International Core Journal of Engineering Volume 12 Issue 6, 2026
ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

Prediction and Interpretability Analysis of Deep Eutectic Solvent Viscosity based on a Stacking Ensemble Model

Zhen Yang, Jin Shu, Yichi Zhang, Xingchi Deng, and Shengtao He

North China University of Science and Technology, Tangshan 063000, China

Abstract

Deep eutectic solvents (DES) are a new type of green solvent with significant application potential in the selective leaching of zinc-containing solid waste via hydrometallurgy. However, traditional design methods rely heavily on trial-and-error experimentation, resulting in low R&D efficiency, and the core physical property (viscosity) that determines mass transfer and reaction kinetics in the system is difficult to estimate accurately. To address this industry challenge, this paper proposes a high-precision viscosity prediction and micro-mechanism analysis framework that integrates multi-dimensional cross-feature engineering, the Whale Optimization Algorithm (WOA), and Stacking ensemble learning. Based on a rigorously cleaned, multi-source standardized dataset, the model employs Extreme Gradient Boosting (XGBoost) and Random Forest (RF)-both globally optimized via WOA-as base learners, combined with a Linear Regression (LR) meta-learner through a two-stage deep integration. Extrapolation evaluation on an independent test set demonstrates that this two-layer architecture effectively overcomes the generalization bottleneck in complex sample chemical spaces, achieving an excellent accuracy with a coefficient of determination (R^2) of 0.8620 and an average absolute relative deviation (AARD) as low as 9.88%, with overall performance significantly outperforming various single-baseline models.Furthermore, this study introduced game-theoretic SHAP analysis, successfully breaking through the "black-box" barrier of deep ensemble models. The research quantitatively confirmed that the drastic changes in DES viscosity essentially stem from strong electrostatic coupling and a dense hydrogen-bond cross-linking network formed after component mixing.The positive synergistic interaction between high steric hindrance in the mixed space and strong electrostatic attraction constitutes the underlying mechanism driving the macroscopic viscosity jump, while the "coupled resonance" of multiple microscopic features at low temperatures serves as the fundamental driving force behind the sharp rise in hydrodynamic viscosity.This study not only provides a breakthrough data-driven paradigm for estimating the physical properties of complex chemical systems, but also offers solid quantitative theoretical guidance for the targeted reverse design of low-viscosity, high-performance green DES solvents for industrial zinc extraction through a multidimensional, mechanism-transparent analysis.

Keywords

Deep Eutectic Solvents (DESs); Viscosity Prediction; Stacking Ensemble Learning; Whale Optimization Algorithm (WOA); SHAP Interpretability Analysis; Hydrometallurgy.

1. Introduction

Zinc is an indispensable strategic base metal for China's national economy and modern manufacturing industry.With the increasing depletion of high-grade primary zinc ores, the recovery and recycling of zinc from bulk zinc-containing industrial solid wastes, such as steelmaking fumes,

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

has become a key pathway to ensuring national resource security and achieving a green, low-carbon industrial transition. However, traditional hydrometallurgy often employs strong acid leaching or ammonia leaching processes, which face technical bottlenecks such as extremely poor selectivity due to the co-dissolution of impurities (e.g., iron, aluminum, etc.), severe secondary pollution, and equipment corrosion.In recent years, deep eutectic solvents (DES)-green ionic liquids formed by the hydrogen-bonded complexation of hydrogen-bond donors (HBDs) and hydrogen-bond acceptors (HBAs)-have garnered significant attention in the field of hydrometallurgical separation. This is due to their advantages of simple preparation, low toxicity, and biodegradability, as well as their extremely high selective dissolution capacity for metal oxides such as ZnO.and have been hailed as the third-generation green solvents [1][2][3].

Table 1. Timeline of DES Physical Property Research and Machine Learning Applications

Time Period	Representative Research Milestones and Key Literature	Core Methods and Scientific Significance	Limitations/Shortcomings
2002–2004	Abbott et al. [6] first proposed the concept of a choline chloride-based eutectic solvent and its application framework.	They established the fundamental experimental framework for DES, systematically documented and emphasized the decisive role of thermophysical properties such as viscosity and electrical conductivity in applications.	Highly reliant on experimental trial and error: no data-driven prediction methods, extremely long R&D cycles, high trial-and-error costs, and inability to pre-evaluate the viscosity of unsynthesized systems.
2018–2020	Methods for estimating physical properties based on the Group Contribution (GC) method and theoretical calculations began to gain widespread adoption.	Combined with quantum chemical descriptors such as COSMO-RS, molecular characteristics of DES were preliminarily quantified.	Extremely poor nonlinear fitting capability: Traditional mathematical or physical models struggle to accurately capture the highly nonlinear relationships resulting from strong hydrogen-bond networks and electrostatic interactions, leading to significant prediction errors.
2021–2022	Mu et al. [4] (*PCCP*, 2022) introduced algorithms such as XGBoost to predict DES viscosity.	It was demonstrated that traditional tree-based models, such as XGBoost, can achieve high-precision predictions of physical properties with only a small amount of experimental data.	Single-model generalization bottleneck: Relying solely on a single tree model (such as a standalone XGBoost) makes it extremely prone to local optima and overfitting in small chemical datasets, and lacks a global intelligent optimization algorithm for hyperparameters.
2023–2024	Mohan et al. [5] (*JCTC*, 2024) conducted large-scale prediction studies on DES viscosity using models such as CatBoost.	Traditional single-layer ensemble machine learning has become the mainstream approach for predicting physical properties in chemical space, effectively enhancing the screening capabilities for high-dimensional chemical components.	Lack of deep model fusion and physical decomposition: Research remains limited to single-layer ensemble learning, lacking a fusion architecture like Stacking that leverages the strengths of multiple models; furthermore, in-depth SHAP visualization analysis of hydrogen-bond-electrostatic coupling mechanisms for individual samples remains insufficient.

Although DES hold great potential for the selective leaching of zinc-containing solid waste, their macroscopic physical properties-particularly viscosity-directly determine mass transfer efficiency,

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

reaction kinetics, and the feasibility of industrial scale-up during the leaching process. Excessively high viscosity significantly increases internal fluid friction, leading to a sharp decline in leaching efficiency.Currently, the design of DES relies heavily on the traditional "trial- -and-error" approach. However, the chemical space formed by theoretically feasible combinations of HBA and HBD is extremely vast. Traditional experimental screening is time-consuming and costly, and it is difficult to fundamentally elucidate the structure-property relationship between microscopic molecular interactions and macroscopic viscosity.

With the convergence of computational chemistry and artificial intelligence, data-driven machine learning (ML) methods have provided a revolutionary approach to the prediction and directed design of the physical properties of DES.In recent years, scholars both domestically and internationally have conducted extensive exploratory research on predicting the macroscopic properties of DES using traditional machine learning algorithms. For example, Mu et al. (2022) [4] utilized single algorithms such as Random Forest (RF), Support Vector Regression (SVR), and XGBoost, combined with the Morgan molecular fingerprint, to construct a viscosity prediction model for DES, confirming the advantages of tree-based ensemble algorithms in terms of prediction accuracy;Subsequently, Mohan et al. (2024) [5] developed a viscosity prediction model based on the CatBoost algorithm using a larger-scale database, further expanding the applicability of machine learning under multi-temperature and multi-molar ratio conditions. Table 1 briefly summarizes the timeline of DES property research and the evolution of traditional machine learning applications in recent years [4-6].

However, a review of current cutting-edge research reveals that when applied to complex chemical systems, these approaches generally face the following two common shortcomings and bottlenecks:

First, the limitations of single-model architectures and generalization bottlenecks. Most existing studies rely on a single estimator algorithm (such as XGBoost or a single RF). Since obtaining precise quantum chemical calculation data is extremely costly, DES datasets typically represent a chemical space characterized by "small sample sizes, nonlinearity, and high noise."Given these high-dimensional characteristics, a single-model approach is highly prone to getting stuck in local optima and faces a severe risk of overfitting, making it difficult to achieve stable, highly robust extrapolation predictions across different combinations of hydrogen bond donors and acceptors.

Second, the lack of physical meaning in features and limitations in deep interpretability. Many existing models over-rely on graph theory topology or pure molecular fingerprints as inputs, lacking quantitative characterizations of the microscopic charge density and hydrogen bond network strength that determine the essence of DES. Although some studies have made preliminary attempts to introduce interpretability methods, most remain at the stage of simply ranking global feature weights.They lack in-depth visual analysis of the intrinsic interaction mechanisms within individual samples and are unable to deeply link machine learning results with underlying quantum chemical mechanisms. Consequently, they cannot directly provide clear reverse-engineering guidance for the "targeted modification of low-viscosity solvents" in hydrometallurgy.

To address the aforementioned research limitations and industry pain points, this paper innovatively proposes a novel research framework titled "Precise Prediction and Interpretability Analysis of Eutectic Solvent Viscosity Based on Stacking Machine Learning Models." Taking a binary DES system as the subject, we first construct a macroscopic viscosity experimental database and a microscopic molecular property dataset, followed by data cleaning and feature engineering;Second, to overcome the limitations of single models, the Whale Optimization Algorithm (WOA) is introduced to perform global intelligent hyperparameter tuning for Random Forests (RF) and XGBoost. These are then used as base learners to construct a two-layer Stacking ensemble model, aiming to leverage meta-learners to integrate the advantages of multiple models and thoroughly overcome the fitting and generalization bottlenecks associated with small-sample chemical datasets;Finally, to address the lack of in-depth mechanism analysis, we introduce SHAP (SHapley Additive Explanations) visualization technology to decompose the physicochemical mechanisms underlying viscosity changes. This study aims to provide quantitative theoretical guidance and a

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

breakthrough model paradigm for the targeted design of low-viscosity, high-performance DES systems for efficient zinc extraction through the synergistic integration of ensemble learning and visual mechanism analysis.

2. Data and Feature Engineering

2.1 Acquisition and Analysis of Raw Datasets

High-quality data with high physicochemical fidelity is the cornerstone for building robust machine learning models. The underlying data is primarily composed of two integrated components: a macroscopic experimental viscosity dataset and a microscopic molecular descriptor dataset.

2.1.1 DES Macroscopic Viscosity Experimental Dataset

This study constructed a comprehensive and highly representative macroscopic experimental database of binary low-melting-point solvents (DES) [7]. This dataset systematically integrates a large volume of rigorously experimentally validated viscosity data for binary systems, comprising a total of 5,790 viscosity test results for HBA-HBD pairs, providing a reliable data benchmark for establishing high-precision machine learning prediction models.

In terms of variable definitions, the dataset identifies three core fundamental variables that govern macroscopic viscosity: First, the component ratio, quantified by molar fractions (X_1 andX_2). This metric directly reflects the packing density of micromolecules within the system and the saturation of the hydrogen-bond cross-linking network;Second is thermodynamic temperature (T), with data primarily distributed between 278 K and 378 K, effectively covering the mainstream process temperature range from ambient-temperature operations to high-temperature metallurgical leaching; third is the physical property target, specifically the experimentally measured absolute viscosity value (in cP) serving as the sole target variable (Target) for machine learning model fitting.All input feature matrices were screened and restructured based on the structure-property relationships surrounding this target variable.

To visually demonstrate the heterogeneity and nonlinear characteristics of the dataset, Table 2 presents a selection of experimental samples from the dataset. As shown in the table, even when choline chloride is fixed as the sole acceptor, the introduction of hydrogen-bonding donors with different chemical properties, combined with variations in molar ratio and temperature, results in significant cross-order-of-magnitude nonlinear changes in the system's macroscopic viscosity.

Table 2. Excerpt of experimental data on the viscosity of binary low-melting-point solvents (DES)

Hydrogen Bond Acceptor (HBA)	Hydrogen Bond Donor (HBD)	HBA molar fraction (X#1)	HBD Molar Fraction (X#2)	Temperature (T/K)	Experimental viscosity (cP)
Choline chloride	Chromic chloride hexahydrate	0.710	0.290	347.65	21.7
Choline chloride	Phenol	0.192	0.808	338.23	97.9
Choline chloride	Ethylamine hydrochloride	0.667	0.333	333.21	50.8
Choline chloride	1,2-Propanediol	0.160	0.840	293.13	29.0

2.1.2 PubChem-Based Dataset of Molecular Properties

To convert macroscopic names into features readable by machine learning models, this study used choline chloride as the primary hydrogen bond acceptor (HBA) and extracted a variety of typical hydrogen bond donors (HBDs), including glycerol, ethylene glycol, tartaric acid, diethylene glycol,oxalic acid, triethylene glycol, and xylitol.By searching the authoritative chemical database

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

PubChem, molecular properties were obtained, including molecular weight (Molecular_Weight), exact mass (Exact_Mass), monoisotopic mass (Monoisotopic_Mass), charge (Charge), octanol-water partition coefficient (XLogP), polar surface area (Polar_Area), complexity (Complexity),number of heavy atoms (Heavy_Atom_Count), and number of rotatable bonds (Rotatable_Bond_Count), as well as the number of hydrogen bond donors (H-Bond_Donor_Count), number of hydrogen bond acceptors (H-Bond_Acceptor_Count), number of covalent units (Covalent_Unit_Count), number of isotopic atoms (Isotopic_Atom_Count),Total Stereocenter Count, Defined Stereocenter Count, and Undefined Stereocenter Count, among others. This approach constructs an initial high-dimensional feature space that comprehensively characterizes the molecular micro-physicochemical environment.Table 3 presents data on selected basic physicochemical properties of some HBAs and HBDs.

Table 3. Basic physicochemical properties of HBAs and selected HBDs

Molecule Name	Constituent Role	Molecular Weight (MW)	Polar Surface Area	XLogP	Hydrogen Bond Donor Count (HBD Count)	Hydrogen Bond Acceptor Count (HBA Count)	Number of Rotatable Bonds
Choline chloride	HBA	139.62	20.30	-3.70	1	2	2
Glycerol	HBD	92.09	60.70	-1.70	3	3	2
Ethylene glycol	HBD	62.07	40.50	-1.40	2	2	1
Tartaric acid	HBD	150.09	115.00	-1.40	4	6	3
Diethylene Glycol	HBD	106.12	50.20	-1.30	2	3	4
……	……	……	……	……	……	……	……

2.2 Data Cleaning and Multidimensional Feature Engineering

Since chemical experiment data often contains noise, missing values, and non-normal distributions, using it directly for model training can easily lead to the curse of dimensionality or cause the model to fail to converge. Therefore, this study implemented a rigorous feature engineering and data cleaning pipeline.

2.2.1 Missing Value Imputation, Outlier Removal, and Target Variable Transformation

First, the validity of all molecules' SMILES strings was verified using the RDKit toolkit, and invalid samples with unparseable structures were removed. For dimensions in the feature matrix with a small number of missing values, reasonable imputation was performed using the K-nearest neighbors (KNN) interpolation method based on feature space similarity.

Second, regarding the target label of viscosity, since macroscopic fluid viscosity varies widely across different formulations and temperatures, it typically exhibits a significantly right-skewed (long-tail) distribution. Directly fitting linear or tree models would result in severe bias in the model's predictions for high-viscosity samples (i.e., the model would be dominated by outliers). This study first employed the interquartile range (IQR) method to remove extreme outliers:

$$IQR = Q_3 - Q_1 \quad (1)$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

outliers outside the interval [Q_1−1.5IQR, Q_3+1.5IQR] are removed. Subsequently, a logarithmic transformation is applied to the viscosity labels:

$$y' = \log_{10}(1 + y) \tag{2}$$

This transformation forces the long-tail data to approximate a normal distribution, significantly improving the model's fitting stability and regression accuracy across the entire viscosity range.

2.2.2 RDKit-Based Molecular Descriptor Feature Extraction

To translate the chemical essence of eutectic solvents into high-dimensional numerical vectors that computers can interpret, we utilized the RDKit toolkit in Python to perform in-depth digital characterization of hydrogen bond acceptors (HBAs) and hydrogen bond donors (HBDs) in binary systems. Compared to traditional group contribution methods, RDKit can extract features across multiple scales, including molecular graph theory, topological manifolds, and three-dimensional geometric configurations.Ultimately, a feature space encompassing six core dimensions was constructed, enabling a comprehensive characterization ranging from basic composition to deep-level electron distribution. This provides a robust data foundation for subsequent precise viscosity predictions and mechanistic analysis.

Regarding the characterization of molecular composition and topological connectivity, descriptors such as exact molecular weight (ExactMolWt) and heavy atom count (HeavyAtomCount) were first extracted to define the basic size and electron density of constituent molecules.Building on this, the Kappa indices (Kappa1–3) and Hall-Kier sorting values were used to quantify the molecular shape and the compactness of spatial arrangement, while connectivity indices (Chi series) were employed to finely characterize the degree of internal branching within the molecule. These topological descriptors reflect the mutual restraining effects of molecules as they move between fluid layers, forming the physical basis for viscosity.Furthermore, the introduction of the molar refractive index (MolMR) provides a comprehensive reflection of the molecule's volume effects and polarity.

Regarding molecular surface properties and physicochemical interactions, the study focused on topological polar surface area (TPSA), which quantifies the contribution of polar regions within molecules (N and O atoms and their bonded hydrogens). High TPSA values typically indicate stronger dipole-dipole interactions, which significantly increase fluid resistance.To characterize the complex microenvironmental effects of DES more precisely, the study also extracted LabuteASA and segmented surface area descriptors (SlogP_VSA, SMR_VSA series). By calculating the van der Waals surface area of atoms and segmenting them based on lipid-water partition coefficients (MolLogP), the distribution of different chemical environments (hydrophilic/hydrophobic regions) on the molecular surface was characterized.This characterization method is crucial for capturing subtle intermolecular electrostatic couplings.

The three-dimensional spatial geometry and charge distribution of a molecule are key factors determining the microscopic "mechanical friction" of viscosity. First, a series of principal moments of inertia (CalcPMI1, CalcPMI2, CalcPMI3) was introduced as core features to represent the magnitude of inertia when a molecule rotates around three orthogonal principal axes.Next, the eccentricity and sphericity index calculated via PMI can accurately distinguish linear, disc-shaped, and spherical molecules, effectively capturing the hindering effect of steric hindrance on the connectivity of hydrogen bond networks.At the electronic level, the maximum partial charge (MaxPartialCharge) of atoms within the molecule and its absolute value were extracted to reflect the strong electrostatic coupling forces resulting from local high charge density.Finally, to address the nature of DES formation, the study rigorously counted the number of hydrogen bond donor and acceptor sites (NumHDonors, NumHAcceptors) and combined this with the number of rotatable bonds (NumRotatableBonds) to characterize molecular dynamic flexibility, thereby constructing a

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

comprehensive descriptive framework spanning from microscopic steric hindrance to electrostatic-van der Waals coupling.

Through the aforementioned systematic extraction process, each HBA-HBD sample pair is mapped into a multiscale physicochemical feature space with up to 208 dimensions. This characterization approach ensures that the Stacking integrated model can deeply understand the patterns of viscosity fluctuations among different components resulting from steric hindrance, hydrogen bond network cross-linking, and electrostatic coupling, laying a solid data foundation for subsequent ultra-high-precision predictions.

2.2.3 Feature Dimension Reduction Based on Multi-Order Filtering Mechanisms

A high-dimensional feature space not only significantly increases computational overhead and training complexity but also exacerbates the "curse of dimensionality" in the sample dataset, substantially raising the risk of model overfitting and leading to a severe decline in model generalization ability. To address this core issue, this study designed and implemented a progressive feature selection strategy. Through a systematic process of filtering from coarse to fine in successive layers, the feature dimension is compressed while maximizing the retention of key predictive information, ultimately constructing a core feature set that balances predictive accuracy and computational efficiency.

First, a zero-variance feature removal operation is performed to directly eliminate all constant features whose values are identical across all samples. Such features contain no discriminative information and make no substantive contribution to model predictions; their early removal effectively reduces redundant computations in subsequent stages.

After completing the initial feature cleaning, the Z-score standardization method is applied to all retained features to perform a uniform scale transformation, mapping them to a standard normal distribution with a mean of 0 and a standard deviation of 1. This transformation effectively eliminates differences in units and numerical scales between features, preventing training bias caused by vastly different feature value ranges. The mathematical expression is:

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma} \tag{3}$$

where$x$ is the original feature value,μ is the mean of that feature in the training set, andσ is the corresponding standard deviation.

Subsequently, Pearson correlation analysis is introduced to detect multicollinearity by calculating the correlation matrix among all features, using the following formula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{4}$$

Here, n represents the total number of samples, x_i and y_i denote the observations of two different features in the i th sample, and $\overline{x}$ and $\overline{y}$ are the sample means of the corresponding features.The correlation coefficient$r$ ranges from$[-1,1]$: when$r > 0$, the variables are positively correlated; when$r < 0$, the variables are negatively correlated. The closer the absolute value$|r|$ is to 1, the higher the degree of linear correlation between the variables. The correlation results for some features are shown in Fig. 1.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

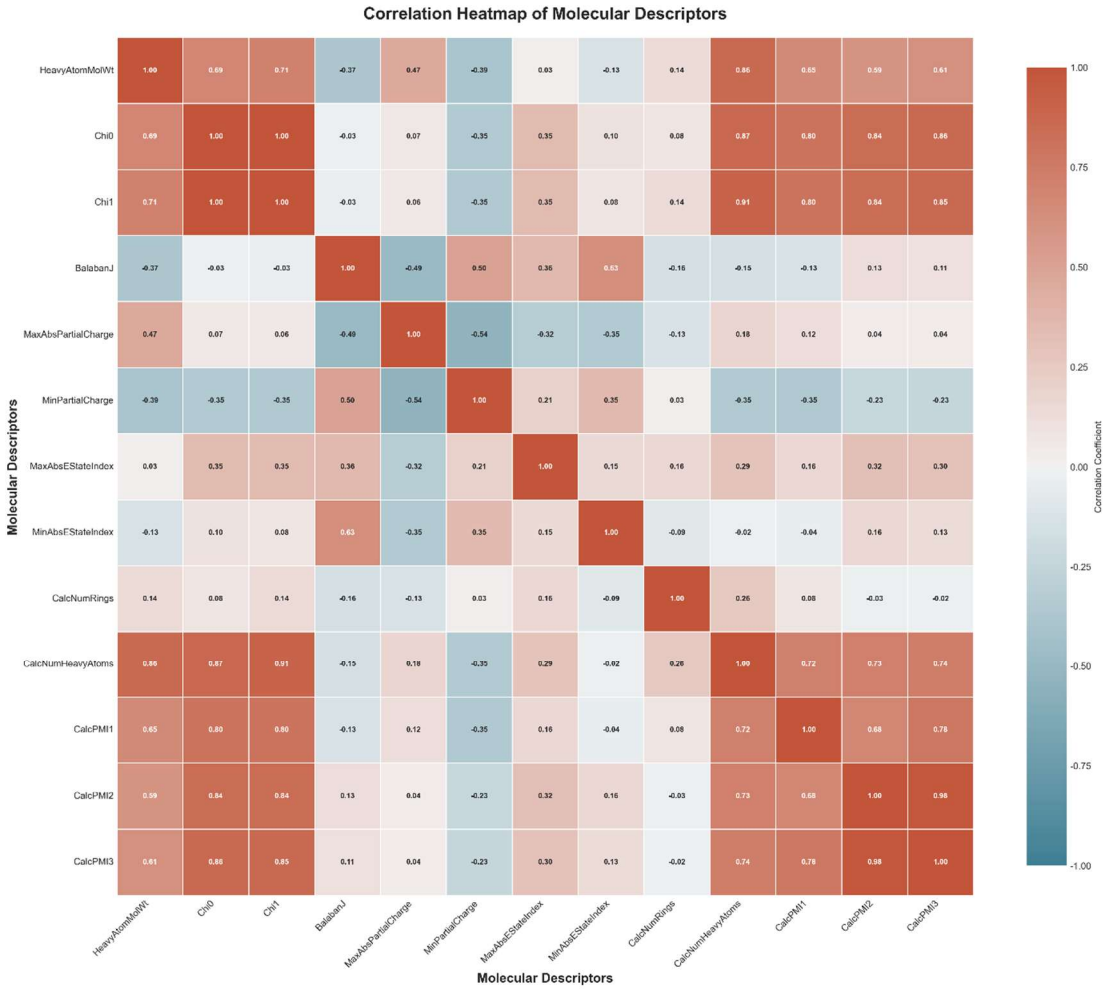

Fig. 1 Heatmap of the correlation matrix for selected molecular properties

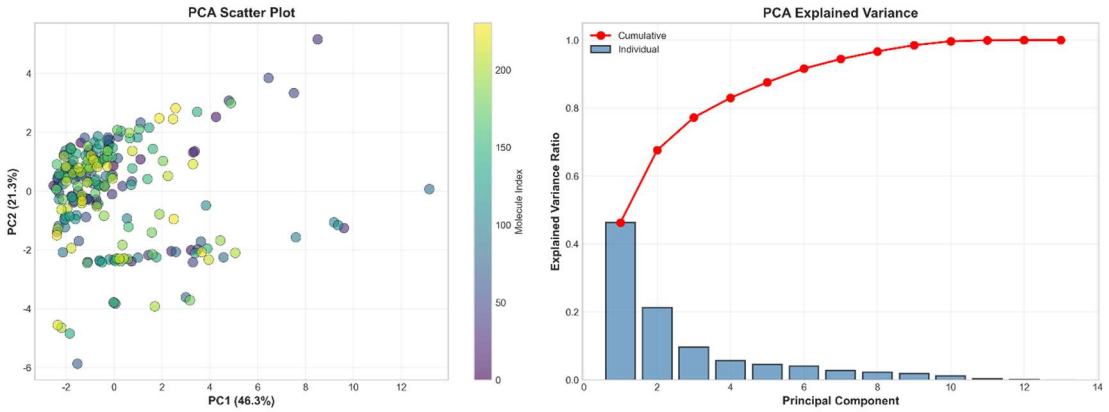

Fig. 2 PCA dimension reduction analysis of core feature variables

For highly redundant feature pairs with an absolute correlation coefficient $|r| > 0.8$, only the feature with higher correlation to viscosity was retained. This approach eliminates obvious linear dependencies among features, optimizes the independence of the model input matrix, reduces the model's sensitivity to noise, and preliminarily simplifies the feature space.

Finally, to further eliminate potential multicollinearity among the remaining features and validate the validity of the feature space after dimensionality reduction, this study introduced Principal Component Analysis (PCA) to perform orthogonal transformation and deep dimensionality reduction on the data. Through linear mapping, PCA projects the original highly correlated multidimensional

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

features onto a new orthogonal coordinate system, generating a series of mutually uncorrelated composite variables, while ranking these principal components according to the principle of variance maximization.

Based on the PCA analysis results of this study, as shown in Fig. 2, the first two principal components explain 46.3% and 21.3% of the total variance in the original data, respectively. The cumulative variance contribution indicates that only a few principal components are sufficient to highly summarize and retain the vast majority of physicochemical information in the original chemical feature space; the first 12 principal components cumulatively explain over 99% of the original information.Through this unsupervised dimensionality reduction strategy, this study thoroughly filtered out redundant background noise while maximizing the preservation of the data's intrinsic structure and distribution differences. Ultimately, it refined a standardized, high-quality, and mutually independent orthogonal feature matrix, providing reliable data inputs for the efficient fitting and stable training of subsequent machine learning models.

3. Model

3.1 Theoretical Algorithms of Base Models and Meta-Models

3.1.1 Random Forest (RF)

Random Forest is a parallel ensemble learning algorithm based on the Bagging concept [8]. When dealing with chemical property data containing noise, RF demonstrates advantages in resisting overfitting and exhibiting strong robustness to outliers.

The RF algorithm uses bootstrap sampling to draw multiple training subsets with replacement from the original dataset and constructs a CART regression tree on each subset. During node splitting, RF introduces a random feature selection mechanism, which involves randomly selecting features within a specified range ($max\_features$) to perform splits that minimize the mean squared error. For regression problems, the final prediction of RF is the arithmetic mean of the outputs from all independent decision trees:

$$H(x) = \frac{1}{K}\sum_{k=1}^{K} h_k\,(x) \tag{5}$$

Here, K represents the total number of decision trees (i.e., the hyperparameter $n\_estimators$), and $h_k(x)$ is the viscosity prediction value for the input vector x from the k th decision tree. By ensemble learning, RF smooths out the variance of individual trees, effectively improving generalization performance.

3.1.2 Extreme Gradient Boosting (XGBoost)

XGBoost is a sequential ensemble algorithm based on Boosting. Its core idea is to fit the residuals generated by the previous rounds model predictions by continuously adding new regression trees [9]. XGBoost performs a second-order Taylor expansion of the loss function of traditional GBDT and incorporates a regularization term to control tree complexity, making it particularly suitable for capturing the extremely complex nonlinear topological relationship between hydrogen bonding interactions and macroscopic viscosity in the DES system.

Its objective function is defined as:

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{6}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

Where l is a differentiable convex loss function (mean squared error is used in this study), $\hat{y}_i^{(t-1)}$ represents the viscosity prediction values from the previous$t-1$ rounds, $f_t(x_i)$ is the output model of the t th tree, and $\Omega(f_t)$ is the regularization penalty term, calculated as:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{7}$$

Here, T represents the number of leaf nodes in the tree, w_j is the prediction weight of the j th leaf node, γ is the threshold difficulty controlling node splitting, and λ is the L2 regularization coefficient. By finely tuning these regularization parameters, overfitting of the model on small-scale chemical datasets can be effectively mitigated.

3.1.3 Linear Regression (LR)

In the second layer of the Stacking ensemble architecture, Linear Regression (LR) is selected as the meta-learner. As a classic and robust parametric statistical model, the core function of LR is to learn and balance the primary prediction values output by the underlying heterogeneous base learners (RF and XGBoost), achieving collaborative compensation of prediction errors through globally optimal weight allocation [10].

The choice of linear regression over more complex nonlinear models as the meta-learner is based on a careful consideration of the trade-off between model complexity and generalization performance.Since the predictions (meta-features) output by the base learners in the Level 0 layer already possess extremely high information purity and exhibit a strong linear correlation with the target variable (viscosity), introducing a high-complexity model at Level 1 would likely cause the model to capture noise within the limited meta-feature space, thereby leading to overfitting. By maintaining a low model capacity, LR effectively ensures the robustness of the Stacking architecture during the secondary learning process.

Assuming the vector of base learner predictions generated by cross-validation at Level 0 is $\hat{\mathbf{y}}_{base} = [\hat{y}_{RF}, \hat{y}_{XGB}]^T$, the final prediction function of the linear regression submodel is expressed as:

$$\hat{y}_{final} = w_1\hat{y}_{RF} + w_2\hat{y}_{XGB} + \beta \tag{8}$$

Where w_i are the weighted coefficients (regression weights) of the corresponding base learner, and β is the model bias term. The meta-learner determines the optimal coefficients by minimizing the mean squared error (MSE) objective function:

$$J(w, \beta) = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_{final,i})^2 \tag{9}$$

Through dynamic trade-offs achieved via linear regression, the Stacking model accurately identifies the predictive strengths of base models across different viscosity ranges. It effectively combines the random variance smoothing capability of RF with the nonlinear residual correction capability of XGBoost, ultimately producing viscosity prediction results that feature both high predictive accuracy and cross-system generalization capabilities.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

3.2 Global Hyperparameter Optimization based on the Whale Optimization Algorithm (WOA)

The performance of machine learning models is highly dependent on hyperparameter settings. Traditional grid search not only incurs significant computational overhead but is also prone to getting stuck in local optima within the continuous, high-dimensional chemical feature space. To address this, this paper introduces a swarm intelligence-inspired algorithm-the Whale Optimization Algorithm (WOA)-to perform adaptive global intelligent optimization of the core hyperparameters for RF and XGBoost [11].

The WOA algorithm simulates the "bubble-net" hunting strategy unique to humpback whales in nature. The algorithm updates the position of solutions (i.e., hyperparameter combinations) in the search space through three core behaviors:

Encircling Prey: Assuming the current optimal individual (the hyperparameter combination with the highest predictedR^2 on the validation set) is the target prey, other individuals converge toward it. The position update formula is:

$$\vec{D} = \left|\vec{C} \cdot \vec{X}^*(t) - \vec{X}(t)\right| \vec{X}(t+1) = \vec{X}^*(t) - \vec{A} \cdot \vec{D} \tag{10}$$

Where t is the current iteration number, $\vec{X}^*(t)$ is the current optimal solution, and $\vec{X}(t)$ is the current individual's position. The calculation of the coefficient vectors $\vec{A}$ and $\vec{C}$ includes linear decay factors to balance global and local search.

Bubble-net Attacking: The algorithm employs a spiral update mechanism to simulate the trajectory of a humpback whale expelling a bubble net and spiraling upward:

$$\vec{X}(t+1) = \vec{D}' \cdot e^{bl} \cdot \cos(2\pi l) + \vec{X}^*(t) \tag{11}$$

Where $\vec{D}'$ represents the distance between the individual and the optimal solution, b is a constant defining the spiral shape, and l is a random number between $[-1,1]$. The algorithm switches randomly between the enveloping mechanism and the spiral model with a 50% probability.

Search for Prey: When the coefficient vector is set to $|\vec{A}| \geq 1$, individuals deviate from the current optimal target and randomly select other individuals in the population as references for position updates. This mechanism endows the algorithm with strong global exploration capabilities, enabling it to escape the trap of local optima.

3.3 Construction of a WOA-Based Stacking Ensemble Model Architecture

To fully explore the deep nonlinear mapping relationship between the microstructural characteristics of deep eutectic solvents (DES) and their macroscopic viscosity, and to overcome the generalization bottlenecks and prediction biases inherent in single machine learning models when handling complex sample data in the chemical engineering field, this study constructs a two-layer Stacking ensemble model architecture driven by the Whale Optimization Algorithm (WOA).The core logic of Stacking lies in achieving a stepwise transfer of knowledge through a multi-level learning mechanism. By utilizing a meta-learner to perform secondary learning and weighted fusion of the prediction results from multiple heterogeneous base learners, it achieves a synergistic effect of residual correction and variance reduction. The overall process is shown in Fig. 3.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

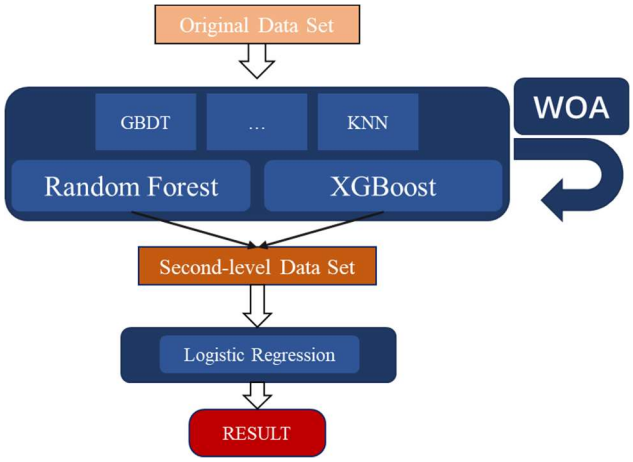

Fig. 3 WOA-based Stacking ensemble learning model

The Stacking architecture designed in this study first employs the WOA algorithm to intelligently pre-optimize the underlying base learners. Since the performance ceiling of the base learners directly determines the final accuracy of the ensemble model, this study uses WOA to simulate the contracting envelope and spiral update mechanisms observed in humpback whales, thereby identifying optimal configurations for Random Forests (RF) and Extreme Gradient Boosting (XGBoost) within the vast hyperparameter search space.This step ensures that every base model input at Level 0 is in an optimal state, thereby reducing system errors caused by hyperparameter misconfiguration at the algorithmic source.

At Level 0 (the base learner layer), this study selected RF and XGBoost as the core prediction units, while gradient-boosted decision trees (GBDT), support vector regression (SVR), and k-nearest neighbors (KNN) were found to have relatively poor performance through experimental testing and were therefore not included in the base models for stacking. RF is a parallel ensemble algorithm based on the Bagging concept. It effectively reduces model variance by constructing a large number of independent decision trees and taking their average, demonstrating strong robustness against random noise in the DES experimental data; In contrast, XGBoost is a serial iterative algorithm based on the Boosting concept. By continuously fitting residuals through second-order Taylor expansions and L2 regularization terms, it can accurately capture the contribution of minute changes in molecular features to viscosity, exhibiting extremely low bias. Stacking these two models with fundamentally different ensemble mechanisms allows for the complementary advantages of “variance reduction” and “bias reduction,” thereby fundamentally enhancing the model’s ability to capture dimensions in complex chemical spaces.

To strictly prevent "data leakage" and ensure that meta-learners can access the true prediction distribution, the Level 0 layer employs a 5-fold out-of-fold (OOF) cross-validation mechanism. The specific implementation process is as follows: the original training set is divided into five subsets. In each iteration, four of these subsets are used to train the base learners, while the remaining validation subset is used for viscosity prediction.After five iterations, the prediction results from all validation subsets are concatenated to generate a new feature vector of the same length as the original dataset. This set of preliminary predictions, outputted by RF and XGBoost, constitutes the “meta-features” for subsequent training. Concurrently, the base models are retrained using the full training set, and inferences are made on an independent test set to generate the meta-feature matrix for the testing phase.

At Level 1 (the meta-learner layer), this study selected Linear Regression (LR), which is highly robust and logically straightforward, as the decision-making core. Since the meta-features output by Level 0 already possess extremely high predictive accuracy, using overly complex nonlinear models as meta-learners would likely lead to overfitting during the learning process in the second layer.The LR

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

model uses the least squares method to learn the contribution weights of RF and XGBoost across different prediction intervals, constructing the following linear weighted equation:

$$\hat{y}_{final} = w_1 \cdot \hat{y}_{RF} + w_2 \cdot \hat{y}_{XGB} + b \tag{12}$$

Here, $\hat{y}_{RF}$ and $\hat{y}_{XGB}$ represent the outputs of the base learners, respectively, while w_1 and w_2 are the optimized combination coefficients. Through this linear weighting, Level 1 intelligently identifies the distribution of strengths across each base model, corrects low-accuracy predictions, and ultimately outputs viscosity prediction results with ultra-high accuracy and stable generalization.

3.4 Performance Evaluation and Validation Metrics

To objectively and quantitatively evaluate the prediction accuracy and generalization capability of the Stacking ensemble model for DES viscosity, this study established a rigorous mathematical evaluation metric system. All models were validated through extrapolation on an independently partitioned test set, with a training-to-test split ratio of 8:2. The specific core evaluation metrics adopted are as follows:

Coefficient of Determination (R^2): Indicates the proportion of variance in the observed values that is explained by the model's predicted values. The closer the value is to 1, the better the model fits the data and the stronger its explanatory power.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{13}$$

Root Mean Squared Error (RMSE): A measure of the absolute magnitude of prediction bias. Because the error is squared, this metric is highly sensitive to outliers with extreme deviations and effectively penalizes prediction points with severe deviations.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{14}$$

Mean Absolute Error (MAE): Reflects the average level of absolute error between predicted and true values. Compared to RMSE, MAE does not amplify errors by squaring them, making it more robust to extreme outliers and providing a more intuitive and accurate measure of the model's average prediction bias across the entire sample.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{15}$$

Average Absolute Relative Deviation (AARD): This metric holds the greatest engineering significance in the estimation of chemical property values. It eliminates the interference caused by the absolute magnitude and units of viscosity across different systems, instead characterizing the relative percentage of prediction error; a smaller value indicates higher reliability of the model in engineering applications.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

$$AARD(\%) = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (16)$$

In Equations (13), (14), (15), and (16), n is the total number of samples in the test set, y_i is the logarithm of the true experimental viscosity from the DES, $\hat{y}_i$ is the logarithm of the viscosity predicted by the model, and $\overline{y}$ is the sample mean of the true values.

4. Experiment

4.1 Evaluation and Selection of Baseline Models

To comprehensively evaluate the suitability of different algorithms for the viscosity prediction task of low-melting-point solvents (DES), five classic machine learning regression models were first constructed as baseline estimators during the experimental phase: Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Regression (SVR), Gradient Boosted Decision Trees (GBDT), and K-Nearest Neighbors Regression (KNN).During the model training phase, 5-fold cross-validation (5-Fold Cross-Validation) was uniformly applied to all baseline models to avoid evaluation randomness caused by a single data split and to ensure that the models possess reliable generalization capabilities.

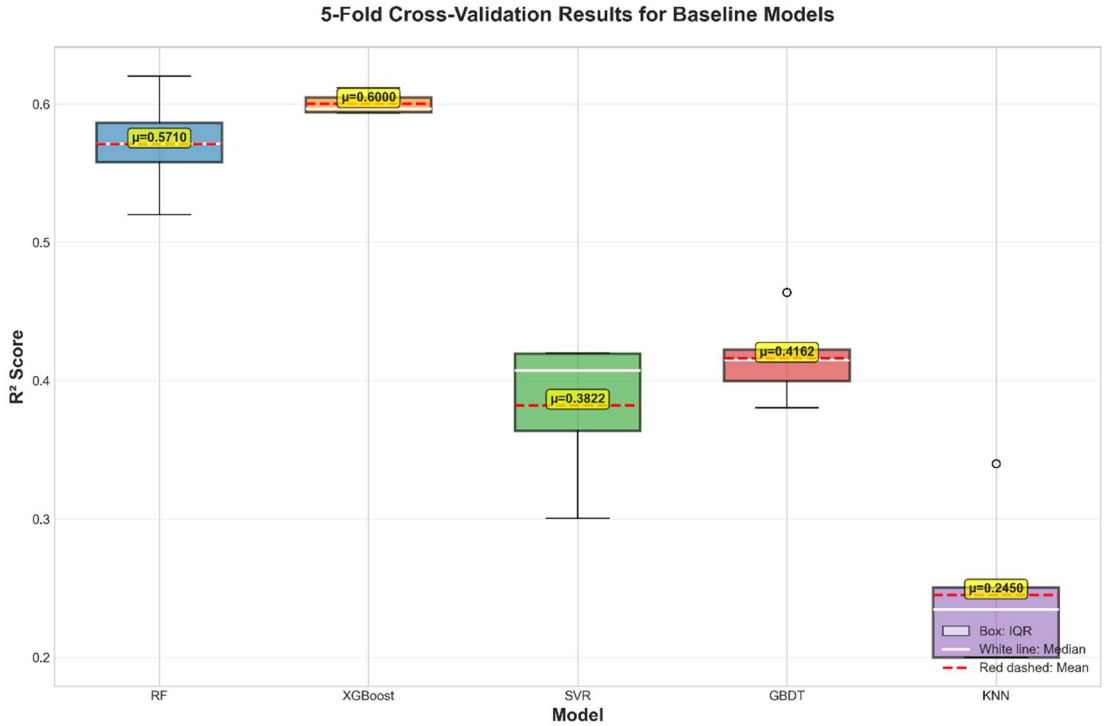

Fig. 4 Box plot of the R^2 distribution for the five base models under 5-fold cross-validation

The box plot shown in Fig. 4 visually illustrates the distribution range and median levels of the coefficient of determination (R^2) for the five models during cross-validation. The evaluation results show that the XGBoost and RF models demonstrated the best predictive performance, with average R^2 values of 0.6000 and 0.5710, respectively, in the 5-fold cross-validation. The data distribution was relatively concentrated, proving that tree-based ensemble algorithms possess excellent fitting capabilities when handling high-dimensional chemical features.In contrast, the fitting performance of SVR and GBDT was moderate, while the KNN model's average R^2 was only 0.2450, making it difficult to effectively capture the complex nonlinear mapping relationships within the complex

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

chemical space of the samples.Furthermore, independent evaluation on the initial test set further indicates that the initial R^2 of the XGBoost model is 0.6114 (with an Average Absolute Relative Deviation, AARD, of 67.44%), while the initial R^2 of the RF model is 0.5830 (with an AARD of 62.25%).These objective metrics establish the central role of XGBoost and RF in the DES viscosity prediction task, thereby identifying the target base models for subsequent hyperparameter optimization.

4.2 WOA-Based Hyperparameter Optimization and a Significant Improvement in Base Model Performance

Given that the hyperparameters of the initial base models were set to their default values, and the prediction accuracy and error control had not yet met industrial application standards, the Whale Optimization Algorithm (WOA) was subsequently introduced to perform global intelligent optimization on the selected RF and XGBoost models. The optimization objective was set as a joint constraint of maximizing the validation set R^2 and minimizing the AARD, in order to avoid the shortcoming of traditional grid search, which is highly prone to getting stuck in local optima.

After multiple spiral iterations and adaptive enveloping optimization by the WOA algorithm, the optimal parameter spaces for both models were finally established. For the RF model, the optimal parameter configuration is: n_estimators=200, max_depth=15, min_samples_split=3, max_features=0.7, etc.;For the XGBoost model, the optimal parameters converged to: n_estimators=300, learning_rate=0.05, max_depth=8, gamma=0.1, reg_lambda=2, etc. The optimized parameter sequences were re-substituted into the models, and comprehensive performance backtesting was conducted on the test set.

Table 4. Comparison of performance metrics for baseline models on the test set after WOA optimization

Model	R^2 Score	RMSE (cP)	MAE (cP)	AARD (%)
XGBoost	**0.8123**	**11.0532**	**7.9542**	**15.42**
RF	0.7456	13.8415	9.6321	17.75
GBDT	0.6892	15.2104	11.0543	21.34
SVR	0.3541	21.6543	16.4320	31.05
KNN	0.1205	25.1028	19.8541	36.82

Table 4 summarizes the key performance metrics of each baseline model on the test set after deep tuning using the Whale Optimization Algorithm (WOA). The data indicates that WOA's global optimization significantly improved the models' generalization ability, with XGBoost and Random Forest (RF) establishing a clear lead.XGBoost ranks first in performance, with its coefficient of determination (R^2) surging to 0.8123, while its RMSE, MAE, and AARD dropped to the lowest values in the group at 11.0532 cP, 7.9542 cP, and 15.42%, respectively. RF follows closely behind, with aR^2 of 0.7456 and an AARD of 17.75%.In contrast, the remaining models (such as KNN, whoseR^2 is only 0.1205) exhibit poor predictive accuracy and struggle to handle complex nonlinear mapping tasks.

The experimental results described above provide a solid foundation for the subsequent development of a stacking ensemble architecture. As a typical representative of boosting-based methods, XGBoost excels at reducing prediction bias; meanwhile, as a parallel ensemble algorithm based on the bagging mechanism, RF effectively reduces model variance and resists data noise.Not only do these two methods rank first and second in terms of prediction accuracy, but the inherent heterogeneity of their underlying algorithms also ensures the diversity of the ensemble model. Therefore, selecting XGBoost and RF as the core base learners maximizes the complementarity of their strengths,

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

effectively approaching the theoretical upper limit of prediction performance while strictly controlling overfitting.

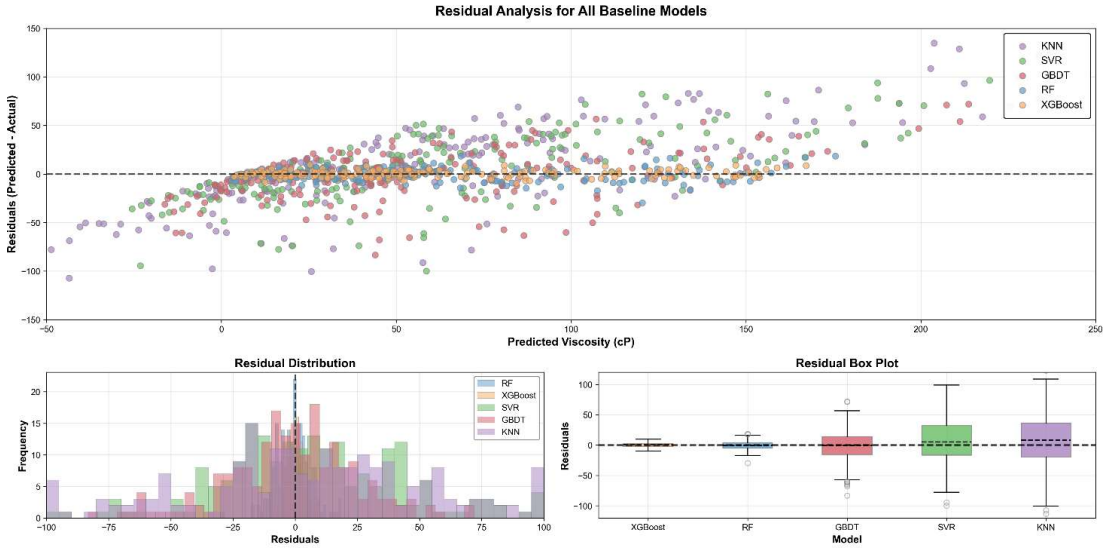

Fig. 5 Analysis of prediction residuals for each base model optimized by WOA

Fig. 5 provides a multidimensional analysis of the error distribution characteristics of each model through residual scatter plots, histograms, and box plots. The results clearly indicate that the prediction errors of the WOA-optimized XGBoost and RF models are significantly compressed near the zero baseline. The histograms exhibit a tall, converging zero-mean normal distribution, with extremely compact boxes, demonstrating excellent global prediction accuracy and stability.In contrast, when dealing with samples of medium to high viscosity, the residual scatter plots of the GBDT, SVR, and KNN models exhibit significant heteroscedasticity characterized by a pronounced "funnel-shaped" divergence, with flat and broad histograms and severely stretched boxes accompanied by a large number of extreme outliers.This stark contrast fully confirms that the WOA-XGBoost model, based on tree-based ensemble methods, possesses exceptional robustness in handling the complex high-dimensional mapping of eutectic solvents and suppressing extreme outliers, whereas algorithms that over-rely on global distance metrics (such as KNN) or traditional boundary fitting (such as SVR) struggle to handle such complex nonlinear prediction tasks.

4.3 Prediction Results of the WOA-Stacking Ensemble Model

Although the WOA-XGBoost model demonstrated good predictive performance as a standalone algorithm, the single-model approach still faces limitations in learning when dealing with feature spaces containing highly complex hydrogen-bond networks and electrostatic coupling effects. To thoroughly overcome this generalization bottleneck, the final experiment implemented a two-layer Stacking ensemble fusion model. This architecture uses WOA-optimized RF and XGBoost as base learners in the Level-0 layer, and selects a linear regression (LR) model-which is logically simple and robust to noise-as the meta-learner in the Level-1 layer, performing secondary learning through 5-fold cross-feature concatenation.

Fig. 6 comprehensively evaluates the overall performance of the WOA-tuned Stacking ensemble model (RF + XGBoost-LR) in the DES viscosity prediction task. In terms of core performance metrics, the fusion model demonstrates excellent predictive accuracy on the independent test set, with a coefficient of determination (R^2) reaching 0.8620, the average absolute relative deviation (AARD) significantly optimized to 9.88%, and the RMSE reduced to 7.0276 cP. All metrics outperform the aforementioned single baseline models, proving the substantial improvement in model generalization capability achieved by the Stacking architecture.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

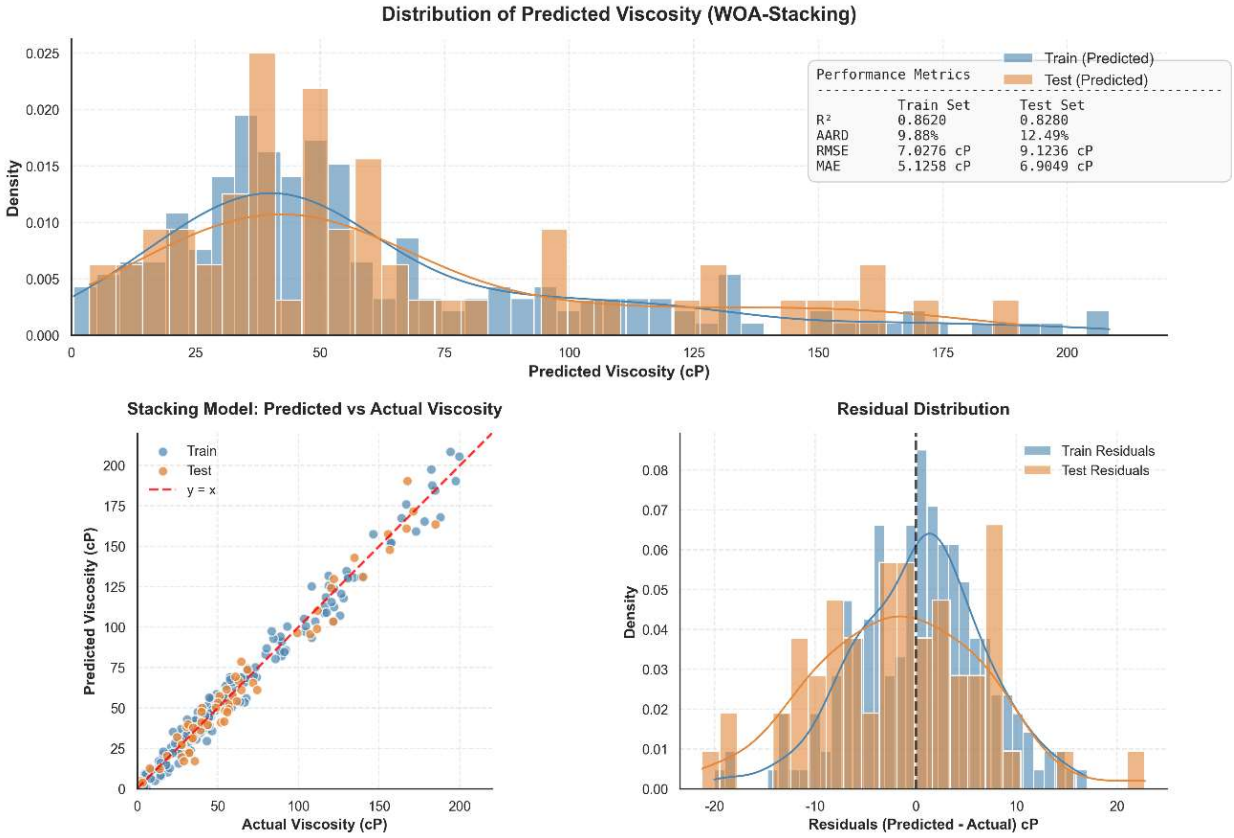

Fig. 6 Analysis of the prediction performance and residual distribution of the WOA-Stacking ensemble model

At the level of viscosity distribution (Distribution of Predicted Viscosity), the predicted results exhibit a significant right-skewed (long-tail) distribution, with the majority of data points concentrated in the common viscosity range of 20–100 cP, while the distribution is sparse in the extremely high viscosity region of 160–200 cP.This distribution pattern closely aligns with the actual physical distribution of low-eutectic solvents, demonstrating the model's ability to accurately capture cross-order-of-magnitude nonlinear mappings.

In terms of regression fit (Predicted vs. Actual Viscosity), the scatter points are highly clustered near the diagonal line ($y = x$). Observation reveals that in the low-to-medium viscosity range, both the training and test sets exhibit extremely high consistency in fit; whereas in the sparse high-viscosity region above 160 cP, although the absolute error fluctuates slightly as viscosity increases, the prediction trend remains robust and shows no systematic bias.

In terms of error robustness (Residual Distribution), the residual distributions of both the training and test sets are highly symmetric and strictly centered at 0, exhibiting good normality. Although the residual distribution of the test set is slightly broader than that of the training set, the overall high convergence ensures that the model maintains extremely low bias even when handling nonlinear fluctuations caused by complex hydrogen-bond networks.In summary, the WOA-Stacking model successfully achieves high-precision mapping and robust generalization across the complex viscosity space of DES.

4.4 Analysis of Model Interpretability and Revealing of Microscopic Mechanisms

Although the WOA-optimized Stacking ensemble model demonstrates excellent performance in terms of predictive accuracy and generalization ability, the "black-box" nature of ensemble learning models obscures the physicochemical mapping relationships between input features and macroscopic viscosity. To overcome this limitation, we introduced the game-theory-based SHAP (SHapley Additive exPlanations) analysis technique to provide post-hoc explanations for the model's

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

predictions. Combined with the reconstructed cross-feature space, the model enables a deep analysis of the fundamental drivers behind the drastic viscosity changes in the DES system from the perspective of microscopic interactions.

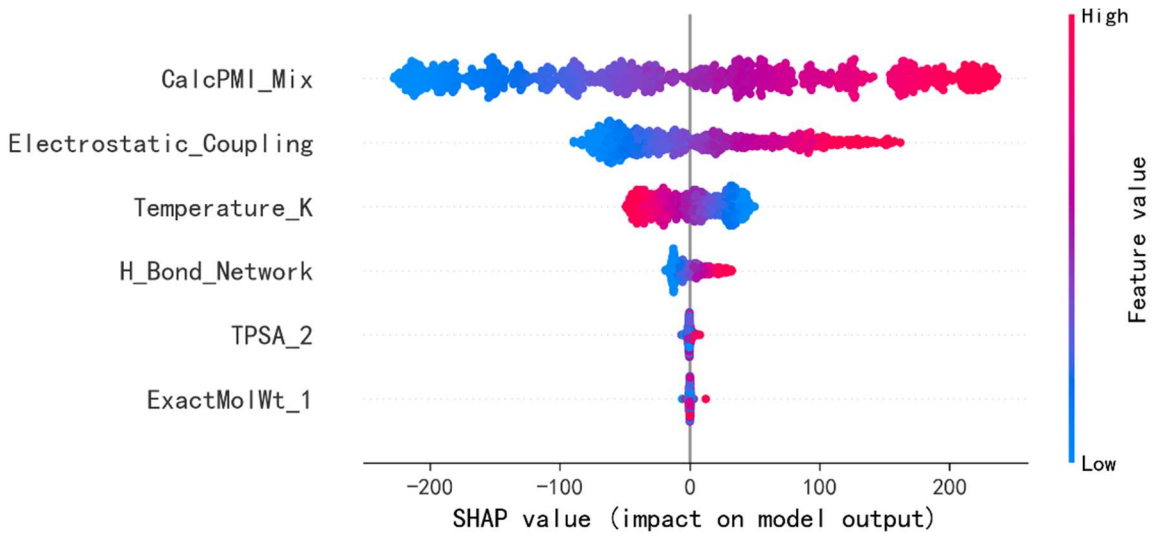

Fig. 7 Scatter plot of feature importance

Fig. 7 shows the feature importance ranking based on the average absolute SHAP value and the sample distribution. This figure reveals, from a global perspective, the magnitude of contributions made by various microscopic physicochemical features to viscosity prediction, as well as their positive and negative driving relationships.

First, the mixed-space steric hindrance (CalcPMI_Mix) ranks first in feature importance, with its red sample points (high feature values) significantly concentrated in the right-hand region where SHAP values are greater than 0.This indicates that when the overall principal moment of inertia of the system increases after mixing HBA and HBD, molecular shapes become more complex and spatial volume expands, leading to a sharp rise in mechanical frictional resistance within the fluid, which in turn significantly drives the increase in macroscopic viscosity.

Second, the electrostatic coupling strength (Electrostatic_Coupling) follows closely behind, exhibiting a strong positive driving effect. Its physical essence lies in the fact that the stronger the electrostatic attraction between ions and polar molecules in the system, the more easily microscopic molecules are “locked” within a local space, with their translational and rotational degrees of freedom significantly restricted, thereby triggering a significant increase in viscosity.Concurrently, the hydrogen bond network (H_Bond_Network) also exhibits a consistent positive synergistic thickening effect, demonstrating that the more hydrogen bond cross-linking sites between components, the denser the resulting three-dimensional network, and the greater the fluid flow resistance.

Furthermore, the temperature (Temperature_K) exhibits a perfect strong negative correlation. The red data points (high temperature) in the figure are all concentrated in the left region where SHAP values are less than 0, which precisely quantifies the Arrhenius law: an increase in system temperature provides a large amount of additional thermal energy, effectively breaking the hydrogen-bond cross-linking network and electrostatic bonds between microscopic molecules, thereby significantly reducing the macroscopic viscosity of the system.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

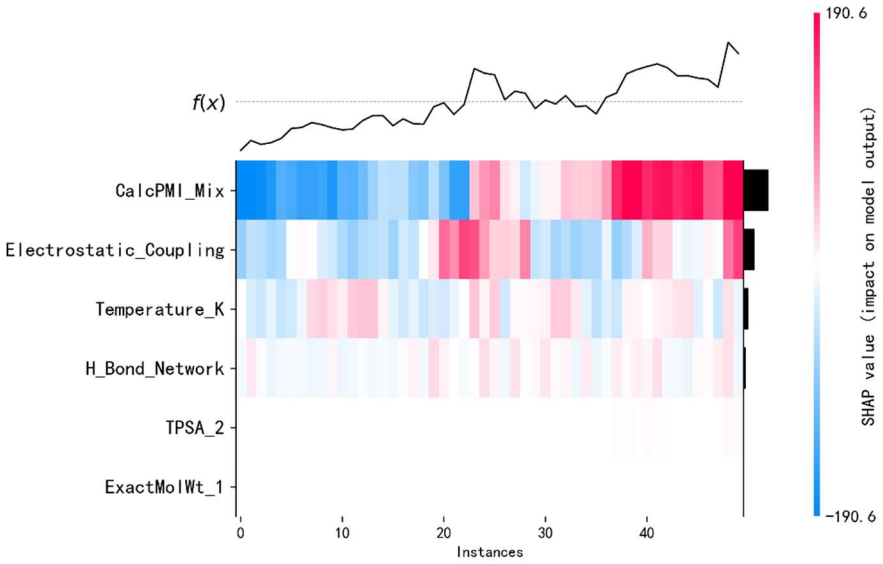

Fig. 8 Heatmap of SHAP influence patterns for typical samples under multi-feature interactions (Top 50 Samples)

Fig. 8 further illustrates the synergistic effects of local interactions among multiple features on individual samples through a heatmap. The horizontal axis represents 50 typical DES samples sorted by predicted viscosity from low to high, while the vertical axis shows feature names. The intensity of the colors indicates the magnitude of the SHAP value's influence on the model output (red indicates positive viscosity increase, blue indicates negative viscosity decrease).

In the region of high-viscosity samples on the right (samples 35–50), a distinct "red resonance" phenomenon is clearly observable. In these extremely high-viscosity systems, both CalcPMI_Mix and Electrostatic_Coupling appear deep red (indicating a very strong positive contribution), accompanied by lighter or blue distributions representing lower temperatures.This localized coupling effect strongly demonstrates that the emergence of extremely high DES viscosity is by no means dominated by a single factor, but rather results from the combined superposition and resonance of a dense hydrogen-bond cross-linking network, strong electrostatic coupling, and significant steric hindrance at lower temperatures.

In summary, the SHAP interpretability analysis confirms at the quantitative level that the dramatic change in DES viscosity essentially stems from the strong hydrogen-bonded network and electrostatic coupling formed after the mixing of HBA and HBD.The synergistic combination of high steric hindrance and strong electrostatic attraction greatly restricts the degrees of freedom of microscopic molecules, enhances the connectivity of the hydrogen bonding network within the system, and constitutes the core underlying microscopic mechanism that drives the macroscopic dynamic viscosity jump in DES.

5. Conclusion

Addressing industry pain points such as the heavy reliance of traditional DES viscosity prediction on experimental trial-and-error, the poor generalization ability of single models, and the lack of microscopic mechanism analysis, this paper innovatively proposes a high-precision prediction and interpretability analysis framework that integrates multi-dimensional feature engineering, the Whale Optimization Algorithm (WOA), and Stacking ensemble learning. During the feature engineering stage, this study extracted microscopic molecular descriptors from multiple scales and designed a progressive dimension reduction strategy incorporating zero-variance removal, standardization, multicollinearity detection, and PCA principal component analysis. This approach thoroughly filtered out redundant noise in the high-dimensional space while retaining core physicochemical information, laying a solid data foundation for the efficient fitting of the underlying base models.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

In terms of predictive model construction and validation, this study overcomes the generalization bottleneck of single estimators. By introducing the swarm intelligence WOA algorithm to perform global intelligent hyperparameter optimization for Extreme Gradient Boosting (XGBoost) and Random Forest (RF), and by employing Linear Regression (LR) as a meta-learner for secondary deep integration, we successfully constructed a two-layer WOA-Stacking architecture that combines the advantages of "variance reduction" and "bias reduction."Extrapolation evaluation on an independent test set demonstrates that this integrated architecture effectively overcomes the overfitting issues common in complex chemical datasets, achieving high-precision mapping of the complex nonlinear viscosity space of DES. It attained an excellent coefficient of determination (R^2) of 0.8620 and an average absolute relative deviation (AARD) as low as 9.88%, with its comprehensive predictive capability and global error control significantly outperforming various single-baseline models.

In terms of in-depth analysis of micro-mechanisms, this study introduced SHAP visualization analysis technology and, in combination with the reconstructed cross-feature space, successfully broke through the "black box" barrier of deep ensemble models. Analysis of global feature importance and local interaction effects has, for the first time at a quantitative level, clearly confirmed that the drastic changes in DES viscosity essentially stem from the strong electrostatic coupling and dense hydrogen-bond cross-linking network formed after the mixing of hydrogen bond acceptors (HBAs) and hydrogen bond donors (HBDs). The study profoundly reveals the underlying microscopic mechanism driving the macroscopic viscosity jump: the positive synergistic superposition of mechanical friction and strong electrostatic attraction (Electrostatic_Coupling) induced by high mixed-space steric hindrance (CalcPMI_Mix) significantly restricts the translational and rotational degrees of freedom of microscopic molecules;In extreme high-viscosity systems, the "coupled resonance" phenomenon-arising from the combined effects of a dense hydrogen-bond cross-linking network, strong electrostatic coupling, and massive steric hindrance at lower temperatures-is the fundamental driving force behind the sharp rise in hydrodynamic viscosity.Simultaneously, the strongly negative temperature-dependent response precisely quantifies the physicochemical mechanism in the Arrhenius law whereby increased temperature weakens molecular constraints.

In summary, this study establishes a data-driven modeling paradigm for estimating the physical properties of complex chemical systems, effectively bypassing the time-consuming and costly nature of traditional trial-and-error methods. Furthermore, through a multidimensional and mechanism-transparent analysis, it provides rigorous quantitative guidance for the targeted reverse design of low-viscosity, high-performance green DES. Ultimately, these insights lay a solid foundation for industrial applications in hydrometallurgy, particularly for the selective leaching of zinc-containing solid waste.

References

[1] Smith, E. L., Abbott, A. P., & Ryder, K. S. (2014). Deep eutectic solvents (DESs) and their applications. *Chemical Reviews*, 114, 11060–11082. https://doi.org/10.1021/cr500238n

[2] Zhang, Q., De Oliveira Vigier, K., Royer, S., et al. (2012). Deep eutectic solvents: syntheses, properties and applications. *Chemical Society Reviews*, 41, 7108–7146. https://doi.org/10.1039/C2CS35178A

[3] Abbott, A. P., Capper, G., Davies, D. L., et al. (2003). Novel solvent properties of choline chloride/urea mixtures. *Chemical Communications*, 1, 70–71. https://doi.org/10.1039/B211970G

[4] Shi, D., Zhou, F., Wu, J., et al. (2022). Deep insights into the viscosity of deep eutectic solvents using an XGBoost-based model with SHapley Additive Explanation. *Physical Chemistry Chemical Physics*, 24, 26029–26036. https://doi.org/10.1039/D2CP03818A

[5] Mohan, M., Jetti, K., Smith, M. D., et al. (2024). Accurate machine learning for the prediction of the viscosities of deep eutectic solvents. *Journal of Chemical Theory and Computation*, 20, 1155–1166. https://doi.org/10.1021/acs.jctc.4c00028

[6] Abbott, A. P., Capper, G., Davies, D. L., et al. (2003). Novel solvent properties of choline chloride/urea mixtures. *Chemical Communications*, 1, 70–71. https://doi.org/10.1039/B211970G

[7] Wu, T., & Lin, X. Q., et al. (2025). Multiscale exploration of informative latent features for accurate deep eutectic solvents viscosity prediction. *AIChE Journal*, 71, e18007. https://doi.org/10.1002/aic.18007

ISSN: 2414-1895 DOI: 10.6919/ICJE.202606_12(6).0007

[8] Xiang, J. Y., Wang, Z. H., & Deng, Y. Y. (2024). A review of machine learning classification based on the Random Forest algorithm. *Research on Artificial Intelligence and Robotics*, 13, 143–152.

[9] Chen, X. H., Hu, Y., Wang, Y. J., et al. (2024). Dam deformation interval prediction model based on XGBoost. *Journal of Hydroelectric Engineering*, 43, 121–136.

[10] Ma, X. W. (2008). Methods for diagnosing multicollinearity in linear regression equations and their empirical analysis. *Journal of Huazhong Agricultural University (Social Sciences Edition)*, 2, 78–81, 85.

[11] Mirjalili, S., & Lewis, A. (2016). The whale optimization algorithm. *Advances in Engineering Software*, 95, 51–67. https://doi.org/10.1016/j.advengsoft.2016.01.008