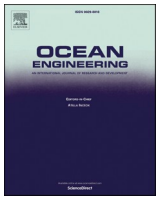
Ocean Engineering 303 (2024) 117679

Contents lists available at ScienceDirect

Ocean Engineering

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



Research paper

A novel dataset and feature selection for data-driven conceptual design of offshore jacket substructures

Han Qian [a,*], Emmanouil Panagiotou [b], Mengyan Peng [a], Eirini Ntoutsi [c], Chongjie Kang [a], Steffen Marx [a]

[a] Institute of Concrete Structures, Technische Universität Dresden, August-Bebel-Straße 30/30A, 01219, Dresden, Germany
[b] Department of Mathematics & Computer Science, Freie Universität Berlin, Arnimallee 7, 14195, Berlin, Germany
[c] Research Institute CODE, Universität der Bundeswehr München, Carl-Wery-Str. 18, 81739, Munich, Germany

ARTICLE INFO

Keywords:
Offshore jacket substructure
Conceptual design
Data-driven method
Machine learning
Dataset
Feature selection

ABSTRACT

Conceptual design is crucial for designing offshore jacket substructures because it sets the direction for the entire design process. Nevertheless, conventional simulation-based optimization methods for jacket conceptual design face challenges, such as high computational costs and restricted optimization objectives. This paper proposes a data-driven method for offshore jacket conceptual design using machine learning (ML). First, a novel dataset of completed and under-construction jackets worldwide was established as the cornerstone of ML. The dataset comprised "in-action" data capturing key structural parameters of jackets and information on design boundary conditions. Subsequently, different features were comprehensively selected to identify and visualize their correlations for an interpretable data-driven design, ensuring the effectiveness of the dataset for training the ML models. Finally, random forest and eXtreme gradient boosting models were trained on the data from the selected feature subsets and then employed to predict individual jacket structural parameters. The predictive performance of the models indicates that the dataset and feature selection can capture the fundamental and shared characteristics of well-designed jackets, thereby improving the accuracy and efficiency of the conceptual design process. This study suggests the potential of a data-driven conceptual design for offshore jacket substructures.

1. Introduction

The offshore wind industry, initiated in Denmark in 1991, has experienced rapid global growth, particularly in Europe, Asia, and North America, with turbine sizes expanding from 450 kW in the 1990s to the current standard of 6–8 MW (Seidel, 2014). The development of turbines of up to 15 MW has driven the evolution of offshore substructures to support larger turbines against increased loads and withstand challenging ocean conditions. Various types of offshore substructures exist, including bottom-fixed substructures such as monopiles, jackets, tripods, gravity-based structures, and floating platforms, each with its own suitable application regions. Monopile is the most common type of offshore substructure primarily used at shallow to medium water depths (Damiani et al., 2016). However, considering the increased rated power of wind turbines and the development of wind farms in deeper waters, jacket substructures are more competitive than other bottom-fixed substructures in the offshore wind industry (Marjan and Huang, 2023). According to a global estimate spanning 2021 to 2025, a minimum of 1083 jacketed turbines will be deployed worldwide. This tendency positions jackets as the second most popular bottom-fixed substructure type after monopiles (Offshore Engineer Magazine, 2020). A jacket is a truss-like lattice structure consisting of welded tubular steel members. Owing to the nature of its topology (e.g., the wide base and multiple legs for support and anchoring) and the relatively small diameter of the tubular members, it has higher stability and load-bearing capacities than other bottom-fixed substructures. In principle, jacket substructures have lower self-weight than other substructures. Therefore, they can use less material to achieve the same strength. This characteristic could result in cost savings in procurement and fabrication (Chen et al., 2016).

The increasing demand for offshore wind energy underscores the growing importance of optimized designs for offshore substructures. In accordance with developed design codes and industry standards such as DNV–OS–J101 (DNV GL, 2014) and IEC 61400–3 (IEC, 2019), the design process of offshore substructures in the offshore wind industry can generally be divided into three sequential phases: conceptual design,

* Corresponding author.
E-mail address: han.qian@tu-dresden.de (H. Qian).

https://doi.org/10.1016/j.oceaneng.2024.117679
Received 24 November 2023; Received in revised form 10 March 2024; Accepted 25 March 2024
Available online 6 April 2024

iterative design, and detailed design. In the conceptual design phase, the design basis is established at the beginning of the process. These include specific design requirements, applicable standards and codes, project descriptions, dimensions, site conditions, and assessment methods. Based on the design information, the initial structural and cost models of the substructure are determined for the subsequent design phase (Stolpe and Sandal, 2018). In the iterative design phase, the ultimate limit state (ULS) and fatigue limit state (FLS) are assessed, refining the geometry of the substructure based on these assessments until all stakeholder requirements are met. The detailed design phase involves preparing the final designs, drawings, and documentation for certification and construction. The conceptual design phase is critical for setting the direction of the subsequent design process. Decisions made during this phase can significantly influence the design efficiency, cost, and technical feasibility of a project. The preliminary topology in the conceptual design phase determines the number of iterations required to refine the structure during the iterative design phase (Seidel, 2010).

The optimization of offshore jacket substructures in the conceptual design phase has recently attracted increasing academic interest. To date, most investigations have focused on simulation-based methodologies. Chew et al. (2015, 2016) introduced a global optimization framework using an analytical gradient-based method to reduce the structural mass of offshore substructures. Oest et al. (2017) applied a similar method to an OC4 reference jacket, achieving a 40% mass reduction. Alternative jacket optimization approaches include genetic algorithm techniques, as presented by AlHamaydeh et al. (2017) using a genetic algorithm with domain trimming, and the use of surrogate models for efficiency, such as the Gaussian process regression (GPR) models incorporated in Häfele et al. (2018). Despite these advances, the feasibility of these methods under diverse designs and conditions requires further investigation. To this end, a jacket-sizing tool was developed by the National Renewable Energy Laboratory (NREL) (Damiani and Song, 2013; Damiani et al., 2017). Recently, multi-objective methods for structural optimization have been proposed to achieve balanced and efficient design solutions (Berger et al., 2021; Mathern et al., 2022).

However, the limitations of these simulation-based methodologies are also evident, including high computational requirements for solving objective functions, limited scope of optimization objectives, simplified assumptions regarding loads and structural code checks, and a lack of robust evaluations of the proposed methods. This study proposes an innovative, transformative shift toward data-driven methodologies using machine learning (ML) for the jacket's conceptual design. The insights gained from the existing designs can contribute to developing robust and reliable offshore jacket substructures. To this end, considering completed and under-construction offshore jackets in operational offshore wind farms is essential because these jackets have already been designed to support real loads in wind farms and have demonstrated adequate load-bearing capacities. In this context, ML techniques offer great potential for learning from data, enabling machines to simulate human brain-like learning and thinking processes and to identify patterns and relationships in large datasets (Alpaydin, 2020). ML techniques are being increasingly explored and applied to challenging tasks in structural designs. They apply not only to automated design processes for individual structural components, such as the prediction of optimum prestressing of concrete members using artificial neural networks (ANN) (Torky and Aburawwash, 2018), but also to large-scale structural designs, such as ML-based assistant for the conceptual design of steel frame halls (Fisch et al., 2023).

Despite their potential benefits, the aforementioned studies offer limited insights into applying ML techniques to the jacket conceptual design. This is primarily attributed to the complex nature of jacket topologies, which present a wide range of design options, along with the challenges in collecting sufficient design data for training ML models. This study aims to bridge this gap by developing a novel and comprehensive dataset of completed and under-construction offshore jackets worldwide. The dataset contains the main structural parameters of jackets and relevant boundary condition variables, such as information on site conditions, rotor-nacelle-assembly (RNA), and tower. This dataset forms the foundation of a data-driven method for offshore jacket conceptual design leveraging ML techniques, such as predictive models of random forest (RF) and eXtreme gradient boosting (XGBoost). Applying the dataset, feature selection based on correlation analysis can identify the most relevant and influential features (also referred to as parameters or variables in this study) within the dataset and uncover explicit and implicit correlations between the structural parameters of jackets and the boundary condition variables. According to the values of the correlation coefficients and structural design requirements, appropriate feature subsets can be selected as inputs for ML-based predictive tasks that aim to estimate various target structural features as outputs in an interpretable data-driven design process.

The remainder of this paper is organized as follows. Section 2 introduces the development of a global dataset of offshore jacket substructures, including the dataset structure, data source, raw data processing methods, and preliminary variable analysis. Section 3 illustrates the feature selection process for identifying correlations between features and extracting influential input features for each target feature. The feasibility of feature selection was evaluated using the RF and XGBoost models to predict the structural parameters of jackets, considering the comparison of selected input features with all input features. Section 4 concludes the paper by addressing the benefits and limitations of the novel dataset and the feature selection process. It also discusses the potential of a data-driven methodology for jacket conceptual design using the dataset with associated feature selection results.

2. Development of the dataset of offshore jackets

One of the main objectives of this study is to establish a comprehensive dataset suitable for the data-driven preliminary conceptual design of offshore jacket substructures. It should provide an in-depth understanding of the interactions between jacket substructures and their corresponding offshore wind farms globally. This dataset has the following characteristics.

1) The dataset consists of two groups of data. One group contained the structural parameters of the jackets, whereas the other group corresponded to the external variables that served as boundary conditions for the conceptual design of offshore jackets.
2) The structural variables of the jackets were selected based on the Level of Detail of 100 (LOD 100). LOD 100 only requires structural variables to form the overall topology of the structure and the initial cost model for the conceptual design (Akadiri et al., 2012).
3) The external variables consider the general influencing factors of jacket design, referred to as conceptual information on the design basis of an offshore wind turbine project.
4) The jacket samples in the dataset were diversely distributed in temporal and spatial dimensions to provide insights into the structural designs under different site conditions and to capture variations and trends in designs over time.

An initial dataset was preliminarily proposed in Qian et al. (2023) and was further improved in this study. First, we introduce the dataset structure and data sources. Subsequently, data are derived and analyzed.

2.1. Dataset structure and data source

Fig. 1 shows the structure of the dataset, and the definitions of the variables in the dataset are listed in Table 1. The starting point of the data collection process was a global search for offshore wind farms where jacket substructures were applied. The collected information for these wind farms encompasses their locations and years of completion, providing valuable context regarding each wind farm's site conditions

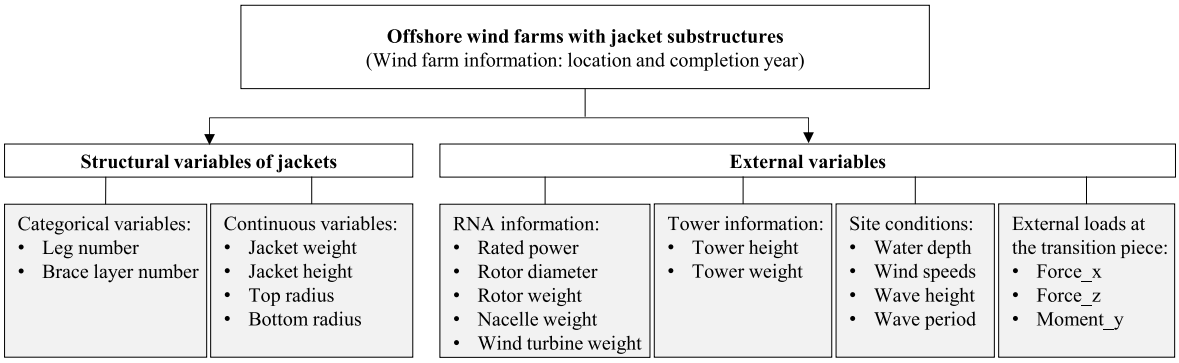


Fig. 1. Main structure of the dataset.

and construction timeline. After identifying these wind farms, the next step involved collecting data related to the structural variables of the jackets and corresponding external variables. The structural variables consisted of discrete variables, including the leg and brace layer numbers, and continuous variables, including the jacket's weight, height, and top and bottom radii. These variables form the overall topology of jackets and offer insights into their physical properties and load-bearing capacities in real load cases. External variables related to the general information of the RNA, tower, site conditions, and main influential external loads at the transition piece (TP) were collected. All these external variables can influence the conceptual design of jackets, as discussed in Section 3. The development of a dataset is an iterative refinement and improvement process that continuously enhances data collection and processing workflows to maintain the quality and integrity of the dataset.

Based on the dataset structure, various data sources were targeted to collect as much comprehensive data as possible. In the initial stage, an online database from the 4C Offshore (2023), which summarizes the general data of global offshore wind energy projects (particularly in Europe), was used to collect wind farm names with jacket substructures worldwide. In addition, more wind farms have been identified through reports on offshore wind energy development in different regions. According to the information on the selected wind farms, the raw data for most of the external variables were obtained from online sources, including the wind-turbine-models (Bauer and Matysik, 2023), Global Wind Atlas (DTU, 2023), and ECMWF Reanalysis v5 (ERA5, 2023) websites. Wind-turbine-models website serves as a vast online database providing the overall information on wind turbines, manufacturers, and models, including rated power, rotor, generator, and tower data. The Global Wind Atlas is a freely accessible web-based application to identify high wind areas suitable for wind power generation worldwide. Additionally, it facilitates preliminary calculations of wind power generation in these identified areas. ERA5 is the fifth-generation ECMWF atmospheric global climate reanalysis, covering the period from January 1940 to the present. It offers hourly estimates of numerous atmospheric, land, and oceanic climate variables and information on the uncertainties of all variables at reduced spatial and temporal resolutions.

In addition to the online sources, the external load of F_z at the TP was calculated according to the weights of the superstructures (RNA and tower), while the external loads of F_x and M_y were determined using the results of aerodynamic simulations of the IEA Wind 15-MW reference turbine (Gaertner, 2020) and the scaling rules proposed by Gasch and Twele (2012). Meanwhile, the available design documents of jacket substructures and project reports on offshore wind farms were searched to collect the jacket structural data. Alternative approaches were adopted when accurate data sources were unavailable. First, information from online public sources, such as actual photos, videos, and jacket news articles, was used to estimate the corresponding missing data. Second, appropriate imputation methods were applied to impute missing data when clear correlations were observed between the variable in question and other variables.

The percentages of data sources for each variable in the dataset are shown in Fig. 2. It is evident that the data sources of wind farms and the structural variables of jackets were collected from the 4C Offshore database, reports, design documents, or estimated through other public sources. Therefore, the reliability of these data was sufficient to conduct further investigations on data-driven structural designs. The raw data of the external variables (rotor diameter, tower height, mean wind speed at different heights, and time series of wave height) were mainly collected from the three aforementioned websites according to the known wind turbine models and locations of the wind farms. The values of external variables with the data source of "calculations" were estimated either using multiple imputation, such as missing data of the weight variables, or using scaling rules and simulation results of reference wind turbines. Notably, the extreme wave height with a return period of 50 years and the associated wave period were derived through statistical approaches based on the corresponding time-series data. The derivation methods of these two wave variables and the variables from the data source of "calculations" are illustrated in Section 2.2.

2.2. Processing of raw data

2.2.1. Missing data on the variables of the wind turbine and tower

Compared with the jacket variables, the variables of the wind turbine and tower were partially estimated, as there were missing data owing to the lack of accurate data sources, as shown in Fig. 2. Methods for handling missing data in datasets, such as listwise deletion, single imputation, and multiple imputation, have been developed and refined over the years (Pigott, 2001). Multiple imputation developed by Rubin (1987), which has a broader range of applications with fewer limitations, was developed to consider the uncertainty in the variables and address the problem of increased noise caused by the first two methods. A popular approach is multiple imputation by chained equations (MICE) and is selected to impute the missing data in this dataset. It predicts missing values using other features/variables from the dataset multiple times to create "complete" datasets. Each complete dataset is then analyzed, and the results are pooled to obtain a final result (Azur et al., 2011). The MICE algorithm is summarized as follows.

1) Fill in all missing values of the target variables in the dataset with a simple imputation, such as using the mean of the non-missing values. Initially, a complete dataset is given, but the filled-in values are just temporary placeholders.
2) Pick one variable (denoted as "var") with missing values and revert the placeholder values for this variable back to missing.
3) Use the complete variables (those without any missing values) to predict the missing values in "var": this is done by treating "var" as the output (or dependent variable) in a regression model and the complete variables as inputs (or independent variables).

Table 1
Descriptions of the variables in the dataset.

Variable name	Abbreviation	Variable Type	Unit	Annotation
Wind farm information				
Location	LC	categorical	–	Wind farm name and country
Completion year	CY	categorical	–	Actual completion year of completed jackets and estimated completion year of under-construction jackets
Structural variables of jackets				
Leg number	LN	discrete	–	Number of jacket legs
Layer number	BN	discrete	–	Number of brace layers (or bays)
Jacket weight	JW	continuous	[t]	Crucial variable in the initial cost model in conceptual design without considering the weight of the foundation below the seabed
Jacket height	JH	continuous	[m]	Vertical height of the jacket between the TP layer and seabed
Top radius	TR	continuous	[m]	Radius of the circumcircle formed by the locations of jacket legs at the TP layer
Bottom radius	BR	continuous	[m]	Radius of the circumcircle formed by the locations of jacket legs at the ground layer of the jacket
External variables				
Rated power	RP	continuous	[MW]	–
Rotor diameter	RD	continuous	[m]	Diameter of the circular area covered by the rotating blades
Rotor weight	RW	continuous	[t]	Total weight of the blades and hub
Nacelle weight	NW	continuous	[t]	–
Wind turbine weight	WW	continuous	[t]	Total weight of the rotor and nacelle
Tower height	TH	continuous	[m]	–
Tower weight	TW	continuous	[t]	–
Water depth	WD	continuous	[m]	Mean water depth in the offshore wind farm
Wind speeds	WS	continuous	[m/s]	Mean wind speed at 10/50/100/150/200 m above the mean sea level (MSL) in the offshore wind farm
Wave height	WH	continuous	[m]	Extreme wave height with a return period of 50 years in the wind farm
Wave period	WP	continuous	[s]	Wave period associated with the extreme wave height
Force_x	F_x	continuous	[kN]	Maximum aerodynamic thrust applied on the rotor and transmitted to TP
Force_z	F_z	continuous	[kN]	Compression force due to the superstructures (RNA and tower)
Moment_y	M_y	continuous	[kNm]	Maximum bending moment due to the thrust at TP

4) Replace the missing values in "var" with the values predicted by the regression model: "var" is a complete variable now, including both the observed and predicted values.
5) Repeat Steps 2–4 for each target variable. The loop through all the variables constitutes one iteration, and at the end of this iteration, all missing values are replaced with predictions that consider the relationships among variables.
6) Repeat the entire iteration (Steps 2–4 for all variables) several times until the imputations converge and the missing data values are refined with new predictions.

In this dataset, the complete and missing data for all instances are visualized in Fig. 3, in which each column represents a variable and each row corresponds to an individual instance in the dataset. Blue squares represent complete data, while white squares indicate missing values. According to the top x-axis displaying the number of complete data entries for each variable, the imputation of missing data was limited to parts of the RNA and tower information. Because the rated power and rotor diameter were complete, they were applied as the initial independent variables. For the imputation of the rotor and nacelle weight, it can be assumed that both variables correlate highly with the rated power and rotor diameter. These correlations can be captured by the regression models in the MICE algorithm, as they are essential parameters in a specific wind turbine model. Subsequently, the wind turbine weight was calculated as the sum of rotor and nacelle weights.

Furthermore, considering the physical design, the tower height is also subject to constraints imposed by the rated power and rotor diameter, that is, it must be larger than the blade length (half of the rotor diameter) and limited by the cost model in terms of the rated power of the wind turbine, whereas the tower height influences the tower weight. Therefore, using the rated power and rotor diameter to estimate the tower height and weight is reasonable. The MICE imputation procedure was conducted in Python via "IterativeImputer" in "sklearn" package (Scikit-learn, 2023), and the density distributions of the collected and imputed data of target variables are presented in Fig. 4. It can be seen that the distributions of variables with missing data remain essentially constant before and after the imputations. Therefore, using MICE to handle the missing data yielded reliable outcomes for the variables related to the wind turbine and tower.

2.2.2. *External loads at the transition piece (the top of jackets)*

Generally, the design of offshore substructures should consider permanent, variable, environmental, accidental, and deformation loads (DNV GL, 2014). Because this study only focused on the preliminary conceptual design of jackets, the following assumptions were made.

1) These loads are limited to the permanent loads of structural members and environmental loads owing to wind and waves.
2) The lack of detailed technical information on real wind turbines in the collected wind farms did not allow for simulations for each wind turbine. Instead, a reference wind turbine model, the IEA Wind 15-MW reference turbine (Gaertner, 2020), was analyzed in essential design load cases (DLCs) using the open-source simulation tool OpenFAST (Jonkman and Sprague, 2017).
3) Only extreme loads in the ultimate limit state (ULS) were derived for the reference wind turbine and converted to actual wind turbines with different rotor diameters using the scaling rules proposed in Gasch and Twele (2012).

In this case, three essential load variables, F_x, F_z, and M_y, were collected into the dataset to consider the influence of the interface loads at the TP owing to the superstructures on the design of the jackets. The load definitions and locations are listed in Table 1 and shown in Fig. 5. Force F_z is the compression force due to the self-weight of the superstructures. Force F_x denotes the extreme shear force owing to the thrust applied to the rotors and transmitted to the interface. The bending moment M_y corresponds to the extreme overturning moment owing to thrust.

The DLCs for the simulation in the ULS should be determined to

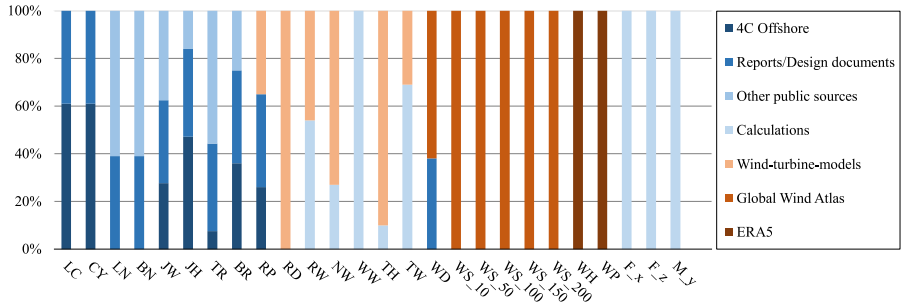


Fig. 2. Distribution of the data sources for the variables in the dataset.

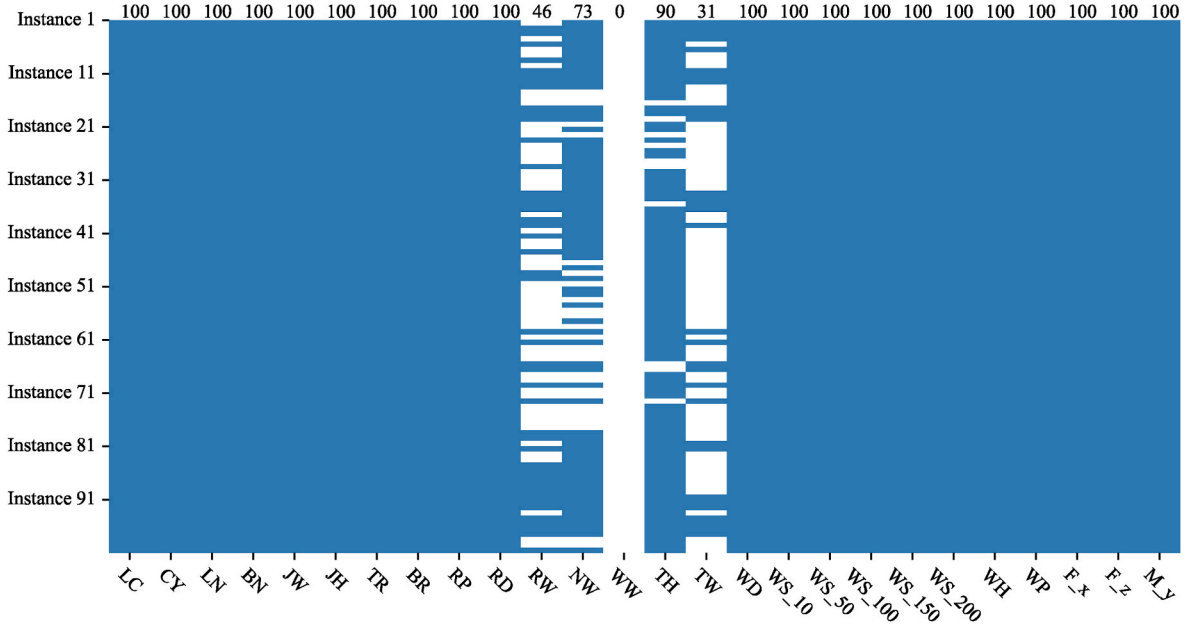


Fig. 3. Distribution of complete and missing data in the raw dataset.

derive the extreme loads at the TP for the reference wind turbine. According to IEC-61400-3 (IEC, 2019), three extreme load cases, DLC 1.5, 1.6, and 6.1, were identified as potentially critical events and were therefore selected for the simulation in this study. DLC 1.5 specifies transient shear cases in the life of an offshore wind turbine, including horizontal positive and negative shear and vertical shear with turned-on and turned-off controllers. DLC 1.6 encompasses the specifications for the ultimate loading arising from normal turbulence (NTM) conditions, in which six wind speeds of 5, 7.5, 10, 13, 18, and 21 m/s at the hub height were considered. Both load cases correspond to the design situations of power production. In the design situation of standing or idling, DLC 6.1 was analyzed to determine ultimate loads in the extreme wind speed model for the wind turbine with a mean yaw misalignment of $\pm 8^{\circ}$. According to the time series simulation results over 10 min under the selected load cases, the extreme loads F_x and M_y at the interface between the tower base and TP for the reference wind turbine are summarized in Table 2. Therefore, the maximum values of F_x and M_y occur in DLC 1.6 with a wind speed of 10 m/s, which are 2744 kN and 304,400 kNm, respectively. Notably, extreme loads in the same DLC can be captured at different time steps.

According to Gasch and Twele (2012), the effects of a change in rotor diameter on the forces at the blade can be determined through the rules of similarity. The results of this study show that the aerodynamic forces increase with the square of the rotor diameter. Therefore, the shear force F_x at the TP of the jackets for wind turbines in the dataset can be derived as

$$F\_x = T = T_0 \bullet \left(RD^2/RD_0^2\right) = F\_x_0 \bullet \left(RD^2/RD_0^2\right) \tag{1}$$

with $F\_x_0 = 2744$ kN and $RD_0 = 240$ m, where T and T_0 are the sum of the thrust applied to the rotor of the target and reference wind turbine, respectively, F_x_0 is the maximum shear force for the reference wind turbine, RD and RD_0 are the rotor diameters of the target and reference wind turbine, respectively.

Subsequently, the overturning moment M_y at the TP of the jackets for the wind turbines in the dataset is derived as

$$\begin{aligned} M\_y &= T \bullet H_{hub} = \left(M\_y_0/H_{hub,0}\right) \bullet \left(RD^2/RD_0^2\right) \bullet H_{hub} \\ &= M\_y_0 \bullet \left(RD^2/RD_0^2\right) \bullet \left(H_{hub}/H_{hub,0}\right) \end{aligned} \tag{2}$$

with $M\_y_0 = 304400$ kNm and $H_{hub,0} = 150$ m, where M_y_0 is the maximum overturning moment for the reference wind turbine, H_{hub} and $H_{hub,0}$ are the hub heights of the target and reference wind turbines.

The compression load F_z should be the sum of the self-weights of the superstructures and can be calculated as

$$F\_z = RW + NW + TW \tag{3}$$

where RW, NW, and TW are the self-weights of the rotor, nacelle, and tower, respectively.

By incorporating the external loads arising from the superstructures, the dataset provides valuable insights into the comprehensive

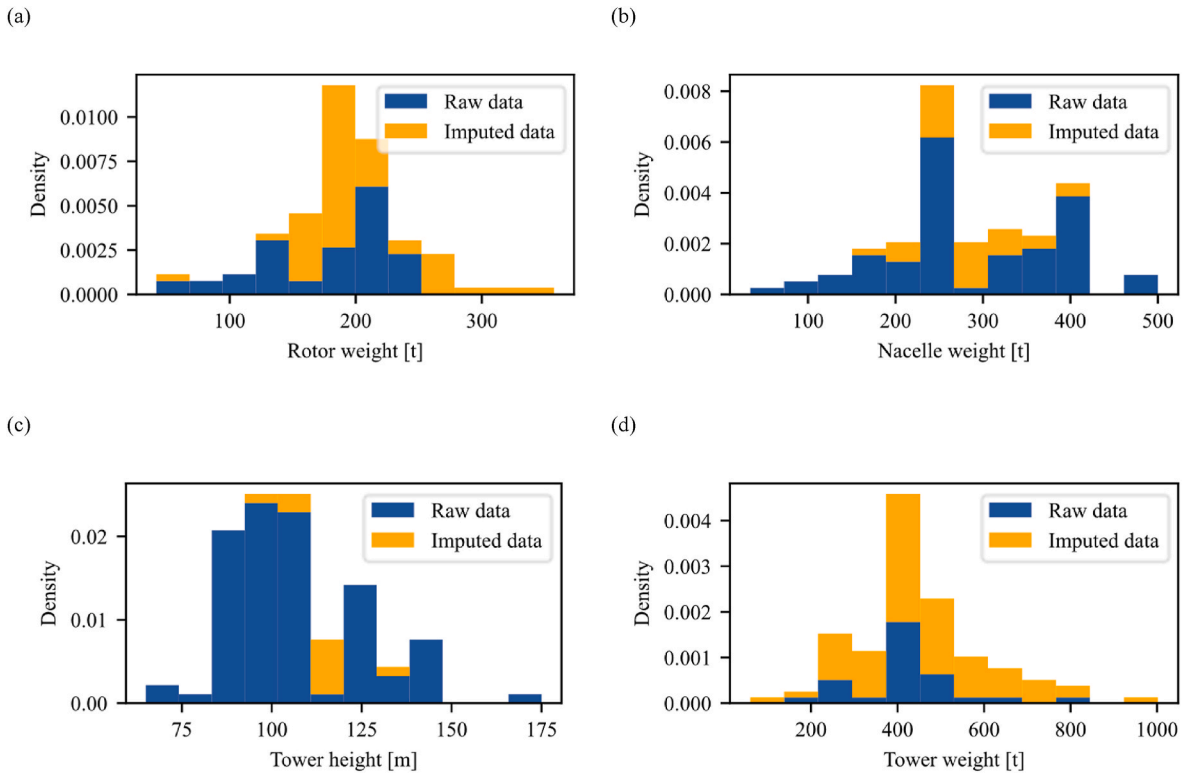


Fig. 4. Density distributions of the collected and imputed data for the (a) rotor weight, (b) nacelle weight, (c) tower height, and (d) tower weight.

interactions of the RNA and tower with jacket substructures. This inclusion serves as a foundation for providing essential design preconditions, helping to simulate the intricate design thinking of structural engineers during the conceptual design phase of jackets. This part of the dataset provides an attempt to translate the understanding of the complex mechanics of the system into practical design considerations for the further investigation of data-driven methods.

2.2.3. Extreme wave height and associated wave period

In this study, a long-term data-based wave characteristic analysis of selected offshore wind farms was conducted following the systematic flowchart shown in Fig. 6. Initially, raw wave data were collected from the ERA5 platform, which provides monthly mean averaged data of the global climate and weather from 1940 onwards. The chosen variable is the "significant height of combined wind waves and swell," with data covering 50 years from 1973 to 2022. The raw data were initially stored in GRIB-format files from the platform and visualized using Panoply software (Schmunk, 2023), allowing for the visualization of monthly wave data at a spatial resolution of 0.5° on longitude and latitude globally. The wave height dataset was exported in the CSV format from Panoply to facilitate the extraction of wave data specific to the desired location of each wind farm. Subsequently, a Python script was implemented to extract the required data at user-defined latitude and longitude coordinates. For each wind farm, 600 significant wave height data points were extracted, representing 50 years of monthly data.

Following the data extraction process, a statistical analysis of the time-series data was performed. In this phase, various statistical distributions were employed to identify the appropriate long-term patterns for wave height and calculate the extreme wave height for each wind farm in the dataset. Analyses of environmental data in offshore wind farms by Hübler et al. (2017) indicate that environmental data can be adapted to fit several statistical models. In this study, Normal, Weibull, Gamma, and Gumbel distributions were used to fit the wave height data, enabling the derivation of probability density functions (PDFs) and cumulative distribution functions (CDFs). The PDFs of the statistical distributions are listed in Table 3.

For each offshore wind farm, the parameters of four statistical distributions were derived with the raw wave height data in Python via the "scipy.stats" package. The goodness of fit of the distributions was evaluated using the Kolmogorov–Smirnov (KS) test, in which P-values were calculated. Subsequently, the optimal distribution of the wave height time-series data in each wind farm was selected to further determine the significant wave height with a certain return period. The significant wave height was derived using the first-order reliability method (FORM) illustrated by Liu and Burcharth (1997). According to IEC 61400–3 (IEC, 2019), the extreme wave height with a return period of 50 years and the associated range of wave period were estimated to assess extreme wave conditions. To obtain the extreme wave height *WH* and the associated range of the wave period *WP*, the following relationships were applied:

$$WH = 1.86\, H_{s,50} \tag{4}$$

$$11.1\sqrt{H_{s,50}/g} \le WP \le 14.3\sqrt{H_{s,50}/g} \tag{5}$$

where $H_{s,50}$ is the significant wave height with a return period of 50 years. In this study, the lower limit of the wave period range was assumed to yield the most severe loading conditions.

2.3. Preliminary variable analysis

2.3.1. Wind farm information

The distribution of the collected offshore wind farms with jackets is shown in Figs. 7 and 8, which describe the locations and completion years of the wind farms. More than half of the collected jackets are distributed in Asia, with the remainder primarily situated across Europe and the United States. The time frame for the completion years of these wind farms spans from 2006 to 2028. Before 2017, relatively few instances of offshore wind farms were constructed using jacket substructures. This primarily occurs because monopiles are optimal for offshore wind farms in shallow water areas. However, the landscape

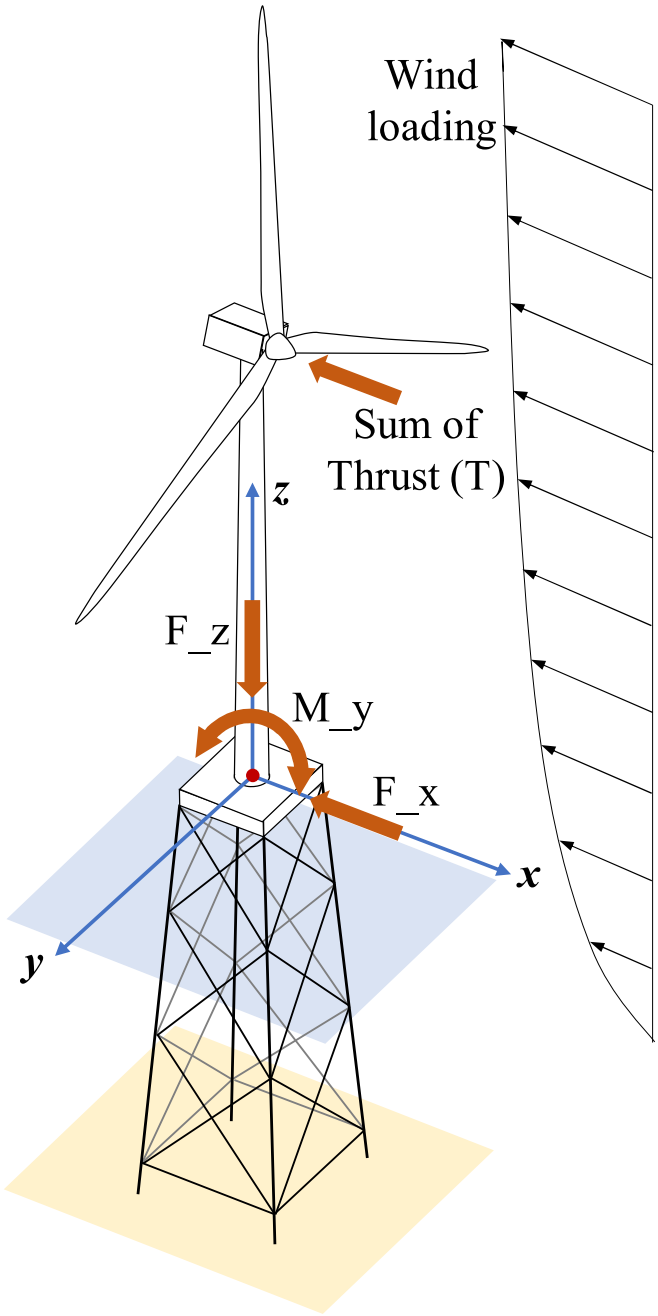


Fig. 5. The sketch of the offshore wind turbine with jacket substructure and the external loads in the tower-base coordinate system.

began to change around 2019 when jacket substructures started gaining traction in wind farms in deeper water areas. According to the GWEC (2022), 2021 is the best year ever for the global offshore wind industry, which is also reflected in the construction of jacket substructures. Additionally, owing to the lack of data after 2022, approximately 20 wind farms employing jacket substructures completed in 2023 and the upcoming years were included in this dataset, as shown by the light blue bars in Fig. 8.

Various types of jackets can be erected within the same wind farms. Therefore, the dataset encompasses 100 instances of offshore jacket substructures distributed across 90 offshore wind farms. Consequently, the instances of jackets within the dataset exhibited rich diversity in both temporal and spatial dimensions. This broad distribution aligns with the foundational requirement of the dataset, which seeks to provide comprehensive insights into structural designs across a spectrum of site conditions while also encapsulating the dynamic variations and evolution of designs over time.

Table 2
Extreme loads F_x and M_y at the interface between the tower base and TP for the reference wind turbine due to selected DLCs.

DLC	Annotation	F_x [kN]	M_y [kNm]
1.5	horizontal positive shear	2522	282,800
1.5	horizontal negative shear	2522	287,800
1.5	vertical shear	2522	304,300
1.5	vertical shear with turned-off controller	2500	303,300
1.6	wind speed: 5 m/s	1591	136,100
1.6	wind speed: 7.5 m/s	2396	226,300
1.6	wind speed: 10 m/s	2744	304,400
1.6	wind speed: 13 m/s	2305	242,600
1.6	wind speed: 18 m/s	2134	190,400
1.6	wind speed: 21 m/s	2193	203,600
6.1	mean yaw misalignment of +8°	2120	118,800
6.1	mean yaw misalignment of - 8°	1832	96,960

2.3.2. *Structural variables of jackets*

As discussed in Section 2.1, in the conceptual design of jacket substructures, structural variables can be categorized into two primary types: continuous and discrete. For continuous variables, it is assumed that there are an infinite number of values within a given range. In contrast, discrete variables have distinct values or categories. Specific visualization charts were employed to provide a comprehensive understanding of the data distribution associated with these two distinct types of variables. For continuous variables, error bar plots were used to clearly represent the average values (mean) and the range (standard deviation) within which most data points were located. For discrete variables, the data were presented using standard bar plots showing the frequency of each distinct value, thus providing insights into the prevalence and distribution of these variables within the dataset. Additionally, the means and standard deviations of the continuous variables spanned significantly diverse ranges, as shown in Table 4. The data were standardized using the min-max scaling method to facilitate a meaningful comparison of their distributions, which were calculated as

$$x' = (x - \min(X))/(\max(X) - \min(X)) \tag{6}$$

where x' is the standardized value, x is the original value, $\min(X)$ and $\max(X)$ are the minimum and maximum values of this variable, respectively. A comparable representation of these distributions is presented in Tables 4 and is graphically depicted in Fig. 9(a). Some noteworthy findings are noted based on the distributions of the structural variables.

Jacket height is essential in the initial conceptual design, as it is usually the first variable determined based on site conditions. According to DNV–OS–J101 (DNV GL, 2014), jacket substructures are well-suited for sites with water depths ranging from 20 to 50 m. The total jacket height is computed as the sum of the water depth and air gap. The air gap is the vertical distance between the mean sea level and the lowest point of the turbine tower, ensuring that the wave action does not directly impact the turbine structure. The air gap size is determined based on the wave analysis for a specific site and could exceed 20 m, depending on the predicted extreme wave heights for the site. Table 4 reveals that the jacket height in the dataset, within a confidence interval of 68.2% (μ-σ to μ+σ), ranges from 40 m to 70 m. This observed range closely aligns with the optimal depth range set by the standard when considering the additional height added by the air gap. This correspondence suggests that the data investigated in the dataset are consistent with recognized industry practices, thus proving the reliability of the dataset.

Concerning the standardized values of continuous variables, the standardized means for all variables hover around the range from 0.44 to 0.51, and their standard deviations were similar, ranging from 0.17 to

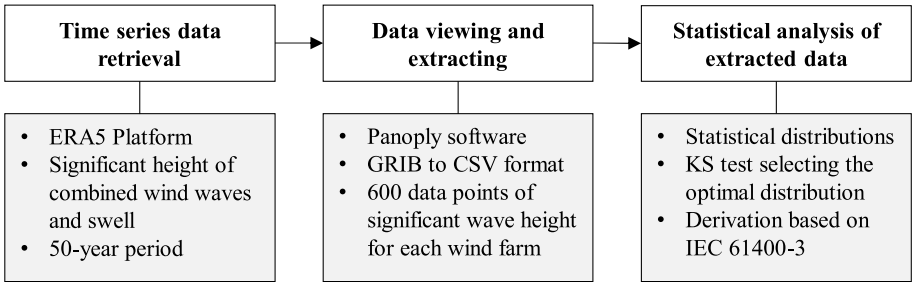


Fig. 6. Flowchart for a long-term data-based wave characteristics analysis.

Table 3
Probability density functions and corresponding parameters of the statistical distributions.

Distribution	PDF	Location parameter	Shape parameter	Scale parameter
Normal	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ	–	σ
Weibull	$f(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}, x \geq 0 \\ 0, x < 0 \end{cases}$	–	k	λ
Gamma	$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$	–	k	θ
Gumbel	$f(x) = \frac{1}{\beta} e^{-(z+e^{-z})}, z = \frac{x-\mu}{\beta}$	μ	–	β

0.23. This close grouping of standardized values suggests consistency and relative similarity in the spread and central tendencies of the jacket structural variables in the dataset. This uniformity is crucial because it ensures that no particular feature dominates the learning process in ML applications owing to its scale or variability. Consequently, ML algorithms can converge faster, produce more generalized models, and offer enhanced interpretability of feature correlations, as discussed in subsequent sections.

Regarding the discrete variables of the brace layer and leg number, the dataset reveals that the brace layer numbers are typically two, three, or four, whereas the leg number should be either three or four, as shown in Fig. 9(b). According to the design codes and standards, the brace layer number is usually influenced by the jacket height owing to the specific design constraints for brace members in jackets. For instance, NORSOK (2004) requires the angle between the brace and the leg to exceed 30°. Four-legged jackets were popular in previous designs owing to their structural symmetry and stability. However, recent studies have suggested that three-legged jackets may be more advantageous under certain boundaries, making it possible to optimize mass-dependent cost models when considering the leg numbers (Häfele et al., 2019). Consequently, there is a growing trend towards designing three-legged jackets for emerging wind farms, although four-legged jackets remain more prevalent in the dataset.

2.3.3. External variables

The dataset contains 18 external variables, the descriptions of which are presented in Table 1. To visualize the distribution of the variables, the histograms and associated density curves generated using the kernel density estimate (KDE) method are shown in Fig. 10. Similar to the structural variables of jackets, different variables exhibited qualitative variations across diverse scales, and the ranges of the variables also showed apparent differences. The external variables were standardized using the min–max scaling method to ensure a consistent comparison. The mean and standard deviation of the standardized data are presented in Fig. 11.

By combining Figs. 10 and 11, specific variable pairs exhibit considerable correlations, as evidenced by the close alignment of their density curves and distributions. Notable correlated pairs include tower height with tower weight, wind speeds at different heights, wave height with wave period, and force_x with moment_y. Some relationships can be explained as inherently physical or empirical among these correlated pairs. For instance, the similarity of tower height and weight makes intuitive sense, as a higher tower would generally require more materials and, hence, would be heavier. The wind speeds at varying heights are correlated, indicating that wind patterns at one height may be predictive of those at other heights, considering the nature of the aerodynamics and interactions among wind currents. Furthermore, some

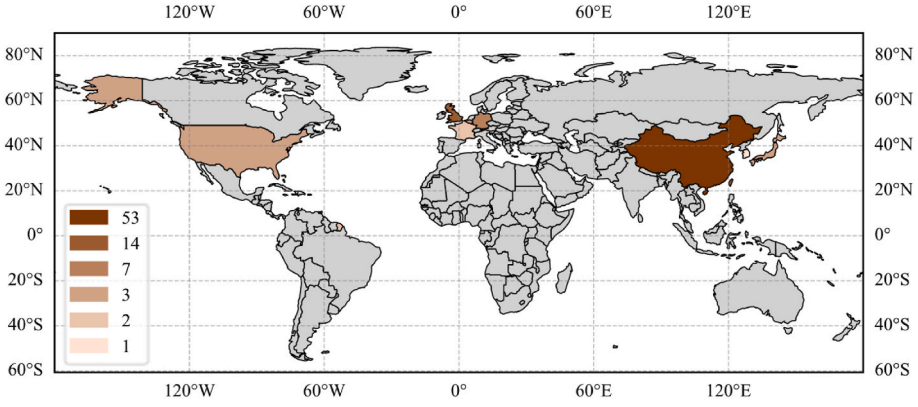


Fig. 7. Spatial distribution (location) of the offshore wind farms with jackets in the dataset distributed worldwide.

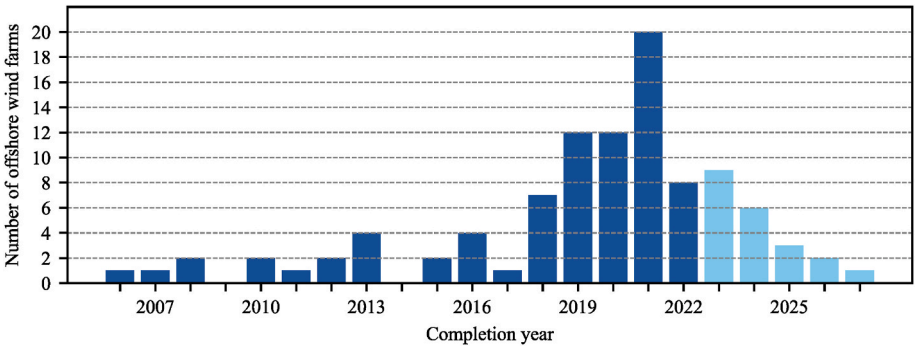


Fig. 8. Temporal distribution (completion year) of the offshore wind farms with jackets in the dataset between 2006 and 2027.

Table 4
Original and standardized mean and deviation values of continuous variables of jackets in the dataset.

Continuous variables	Original mean μ	Original deviation σ	Standardized mean	Standardized deviation
Jacket weight	1101.67 t	361.83 t	0.44	0.23
Jacket height	55.53 m	15.96 m	0.49	0.21
Bottom radius	18.77 m	3.90 m	0.51	0.20
Top radius	9.30 m	2.50 m	0.45	0.17

other correlations are mathematically derived. For instance, the wave period values are not independent measurements but are calculated from extreme wave heights. This derivative relationship is outlined in Equation (5), which employs well-established principles from standards.

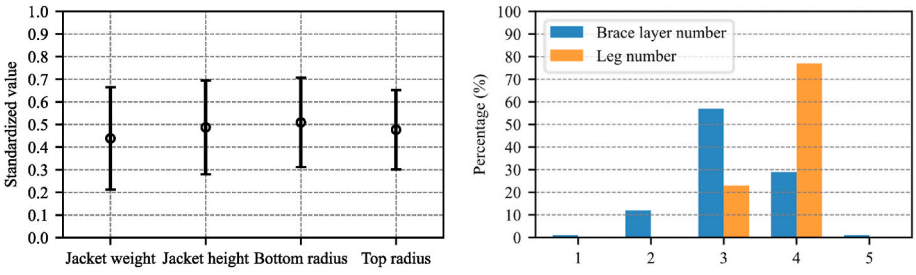
Understanding these correlations offers insights into the interaction of variables that can assist in decision-making processes while designing offshore substructures and is also crucial when preparing data for ML applications. In ML, correlated input features can sometimes introduce redundancy, which can affect the prediction performance of the trained ML models. This issue should be appropriately addressed through further feature analyses, allowing for the establishment of more robust and efficient ML models. This is one of the most important motivations for feature selection, as discussed in Section 3.

3. Feature selection

To perform the data-driven conceptual design of jackets, each target feature of the structural variables of jackets must find appropriate subsets of features as inputs to train ML models and perform predictions. Feature selection plays a pivotal role in enhancing the understandability of the data-driven method for jacket design. On the one hand, it simplifies the ML models and improves the interpretability of the method by highlighting the most influential and relevant features. On the other hand, it reduces overfitting owing to the small size of the dataset and relatively more features, ensuring more reliable and generalizable results (Haury et al., 2011). More specifically, by eliminating redundant input features in the dataset, feature selection can facilitate more precise data visualization and improve predictive performance. Meanwhile, by extracting appropriate subsets of input features, it aids in deriving meaningful and domain-specific insights that give designers a clearer picture of which features should be considered in the ML field and assists them in determining whether other necessary influencing factors should also be considered in the engineering domain for subsequent data-driven conceptual design. A flowchart of the feature selection is shown in Fig. 12, which consists of four main steps.

1) Remove the redundant external features.
2) Define the prediction order of the target features of the jackets and all candidate input features for each target feature.
3) Determine the subsets of input features for each target feature.
4) Check the feasibility of the selected subsets of input features using ML models.

The redundancy of the external features in the dataset in the first step was identified by applying Spearman's correlation to measure the linear or nonlinear correlations between two external features. High correlation values indicate redundancy. According to the correlation coefficient values visualized in the heat map, one of the external features in the feature pairs with large correlation values should be removed. Subsequently, the prediction order of the target features and all candidate input features for each target feature were defined based on the structural design requirements of the industrial projects and standards. The candidate input features consist of external features without redundancy and specific structural features. Some structural features of the jackets can be regarded as input features for predicting other structural features. Subsequently, Spearman's correlation is applied again in the third step to identify the input features that are highly correlated with each target feature and to determine the corresponding subsets of input features. Finally, the RF and XGBoost models were trained using the data of selected subsets of input features (experimental groups) and all external features (control groups), respectively. The feasibility of feature



(a) Mean and standard deviation of continuous variables (b) Percentage distribution of discrete variables

Fig. 9. Distribution of structural variables of jackets in the dataset.

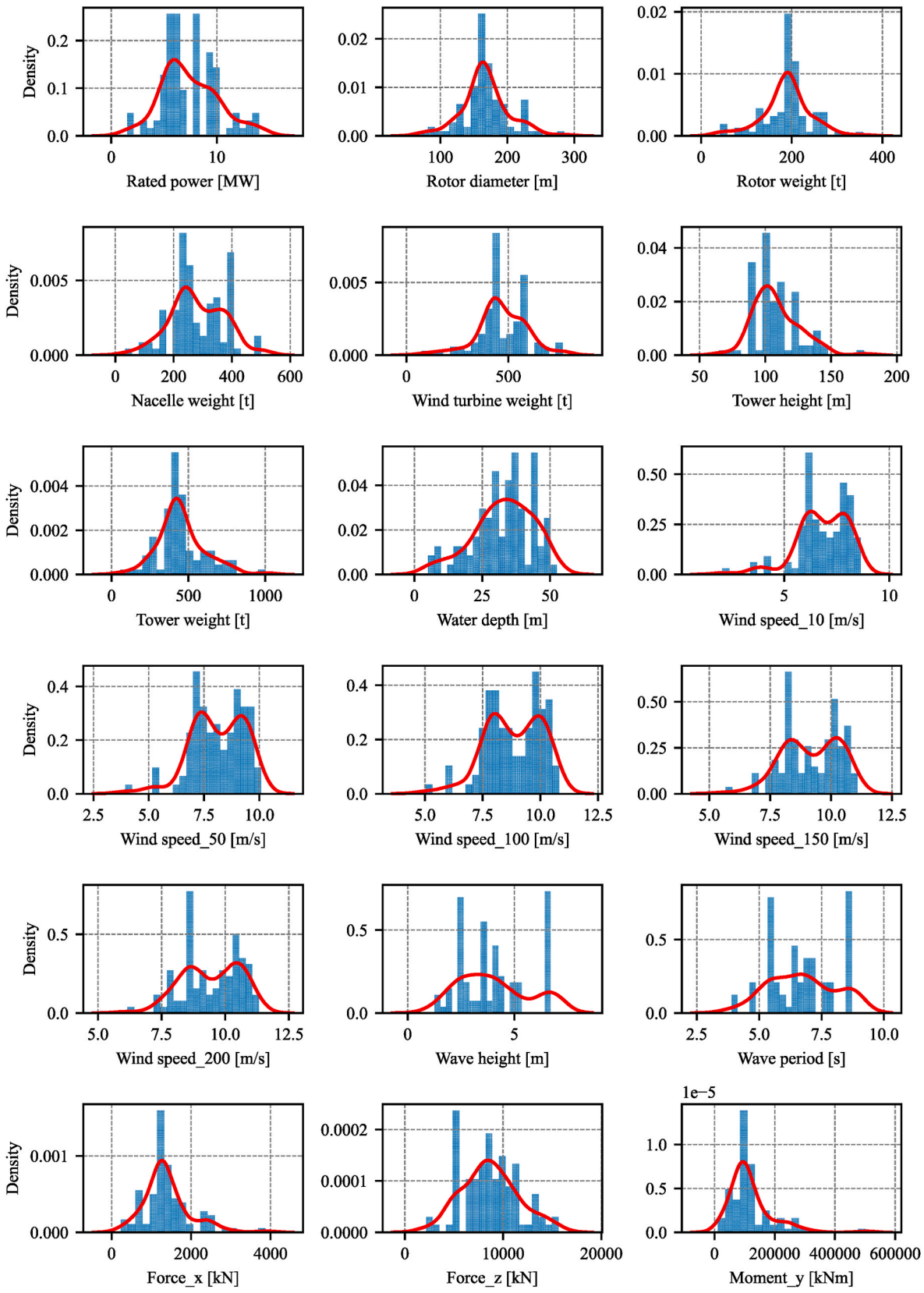


Fig. 10. Histograms of the external variables.

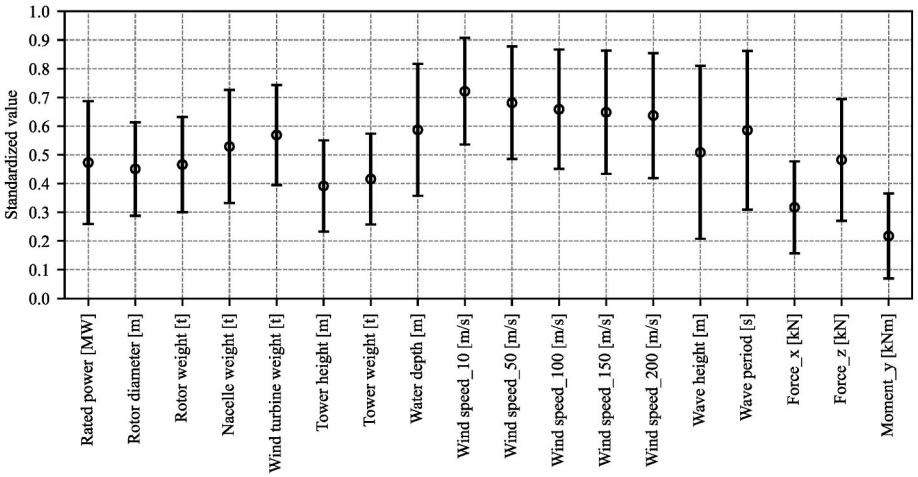


Fig. 11. Mean and standard deviation of standardized external variables.

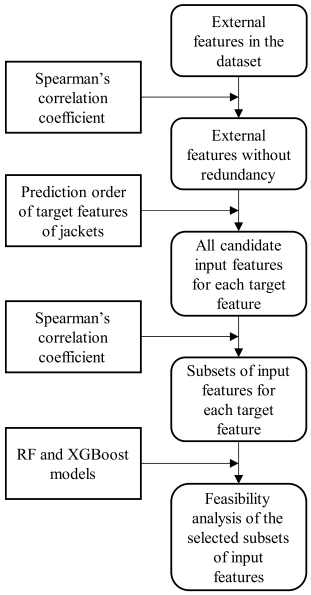


Fig. 12. Flowchart of feature selection.

selection can be evaluated by comparing the predictive performances of the ML models trained using two different sets of input features. It is worth noting that this feasibility analysis can also be regarded as a preliminary case study for the data-driven conceptual design of jackets, which is valuable for future ML-based jacket design.

3.1. Methodology

3.1.1. Spearman's correlation coefficient

Spearman's correlation coefficient (SCC), denoted as r_s, is a nonparametric measure of rank correlation. It assesses the extent to which the relationship between two variables can be described using a monotonic function. Unlike Pearson's correlation, which requires linear relationships and interval data, Spearman's correlation works with ordinal data and is robust to nonlinear relationships (Sprent and Smeeton, 2016). Furthermore, SCC is appropriate for continuous and discrete ordinal variables (Lehman et al., 2013), making it ideal for analyzing the correlations between features in the dataset with both data types in this study.

For a sample of two random variables (X, Y) of size n, the n raw values X_i and Y_i are converted to ranks R (X_i) and R (Y_i), and r_s, the value of SCC, is computed as follows:

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \mathrm{cov}(\mathrm{R}(X), \mathrm{R}(Y)) \,/\, \left(\sigma_{\mathrm{R}(X)} \bullet \sigma_{\mathrm{R}(Y)}\right) \tag{7}$$

where $\rho_{\mathrm{R}(X),\mathrm{R}(Y)}$ denotes the usual Pearson correlation coefficient applied to the rank variables (R(X), R(Y)), $\sigma_{\mathrm{R}(X)}$ and $\sigma_{\mathrm{R}(Y)}$ are the standard deviations, and cov (R(X), R(Y)) is the covariance of the rank variables, which is computed by:

$$\mathrm{cov}(\mathrm{R}(X), \mathrm{R}(Y)) = \mathbb{E}\left[\left(R(X) - \mu_{\mathrm{R}(X)}\right)\left(R(Y) - \mu_{R(Y)}\right)\right] \tag{8}$$

where $\mu_{\mathrm{R}(X)}$ and $\mu_{\mathrm{R}(Y)}$ are the mean values of the rank variables (R(X), R (Y)), and $\mathbb{E}$ is the expectation. If there are no repeated data values, $r_s = \pm 1$ occurs when each variable is an entirely monotonic function of the other, even if their relationship is nonlinear. The SCC sign indicates the direction of association between X (independent variable) and Y (dependent variable). If Y tends to increase when X increases, r_s is positive. Conversely, if Y decreases as X increases, r_s becomes negative. $r_s = 0$ indicates that Y has no tendency to either increase or decrease as X increases. This study computed the SCC matrices in Python via the "pandas" package.

3.1.2. RF and XGBoost models

Both RF and XGBoost models are used for supervised learning problems, where a target feature can be predicted using multiple input features. They are ideal for feasibility evaluation in the feature selection process because of their ability to handle high-dimensional data and their robustness against overfitting. Therefore, they are suitable for analyzing complex datasets with multiple features, such as the one in this study, and ensuring more reliable performance evaluations. These models are efficient and accurate for various predictive tasks and provide valuable insights into feature importance, aiding in the evaluation of the selected feature subsets in the previous steps. Their versatility in handling both regression and classification problems provides flexibility in the types of predictions for different target structural features in jacket designs, ensuring a comprehensive assessment of the feasibility of the selected features.

RF is an ensemble ML algorithm that operates by constructing several decision trees with bagging (also known as bootstrap aggregating) in the training phase and outputting class prediction by majority voting in classification problems or mean prediction in the regression problems of individual trees (Breiman, 2001). A fundamental concept is that a group of weak learners (decision trees) combine to form a strong learner. The randomness injected in the model-building process helps in improving the accuracy of the model and reducing overfitting (Ho, 2002). The RF procedure is as follows.

1) Initialize the model and randomly sample: All hyperparameters should be defined when initializing the model. Multiple training sets are created through bagging (sampling with replacement) the processed dataset after feature selection.
2) Construct and train decision trees: Each training set is used to build a decision tree. When splitting a node during the construction of the tree, only a random subset of the features is considered through bagging.
3) Predict target features: Each tree in the forest outputs a class prediction for classification. The class with the majority vote becomes the model's prediction, and the model's output is the mean of the predictions from all trees.

The bagging technique is key to the RF algorithm and is utilized twice in the RF procedure: random sampling with replacement for both training sets and features. This implies that each tree knows only the data associated with a small constant number of features and a variable number of samples less than or equal to that of the original dataset. Consequently, decision trees are more likely to return a broader range of prediction answers learned from more diverse knowledge, resulting in a more robust predictive performance and less overfitting in RF.

XGBoost is a highly efficient and scalable implementation of the gradient boosting algorithm that works by sequentially adding classification and regression trees (CART), each of which corrects its predecessor's errors. According to the XGBoost Documentation (XGBoost, 2022), a CART slightly differs from decision trees in RF, in which the leaf only contains decision values. In CART, a real score is associated with each leaf, providing richer interpretations beyond the classification. Unlike bagging in RF, overfitting issues in XGBoost can be prevented by adding regularizations controlling the complexity of CART into the objective functions as follows:

$$\mathrm{obj}(\theta) = \sum_{i=1}^{n} l(y_i, \widehat{y}_i) + \sum_{k=1}^{K} \omega(f_k) \tag{9}$$

where θ denotes the parameters to be optimized in the model, $l(y_i, \widehat{y}_i)$ is a differentiable convex loss function that measures the difference between the target y_i and the corresponding prediction $\widehat{y}_i$ of the i-th instance in the training set, and $\omega(f_k)$ is the regularization of the function of the k-th CART. The general procedure for applying XGBoost involves the following key steps:

1) Initialize the model: All hyperparameters, including regularization parameters, should be defined when initializing the model. The regularization parameters ensure consistency in how the algorithm penalizes the model complexity throughout the training process. It begins with an initial prediction, such as the mean of the target for regression or the log odds for classification.
2) Build CART sequentially: In each iteration, add a new CART, where each tree targets the residuals and errors of the previous trees. At the t-th iteration, the substeps are as follows:
   a) Calculate the Gradient and Hessian, and apply regularization: For each instance, compute the Gradient g_i and Hessian h_i of the loss function with respect to the current prediction and include regularization in this computation to balance the model complexity:

$$\mathrm{obj}^{(t)} = \sum_{i=1}^{n} \left[l\left(y_i, \widehat{y}_i^{(t-1)} + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \omega(f_t) \right. \tag{10}$$

   b) Create a new tree: A new tree is built based on the calculated Gradients and Hessians to minimize the loss function.
   c) Update the model: After the tree is added, the prediction of the model is updated by adding the output of the new tree, and scaled by the learning rate.
3) Repeat the process: The iteration continues until a stopping criterion is met, such as the maximum number of trees or no further improvement in the loss function.
4) Output the final trained model and predict the target features: The final prediction is the aggregate of the predictions from all the individual trees in the trained model, that is, the majority voting for classification and mean prediction for regression.

For more details on these two ML models, the interested reader is referred to Breiman (2001) and Chen and Guestrin (2016), which involve comprehensive research and interpretation of two models. The RF and XGBoost models in Python were implemented in this study by using the "RandomForestRegressor" and "RandomForestClassifier" from the "sklearn" package (Scikit-learn, 2023) for RF and the "xgboost" package (Cho and Yuan, 2021) for XGBoost.

3.2. Correlation analysis and feature selection results

The SCC matrix for all external features is presented in the heat map in Fig. 13. Because the definition of redundant features in a dataset varies with specific problems and objectives, a definitive threshold for redundancy based on the SCC has not been universally established. This study defined a conservative threshold of 0.9 to identify redundancy among features. This value was chosen to balance minimizing redundancy and preserving vital information in the dataset, ensuring that essential features were retained while reducing the risk of including highly correlated external features. The redundant feature pairs are listed in Table 5.

The SCC values of all wind speed pairs exceeded 0.9, which is consistent with the trends observed in Figs. 10 and 11. Because the wind effects at greater heights that indirectly influence jacket design are already considered in the external loads Force_x and Moment_y, the wind speeds above 10 m can be considered redundant. In contrast, a wind speed of 10 m was crucial for measuring the wind effects acting directly on the jacket. In addition to wind speeds, ten more feature pairs exhibited redundancy elements. To identify redundant features within the ten pairs, those that recur most frequently were likely candidates. In this context, Force_x and Moment_y, appearing in six of the ten pairs, are deemed redundant. Theoretically, both features were derived by following the scaling rules outlined in Equations (1) and (2), which make it expected and reasonable that they are highly correlated with the rotor diameter and tower height.

Similarly, the tower weight, which appears twice in ten pairs, is identified as a redundant feature that has an essentially monotonous relationship with the tower height owing to the constant gradient of the material demand with respect to the tower cross-section along the tower height. Concerning the pair of wind turbine and nacelle weights, the wind turbine weight consists of the combined weight of the nacelle and rotor. Given that the nacelle weight significantly surpasses the rotor weight, the wind turbine weight significantly overlaps with the nacelle weight and is thus redundant. As for the wave period and wave height, the wave period was calculated from the wave height using Equation (5). This relationship establishes a significant correlation between the two features, categorizing the wave period as redundant. Based on Spearman's correlation analysis, nine external features were removed owing to redundancy, whereas the remaining external features were retained in the subsequent steps for feature selection.

The retained external features and all structural features were considered for the final feature selection because certain structural features are necessary inputs for predicting other structural features. The prediction order of the target features was defined according to the conceptual design requirements of industrial projects and standards. The jacket height largely depends on the site water depth. Therefore, it should be the initial predictive target. Subsequently, the bottom and top radii were designed and predicted based on various influencing factors, such as loads at the TP from the wind turbine and tower, wave and

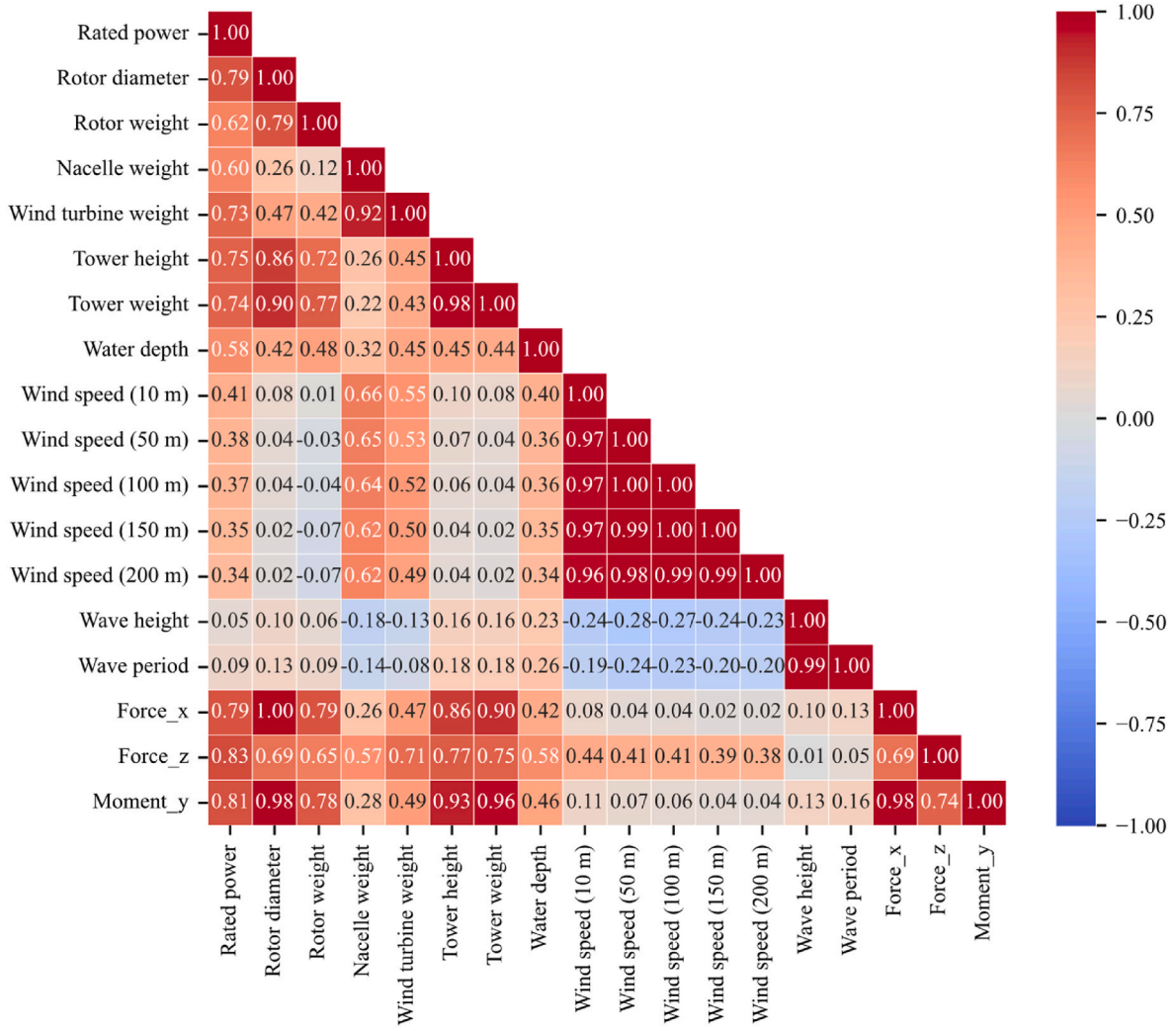


Fig. 13. Heat map of SCC values between the two external features.

Table 5
Pairs of external features with SCC values ≥ 0.90.

External feature 1	External feature 2	SCC value	Redundant feature
Tower weight	Rotor diameter	0.90	Tower weight
Force_x	Rotor diameter	1.00	Force_x
Moment_y	Rotor diameter	0.98	Moment_y
Wind turbine weight	Nacelle weight	0.92	Wind turbine weight
Tower weight	Tower height	0.98	Tower weight
Moment_y	Tower height	0.93	Moment_y
Force_x	Tower weight	0.90	Force_x
Moment_y	Tower weight	0.96	Moment_y
Wind speed (*i* m)	Wind speed (*j* m)	≥0.96	Wind speed (50/100/150/200 m)
Wave period	Wave height	0.99	Wave period
Moment_y	Force_x	0.98	Moment_y or Force_x

current loads, and mechanical and constructional properties of the overall structural framework. Once the general geometrical parameters of the jacket are determined, an approximate estimation of the jacket weight can be performed for conceptual design because the material and commonly used tube models are already standardized in the industry. Furthermore, the layer and leg number are essential for topology determination, both of which are influenced by specific structural features and should be determined in the end. As mentioned in Section 2.3.2, the brace layer number depends on the jacket height owing to the design constraints in jackets, whereas the leg number of the jacket could be correlated with the mass-dependent cost model. The final prediction order for the target features of jackets is listed in Table 6.

Correlation analysis between target features and all candidate input features was conducted using Spearman's correlation. The heat map of the SCC matrix is shown in Fig. 14, in which the first six are target features, and the rest are retained external features. According to the prediction order, the target features that are higher in order can serve as

Table 6
Prediction order and results of feature selection.

Prediction order	Target feature	Selected input feature subset	Corresponding SCC values
No. 1	Jacket height	['WD', 'WS', 'NW', 'RP', 'F_z', 'TH', 'RD', 'RW']	[0.63, 0.63, 0.56, 0.50, 0.48, 0.26, 0.23, 0.22]
No. 2	Bottom radius	['TH', 'F_z', 'RW', 'RD', 'RP', 'WD', 'JH', 'NW']	[0.57, 0.55, 0.53, 0.50, 0.48, 0.41, 0.39, 0.22]
No. 3	Top radius	['BR', 'WS', 'JH', 'NW', 'F_z', 'RP', 'RD', 'TH']	[0.45, 0.40, 0.37, 0.35, 0.30, 0.27, 0.24, 0.21]
No. 4	Jacket weight	['RW', 'F_z', 'RP', 'BR', 'TH', 'RD', 'WD', 'JH']	[0.73, 0.71, 0.69, 0.67, 0.65, 0.64, 0.62, 0.55]
No. 5	Layer number	['JH', 'WS', 'WD', 'NW', 'JW', 'RP']	[0.42, 0.41, 0.37, 0.24, 0.22, 0.12]
No. 6	Leg number	['JH', 'TR', 'WS', 'NW', 'RP', 'BR']	[-0.44, −0.35, −0.30, −0.26, −0.20, −0.16]

inputs for the later predicted ones. For example, jacket height can be applied as an input for all other target features. Based on the absolute SCC values in the matrix, the candidate features in the top eight pairs were selected for continuous target features, whereas discrete target features required the top six. The difference in the number of selected input features can be attributed to the nature of different types of predictive tasks. Regression tasks with continuous targets typically require more predictors to capture subtle variations in data. By contrast, classification tasks with discrete targets often require fewer predictors to differentiate between distinct classes effectively. The finalized input feature subsets selected for each target feature according to the selection rules are listed in Table 6.

3.3. Feasibility analysis using RF and XGBoost models

To evaluate the feasibility of the selected input feature subsets listed in Table 6, the RF and XGBoost models played a crucial role in the last step of feature selection. These two universal ML algorithms were used to assess the predictive strength and overall effectiveness of the selected input features for each target feature because they demonstrated high robustness and accuracy in complex tabular data scenarios for various prediction tasks.

The continuous target features of the jacket height, bottom radius, top radius, and jacket weight were predicted using regression models ("RandomForestRegressor" and "XGBRegressor" from Python packages). In contrast, the discrete target features of the layer number and the leg number were predicted using classification models ("RandomForestClassifier" and "XGBClassifier"). Each target feature has two sets of input features, one of which is the selected subset in Table 6 and the other includes all the external features.

Prior to model training, data preprocessing was performed. First, the raw data in the dataset were standardized using the min–max scaling method outlined in Equation (6). Outliers were detected and eliminated using box plots, with the threshold set to 1.5 times the interquartile range (1.5 IQR). After processing raw data in Section 2.2, all instances in the dataset were complete. The dataset was then randomly split into training set (80 % of all 100 instances) and testing set (20 % of all 100 instances) to prevent overfitting of the models. The hyperparameter max. Number of trees (n_estimator in the models) was tuned within a range of [2, 20], whereas the other hyperparameters were set as default.

For each target feature, the RF and XGBoost models with varying n_estimator values were trained five times to enhance the robustness of the results. The predictive performances of the trained models were tested using the testing sets. Regression models were assessed using the coefficient of determination (R^2), and classification models were evaluated based on the F_1 score. Both metrics are used to measure the accuracy of the prediction and range from zero to one. Higher values of both metrics indicate better predictive performance of the model. The mean μ and standard deviation σ of the metrics for all RF and XGBoost models with the same n_estimator trained over five times are

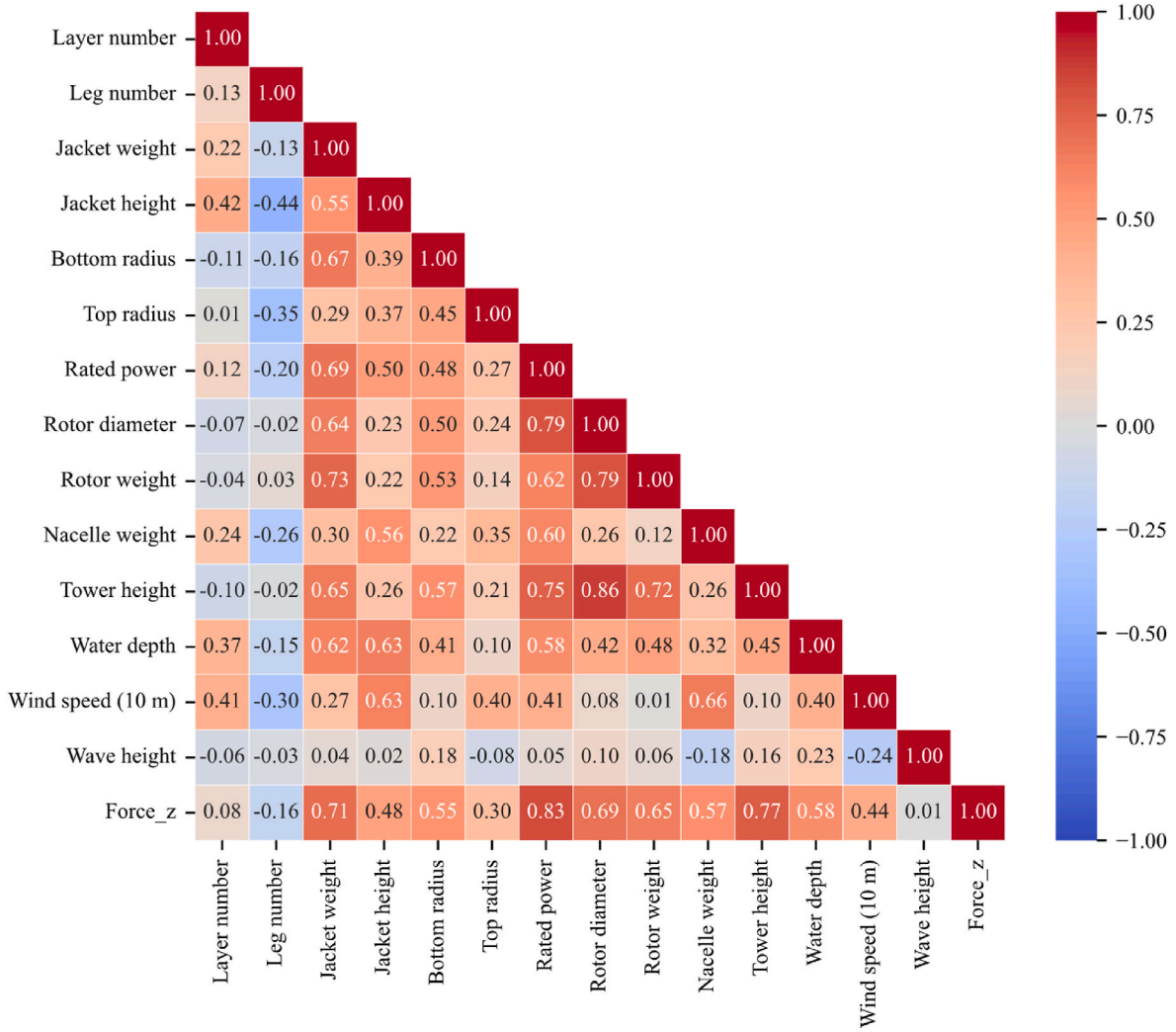


Fig. 14. Heat map of SCC values between candidate input features and target features.

summarized in Table 7. The models trained on the selected input features exhibit better predictive performance than those trained on all external features. For each target feature, the means of the metrics for both RF and XGBoost models showed a notable increase following feature selection, whereas the standard deviations of the metrics after feature selection were generally lower than those before feature selection. This indicates that feature selection enhances the accuracy of predictions and improves the stability of the trained models.

Both ML models generally exhibited comparable predictive performances, with RF showing a slight edge over XGBoost. Visualizations of the prediction of the RF models trained on feature sets before and after feature selection are presented in Figs. 15 and 16. One of the five trained RF models was selected for comparison for each target feature. For continuous target features, scatter plots were used to show the proximity between the predicted and real data. For discrete target features, receiver operating characteristic (ROC) curves were used to assess the ability of the model to differentiate between classes (Hand, 2009).

In Fig. 15, the scatter plots compare the predicted standardized values of the continuous target features with the corresponding real standardized values based on the two sets of input features. Ideally, if the predictions are perfect, all the points lie on the dashed diagonal line, representing the point where the predicted value equals the real value. The closer a point is to the line, the higher the accuracy of this prediction. The distribution of points indicates the performance of the model based on each set of input features with potentially different levels of accuracy and bias. From the four scatter plots, the blue points for the selected feature subset are more tightly clustered around the diagonal line than the red points for the set of all external features. This suggests that the selected feature subsets generally provide predictions that are closer to the real values, indicating a better performance in predicting the target features and matching the results in Table 7.

In Fig. 16, the ROC curves help determine how well the classification models can distinguish between the classes of discrete target features using two sets of input features. ROC curves were plotted with the True Positive Rate (TPR) on the y-axis and the False Positive Rate (FPR) on the x-axis. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the prediction. The area under the curve (AUC) measures the classification ability of a model. An AUC of 0.5 suggests no discriminative ability (equivalent to random guessing), while an AUC of 1.0 indicates perfect classification. Fig. 16(a) shows six ROC curves for the multi-class classification of the layer number, in which the solid blue, red, and orange curves represent the selected feature subset for the three classes, while the dashed lines of the same color represent all external features for each corresponding class. The AUCs for the RF models using all external features (i.e., 0.997, 0.956, and 0.964) are slightly lower than those using the selected feature subset (i.e., 1.000, 0.992, and 0.994), suggesting that the selected feature subset provides better performance for predicting the layer number based on the given dataset. Similarly, Fig. 16(b) shows the ROC curves for the binary classification of leg numbers using two sets of input features. The selected feature subset (solid blue curve) exhibits a slightly higher AUC than the set of all external features (dashed red curve), suggesting that it performs better at predicting the correct leg number.

According to Table 7, Figs. 15 and 16, the selected feature subsets generally provide better predictive performance for the ML models than the full set of external features, proving their feasibility in both regression and classification tasks based on the developed dataset.

4. Conclusions

This study demonstrated a novel and meaningful shift from conventional design methodologies to a data-driven approach using trained ML models in the conceptual design of offshore jacket substructures, automatically predicting jacket structural parameters based on specific boundary condition inputs. Establishing a comprehensive dataset for jackets and the subsequent feature selection process, including the feasibility evaluation of selected input feature subsets, represent the interconnected components of the innovative data-driven approach. The established dataset lays the foundation for training ML models, such as RF and XGBoost, during the entire process. Feature selection enhances the interpretability of the approach and ensures that the data-driven method aligns with the critical requirements inherent in physical-based structural designs. This harmonization of data-driven insights with conventional structural design considerations suggests more accurate, efficient, and reliable offshore jacket design strategies.

The dataset was compiled from completed and under-construction offshore jackets worldwide. Focusing on the key structural parameters and design boundary conditions aligned with LOD 100 ensured that the dataset contained the necessary details for conceptual design without overwhelming complexity. This balance is essential for practical applications because it represents various real-world scenarios in the conceptual design phase. The broad scope of the dataset provided a comprehensive perspective on actual jacket designs, enabling more informed and robust design decisions.

Building on this dataset, the proposed feature selection process enhances the understandability and applicability of the data-driven method for the jacket conceptual design. By systematically identifying the most influential and relevant features using Spearman's correlation, feature selection not only simplifies the models, making them more interpretable, but also addresses the challenges of overfitting and ensures more generalized predictive outcomes. Furthermore, the feasibility of the selected feature subsets was assessed using the RF and XGBoost models. These models were trained on both the selected feature subsets and all external features, allowing for a comparative analysis of their predictive performance. The results of this analysis not only demonstrate the effectiveness of the selected feature subsets in accurately predicting jacket structural parameters but also validate the practicability of the developed dataset for data-driven conceptual design.

However, the limitations of current investigations on data-driven design approaches are apparent. There are two main boundary conditions in the dataset. First, the development and utilization of offshore jackets have gained traction only in the last decade. Therefore, the sample size available for the dataset is constrained. Currently, it

Table 7
Mean μ and standard deviation σ of R^2 of regression models for continuous target features and F_1 of classification models for discrete target features.

Target feature	RF					XGBoost				
	n_estimator	Selected features		All external features		n_estimator	Selected features		All external features	
		μ	σ	μ	σ		μ	σ	μ	σ
Jacket height	4	0.869	0.081	0.798	0.065	7	0.808	0.027	0.576	0.176
Bottom radius	7	0.744	0.067	0.613	0.172	7	0.687	0.066	0.480	0.077
Top radius	6	0.794	0.065	0.550	0.127	13	0.793	0.085	0.660	0.097
Jacket weight	11	0.895	0.050	0.814	0.084	20	0.866	0.034	0.764	0.065
Layer number	17	0.924	0.031	0.728	0.175	23	0.849	0.013	0.716	0.111
Leg number	5	0.892	0.035	0.641	0.160	17	0.885	0.071	0.649	0.135

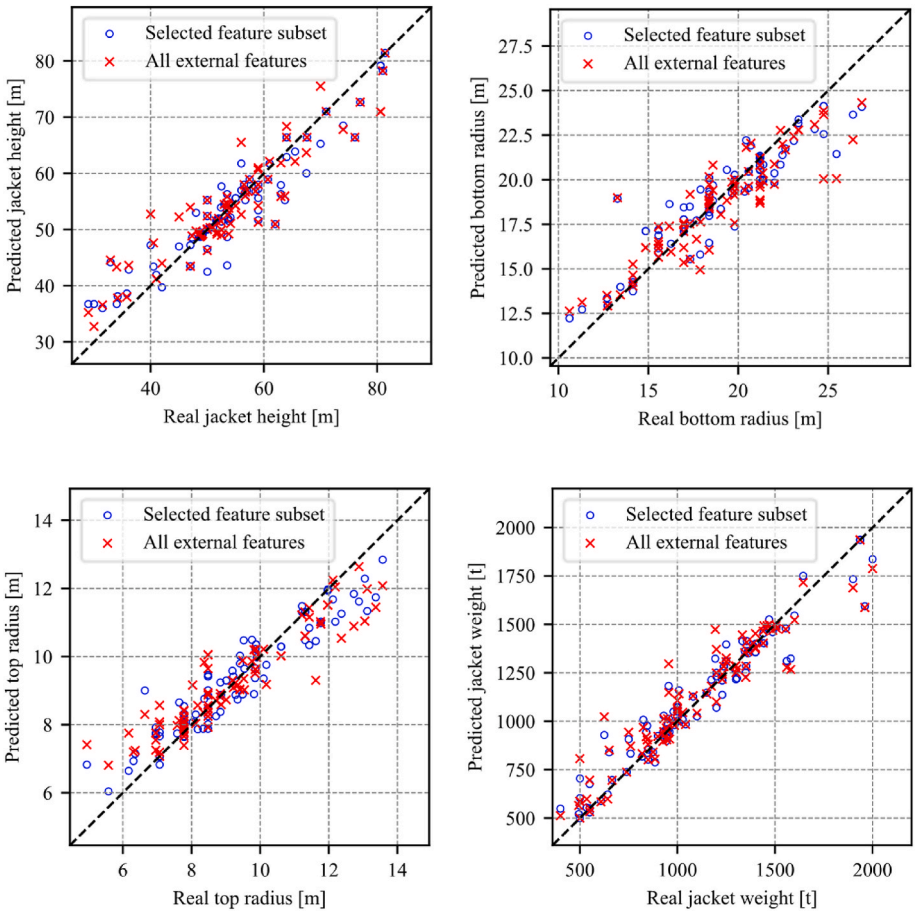
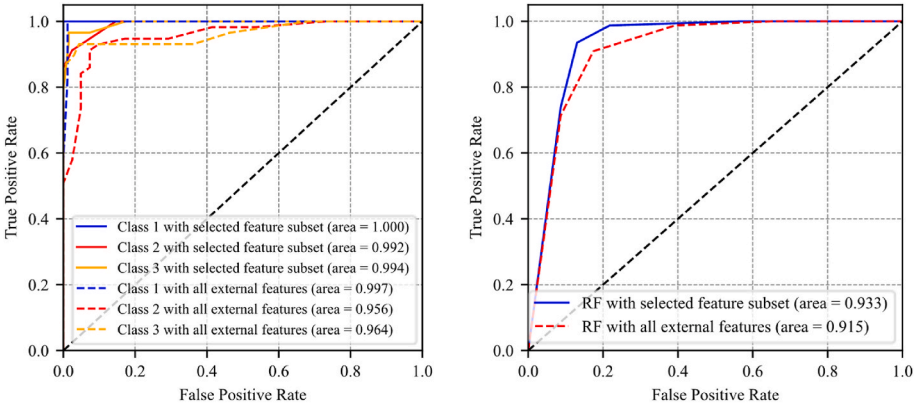


Fig. 15. Diagrams for comparing predicted and real data for continuous target features with respect to two sets of input features.



(a) ROC curves for the layer number (b) ROC curves for the leg number

Fig. 16. ROC curves for the discrete target features with respect to two sets of input features.

contains 100 instances of jacket substructures. Furthermore, because this study focuses only on the conceptual design phase of jackets, the structural parameters of jackets in this dataset primarily encompass general information that forms the topology and initial cost model of the jackets. Further detailed structural parameters, such as the cross-section sizes of the tubular members, will be determined through structural assessments and code checks in the following iterative design phase, which are not considered in this study.

The accuracy and applicability of the model to a broader range of design scenarios can be enhanced by expanding and refining the dataset with additional instances and more diverse data from new offshore wind projects because the accuracy and stability of the trained ML models are highly dependent on the size of the dataset. Generative models can be applied to generate synthetic jacket instances according to real jackets in the dataset, thereby expanding the dataset.

In addition, it is recommended that more advanced feature selection

methods and ML techniques can be integrated into the data-driven approach to capture essential features more robustly and handle more complex design scenarios with greater efficiency.

Furthermore, this approach can be extended to other types of offshore substructures, broadening its impact beyond jacket substructures. The principles and methodologies developed in this study can serve as prototypes for similar advancements in other areas of offshore engineering.

CRediT authorship contribution statement

Han Qian: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Emmanouil Panagiotou:** Visualization, Software, Methodology. **Mengyan Peng:** Writing – review & editing, Methodology, Data curation. **Eirini Ntoutsi:** Resources, Funding acquisition. **Chongjie Kang:** Writing – review & editing, Visualization, Validation, Conceptualization. **Steffen Marx:** Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgement

The authors want to acknowledge the support of the Collaborative Research Center 1463 (CRC1463) “Integrated Design and Operation Methodology for Offshore Megastructures” (project number 434502799) funded by the German Research Foundation (DFG) and the mdfBIM + project (project number 19FS2021C) funded by the Federal Ministry for Digital and Transport (BMDV).

References

4C Offshore, 2023. 4C Offshore Construction and Maintenance Vessel Online Data-Base. 4C Offshore Ltd. https://www.4coffshore.com/.
Akadiri, P.O., Chinyio, E.A., Olomolaiye, P.O., 2012. Design of A Sustainable building: a conceptual framework for implementing sustainability in the building sector. Buildings 2 (2), 126–152. https://doi.org/10.3390/buildings2020126.
AlHamaydeh, M., Barakat, S., Nasif, O., 2017. Optimization of support structures for offshore wind turbines using genetic algorithm with domain-trimming. Math. Probl Eng. 1–14. https://doi.org/10.1155/2017/5978375, 2017.
Alpaydin, E., 2020. Introduction to Machine Learning, fourth ed. The MIT Press.
Azur, M.J., Stuart, E.A., Frangakis, C., Leaf, P.J., 2011. Multiple imputation by chained equations: what is it and how does it work? Int. J. Methods Psychiatr. Res. 20 (1), 40–49. https://doi.org/10.1002/mpr.329.
Bauer, L., Matysik, S., 2023. The big portal for wind energy, wind-turbine-models. https://www.wind-turbine-models.com/.
Berger, R., Bruns, M., Ehrmann, A., Haldar, A., Häfele, J., Hofmeister, B., et al., 2021. EngiO – Object-oriented framework for engineering optimization. Adv. Eng. Software 153, 102959. https://doi.org/10.1016/j.advengsoft.2020.102959.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.
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.
Chen, I., Wong, B., Lin, Y., Chau, S., Huang, H., 2016. Design and analysis of jacket substructures for offshore wind turbines. Energies 9 (4), 264. https://doi.org/10.3390/en9040264.
Chew, K., Tai, K., Ng, E.Y.K., Muskulus, M., 2015. Optimization of offshore wind turbine support structures using an analytical gradient-based method. Energy Proc. 80, 100–107. https://doi.org/10.1016/j.egypro.2015.11.412.
Chew, K., Tai, K., Ng, E.Y.K., Muskulus, M., 2016. Analytical gradient-based optimization of offshore wind turbine substructures under fatigue and extreme loads. Mar. Struct. 47, 23–41. https://doi.org/10.1016/j.marstruc.2016.03.002.
Cho, H., Yuan, J., 2021. DMLC/XGBoost gradient boosting framework. https://github.com/dmlc/xgboost/.
Damiani, R., Dykes, K., Scott, G., 2016. A comparison study of offshore wind support structures with monopiles and jackets for U.S. waters. J. Phys.: Conf. Ser. 753, 92003 https://doi.org/10.1088/1742-6596/753/9/092003.
Damiani, R., Ning, A., Maples, B., Smith, A., Dykes, K., 2017. Scenario analysis for techno-economic model development of U.S. offshore wind support structures. Wind Energy 20 (4), 731–747. https://doi.org/10.1002/we.2021.
Damiani, R., Song, H., 2013. A jacket sizing tool for offshore wind turbines within the systems engineering initiative. In: OTC Offshore Technology Conference. Houston, Texas, USA, 06.05.2013 - 09.05.2013.
DNV, G.L., 2014. DNV-OS-J101–Design of Offshore Wind Turbine Structures. DNV GL, Oslo, Norway.
DTU, 2023. Wind Data Searching Tool, GWA 3.0. Global Wind Atlas. https://globalwindatlas.info/en/.
ERA5, 2023. Advancing Global NWP through International Collaboration. https://www.ecmwf.int/.
Fisch, R., Stecker, E., Kraus, M.A., 2023. Maschinelles Lernen beim Entwurf und der Bemessung von Stahlrahmenhallen. Stahlbau. https://doi.org/10.1002/stab.202200054. Article stab.202200054.
Gaertner, E., 2020. Definition of the IEA Wind 15-megawatt Offshore Reference Wind Turbine Tech. National Renewable Energy Laboratory, Golden, CO. Rep. NREL/TP-5000-75698.
Gasch, R., Twele, J., 2012. Wind power Plants. Fundamentals, Design, Construction and Operation, second ed. Springer, Springer Heidelberg Dordrecht London New York.
GWEC, 2022. Global offshore wind report 2022. https://gwec.net/wp-content/uploads/2022/06/GWEC-Global-Offshore-Wind-Report-2022.pdf.
Häfele, J., Damiani, R.R., King, R.N., Gebhardt, C.G., Rolfes, R., 2018. A systematic approach to offshore wind turbine jacket predesign and optimization: geometry, cost, and surrogate structural code check models. Wind Energ. Sci. 3 (2), 553–572. https://doi.org/10.5194/wes-3-553-2018.
Häfele, J., Gebhardt, C.G., Rolfes, R., 2019. A comparison study on jacket substructures for offshore wind turbines based on optimization. Wind Energ. Sci. 4 (1), 23–40. https://doi.org/10.5194/wes-4-23-2019.
Hand, D.J., 2009. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach. Learn. 77 (1), 103–123. https://doi.org/10.1007/s10994-009-5119-5.
Haury, A.C., Gestraud, P., Vert, J.P., 2011. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PLoS One 6 (12), e28210. https://doi.org/10.1371/journal.pone.0028210.
Ho, T., 2002. A data complexity analysis of comparative Advantages of decision forest Constructors. Pattern Anal. Appl. 5 (2), 102–112. https://doi.org/10.1007/s100440200009.
Hübler, C.J., Gebhardt, C.G., Rolfes, R., 2017. Development of a comprehensive database of scattering environment conditions and simulation constraints for offshore wind turbines. Wind Energ. Sci. https://doi.org/10.15488/4806.
IEC, 2019. Wind energy generation systems-Part 3-1: design requirements for fixed offshore wind turbines. International Standard IEC 61400–61403.
Jonkman, J., Sprague, M., 2017. OpenFAST in GitHub. https://github.com/openfast/.
Lehman, A., O’Rourke, N., Hatcher, L., Stepanski, E., 2013. JMP for Basic Univariate and Multivariate Statistics: Methods for Researchers and Social Scientists. Sas Institute.
Liu, Z., Burcharth, H.F., 1997. Design wave height related to structure Lifetime. In: Edge, Billy L. (Ed.), Coastal Engineering 1996. 25th International Conference on Coastal Engineering. Orlando, Florida, United States, September 2-6, 1996. American Society of Civil Engineers, New York, NY, pp. 2560–2572.
Marjan, A., Huang, L., 2023. Topology optimisation of offshore wind turbine jacket foundation for fatigue life and mass reduction. Ocean Eng. 289, 116228 https://doi.org/10.1016/j.oceaneng.2023.116228.
Mathern, A., Penadés-Plà, V., Armesto Barros, J., Yepes, V., 2022. Practical metamodel-assisted multi-objective design optimization for improved sustainability and buildability of wind turbine foundations. Struct. Multidiscip. Optim. 65 (2) https://doi.org/10.1007/s00158-021-03154-0.
NORSOK, 2004. Standard, design of steel structures. N-004. Rev. 2.
Oest, J., Sørensen, R.T., Overgaard, L.C., Lund, E., 2017. Structural optimization with fatigue and ultimate limit constraints of jacket structures for large offshore wind turbines. Struct. Multidiscip. Optim. 55 (3), 779–793. https://doi.org/10.1007/s00158-016-1527-x.
Offshore Engineer Magazine, 2020. Rystad: 2020 Biggest Year for Offshore Wind Jackets. https://www.oedigital.com/news/475622-rystad-2020-biggest-year-for-offshore-wind-jackets/.
Pigott, T.D., 2001. A review of methods for missing data. Educ. Res. Eval. 7 (4), 353–383. https://doi.org/10.1076/edre.7.4.353.8937.
Qian, H., Panagiotou, E., Marx, S., Ntoutsi, E., 2023. Data-based conceptual design of offshore jackets using a self-developed database. ISOPE International Ocean and Polar Engineering Conference, S. ISOPE-I-23-154.
Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Inc, Hoboken, NJ, USA.
Schmunk, R., 2023. Panoply netCDF, HDF and GRIB data Viewer v5. https://www.giss.nasa.gov/tools/panoply/.
Scikit-learn, 2023. Scikit-learn: machine learning in Python. https://scikit-learn.org/stable/.
Seidel, M., 2010. Design of support structures for offshore wind turbines - Interfaces between project owner, turbine manufacturer, authorities and designer. Stahlbau 79 (9), 631–636. https://doi.org/10.1002/stab.201001362.

Seidel, M., 2014. Substructures for offshore wind turbines-Current trends and developments. Festschrift Peter Schaumann 363–368. https://doi.org/10.2314/GBV:77999762X.
Sprent, P., Smeeton, N.C., 2016. Applied Nonparametric Statistical Methods. CRC press.
Stolpe, M., Sandal, K., 2018. Structural optimization with several discrete design variables per part by outer approximation. Struct. Multidiscip. Optim. 57 (5), 2061–2073. https://doi.org/10.1007/s00158-018-1941-3.
Torky, A.A., Aburawwash, A.A., 2018. A Deep learning approach to automated structural engineering of prestressed members. IJSCER 347–352. https://doi.org/10.18178/ijscer.7.4.347-352.
XGBoost, 2022. XGBoost Documentation. https://xgboost.readthedocs.io/en/stable/.