scientific reports

Check for updates

OPEN

Machine learning-driven stability analysis of eco-friendly superhydrophobic graphene-based coatings on copper substrate

Himanshu Prasad Mamgain[1]✉, Maria Vittoria Diamanti[2], Pravat Ranjan Pati[3], M. E. Mohamed[4], Jitendra Kumar Pandey[5]✉, Nitin Bhardwaj[6], Ankit Vasudeva[7] & Mohammad Kanan[8,9]✉

This study inspects the integration of machine learning (ML) techniques with materials science to develop durable, eco-friendly superhydrophobic (SHP) graphene-based coatings for copper. We employed various ML and regression models, including XGBoost, polynomial regression models, Random Forest (RF), K-Nearest Neighbours (KNN), and Support Vector Regression (SVR), to predict the stability of the contact angle (CA) under different stress conditions, such as NaCl immersion, abrasion cycles, tape peeling tests, sand impact, and open-air exposure. Our findings demonstrate that ensemble learning models, particularly XGBoost and Random Forest, outperform traditional regression techniques by effectively capturing nonlinear dependencies between stress parameters and CA retention. Higher-order polynomial regression models also exhibit strong predictive accuracy, making them well-suited for conditions where CA follows a well-defined trend. In contrast, SVR and KNN show limited generalization due to their sensitivity to hyperparameter selection and local interpolation effects, leading to weaker performance in datasets with high variability. ML-based algorithms predict CA values for tested coatings at longer term with respect to experimental tests, and underlined the beneficial effect of graphene incorporation in the coatings to extend the service life and preserve superhydrophobicity, overall reflecting the material's resilience under mechanical stress. The study highlights the importance of advanced predictive models, such as higher-degree polynomial regression and XGBoost, in capturing the complex relationships between variables influencing coating stability. Additionally, the integration of these models significantly accelerates the design and analysis process by reducing the reliance on time-consuming experimental testing.

Keywords Superhydrophobic, ML, Stability, Corrosion resistant, Wettability

Superhydrophobic (SHP) surfaces, characterized by a sliding angle (SA) less than 10° and a contact angle (CA) greater than 150°[1], have gained significant attention as their remarkable properties and broad applications, including drag reduction[2,3], self-cleaning[4–7], corrosion resistance[8], and water/oil separation[9–11]. Inspired by natural phenomena such as lotus leaves, SHP surfaces are designed by achieving a rough surface texture combined with surface chemistry modification using low-energy coatings. Various fabrication techniques like chemical vapor deposition[12], sol–gel[13], chemical etching[14], and electrodeposition[15] have been developed to create SHP surfaces. However, most methods face limitations due to their complexity, high cost, and specialized equipment requirements. Electrodeposition has emerged as a simple and scalable technique for producing

[1]Department of Physics, Applied Science, School of Advanced Engineering, UPES, Dehradun 248007, Uttarakhand, India. [2]Department of Chemistry, Materials and Chemical Engineering "Giulio Natta", Politecnico di Milano, Milan, Italy. [3]Department of Mechanical Engineering, Graphic Era (Deemed to be University), Dehradun 248002, Uttarakhand, India. [4]Chemistry Department, Faculty of Science, Alexandria University, Alexandria, Egypt. [5]Himalayan Institute for Learning and Leadership (HILL), UPES Dehradun, Dehradun 248002, Uttarakhand, India. [6]Lovely Professional University, Phagwāra, India. [7]Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura 140401, Punjab, India. [8]Department of Industrial Engineering, College of Engineering, University of Business and Technology, Jeddah 21448, Saudi Arabia. [9]Department of Mechanical Engineering, College of Engineering, Zarqa University, Zarqa, Jordan. ✉email: himanshuhm1111@gmail.com; Jeetusnu@gmail.com; m.kanan@ubt.edu.sa

SHP coatings on conducting polymers, metal oxides, and metals. This method allows for precise control and cost-effective fabrication of robust coatings, and is particularly relevant for copper substrates, which are widely utilized in numerous applications – e.g., heat conductors, electrical power lin es, and water pipelines. As copper is particularly prone to corrosion in solutions containing chlorides, enhancing the corrosion resistance of copper substrates through SHP coatings is critical for extending their lifespan. Nickel, known for its hardness and corrosion resistance, has been effectively used as a coating material on copper, offering additional benefits when integrated with SHP properties[16,17]. Despite these advantages, SHP surfaces often suffer from susceptibility to external damage, low mechanical stability caused by wear and abrasion, which deteriorate the surface micro/nanostructures, and consequent fast loss of superhydrophobicity, particularly in corrosive environments. Indeed, exposure to corrosive agents such as chloride ions, as well as acidic or alkaline environments, accelerates the degradation of the coatings, reducing their effectiveness in enhancing corrosion resistance. Furthermore, thermal and environmental factors, such as UV exposure and temperature fluctuations, also pose challenges to the long-term stability and durability of SHP surfaces[6].Addressing these challenges requires the development of mechanically and chemically stable SHP coatings[18]. In this frame, large attention has been dedicated to the incorporation of graphene in anti-corrosion coatings due to its strength, hydrophobicity, and chemical inertness. However, challenges such as poor substrate adhesion and ineffective hydrophobicity necessitate graphene modification, often achieved through doping with metals or non-metals.

In our previous study, Ni films and Ni-graphene composite coatings were fabricated on copper substrates using the electrodeposition technique, followed by treatment with myristic acid, a sustainable low-energy compound, to create SHP coatings[19,20]. Wettability, long term durability, corrosion resistance, mechanical stability, and chemical stability were evaluated for the fabricated coatings[18]. To address stability challenges, machine learning (ML) techniques are employed in this investigation. ML models, including classification, regression, and clustering algorithms, are utilized to predict key coating properties such as long-term durability, corrosion resistance, wettability, chemical and mechanical stability. These models analyse experimental data to uncover critical patterns and dependencies between material properties, environmental conditions, and performance metrics. The SHP coatings stability is influenced by multiple factors, including the physicochemical properties of the surface, environmental exposure conditions, and mechanical stresses experienced during operation. Traditional experimental methods utilized to estimate these factors are often labour-intensive, time-consuming, and costly. To address these challenges, ML can become an influential method for forecasting and assessing the stability of SHP coatings. By utilizing data-driven models, ML can reveal complex relationships between coating formulations, processing parameters, and their performance characteristics[21–23]. Despite the growing interest in SHP coatings, only a few studies were performed on ML to predict the long-term stability and anticorrosion behavior of these coatings. Past research has primarily focused on developing SHP coatings with myristic acid (MA) on various substrates, such as aluminum, copper and Steel[1,24,25]. However, there has been a lack of comprehensive studies analysing the effects of process parameters on coating responses. In some cases, researchers have utilized ML techniques such as random forests (RF)[26], artificial neural networks (ANN)[27], support vector machines (SVM)[28], extra trees (ET)[29], particle swarm optimization (PSO)[30], and genetic algorithms (GA)[31] to predict the outcomes of process parameters. Barai et al. used ANN models to predict the anticorrosion efficiency of SHP coatings and validated their predictions against experimental data, achieving highly accurate results[32]. Such studies highlight the potential of ML to optimize process parameters, reduce experimental workload, and enhance the understanding of SHP coating performance, particularly in terms of durability and anticorrosion capabilities. To show the potential of ML, water contact angle data from a previous study [33] were used to validate the ML algorithms developed in this article.

This study builds upon previously published experimental data but introduces a new and significant contribution through the application of machine learning (ML) techniques to analyze and predict the stability of superhydrophobic graphene-based coatings. The novelty lies in the use of multiple ML and regression models such as XGBoost, KNN, Random Forest, SVR, and polynomial regression to assess contact angle (CA) degradation under various environmental and mechanical stress conditions. Unlike prior work, this approach provides a comparative evaluation of model performance tailored to specific degradation mechanisms and reveals valuable insights into the nonlinear and long-term behavior of the coatings. The predictive modeling confirms that graphene incorporation (Ni-G-MA) enhances long-term superhydrophobicity, particularly under mechanical stress, thus offering a new perspective not addressed in earlier studies.

This study advances the current state of research by not only applying a wide range of machine learning (ML) models to predict the degradation of superhydrophobic coatings, but also by offering a comparative, condition-specific evaluation of these models under various real-world stress scenarios for hydrophobicity predication these ML model is never used before this is the first time ML is using for durability predication in such as sand impact, tape peeling, abrasion, and long-term exposure. While previous studies have applied ML in this domain, they often focus on a single model or lack detailed differentiation across degradation mechanisms.

Experimental details

Substrate production and characterization

The working electrode used in the previous work from where data were extracted [33] is a copper plate of dimensions 20 mm × 10 mm × 3 mm. The chemicals used for substrate preparation and coating by electrodeposition include anhydrous ethanol, boric acid, sulfuric acid, nickel chloride hexahydrate, sodium hydroxide, nickel sulfate, and myristic acid. We here briefly report the preparation procedures used. Prior to electrodeposition, copper substrates were sequentially polished using silicon carbide (SiC) abrasive papers ranging from grade 150 to 800, followed by ultrasonic cleaning in a soap solution for 10 min and brief immersion in 0.5 M H_2SO_4 for 1 min. The substrates were then rinsed with distilled water. The electrodeposition bath consisted of $NiSO_4$ (176 g/L), $NiCl_2{\cdot}6\ H_2O$ (40 g/L), and H_3BO_3 (60 g/L). Electrodeposition of nickel (Ni) films was carried out

at an applied potential of 8.75 V using a platinum rod as the anode and the copper substrate as the cathode. For the fabrication of nickel-graphene (Ni-G) films, a graphite rod was used as the anode, and electrochemical exfoliation of graphene was simultaneously achieved at 10.0 V. The exfoliation process was facilitated by the generation of oxygen and hydroxyl radicals from water, enhancing intercalation and delamination of graphene sheets. After deposition, both Ni and Ni-G films were rinsed with distilled water and dried at room temperature for 24 h. Surface modification was performed by immersing the dried films in 0.01 M myristic acid for 15 min, followed by ethanol rinsing and air drying. The resulting films, Ni-MA and Ni-G-MA, were subjected to further characterization and stability analysis[33].Samples were then subjected to mechanical and chemical stress, and coating stability and adhesion were tested through wettability measurements, as previously reported.

Machine learning framework for wettability of Ni-Ma-G coating

In this study, a ML framework is developed to predict the wettability of Ni-MA and Ni-G-MA coatings, utilizing experimental data CA measurements. The goal is to build a predictive model that captures the relationship between various coating characteristics and external factors that influence wettability, including environmental conditions and processing parameters.

The ML model incorporates a range of input features, including the coating deposition parameters and testing conditions such as abrasion cycles, tape peeling cycles, immersion in NaCl, and atmospheric exposure. CA is the primary target variable for the model, as it is the key indicator of coatings SHP properties. Data pre-processing steps were used to confirm the reliability of the model. These steps included normalization, feature selection, and outlier detection. A variety of ML algorithms were tested to determine their suitability for predicting the wettability of experimentally analysed coatings. Among these algorithms, artificial neural networks (ANN), random forests (RF) and support vector machines (SVM), XGBoost[34] and regression models[35] were evaluated for their ability to accurately predict CA and SA values under various conditions. To determine the model performance and ensure generalization to unseen data, cross-validation techniques were employed. These techniques help prevent overfitting and ensure that the model remains accurate when applied to new, unseen experimental data. Once trained, the ML model can predict the wettability characteristics of coatings under different experimental conditions, such as changes in temperature, voltage, pH levels, immersion time in NaCl solution, and number of abrasion cycles, referring to all mechanical damage tests performed (scratch, tape, sand impact). Additionally, to enhance the model performance, data augmentation techniques based on the introduction of Gaussian noise were applied. These techniques were compared with models trained on the original dataset to determine the effect of data augmentation on the model accuracy and generalization capability.

As above mentioned, the study also delves into the relationship between CA and key environmental factors, including immersion time in NaCl solution, long term durability in atmospheric exposure, pH levels, and the number of abrasion cycles. By examining these relationships, the study objective is to understand the factors that govern the wettability behavior of coatings in different conditions and to predict its changes under varying conditions. A regression model was fitted to further explore the nonlinearities and higher-order interactions between the environmental variables and the wettability of the coatings.

Machine learning using random forest and XGBoost

The XGBoost model, introduced by Chen et al.[34], is one of the most advanced and widely used ML algorithms given its exceptional efficiency and performance. At its core, XGBoost leverages the concept of gradient boosting, which is a sequential group learning technique. It constructs multiple "weak learners" (mainly decision trees) in a stepwise manner, each new model seeks to rectify the mistakes of its predecessors. This iterative process enhances the accuracy of predictions by focusing on the residuals (the errors) left by earlier trees.

One of the standouts feature of XGBoost is its ability to handle overfitting through the use of regularization. Unlike traditional gradient boosting methods, which may suffer from overfitting due to the accumulation of many weak models, XGBoost incorporates regularization terms into its objective function. This allows the model to control its complexity, balancing between fitting the training data well and maintaining the ability to generalize to new data. The objective function that XGBoost minimizes is a combination of two components: the prediction error and a regularization term that penalizes overly complex models, preventing overfitting. This is mathematically expressed as[34]

$$L = \sum_{i}^{l} (y_i^{\cdot} y_i) + \sum_{k}^{\Omega} \Omega\ (f_k), \quad \text{where} \quad \Omega\ (f_k) = \gamma_{\ T} + \frac{1}{2}\lambda\ ||\omega||^2 \tag{1}$$

Where L is the loss function, $y_i$is the actual value and $y_{i_1}^{\cdot}$ is the predicted value, $\Omega\ (f_k)$is the regularization term, which controls the complexity of the model, $\gamma_{\ T}$ is a parameter that penalizes the number of trees in the model. ω represents the weights of the model and λ is the regularization parameter that controls the weight of the L_2 norm (ridge regression-like penalty) on the model parameter By minimizing this objective function, XGBoost not only improves predictive performance, but also ensures robustness by preventing overfitting.

In contrast, the Random Forest (RF) model operates based on an entirely different philosophy. Rather than optimizing a single function, RF combines the predictions of multiple individual decision trees, each trained on a random portion of the dataset. Each tree in the forest generates its own prediction for a given input, and the final prediction is obtained by averaging the predictions of all trees in the ensemble. Considering the collection of decision trees, denoted as $\{T_1, T_2, \ldots, T_n\}$, in the random forest group, each tree, T_i generates a prediction y_i, for a given input x. In a regression problem, the random forest ensemble final prediction, y, is usually obtained by taking the average of the predictions made by all the trees.

$$y = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2}$$

The technique of averaging serves to mitigate the impact of individual tree projections and generates a more consistent and resilient overall prediction[35].

Linear regression model
Linear Regression is a supervised learning algorithm used for predicting continuous numerical values. It assumes a linear relationship between the independent variable(s) (features, X) and the dependent variable (target, y). The goal is to find the best-fit line that minimizes the difference between actual and predicted values.

Polynomial regression models
Polynomial regression extends linear regression by modelling the relationship between independent variables (predictors) and the dependent variable (response) as an nth-degree polynomial. This method is especially effective for capturing non-linear relationships within the data, which can provide a more accurate fit for complex patterns that linear models might fail to capture.

In the context of this study, polynomial regression is used to model the wettability of the Ni-MA and Ni-G-MA coating by considering the relation between the environmental and processing factors used in stressing the coatings and the resulting CA.

The polynomial regression equation with one independent variable can be expressed in the general form:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \in \tag{3}$$

Here, y is the dependent variable, x is the independent variable, β_0, β_1, β_2, ... β_n are the coefficients of the polynomial terms, $\in$is the error term and n is the degree of the polynomial.

The complexity of the relationship that the model can capture is determined by the degree of the polynomial. Polynomial regression models are commonly fitted using the method of least squares estimation, which aims to minimize the total sum of the squared discrepancies between the predicted and observed values. The polynomial terms coefficient is computed to optimize the fit to the data. The coefficients of the polynomial terms quantify the impact of each degree of the independent variable on the dependent variable.

The coefficient β_1 indicates the linear relation between the independent and dependent variables, β_2 captures the quadratic relationship, and so on for higher-order terms[36].

Support vector model (SVR)
Support Vector Regression (SVR) is a ML model designed for regression tasks, aiming to predict continuous values. It builds upon the principles of Support Vector Machines (SVM), which are primarily utilized for classification, but are adapted in SVR to forecast real-valued outputs. The fundamental concept of SVR involves identifying a function that accurately represents the data while maintaining a specified margin of tolerance (ε), meaning that deviations within this margin are not penalized. The model works by identifying support vectors, which are the data points closest to the regression function and optimizing the function to be as flat as possible while minimizing prediction errors. To address non-linear relationships, SVR employs kernel functions, including radial basis function (RBF), polynomial, and linear kernels. These functions transform the data into higher-dimensional spaces, enabling the identification of linear relationships. The SVR is highly effective in high-dimensional spaces and demonstrates robustness against overfitting. However, it can be computationally intensive, particularly when dealing with large datasets. Additionally, it necessitates meticulous tuning of hyperparameters, including the regularization parameter (C) and ε. Despite these challenges, SVR is widely used in applications like time series forecasting and stock market prediction, where the relationships between variables are often complex and non-linear[37].

The goal of SVR is to find a function f(x) that predicts the target variable with minimal deviation from the actual data points, within a margin. The general form of the regression function is:

$$f(x) = w.\theta(x) + b \tag{4}$$

Where w is the weight vector (the coefficients of the regression model), $\theta(x)$ is the mapping function that transforms the data into a higher-dimensional space through the kernel function and b is the bias term (the intercept).

Multilayer perceptron (MLP)
A Multilayer Perceptron (MLP) is a sophisticated type of artificial neural network constructed for supervised learning tasks, including classification and regression. It comprises several layers of interconnected nodes, or neurons, where each neuron in one layer is linked to every neuron in the subsequent layer. The architecture comprises an input layer responsible for data reception, one or more hidden layers dedicated to information processing, and an output layer that produces the final predictions. In the hidden layers, each neuron takes inputs from the preceding layer, calculates a weighted sum, incorporates a bias, and applies an activation function—like ReLU or sigmoid—to introduce non-linearity. The output layer then produces the final prediction using an appropriate activation function, like softmax for classification tasks or linear activation for regression. During the training phase, the network learns its weights and biases through backpropagation, where the discrepancy between predicted and actual outputs is sent back through the network. Gradient descent is employed to adjust the weights, aiming to minimize this error effectively. The process involves forward propagation, loss calculation,

and backward propagation, which are repeated until the model converges. The MLP ability to model complex relationships comes from its use of multiple hidden layers, enabling it to capture non-linear patterns in data[38].

Experimental dataset

The electrodeposited coatings on copper exhibited distinct micro-nano structures, with Ni-MA forming dendritic patterns and Ni-G-MA displaying a cauliflower-like morphology). SEM images of the coatings are reported with permission from [33] in Supplementary Figures S1. Contact angle (CA) measurements revealed that both Ni-G-MA and Ni-MA achieved superhydrophobicity (162° and 159°, respectively), compared to pure copper (58°) and nickel-coated copper (24°), with Ni-MA forming dendritic patterns and Ni-G-MA displaying a cauliflower-like morphology. WCA images of the coatings are reported with permission from[33] in Supplementary Figures S2. This behavior aligns with the Cassie-Baxter model, where micro- and nano-scale structuring enhances WCA.

Machine learning results

Linear regression, xgboost, random forest

This study investigates the correlation between superhydrophobicity and several influencing factors using machine learning models. The input variables included: number of days immersed in NaCl solution, pH levels, environmental exposure duration, number of tape peeling cycles, sand impact and number of abrasion cycles. The contact angle (CA) served as the sole output variable.

Various ML algorithms were applied to analyse the data, including XGBoost, Random Forest (RF), polynomial regression, multilayer perceptron (MLP), gradient boosting, support vector regression (SVR), and k-nearest neighbours (KNN). The experimental datasets taken from the previous research[1], and detailed in Tables S1–S12, were used to train separate models for each dataset. Additionally, the importance of each input variable in predicting the CA was evaluated and discussed, highlighting their influence on the SHP performance of the coatings[21]. Linear regression is initially used to capture any linear trends, providing a simple and direct approach to modelling the data. Polynomial regression models are then applied to better capture any nonlinear patterns, offering flexibility to model more complex relationships between CA and abrasion cycles. Additionally, we incorporate two advanced machine learning models: XGBoost, particularly effective in handling large datasets and complex relationships, and Random Forest, which combines predictions if multiple decision trees to improve accuracy.

The study initially utilized the dataset related to abrasion cycles in Tables S1, S2. The objective was to develop regression models using the XGBoost and Random Forest (RF) algorithms to predict CA based on the provided input. The dataset was randomly divided into training and testing subsets using an 80:20 split. This approach ensured that a substantial portion of the data was available for model training while preserving a separate subset for evaluation. Both the XGBoost and Random Forest (RF) models were trained using the training dataset, and their performance was evaluated with the test dataset.

Figure 1a represents the distribution of experimental values (in blue) and values predicted using linear regression, Random Forest and XG Boost ML models for Ni-MA coatings (red, green and yellow, respectively). Key performance metrics were then extracted and reported in Table 1, which included the mean squared error (MSE) and R-squared (R^2) values. The MSE measured the average squared differences between the predicted and actual CA values, serving as an indicator of prediction accuracy, while, the R^2 value represented the proportion of variance in CA that the model explained. Performance metrics for both models are summarized in Table 2, highlighting their effectiveness in forecasting the CA.

The Linear Regression model exhibited excellent performance, with an R^2 of 0.9979 and a low MSE of 0.229 on the testing data, indicating a near-perfect fit and a strong predictive power. The training data results were also impressive, with an R^2 of 0.9846 and an MSE of 1.118, suggesting that while the model performed well, it slightly underperformed on the training data compared to the testing set. This highlights the model ability to

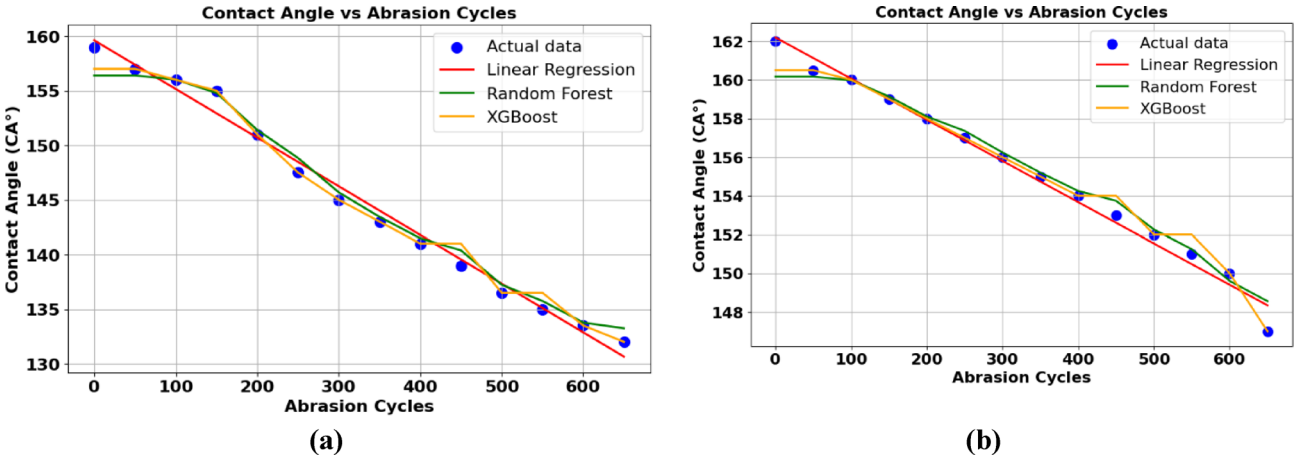

Fig. 1. Distribution of predicted and measured contact angle values from linear regression, random forest and Xgboost for (**a**) Ni-MA and (**b**) Ni-G-MA.

Model	MSE (testing)	R^2 (testing)	MSE (training)	R^2 (training)
Linear regression	0.2290	0.9979	1.1180	0.9846
Random forest	3.010	0.9719	0.4900	0.9933
XG boost	3.4182	0.9690	0.1000	0.9990

Table 1. MSE and R^2 value by linear regression, XG boost model and random forest model for Ni-MA.

Model	MSE	R^2
Polynomial regression (Order 2)	0.2993	0.9869
Polynomial regression (Order 3)	0.0428	0.9981
Polynomial regression (Order 4)	1.0509	0.9541
SVR	6.6876	0.7078
KNN	2.3056	0.8993

Table 2. MSE and R^2 value by polynomial regression of different orders, SVR and KNN models for Ni-G-MA.

Model	MSE (Testing)	R^2 (Testing)	MSE (Training)	R^2 (Training)
Linear regression	0.1557	0.9932	0.2732	0.9836
Random forest	1.3258	0.9421	0.2805	0.9832
XG Boost	1.4181	0.9380	0.0001	0.9999

Table 3. MSE and R^2 value by linear regression, XG boost model and random forest model for Ni-MA-G.

generalize effectively to unseen data while maintaining a good fit to the training data. In contrast, the Random Forest model showed a slightly higher MSE of 3.00976 on the testing data and an R^2 of 0.9719. While this is still a strong performance, it is less accurate than Linear Regression in terms of generalization to the testing data. However, the model performed well on the training data, with an MSE of 0.490 and an R^2 of 0.9933, suggesting that it was able to fit the training data very well. The difference between the training and testing performance indicates a potential overfitting of the Random Forest model to the training data. The XGBoost model, although powerful, demonstrated the highest testing MSE of 3.4182 and an R^2 of 0.9690, indicating slightly lower accuracy compared to both Linear Regression and Random Forest. On the training data, XGBoost performed exceptionally well, with an MSE of 0.100 and an R^2 of 0.999, suggesting that the model closely fits the training data. However, similar to Random Forest, this discrepancy in performance between training and testing data points to potential overfitting. Overall, Linear Regression provided the best balance of accuracy and generalization, with superior performance on both training and testing datasets. While Random Forest and XGBoost also demonstrated strong predictive capabilities, they showed a tendency to overfit the training data, leading to slightly reduced performance on the testing data.

Similar considerations can be drawn on data related to Ni-G-MA coatings, reported in Fig. 1b, whose figures or merit are summarized in Table 3. Indeed, also in this case Linear Regression achieved the highest testing R^2 with a value of 0.9932 and a low MSE of 0.1557, indicating strong generalization and accuracy. Random Forest had a slightly lower testing R^2 of 0.9421 with a higher MSE of 1.3258, suggesting it made larger prediction errors but still performed well. XGBoost showed the lowest testing R^2 of 0.9380 and the highest MSE of 1.4181, indicating higher error rates and potential overfitting, despite achieving perfect accuracy on the training data. Overall, Linear Regression outperformed the other models in terms of both accuracy and generalization.

To evaluate the predictive performance and robustness of the selected machine learning models, 5-fold cross-validation was performed using R^2 (coefficient of determination) and mean squared error (MSE) as evaluation metrics. Among the tested models, linear regression demonstrated the highest accuracy, achieving an average R^2 of 0.9616 with a low standard deviation of 0.0504, indicating consistent performance across folds. It also yielded the lowest average MSE of 1.3047, confirming its suitability for modeling degradation trends that follow a relatively linear behavior. The Random Forest model also performed well, with an average R^2 of 0.9557 and MSE of 2.1399, though slightly less accurate and stable than linear regression, possibly due to sensitivity to data noise or moderate overfitting. In contrast, XGBoost showed comparatively lower performance, with an average R^2 of 0.8697 and a higher MSE of 5.4084. Additionally, its larger standard deviations in both R^2 (0.0849) and MSE (1.9493) suggest higher variability and reduced generalization capability. These results indicate that while ensemble models like Random Forest and XGBoost offer flexibility, simpler models such as linear regression may be more effective when the underlying data trends are predominantly linear, offering more reliable and interpretable predictions.

Model	MSE (testing)	R^2 (testing)	MSE (training)	R^2 (training)
Linear regression	0.2100	0.9981	1.1028	0.9853
Random forest	2.2338	0.9797	0.5033	0.9933
XG boost	3.1506	0.9714	0.0059	0.9999

Table 4. MSE and R^2 value by linear regression, XG boost model and random forest model for Ni-MA after data augmentation.

Abrasion length (mm)	Experimental CA (°)	Predicted CA (°) (linear regression)	Predicted CA (°) (random forest)	Predicted CA (°) (XGBoost)
0	159	159	156	156
50	157	157	156	156
100	156	155	156	155
150	155	152	154	155
200	151	150	151	150
250	147	148	148	147
300	145	146	145	145
350	143	144	143	143
400	141	141	141	141
450	139	139	140	141
500	136	137	137	136
550	135	135	135	136
600	133	132	133	133
650	132	131	131	132

Table 5. Comparison between experimental and predicted CA by different ML models for Ni-MA after data augmentation.

Data augmentation for XGBoost and RF

Data augmentation was applied to the XGBoost and Random Forest models to mitigate overfitting and improve model performance. Among the different augmentation techniques available, Gaussian noise augmentation was selected for its simplicity, effectiveness, and ability to improve model performance. By adding small, controlled random perturbations, it helps prevent overfitting and increases the model exposure to varied data without altering the underlying structure. Unlike methods like SMOTE or jittering, Gaussian noise is easy to apply to both features and targets in regression tasks. It offers precise control over the noise level, allowing for fine-tuning based on model needs. This approach is computationally efficient, scalable, and ideal for situations where generating large datasets is impractical. Gaussian Noise Augmentation involves augmenting a given data point or feature vector, represented as x, by introducing Gaussian noise. Mathematically, this augmentation process can be expressed as[39],

$$x_{augmt} = x + \in \tag{5}$$

Where:

x_{augmt} : represent the augmented data point.

x : is the original data point.

$\in$: is a random sample drawn from a Gaussian distribution with a mean of 0 and a specified standard deviation.σ

The Gaussian distribution is defined as,

$$p(\in) = \frac{1}{sqrt(2\pi\sigma^2)} e^{-\frac{\in^2}{2\sigma^2}} \tag{6}$$

The parameter σ specifies the level of noise that is added using Gaussian Noise Augmentation. Smaller values of σ indicate reduced noise, whereas larger values of σ result in more pronounced noise. In this study, a mean of 0 and a variance of 0.2 are used as Gaussian distribution parameters.

As reported in Tables 4 and 5, data augmentation using Gaussian noise resulted in improved R^2 scores and reduced MSE for all three models, especially in the testing data. This suggests that augmenting the dataset with Gaussian noise helped the models generalize better by exposing them to a wider variety of data points, thus reducing overfitting. The slight improvements in the testing performance for Linear Regression and Random Forest highlight the benefits of adding augmented data, particularly when real-world data is limited or costly to acquire. XGBoost showed a more dramatic improvement in training performance, indicating that it might

Model	MSE (testing)	R^2 (testing)	MSE (training)	R^2 (training)
Linear regression	0.1582	0.9933	0.3768	0.9769
Random forest	1.3125	0.9444	0.2346	0.9856
XG boost	1.4085	0.9403	0.000779	0.9999

Table 6. MSE and R^2 value by linear regression, XG boost model and random forest model for Ni-G-MA after data augmentation.

Abrasion length (mm)	Experimental CA (°)	Predicted CA (°) (linear regression)	Predicted CA (°) (random forest)	Predicted CA (°) (XGBoost)
0	162	162	160	160
50	160	161	160	160
100	160	160	160	160
150	159	159	159	159
200	158	158	158	158
250	157	158	157	157
300	156	157	157	156
350	155	155	156	155
400	154	155	155	154
450	153	154	154	154
500	152	152	154	152
550	151	150	152	152
600	150	149	151	` 150
650	147	148	150	149

Table 7. Comparison between experimental and predicted CA by different ML models for Ni-G-MA after data augmentation.

have been overfitting to the original data, and Gaussian noise helped alleviate that issue, leading to better generalization on unseen data.

The Table 6 show that Linear Regression outperforms both Random Forest and XGBoost in terms of generalization, with the lowest testing MSE (0.1582) and a high R^2 (0.9933). This indicates that it provides accurate predictions without overfitting. In contrast, Random Forest and XGBoost exhibit higher testing MSE values (1.3125 and 1.4085, respectively) and lower testing R^2, suggesting they struggle more with generalization. Notably, XGBoost has an almost perfect training R^2 (0.9999) and an extremely low training MSE (0.000779), indicating overfitting, where the model memorizes training data but does not perform as well on new data.

Table 7 Figure S3 further confirms these trends, as Linear Regression closely follows the experimental CA values, while Random Forest and XGBoost show minor deviations, particularly at higher abrasion lengths. For example, at 650 mm, the experimental CA is 147, but Random Forest predicts 150, and XGBoost predicts 148, highlighting their slightly reduced accuracy. These results suggest that while all models perform reasonably well, Linear Regression remains the most reliable choice for this dataset, balancing both accuracy and generalization effectively.

To assess the generalization performance of the models, 5-fold cross-validation was conducted using mean squared error (MSE) and R^2 as evaluation metrics. Linear Regression outperformed the other models, achieving the lowest cross-validated MSE of 0.5355 and the highest R^2 of 0.9704, indicating excellent predictive accuracy and minimal error on unseen data. Random Forest followed with a higher MSE of 1.3847 and an R^2 of 0.9235, reflecting good but slightly less consistent performance. XGBoost showed the highest MSE at 1.8816 and the lowest R^2 of 0.8961 among the three, suggesting that while it can capture complex patterns. Overall, the results highlight that linear regression is the most suitable model for this dataset, likely due to the largely linear nature of contact angle degradation trends under the tested conditions.

Other ML models

Other models were applied to NI-MA and NI-G-MA CA values on different abrasion cycles to capture the corelation between input and output. Result of MSE and R^2 are listed in Table 8, Fig. 2.

Table 8 summarizes the performance of various regression models in predicting the contact angle (CA) variation with abrasion length for Ni-MA coatings. Among them, the third-order polynomial regression demonstrated the highest accuracy, with the lowest mean squared error (MSE: 0.2988) and the highest coefficient of determination (R^2: 0.9973), indicating a near-perfect fit. The second-order polynomial model also performed well (MSE: 1.8579, R^2: 0.9831), capturing the overall trend effectively. However, the fourth-order polynomial exhibited a slightly higher error (MSE: 2.1525, R^2: 0.9805), possibly due to overfitting. In contrast, the support vector regression (SVR) model showed poor predictive capability (MSE: 61.6268, R^2: 0.4409), likely due to its sensitivity to the dataset's complexity. The k-nearest neighbours (KNN) model performed better (MSE: 3.7870,

Model	MSE	R^2
Polynomial regression (Order 2)	1.8579	0.9831
Polynomial regression (Order 3)	0.2988	0.9973
Polynomial regression (Order 4)	2.1525	0.9805
SVR	61.6268	0.4409
KNN	3.7870	0.9656

Table 8. MSE and R^2 value by Poly regression (order 2, 3, 4), SVR, and KNN model for Ni-MA.

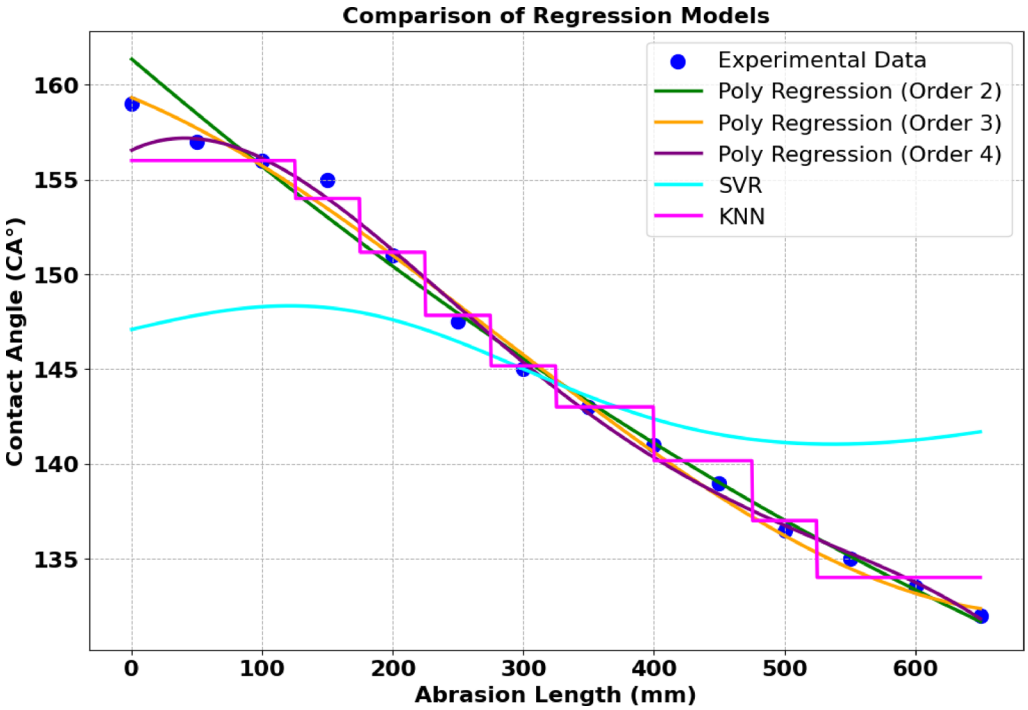

Fig. 2. Distribution of predicted and measured contact angle values from polynomial regression order 2,3,4, SVR and KNN models for Ni-MA.

R^2: 0.9656) but still lagged behind polynomial regression. KNN's performance is influenced by the choice of neighbours (k) and its reliance on local interpolation. Since CA variation with abrasion length follows a continuous nonlinear trend, KNN's distance-based approach may fail to generalize well across the entire dataset, leading to inconsistencies. Moreover, KNN is sensitive to noise, and the presence of abrupt changes in CA values can distort its predictions while SVR relies on finding a hyperplane that maximizes the margin within a specified tolerance (epsilon-tube) around the data points, which makes it highly sensitive to the choice of hyperparameters such as kernel type, regularization parameter (C), and epsilon value. If these parameters are not carefully tuned, the model may struggle to capture highly nonlinear patterns, as seen in CA variation with abrasion length. Additionally, SVR assumes that small deviations within the epsilon range are not significant, which might limit its accuracy in datasets where small variations are critical.

In the Table 2 the second-order polynomial regression model achieved a good fit with an MSE of 0.2993 and an R^2 of 0.9869, capturing most of the data variation. The third-order polynomial regression performed the best, with an exceptionally low MSE of 0.0428 and a high R^2 of 0.9981, showing high predictive accuracy. In contrast, the fourth-order polynomial regression showed a higher MSE of 1.0509 and a lower R^2 of 0.9541, suggesting overfitting as the model increased complexity. The Support Vector Regression (SVR) model exhibited the highest MSE of 6.6876 and an R^2 of 0.7078, indicating poor performance likely caused by its sensitivity to noise and lack of proper tuning. The K-Nearest Neighbours (KNN) model showed moderate results, with an MSE of 2.3056 and an R^2 of 0.8993, which was better than SVR but still behind the polynomial models.

In Fig. 3 the CA distributions obtained using polynomial regression order 2,3,4, as well as the models and KNN, are presented for Ni-G-MA. Overall, the third-order polynomial regression model outperformed the others. Indeed, the polynomial regression model, especially of order 3, effectively captured the nonlinear relationships within the data, which is crucial for the accurate prediction of contact angles. Compared to other models, Polynomial Regression (Order 3) demonstrated a better balance between bias and variance, as it did not overfit to the training data, unlike models such as XGBoost, which showed overfitting with a perfect training R^2 but a poor R^2 on testing data. While Linear Regression also performed well, it did not fully capture

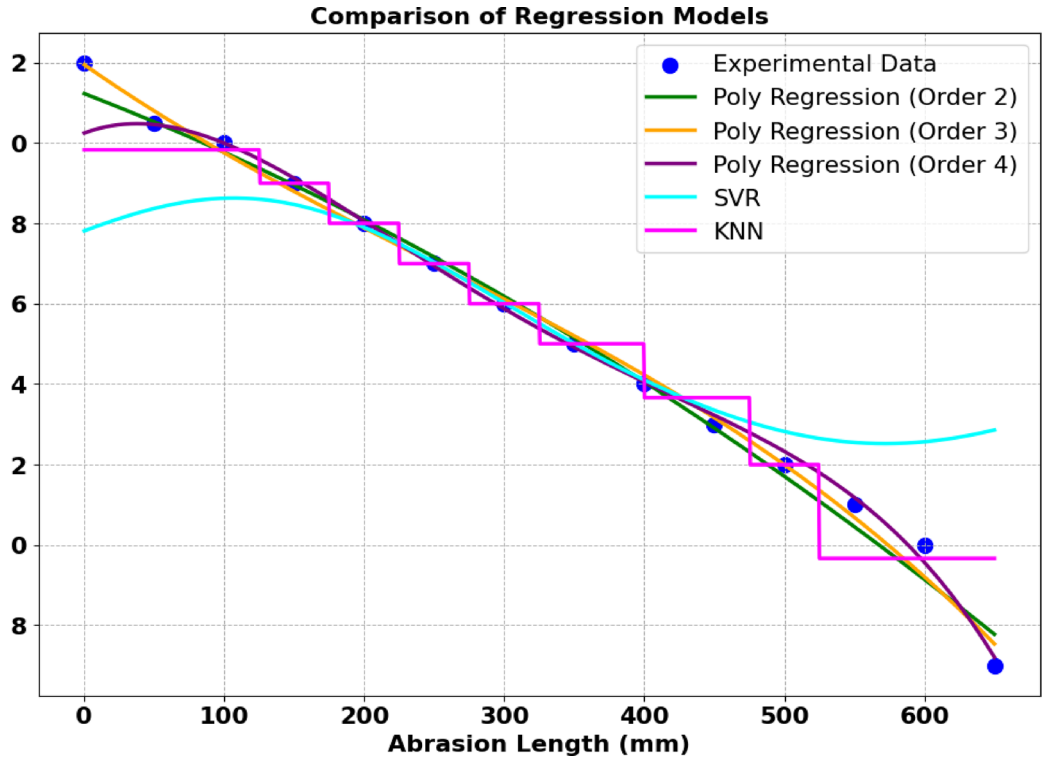

Fig. 3. Distribution of predicted and measured contact angle values from polynomial regression order 2,3,4, SVR and KNN models for Ni-G-MA.

the complexities of the dataset, and models like SVR and KNN yielded lower performance in terms of R^2 and MSE. Therefore, Polynomial Regression (Order 3) stands out due to its ability to model the data underlying patterns more effectively and with greater accuracy. Polynomial regression, particularly the second- and third-order models, performed best because they effectively captured the dataset's nonlinear pattern without excessive complexity. The fourth-order polynomial suffered from overfitting. SVR performed the worst, likely due to improper tuning and sensitivity to noise, while KNN showed decent but not outstanding performance due to its reliance on local neighbourhood relationships rather than an explicit global function.

The performance of various regression models was evaluated using cross-validation to predict the target output. Among the polynomial models, the third-degree polynomial (Poly3) exhibited the best performance with the lowest mean squared error (MSE) of 0.6649 and the highest coefficient of determination (R^2) of 0.9918, indicating a strong fit to the data. In comparison, Poly2 and Poly4 showed slightly higher MSE values of 1.2249 and 1.2924, with corresponding R^2 values of 0.9849 and 0.9840, respectively. The k-nearest neighbors (KNN) model also demonstrated good predictive accuracy with an MSE of 3.5675 and an R^2 of 0.9559. However, the support vector regression (SVR) model yielded significantly poorer results, with a high MSE of 46.4255 and a low R^2 of 0.4262, indicating limited suitability for this dataset. Overall, the Poly3 model emerged as the most reliable among the tested algorithms.

Tape peeling and chemical stability datasets

The dataset presented in Table S3, Table S4 was analysed to investigate the relationship between the contact angle (CA) and tape peeling cycles. To achieve this, various modelling techniques were applied, including linear regression, polynomial regression, XGBoost, and Random Forest, as illustrated in Fig. 6. For capturing potential nonlinear relationships, polynomial regression models of degrees 2 through 4 were utilized, offering enhanced flexibility to accurately represent complex patterns within the data. Advanced ML models, such as XGBoost and Random Forest, were also employed as their ability to handle intricate, nonlinear associations effectively. These approaches provided a robust framework to complement traditional regression methods, enabling a more comprehensive understanding of the underlying trends in the dataset.

Table 9 presents the Mean Squared Error (MSE) and R^2 values for different regression models applied to Ni-MA data. The results indicate that the XGBoost model outperforms all others, achieving the lowest MSE (0.0001) and the highest R^2 (0.9999), demonstrating its superior predictive accuracy. Random Forest and higher-order Polynomial Regression models also exhibit strong performance, with R^2 values exceeding 0.99, while Linear Regression has the highest MSE (0.4090), suggesting a weaker fit. The trend highlights the advantage of non-linear models in capturing complex relationships within the data, with XGBoost providing the most precise predictions (Fig. 4).

In Table 10; Fig. 5, the results illustrate the effectiveness of various models in predicting the CA based on tape-peeling cycles. Linear regression performed well, with an MSE of 0.1161 and an R^2 of 0.9931, capturing

Model	MSE	R^2
Linear regression	0.4090	0.9715
Random forest	0.07834	0.9945
XG Boost	0.0001	0.9999
Polynomial regression (Order 2)	0.1649	0.9885
Polynomial regression (Order 3)	0.0886	0.9938
Polynomial regression (Order 4)	0.0629	0.9956

Table 9. MSE and R^2 value by linear regression, XG boost model and random forest model for Ni-MA.

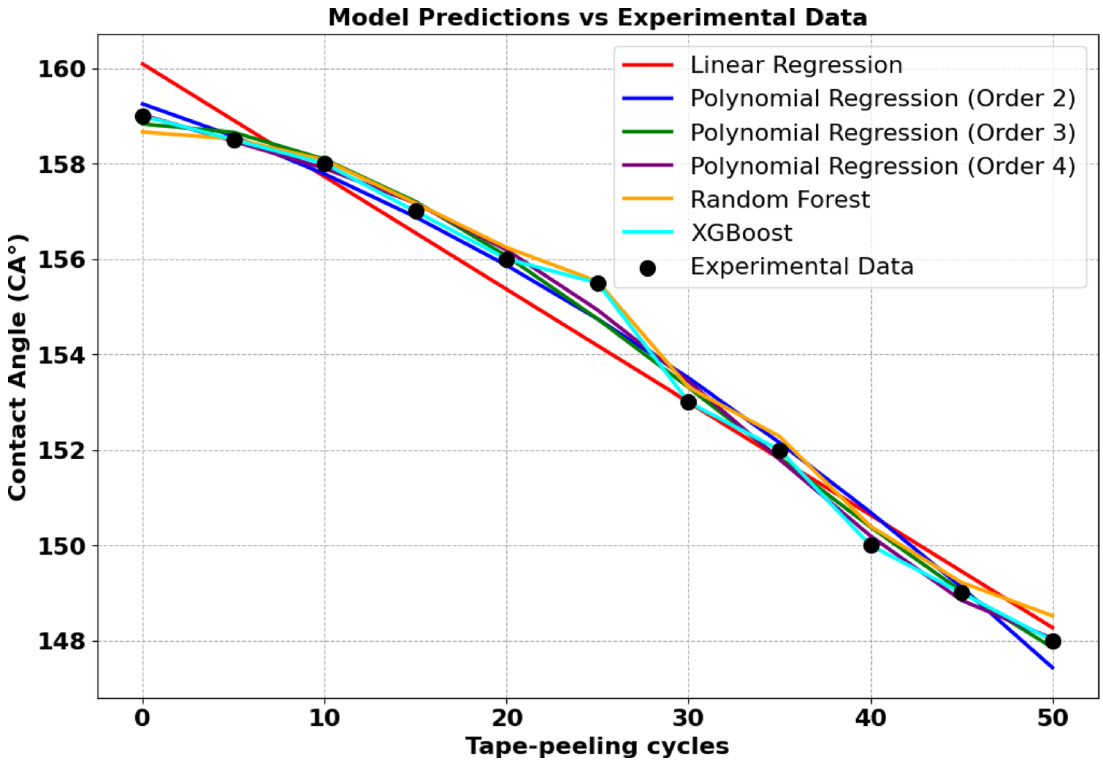

Fig. 4. Distribution of predicted and measured contact angle values from linear regression XGBoost, Random Forest, poly order 2,3,4 for Ni-MA.

Model	MSE	R^2
Linear regression	0.1161243	0.993144
Random Forest	0.1561192	0.990783
XG Boost	0.000988	0.999
Polynomial Regression (Order 2)	0.1158841	0.993158
Polynomial Regression (Order 3)	0.06209175	0.996334
Polynomial Regression (Order 4)	0.06201943	0.996338

Table 10. MSE and R^2 value by linear regression, XG boost model and random forest model for Ni-G-MA.

the overall linear trend in the data. However, nonlinear models offered better flexibility in identifying subtle patterns. Polynomial regression models demonstrated significant improvements, with the second-order model achieving an MSE of 0.1159 and an R^2 of 0.9932, while the third- and fourth-order models further enhanced predictions, both achieving an MSE of 0.0620 and an R^2 of 0.9963. Advanced machine learning techniques were also applied, with Random Forest achieving an MSE of 0.1561 and an R^2 of 0.9908, effectively modelling the dataset's nonlinear features. The XGBoost model outshone all others, achieving an outstandingly low MSE of 0.00098 and an R^2 of 0.999, underscoring its ability to model complex relationships with exceptional accuracy

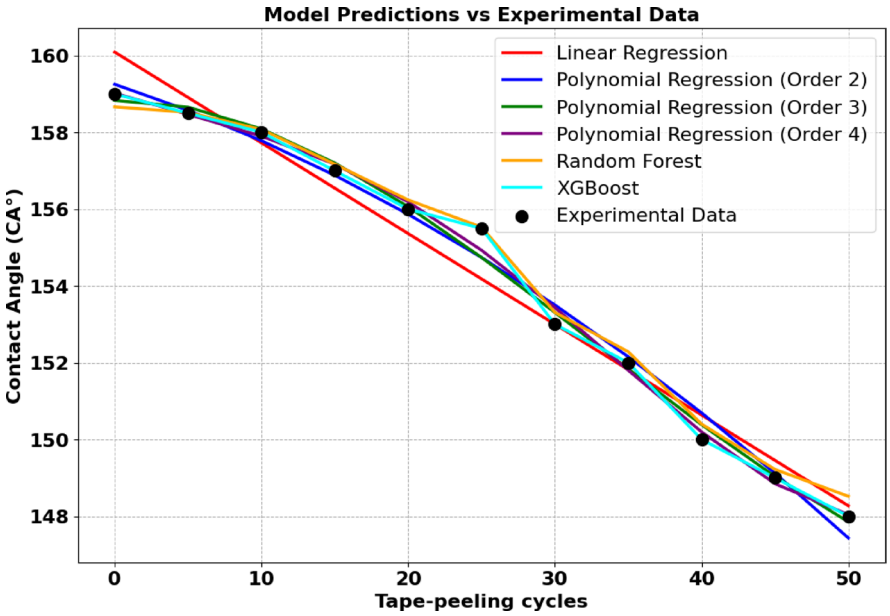

Fig. 5. Distribution of predicted and measured contact angle values from linear regression XGBoost, Random Forest, poly order 2,3,4 for Ni-G-MA.

Model	MSE	R^2
Linear regression	0.1336	0.9863
Random Forest	0.3162	0.9675
XG Boost	1.4160	0.8544

Table 11. MSE and R^2 value by linear regression, XG boost model and random forest model for Ni-MA.

and reliability. While polynomial regression and Random Forest provided strong performance, XGBoost proved to be the most accurate and versatile model.

The comparative evaluation of regression models using 5-fold cross-validation revealed that polynomial regression models outperformed other approaches in terms of prediction accuracy. Specifically, the fourth-order polynomial regression model achieved the best performance, with the lowest mean squared error (MSE) of 0.1674 and the highest coefficient of determination (R^2) of 0.9884. This was closely followed by the third-order polynomial model (MSE: 0.1951, R^2: 0.9864), indicating that increasing the polynomial order up to a certain extent improves model fit. The second-order polynomial and linear regression models also showed strong performance, with R^2 values of 0.9704 and 0.9640, respectively. In contrast, ensemble models such as Random Forest and XGBoost demonstrated relatively lower predictive accuracy, with MSE values of 0.7424 and 1.4772, and corresponding R^2 values of 0.9484 and 0.8972. These findings confirm the effectiveness of polynomial regression, particularly of orders three and four, in capturing the underlying patterns in the dataset.

In the dataset presented in Table S5, S6 we begin the analysis by plotting the CA against the number of sand impact cycles. To show the relationship between these variables, we employ a combination of polynomial regression, linear regression, XGBoost, and Random Forest models.

Table 11 presents the Mean Squared Error (MSE) and R^2 values for Ni-MA modeling using Linear Regression, Random Forest, and XGBoost. The Linear Regression model demonstrates the best performance with the lowest MSE (0.1336) and highest R^2 (0.9863), indicating strong predictive accuracy and minimal error. The Random Forest model shows slightly higher error (MSE = 0.3162) and a lower R^2 (0.9675), suggesting a moderate decrease in performance. In contrast, the XGBoost model exhibits significantly higher MSE (1.4160) and lower R^2 (0.8544), indicating a poorer fit, due to overfitting, hyperparameter tuning issues.

Figure 6 presents a scatter plot comparing the predicted and experimental contact angles (CA°) as a function of the number of sand impact cycles. The black dots represent experimental data, while predictions from Linear Regression (LR), Random Forest (RF), and XGBoost models are indicated by red crosses, green crosses, and blue crosses, respectively.

The plot shows that Linear Regression and Random Forest predictions closely align with the experimental values, particularly at lower sand impact cycles, reinforcing their superior performance (as indicated by the lower MSE and higher R^2 values in Table 11). In contrast, the XGBoost model exhibits greater deviations, especially at higher impact cycles,

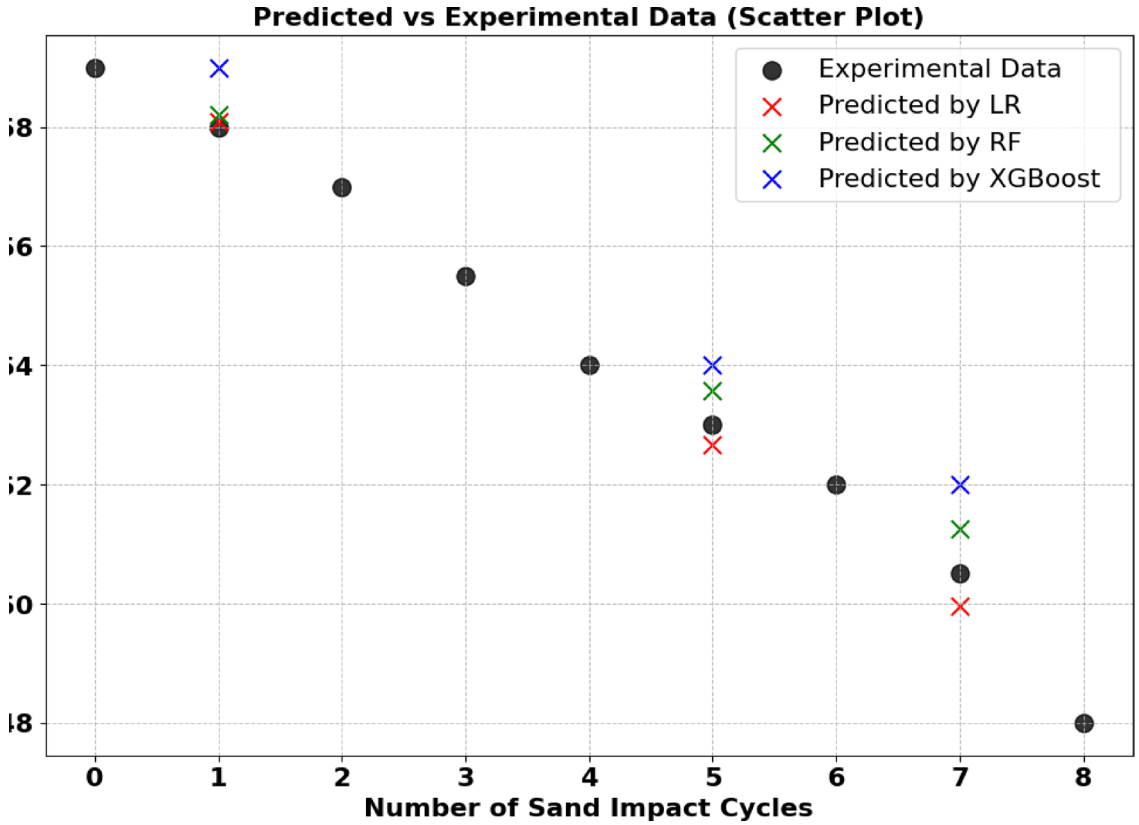

Fig. 6. Distribution of predicted and measured contact angle values from linear regression XGBoost, Random Forest, for Ni-MA.

Model	MSE	R^2
Linear regression	0.1250	0.9945
Random Forest	1.0448	0.9537
XG Boost	3.1234	0.8616

Table 12. MSE and R^2 value by linear regression, XG boost model and random forest model for Ni-G-MA.

In Table 12; Fig. 7, the Linear Regression model is the best, with the lowest MSE (0.1250) and highest R^2 (0.9945), showing it fits the data most accurately. The other models, Random Forest and XGBoost, have higher MSE and lower R^2 values, suggesting less accurate predictions.

The 5-fold cross-validation analysis further highlights the superiority of linear regression for this dataset. Linear regression achieved the lowest mean squared error (MSE) of 0.2820 and the highest coefficient of determination (R^2) of 0.9826, indicating an excellent fit and strong predictive capability. In contrast, ensemble models such as Random Forest and XGBoost showed relatively lower performance, with MSE values of 1.6400 and 2.2089, and R^2 values of 0.8987 and 0.8635, respectively. Although ensemble techniques are generally robust for complex, non-linear data, in this case, the simplicity of linear regression appears to better capture the underlying relationships, likely due to the dataset's linear or near-linear characteristics.

In Table S7, S8 we analyse the relationship between the CA and number of days in NaCl (0.5 M) solution by plotting the data and applying again various machine learning models.

Cross-validation (5-fold) was conducted to evaluate the performance and generalization capability of different regression models. Linear regression exhibited the best performance among the evaluated models, with a mean squared error (MSE) of 0.1887 ± 0.1297. Random Forest and XGBoost showed significantly higher errors, with MSE values of 2.1400 ± 1.3269 and 3.8006 ± 0.9731, respectively.

Figure. 8 illustrates the variation of contact angle (CA°) over time of immersion in NaCl (days) using actual data (black dots) and predictions from three models: Linear Regression (blue line), Random Forest (orange line), and XGBoost (green line). The Linear Regression model follows a smooth, linear trend, closely approximating the overall experimental pattern. The Random Forest and XGBoost models, in contrast, exhibit step-like predictions, characteristic of tree-based models. While Random Forest predictions align relatively well with actual data, XGBoost shows larger deviations, particularly at certain intervals, reinforcing its higher MSE

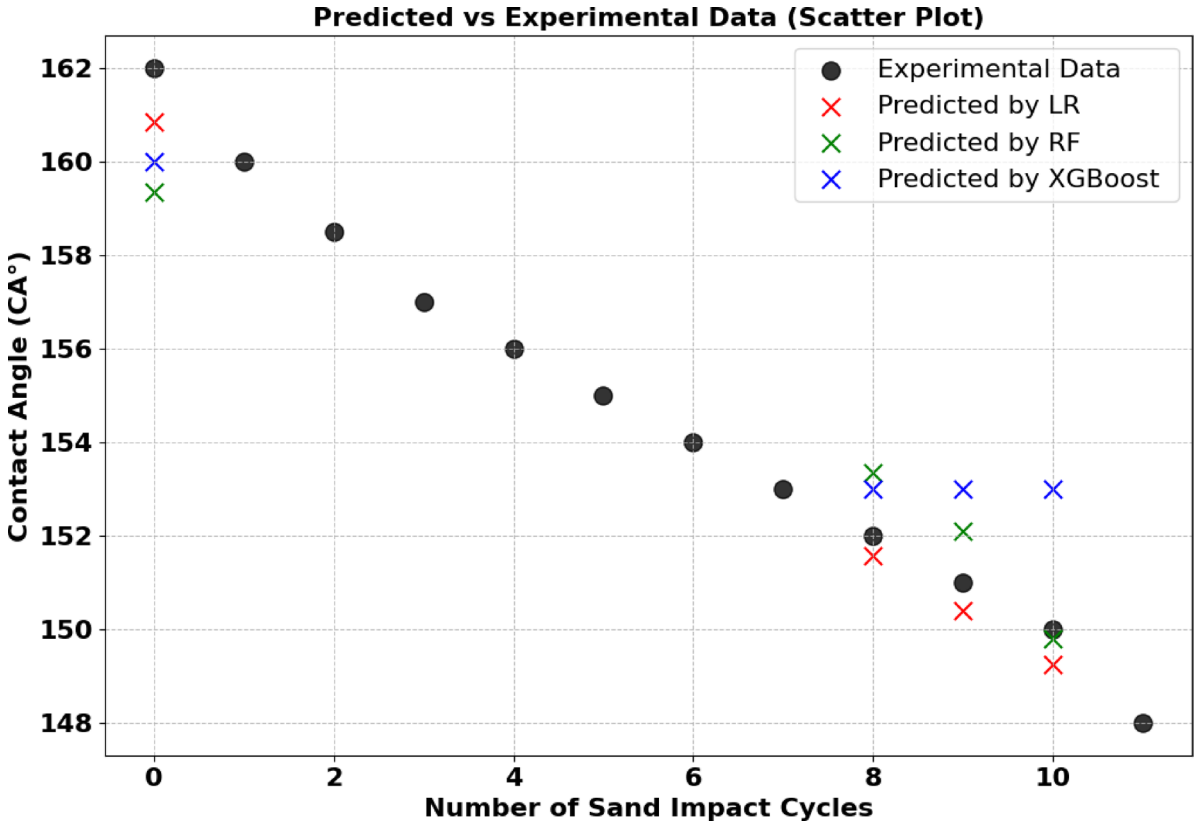

Fig. 7. Distribution of predicted and measured contact angle values from linear regression XGBoost, Random Forest, for Ni-G-MA.

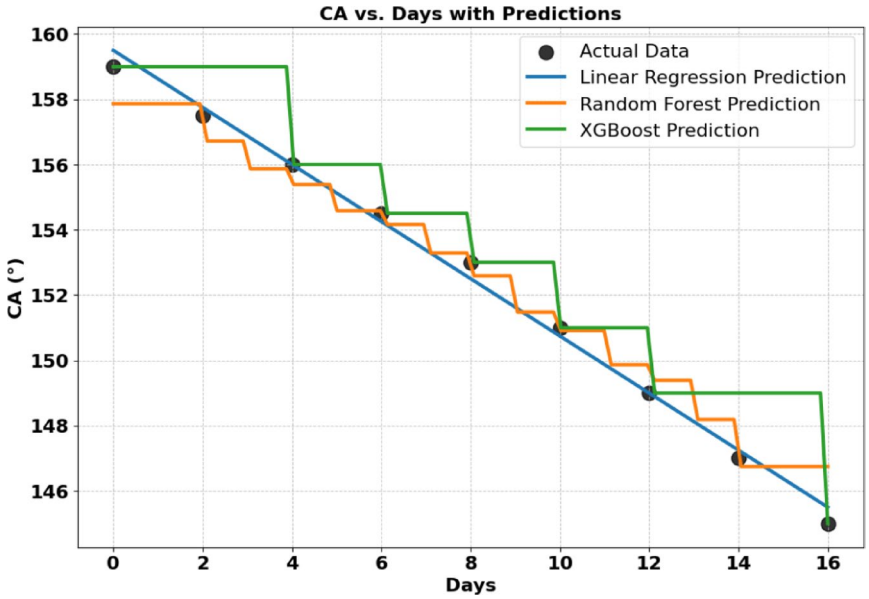

Fig. 8. Distribution of predicted and measured contact angle values from linear regression, XGBoost, Random Forest, for Ni-MA.

(3.1235) and lower R^2 (0.8867) Table 13. This suggests that XGBoost struggles with the dataset's characteristics, potentially due to overfitting or sensitivity to variations in the data.

In Table 14; Fig. 9, considering the same phenomenon on coating Ni-G-MA, the Linear Regression model is the best, with the lowest MSE (0.1250) and highest R^2 (0.9945), showing it fits the data most accurately. The

Model	MSE	R^2
Linear regression	0.0625	0.9977
Random Forest	0.7729	0.9720
XG Boost	3.1235	0.8867

Table 13. MSE and R^2 value by linear regression, XG boost model and random forest model for Ni-MA.

Model	MSE	R^2
Linear regression	0.1250	0.9945
Random Forest	1.0448	0.9537
XG Boost	3.1234	0.8616

Table 14. MSE and R^2 value by linear regression, XG boost model and random forest model for Ni-G-MA.

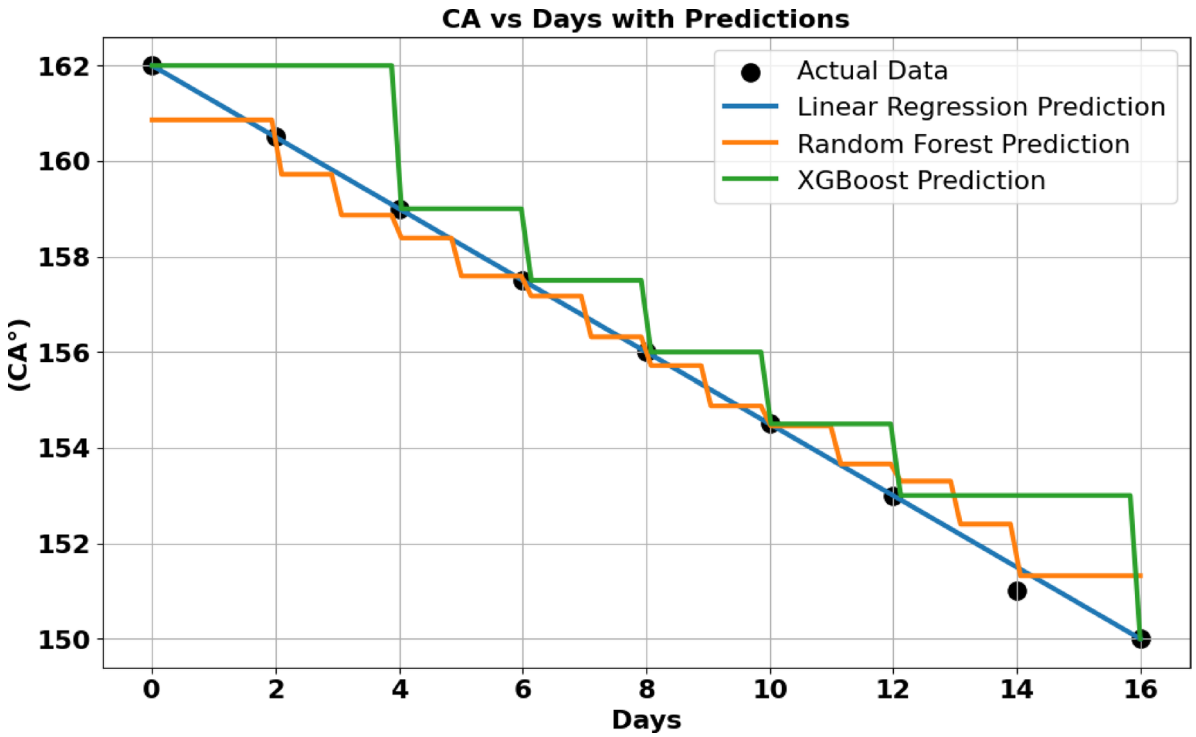

Fig. 9. Distribution of predicted and measured contact angle values from linear regression XGBoost, Random Forest, for Ni-G-MA.

other models, Random Forest and XGBoost, have higher MSE and lower R^2 values, suggesting less accurate predictions.

ML models were also applied to data in Table S9, S10 to find out correlation between CA and number of days of exposure in open environment.

Table 15; Fig. 10 compare the performance of various models in predicting contact angle values after outdoor exposure in coating Ni-MA. The Polynomial Regression models of order 2 and 3 achieve the best results with the lowest MSE (0.1125) and highest R^2 (0.9854), indicating excellent predictive accuracy. Random Forest also performs well (MSE = 0.3085, R^2 = 0.9601), outperforming Linear Regression (MSE = 0.4228, R^2 = 0.9452) and XGBoost (MSE = 0.6668, R^2 = 0.9137). The fourth-order polynomial model (MSE = 0.2446, R^2 = 0.9683) shows slight overfitting, suggesting that a second- or third-order polynomial is optima for this dataset.

In Table 16 Fig. 11, when applying the models to outdoor exposure of coating Ni-G-MA, the Polynomial Regression of Order 3 performs the best, with the lowest MSE (0.0393) and the highest R^2 (0.9949). This indicates a very close fit to the data, capturing the underlying relationship with high accuracy. Polynomial Regression of higher orders (4) and other models, such as Random Forest and XGBoost, show slightly worse performance, likely due to overfitting (in the case of higher polynomial orders) or a less optimal fit compared to the cubic

Model	MSE	R^2
Linear regression	0.4228	0.9452
Random Forest	0.3085	0.9601
XG Boost	0.6668	0.9137
Poly Order 2	0.1125	0.9854
Poly Order 3	0.1125	0.9854
Poly Order 4	0.2446	0.9683

Table 15. MSE and R^2 value by linear regression, XG boost model and random forest model for Ni-MA.

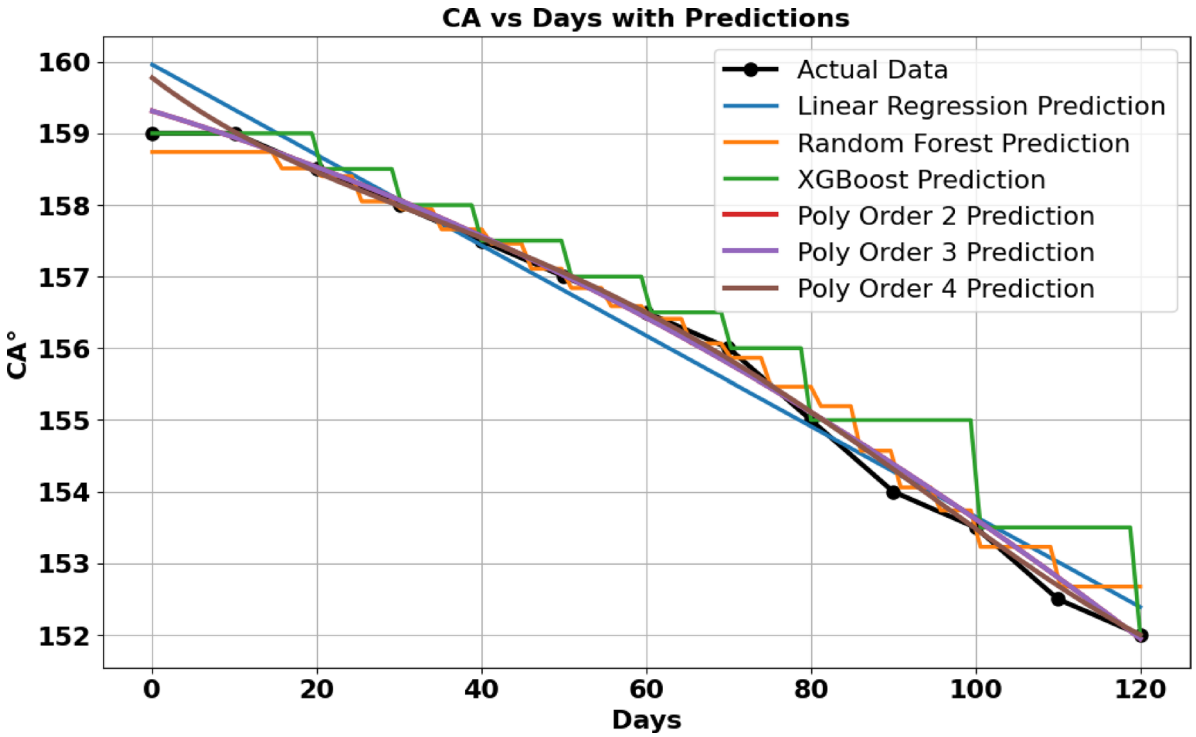

Fig. 10. Distribution of predicted and measured contact angle values from linear regression XGBoost, Random Forest, for Ni-MA.

Model	MSE	R^2
Linear regression	0.5317	0.9311
Random Forest	0.1792	0.9768
XG Boost	0.4168	0.9460
Poly Order 2	0.4358	0.9436
Poly Order 3	0.0393	0.9949
Poly Order 4	0.4466	0.9422

Table 16. MSE and R^2 value by linear regression, XG boost model and random forest model for Ni-G-MA.

model. The Linear Regression also performs reasonably well, but its MSE and R^2 indicate it doesn't capture the more complex relationship as effectively as the cubic polynomial.

A detailed 5-fold cross-validation was performed to assess the predictive accuracy and stability of multiple regression models. Among all models, polynomial regression of order 2 demonstrated the best performance, achieving the lowest mean squared error (MSE) of 0.0566 ± 0.0506 and the highest mean R^2 value of 0.9866 ± 0.0081, indicating excellent fit and low variance across folds. Polynomial regressions of orders 3 and 4 followed closely, with MSEs of 0.0601 and 0.0771 and R^2 values of 0.9797 and 0.9796, respectively, suggesting that higher-order terms do not significantly enhance model performance beyond the second order. Among the non-polynomial models, Random Forest outperformed both Linear Regression and XGBoost, achieving an MSE

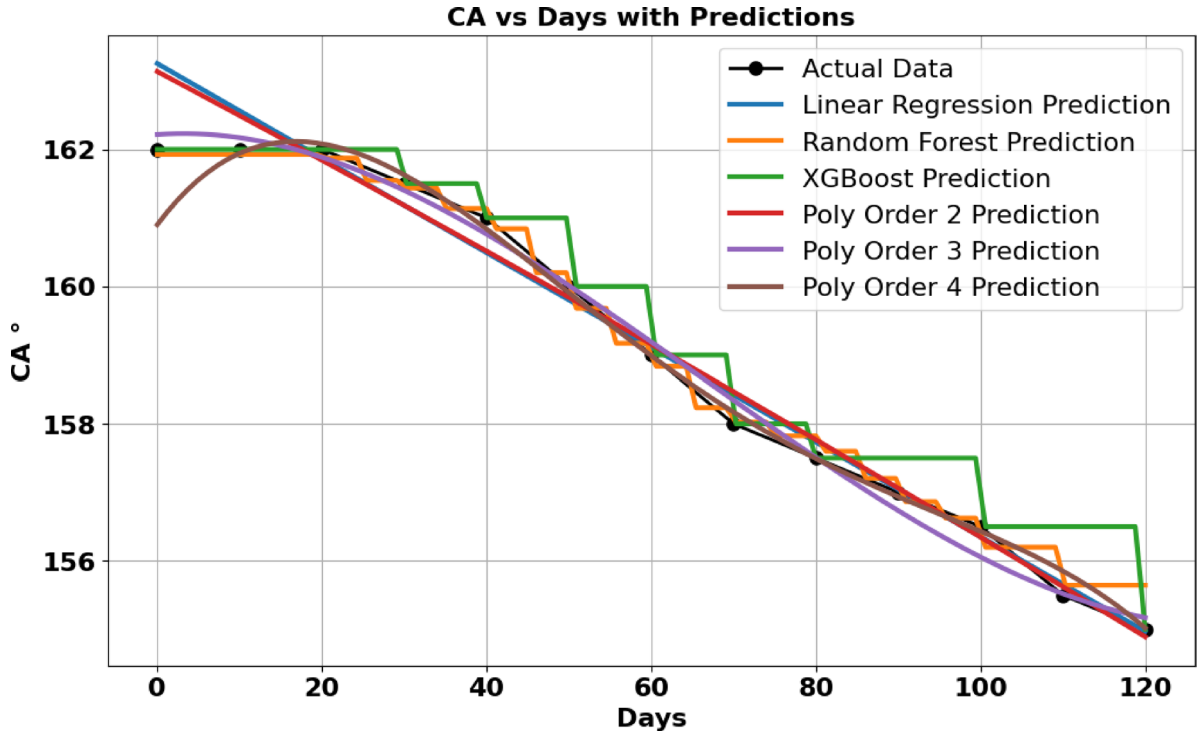

Fig. 11. Distribution of predicted and measured contact angle values from linear regression XGBoost, Random Forest, for Ni-G-MA.

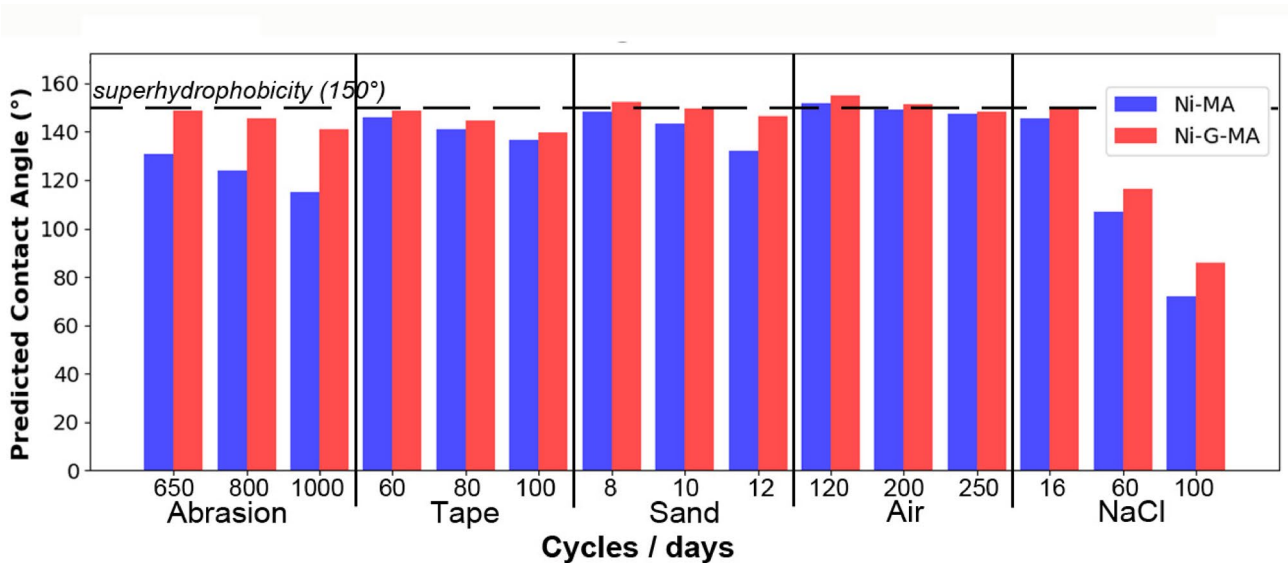

Fig. 12. Predicted contact angle values from linear for Ni-MA and Ni-G-MA.

of 0.2291 ± 0.1916 and R^2 of 0.9333 ± 0.0385. Linear Regression yielded a mean MSE of 0.1855 ± 0.1269 and R^2 of 0.9025 ± 0.0902, while XGBoost showed the lowest performance with an MSE of 0.3667 ± 0.1869 and R^2 of 0.8094 ± 0.1477. Overall, these results clearly indicate that second-order polynomial regression offers the best balance of accuracy and generalizability for this dataset.

Eventually, concerning data related to coatings immersion in different pH and reported in Table S11, S12, all models struggle to predict contact angle (CA) at different pH levels, with high MSE values and mostly negative R^2 scores, likely due to the small dataset failing to capture underlying patterns. Among them, the fourth-order polynomial model (Poly Order 4) performs relatively better, achieving the lowest MSE (0.4118) and the only positive R^2 (0.5882). However, despite its improved fit compared to other models, its predictive accuracy remains suboptimal, suggesting that the dataset lacks sufficient variation for reliable trend identification Figure S4.

Stability prediction on superhydrophobic coatings

In this study, we employed a variety of machine learning (ML) and regression models, including XGBoost, K-Nearest Neighbours (KNN), Random Forest (RF), Support Vector Regression (SVR), and multiple polynomial regression models, to predict the stability of the contact angle (CA) under different environmental and mechanical stress conditions. These conditions included immersion in NaCl solution, varying abrasion cycles, tape peeling tests, sand impact, and open-air exposure. Our analysis revealed that the best-performing model varied depending on the specific degradation mechanism. For abrasion cycles and tape peeling tests, linear regression performed best, as the CA degradation followed a relatively linear trend due to gradual surface wear and adhesive forces. XGBoost also demonstrated strong predictive capability in these cases, benefiting from its ability to detect slight nonlinearities while remaining robust against noise. However, for sand impact and open-air exposure, where the CA degradation followed a nonlinear trend, third-order polynomial regression (cubic regression) outperformed other models. The nature of sand impact resulted in an initial sharp decline in CA followed by stabilization, while open-air exposure led to a time-dependent degradation influenced by oxidation and contamination effects. Higher-order polynomial regression effectively captured these nonlinear trends, whereas simpler models like linear regression and SVR failed to generalize well. Among all models, XGBoost consistently delivered strong performance across multiple conditions, as it could handle both linear and nonlinear relationships while effectively reducing overfitting. In contrast, SVR and KNN exhibited weaker predictive capabilities. SVR struggled due to its sensitivity to hyperparameter tuning, leading to suboptimal performance in highly nonlinear cases such as sand impact and open-air exposure. KNN, being a distance-based algorithm, was less effective at capturing long-term trends and tended to overfit localized variations, making it unsuitable for extrapolating CA degradation over extended periods.

Using the best-performing model in each case, the contact angle (CA) values were predicted under different untested conditions, as shown in Figure. 12. These predicted values closely match the experimental results until available (Figure S5, S6 Table S13), confirming the model reliability in estimating CA under different environmental and mechanical conditions. Moreover, these values allow to draw conclusions on the long-term behavior of these coatings, indicating that Ni-G-MA overperforms Ni-MA. to the presence of graphene not only in initial conditions, but also by preserving, or approaching, superhydrophobicity on a longer run even when subjected to strong mechanical stress. The only clear point of failure of superhydrophobicity for both coatings was for prolonged contact with salty water, which indeed is one of the major concerns in the field of coatings durability.

Based on the predicted contact angle data), NaCl immersion emerged as the most critical degradation factor, especially over prolonged exposure (100 days), where both Ni-MA and Ni-G-MA coatings experienced a substantial drop in contact angle falling below the 150° superhydrophobic threshold. In contrast, air storage, sand impingement, and tape peeling exhibited relatively minor influence, with contact angles consistently remaining above 150°, indicating better stability under those conditions. Notably, abrasion resistance was improved in Ni-G-MA compared to Ni-MA, suggesting that graphene incorporation enhanced mechanical robustness. This analysis highlights chemical durability particularly in corrosive environments as the most significant challenge to maintaining long-term superhydrophobicity, and has been discussed in detail in the revised manuscript.

Linear Regression was chosen as a baseline due to its simplicity and interpretability. Polynomial Regression (orders 2–4) was applied to capture potential non-linear relationships between the input features and the predicted contact angle. Random Forest and XGBoost, both ensemble-based models, were selected for their ability to model complex, non-linear interactions and their robustness to overfitting. These models are widely used in materials science and surface engineering for predictive tasks due to their high accuracy and feature importance analysis capabilities.

Conclusion

This study illustrates the integration of machine learning (ML) with materials science to develop durable, eco-friendly superhydrophobic (SHP) graphene-based coatings for copper. The choice of coatings to be used in this study fell on Ni-graphene composite films modified with myristic acid, which resulted in exceptional hydrophobicity, achieving a contact angle (CA) of up to 162°, while maintaining impressive stability under mechanical, chemical, and environmental stresses. Various ML models, including polynomial regression, XGBoost, Random Forest, Linear Regression, Support Vector Machine (SVM), and k-Nearest Neighbours (KNN), were employed to predict the performance of the coatings. Among these, the polynomial regression (Order 3), XGBoost, and Linear Regression models showed superior performance, closely aligning with experimental data even considering long term performances expressed in abrasion resistance and in exposure to potentially aggressive environments. The data augmentation techniques employed helped reduce overfitting, improving the models' ability to generalize and enhance prediction accuracy. By minimizing the need for extensive experimental testing, ML models provided precise performance forecasting, significantly accelerating the design and analysis process.

Data availability

Data analysed and code is available [https://github.com/himanshuhm1111/Zn-MA/tree/main](https:/github.com/himanshuhm1111/Zn-MA/tree/main).

Received: 5 June 2025; Accepted: 29 August 2025
Published online: 03 October 2025

References

1. Athulya, V., Vanithakumari, S. C., Shankar, A. R. & Ningshen, S. Electrodeposition of myristate based superhydrophobic coatings on steel with enhanced corrosion resistance and self-cleaning property. *Surf. Coat. Technol.* **489**, 131114. https://doi.org/10.1016/j.surfcoat.2024.131114 (2024).
2. Deng, Y. et al. A facile method for constructing scalable and low-cost superhydrophobic coating with anti-corrosion and drag-reduction properties. *Ind. Crops Prod.* **216**, 118732. https://doi.org/10.1016/j.indcrop.2024.118732 (2024).
3. Yin, Z. et al. A multifunctional and environmentally safe superhydrophobic membrane with superior oil/water separation, photocatalytic degradation and anti-biofouling performance. *J. Colloid Interface Sci.* **611**, 93–104. https://doi.org/10.1016/j.jcis.2021.12.070 (2022).
4. Anjum, A. S., Sun, K. C., Ali, M., Riaz, R. & Jeong, S. H. Fabrication of coral-reef structured nano silica for self-cleaning and super-hydrophobic textile applications. *Chem. Eng. J.* **401**, 25859–25859 (2020).
5. Li, M. et al. Facile fabrication of superhydrophobic and photocatalytic self-cleaning flexible strain sensor membrane for human motion. *Sens. Actuators Phys.* **363**, 114750. https://doi.org/10.1016/j.sna.2023.114750 (2023).
6. Deng, Y. et al. Delaying frost formation by controlling surface chemistry of ZnO-coated 304 stainless steel surfaces. *Colloids Surf. Physicochem Eng. Asp* **696**, 134375. https://doi.org/10.1016/j.colsurfa.2024.134375 (2024).
7. Yin, Z. et al. Superhydrophobic Photocatalytic Self-Cleaning Nanocellulose-Based Strain Sensor for Full-Range Human Motion Monitoring. *Adv. Mater. Interfaces* **10**(33), 2300350. https://doi.org/10.1002/admi.202300350 (2023).
8. Abiola, O. K. & Tobun, Y. Cocos nucifera L. water as green corrosion inhibitor for acid corrosion of aluminium in HCl solution. *Chin. Chem. Lett.* **21**, 12, 1449–1452. https://doi.org/10.1016/j.cclet.2010.07.008 (2010).
9. Cho, E. C. et al. Robust multifunctional superhydrophobic coatings with enhanced water/oil separation, self-cleaning, anti-corrosion, and anti-biological adhesion. *Chem. Eng. J.* **314**, 347–357 (2017).
10. Zha, Q. et al. Facile construction of multifunctional 3D smart MOF-based polyurethane sponges with photocatalytic ability for efficient separation of oil-in-water emulsions and co-existing organic pollutant. *Chem. Eng. J.* **490**, 151747. https://doi.org/10.1016/j.cej.2024.151747 (2024).
11. Yang, G. et al. Robust mussel-inspired LBL carbon nanotube-based superhydrophobic polyurethane sponge for efficient oil–water separation utilizing photothermal effect. *Fuel* **381**, 133353. https://doi.org/10.1016/j.fuel.2024.133353 (2025).
12. Siddiqui, A. R., Maurya, R., Katiyar, P. K. & Balani, K. Superhydrophobic, self-cleaning carbon nanofiber CVD coating for corrosion protection of AISI 1020 steel and AZ31 magnesium alloys. *Surf. Coat. Technol.* **404**, 26421–26421 (2020).
13. Albayrak, S. et al. The Investigation of Hybrid and Layered Ha/Ta2o5 Sol-Gel Composite Coating on Az91 Mg Alloy, *SSRN Electron. J.* (2023). https://api.semanticscholar.org/CorpusID:256654717
14. Jameei Rad, P., Aliofkhazraei, M. & Darband, G. B. Ni-W nanostructure well-marked by Ni selective etching for enhanced hydrogen evolution reaction. *Int. J. Hydrog Energy*. **44**, 2, 880–894. https://doi.org/10.1016/j.ijhydene.2018.11.026 (2019).
15. Elias, L. & Chitharanjan Hegde, A. Electrodeposition of laminar coatings of Ni–W alloy and their corrosion behaviour. *Surf. Coat. Technol.* **283**, 61–69. https://doi.org/10.1016/j.surfcoat.2015.10.025 (2015).
16. Cao, J. et al. In-situ fabrication of superhydrophobic surface on copper with excellent anti-icing and anti-corrosion properties. *Mater. Today Commun.* **33**, 104633. https://doi.org/10.1016/j.mtcomm.2022.104633 (2022).
17. Li, H., Sun, Y., Wang, Z. & Wang, S. Constructing Superhydrophobic Surface on Copper Substrate with Dealloying-Forming and Solution-Immersion Method. *Materials* **15**(14), 4816. https://doi.org/10.3390/ma15144816 (2022).
18. Han, J. et al. A smart electroplating approach to fabricate mechanically robust and fluorine-free Ni-W alloys based superhydrophobic coating on al alloy. *Vacuum* **217**, 112501. https://doi.org/10.1016/j.vacuum.2023.112501 (2023).
19. Qiao, M., Ji, G., Lu, Y. & Zhang, B. Sustainable corrosion-resistant superhydrophobic composite coating with strengthened robustness. *J. Ind. Eng. Chem.* **121**, 215–227. https://doi.org/10.1016/j.jiec.2023.01.025 (2023).
20. Mahdi. Ebrahimi, A., Bayat, S. R., Ardekani, E. S., Iranizad, & Moshfegh, A. Z. Sustainable superhydrophobic branched hierarchical ZnO nanowires: Stability and wettability phase diagram. *Appl. Surf. Sci.* **561**, 150068. https://doi.org/10.1016/j.apsusc.2021.150068 (2021).
21. Prasad Mamgain, H., Pal, R., Kumar, S., Brajpuriya, R. & Pandey, J. K. Machine Learning-Based stability prediction and analysis of polypropylene Cu-MA superhydrophobic coating on the aluminum substrate. *J. Phys. Chem. C.* **128** (40), 17184–17195. https://doi.org/10.1021/acs.jpcc.4c05934 (2024).
22. Wang, Q., Dumond, J. J., Teo, J. & Low, H. Y. Superhydrophobic polymer topography design assisted by machine learning algorithms. *ACS Appl. Mater. Interfaces*. **13**, 25, 30155–30164. https://doi.org/10.1021/acsami.1c04473 (2021).
23. Usman, J. et al. Design and machine learning prediction of in situ grown PDA-Stabilized MOF (UiO-66-NH2) membrane for Low-Pressure separation of emulsified oily wastewater. *ACS Appl. Mater. Interfaces*. **16**, 13, 16271–16289. https://doi.org/10.1021/acsami.4c00752 (2024).
24. Elhady, S., Zaki, E. G., El-Azabawy, O. E. & Fahim, I. S. Electrochemical evaluation of green corrosion inhibitor based on ground coffee waste in Petroleum fields. *Results Eng.* **21**, 101880. https://doi.org/10.1016/j.rineng.2024.101880 (2024).
25. Li, M. et al. Myristic acid-tetradecanol-capric acid ternary eutectic/SiO2/MIL-100(Fe) as phase change humidity control material for indoor temperature and humidity control. *J. Energy Storage* **74**, 109437. https://doi.org/10.1016/j.est.2023.109437 (2023).
26. Fawagreh, K., Gaber, M. M. & Elyan, E. Random forests: from early developments to recent advancements. *Syst. Sci. Control Eng.* **2**, 1. https://doi.org/10.1080/21642583.2014.956265 (2014).
27. Tafarroj, M. M., Mousavi Ajarostaghi, S. S., Ho, C. J. & Yan, W. M. Artificial neural network approaches for predicting the heat transfer in a Mini-Channel heatsink with alumina/water nanofluid. *J. Heat. Mass. Transf. Res.* **11** (1), 75–88. https://doi.org/10.22075/jhmtr.2024.32947.1520 (2024).
28. Chen, Y., Xiao, C., Yang, S., Yang, Y. & Wang, W. Research on long term power load grey combination forecasting based on fuzzy support vector machine. *Comput. Electr. Eng.* **116**, 109205. https://doi.org/10.1016/j.compeleceng.2024.109205 (2024).
29. John, V., Liu, Z., Guo, C., Mita, S. & Kidono, K. Real-Time lane Estimation using deep features and extra trees regression. In *Image and Video Technology* (eds Bräunl, T. et al.) 721–733 (Springer International Publishing, 2016).
30. Cuong-Le, T. et al. An efficient approach for damage identification based on improved machine learning using PSO-SVM. *Eng. Comput.* **38** (4), 3069–3084. https://doi.org/10.1007/s00366-021-01299-6 (2022).
31. Paszkowicz, W. Genetic algorithms, a Nature-Inspired tool: survey of applications in materials science and related fields. *Mater. Manuf. Process.* **24**, 2, 174–197. https://doi.org/10.1080/10426910802612270 (2009).
32. Barai, B. et al. Enhancing corrosion resistance of biodegradable magnesium alloys through hydrophobic surface modification: experimental analysis and ANN modeling. *Surf. Rev. Lett.* https://doi.org/10.1142/S0218625X24500938 (2024).
33. Ragheb, D. M., Abdel-Gaber, A. M., Mahgoub, F. M. & Mohamed, M. E. Eco-friendly method for construction of superhydrophobic graphene-based coating on copper substrate and its corrosion resistance performance. *Sci. Rep.* **12**(1), 17929. https://doi.org/10.1038/s41598-022-22915-5 (2022).
34. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '16. Association for Computing Machinery, 785–794. (New York, NY, USA, 2016). https://doi.org/10.1145/2939672.2939785
35. Kazemi, M. M. K., Nabavi, Z. & Armaghani, D. J. A novel hybrid XGBoost methodology in predicting penetration rate of rotary based on Rock-Mass and material properties. *Arab. J. Sci. Eng.* **49** (4), 5225–5241. https://doi.org/10.1007/s13369-023-08360-0 (2024).

36. Swischuk, R., Mainini, L., Peherstorfer, B. & Willcox, K. Projection-based model reduction: formulations for physics-based machine learning. *Comput. Fluids*. **179**, 704–717. https://doi.org/10.1016/j.compfluid.2018.07.021 (2019).
37. Tran, N. K., Kühle, L. C. & Klau, G. W. A critical review of multi-output support vector regression. *Pattern Recognit. Lett.* **178**, 69–75. https://doi.org/10.1016/j.patrec.2023.12.007 (2024).
39. Zhang, C. et al. Enhancing state of charge and state of energy estimation in Lithium-ion batteries based on a TimesNet model with Gaussian data augmentation and error correction. *Appl. Energy* **359**, 122669. https://doi.org/10.1016/j.apenergy.2024.122669 (2024).
38. Zhang, C. et al. Enhancing state of charge and state of energy estimation in Lithium-ion batteries based on a TimesNet model with Gaussian data augmentation and error correction. *Appl. Energy* **3589**, 129. (2021).

Acknowledgements
We express our heartfelt gratitude to the University of Petroleum and Energy Studies (UPES), Research and Development department UPES, HILL UPES for their support and guidance.

Author contributions
H.P.M. conceptualized the study and performed the machine learning modeling and data analysis. M.V.D. supervised on the ML P.R.P. contributed to materials synthesis and experimental validation. M.E.M. assisted in mechanical durability testing and corrosion analysis. J.K.P. and M.K. co-supervised the entire study and contributed to the interpretation of results. N.B. supported data curation and visualization. A.V. assisted in model tuning and regression validation. H.P.M. wrote the main manuscript text.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-18155-y.

Correspondence and requests for materials should be addressed to H.P.M., J.K.P. or M.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.