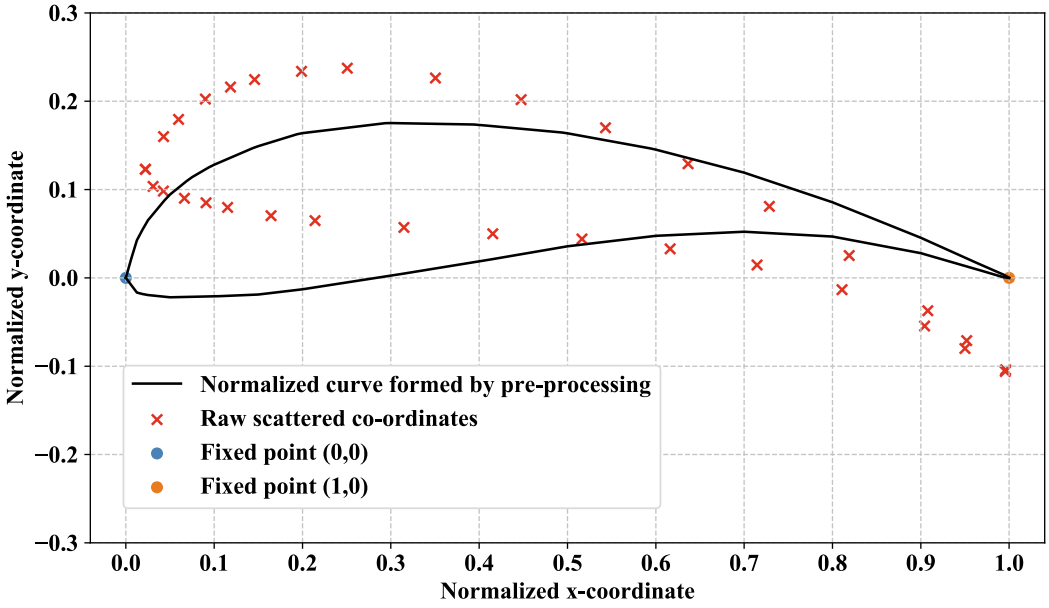

Fig. 1. Illustration of the reformulation on raw airfoil coordinates [30] (No. 00599, Airfoil name: GOE 244 (MVA PR.4)).

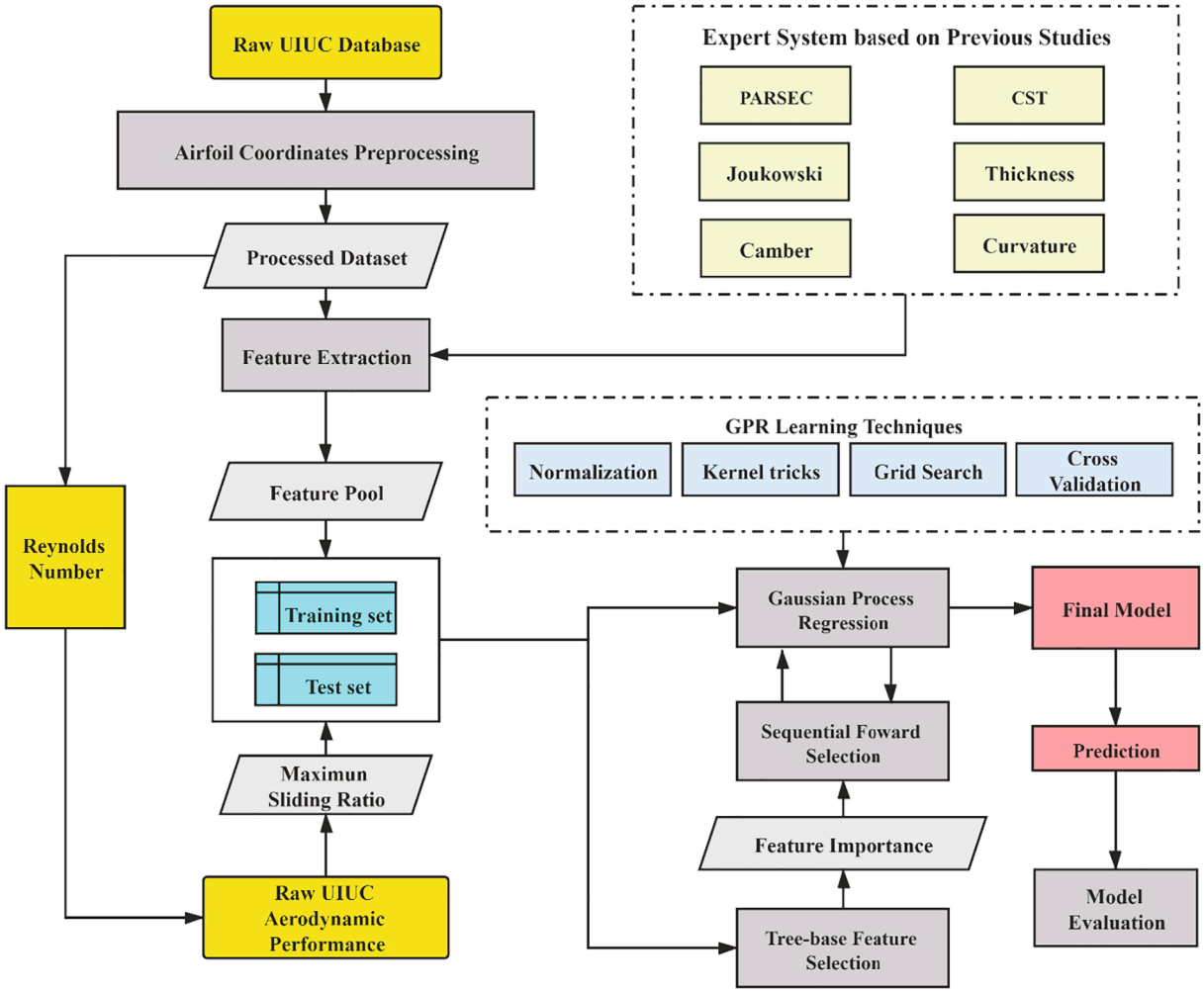

Fig. 2. Flowchart of the machine learning framework.

Table 2
Geometrical feature pool of airfoil for machine learning.

Category	Feature description	Symbol
Global feature	Max thickness	T_{max}
	Max thickness location	$X_{T_{max}}$
	Max camber	C_{max}
	Max camber location	$X_{C_{max}}$
	Total mean thickness	$\overline{T}$
	Total mean camber	$\overline{C}$
Local feature	Mean thickness at leading part	$\overline{T_{lead}}$
	Mean thickness at middle part	$\overline{T_{mid}}$
	Mean thickness at trailing part	$\overline{T_{trail}}$
	Top-most surface curvature	K_{top}
	Bot-most surface curvature	K_{bot}
Single feature	Thickness at 1/4 chord	$T_{c/4}$
	Thickness at 1/2 chord	$T_{c/2}$
	Thickness at 3/4 chord	$T_{3c/4}$
	Camber at 1/4 chord	$C_{c/4}$
	Camber at 1/2 chord	$C_{c/2}$
	Camber at 3/4 chord	$C_{3c/4}$

quantified importance of each feature from the feature pool. Based on this, the sequential forward selection (SFS) was then implemented to further determine the sub-feature set for improving the prediction accuracy of the machine learning model: GPR.

It is worth noting that such hydration is a trade-off, making their respective advantages complementary to each other. On the one hand, as a low budget warm-up, the RF algorithm can provide a filtrating routine, along which the SFS would smoothly proceed. Next, in return, the SFS method will re-check this order step by step through the coupled iteration with GPR, and improve the sub-feature set performance as a result.

3.2.1. *Random forest feature selection*

The core idea of tree-based feature filters is to eliminate unwanted features according to their importance. As a typical representative, the RF is an ensemble of a number of randomized basic selectors, known as the decision tree (DT). To be the milestone of such algorithm, the Classification and Regression Trees (CART) algorithm that originally developed by Breiman et al. [33] is adopted in this work.

Given a training set$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2) \cdots (\boldsymbol{x}_{N_{trn}}, y_{N_{trn}})\}$, where$\boldsymbol{x}_i = (x_i^{(1)}, x_i^{(2)} \cdots x_i^{(n)})$, the construction of DT regression is to find pairs of (j, s) from treetop to the root to minimize the loss [33]:

$$L_t = \sum_{x_i \in R_1(j,s)} (y_i - \widehat{c_1})^2 + \sum_{x_i \in R_2(j,s)} (y_i - \widehat{c_2})^2, \tag{4}$$

where t is the tree node index;$R_1(j,s) = \{x_i^{(j)} | x_i \le s\}$,$R_2(j,s) = \{x_i^{(j)} | x_i > s\}$ are sub-regions of x set under the current node; $\widehat{c_1} = \sum_{x_i \in R_1(j,s)} y_i / N_1$, $\widehat{c_2} = \sum_{x_i \in R_2(j,s)} y_i / N_2$ are the

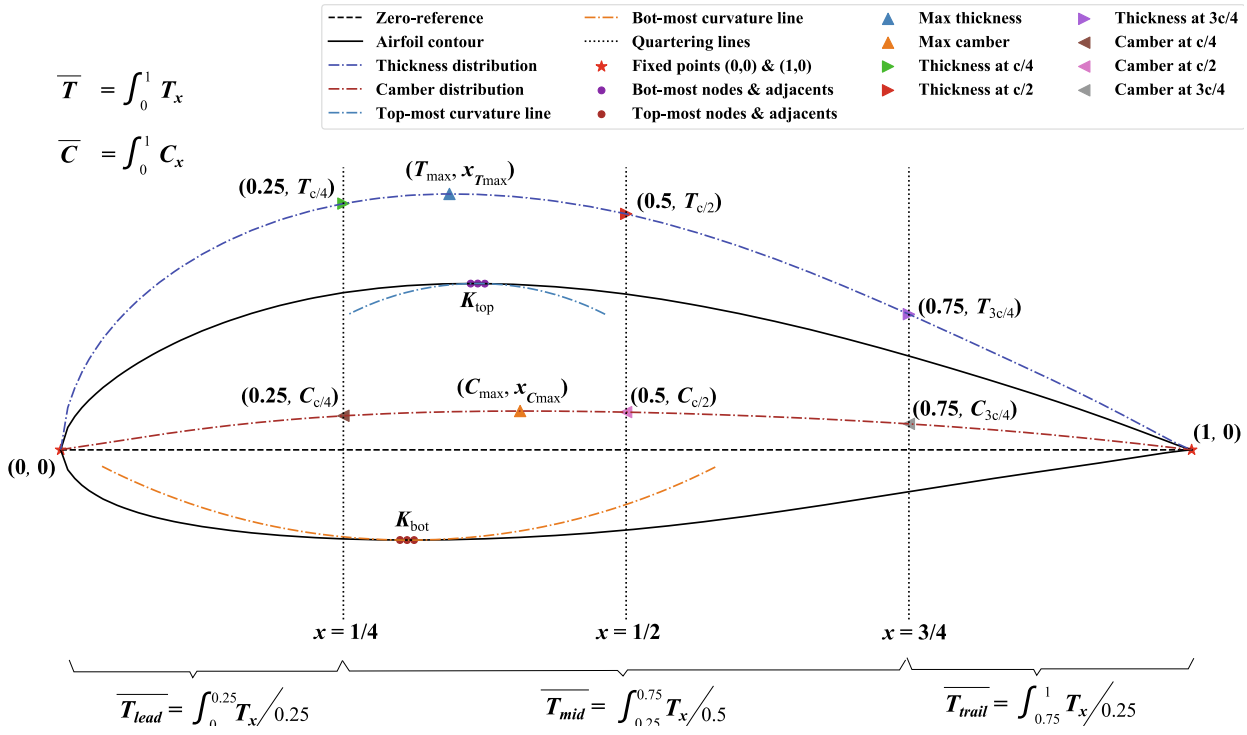

Fig. 3. Illustration of geometrical features in feature pool (No. 00496; Airfoil name: FX 83-W-227).

averaged y values in each region.

Once the regression tree has been fitted, the permutation importance $I^{(j)}$ of feature j is measured as follows: (i). Shuffle$x^{(j)}$and keep other columns unchanged; (ii). Make predictions using the shuffled dataset; and (iii). Calculate the deterioration of the loss functions by using shuffled dataset from the untreated one [13].

$$I^{(j)} = \sum_{t=1}^{T} L_t^{(j)} - \sum_{t=1}^{T} L_t, \tag{5}$$

where$\sum_{t=1}^{T} L_t^{(j)}$ denotes the total loss function using the dataset with j^{th} feature shuffled; T is the total number of tree nodes.

As an ensemble, to control over-fitting by using just a single tree, the RF fits a number of randomized decision trees on various sub-samples of the training dataset through averaging. Therefore, the final feature importance is denoted by [34]:

$$\overline{I^{(j)}} = \sum_{n_t=1}^{N_t} I_{n_t}^{(j)} \Big/ N_t, \tag{6}$$

whereN_t is the number of trees in forest, set to be 200 in this work.

3.2.2. Sequential forward selection

Starting from the empty set, the SFS wraps the estimator to examine the feature performance in succession. However, this method is intractable for high-dimensional problems because of the considerable amount of evaluations to be carried out [35]. For instance, the total evaluations for current feature pool should become17! ≈ 3.5trillions, which is impossible to accomplish with GPR as the predictor.

To alleviate this problem, inspired by the concept of "Incremental Wrapper-based Subset Selection (IWSS)" [36], the idea is to connect the wrapper with the tree-based filter (Section 3.2.1) and in this paper, it is to let the SFS guided by the feature importance rank (Fig. 4). Here, the target function$J(\boldsymbol{\omega})$ to evaluate the model performance is defined as follows:

$$J(\boldsymbol{\omega}) = P(\boldsymbol{\omega}) \cdot Q(\boldsymbol{\omega})^{-\lambda_\sigma} \tag{7}$$

where $\boldsymbol{\omega}$ denotes the feature set through iteration; $P(\boldsymbol{\omega})$ and $Q(\boldsymbol{\omega})$ are the prediction accuracy estimators on target airfoil samples from pointed and interval aspects of view (Section 3.2.2); $\lambda_\sigma > 0$ is the weight item.

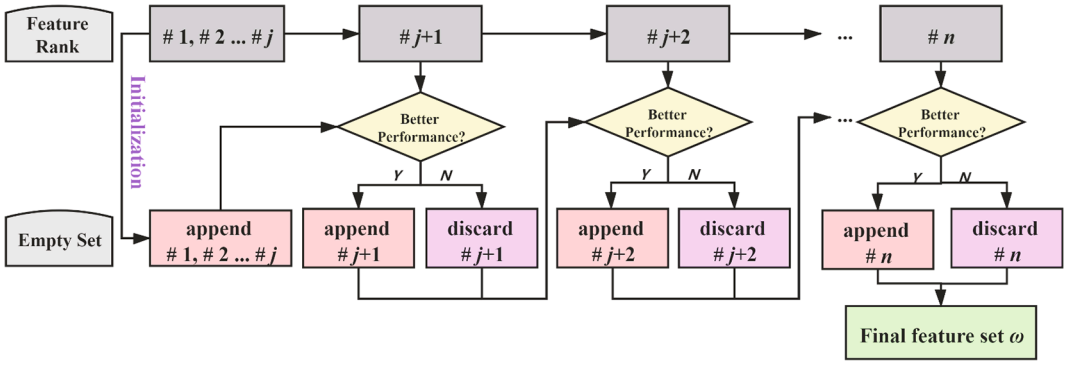

Fig. 4. Flowchart of the hybrid feature selection with feature rank provided by RF and wrapper conducted through SFS.

3.3. Gaussian process regression

GPR is adopted as the maximum lift-to-drag ratio predictor of this work, owing to its good extensibility, strong compatibility with moderate dataset and particular probabilistic outcomes for users to evaluate [14]. The theorem and the current application of GPR will be illustrated in following sub-sections.

3.3.1. Basic equations

Given a training set$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2) \cdots (\boldsymbol{x}_{N_{trn}}, y_{N_{trn}})\}$, the task of GPR is to forecast the $\widehat{y}^*$ given the new input$\boldsymbol{x}^*$. Based on Bayesian learning theory, GPR assumes that any finite set of $\boldsymbol{y}_m = \begin{bmatrix} \boldsymbol{y}_{trn} \\ \widehat{y}^* \end{bmatrix}$ obeys a joint normal distribution with prior zero mean and covariance kernel$\boldsymbol{k}$ [14]:

$$\begin{bmatrix} \boldsymbol{y}_{trn} \\ \widehat{y}^* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \boldsymbol{k}(\boldsymbol{x}_{trn}, \boldsymbol{x}_{trn}) & \boldsymbol{k}(\boldsymbol{x}_{trn}, \boldsymbol{x}^*) \\ \boldsymbol{k}(\boldsymbol{x}^*, \boldsymbol{x}_{trn}) & k(\boldsymbol{x}^*, \boldsymbol{x}^*) \end{bmatrix}\right) \tag{8}$$

this probabilistically gives the distribution of $\widehat{y}^*$ follows [37]:

$$\widehat{y}^*|_{\boldsymbol{x}^*, \boldsymbol{x}_{trn}, \boldsymbol{y}_{trn}} \sim \mathcal{N}(\widehat{y}^*, \sigma^*) \tag{9}$$

$$\widehat{y}^* = \boldsymbol{k}(\boldsymbol{x}^*, \boldsymbol{x}_{trn}) \boldsymbol{k}(\boldsymbol{x}_{trn}, \boldsymbol{x}_{trn})^{-1} \boldsymbol{y}_{trn} \tag{10}$$

$$\sigma^* = k(\boldsymbol{x}^*, \boldsymbol{x}^*) - \boldsymbol{k}(\boldsymbol{x}^*, \boldsymbol{x}_{trn}) \boldsymbol{k}(\boldsymbol{x}_{trn}, \boldsymbol{x}_{trn})^{-1} \boldsymbol{k}(\boldsymbol{x}_{trn}, \boldsymbol{x}^*) \tag{11}$$

where σ^* is the standard deviation of$\widehat{y}^*$; the dimension of $\boldsymbol{k}(\boldsymbol{x}_{trn}, \boldsymbol{x}^*)$ equals to $N_{trn} \times 1$ for a single test sample of airfoil (i.e. $\widehat{y}^*$ is a scalar), similarly for other kernel matrix.

3.3.2. Performance criteria

Generally, the coefficient of determination (R^2) has been widely used to evaluate the regression models [38]. However, this score is feature-dependent and neglects bias in samples' variance. Therefore, from the view of dataset expansion and unbiased estimation, the explained variance score (EVS) is used to evaluate the performance [13]:

$$P(\boldsymbol{\omega}) = EVS(\boldsymbol{y}, \widehat{\boldsymbol{y}}(\boldsymbol{\omega})) = 1 - \frac{Var(\boldsymbol{y} - \widehat{\boldsymbol{y}}(\boldsymbol{\omega}))}{Var(\boldsymbol{y})} \tag{12}$$

where $\boldsymbol{y}$ and $\widehat{\boldsymbol{y}}$ are the real and the predicted maximum C_l/C_d of airfoil samples;$Var(\cdot)$ denotes the variance.

In addition, to evaluate the GPR point-prediction from the degree of dispersion and its stability, another criteria,$Q(\boldsymbol{\omega})$ is defined as follows:

$$Q(\boldsymbol{\omega}) = \overline{\sigma_{set}}(\boldsymbol{\omega}) = \sum_{m=1}^{N_{set}} \sigma_m^*(\boldsymbol{\omega}), \tag{13}$$

whereN_{set} denotes the size of the training or test set; $\sigma_m^*(\boldsymbol{\omega})$ is the standard deviation of prediction distribution of the m^{th} sample given $\boldsymbol{\omega}$ as the feature set.

3.3.3. Learning techniques

Several learning techniques are used to enhance the accuracy of prediction and strengthen the outcome reliability, including:

(1) Normalization. Min-Max normalization was adopted on both input features and their labels of max lift-drag ratio.
(2) Kernel production. The production of the constant kernel and RBF kernel [14] was applied during GPR process.
(3) Grid hyper-parameter searching. The hyper-parameters of GPR, including the noise level (α_0), the RBF length scale (l_{rbf}) and the constant value (η) of constant kernel, were systematically tuned through grids, with their distributions shown in Fig. 5 as below:

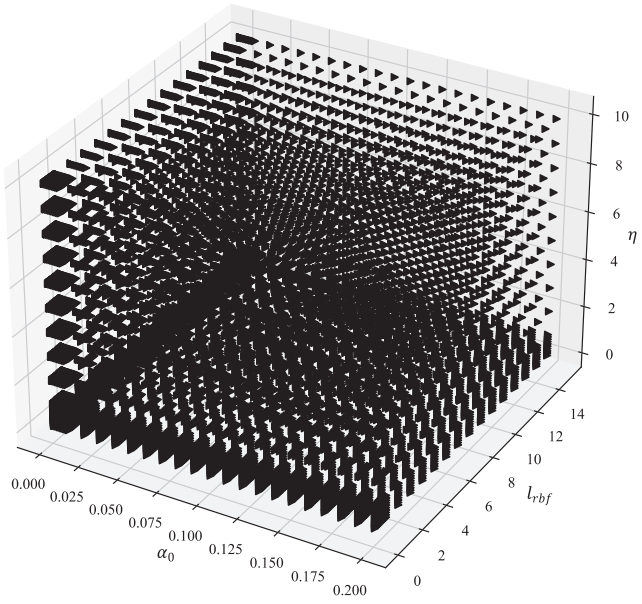

Fig. 5. Hyper-parameters searching grids.

(4) Cross-validation. The shuffle-split technique was used for cross-validation of the model, and in this work, such process was packaged with grid searching [13].

4. Results and discussion

In this section, the numerical results will be progressively illustrated from three perspectives. In Section 4.1, the process of feature selection and model training will be presented and discussed. In Section 4.2, the performance of the maximum lift-to-drag ratio predictions over a group of airfoils (i.e. the dataset) will be comparatively analyzed, and this section has two branches: one for point estimate accuracy and the other is about the interval prediction accuracy measured by the mean standard deviation. Furthermore, in Section 4.3, there is a case study on a single airfoil to demonstrate how to apply this work in real application.

4.1. Feature selection & model training

Based on the proposed framework, the processes of feature selection (Section 4.1.1) and model training (Section 4.1.2) are first illustrated in this part. Also, in Section 4.1.3, the computation time will be addressed.

4.1.1. Random forest filter: Sorted feature importance

The hyper-parameters for RF structure are shown in Table 4. For every single regression tree among the forest, its depth has been expanded as far as possible, until that the minimum split number equals to 2 [15]. Moreover, a relative large number ofN_t and$N_{shuffle}$ have been set to enhance the robustness of the whole forest.

The importance rank of the geometric features of the feature pool is plotted in Fig. 6. As the chart shows, the values of the max thickness and the max camber location are considerably higher than any other features. Noticeably, this is perfectly consistent with the naming principle

Table 4
Hyper-parameters [13] setting of RF algorithm.

Symbol	N_t	$N_{shuffle}$	S_{split}
Definition	Number of trees in the forest	Times of re-shuffle to compute the feature importance	Minimum number of samples required to split an internal node
Value	200	100	2

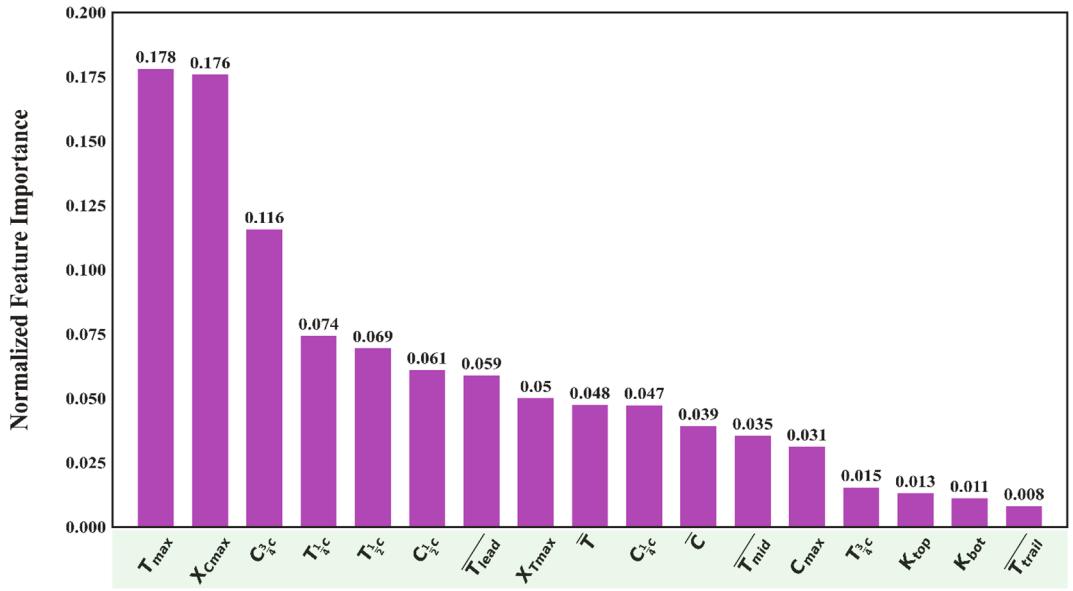

Fig. 6. Normalized feature importance of the hybrid feature pool through RF model.

of the well-known NACA 5–digit airfoils (also a part of the dataset) [39], which not only statistically but also mechanically explained their significance for maximum lift-to-drag ratio prediction.

Meanwhile, according to Fig. 6, control-point based features also have a large importance, where the 3/4 camber, 1/4 thickness, and mid-section values occupied the rank from third to sixth. This outcome can partially attribute to that once these points are determined, the overall outline of the airfoils will be formed. Given the current dataset has excluded airfoils with sharp concaves in pre-treatment, their aero-performance will hopefully be well-forecasted through such variables.

On the reverse side, the curvatures on topmost and bottommost surfaces have indicated a minor correlation to the max C_l/C_d output, which is contradicted to some of the design parameters in PARSEC [20]. A possible reason for this is that compared with the thickness and camber, these curvatures are even detailed geometric values for the profile. Hence, given such a large distribution of airfoils at present, these parameters may pose an inferior correlation on aero-performance than they did in airfoil set where profiles are generally more similar to each other.

4.1.2. Training process of sequential forward selection-Gaussian process regression

According to the above feature significance, the SFS-GPR training process was conducted. The "max thickness", which occupied the largest feature importance, was used to initialize the GPR model. Through the grid search technique, the main model parameters are shown in Table 5 as below. The first three parameters were used to construct the kernel function of GPR and the last parameter defined the weight item in the target function of SFS (Formula 7).

Fig. 7 illustrates the changes of the objective function during the SFS-GPR process on training airfoils group, where the dot means that the corresponding feature was preserved while triangle denotes the deprecation. As it can be found, most of the high-rank features were preserved, except for the max thickness position. The last three features were all deprecated, since appending neither of them as the feature could additionally enhance the overall performance. As a comparison, the corresponding changes on testing airfoils are shown in Fig. 8. It should be noted that these samples had never been exposed to the model during the SFS-GPR process.

In Figs. 8 and 9, a similar tendency can be found between the two curves. The deprecated features that lowered down the performance on training airfoils also reduced the objective function value on test samples. At the same time, a "fast followed by slow" trend of increase can be observed in curves for both training and testing samples, indicating that the degree of improvement of GPR accuracy would gradually drop as its feature dimension increases. The main difference between the curves is on their beginning part, where the objective function value is particularly low on testing samples. One reason for this may be the number of test samples were too small to fit the kernel while the input features has such low dimension. Afterwards, when more features participated into the input set, the objective function values on training and testing set became closer.

Table 5
Parameters [14] setting of SFS-GPR.

Symbol	α_0	l_{rbf}	η	λ_σ
Value	0.001	0.6	0.1	0.05

4.1.3. Computational cost

For machine learning approaches, although it is quite rapid (less than a second) to make the prediction when the model was established, it is important to address the training cost of the model. The whole model was trained through Python 3.8 platform on a domestic server: Inter (R) Xeon (R) CPU E5-2673 v4 @ 2.30 GHz, and the whole code was developed based on scikit-learn packages [13]. The computation time is listed as follows in Table 6.

As shown in Table. 6, based on the current dataset and the proposed framework, apart from the neglectable time for prediction, the total training time is less than 7 min. Hence, it is acceptable for further model update.

4.2. Prediction performance on airfoil groups

In this section, aiming on the training and testing sets of airfoils, the prediction performance from the proposed model will be discussed. The discussion will be presented in two aspects: the regression accuracy for point estimate (Section 4.2.1) and the concentration degree for interval estimate (Section 4.2.2). Furthermore, the sensitivity of the weight item λ_σ on prediction results will be presented in Section 4.2.3.

4.2.1. Regression accuracy for point estimate

To assess the point-prediction performance on regression task, the

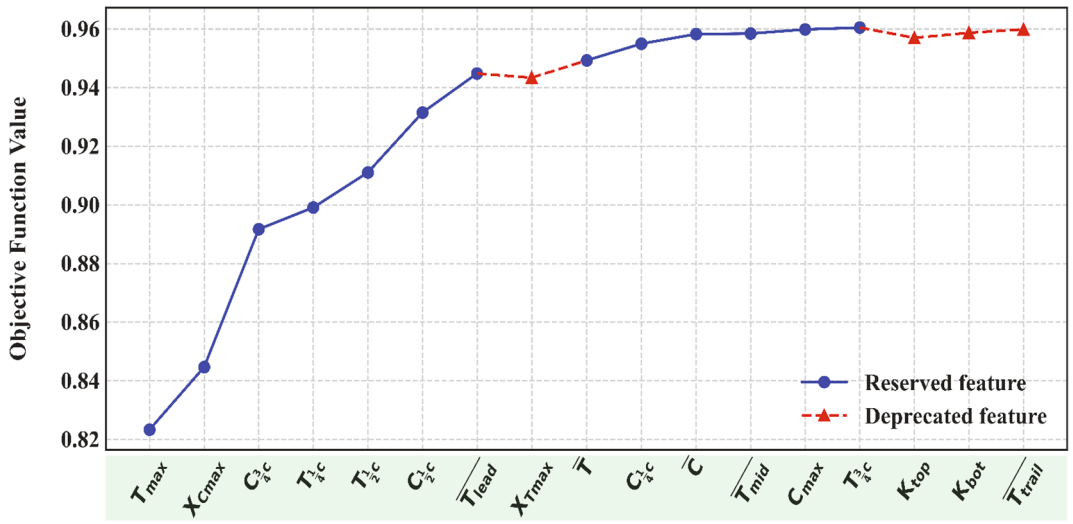

Fig. 7. Changes of objective function during SFS-GPR process (on training airfoils).

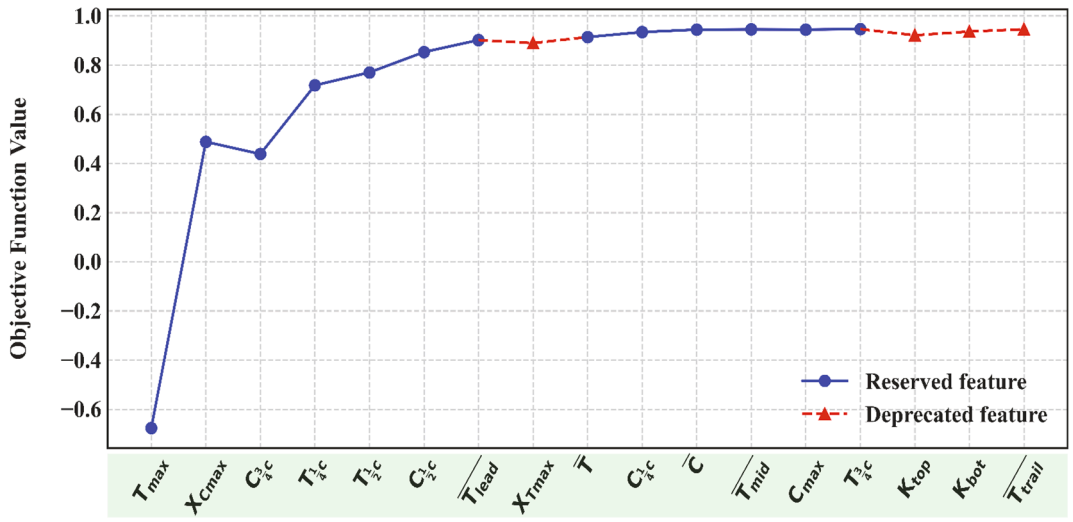

Fig. 8. Changes of objective function during SFS-GPR process (on testing airfoils).

EVS values (Formula 12) on the target airfoil groups are used in this work. First, the scatter-plots between the observed maximum lift-to-drag ratios and their point estimates for both training and testing airfoils are shown in Fig. 9. In addition, the scatter-plots and EVS values of the full-feature GPR model and the ANN model are also shown aside as comparison.

As shown in Fig. 9 (a) to (c), on training airfoil group, the point-prediction performances among different models are comparable, with their EVS values all above 0.9. For GPR model with selected feature set, the value is 0.952, which is slightly lower than all-feature GPR model (0.961) but appreciably higher than the ANN model (0.911). The difference of the model performance can also be told from the scatter distribution around the isoline. The width of the current scatter-bar is mildly broader than all-feature GPR model while narrower than that of the ANN. It should be noted that, this ANN model has been carefully designed, adopting a 4-hidden-layer structure and over 650 parameters updated by Adam [40] optimizer at a learning rate of $5e^{-4}$.

However, for the testing airfoils (Fig. 9 (d) to (f)), the point-prediction performance of the proposed method has an evident superiority over the other models. As in Fig. 9 (e), a large overfitting can be found in all-feature GPR model, whose EVS value for testing set dramatically drops from 0.961 to 0.858. Similar in ANN model, the accuracy matrix drops to 0.879 on testing airfoil group. While by using the GPR model with selected feature set, the prediction accuracy maintained itself at a high value of 0.941.

In addition, Fig. 10 shows the variation of predicted maximum lift-to-drag ratio on testing set together with the UIUC observations [30]. As it can be found, the curves of the current model (black line) can better fit the changes of the observations (red line) almost in every testing airfoil. While using all-feature GPR and ANN models, some large prediction bias can be found, such as the marked airfoil samples within blue rectangles of Fig. 10. This indicates that the proposed algorithm can reflect a closer relationship between the airfoil geometry characters (i.e. the model inputs) and the target maximum lift-to-drag coefficient (i.e. the model output).

4.2.2. *Regression accuracy for interval estimate*

Apart from the point estimate for the maximum sliding ratio by using the mean predictions from GPR, the standard deviation (std.) is another important outcome and a performance criterion when using such model. It not only reflects to what extent the user could believe the point estimate, but also contains a confidence interval to offer more design reference from the model. In this section, aiming at groups of airfoils (i.e. training set and testing set), this parameter is assessed first through

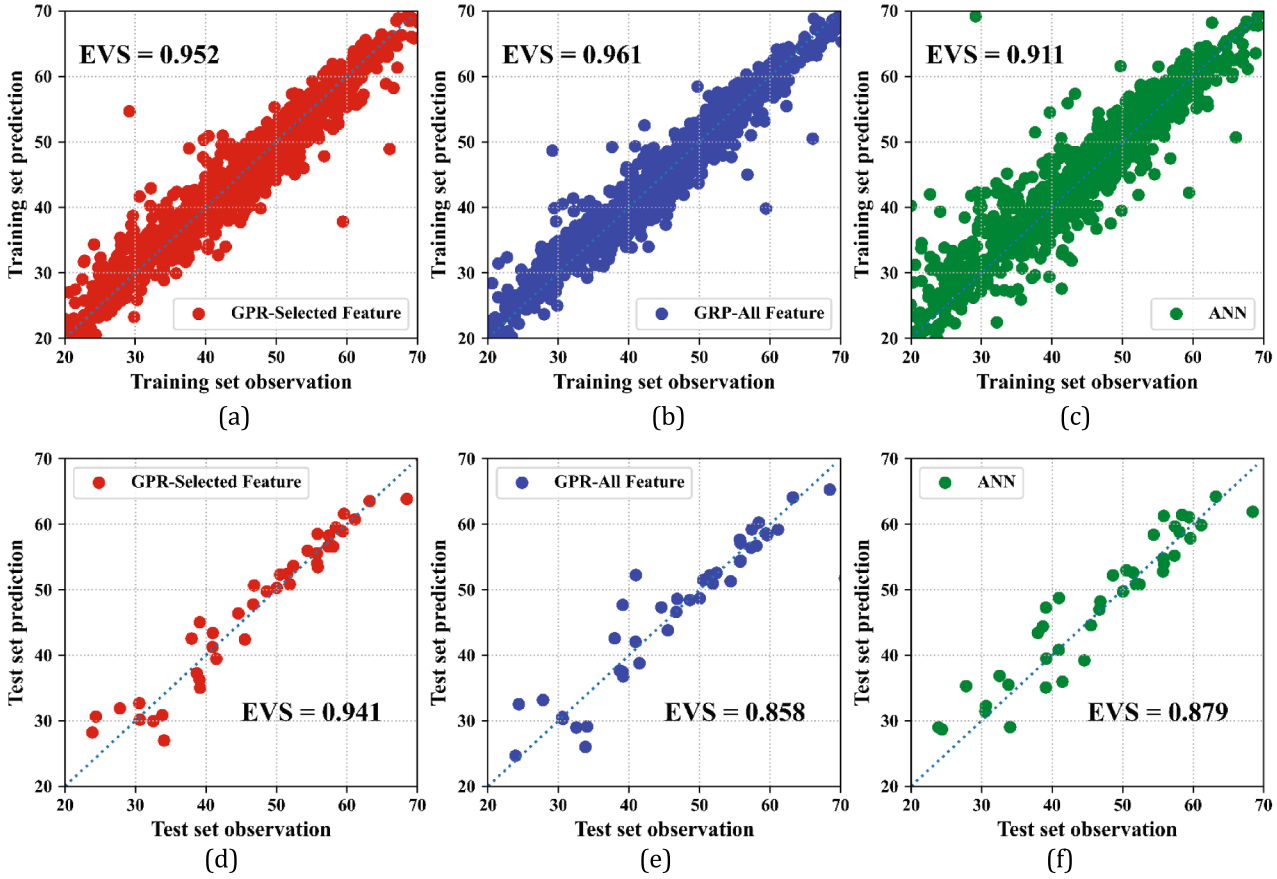

Fig. 9. Performance comparison between current model and other models for the point estimate of max C_l/C_d on both training and testing airfoil groups.

Table 6
Computational time during model training and prediction.

Procedure		Computational time
RF (with settings in Table 4)		2 min 34 s
SFS-GPR (with settings in Table 5)	Initialization	9.26 s
	Average time per step	15.22 s
	Total time	4 min 19 s
Total training time		6 min 53 s
Prediction time		less than a second

their overall statistical values.

As shown in Fig. 11, the mean and median std. of on both training and testing airfoil sets decreased after the feature selection process. For the median std., the charts are similar between training and testing set, where the current std. is about 20% cut-down from the all-feature GPR result. However, impressively, for the mean std. on testing airfoils, the value dramatically dropped from 1.502 to 0.894. Meanwhile, the difference between the mean std. on training and testing airfoils have largely shrunken, from 0.460 to 0.063. This implies that the feature selection can make the model more generalized to largely improve the prediction accuracy and the confidence for unseen samples.

Fig. 12 shows a comparison of the detailed interval predictions on testing airfoils by using GPR models without (Fig. 12(a)) or with (Fig. 12 (b)) current feature selection. As it can be found, no matter using all-feature or selected-feature as input, the observation value could mostly drop into the range with 98.7% confidence level ($\pm 2.5\sigma$) around the point-prediction. However, the proposed GPR model can do better to offer a more accurate interval prediction form the following aspects.

First, if its point estimate is considerably different from the all-feature GPR model (marked by triangles, Fig. 12), its value performs closer to the observation, which gives a more precise range of interval prediction. Second, if its point estimate is similar to the all-feature GPR

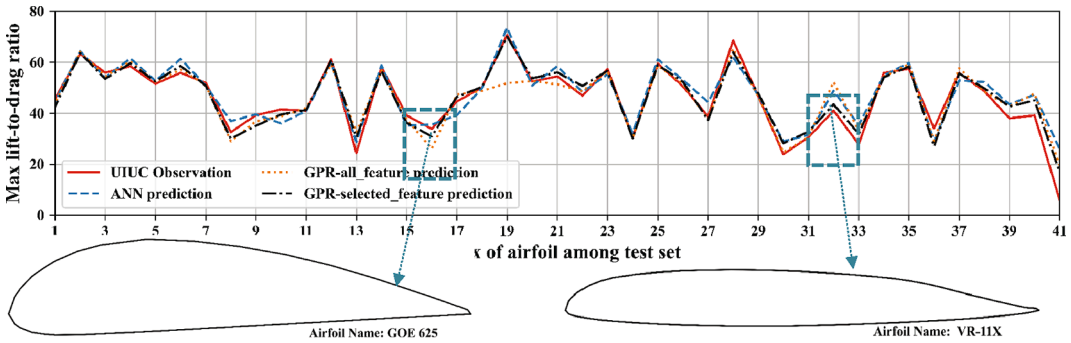

Fig. 10. Comparison of the max lift-to-drag ratio prediction variation on testing airfoils using different models [30].

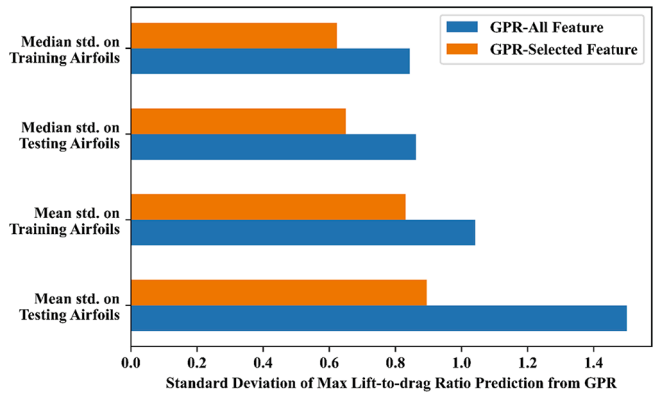

Fig. 11. Comparison of the median and mean std. of GPR predictions between all-feature model and current model on both training and testing set.

model (marked within dashed circles, Fig. 12), the prediction range under the same interval prediction has also largely shrunken but also successfully covered the observation. In another word, by using the current GPR model, the same prediction interval would have a higher confidence level than naïve GPR model if its margin has just covered the real observations. These two points together confirmed the effectiveness of the current feature selection algorithm and shows the priority of the proposed model.

4.2.3. Sensitivity of λ_σ

As a key parameter for current SFS-GPR process that controls the weight allocation between two accuracy criteria (Formula 12, 13) from "point" and "interval" views, the value of λ_σ in Formula 7 has been carefully tuned in this work. In this section, its sensitivity on final model performance will be further discussed.

The current λ_σ is denoted asλ_{σ_0}, which equals to 0.05 (Table 5). To investigate the parameter influence, target functions with different λ_σ values, equaling to 4, 2, 1.5, 1, 0.5 and 0.25 times ofλ_{σ_0}, were respectively conducted. According to Formula 7, the larger value ofλ_σ, the greater the weight of assessing from the interval view would be.

In Fig. 13, the final EVS on training samples first gradually decreases as the λ_σ keeps increasing from $0.25\lambda_{\sigma_0}$ toλ_{σ_0}, and stays flat afterwards. However on the testing set, despite the EVS value experiences an initial growing, reaching the maximum at 0.941 when λ_σ is in the range of 1.0 to 1.5 times ofλ_{σ_0}, it immediately drops to 0.901 when the parameter continues enlarging to $2.0\lambda_{\sigma_0}$ or more. Therefore, the present λ_{σ_0} gives a local peak value assessed from the point estimate.

In Fig. 14, the final mean standard deviations among training and testing samples are provided under various λ_σ values. As it can be found, the dispersion degrees of the prediction are relatively high for the boundary values of λ_σ, where the std. values for testing airfoils are larger than one. However, such value is gradually becoming small when λ_σ approaching the current value of λ_{σ_0}, representing that the distribution of the predictions is more concentrated and the predicting range with the same confidence interval is narrowing. Hence, this gives another local optima at the current λ_{σ_0} from the view of interval prediction.

4.3. Case study on a single airfoil

The above discussion has presented the merits of this work on a group of airfoils. In order to illustrate the routine of the framework for real application of wind turbine design, a case study on a specific airfoil, NACA 64_3–218, is presented in this section. Such airfoil belongs to the NACA 6-digit series and it was considered as a potential candidate for large wind turbine blades [41].

4.3.1. Instantiation

Given the geometry of NACA 64_3–218, through the pre-processing, the airfoil is firstly normalized from the raw coordinates along its outline (Section 2.2). In Fig. 15, the prepared profile is plotted, which has a unit chord length and continuous shape curves.

Then, based on the discrete coordinates from Fig. 15, the hybrid feature pool (Table 2) is formed using the Formula 1–3. Table 7 shows all 17 features corresponding to the NACA 64_3–218 airfoil. Its max thickness equals to 17.95% and is located at 34.92% chord from the leading edge. The max camber of the airfoil is only 1.09%, reflecting an overall symmetricity about the zero x-cord.

Next, based on the hybrid selection methods, including the RF in Section 3.2.1 and SFS method in Section 3.2.2, the redundant features of Table 6 are eliminated. According to the results from Fig. 7, the deprecated features are$x_{T_{max}}$,K_{top},K_{bot} and $\overline{T_{trail}}$. Hence, the remaining 13 features are used for the final regression task.

As shown in Fig. 16, predicted by current GPR model, the estimation of its maximum lift-to-drag ratio is a Gaussian distribution. To be the

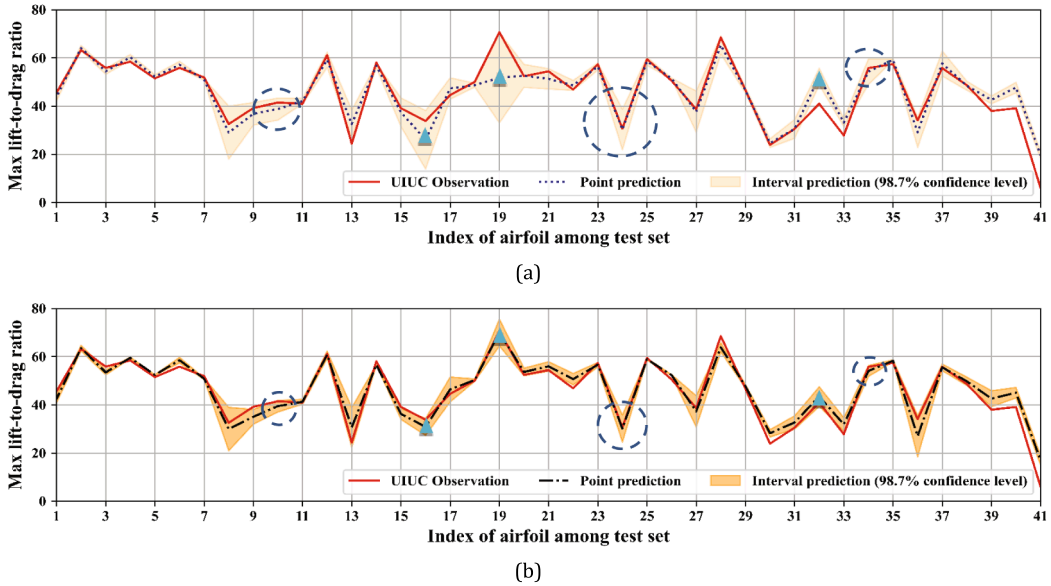

Fig. 12. Comparison of interval predictions of max lift-to-drag ratio on testing airfoils using: (a) non-selected features (b) selected features.

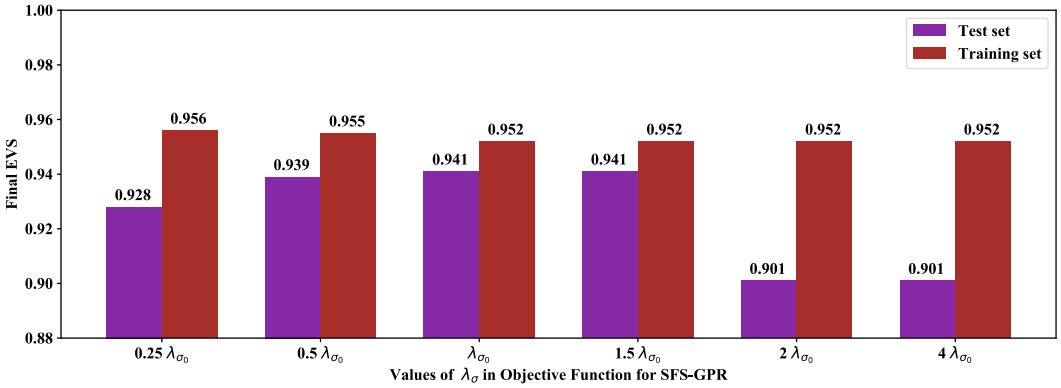

Fig. 13. Comparison of the final EVS using target functions with different.λ_σ

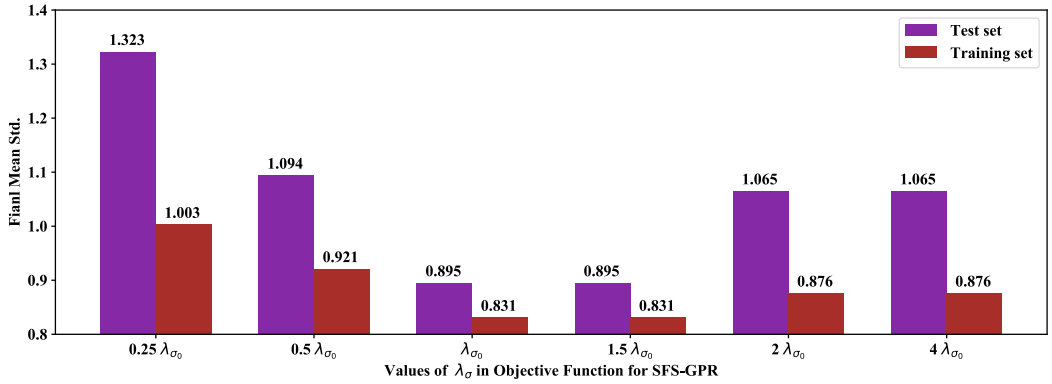

Fig. 14. Comparison of the final mean std. using target functions with different.λ_σ

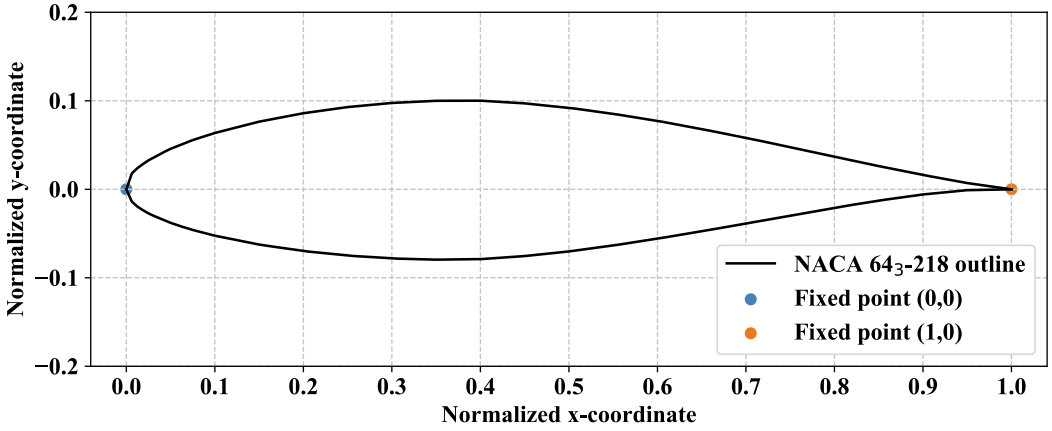

Fig. 15. Normalized outline of NACA 64$_3$–218 for case study [30].

most straightforward usage, the designer could directly adopt the mean value μ of 39.45 as the point estimate. However, when making interval predictions, the other appealing outcome of GPR, standard deviationσ, would be taken into consideration (Section 4.3.2).

4.3.2. Annotation for interval prediction

Different from other machine learning algorithms, such as neural network that can only provide a single-point estimation [42], the GPR has its unique characteristic of interval prediction according to the empirical rule for Gaussian distribution [14].

As an annotation for the case of NACA 64$_3$-218 airfoil, the results from Fig. 16 told us that if the current model works: based on the current dataset and feature pool, there is 68.26% confidence for the given NACA 64$_3$–218 airfoil, its max C_l/C_d ratio under Reynolds number of 100,000 will fall into the interval from 38.64 to 40.26 ($\mu \pm \sigma$). For such prediction interval, the margin error is merely 0.81. As the prediction becomes more conservative, the confidence level is higher and the margin error is becoming larger. Table 8 shows other common intervals together with the referred observation from UIUC test [30].

According to Table 8, under the Reynolds number of 100,000, the referred maximum lift-to-drag ratio of NACA 64$_3$–218 is equal to 41.47 If the designer purely adopted the point estimate of 39.45, there would be a downward bias of 2.02. Nevertheless, by gradually considering different levels of confidence interval from low to high, one could take

Table 7
Geometric Features of NACA 64_3–218 for case study.

T_{max}	0.1795	K_{top}	1.0797
$x_{T_{max}}$	0.3492	K_{bot}	5.6615
C_{max}	0.0109	$T_{c/4}$	0.1678
$x_{C_{max}}$	0.5000	$T_{c/2}$	0.1622
$\overline{T}$	0.1115	$T_{3c/4}$	0.0773
$\overline{C}$	0.0079	$C_{c/4}$	0.0090
$\overline{T_{lead}}$	0.1165	$C_{c/2}$	0.0109
$\overline{T_{mid}}$	0.1483	$C_{3c/4}$	0.0090
$\overline{T_{trail}}$	0.0321		

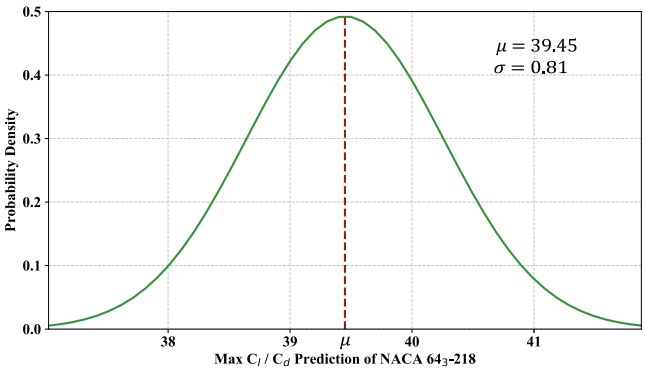

Fig. 16. Gaussian distributed max C_l/C_d prediction for NACA 64_3–218 airfoil under $Re = 100{,}000$ with interval range from $\mu - 3\sigma$ to.$\mu + 3\sigma$.

Table 8
Summary of GPR predictions for the case study on NACA 64_3–218 airfoil.

Point estimateμ		39.45	
Standard deviationσ		0.81	
Referred observation [30]		41.47	
Interval range	Prediction range	Confidence	Margin error
$\mu \pm \sigma$	[38.64, 40.26]	68.26%	0.81
$\mu \pm 1.5\sigma$	[38.23, 40.66]	86.63%	1.22
$\mu \pm 2\sigma$	[37.83, 41.07]	95.44%	1.62
$\mu \pm 2.5\sigma$	[37.43, 41.48]	98.75%	2.03
$\mu \pm 3\sigma$	[37.02, 41.88]	99.73%	2.43

precautions against such risk in advance. In this case, the margin error 2.03 of 98.75% confidence has successfully covered such bias, offering a conservative but also accurate interval prediction [42].

Further in applications for real wind turbine system, the above information could be helpful in its early design stage, such as in determining the range of designed tip speed ratio and consequently, the specified rotational speed of the rotor [2] for the target wind turbine that adopts such airfoil as its blade shape. Moreover, by choosing desired confidence level that takes potential risks into consideration based on the historical data, the corresponding interval prediction can additionally offer a more conservative and accurate reference for the designed working range.

5. Conclusion and future work

In this paper, an original machine learning framework for airfoil maximum lift-to-drag ratio prediction is proposed. When applied to an airfoil with specific geometry, the model was able to provide accurate prediction results from both point and interval perspectives under a turbulent Reynolds number of 100,000.

The framework establishment process, numerical measurements, multi-aspect comparison and the parameter sensitivity analysis were enough to prove that the model here presented is better developed than parallel models. Based on this, several conclusions can be summarized.

First, the proposed model has originally synthesized the geometric features types extracted from previous studies to form a hybrid potential feature pool. An original hybrid feature selection was then conducted to dig the sub feature set of the hybrid pool for regression. Second, the proposed framework was trained and tested on the widely used UIUC dataset [30], known for providing a high-precision prediction for both assessments: point and interval ones. For the group of airfoils that underwent testing, the explained variance score of the current model reached 0.941 and the mean standard deviation was only 0.895, which are considered impressive results.

Thirdly, the feasibility of this work and its easy-to-implement features were confirmed through a case study based on a single airfoil. All of the potential applications of the novel model in this specific engineering field usage have been clearly presented. At the same time, the unique probabilistic prediction obtained from the GPR results has also been annotated in detail to illustrate its value for engineering reference.

This work leaves room for other scholars in the field to make improvements and new related discoveries. The following aspects require content can be further investigation: (1).The airfoil dataset could be further updated or expanded to make the new model more sophisticated; (2). Different Reynolds numbers may be considered and their influence on prediction performance further analyzed; (3). More features of the airfoil could be added into the hybrid feature pool; (4). The SFS-GPR procedure could also be improved to become more flexible when conducting the process of feature selection, so it does not strictly rely on the impotence rank provided by Random Forest algorithm.

CRediT authorship contribution statement

Yaoran Chen: Investigation, Methodology, Software, Data curation. **Zhikun Dong:** Methodology. **Jie Su:** Methodology, Revision. **Yan Wang:** Methodology. **Zhaolong Han:** Methodology, Revision, Supervision. **Dai Zhou:** Conceptualization, Supervision. **Yongsheng Zhao:** Supervision. **Yan Bao:** Supervision.

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.

Acknowledgements

The financial supports from the National Natural Science Foundation of China (Nos. 42076210, 51879160, 51809170, U19B2013, and 11772193), and Innovation Program of Shanghai Municipal Education Commission (No.:2019-01-07-00-02-E00066) are gratefully acknowledged. This research is also sponsored in part by Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (No. TP2017013), “Shuguang Program” supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission (No.19SG10) and the Oceanic Interdisciplinary Program of Shanghai Jiao Tong University (No.SL2020PT201).

References

[1] Ayodele TR, Jimoh A, Munda JL, Agee J. Challenges of grid integration of wind power on power system grid integrity: A review. World 2020;3(6):618–26.
[2] Sogukpinar H, Bozkurt I. Numerical calculation of aerodynamics wind turbine blade S809 airfoil and comparison of theoretical calculations with experimental measurements and confirming with NREL data. In: Conference Numerical calculation of aerodynamics wind turbine blade S809 airfoil and comparison of

theoretical calculations with experimental measurements and confirming with NREL data, vol. 1935. AIP Publishing LLC, p. 020004.
[3] Hann R, Hearst RJ, Sætran LR, Bracchi T. Experimental and numerical icing penalties of an S826 airfoil at low Reynolds numbers. Aerospace 2020;7(4):46.
[4] Sheldahl RE, Klimas PC, Feltz LV. Aerodynamic performance of a 5-metre-diameter Darrieus turbine with extruded aluminum NACA-0015 blades. National Technical Information Service USA 1980.
[5] Apsley D, Stansby P. Unsteady thrust on an oscillating wind turbine: Comparison of blade-element momentum theory with actuator-line CFD. Journal of Fluids and Structures 2020;98:103141.
[6] Ilie M, McAfee J. Aerodynamics and design of vertical-axis wind turbine; numerical studies using LES and IDDES. Conference Aerodynamics and design of vertical-axis wind turbine; numerical studies using LES and IDDES. p. 2693.
[7] He J, Jin X, Xie S, Cao L, Wang Y, Lin Y, et al. CFD modeling of varying complexity for aerodynamic analysis of H-vertical axis wind turbines. Renewable Energy 2020;145:2658–70.
[8] Hui X, Bai J, Wang H, Zhang Y. Fast pressure distribution prediction of airfoils using deep learning. Aerospace Science and Technology 2020;105:105949.
[9] Oh S. Comparison of a response surface method and artificial neural network in predicting the aerodynamic performance of a wind turbine airfoil and its optimization. Applied Sciences 2020;10(18):6277.
[10] Wen H, Sang S, Qiu C, Du X, Zhu X, Shi Q. A new optimization method of wind turbine airfoil performance based on Bessel equation and GABP artificial neural network. Energy 2019;187:116106.
[11] Viquerat J, Hachem E. A supervised neural network for drag prediction of arbitrary 2D shapes in laminar flows at low Reynolds number. Computers & Fluids 2020;210:104645.
[12] Chen Y, Dong Z, Wang Y, Su J, Han Z, Zhou D, et al. Short-term wind speed predicting framework based on EEMD-GA-LSTM method under large scaled wind history. Energy Conversion and Management 2021;227:113559.
[13] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research 2011;12:2825–30.
[14] Rasmussen CE. Gaussian processes in machine learning. Conference Gaussian processes in machine learning. Springer, p. 63–71.
[15] Liu X, Zhu Q, Lu H. Modeling multiresponse surfaces for airfoil design with multiple-output-Gaussian-process regression. Journal of Aircraft 2014;51(3):740–7.
[16] Cai H, Jia X, Feng J, Li W, Hsu Y-M, Lee J. Gaussian Process Regression for numerical wind speed prediction enhancement. Renewable Energy 2020;146:2112–23.
[17] Mathioudakis N, Panagiotou P, Kaparos P, Yakinthos K. A Genetic Algorithm based Method for the Airfoil Optimization of a Tactical Blended-Wing-Body UAV. Conference A Genetic Algorithm based Method for the Airfoil Optimization of a Tactical Blended-Wing-Body UAV. IEEE, p. 1582-9.
[18] Haryanto I, Utomo TS, Sinaga N, Rosalia CA, Putra AP. Optimization of maximum lift to drag ratio on airfoil design based on artificial neural network utilizing genetic algorithm. Conference Optimization of maximum lift to drag ratio on airfoil design based on artificial neural network utilizing genetic algorithm, vol. 493. Trans Tech Publ, p. 123-8.
[19] Ma N, Lei H, Han Z, Zhou D, Bao Y, Zhang K, et al. Airfoil optimization to improve power performance of a high-solidity vertical axis wind turbine at a moderate tip speed ratio. Energy 2018;150:236–52.
[20] Sobieczky H. Parametric airfoils and wings. In: Recent development of aerodynamic design methodologies: Springer; 1999. p. 71-87.
[21] Chen J, Pan X, Wang C, Hu G, Xu H, Liu P. Airfoil parameterization evaluation based on a modified PARASEC method for a H-Darrious rotor. Energy 2019;187:115910.
[22] Hansen T. Airfoil optimization for wind turbine application. Wind Energy 2018;21(7):502–14.
[23] Kostas K, Amiralin A, Sagimbayev S, Massalov T, Kalel Y, Politis C. Parametric model for the reconstruction and representation of hydrofoils and airfoils. Ocean Engineering 2020;199:107020.
[24] Joukowsky N. Über die konturen der Tragflächen der Drachenflieger. Zeitschrift für Flugtechnik und Motorluftschiffahrt 1910;1(22):281–5.
[25] Chen H, He L, Qian W, Wang S. Multiple aerodynamic coefficient prediction of airfoils using a convolutional neural network. Symmetry 2020;12(4):544.
[26] Yu B, Xie L, Wang F. An improved deep convolutional neural network to predict airfoil lift coefficient. In: Conference An Improved Deep Convolutional Neural Network to Predict Airfoil Lift Coefficient. Springer, p. 275-86.
[27] Bellman M, Straccia J, Morgan B, Maschmeyer K, Agarwal R. Improving genetic algorithm efficiency with an artificial neural network for optimization of low Reynolds number airfoils. In: Conference Improving genetic algorithm efficiency with an artificial neural network for optimization of low Reynolds number airfoils. p. 1096.
[28] Andrés E, Salcedo-Sanz S, Monge F, Pérez-Bellido A. Efficient aerodynamic design through evolutionary programming and support vector regression algorithms. Expert Systems with Applications 2012;39(12):10700–8.
[29] Kar M, Dewangan L. Univariate feature selection techniques for classification of epileptic EEG Signals. In: Advances in Biomedical Engineering and Technology: Springer; 2021. p. 345-65.
[30] Selig MS. UIUC airfoil data site; 1996.
[31] Noyes C, Loth E, Martin D, Johnson K, Ananda G, Selig M. Extreme-scale load-aligning rotor: To hinge or not to hinge? Applied Energy 2020;257:113985.
[32] Karthikeyan N, Murugavel KK, Kumar SA, Rajakumar S. Review of aerodynamic developments on small horizontal axis wind turbine blade. Renewable and Sustainable Energy Reviews 2015;42:801–22.
[33] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC Press; 1984.
[34] Breiman L. Random forests. Machine Learning 2001;45(1):5–32.
[35] Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence 1997;97(1-2):273–324.
[36] Bermejo P, Gámez JA, Puerta JM. Incremental wrapper-based subset selection with replacement: An advantageous alternative to sequential forward selection. In: Conference Incremental wrapper-based subset selection with replacement: An advantageous alternative to sequential forward selection. IEEE, p. 367-74.
[37] Myung IJ. Tutorial on maximum likelihood estimation. Journal of Mathematical Psychology 2003;47(1):90–100.
[38] Hu G, Kwok K. Predicting wind pressures around circular cylinders using machine learning techniques. Journal of Wind Engineering and Industrial Aerodynamics 2020;198:104099.
[39] Abbott IH, Von Doenhoff AE, Stivers Jr LS. Summary of airfoil data 1945.
[40] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980; 2014.
[41] Timmer W. An overview of NACA 6-digit airfoil series characteristics with reference to airfoils for large wind turbine blades. In: Conference An overview of NACA 6-digit airfoil series characteristics with reference to airfoils for large wind turbine blades. p. 268.
[42] Sarajcev P, Jakus D, Mudnic E. Gaussian process regression modeling of wind turbines lightning incidence with LLS information. Renewable Energy 2020;146:1221–31.