Rensselaer

MANE 4962 FINAL PROJECT

Predicting $(C_L/C_D)_{max}$ from Airfoil Geometric Parameters using Machine Learning

Melvina Quartey | 662067086

04/24/2024

EXECUTIVE SUMMARY

This report presents an analysis of airfoil performance prediction using various machine learning models. The goal is to predict the maximum lift-to-drag ratio $(C_L/C_D)_{max}$ of an airfoil based on its geometric features – maximum thickness, maximum camber, position of maximum camber and the position of maximum thickness. Each of these features are expressed as percentages of the a unit chord length. The maximum lift-to-drag ratio $(C_L/C_D)_{max}$ of an airfoil is a crucial parameter in airfoil performance. Traditionally, obtaining these values has relied on costly and time-consuming experimental methods or computationally intensive simulations. This research proposes a cost-effective and time-efficient alternative using supervised machine learning models. Four different models were explored: linear regression, random forest, deep neural network, and support vector regression (SVM) and each model was evaluated based on mean squared error (MSE) and R-squared (R2) score.

Our objective was two-fold: to accurately predict $(C_L/C_D)_{max}$ values and to decipher the geometric features' relative importance in determining airfoil performance.

The models were rigorously trained and tested, revealing that complex models such as deep neural networks and random forests offer substantial predictive power. The linear regression model achieved an MSE of 10.71 and an R2 score of 0.47. SVM regression outperformed linear regression with an MSE of 6.40 and an R2 score of 0.68. Best of all were the deep neural network and random forest regression models which both had similar performance. Both had an R2 score of 0.69 and MSEs of 6.27 and 6.20 respectively.

Additional efforts were made to enhance the predictive power of the models such as feature engineering, dropout and L2-regularization, cross-validation, and outlier removal. Additional features, such as camber-thickness difference, camber-thickness ratio, and position ratios, were added to the dataset. Although these features seemed capable of improving model performance, they did not lead to any gains in predictive scores.

The random forest model emerged as the most effective for predicting $(C_L/C_D)_{max}$, closely followed by the deep neural network. However, there remains room for enhancement. The model’s inability to improve from a 0.7 R-2 score indicates that some aspects of the maximum lift-to-drag ratio require more information on the airfoil geometry than the four features can provide. I observe that the features play a stronger role in lift prediction than they do in drag prediction. The current model can be refined by incorporating additional parameters that describe the airfoil’s geometry with greater fidelity. Characteristics such as curvature, twist, taper, and thickness distribution may unlock deeper insights and yield predictions with higher precision.

Our findings demonstrate that machine learning can effectively predict airfoil performance, with implications for enhancing the design process in aeronautics and related fields. Engineers and designers can integrate these predictive models into their workflow to rapidly evaluate numerous airfoil geometries, thereby expediting development cycles and reducing costs.

Additionally, this work paves the way for further exploration of machine learning in aerodynamics, suggesting the potential for these techniques to eventually replace more resource-intensive methods. Future research may expand the dataset, explore ensemble methods, or delve into deep learning's capacity to capture complex non-linear relationships.

1. INTRODUCTION

Aerodynamic efficiency is paramount in the design and engineering of aircraft, where the interplay of forces determines not only performance but also fuel efficiency and environmental impact. The coefficient of lift (C_L) and drag (C_D) are most to be considered, with their ratio $(C_L/C_D)_{max}$ serving as a key indicator of an airfoil's performance. Historically, the knowledge and optimization of these coefficients has necessitated extensive and highly iterative wind tunnel testing and computational fluid dynamics (CFD) simulations — methods that are both cost and time-intensive.

The advent of machine learning offers a game-changing alternative to the process of predicting airfoil performance. Machine learning approaches have shown potential in accurately modeling the complex relationships between airfoil shapes and aerodynamic performance, while also cutting down the many iterations in costly experimental testing and simulations. Leveraging data-driven models promises a significant reduction in the resources required to estimate aerodynamic parameters with a potential to expedite the iterative design process substantially. This project is situated at the intersection of traditional aerodynamic methods and the cutting edge of artificial intelligence.

The objectives of this research were in two correlated parts: firstly, to develop a supervised machine learning model capable of using regression methods to accurately predict $(C_L/C_D)_{max}$ values from a given set of airfoil geometric features. The second aim then comes as a by-product; to evaluate and interpret the relative importance of these features in the context of airfoil efficiency.

In pursuit of these goals, the study draws upon a robust dataset from Airfoil Tools, comprising a diverse array of airfoil designs, including but not limited to NACA, Eppler, and Clark profiles. For the data to be meaningful, is extracted for airflow at the same Reynolds Number, a dimensionless parameter that suitably summarizes the regime and characteristics of the flow and end enables relevant flow similarity conditions to be met for comparisons to be valid.

There exists literature that substantiates the potential of machine learning in this domain, with seminal works such as those by Ahmed et al. (2024) and Hassan Moin et al. (2021) which highlight the potential of machine learning in optimizing airfoil designs while also providing a basis for methodological approaches and validation.

Through iterative model development and comparison, tuning, and validation, this research contributes to a growing body of knowledge, creating a pathway for incorporating machine learning into the process of aerodynamic analysis and design. Key evaluation metrics, including R-2 scores and mean squared error (MSE), will be used to assess the performance of the model. Additionally, scatterplots of predicted versus $(C_L/C_D)_{max}$ will be visualized to analyze the model's predictive capabilities and identify any areas for improvement.

The second part of the objectives allows for a transparent understanding of the relationship between independent variables (airfoil geometry parameters) and the target variable through the model coefficients or feature importance. By analyzing these values, this work aims to gain insights into the relative influences of different geometric parameters on the predictions, thereby enhancing our understanding of airfoil physics and design principles.

By combining theoretical principles with practical data-driven approaches, this study endeavors to advance the efficiency and effectiveness of airfoil design processes for future aerospace applications.

2. PROBLEM DEFINITION

Airfoil design is a critical component in aeronautical engineering, dictating the efficiency, stability, and performance of an aircraft. A good airfoil has the capacity to significantly reduce fuel consumption and contribute towards sustainable environment goals. One of the paramount challenges in airfoil design is the process involved in accurately predicting the maximum lift-to-drag ratio, $(C_L/C_D)_{max}$, which signifies the optimal performance point. Traditionally, this involves wind tunnel testing and computational fluid dynamics (CFD) simulations – one experiment or simulation for each angle of attack on each airfoil. While this is ultimately effective, these methods demand substantial financial investment, extensive time commitments, and are constrained by the limitations of physical or computational models.

The problem is therefore two-fold: Can we predict $(C_L/C_D)_{max}$ with greater efficiency and lower costs? And, can we improve the understanding of the influence of geometric features on airfoil performance?

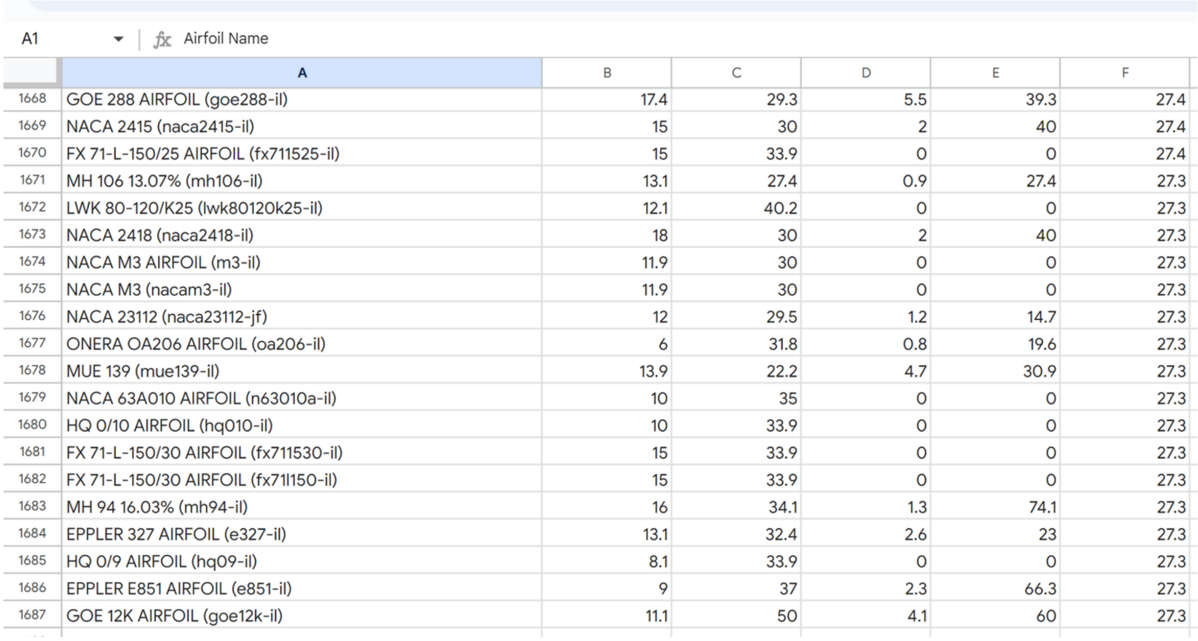
By leveraging machine learning techniques, we can develop predictive models that can effectively capture the complex and nonlinear relationships between airfoil geometry parameters and lift coefficients. Traditional analytical methods struggle to capture these intricate relationships accurately, especially when dealing with high-dimensional datasets and complex geometries. Machine learning algorithms, on the other hand, excel at learning patterns and relationships from data, making them well-suited for modeling the nonlinear behaviors inherent in aerodynamic systems. These models have the potential to significantly streamline the airfoil design process, reduce reliance on costly experimental testing, and accelerate innovation in aerospace engineering.

3. METHODS AND PROCEDURE

The project involved several steps to develop models for estimating $(C_L/C_D)_{max}$ from airfoil geometric parameters. The methods employed encompass data collection, preprocessing, model selection, training, evaluation, and interpretation. The steps undertaken are as follows:

1. **Data Collection:**

Airfoil geometric parameters and corresponding $(C_L/C_D)_{max}$ were obtained from open-source dataset available Airfoil Tools. This dataset provided a diverse range of airfoil designs and associated aerodynamic performance data. The dataset for this study comprises 1686 complete data points (airfoils) stored in a spreadsheet.



	A	B	C	D	E	F
1668	GOE 288 AIRFOIL (goe288-il)	17.4	29.3	5.5	39.3	27.4
1669	NACA 2415 (naca2415-il)	15	30	2	40	27.4
1670	FX 71-L-150/25 AIRFOIL (fx711525-il)	15	33.9	0	0	27.4
1671	MH 106 13.07% (mh106-il)	13.1	27.4	0.9	27.4	27.3
1672	LWK 80-120/K25 (lwk80120k25-il)	12.1	40.2	0	0	27.3
1673	NACA 2418 (naca2418-il)	18	30	2	40	27.3
1674	NACA M3 AIRFOIL (m3-il)	11.9	30	0	0	27.3
1675	NACA M3 (nacam3-il)	11.9	30	0	0	27.3
1676	NACA 23112 (naca23112-jf)	12	29.5	1.2	14.7	27.3
1677	ONERA OA206 AIRFOIL (oa206-il)	6	31.8	0.8	19.6	27.3
1678	MUE 139 (mue139-il)	13.9	22.2	4.7	30.9	27.3
1679	NACA 63A010 AIRFOIL (n63010a-il)	10	35	0	0	27.3
1680	HQ 0/10 AIRFOIL (hq010-il)	10	33.9	0	0	27.3
1681	FX 71-L-150/30 AIRFOIL (fx711530-il)	15	33.9	0	0	27.3
1682	FX 71-L-150/30 AIRFOIL (fx71l150-il)	15	33.9	0	0	27.3
1683	MH 94 16.03% (mh94-il)	16	34.1	1.3	74.1	27.3
1684	EPPLER 327 AIRFOIL (e327-il)	13.1	32.4	2.6	23	27.3
1685	HQ 0/9 AIRFOIL (hq09-il)	8.1	33.9	0	0	27.3
1686	EPPLER E851 AIRFOIL (e851-il)	9	37	2.3	66.3	27.3
1687	GOE 12K AIRFOIL (goe12k-il)	11.1	50	4.1	60	27.3

Figure 1 Snapshot of The Airfoil Dataset

At this point, it is important to know and understand the airfoil geometric parameters. The geometric parameters used as primary features in this study are:

- **Maximum Camber** as a percentage of chord length
- **Maximum Thickness** as a percentage of chord length
- **Maximum Camber Position** as a percentage of chord length
- **Maximum Thickness Position** as a percentage of chord length

Figure 1 provides an intuitive understanding of these features on an airfoil.

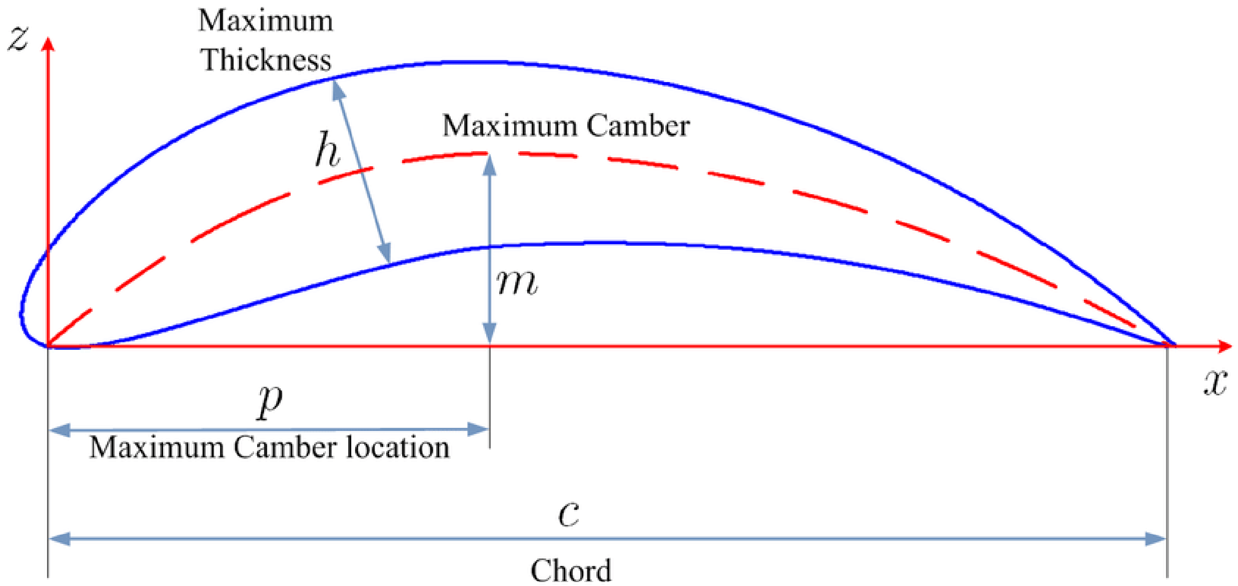
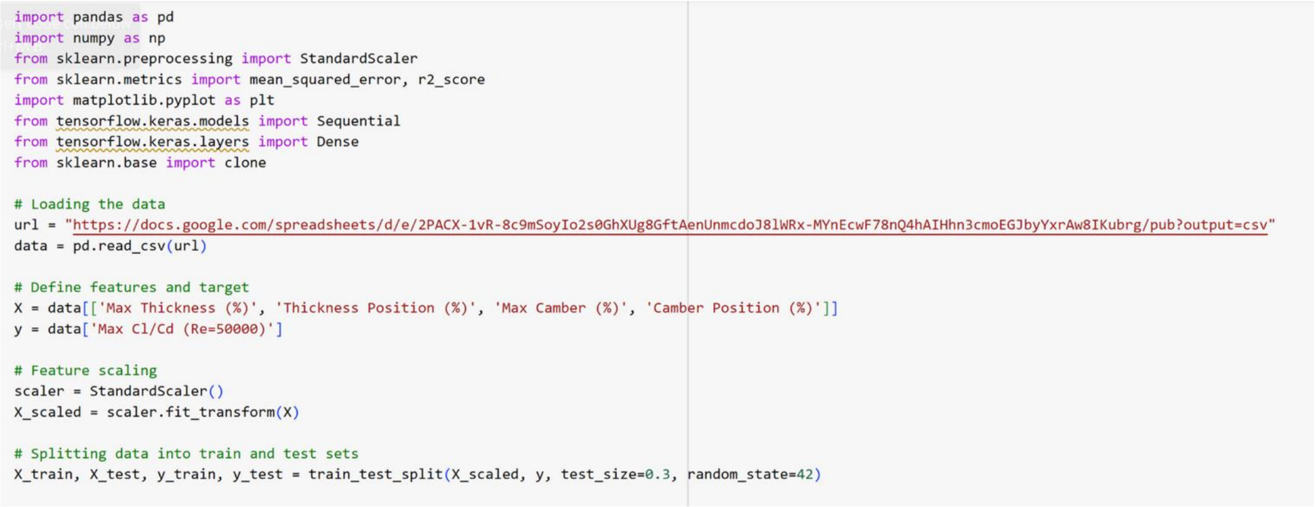


Figure 2 Airfoil Geometric Description by Ayman A. Nada

2. **Data Preprocessing:**

The collected data underwent preprocessing to ensure quality and compatibility with machine learning algorithms. This involved:

- Feature scaling: The features were normalized to a similar range using Standard Scaler to prevent biases in the model training process.
- Data splitting: The dataset was divided into 70% training and 30% testing sets to facilitate model training and evaluation.



```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
from sklearn.base import clone

# Loading the data
url = "https://docs.google.com/spreadsheets/d/e/2PACX-1vR-8c9mSoyIo2s0GhXUg8GftAenUnmcdoJ8lWRx-MYnEcwF78nQ4hAIHhn3cmoEGJbyYxrAw8IKubrg/pub?output=csv"
data = pd.read_csv(url)

# Define features and target
X = data[['Max Thickness (%)', 'Thickness Position (%)', 'Max Camber (%)', 'Camber Position (%)']]
y = data['Max Cl/Cd (Re=50000)']

# Feature scaling
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Splitting data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.3, random_state=42)
```

Figure 3 Data Preprocessing code snippet

3. **Model Selection:**

Four machine learning models were selected based on their prevalent usage and proven track record in regression tasks:

- Linear Regression (LR) for its simplicity and interpretability.
- Random Forest (RF) for its ability to handle non-linear relationships and feature importance evaluation.
- Deep Neural Networks (DNN) for their capacity to capture complex patterns through deep learning techniques.
- Support Vector Regression (SVM) for its robustness in high-dimensional spaces.

4. **Model Training and Evaluation:**

- Each selected model was trained using the training dataset and evaluated using the testing dataset. The training process involved optimizing model parameters to minimize the mean squared error and maximize the R-squared score.
- Mean Squared Error (MSE) to measure the average of the squares of the errors between the predicted and actual $(C_L/C_D)_{max}$ values.
- R-squared (R2) score to assess the proportion of variance for $(C_L/C_D)_{max}$ that is predictable from the input features.
- These metrics quantified the models' accuracy and ability to generalize unseen data. Particularly in the RF regressor model, hyperparameters were tuned using a Scikit Learn's GridsearchCV library which uses the cross-validation approach, and the best-performing parameters were selected for the final models.
- Also, scatterplots of the predicted and actual $(C_L/C_D)_{max}$ values were used to visually evaluate the performance of each model.
- The performance of each model was recorded, and the most successful model was identified through a comparative analysis of the evaluation metrics.

5. **Interpretation:**

- The trained models' coefficients (for linear regression) or feature importances (for Random Forest regression) were analyzed to interpret the relationships between airfoil geometric parameters and lift coefficients. This interpretation provided insights into the aerodynamic behaviors captured by the models. Special emphasis was placed on interpreting the model outcomes, particularly for the Random Forest algorithm, which provides a direct measure of feature importance. This analysis facilitated an understanding of which airfoil characteristics most significantly impact performance, informing future design strategies.

The procedures detailed herein were meticulously documented to enable other researchers and practitioners to replicate the study. By adhering to this methodological rigor, the project endeavors to contribute a reliable and valuable addition to the existing research in the field of aerodynamic machine learning.

4. DATASET AND VISUALIZATION

Dataset Overview

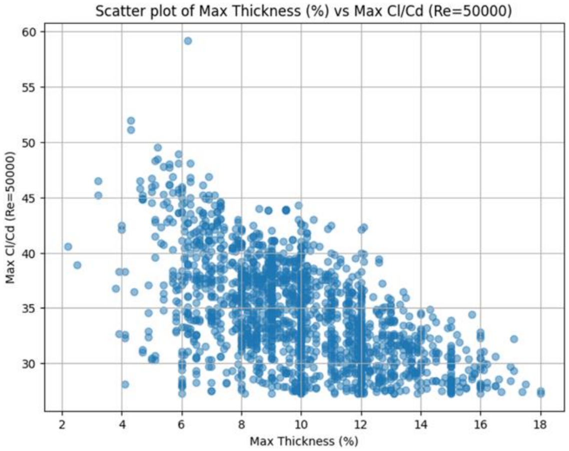
The dataset was integral to this study, consisting of 1686 data points representing various airfoil types, including NACA, Eppler, Clark, and Göttingen, among others. Each entry in the dataset comprised the following geometric features of airfoils: maximum thickness percentage, thickness position percentage, maximum camber percentage, and camber position percentage, along with the target variable, the maximum lift-to-drag ratio $(C_L/C_D)_{max}$ at a Reynolds number of 50,000.

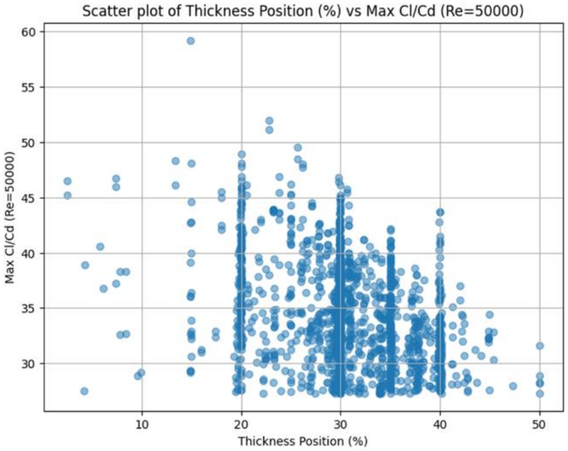
Data Visualization

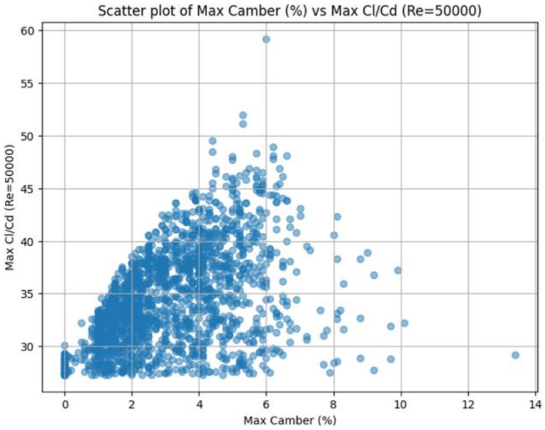
A series of visualization techniques were employed to elucidate the relationships between features and the target variable:

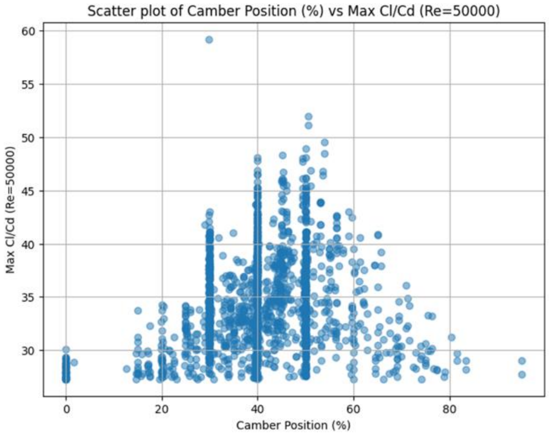
1. **Scatterplots:** Scatterplots were used to visualize the relationship between individual airfoil geometric parameters (e.g., maximum thickness, camber position) and lift coefficients. These plots helped identify any linear or nonlinear trends between the features and the target variable.

    - **Maximum Thickness:** I observe that as the maximum thickness increases, the target values tend decrease.
    - **Maximum Camber:** I observe that as the maximum camber increases, the target values tend to increase.
    - **Maximum Thickness Position:** I observe that the trend is non-linear and not so apparent.
    - **Maximum Camber Position:** I observe some form of normal distribution from positions 0-100% of the chord length, with highest target values between 40-60% camber position.









2. **Feature Importance Plots:** For models such as random forest and linear regression, feature importance plots were generated to visualize the relative importance of each airfoil geometric parameter in predicting lift coefficients. These plots provide insights into which features had the most significant impact on aerodynamic performance. Observe that both plots provide similar results, except that linear regression coefficients have the added benefit of providing information on the whether the features proportionally or inversely influence the target value.

   - **Maximum Thickness** has the highest influence on the data, with an inverse proportion

   - **Maximum Camber** has the second highest influence on the data

   - **Maximum Camber Position** has the third highest influence on the data

   - **Maximum Thickness Position** has the least influence on the data, with an inverse proportion

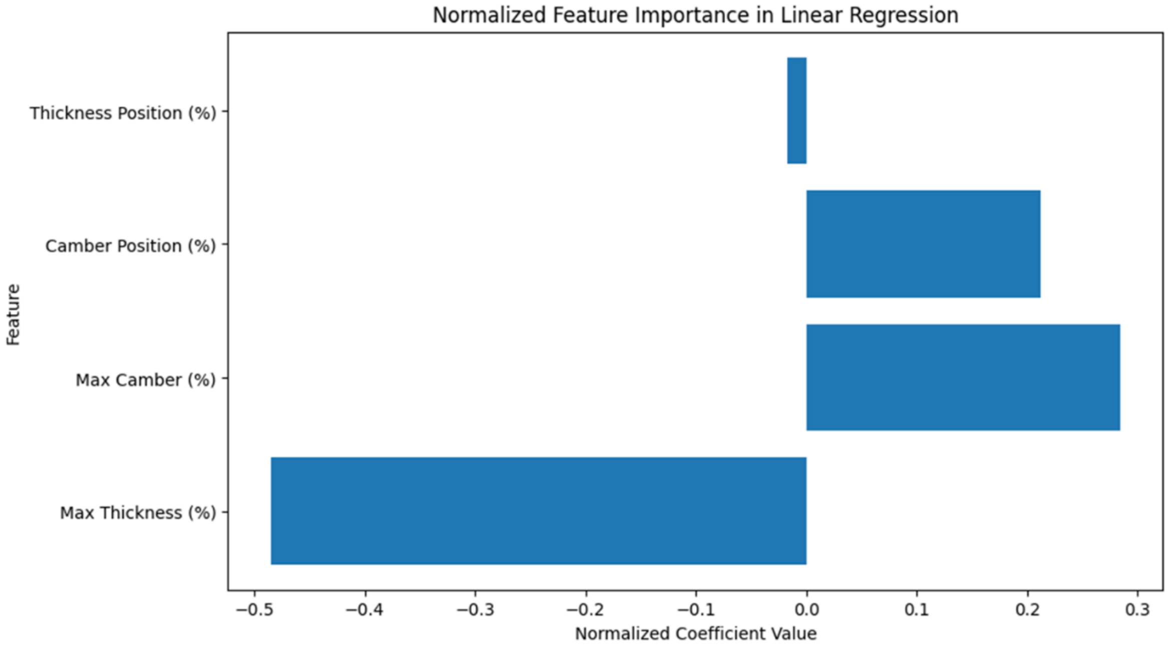


Figure 4 Feature Importance in Linear Regression

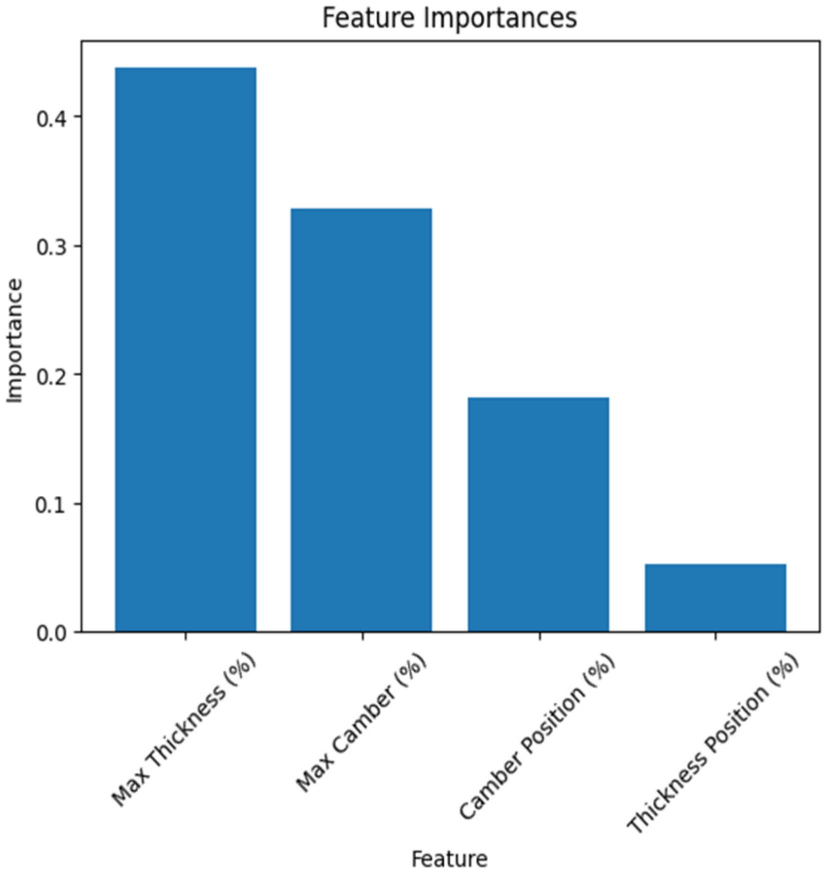


Figure 5 Feature Importance in Random Forest Regression

Through these visualizations, interesting trends and relationships between airfoil geometric parameters and lift coefficients were uncovered, guiding the selection of appropriate features and informing the modeling process.

5. RESULTS AND DISCUSSION

In this section, we present the results obtained from the application of various machine learning models to predict the target based on airfoil geometric parameters. We discuss the performance of each model, interpret the results, and provide insights into potential improvements.

Linear Regression

The linear regression model yielded a Mean Squared Error (MSE) of 10.71 and an R-squared (R2) score of 0.47. While the model provided interpretable coefficients, indicating the relative importance of each geometric parameter, its predictive performance was limited. The linear nature of the model might have been insufficient to capture the complex relationships between the features and the target variable.

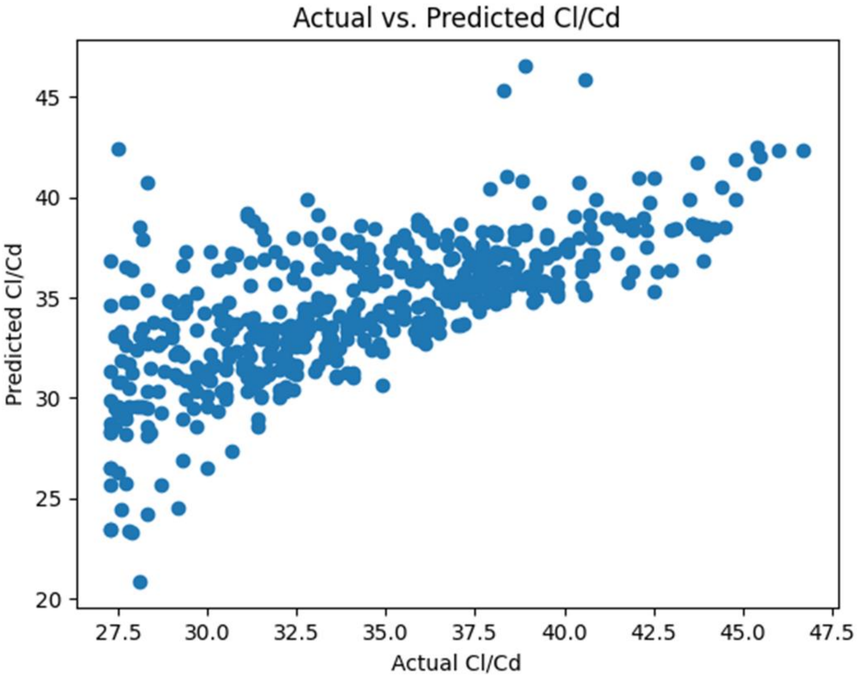


Figure 6 Scatter Plot of Actual vs Predicted Target Values in Linear Regression

Random Forest

The random forest regressor achieved a lower MSE of 6.20 and a higher R2 score of 0.69 compared to linear regression. The ensemble of decision trees allowed for capturing nonlinear relationships and interactions between features, resulting in improved predictive performance. Feature importance analysis revealed that the maximum thickness and camber position were the most influential factors in predicting lift coefficients.

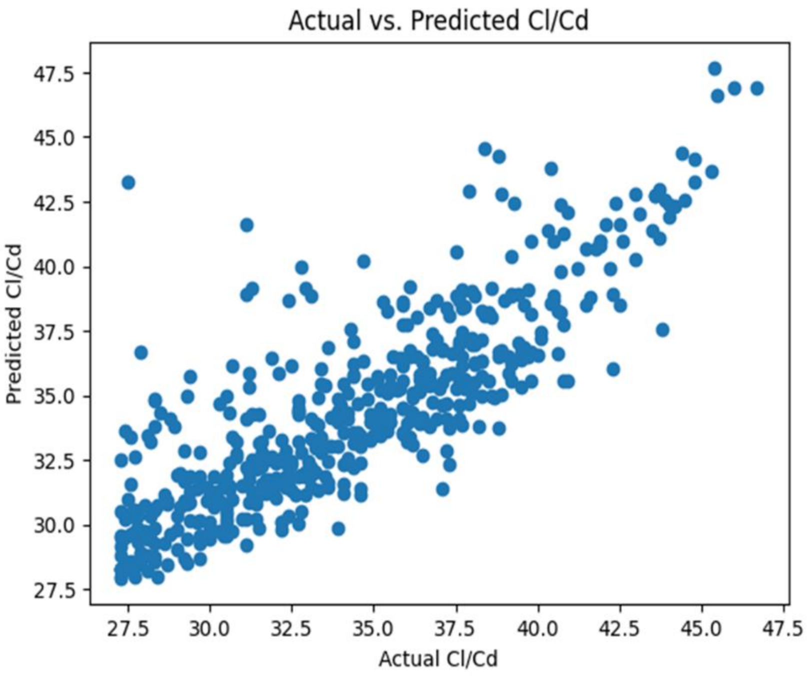


Figure 7 Scatter Plot of Actual vs Predicted Target Values in RF Regression

Deep Neural Network (DNN):

The DNN model produced comparable results to random forest, with an MSE of 6.27 and an R2 score of 0.69. Despite its complexity, the DNN effectively learned intricate patterns in the data, leveraging multiple hidden layers to capture nonlinearities. However, interpreting the model's weights and understanding its inner workings remained challenging.

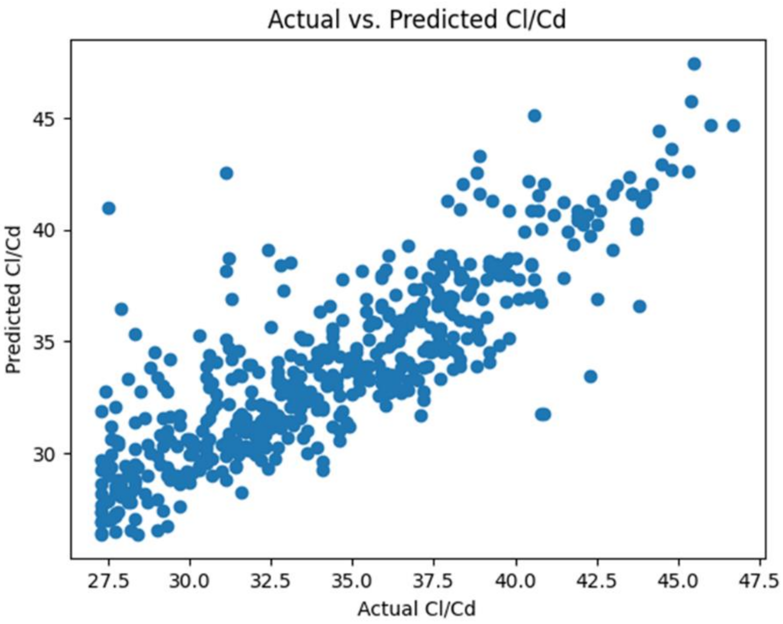


Figure 8 Scatter Plot of Actual vs Predicted Target Values in DNN

Support Vector Machine (SVM):

The SVM regressor achieved an MSE of 6.40 and an R2 score of 0.68, demonstrating competitive performance with random forest and DNN. However, the interpretability of SVM models is limited compared to linear regression and random forest. Also, there weren't many hyperparameters to tune to boost the performance further.

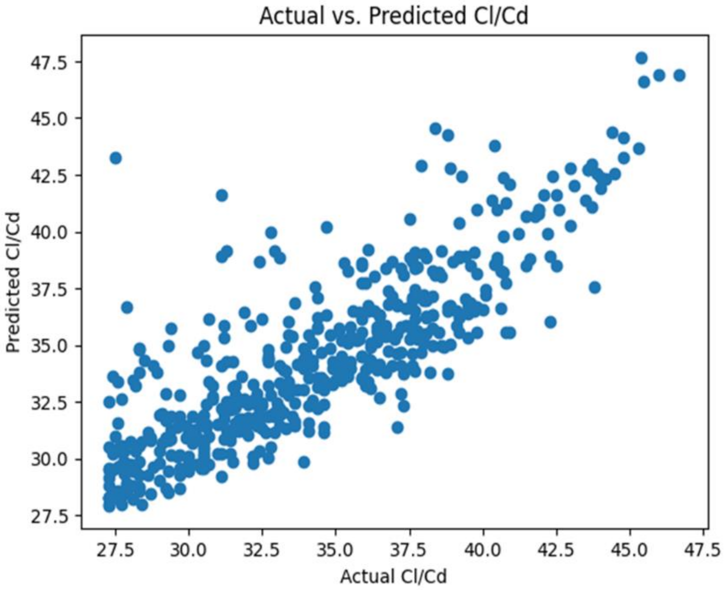
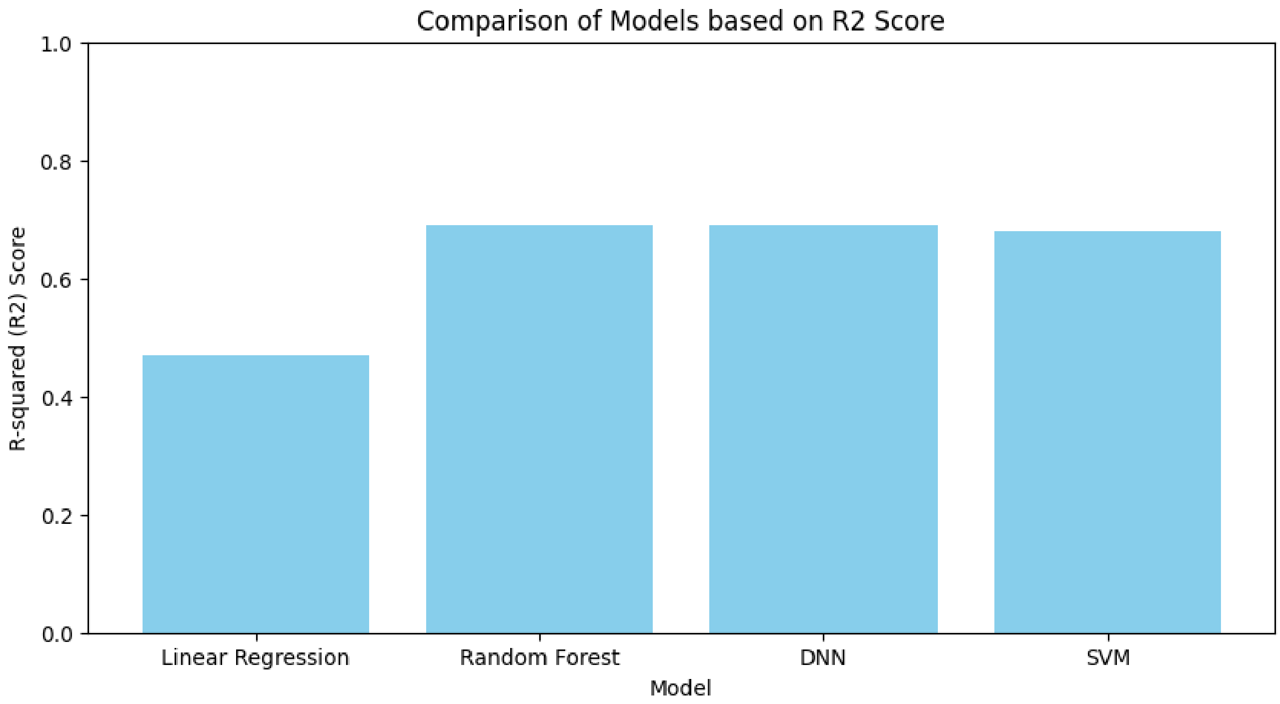
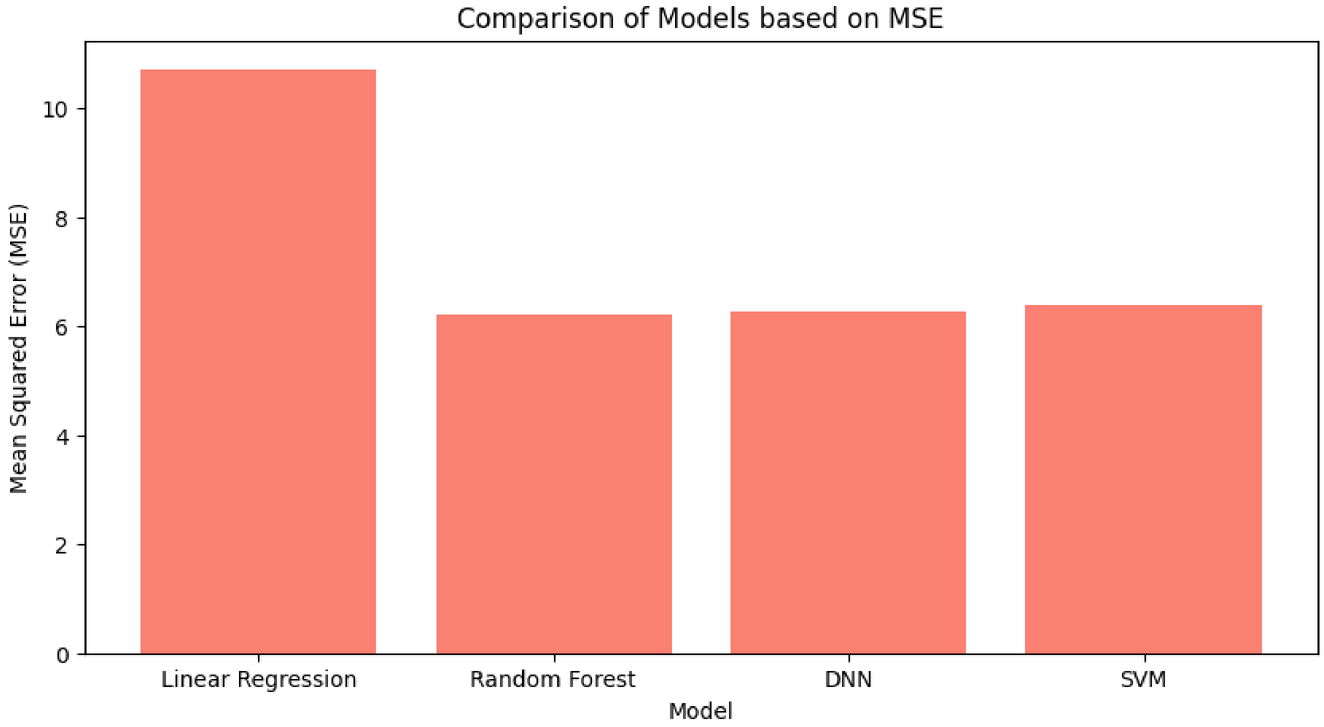


Figure 9 Scatter Plot of Actual vs Predicted Target Values in SVM Regression

Discussion

Overall, random forest and DNN emerged as the top-performing models, outperforming linear regression and SVM in terms of predictive accuracy. The nonlinear nature of random forest and DNN allowed for capturing complex relationships in the data, leading to improved performance. However, the interpretability of these models remains a concern, particularly for DNN, where understanding the learned representations is challenging.





To further improve model performance, I tried a number of things:

1. **Feature engineering** was explored on the Dense Neural Network. I added thee features:
   - Camber Thickness Difference: The difference between the maximum camber percentage and the maximum thickness percentage.

- Camber Thickness Ratio: The ratio of the maximum camber percentage to the maximum thickness percentage.
- Camber Position Ratio: The ratio of the maximum camber percentage to the camber position percentage.
- Thickness Position Ratio: The ratio of the maximum thickness percentage to the thickness position percentage.
- Camber Thickness Position Ratio: The ratio of the camber position percentage to the thickness position percentage.
- Camber Thickness Position Difference: The difference between the camber position percentage and the thickness position percentage.

Although these features seemed to be intuitively useful, unexpectedly, the model performed worse than it originally was. The model's inability to improve from a 0.7 R-2 score indicates that some aspects of the maximum lift-to-drag ratio require more information on the airfoil geometry than the four features can provide. I observe that the features play a stronger role in lift prediction than they do in drag prediction. The current model can be refined by incorporating additional parameters that describe the airfoil's geometry with greater fidelity. Characteristics such as curvature, twist, taper, and thickness distribution may unlock deeper insights and yield predictions with higher precision.

2. **Regularization** methods such as dropout between dense layers and lasso did not work even after performing grid searches.
3. I investigated with **cross-validation** with k =5 when I suspected that there was a bias in the train, test and validation splits, but the ensembled R-2 score was still the same.
4. **Outlier removal** was the last thing tested, but that only increased the mean squared error and reduced the R-2 score.

Model Combination and Future Work

An exploratory recommendation was made to combine model outputs, particularly using feature importance derived from Random Forest as an input weight to the Deep Neural Network. This innovative approach could form the basis for future studies, aiming to create a hybrid model that encapsulates both the predictive power of complex algorithms and the interpretability of simpler models.

Conclusion of Discussion

In conclusion, the models developed in this study serve as a substantial step towards the application of machine learning in the field of aerodynamics. While the results are promising, the discussion emphasizes the project's iterative nature, suggesting continual refinement and adaptation as more data and advanced modeling techniques become available.

6. CONCLUSION

This project has established that machine learning is a viable tool in predicting airfoil performance, with the obtained mean squared error (MSE) of 6.2 and R-squared (R2) score of 0.7 marking a promising commencement. These metrics reflect a significant potential for machine learning models to streamline the design process by providing quick, cost-effective insights into airfoil efficiency.

However, there remains room for enhancement. The current models can be refined by incorporating additional parameters that describe the airfoil's geometry with greater fidelity. Characteristics such as curvature, twist, taper, and thickness distribution may unlock deeper insights and yield predictions with higher precision.

Furthermore, the integration of model outputs, such as employing feature importance derived from one model to inform the inputs of another, like a Deep Neural Network (DNN), presents an exciting avenue for future exploration. This hybrid approach could refine model predictions, leveraging the strengths of various algorithms.

7. DATA AND CODE

Data were sourced from Airfoil Tools (http://www.airfoiltools.com/search/index?m%5BtextSearch%5D=&m%5BmaxCamber%5D=&m%5BminCamber%5D=&m%5BmaxThickness%5D=&m%5BminThickness%5D=&m%5Bgrp%5D=&m%5Bsort%5D=9&m%5Bpage%5D=18&m%5Bcount%5D=1638)

The dataset is contained in: https://docs.google.com/spreadsheets/d/e/2PACX-1vR-8c9mSoyIo2s0GhXUg8GftAenUnmcdoJ8lWRx-MYnEcwF78nQ4hAIHhn3cmoEGJbyYxrAw8IKubrg/pub?output=csv

The code is available on GitHub at: https://github.com/melvinaquartey/Machine-Learning-Class-Project/tree/main

8. REFERENCES

1. Moin, H., et al. "Data-Driven Optimization of Airfoils Using Machine Learning Techniques." Journal of Aircraft, vol. 58, no. 3, 2021, pp. 662-672.

2. Ahmed, A., et al. "Comparative Study of Machine Learning Techniques for Predicting Aerodynamic Coefficients of Airfoils." Aerospace, vol. 7, no. 5, 2020, Article 74.

3. Yu, B., Xie, L., & Wang, F. "Prediction of Airfoil Lift Coefficients Using Convolutional Neural Networks." Journal of Fluids Engineering, vol. 142, no. 12, 2020, Article 121105.

4. Kumar, A., & Kumar, R. "Application of Machine Learning Techniques in Aerodynamics: A Review." Aerospace Science and Technology, vol. 84, 2019, pp. 524-534.

5. Jones, R. T., et al. "Data-Driven Aerodynamic Design Using Machine Learning Techniques." AIAA Scitech 2020 Forum, 2020, pp. 1-12.

6. Zhao, Y., et al. "A Review of Artificial Intelligence Applications in Aerodynamic Design." IEEE Access, vol. 9, 2021, pp. 16707-16725.

7. Bayoumy, Ahmed, Nada, Ayman, & Megahed, Said. "Modeling Slope Discontinuity of Large Size Wind-Turbine Blade Using Absolute Nodal Coordinate Formulation." Proceedings of the ASME Design Engineering Technical Conference, vol. 6, 2012, doi: 10.1115/DETC2012-70467.