Strength Prediction in UHPC with XGBoost Model and Shapley Algorithm Interpretation

Guanzhong Wu[1,2], Hairui Gou[2], Qingzhao Ren[2], Ran Tang[3,a,*], Peng Feng[3,b,*]

[1]State key laboratory of Geohazard Prevention and Geoenvironment Protection, Chengdu University of Technology, Chengdu, China

[2]China Railway Cultural Heritage rehabilitation Technology Innovation Co., Ltd., Chengdu, Sichuan, China;

[3]School of Architecture and Civil Engineering, Chengdu University, Chengdu, Sichuan, China.

*Corresponding author:[a] 546488149@qq.com; [b] fengpengfx@163.com

Abstract. This study employs the XGBoost regression model to predict the strength of UHPC and utilizes the Shapley algorithm to interpret the model's predictions, revealing the impact of various feature parameters. The results demonstrate that the XGBoost regression model effectively fits the data and possesses strong predictive capabilities. Furthermore, the interaction between silica fume and cement significantly influences the model predictions. Additionally, using tools such as the Shapley heatmap, the study analyzes the model's characteristics and finds that only a subset of samples have Shapley values below the mean, indicating the dataset contains relatively few high-quality samples. Through the Shapley algorithm, the optimal range for silica fume quantity is determined to be between 0 and 320 kg. This research validates the effectiveness of the XGBoost regression model for predicting UHPC strength and enhances model interpretability using the Shapley algorithm.

Keywords: UHPC；Machine learning；XGBoost；Strength prediction；Shapley algorithm

1. Introduction

Ultra-High Performance Concrete (UHPC) represents a novel cementing material achieving significant performance advancements over traditional concrete, facilitated by meticulous design and utilization of high-performance constituents. For the determination of UHPC strength, conventional laboratory methods often necessitate resources for experimentation and data analysis, whereas machine learning algorithms offer distinct advantages. Machine learning models can forecast the performance metrics of UHPC, thereby reducing the experimental costs and time investments. Many scholars have predicted the mechanical properties of various concretes via different machine learning algorithms. Mohamed Abdellatief et al. [1] employed Random Forest, Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost) to predict the compressive strength of Ultra-High Performance Geopolymer Concrete. Dinesh et al. [2] revealed the primary factors on the compressive strength of UHPGC using machine learning models, optimizing formulations for green building materials, and concluded that XGBoost, Support Vector Machine (SVM) and SVR are proficient in predicting shear strength. Huang et al. [3] proposed a corrosion-resistant steel-reinforced concrete strength prediction model based on SHapley Additive exPlanations (SHAP) values, integrating several ML algorithms with experimental data from 166 interfaces of corroded steel-reinforced concrete. Yuan et al. [4] explored self-healing capabilities of engineered cementing composites using various machine learning models and employed the SHAP algorithm to analyze the impact of input parameters on output parameters. Khan et al. [5] utilized the machine learning algorithms to identify the primary factors on predicting the compressive strength of Reactive Powder Concrete, demonstrating the effectiveness of multi-layer stacking models in strength prediction. Pakzad et al. [6] predicted the splitting tensile strength of Steel Fiber Reinforced Concrete (SFRC) using multiple machine learning models, highlighting Support Vector Regression as the optimal performer.

So far, the effectiveness of XGBoost for predicting the UHPC strength has not been demonstrated. This study utilize the XGBoost algorithm for strength prediction in UHPC, and interpreted the results via SHapley Additive exPlanations, which have never been studied.

2. Methodlogy

2.1 XGBoost algorithm

The Extreme Gradient Boosting (XGBoost) algorithm primarily optimizes the Gradient Boosting Decision Trees, the fundamental principles of which can be illustrated in Fig. 1a. The objective function of XGBoost encompasses loss functions and regularization terms, aiming to enhance the predictive accuracy and prevent overfitting[7]. Employing classification and regression trees as the basic learners, the algorithm constructs decision trees by minimizing the objective function, where the algorithm selects the optimal feature for partitioning for each node within XGBoost.

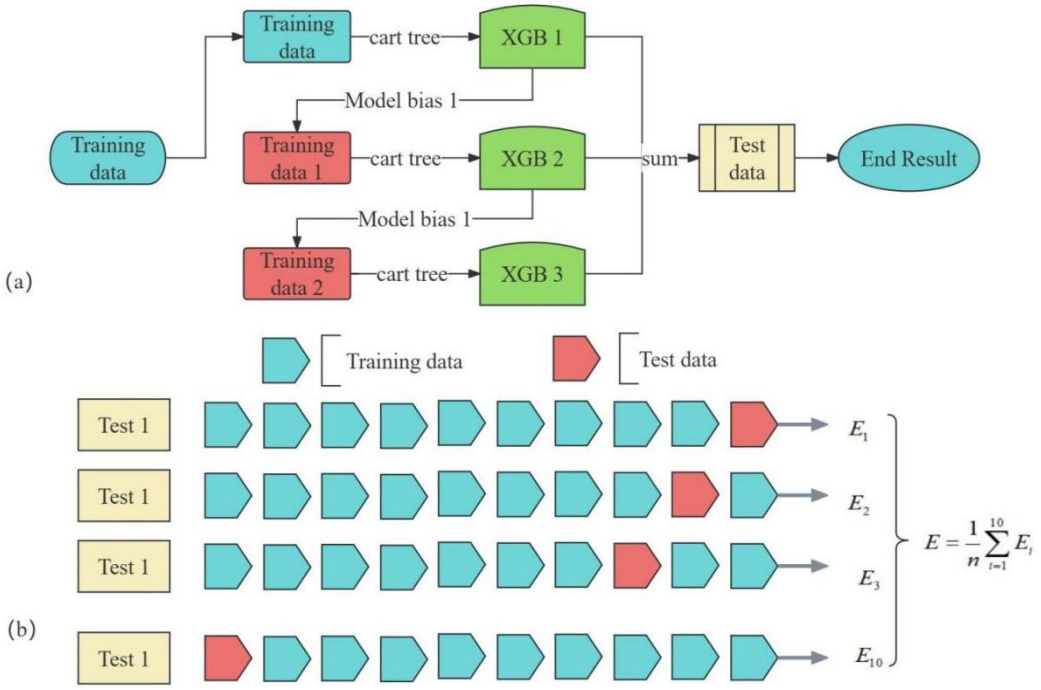

Fig. 1. Schematic diagram of (a) the XGBoost algorithm ,and (b) 10-fold cross-validation

2.2 Cross-validation

Cross-validation is widely employed as an evaluation technique in the field of machine learning, aimed at assessing the generalization ability of the tested models[8]. Ten-fold cross-validation (k = 10) is regarded as an effective means to evaluate model generalization ability due to its unique data partitioning approach[9]. Within each fold, the randomness of data ensures variability in the model performance, reflecting the behaviors across different data distributions. Averaging the results from all k validations can provide a comprehensive and balanced assessment of the model performance. The adoption of repeated validation enhances the resolution in evaluation outcomes, with each validation being independent on distinct data subsets, thereby potentially revealing performance fluctuations, as depicted in Figure 1b.

2.3 Model assessment

Model evaluation metrics serve as critical standards for assessing the model performance including accuracy, precision, and recall. Additionally, the balance and consistency between these metrics

must be considered to ensure the reliability and accuracy of evaluation results. In this study, model training effectiveness is evaluated using four evaluation metrics involving coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE)[10].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}; RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}\ ; MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|\ ; MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

where y_i and $\hat{y}_i$ represent the prediction value and true value of the model, respectively.

2.4 SHAP algorithm

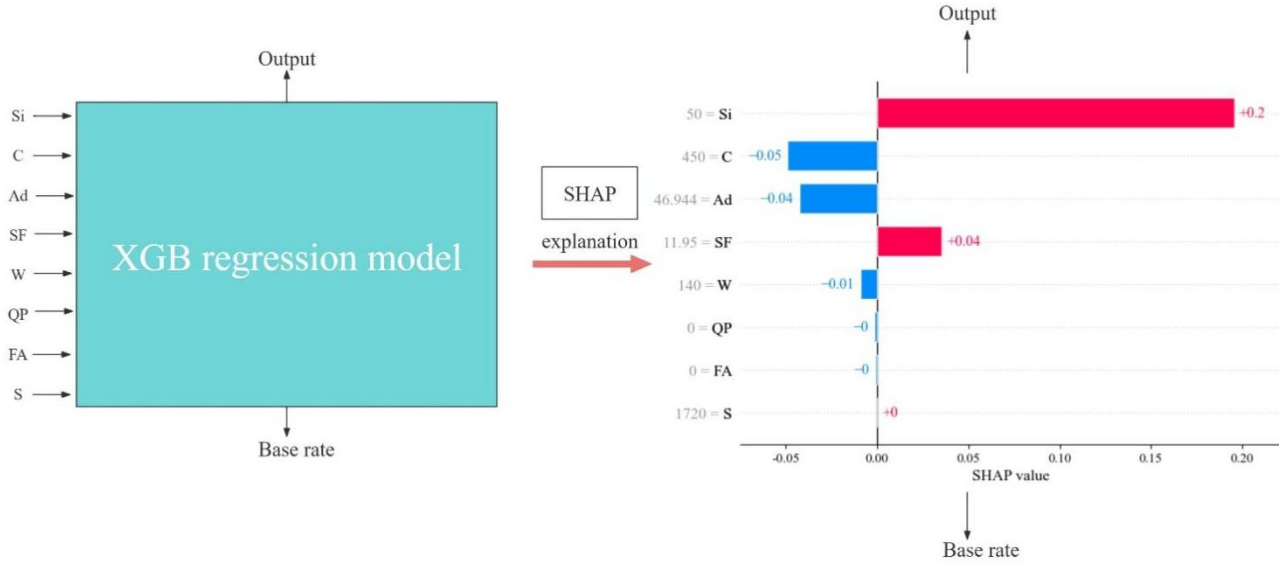

Fig. 2 Shapley Additive exPlanations of the XGBoost model

Shapley Additive exPlanations (SHAP) is a Python-based model interpretation toolkit[11]. The fundamental principle derives from the contributions of input features, calculating all feature combinations and the marginal contributions of each feature within these combinations to determine the Shapley value of each feature. The general procedure of the Shapley algorithm involves initially computing Shapley values to assess the importance of each feature within the sample, measuring their impact on model training. As depicted in Figure 2, the XGB regression model used in this study utilizes SHAP explanations to visually display the Shapley values of each feature, thereby quantifying their impact on model training more intuitively.

3. Data collection and processing

The datasets in this study originates from Abuodeh et al. [12] encompassing eight input parameters: cement (C), silica fume (Si), fly ash (FA), sand (S), steel fiber (SF), quartz powder (QP), water (W), and admixture (Ad), along with one output parameter, UHPC compressive strength (fc). Bootstrap resampling was employed to expand the datasets from 110 original samples to 1099 samples. Figure 3a displays the violin plot of the expanded datasets. The violin plot illustrates the dataset after expansion, where wider sections indicate higher probability of observed values, while narrower sections correspond to lower probabilities. Fig. 3b presents a heatmap of feature correlations, where the sizes of pie and square shapes in the heatmap illustrate the magnitude of correlation coefficients more intuitively than color intensity.

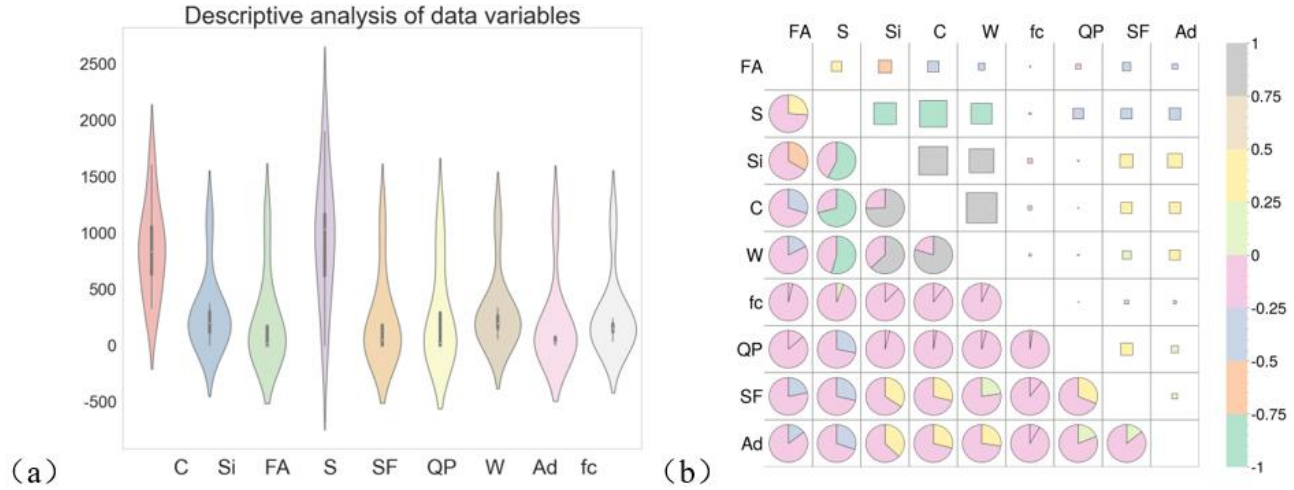

Fig. 3 (a) The violin plot of the expanded datasets, and (b) the violin plot of the expanded datasets

4. Results

4.1 The prediction of LightGBM regression model

Figure 4(a) shows the the prediction of LightGBM regression model. Red points represent instances tested on the training set, while the blue sample points denote predictions. It is evident that these points are distributed closely around the diagonal line, indicating minimal deviation between the training and predicted values. Moreover, the range of predicted values for both training and testing sets is comparable, suggesting consistent predictive performance across different datasets. Figure 4(b) displays the model evaluation metrics of the predicted results, where the XGBoost regression model exhibits superior performance across all metrics, boasting a high coefficient of determination (R^2 = 0.966), low root mean square error (RMSE = 0.036), minimal mean absolute error (MAE = 0.011), and negligible mean absolute percentage error (MAPE = 0.064).

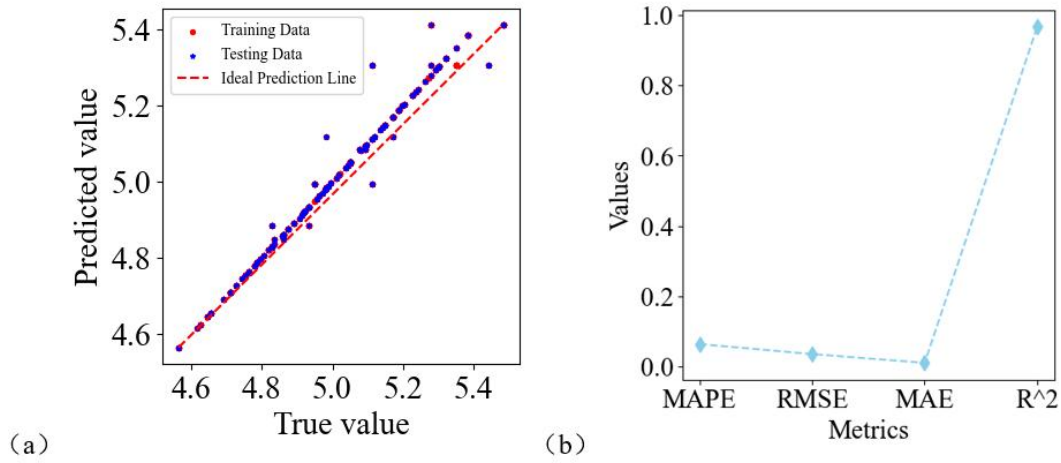

Fig. 4 (a) The fitting results of the XGBoost regression model prediction, and (b) the model evaluation metrics.

4.2 Model evaluation based on Shapley interpretation algorithm

For the XGBoost regression model, SHAP analysis provides a more intuitive depiction of each feature during the prediction process. Figure 5(a) showcases the global feature importance plot, emphasizing the crucial significance of silica fume and cement dosage in training the XGBoost

regression model. This underscores their substantial impact on predicting outcomes within the UHPC mixture. In Figure 5(b), the importance distribution plot reveals that silica fume dosage positively influences the training effectiveness more significantly than cement dosage, which exhibits a slightly negative effect, aligning with the linear relationship of cement in the UHPC manufacturing process. Figure 5(c) displays the interaction SHAP summary plot, demonstrating varying impacts on the XGBoost regression model under different feature interactions, with notable significance observed in interactions involving cement. Figure 5(d) illustrates the single sample decision plots, offering visual insights into the extent of feature impacts and variations in model predictions. Figure 5(e) presents the multi-sample decision plot, where each line converges at the predicted value of the corresponding observation on the x-axis, illustrating each feature's contribution to overall predictions. The multi-sample decision plots, akin to heatmaps, convert logarithmic odds into probabilities, converging observations at the bottom to 1.0. Figure 5(f) depicts the SHAP summary heatmap, highlighting that among the dataset, only the last 50 samples have SHAP values summed for f(x) below the average line, suggesting they are not optimal samples. The XGBoost regression model identifies the optimal range for silica fume quantity as 0 to 320 kg.

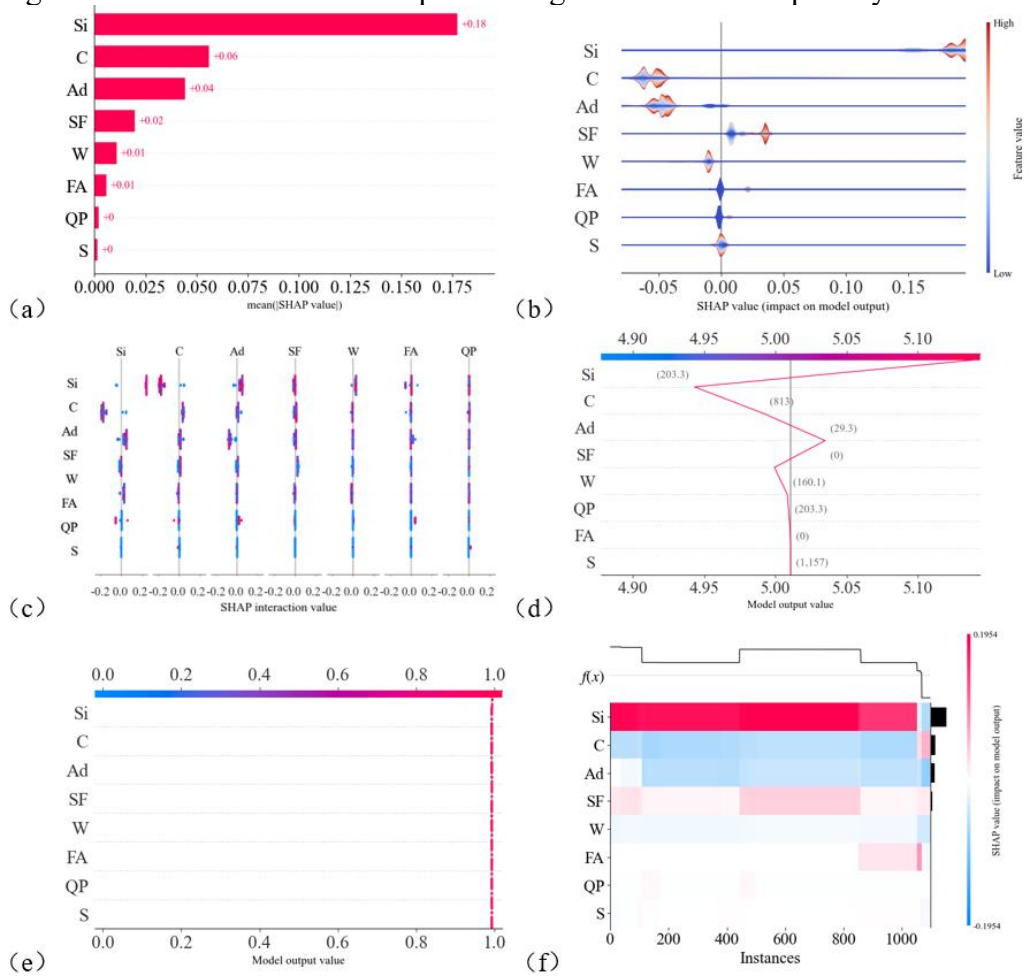

Fig. 5 (a) Feature importance diagram, (b) global feature importance diagram, (c) interactive SHAP-summary diagram, (d) single sample decision diagram, (e) multiple eigenvalues decision diagram, and (f) shapley heat map

Summary

In this study, a datasets comprising 110 samples of UHPC was collected and expanded to 1099 samples using Bootstrap resampling. Evaluation of four performance metrics across three models indicates good fitting to the samples. Through comprehensive SHAP analysis of the XGBoost regression model, it was observed that the quantity of silica fume positively influences model training effectiveness to a greater extent, whereas cement exhibits a stronger negative impact. The interaction effects between silica fume and cement are most pronounced in SHAP analysis, with

silica fume and cement exhibiting the highest single-feature importance levels for the model. The SHAP algorithm effectively reconstructs the decision process of the model predicting the compressive strength of UHPC, identifying optimal sample selections to elucidate the optimal ranges of feature parameters.

Acknowledgements

This work was supported in part by the Research projects of C.R.E.C (2020-KJ001-Z001-A1; 2020-YD-302; 2021-key point-34, 2022-key point-01, 2023-major-02).

References

[1] Mohamed Abdellatief, Youssef M. Hassan, Mohamed T. Elnabwy, Leong Sing Wong, Ren Jie Chin, Kim Hung Mo. Investigation of machine learning models in predicting compressive strength for ultra-high-performance geopolymer concrete: A comparative study[J]. Construction and Building Materials, 2024,436:136884.

[2] Dinesh, Rahul Prasad. Predictive models in machine learning for strength and life cycle assessment of concrete structures[J]. Automation in Construction, 2024,162:105412.

[3] Tao Huang, Tingbin Liu, Yan Ai, et al. Modelling the interface bond strength of corroded reinforced concrete using hybrid machine learning algorithms[J]. Journal of Building Engineering, 2023,74:106862.

[4] Xiongzhou Yuan, Qingyu Cao, Nasir Amin M, et al. Predicting the crack width of the engineered cementitious materials via standard machine learning algorithms[J]. Journal of Materials Research and Technology, 2023,24:6187-6200.

[5] M. Iqbal Khan, Yassir M. Abbas. Intelligent data-driven compressive strength prediction and optimization of reactive powder concrete using multiple ensemble-based machine learning approach[J]. Construction and Building Materials, 2023,404:133148.

[6] Seyed Soroush Pakzad, Mansour Ghalehnovi, Atiye Ganjifa. A comprehensive comparison of various machine learning algorithms used for predicting the splitting tensile strength of steel fiber-reinforced concrete[J]. Case Studies in Construction Materials, 2024,20:e03092.

[7] Resul Özdemir., Murat Taşyürek and Veysel Aslantaş, Improved Marine Predators Algorithm and Extreme Gradient Boosting (XGBoost) for shipment status time prediction. Knowledge-Based Systems, 2024. 294: p. 111775.

[8] Yuanchao Xu, Xiaopeng Kong, Zhiming Cai, Cross-validation strategy for performance evaluation of machine learning algorithms in underwater acoustic target recognition. Ocean Engineering, 2024. 299: p. 117236.

[9] Seyed Matin Malakouti., Babysitting hyperparameter optimization and 10-fold-cross-validation to enhance the performance of ML methods in predicting wind speed and energy generation. Intelligent Systems with Applications, 2023. 19: p. 200248.

[10] Ashkan Safari., Hybrid emerging model predictive data-driven forecasting of three-phase electrical signals of photovoltaic systems using GBR-LSTM. e-Prime - Advances in Electrical Engineering, Electronics and Energy, 2024. 8: p. 100549.

[11] Changlan Yang, Xuefeng Guan, Qingyang Xu, Weiran Xing, Xiaoyu Chen, Jinguo Chen, Peng Jia et al., How can SHAP (SHapley Additive exPlanations) interpretations improve deep learning based urban cellular automata model? Computers, Environment and Urban Systems, 2024. 111: p. 102133.

[12] Abuodeh O R, Abdalla J A, Hawileh R A. Assessment of compressive strength of Ultra-high Performance Concrete using deep machine learning techniques[J]. Applied Soft Computing, 2020,95:106552.