



Machine Learning Surrogate Modelling of Aerodynamic Efficiency (Cl/Cd)

Arif Ali Rind¹, Muhammad Ramzan Luhur², Muhammad Kashif Abbasi³, Aijaz Abbasi⁴, Sajjad Bhangwar⁵

¹ Mechanical Department, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Sindh, Pakistan

Email: arifali@quest.edu.pk

² Mechanical Department, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Sindh, Pakistan.

Email: luhur@quest.edu.pk

³ Industrial and Manufacturing Engineering, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Sindh, Pakistan.

Email: muhammadkashif@quest.edu.pk

⁴ Mechanical Department, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Sindh, Pakistan.

Email: aijaz-abbasi@quest.edu.pk

⁵ Mechanical Department, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Sindh, Pakistan.

Email: sajjadbhangwar@quest.edu.pk

ARTICLE INFO

Submission:

August 28, 2026

Revised:

September 09, 2026

Accepted:

September 25, 2026

Available Online:

October 05, 2026

Keywords:

airfoils, SG6043,
aerodynamic efficiency,

Corresponding author:

Muhammad Ramzan Luhur

Email: luhur@quest.edu.pk

ABSTRACT

High-fidelity aerodynamic evaluation of airfoils is computationally expensive for multidimensional design and operating-condition studies. This study develops a comparative and explainable machine-learning surrogate framework to predict the aerodynamic efficiency lift to drag ratio (CL/CD) of the SG6043 airfoil family and its modified variants. A dataset of 924 observations covering 26 airfoil configurations was developed over Reynolds numbers of ($1.0 \times 10^5 - 6.0 \times 10^5$), angles of attack of $2.5-10^\circ$, thickness ratios of $5.00 - 10.02\%$, and camber ratios of $5.00 - 9.00\%$. Multiple Linear Regression (MLR), Random Forest (RF), Artificial Neural Network (ANN), and Extreme Gradient Boosting (XGBoost) were evaluated using an 80:20 train- test split. The ANN, comprising two hidden layers of 100 and 50 neurons, achieved the best performance with ($R^2 = 0.9967$), RMSE = 2.246, MAE = 1.457, and MAPE = 1.68%, compared with ($R^2 = 0.8462$) and MAPE = 12.61% for MLR. RF and XGBoost achieved ($R^2 = 0.9929$) and 0.9946, respectively. SHAP analysis identified Reynolds number as the dominant predictor, followed by camber, angle of attack, and thickness. The ANN exhibited negligible prediction bias (0.179) and 94.59% empirical coverage for the nominal 95% prediction interval. The results demonstrate that ANN-based surrogate modelling can accurately capture the nonlinear (CL/CD) response within the investigated SG6043 design space, providing an efficient and interpretable approach for rapid airfoil screening, sensitivity analysis, and aerodynamic optimization.

Introduction

A Small horizontal-axis wind turbines generally operate at low Reynolds numbers, where viscous effects, boundary-layer transition and flow separation strongly influence aerodynamic performance. Giguère and Selig [1] developed an extensive experimental database for low-Reynolds-number airfoils and highlighted the importance of airfoil selection for small wind turbines. Their subsequent development of dedicated SG60XX profiles demonstrated the potential of specially designed airfoils to improve lift-to-drag performance over Reynolds numbers of approximately Literature Review Small horizontal-axis wind turbines generally operate at low Reynolds numbers, where viscous effects, boundary-layer transition and flow separation strongly influence aerodynamic performance. Giguère and Selig [1] developed an extensive experimental database for low-Reynolds-number airfoils and highlighted the importance of airfoil selection for small wind turbines. Their subsequent development of dedicated SG60XX

profiles demonstrated the potential of specially designed airfoils to improve lift-to-drag performance over Reynolds numbers of approximately [2]. Selig and McGranahan [3] further provided experimental lift, drag and moment data for several small-wind-turbine airfoils at Reynolds numbers between confirming the strong dependence of aerodynamic behaviour on operating conditions. Singh et al. [4] also demonstrated the importance of geometric design in maintaining attached flow and improving low-Reynolds-number performance. These studies establish Reynolds number, angle of attack and airfoil geometry as key variables governing aerodynamic efficiency. The SG6043 profile has subsequently received considerable attention for small horizontal-axis wind turbines. Shin and Kim [5] evaluated SG6043 at low Reynolds number and reported favourable aerodynamic performance for small-wind-turbine applications. Noronha and Krishna [6] compared several candidate airfoils using QBlade and CFD over different Reynolds numbers and angles of attack and identified SG6043 as a suitable profile for low-speed small wind turbines. More recently, Davari et al. [7] investigated SG6043 over Reynolds numbers from and examined modifications based on thickness-to-camber ratio, demonstrating that geometric modification can substantially change aerodynamic efficiency. Rind et al. [8] further investigated systematic SG6043 modifications for small horizontal-axis wind turbines, reinforcing the sensitivity of to airfoil geometry and operating conditions. These studies provide a strong aerodynamic basis for investigating systematically modified SG6043 profiles. Computational tools have substantially reduced the cost of aerodynamic analysis compared with repeated experimental testing. XFOIL, developed by Drela [9], combines panel-method analysis with viscous boundary-layer and transition modelling and has been widely used for low-Reynolds-number airfoil analysis. QBlade, developed by Marten et al. [10], extends airfoil analysis to wind-turbine blade and rotor simulations. However, evaluating a large number of geometries across multiple Reynolds numbers and angles of attack can still require considerable computational effort. Surrogate modelling provides an alternative by learning the aerodynamic response from an existing database and subsequently replacing repeated aerodynamic evaluations with inexpensive predictions. Forrester et al. [11] and Queipo et al. [12] established the broader engineering framework for surrogate-based modelling and optimization, while Chung and Alonso [13] demonstrated its application to aerodynamic airfoil optimization. The recent development of machine learning has expanded the capabilities of aerodynamic surrogate modelling. Du et al. [14] developed neural-network-based surrogate models for rapid airfoil design optimization and demonstrated that aerodynamic quantities could be predicted sufficiently accurately to support rapid optimization. Li et al. [15] reviewed machine-learning applications in aerodynamic shape optimization and identified rapid aerodynamic analysis, compact design representations and efficient optimization as major application areas. Teimourian et al. [16] specifically compared several machine-learning algorithms for predicting airfoil lift-to-drag ratio, demonstrating the applicability of nonlinear learning algorithms to aerodynamic performance prediction. Kieszek et al. [17] similarly developed a feed-forward ANN surrogate for predicting airfoil aerodynamic coefficients and used the model for subsequent optimization. These studies indicate that machine-learning models can provide efficient alternatives to repeated aerodynamic simulations, particularly when nonlinear interactions exist between geometry and flow conditions. Different machine-learning algorithms offer complementary capabilities for surrogate modelling. Random Forest [18] provides nonlinear ensemble regression through multiple decision trees, while gradient boosting [19] and XGBoost [20] provide increasingly flexible boosted-tree formulations. Artificial neural networks provide another nonlinear approximation framework and are capable of representing complex interactions between multiple continuous and categorical variables. The present study therefore compares Multiple Linear Regression, Random Forest, ANN and XGBoost rather than assuming that a particular algorithm is universally appropriate. This comparative approach is applied to 924 observations covering 26 SG6043-family variants, Reynolds numbers from 100,000 to 600,000, angles of attack from 2.5° to 10° , and systematic variations in thickness and camber. Beyond predictive accuracy, interpretation of machine-learning models is important in aerodynamic applications. Lundberg and Lee [21] introduced SHAP as a unified framework for attributing model predictions to individual input variables. In the present study, SHAP is used to interpret the selected ANN surrogate and identify the relative contribution of Reynolds number, camber, angle of attack, thickness and airfoil identity.

Dataset Description

The dataset comprises 924 rows, each corresponding to a unique combination of airfoil geometry and flow condition, drawn from 26 distinct airfoil variants within the SG6043 family (including the baseline SG6043 profile and 25 systematically modified derivatives, denoted by thickness/camber modification codes such as SG6043M6-8, SG6043M9-5, etc.). Table 1. summarises the feature set and the prediction target.

Table 1: Feature and Target Summary

#	Variable	Type	Role	Description
1	Airfoil	Categorical (26 levels)	Feature	Airfoil geometry identifier / family variant
2	Reynolds Number	Numeric	Feature	Flow Reynolds number (100,000 – 600,000)

3	AOA (Deg)	Numeric	Feature	Angle of attack, degrees (2.5 – 10.0)
4	Thickness (%)	Numeric	Feature	Maximum thickness as % of chord (5.00 – 10.02)
5	Camber (%)	Numeric	Feature	Maximum camber as % of chord (5.00 – 9.00)
6	CL/CD	Numeric	Target	Lift-to-drag ratio (aerodynamic efficiency)

Descriptive statistics for the four numeric features and the target variable, computed over all 924 observations, are reported in Table 2. The target variable CL/CD spans a wide range (26.84 to 193.44), with a mean of 124.26 and a coefficient of variation of 32.3%, indicating substantial and well-distributed variability suitable for regression modelling. None of the numeric features exhibit missing values or duplicate feature rows, and the Airfoil categorical column contains no dominant class (36 rows for most levels, with the most frequent airfoil, SG6043, appearing 36 times out of 924).

Table 2: Descriptive Statistics of Numeric Features and Target

Variable	N	Mean	Median	SD	Min	Max	CV (%)	Skew
Reynolds Number	924	352,164.50	400,000.00	169,588.35	100,000	600,000	48.16	-0.01
AOA (Deg)	924	4.45	4.00	1.49	2.50	10.00	33.38	0.94
Thickness (%)	924	7.12	7.00	1.51	5.00	10.02	21.18	0.07
Camber (%)	924	6.93	7.00	1.42	5.00	9.00	20.46	0.09
CL/CD (target)	924	124.26	130.99	40.18	26.84	193.44	32.34	-0.58

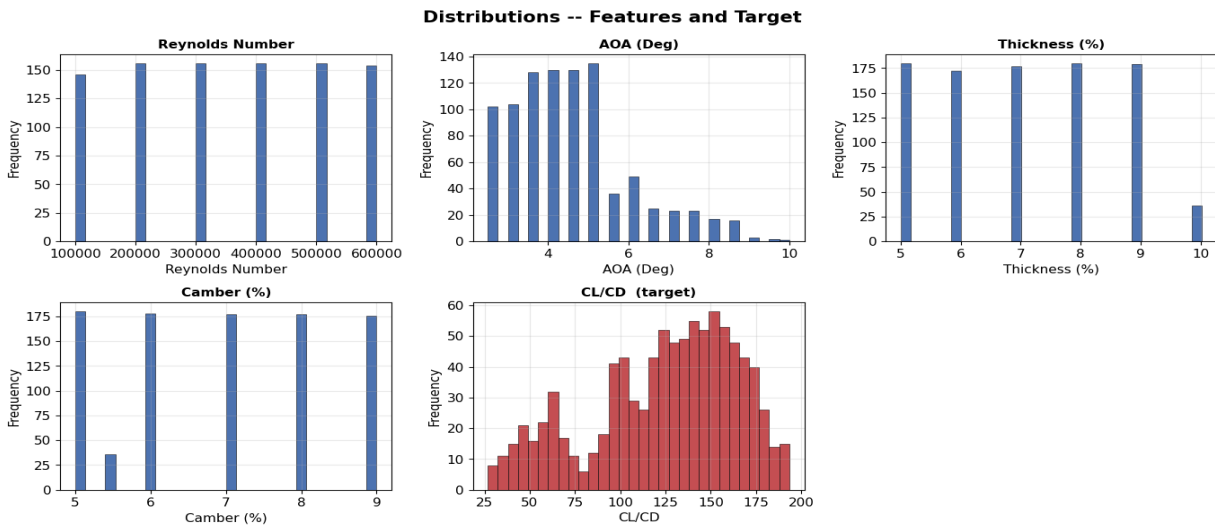


Figure 1: Distributions of the four numeric features and the target variable (CL/CD) across all 924 observations.

The Reynolds Number and the two geometric parameters (Thickness, Camber) were sampled at approximately six discrete design levels each, producing the near-uniform, multi-modal histograms visible in Figure 1.1. AOA was sampled more densely in the 2.5–5.5° range, reflecting the practical operating envelope of interest for the airfoil family under study. The target distribution, CL/CD, is unimodal and mildly left-skewed (skewness = -0.58), consistent with diminishing aerodynamic efficiency gains at higher angles of attack as the airfoils approach stall-adjacent conditions.

Correlation Structure

Pairwise Pearson correlation coefficients between the numeric features and the target are shown in Figure 2. Reynolds Number shows the strongest linear association with CL/CD ($r = 0.91$), followed by a moderate negative correlation with AOA ($r = -0.63$). Thickness and Camber show weak linear correlations with the target ($r = -0.08$ and $r = 0.12$, respectively), suggesting that their influence on CL/CD, if present, is non-linear or conditional on the other operating parameters rather than expressible as a simple monotonic trend.

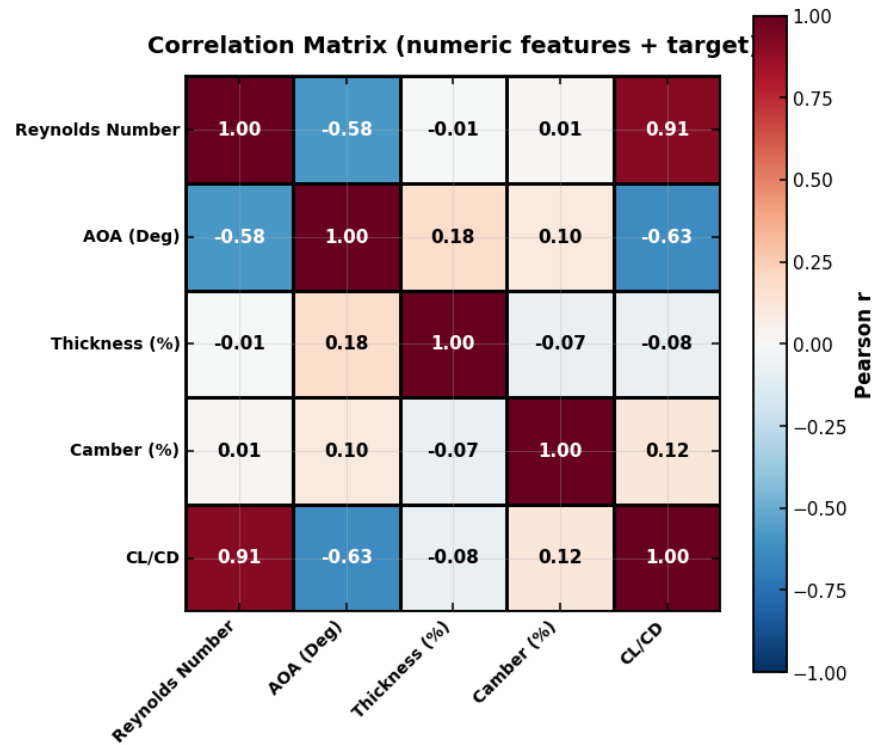


Figure 2: Pearson correlation matrix for the four numeric features and the target (CL/CD).

It is worth flagging this correlation structure explicitly for downstream interpretation: a Reynolds-Number-CL/CD correlation of 0.91 is considerably stronger than is typical for pure Reynolds-number effects on lift-to-drag ratio in classical airfoil aerodynamics, where AOA and camber usually dominate CL, and skin-friction/transition behaviour usually dominates CD to a comparable or lesser degree. This raises the possibility that Reynolds Number in this dataset is partially confounded with, or acting as a proxy for, another varying condition in the underlying case matrix (for example, if certain Reynolds Number levels were paired systematically with particular AOA sweeps or particular airfoil sub-groups during dataset generation). This does not invalidate the surrogate models built on the data they remain statistically valid interpolators of the sampled design space but it does mean that the feature-importance and SHAP attributions presented in Section 4.4 should be read as importance within this dataset's sampling scheme, and revisited once the aerodynamic case-generation matrix is independently confirmed to be free of hidden confounds.

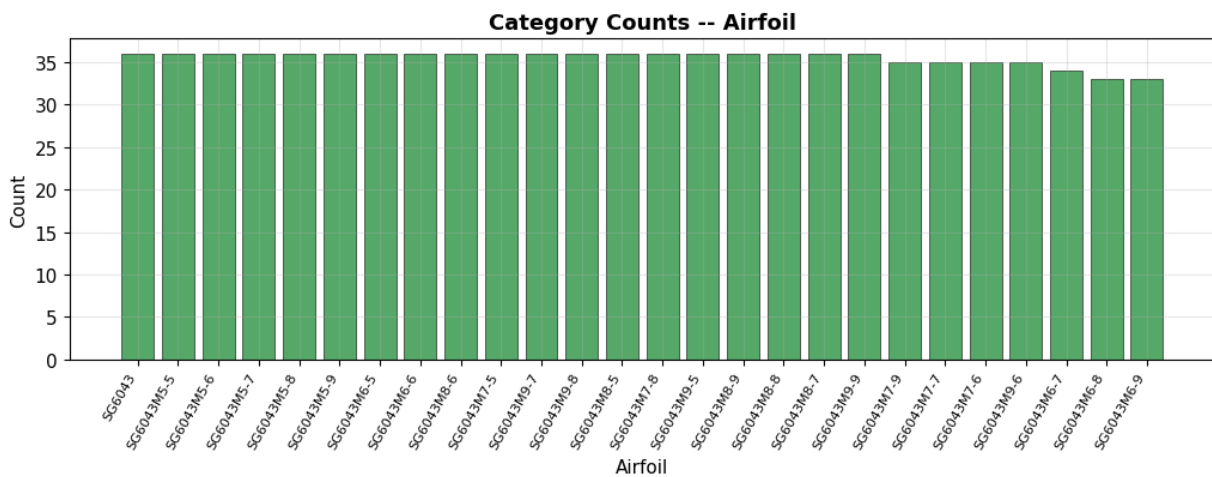


Figure 3: Sample count per airfoil variant (26 levels of the categorical Airfoil feature).

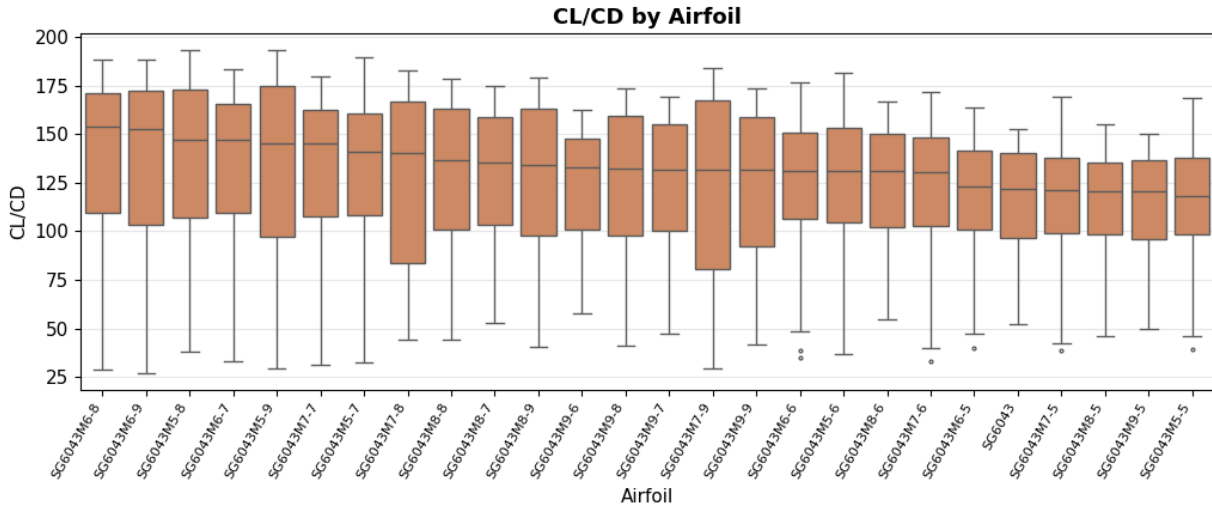


Figure 4: Distribution of CL/CD across the 26 airfoil variants, ordered by descending median efficiency.

Figure 4 shows that median CL/CD varies by roughly a factor of 1.3 across airfoil variants (from approximately 118 to 153), with substantial within-airfoil spread driven by the Reynolds Number and AOA sweep applied to each geometry. This confirms that the categorical Airfoil identity carries genuine, if secondary, predictive signal beyond the four continuous parameters, justifying its inclusion as a one-hot-encoded feature block in the models below.

Research Methodology

Surrogate Model Development Methodology

Data Partitioning

The dataset (924 rows) was partitioned into training and test subsets using an 80/20 random split (`train_test_split`, `scikit-learn`), with a fixed random seed (`random_state = 42`) to ensure reproducibility. This yielded 739 training observations and 185 held-out test observations. The test set was not used in any form during model fitting, hyperparameter specification, or feature preprocessing; it is used exclusively for the performance evaluation reported in Section.

Feature Encoding and Scaling

The single categorical feature, Airfoil, was one-hot encoded (26 binary indicator columns, unknown categories ignored at inference time), producing an encoded feature matrix of width 30 (4 continuous + 26 one-hot categorical columns). Two parallel preprocessing pipelines were fitted on the training partition only and then applied to both partitions:

- Scaled pipeline (`StandardScaler` on numeric columns): used for models sensitive to feature scale — MLR and ANN.
- Unscaled/passthrough pipeline: used for tree-based ensembles RF and XGBoost for which monotone feature transformations do not affect split quality.

For the ANN and XGBoost estimators, the target variable itself was additionally standardised via a `TransformedTargetRegressor` wrapper (fit on the training target only, inverse-transformed automatically at prediction time). This is a standard precaution when a target's numerical scale is far from unit order, and as implemented here costs nothing in cases such as this one where CL/CD is already reasonably well-scaled; it guards against optimiser sensitivity to target scale without altering the interpretation of predicted values.

Model Specifications

Table 3: Documents the exact hyperparameters used for each of the four estimators, as configured in the modelling pipeline.

Model		Implementation	Key hyperparameters	Scaling required
MLR		sklearn LinearRegression	Ordinary least squares, no regularisation	Yes (numeric features standardised)
Random Forest (RF)	Forest	sklearn RandomForestRegressor	n_estimators = 300; random_state = 42; all other parameters at scikit-learn defaults	No
Artificial Neural Network (ANN)	Neural	sklearn MLPRegressor	Hidden layers = (100, 50); max_iter = 3000; early_stopping = True; n_iter_no_change = 15; target standardised via TransformedTargetRegressor	Yes
XGBoost		xgboost XGBRegressor	n_estimators = 300; learning_rate = 0.05; max_depth = 6; subsample = 0.9; colsample_bytree = 0.9; target standardised via TransformedTargetRegressor	No

All models share the same random seed (42) for the train/test split and for any stochastic component of model fitting, ensuring that performance differences reported below are attributable to model structure rather than sampling variation.

Evaluation Metrics

Four standard regression metrics were computed on both the training and test partitions for every model:

- Coefficient of determination (R^2) proportion of target variance explained by the model.
- Root Mean Squared Error (RMSE) average magnitude of prediction error, in the same units as CL/CD, with larger errors penalised quadratically.
- Mean Absolute Error (MAE) average absolute magnitude of prediction error, in the same units as CL/CD.
- Mean Absolute Percentage Error (MAPE, %) average absolute error normalised by the observed value, expressed as a percentage.

Test-set metrics (rather than training-set metrics) are treated as the primary basis for model selection and comparison throughout this chapter, since training performance alone does not indicate generalisation and can be inflated by overfitting, particularly for the more flexible learners (RF, ANN, XGBoost).

Results

Model Performance Comparison

Table 3 reports training and test performance for all four models. All non-linear models substantially outperform the linear baseline, and all achieve test-set R^2 above 0.99, indicating that the ANN, RF, and XGBoost surrogates recover the underlying input-output mapping with high fidelity across the sampled design space.

Table 4: Training and Test Set Performance Metrics

Model	Train R^2	Test R^2	Test RMSE	Test MAE	Test MAPE (%)
MLR	0.8750	0.8462	15.2241	11.0919	12.6096
Random Forest	0.9986	0.9929	3.2797	2.2320	2.2051
ANN	0.9986	0.9967	2.2459	1.4568	1.6806
XGBoost	0.9993	0.9946	2.8657	1.7258	1.8599

The Multiple Linear Regression baseline achieves a test R^2 of 0.846, reflecting the fact that CL/CD is not a purely linear function of the four continuous predictors — a conclusion already anticipated by the weak linear correlations of Thickness and Camber with the target reported in Section 4.2. The three non-linear learners close this gap substantially: Random Forest reaches a test R^2 of 0.993, XGBoost reaches 0.995, and the Artificial Neural Network achieves the highest test R^2 of all models evaluated, 0.997, with the lowest test RMSE (2.246), lowest MAE (1.457), and lowest MAPE (1.68%) of the four models. On this basis, the ANN was selected as the best-performing surrogate model and is carried forward for the interpretability and uncertainty analyses in Sections 4.4.2 and 4.4.3.

A further observation supporting the robustness of the ANN result is the small gap between its training R^2 (0.9986) and test R^2 (0.9967) — a difference of only 0.0019 — indicating that the network's generalization to unseen combinations of Reynolds Number, AOA, thickness, camber, and airfoil identity is not achieved at the cost of overfitting. XGBoost shows a comparable pattern (train 0.9993, test 0.9946, gap 0.0047), while Random Forest shows the smallest train–test gap of the ensemble methods in absolute R^2 terms (0.9986 to 0.9929).

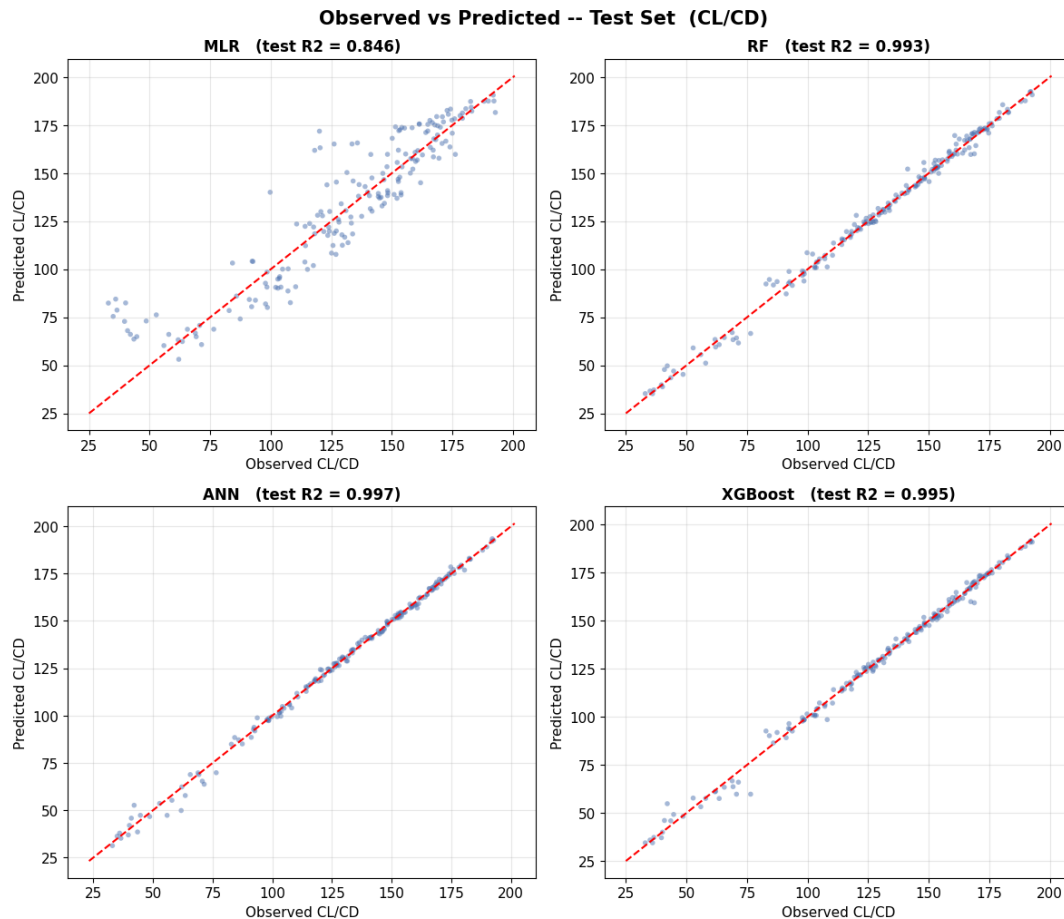


Figure 4: Observed vs. predicted CL/CD on the held-out test set (185 samples) for all four models. The dashed red line is the 1:1 identity line; tighter clustering along this line indicates better predictive accuracy.

Figure 4 visualizes the observed-versus-predicted scatter for all four models on the test set. The MLR panel shows a clear pattern of biased, high-variance scatter around the identity line particularly at low CL/CD values, where the linear model both under- and over-predicts noticeably consistent with its lower R^2 and higher RMSE. In contrast, the RF, ANN, and XGBoost panels all show tight clustering along the 1:1 line across the full observed range (approximately 25 to 195), with no visually apparent systematic bias, heteroscedasticity, or region-specific breakdown in accuracy. The ANN panel in particular shows the least scatter of the four models at both the low and high ends of the CL/CD range.

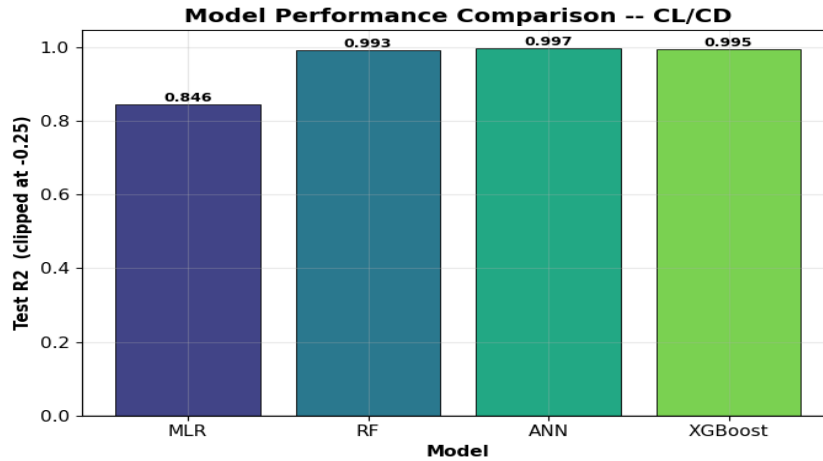


Figure 5: Test-set R^2 comparison across the four surrogate models.

Feature Importance and Explainability (SHAP Analysis)

Because the selected best model (ANN) does not expose a native feature-importance attribute (unlike tree-based models, which provide impurity- or gain-based importance directly), SHAP (SHapley Additive exPlanations) values were computed to attribute the ANN's predictions to individual input features. SHAP values were estimated on a representative sample of 300 test-set observations, using a model-agnostic Kernel/Deep SHAP explainer appropriate for the fitted MLPRegressor.

Table 5: Top 10 Features by Mean Absolute SHAP Value (ANN model)

Rank	Feature	Mean SHAP
1	Reynolds Number	29.895
2	Camber (%)	4.359
3	AOA (Deg)	4.158
4	Thickness (%)	3.176
5	Airfoil_SG6043M8-6	0.781
6	Airfoil_SG6043M6-6	0.675
7	Airfoil_SG6043M6-7	0.668
8	Airfoil_SG6043M7-7	0.588
9	Airfoil_SG6043M9-8	0.503
10	Airfoil_SG6043M5-7	0.491

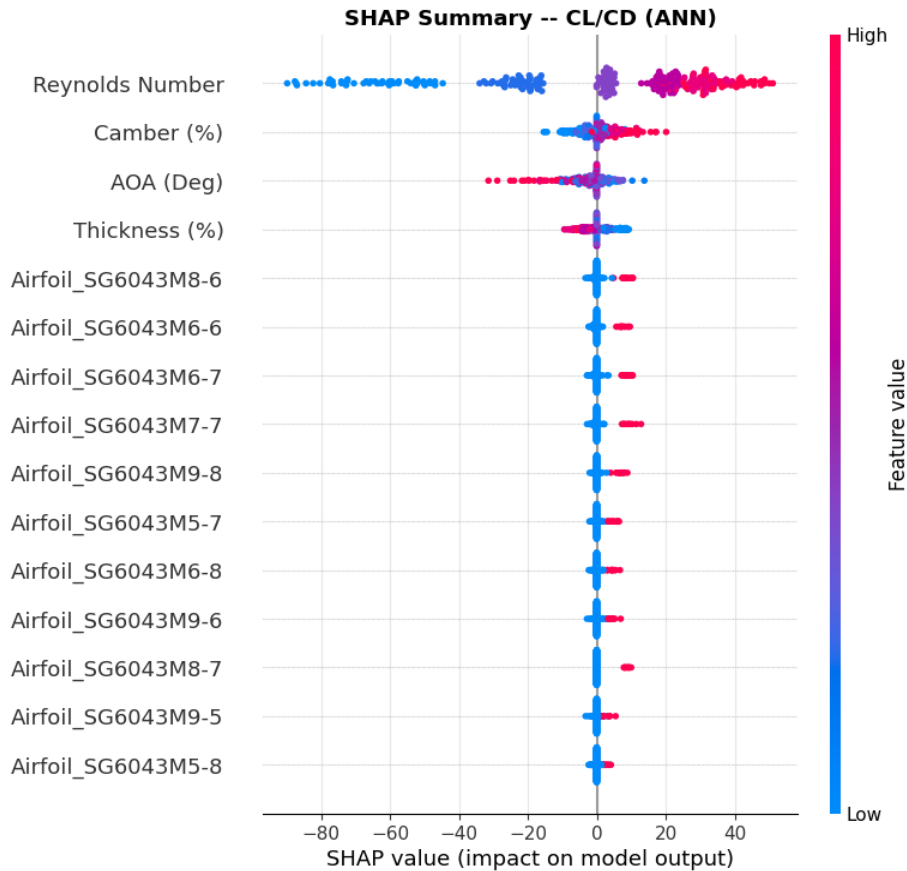


Figure 6: SHAP summary plot for the ANN surrogate model. Each point is one test-set observation; horizontal position shows the SHAP value (impact on predicted CL/CD); colour shows the feature's own value (red = high, blue = low).

Reynolds Number dominates the SHAP attribution by nearly an order of magnitude (mean $|\text{SHAP}| = 29.9$) relative to every other feature, mirroring the strong linear correlation already observed in Section 4.2. High Reynolds Number values (red points, right-hand cluster in Figure 6) push predicted CL/CD strongly upward, while low Reynolds Number values (blue points) push it strongly downward – consistent with the physical expectation that boundary-layer transition and reduced relative viscous drag at higher Reynolds Number improve lift-to-drag performance, though the magnitude of this effect in the present dataset (as discussed in Section 4.2) may also reflect the sampling structure of the underlying case matrix rather than aerodynamics alone.

Camber and AOA follow as the second- and third-most influential features (mean $|\text{SHAP}|$ of 4.36 and 4.16, respectively), with Thickness fourth (3.18). Individually, no single airfoil-identity indicator column contributes more than approximately 2.6% of the Reynolds Number's importance, confirming that the continuous flow/geometry parameters not airfoil identity are the dominant physical drivers of CL/CD in this surrogate. The airfoil-specific indicators nonetheless collectively account for the residual airfoil-to-airfoil spread visible in Figure 3.3, and their presence in the top-10 ranking (five of the ten most important features are individual airfoil levels) shows that shape family membership retains a small but non-negligible marginal effect after conditioning on flow state.

Residual Diagnostics and Prediction Uncertainty

Residual diagnostics for the selected ANN surrogate were computed on the test set ($n = 185$). The residual mean (bias) was 0.179, close to zero and indicating negligible systematic over- or under-prediction. The residual standard deviation was 2.245, consistent with the RMSE reported in Table 4.4. A 95% prediction interval was constructed as $\pm 1.96\sigma$ around each point prediction, yielding a half-width of 4.400 CL/CD units; empirical coverage of this interval on the test set was 94.59%, closely matching the nominal 95% target and supporting the appropriateness of a Gaussian-residual assumption for uncertainty quantification with this model.

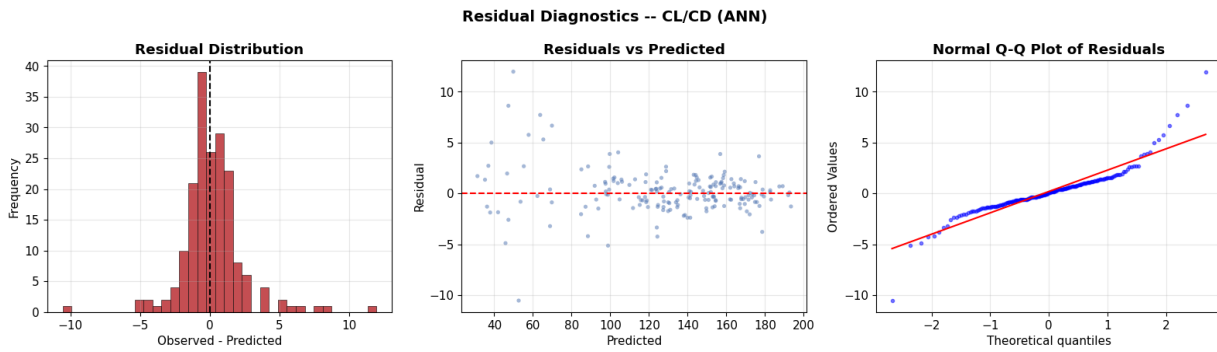


Figure 7: Residual diagnostics for the ANN surrogate model: (left) histogram of residuals (observed minus predicted); (centre) residuals plotted against predicted CL/CD; (right) normal quantile–quantile plot of residuals.

The residual histogram (Figure 7, left) is approximately symmetric and centred near zero, with a small number of outlying residuals beyond ± 8 CL/CD units. The residuals-versus-predicted panel (centre) shows no strong funnel or curvature pattern across the predicted range (roughly 30 to 195), indicating that prediction error is reasonably homoscedastic and does not degrade systematically at either extreme of the aerodynamic efficiency range. The normal Q-Q plot (right) tracks the reference line closely through the central mass of residuals, with mild deviation in the tails consistent with the handful of larger residuals visible in the histogram and indicating slightly heavier-than-Gaussian tails, a common and generally benign feature of neural-network residuals on finite test samples.

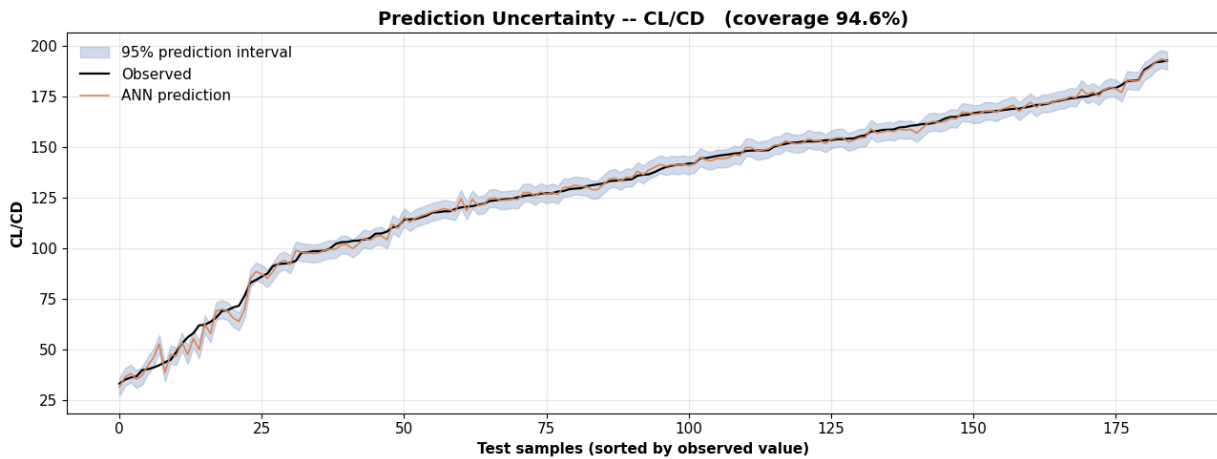


Figure 8: Observed CL/CD, ANN point predictions, and the associated 95% prediction interval across the test set, sorted by ascending observed value (coverage = 94.6%).

Figure 8 confirms that the ANN prediction band tracks the observed CL/CD trend closely across its full range, from the lowest-efficiency cases (CL/CD $\approx$ 30) to the highest (CL/CD $\approx$ 195), with the shaded 95% interval containing the true observation in all but a small number of cases concentrated in the lower-CL/CD region consistent with the slightly heavier tail behaviour noted in the Q-Q plot above.

Discussion

The results in Section 4.4 demonstrate that a modestly sized artificial neural network (two hidden layers, 100 and 50 units) can act as an accurate, low-cost surrogate for the aerodynamic efficiency of the SG6043 airfoil family and its systematic geometric derivatives, achieving a test-set R^2 of 0.997 and a mean absolute percentage error of 1.68% an order-of-magnitude improvement in fidelity over the linear baseline (test $R^2 = 0.846$, MAPE = 12.61%). This confirms that the mapping from (Airfoil, Reynolds Number, AOA, Thickness, Camber) to CL/CD contains non-linear and/or interaction structure that a purely linear model cannot capture, but that this structure is nonetheless learnable from a moderately sized dataset (739 training rows) without evidence of overfitting.

The ensemble tree-based methods (RF and XGBoost) perform only marginally behind the ANN (test R^2 of 0.993 and 0.995 respectively) and offer two practical advantages the ANN does not: native, computationally cheap feature-importance estimates (removing the need for a separate SHAP pass) and typically faster training and inference at this dataset scale. Where interpretability speed or deployment simplicity is prioritised over the last increment of accuracy, XGBoost is a strong secondary candidate; where maximum point-prediction accuracy is the priority, as adopted here, the ANN is preferred.

The SHAP-based explainability analysis (Section 4.4.2) is the most important result to interpret cautiously. It correctly identifies Reynolds Number as the dominant driver of predicted CL/CD in this dataset, but the unusually strong linear correlation between Reynolds Number and the target ($r = 0.91$, Section 4.2) well above what is typically observed for Reynolds-number effects alone in subsonic airfoil aerodynamics suggests that this dataset's case-generation matrix may not vary Reynolds Number, AOA, and geometry fully independently of one another. Before these importance rankings are reported as a general physical conclusion about the SG6043 family (rather than as a description of what the model learned from this particular sample), it is recommended that:

- The underlying case-generation matrix (whether from panel-method, RANS, or experimental sources) be audited to confirm that Reynolds Number was not systematically co-varied with AOA sweep range or with a specific airfoil subgroup during data generation.
- A variance-inflation-factor (VIF) or partial-correlation check be performed between Reynolds Number and the other three continuous features to quantify any residual confounding not visible in the pairwise Pearson matrix of Figure 4.2.
- If confounding is confirmed, the SHAP analysis be re-run on a re-balanced or stratified resampling of the dataset that decorrelates Reynolds Number from AOA and airfoil identity, to isolate each feature's independent physical contribution.

This caveat does not undermine the predictive validity of the four surrogate models presented: all were trained and evaluated using a proper held-out test split, and their reported R^2 , RMSE, MAE, and MAPE values are legitimate measures of how well each model interpolates within the sampled design space. The caveat applies specifically to causal or physical interpretation of the feature-importance ranking, which should be read as importance-within-this-dataset rather than an independently validated statement about airfoil aerodynamics in general.

The residual and uncertainty analysis (Section 4.4.3) provides additional confidence in the ANN surrogate's practical usability: near-zero bias, close-to-nominal 95% prediction interval coverage (94.6% against a 95% target), and no strong evidence of heteroscedasticity across the observed CL/CD range together indicate that a simple Gaussian-residual prediction interval is an adequate first-order uncertainty quantification method for this model, suitable for use in downstream design-space screening or optimisation workflows where a calibrated confidence band around each point prediction is required.

Summary

This chapter developed and benchmarked four machine-learning surrogate models — Multiple Linear Regression, Random Forest, an Artificial Neural Network, and XGBoost — for predicting the aerodynamic efficiency ratio CL/CD of the SG6043 airfoil family and its 25 geometric derivatives, using a dataset of 924 observations spanning Reynolds Number, angle of attack, thickness, and camber. The key findings are summarised below:

- An 80/20 train-test split (739/185 observations, seed = 42) with one-hot encoding of the categorical Airfoil feature and standardisation appropriate to each estimator was used consistently across all four models.
- The Artificial Neural Network (two hidden layers, 100 and 50 neurons) achieved the best generalisation performance, with test $R^2 = 0.9967$, RMSE = 2.246, MAE = 1.457, and MAPE = 1.68%, and was adopted as the operative surrogate for this study.
- XGBoost (test $R^2 = 0.9946$) and Random Forest (test $R^2 = 0.9929$) performed nearly as well and remain viable alternatives, particularly where native feature-importance output or faster inference is preferred over marginal accuracy gains.
- The linear baseline (MLR, test $R^2 = 0.8462$) confirms that CL/CD is not a purely linear function of the four continuous predictors, motivating the use of non-linear surrogate architectures.

- SHAP analysis of the ANN identified Reynolds Number as the dominant predictor of CL/CD (mean |SHAP| = 29.9, roughly seven times the next most important feature, Camber), followed by Camber, AOA, and Thickness, with individual airfoil-identity indicators contributing smaller, secondary effects though this ranking should be validated against the dataset's case-generation matrix before being generalised as a physical conclusion (Section 4.5).
- Residual diagnostics showed near-zero bias (0.179), a 95% prediction interval with 94.6% empirical coverage, and no strong evidence of heteroscedasticity, supporting the use of the ANN surrogate together with a simple Gaussian prediction interval for rapid, uncertainty-aware CL/CD estimation across the sampled design space.

Taken together, these results establish that a compact ANN surrogate can replace repeated high-fidelity aerodynamic evaluation for this airfoil family within the sampled design envelope, at a fraction of the computational cost, while the identified data-provenance caveat around Reynolds Number defines a clear, actionable next step before the model's feature-importance conclusions are extended beyond this dataset.

References

1. Giguère, P., Selig, M.S. (1997). Low Reynolds number airfoils for small horizontal axis wind turbines. *Wind Engineering*, 21(6), 367–380. Illinois Experts
2. Giguère, P., Selig, M.S. (1998). New airfoils for small horizontal axis wind turbines. *Journal of Solar Energy Engineering*, 120(2), 108–114. <https://doi.org/10.1115/1.2888052>. Illinois Experts
3. Selig, M.S., McGranahan, B.D. (2004). Wind tunnel aerodynamic tests of six airfoils for use on small wind turbines. *Journal of Solar Energy Engineering*, 126(4), 986–1001. <https://doi.org/10.1115/1.1793208>. Illinois Experts
4. Singh, R.K., Ahmed, M.R., Zullah, M.A., Lee, Y.H. (2012). Design of a low Reynolds number airfoil for small horizontal axis wind turbines. *Renewable Energy*, 42, 66–76. <https://doi.org/10.1016/j.renene.2011.09.014>. IDEAS/RePEc
5. Shin, P., Kim, K. (2020). Aerodynamic performance prediction of SG6043 airfoil for a horizontal-axis small wind turbine. *Journal of Physics: Conference Series*, 1452, 012018. <https://doi.org/10.1088/1742-6596/1452/1/012018>.
6. Noronha, N.P., Krishna, M. (2021). Aerodynamic performance comparison of airfoils suggested for small horizontal axis wind turbines. *Materials Today: Proceedings*, 46, 2450–2455. <https://doi.org/10.1016/j.matpr.2021.01.359>. ScienceDirect
7. Davari, H.S., Abdollahzadeh Jamalabadi, M.Y., Davari Germi, M.S. (2025). The SG6043 airfoil optimization for low Reynolds number applications in wind turbines. *Mechanical Engineering Advances*. <https://doi.org/10.59400/mea2486>.
8. Rind, A.A., Luhur, M.R., Manganhar, A.L., Ghoto, S.M., Bhangwar, S. (2026). Design optimization of a low Reynolds number airfoil SG6043 for small horizontal axis wind turbines. *Wind*, 6(2), 20. <https://doi.org/10.3390/wind6020020>. MDPI
9. Drela, M. (1989). XFOIL: An analysis and design system for low Reynolds number airfoils. In T.J. Mueller (Ed.), *Low Reynolds Number Aerodynamics*. Springer, pp. 1–12.
10. Marten, D., Wendler, J., Pechlivanoglou, G., Nayeri, C.N., Paschereit, C.O. (2013). QBlade: An open source tool for design and simulation of horizontal and vertical axis wind turbines. *International Journal of Emerging Technology and Advanced Engineering*, 3, 264–269.
11. Forrester, A.I.J., Sóbester, A., Keane, A.J. (2008). *Engineering Design via Surrogate Modelling: A Practical Guide*. Wiley.
12. Queipo, N.V., Haftka, R.T., Shyy, W., Goel, T., Vaidyanathan, R., Tucker, P.K. (2005). Surrogate-based analysis and optimization. *Progress in Aerospace Sciences*, 41(1), 1–28.
13. Chung, H.S., Alonso, J.J. (2007). A surrogate-model based multidisciplinary shape optimization method with application to a 2D subsonic airfoil. *Computers & Fluids*, 36(3), 520–529.
14. Du, X., He, P., Martins, J.R.R.A. (2021). Rapid airfoil design optimization via neural networks-based parameterization and surrogate modeling. *Aerospace Science and Technology*, 113, 106701. <https://doi.org/10.1016/j.ast.2021.106701>. ScienceDirect

15. Li, J., Du, X., Martins, J.R.R.A. (2022). Machine learning in aerodynamic shape optimization. *Progress in Aerospace Sciences*, 134, 100849. <https://doi.org/10.1016/j.paerosci.2022.100849>. DOI
16. Teimourian, A., Rohacs, D., Dimililer, K., Teimourian, H., Yildiz, M., Kale, U. (2024). Airfoil aerodynamic performance prediction using machine learning and surrogate modeling. *Heliyon*, 10(8), e29377. <https://doi.org/10.1016/j.heliyon.2024.e29377>. ScienceDirect
17. Kieszek, R., Majcher, M., Syta, B., Kozakiewicz, A. (2024). Feed-forward artificial neural network as surrogate model to predict lift and drag coefficient of NACA airfoil and searching of maximum lift-to-drag ratio. *Journal of Theoretical and Applied Mechanics*, 62(3), 521-534.
18. Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32. <https://doi.org/10.1023/A:1010933404324>.
19. Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232.
20. Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. <https://doi.org/10.1145/2939672.2939785>. DOI
21. Lundberg, S.M., Lee, S.I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30, 4765-4774.



2026 by the authors; Journal of Global Social Transformation (JGST). This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).