Ship Residuary Resistance Prediction with Machine Learning

Poorya Khorsandy[1] Seyed Saeed Hayati[2]

[1,2]Department of Maring Engineering, Khorramshahr University of Marine Science and Technology, Khorramshahr, Iran.

***Abstract*—This paper explores machine learning techniques as cost-effective, accurate, and relatively fast alternative to traditional methods for yacht residuary resistance calculation. Currently, solving CFD equations is a well-known method in residuary resistance calculation. However, solving CFD equations requires an iterative process which is very time consuming and needs high performance processing units in addition. Instead of solving CFD equation for any ship geometry, we proposed to apply machine learning on present data obtained from previous CFD-based calculation or measurements of real models. We used Delft yacht hydrodynamics dataset obtained from 308 experiments in calculating residuary resistance corresponding to different ship parameters. This dataset is divided into training and test data and the results of several learning-based method are compared in terms of RMSE. Our results demonstrate that XGBoost regressor is superior that others with RMSE=0.54**

Keywords—ship residuary resistance, machine learning, Delft dataset, regression.

I. INTRODUCTION

Accurate ship residuary resistance prediction is vital in naval architecture. Naval architecture and maritime engineering strive to enhance waterborne vessel efficiency and performance, with ship residuary resistance prediction being a pivotal aspect influencing design, fuel efficiency, and operational costs [1]. Traditional methods involve resource-intensive computational fluid dynamics (CFD) simulations or costly physical experiments. In this research several machine learning-based methods are investigated as a tool for accurate ship residuary residuary resistance prediction for an unseen ship geometry by using available data obtained from previous physical experiments or simulations. Fig.1 shows the basics of ship residuary resistance prediction using learning-based methods.

Utilizing the Delft yacht hydrodynamics dataset, which comprises 308 experiments across 22 hull forms, our study focuses on key attributes such as buoyancy position, Froude number, Length-Displacement Ratio, Beam-Draught ratio, Length-Beam ratio, and Prismatic Coefficient. These attributes are employed as the input feature vector in our analysis.We detail machine learning algorithms, preprocessing steps, and present results showcasing superior accuracy and efficiency. The paper explores insights from the analysis, emphasizing the potential of machine learning to revolutionize naval architecture, concluding with implications and avenues for future research in ship hydrodynamics and machine learning.

II. BACKGROUND OF RESEARCH

Before delving into the intricacies of ship residuary resistance prediction and the application of machine learning and CFD, it is crucial to establish a foundational understanding of the key technologies shaping this research. This section provides a concise overview of Artificial Intelligence (AI), Machine Learning, Deep Learning, and their relevance and their significance in overcoming challenges, especially in scenarios where the number of available data is limited. Application of CFD in ship residuary resistance calculation and comparison with AI-based methods are also given in this section.

A. Artificial Intelligence (AI)

AI is a multidisciplinary field of computer science and engineering that focuses on the creation of intelligent systems capable of performing tasks that typically require human intelligence. These tasks include problem-solving, decision-making, natural language understanding, and pattern recognition. AI systems aim to mimic human cognitive functions and adapt to new information and situations [2].

Machine Learning is a subset of AI that provides systems with the ability to learn and improve from experience, without being explicitly programmed. Machine learning algorithms use data to identify patterns, make predictions, and optimize performance on specific tasks. It encompasses a wide range of techniques, including supervised learning (where models are trained on labeled data), unsupervised learning (where models identify patterns without labeled data), and reinforcement learning (where agents learn through interactions with an environment) [3].

Deep Learning is a specialized subfield of machine learning that focuses on artificial neural networks, inspired by the structure and function of the human brain. Deep learning models, particularly deep neural networks, consist of multiple layers of interconnected artificial neurons (nodes). These networks can automatically extract and learn hierarchical features from data, making them particularly powerful for tasks such as image and speech recognition, natural language processing, and complex pattern recognition [4].

In ship residuary resistance prediction, where extensive data is often a challenge, machine learning methods offer a tailored solution. Unlike traditional rule-based or physics-based methods that demand large datasets and specialized knowledge, several regression methods applied in this paper excels in low data scenarios.

These machine learning models exhibit a unique capability to discern intricate patterns within modest datasets with high accuracy. Relying on data-driven principles, they extract valuable insights even when faced with limited information [5]. This adaptability is crucial when dealing with real-world datasets that may be limited in size or complexity.

Our research justifies the selection of XGBoost regressor [6] by proving its effectiveness in analyzing the Delft yacht hydrodynamics dataset [7]. Despite the dataset's modest size, XGBoost excels in accurately predicting ship residuary resistance. This adaptability and data-driven approach position it as a compelling choice to address the challenges in low data scenarios for ship residuary resistance prediction.

Our research supports the choice of XGBoost regressor by proving its effectiveness in analyzing the Delft yacht hydrodynamics dataset, Despite the dataset's limited size, XGBoost excels in identifying significant patterns and making precise predictions for ship resistance. This adaptability, grounded in a data-driven approach, establishes it as a compelling solution to tackle challenges in scenarios with limited data for ship residuary resistance prediction. Ship resistance computation with CFD

CFD is a specialized field in naval architecture and engineering that focuses on the numerical simulation and analysis of fluid flow interactions with solid structures, particularly in the context of ship hydrodynamics [8]. The primary objective of CFD is to model and understand the complex behaviors of fluid dynamics around ships. CFD is grounded in the application of the Navier-Stokes equations given in (1) , which describe the fundamental principles governing fluid flow [9].

(1)

In (1), represents the density of the fluid. is the material derivative of velocity, describing the acceleration of the fluid at a specific point. represents the pressure gradient, indicating how pressure changes in space. *represents the divergence of the stress tensor . accounting for viscous stresses in the fluid. represents the force per unit volume due to gravity acting on the fluid.*

By solving these equations numerically, CFD simulations provide insights into phenomena such as resistance, drag, and flow patterns around a ship's hull. CFD is renowned for its ability to deliver highly precise and detailed results, allowing for a comprehensive understanding of the forces exerted on a ship, however, successful implementation of CFD requires significant computational resources and expertise in fluid mechanics and numerical methods [10].

The simulation process in CFD involves solving complex mathematical equations iteratively to capture the dynamic fluid-structure interactions accurately. Although CFD is very accurate, it is also very time consuming which prevent it to be a multilateral method for ship resistant calculation. The main contribution of this paper is developing a fast method with an accuracy close to CFD , bringing efficiency to ship residuary resistance prediction.

B. *Machine Learning for Ship residuary Resistance Prediction*

In the pursuit of efficient ship residuary resistance prediction, machine learning emerges as an alternative to conventional methods, with high precision and reduced computational demands. Significant strides have been made in harnessing artificial neural networks (ANNs) for ship residuary resistance prediction. ANNs showcase an exceptional ability to capture intricate nonlinear relationships within the data, providing reliable predictions while substantially reducing computational costs compared to traditional Computational Fluid Dynamics (CFD) simulations.

Similarly, the application of support vector machines (SVMs) for ship residuary resistance prediction has gained traction. The competitive accuracy of SVMs and their effectiveness in handling high-dimensional data, particularly the geometric attributes of ship hulls, highlight their valuable contributions.

Further advancements in deep learning have left a lasting impact. The use of convolutional neural networks (CNNs) to extract features from ship hull images has resulted in remarkably accurate predictions of resistance. This innovative approach underscores the potential synergy between machine learning and image analysis in the intricate field of ship hydrodynamics, paving the way for new possibilities in enhanced predictive capabilities.

C. *CFD VS. MACHINE LEARNING*

Here, some of the main advantages and challenges in CFD-based and machine learning-based ship residuary resistance calculation are given respectively.

The main advantage of CFD-based method is its precision. CFD provides highly accurate and detailed results, offering a comprehensive understanding of fluid dynamics around a ship's hull. In addition, CFD provides In-Depth analysis. It enables a thorough examination of hydrodynamic aspects, including resistance, drag, and flow patterns.

The main challenge using CFD is its computational burden due to its iterative process for solving the equations. This makes CFD a resource-intensive and time-consuming method for ship residuary resistance calculation. Successful CFD implementation requires expertise in fluid mechanics and numerical methods which limits its accessibility. The above-mentioned challenges in CFD lead to financial considerations due to the need for high-performance processors and specialized skills.

A notable advantage of machine learning methods is adaptability. Machine learning models, such as the regression model employed in this research, excel in low data scenarios, adapting well to limited datasets. In addition, machine learning methods are very efficient. They offer rapid predictions, making them time-efficient and suitable for scenarios where quick assessments are crucial. In comparison with CFD, machine learning methods are cost-effective, particularly in scenarios with limited resources, as they do not require the same level of computational infrastructure as in CFD.

There are also challenges in machine learning methods. Machine learning models might lack the interpretability of CFD results, making it difficult to understand the underlying physical mechanisms.

In summary, while CFD offers unparalleled precision, machine learning presents a cost-effective and efficient alternative, particularly in scenarios with limited resources.

In the next section, Delft dataset is introduced and more details about the input and target parameters are give.

III. DELFT DATASET DESCRIPTION

The Delft yacht hydrodynamics dataset serves as the foundational source of information for this research, comprising a comprehensive collection of full-scale experiments conducted at the Delft Ship Hydromechanics Laboratory. These experiments were specifically designed to investigate and understand ship resistance, providing a rich and diverse set of data for our analysis.

The dataset encompasses a total of 308 full-scale experiments, providing a substantial volume of data for analysis. Each experiment corresponds to a unique combination of hull parameters and conditions.

The experiments include 22 distinct hull forms, all derived from a parent hull form closely related to the 'Standfast 43' design by Frans Maas. These variations in hull geometry coefficients and the Froude number offer a diverse set of scenarios to evaluate the predictive capabilities of machine learning models.

The dataset includes several key attributes, that are essential for the analysis and prediction of ship resistance. These parameters which called features in machine learning literature are as follows.

Longitudinal Position of the Center of Buoyancy (LC): It is a dimensionless attribute indicating the longitudinal location of the center of buoyancy on the ship.

Prismatic Coefficient (PC): This feature is an important hull design parameter that characterizes the shape of the hull. It is dimensionless and provides insights into the vessel's form.

Length-Displacement Ratio (L/D): This dimensionless ratio describes the relationship between the ship's length and displacement, providing information about its size and shape.

Beam-Draught Ratio (B/Dr): This is also a dimensionless parameter that reveals the proportions of the ship's beam (width) to its draught (depth).

Length-Beam Ratio (L/B): This dimensionless feature represents the ratio of the ship's length to its beam, providing insights into the vessel's proportions.

Froude Number (Fr): Froude number characterizes the flow regime around the ship. It is a critical parameter for understanding hydrodynamic behavior.

The target parameter is Residuary Resistance (Rr) per unit weight of displacement. This dimensionless variable quantifies the resistance experienced by the ship and it is considered as the target variable in our experiments in ship resistance prediction using machine learning models.

IV. DATA PREPROCESSING

Data preprocessing is a crucial step in ensuring the quality and reliability of our ship resistance prediction models. Here, we outline the key elements of data preprocessing.

A. Missing Data Analysis

In data analysis, "missing data" refers to the absence of values or information for certain data points within a dataset. Missing data can occur for various reasons, such as data collection errors, incomplete records, or the omission of certain attributes during data collection. In our rigorous data analysis process, we meticulously examined the dataset for the presence of missing values. We are pleased to report that no missing data was identified across any attributes within the entire Delft dataset. This remarkable absence of missing values ensures that our analysis is based on complete and reliable data, eliminating the need for imputation or data recovery techniques.

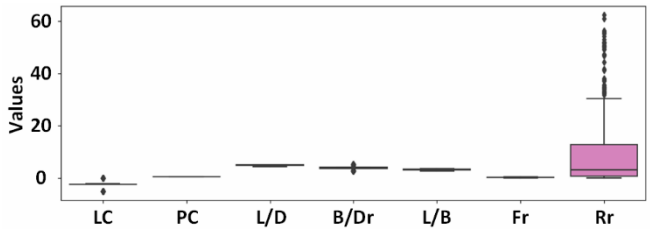


Fig.1: Outlier analysis for input features and target parameter.

B. Outlier Analysis

Outliers, in the context of data analysis, are data points that deviate significantly from the typical or expected patterns in a dataset. They are observations that are either exceptionally high (positive outliers) or exceptionally low (negative outliers) compared to the majority of the data. Outliers can arise due to various reasons, such as measurement errors, data entry mistakes, or genuine rare events.

In our analysis, we conducted an outlier analysis using a common outlier detection technique known as a boxplot [11] for all attributes within the dataset. The result of outlier analysis for input features and target parameter is shown in Fig. 1. We can see that only 'Rr' has an outlier, and the other attributes have no outliers. We can't change 'Rr' because it affects our result. As shown in Fig. 1, the outlier in Rr is high/low!?

C. Skewness Assessment

Skewness is a statistical measure that quantifies the asymmetry in the distribution of data. In essence, it describes the tendency of data to be skewed or biased towards one tail of the distribution, whether it's the left (negative skew) or the right (positive skew). Understanding skewness is essential in data analysis because it impacts the interpretation of data and the performance of predictive models.

Many statistical models, including linear regression, assume that data is normally distributed. Skewed data violates this assumption and can lead to inaccurate model predictions. Skewed data can make it challenging to interpret statistical results. It can affect measures like means and medians, which are commonly used to summarize data. Skewed data can negatively impact machine learning algorithms, as they may struggle to capture patterns in highly skewed data.

In our dataset, we conducted a thorough assessment of skewness across all attributes. It is notable that, while most attributes exhibited normal skewness, we observed non-normal skewness specifically in the . Given that predicting ship residuary resistance is the primary goal of our analysis, we acknowledge the presence of skewness in this attribute but emphasize that it cannot be altered or corrected. This is

because the skewness in the " residuary Resistance" attribute is inherent to the nature of residuary resistance measurements and represents real-world variations that our analysis seeks to capture accurately. Therefore, in our analysis, we accept the presence of non-normal skewness in the attribute as a characteristic of the data that aligns with the objectives of our research. While skewness assessment is crucial in many scenarios, our specific focus on ship resistance prediction prioritizes the accuracy of predictions over normality in the distribution of resistance values.

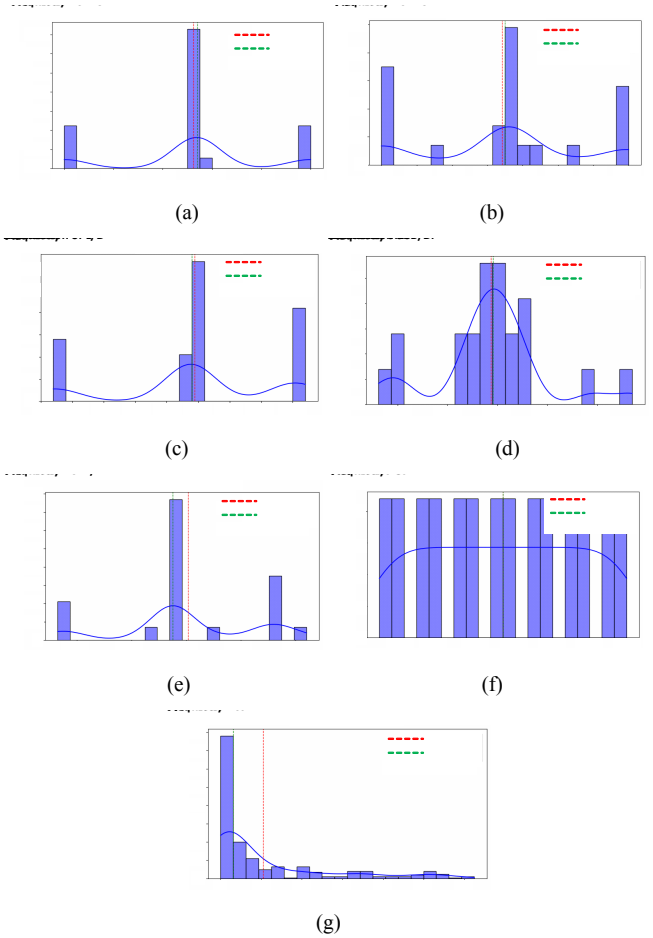


Fig.2: Skewness assessment of (a) LC, (b) PC, (c) L/D, (d) B/Dr, (e) L/B, (f) Fr, (g) Rr.

D. *Hyperparameter Tuning*

Hyperparameter tuning is a critical aspect of optimizing machine learning models for predictive accuracy and performance. Hyperparameters are configuration settings that are not learned from the data but are set prior to training the model. Properly tuned hyperparameters can significantly impact the model's ability to capture complex relationships in the data and make accurate predictions.

Grid search [12] is a systematic approach to hyperparameter tuning that involves evaluating multiple combinations of hyperparameters to identify the best-performing set. It works by specifying a range of values or discrete choices for each hyperparameter, and then exhaustively evaluating all possible combinations of these values.

For each hyperparameter, you define a range or a set of possible values to consider. Grid search systematically evaluates every combination of hyperparameter values within the defined ranges. For each combination, the model's performance is assessed using a chosen evaluation metric, such as Root Mean Squared Error (RMSE). The combination of hyperparameters that results in the best performance metric (e.g., lowest RMSE) is selected as the optimal set of hyperparameters. Calculation of RMSE is obtained follows:

(2
)

where, is the number of data points, is the -th measurement, and is its corresponding prediction.

In our research, we applied grid search to fine-tune the hyperparameters of each machine learning algorithm used in our ship residuary resistance prediction models. This involved defining ranges or possible values for hyperparameters specific to each algorithm, such as learning rates, maximum tree depths, or regularization strengths. Grid search systematically explored these hyperparameter combinations, allowing us to identify the set of hyperparameters that yielded the highest predictive accuracy for each algorithm. This optimization process is essential to ensure that our machine learning models are configured optimally for the specific task of ship residuary resistance prediction. By employing grid search, we aimed to enhance the performance of each algorithm, ultimately leading to more accurate predictions of ship resistance, contributing to the overall success of our analysis.

V. MACHINE LEARNING ALGORITHMS AND TECHNIQUES

In this section, we delve into the machine learning algorithms and techniques employed for predicting ship resistance. Each algorithm brings unique characteristics to the task, contributing to our comprehensive analysis.

In our methodology, an essential step in the machine learning process is the careful partitioning of our dataset into training and testing sets. This partitioning is critical for evaluating the performance and generalizability of our machine learning models.

We adopted a standard data splitting ratio, with 70% of the dataset allocated for training purposes and the remaining 30% reserved for testing. This ratio, which divides the data into a 70-30 split, strikes a balance between training the model on a substantial portion of the data and having a sufficiently large test set to assess its predictive capabilities effectively.

This data splitting strategy ensures that our machine learning models are trained on a diverse and representative subset of the data, allowing them to learn underlying patterns and relationships. Simultaneously, the independent testing set serves as a robust evaluation benchmark, enabling us to measure the models' performance on unseen data accurately.

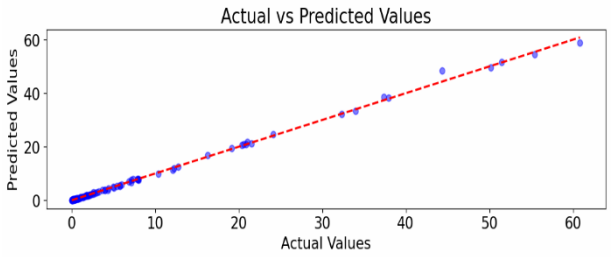
A. *XGBoost Regressor*

XGBoost is a powerful ensemble learning technique that builds multiple decision trees sequentially. It begins by constructing a simple decision tree and then corrects its errors in subsequent trees. This iterative process combines the predictions from multiple trees to create a robust predictive model. XGBoost is chosen for its ability to handle complex relationships in the data and its excellent predictive performance. Its sequential nature allows it to adapt well to

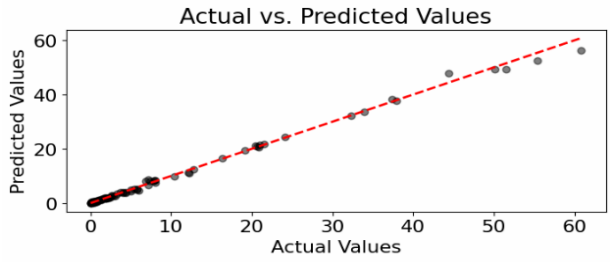
different datasets, making it suitable for ship residuary resistance prediction.

B. *CatBoost Regressor*

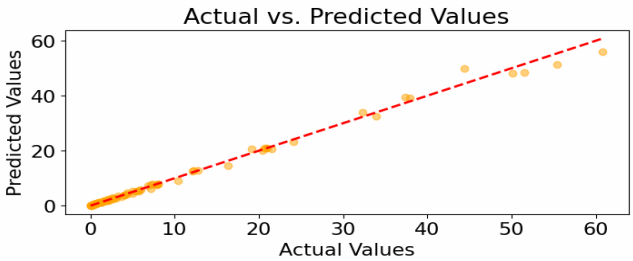
CatBoost [13] is a gradient boosting algorithm that excels in handling categorical features. It employs an ordered boosting technique, which optimally integrates trees into the model, reducing overfitting while enhancing predictive accuracy. CatBoost's ability to work with categorical data is advantageous in ship residuary resistance prediction, where certain attributes may be categorical. Its ordered boosting strategy provides a robust approach to modeling the data.



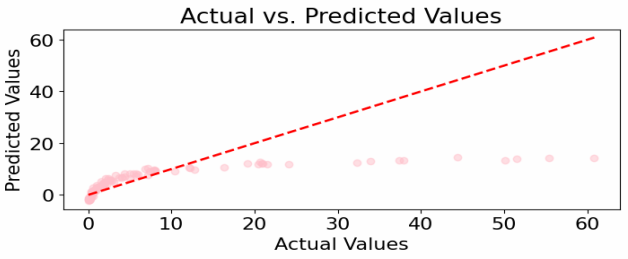
(a)



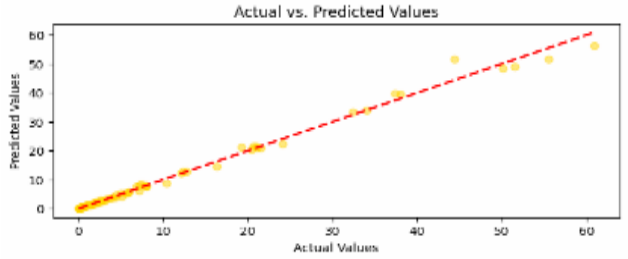
(b)



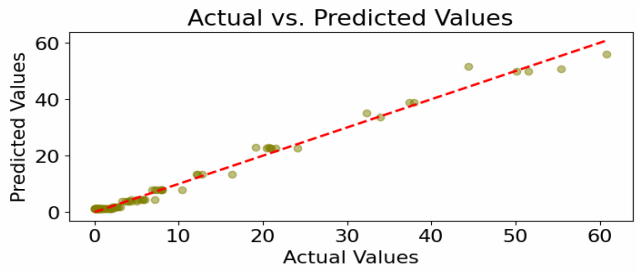
(c)



(d)



(e)



(f)

Fig. 3: Comparing ship residuary resistance prediction with real data of Dleft dataset. (a) XGB regressor, (b) CatBoost, (c) ExtraTree, (d) SVR, (e) RandomForest, (f) Adaboost.

C. *ExtraTrees Regressor*

ExtraTrees [14] is an ensemble method that builds decision trees with random features and thresholds. It introduces an additional layer of randomness by selecting the best split from a random subset of features. ExtraTrees Regressor is chosen for its capacity to mitigate overfitting and produce stable predictions. This randomness in feature selection enhances model diversity, which leads to improving accuracy.

D. *Support Vector Regression (SVR)*

SVR [15] extends Support Vector Machines (SVMs) to regression tasks. It seeks to find a hyperplane that best fits the data, with a margin that minimizes the error between predicted and actual values. SVR is a well-established regression technique suitable for modeling complex relationships. Its ability to handle non-linear data patterns makes it a valuable addition to our analysis.

E. *RandomForest Regressor*

RandomForest [16] is an ensemble learning method that constructs a multitude of decision trees and combines their predictions. It introduces randomness through bootstrap sampling and random feature selection. RandomForest regressor is selected for its ability to mitigate overfitting and deliver stable predictions. The ensemble of trees provides robust modeling capabilities.

F. *AdaBoost Regressor*

AdaBoost [17] is an ensemble technique that iteratively adjusts the weights of misclassified samples to improve model performance over iterations. AdaBoost regressor is chosen for its adaptability and capacity to enhance model accuracy through iterative training. It excels in scenarios where data may be noisy or complex.

VI. EXPERIMENTAL RESULTS

In this section, we present the results of our ship residuary resistance prediction models, showcasing their performance based on the RMSE metric. RMSE quantifies the average error between predicted and actual ship residuary resistance values, providing a comprehensive measure of predictive accuracy. The result of using machine learning methods mentioned in previous section are given in Fig.3.

In Fig. 3, the target parameter () predicted by different machine learning methods are compared with actual values obtained in real experiments in Delft dataset. It worth mentioning that predicted values for in Fig.3 are corresponding to input parameters that were not used in raining phase. As shown in Fig. 3, the best prediction of

values are given by XGBoost regressor and the worst prediction is obtained from SVR. Comparison of different machine learning method used in this paper in terms of RMSE are given in Table 1.

Table. 1: RMSE results of machine learning methods.

Method	**XGBoost**	**CatBoost**	**ExtraTrees**
RMSE	0.54	0.82	1.06
Method	**SVR**	**RandomForest**	**AdaBoost**
RMSE	10.62	1.18	1.4

As shown in Fig.3 and Table.1, The XGBoostRegressor model stood out with the lowest RMSE of 0.54, showcasing exceptional predictive accuracy. This is attributed to XGBoost's robustness in handling complex data relationships, especially in scenarios with limited data, such as ship residuary resistance prediction.

VII. CONCLUSION

In this research, it is proposed to use machine learning methods to predict ship residuary resistance instead of using time-consuming CFD equations or conducting costly physical experiments. Here, several machine learning methods are applied on Delft yacht hydrodynamic dataset. The dataset is obtained by using several experiments with real ship models and precisely measuring the residuary resistance corresponding to different ship parameters. We use 70% of the dataset in learning phase and remaining 30% in test phase. Our experimental results show that XGBoost regressor has best performance with RMSE = 0.54. It should be noted that the ship residuary resistance results obtained in test phase of other machine learning method is also acceptable except for SVR with RMSE = 10.62.

REFERENCES

[1] J. Carlton, Marine propellers and propulsion. Butterworth-Heinemann, 2018.

[2] S. J. Russell and P. Norvig, Artificial intelligence a modern approach. London, 2010.

[3] C. M. Bishop and N. M. Nasrabadi, Pattern recognition and machine learning, vol. 4, no. 4. Springer, 2006.

[4] Y. Bengio, I. Goodfellow, and A. Courville, Deep learning, vol. 1. MIT press Cambridge, MA, USA, 2017.

[5] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction, vol. 2. Springer, 2009.

[6] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[7] M. Kelly, R. Longjohn, and others, "UCI machine learning repository." https://archive.ics.uci.edu, 2007.

[8] J. D. Anderson and J. Wendt, Computational fluid dynamics, vol. 206. Springer, 1995.

[9] F. M. White and J. Majdalani, Viscous fluid flow, vol. 3. McGraw-Hill New York, 2006.

[10] H. K. Versteeg and W. Malalasekera, An introduction to computational fluid dynamics: the finite volume method. Pearson education, 2007.

[11] J. W. Tukey and others, Exploratory data analysis, vol. 2. Reading, MA, 1977.

[12] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization.," J. Mach. Learn. Res., vol. 13, no. 2, 2012.

[13] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," Adv. Neural Inf. Process. Syst., vol. 31, 2018.

[14] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Mach. Learn., vol. 63, pp. 3–42, 2006.

[15] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," Stat. Comput., vol. 14, pp. 199–222, 2004.

[16] A. Liaw, M. Wiener, and others, "Classification and regression by randomForest," R news, vol. 2, no. 3, pp. 18–22, 2002.

[17] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," Adv. Neural Inf. Process. Syst., vol. 9, 1996.