Machine Learning-Based Prediction of Coordination Number from EXAFS Spectra

Authors: Zeng Haitao, Hu Longfei, Yao Tao

Date: 2025-10-30T00:00:00+00:00

Abstract

X-ray Absorption Fine Structure (XAFS) is an important structural analysis technique widely employed for investigating the oxidation states, coordination environments, and neighboring atomic characteristics of amorphous materials and disordered systems. However, due to the complexity of XAFS spectra, their interpretation relies on experienced researchers and is prone to inaccuracies. This study utilizes machine learning methods, specifically neural networks, bagged decision tree models, and random forest models, to analyze XAFS data for predicting the coordination numbers of absorbing elements. The research collected EXAFS data for fourth-period transition metal elements from the Materials Project database, covering various coordination environments, with a total of 13,374 valid data points. The results demonstrate that both neural network and random forest models achieve high accuracy in predicting coordination numbers. By enhancing the generalization capability and interpretability of the models, this study provides a more efficient and reliable method for XAFS data analysis.

Full Text

Machine Learning Approach for Predicting Coordination Numbers from EXAFS Spectra

Haitao Zeng[1], Longfei Hu[1], Tao Yao[1]
[1]National Synchrotron Radiation Laboratory, University of Science and Technology of China, Hefei 230029, China

Abstract

[Background] X-ray Absorption Fine Structure (XAFS) is a vital technique for structural analysis, widely employed to investigate the oxidation state, coordination environment, and neighboring atom properties of amorphous materials and

disordered systems. However, the complexity of XAFS spectra often requires interpretation by experienced researchers, which can still lead to inaccuracies. [Purpose] This study aims to use machine learning approaches to analyze XAFS data and predict the coordination number of absorbing atoms. [Methods] First, a dataset of 13,374 valid EXAFS spectra of fourth-period transition metal elements was sourced from the Materials Project database. Second, this data was utilized to train three machine learning models: neural networks, bagging models, and random forest models. Finally, these models were applied to predict coordination numbers of the absorbing atoms in the spectra. [Results] The study achieves an average prediction accuracy of approximately 70%. Feature importance analysis reveals that data points within $R < 0.3$ nm are critical for predictions, consistent with the prominence of short-range atomic interactions in EXAFS theory. [Conclusions] This research enhances the efficiency and reliability of XAFS data analysis by improving model generalizability and interpretability.

Key Words: EXAFS, Coordination Number, Bagging, Random Forest

Supported by National Natural Science Foundation of China (No. 12025505)

First author: ZENG Haitao, male, born in 1997, graduated from University of Science and Technology of China in 2020, master student, focusing on nuclear science and technology

Corresponding author: YAO Tao, E-mail: yaot@ustc.edu.cn

Introduction

X-ray Absorption Fine Structure (XAFS) can be divided into two main regions: X-ray Absorption Near Edge Structure (XANES) and Extended X-ray Absorption Fine Structure (EXAFS). XAFS is highly sensitive to the oxidation state of absorbing atoms, their coordination environment, and the types and distances of neighboring atoms, making it widely applicable across physics, chemistry, materials science, and biology. However, due to the complexity of XAFS spectra, their interpretation typically relies on experienced researchers, and even then, inaccurate conclusions may be drawn [1]. Therefore, developing more efficient and reliable analytical methods to improve the accuracy of XAFS data interpretation has become an important research direction.

Machine learning has emerged in recent years as a significant data-driven research method in scientific research. Neural networks, in particular, have attracted attention for their high prediction accuracy and have been widely applied to EXAFS analysis and XANES feature extraction. For example, Tetef et al. (2021) demonstrated that unsupervised methods such as t-distributed Stochastic Neighbor Embedding (t-SNE) and Variational Autoencoders (VAE) can effectively classify sulfur organic compounds based on XANES and valence-to-core X-ray emission spectroscopy (VtC-XES), revealing detailed chemical

properties such as oxidation state and aromaticity [2]. Timoshenko et al. (2018) developed a neural network-based method to directly extract radial distribution functions (RDF) from EXAFS spectra without relying on prior structural assumptions, successfully applying it to analyze high-temperature bcc-fcc phase transitions in iron (Fe) and extending it to cobalt (Co) and nickel (Ni) systems [3].

However, existing studies are often limited to simple material systems analyzing specific materials or categories under varying parameters (such as temperature changes), which faces challenges in model generalization. Furthermore, due to the complexity and "black box" nature of neural networks, researchers often struggle to interpret the specific basis for model-generated parameters and cannot intuitively understand the decision-making process. This lack of interpretability limits the application of neural networks in research scenarios requiring explicit feature importance analysis [4]. To improve the interpretability of machine learning models, researchers have developed various techniques, including Ridge Regression (RR) [5], LASSO regression [6], the SISSO method [7], and Decision Trees (DT) [8]. Among these, decision trees have gained widespread attention due to their simple structure and high interpretability. Additionally, based on Ensemble Learning (EL) [9], researchers have further developed Bagging [10], Random Forest (RF) [11], and Boosting [12] methods. Decision tree models have already demonstrated potential in XAFS spectroscopy research. For instance, Torrisi et al. (2020) used random forest machine learning models to analyze XANES spectra, predicting coordination numbers, nearest-neighbor distances, and Bader charges of absorbing atoms in transition metal oxides, and improved model interpretability through multi-scale polynomial featurization to reveal key spectral regions related to different properties [13].

Addressing the two core constraints of machine learning in XAFS analysis—model generalization difficulties and lack of interpretability—this study constructed a large-scale EXAFS dataset covering multiple fourth-period transition metal elements across different material categories and diverse coordination environments using the Materials Project database. Based on this dataset, the developed neural network and random forest models demonstrated generalization capabilities for predicting coordination numbers with high accuracy. Additionally, interpretability analysis of the random forest model successfully identified the spectral feature region that determines coordination numbers: the low-R region after Fourier transformation. This study provides a scalable, highly generalizable XAFS data analysis tool, laying a solid foundation for high-throughput, automated XAFS structural analysis across broad material systems, with significant methodological value and practical prospects.

The workflow of this study is illustrated in Figure 1 [Figure 1: see original paper]. We collected a total of 15,498 XAFS data entries for fourth-period transition metal elements from the Materials Project, retaining 13,374 entries with integer coordination numbers after screening. Subsequently, we transformed these XAFS data from E-space to R-space and imported them into fully-connected

neural networks and decision tree-based bagging and random forest ensemble learning models. For each element, we trained and evaluated models separately.

1.1 X-ray Absorption Fine Structure (XAFS) and Extended X-ray Absorption Fine Structure (EXAFS)

The XAFS phenomenon was first discovered and scientifically documented in experiments by Friche and Hertz in 1920. A breakthrough in XAFS research emerged from the innovative method proposed by Sayers, Stern, and Lytle—using Fourier transform techniques to effectively analyze the wave vector space of EXAFS oscillation signals [14]. This milestone work marked the entry of XAFS technology into the quantitative analysis domain and initiated its systematic development for materials characterization [15].

XAFS technology provides detailed structural information about the local chemical environment around absorbing atoms. In recent years, atomic-level local property analysis based on XAFS has become an important approach in materials science and catalysis research. Specifically, XAFS techniques have been applied to investigate the coordination environment of metal clusters, explore local structural distortions in ferroelectric materials, and analyze catalytic mechanisms in single-atom catalysts. Yang et al. used EXAFS to study the adsorption of As(V) on goethite and hematite, revealing the molecular structure of complexes formed between As and iron oxides [16]. Chen et al. designed a fluorescence gas ionization chamber detector for synchrotron radiation XAFS fluorescence measurements, exceeding the performance metrics of imported equipment [17].

XAFS spectra are obtained by measuring the variation of a material' s X-ray absorption coefficient with incident photon energy. These spectra are typically divided into two characteristic regions: X-ray Absorption Near Edge Structure (XANES) and Extended X-ray Absorption Fine Structure (EXAFS). The theoretical expression for EXAFS can be written as:

$$\chi(k) = S_0^2 \sum_j \frac{N_j f_j(k)}{kR_j^2} e^{-2R_j/\lambda(k)} \sin[2kR_j + \delta_j(k)]$$

where k represents the photoelectron wave vector, N_j denotes the coordination number of the j-th coordination shell, R_j represents the average distance from the absorbing atom to neighboring atoms in the j-th shell, σ_j^2 characterizes the disorder of the j-th shell, $f_j(k)$ represents the scattering amplitude of neighboring atoms for photoelectrons, $\delta_j(k)$ denotes the phase shift, $\lambda(k)$ represents the mean free path of photoelectrons, and S_0^2 is the amplitude reduction factor.

The above theoretical expression indicates that Fourier transform can convert the EXAFS oscillation function from k-space to R-space, thereby obtaining radial distribution information about the local structure around absorbing atoms.

However, since the EXAFS formula contains multiple coupled variable parameters and possible superposition effects from multiple scattering paths, precise structural analysis relying solely on EXAFS spectra faces severe ill-posed problems. Therefore, in practical EXAFS data analysis, researchers typically adopt a "fingerprint comparison" approach, obtaining relative models of material local structures by comparing experimental data with reference spectra from known structures or theoretical simulations.

Although directly inverting all structural parameters from XAFS spectra presents inherent difficulties, calculating theoretical XAFS spectra based on known structural models has become a reliable research method. In XAFS research, FEFF [18] and FDMNES [19] are two widely adopted computational software packages. Both are based on first-principles (ab initio) calculation methods, simulating XAFS spectral features by solving the Schrödinger equation. FEFF employs Real-Space Multiple Scattering (RSMS) theory, while FDMNES is based on the Finite Difference Method (FDM), providing reliable theoretical references for experimental data analysis.

1.2 Larch [20]

Larch is a comprehensive toolkit specifically designed for XAFS and related spectroscopic data analysis. Compared to the IFEFFIT software package developed in FORTRAN, Larch' s reconstruction in Python significantly enhances large-scale data processing capabilities and data visualization functions. In this study, we utilized core functions from the Larch toolkit for data processing: the autobk function for background subtraction and normalization, converting EXAFS spectra from E-space to k-space; and the xftf function for performing Fourier Transform (FT) to convert EXAFS spectra from k-space to R-space.

1.3 Coordination Number

Coordination number is a key structural parameter describing the local coordination environment of atoms in materials and is crucial for studying material properties. In traditional research, coordination numbers were typically determined through intuitive judgment of coordination situations for individual atoms in materials. However, with the development of high-throughput materials computation technology and the generation of large-scale datasets, developing reliable and automated coordination number calculation algorithms has become an urgent research need.

Currently, various coordination number calculation methods have been proposed in materials science [21]. Among them, simpler algorithms are based on empirical parameters for interatomic distances and related tolerances, such as BrunnerNN [22] and MinimumOKeeffeNN [23]. These methods determine coordination numbers by comparing actual interatomic distances in models with their tolerance values. Although these algorithms are relatively simple and intuitive, they are sensitive to minor atomic perturbations and changes in empirical

tolerance parameters, where small atomic variations may lead to different coordination number assignments.

The VoronoiNN algorithm proposed by O' Keeffe [24] provides an alternative solution. Based on geometric principles, this algorithm employs Voronoi decomposition [25] to treat the environment around a central atom as a polyhedron and further performs weighted processing according to spatial angles to determine the number of neighboring atoms. Similarly, this study uses the CrystalNN [26] algorithm. The CrystalNN method is also based on Voronoi decomposition and is widely recognized for its high computational accuracy. This method calculates probabilities of possible coordination environments through Voronoi decomposition and ultimately selects the result with the highest probability as the atom' s coordination environment. We systematically calculated coordination numbers of absorbing atoms in various material systems using the CrystalNN algorithm from the pymatgen materials computation software package [27].

1.4 Neural Networks

As an important branch of machine learning, neural networks (deep learning) have become one of the core areas in current scientific research and technological applications due to their excellent performance in regression and classification tasks. Neural network models achieve feature extraction and pattern recognition through numerous adjustable parameters that are automatically optimized during training via backpropagation algorithms. Researchers primarily optimize model performance by tuning hyperparameters including batch size, learning rate, and network layer structure.

A single hidden layer feedforward neural network model can be formally defined by the following mathematical expression:

$$f(X) = \beta_0 + \sum_k \beta_k g\left(w_{k0} + \sum_j w_{kj} X_j\right)$$

where $X = (X_1, X_2, \ldots, X_p) \in \mathbb{R}^p$ represents a p-dimensional input feature vector. Model parameters include: w_{kj} characterizing connection weights from the j-th feature in the input layer to the k-th neuron in the hidden layer, β_k representing weight coefficients from the k-th neuron in the hidden layer to the output layer, and w_{k0} and β_0 being bias terms for corresponding linear transformations.

Although this model architecture appears as a linear superposition in form, nonlinear activation functions $g(\cdot)$ must be introduced to effectively approximate nonlinear functional relationships in real systems. In machine learning, the Sigmoid function and ReLU (Rectified Linear Unit) [28] function are two classic activation functions. The Sigmoid function has S-shaped saturation characteristics with an output range of (0,1), which made it widely used in early neural network research. However, this function suffers from vanishing gradients when

input values approach positive or negative infinity, where derivatives approach zero, significantly reducing parameter update efficiency during backpropagation. In contrast, the ReLU function is defined as $g(x) = \max(0, x)$, and its piecewise linear characteristics ensure a constant derivative value in the positive interval, effectively alleviating the vanishing gradient problem. Moreover, the computational complexity of the ReLU function is significantly lower than that of the Sigmoid function, enabling it to demonstrate superior computational efficiency and convergence characteristics in deep neural network training.

1.5 Decision Trees

Although neural networks offer significant advantages in prediction performance, their inherent "black box"characteristics result in insufficient model interpretability. Specifically, the mechanism for setting numerous neuron parameters in neural networks is difficult to interpret clearly, prompting researchers to seek other auxiliary analysis tools with stronger interpretability. In this study, we adopt decision trees as an analysis tool due to their intuitive model structure and ease of interpretation. Through decision trees, researchers can clearly identify associations between data features and results. However, traditional decision tree methods exhibit obvious deficiencies in prediction accuracy. To overcome this limitation, researchers have developed various ensemble learning methods, including Bagging, Random Forest, and Boosting. While these ensemble methods significantly improve prediction performance, they sacrifice the intuitive interpretability of single decision trees. Nevertheless, by analyzing the Gini Index, we can still effectively evaluate the importance of each feature in classification decisions.

The modeling mechanism of decision trees follows a recursive feature space partitioning strategy. The algorithm decomposes a p-dimensional feature space (assuming input samples are p-dimensional feature vectors) into several mutually exclusive geometric subspaces through preset tree structure hyperparameters (such as maximum depth) and supervised learning to extract optimal splitting criteria from training datasets. Theoretical studies show that this model exhibits excellent classification performance on datasets with explicit hyperplane partitioning characteristics, but its classification effectiveness decreases significantly when dealing with nonlinearly separable data or classification problems with complex decision boundaries. However, empirical studies demonstrate that ensemble learning algorithms based on decision tree architecture show superior prediction accuracy compared to neural network models when processing tabular data. Taking the XGBoost algorithm as an example, this optimized gradient boosting framework achieves significant performance breakthroughs in multiple benchmark tests through iterative decision tree integration and regularization strategies [29].

1.6 Materials Databases

In machine learning research for materials science, constructing robust predictive models heavily depends on support from large-scale, high-quality datasets [30]. Current mainstream materials informatics platforms include Open Catalyst [31], Materials Project [32], and the Open Quantum Materials Database (OQMD) [33]. This study selected Materials Project as the data source primarily based on the following considerations: the database integrates structural information for over 160,000 materials and provides XAFS spectral data based on FEFF theoretical calculations, which is valuable for establishing structure-property relationships. Moreover, its open Python Application Programming Interface (API) significantly improves data acquisition efficiency and enables seamless integration with high-throughput computational workflows. Data acquisition was performed through the pymatgen materials analysis toolkit, and all processed structured data were stored in a MongoDB document database for efficient retrieval.

According to literature referenced in the official Materials Project documentation, convergence checks and optimization tests were performed on input fields in FEFF9. In convergence checks, researchers varied the rfms1 value from 0.2 nm to 0.8 nm in 0.1 nm increments to change the self-consistent potential automatic calculations in FEFF. The self-consistent potential is controlled by the Self-Consistent Field (SCF) card. Simultaneously, the rfms value was varied from 0.3 nm to 1.1 nm in 0.1 nm increments to determine parameters for the Full Multiple Scattering (FMS) card.

2 Dataset and Preprocessing

The data in this study were sourced from the Materials Project database, where we systematically retrieved materials containing fourth-period transition metals and extracted EXAFS calculation data. All EXAFS data were generated through the FEFF computational software package. To ensure data quality, we excluded samples where different absorbing atoms with varying coordination numbers existed within the same material, as such data could interfere with coordination number classification.

In this study, we used theoretically calculated EXAFS data from the Materials Project database rather than noise-added calculated EXAFS data. According to research by Paolo Ghigna [34] and Nicholas Marcella [35] et al., the impact of adding noise to calculated data is smaller than that of systematic factors (such as background subtraction and other data processing steps). Therefore, we could directly use the calculated EXAFS data for machine learning.

The database contains materials from various complex systems, such as oxides, sulfides, and alloys. While this enhances model generalization capability, it poses challenges for classification. Moreover, overly complex classifications dilute prediction accuracy for individual categories. Therefore, this study focuses solely on classifying absorbing atoms.

For periodic EXAFS data, we employed Fourier transform for preprocessing. For large-scale data processing, we used the Larch toolkit combined with Python programming language to achieve automated processing. Specifically, we utilized Larch' s autobk function to convert raw E-space data to k-space data, followed by the xftf function to Fourier transform k^2-weighted k-space data into R-space data.

For R-space data processing, we extracted intensity values at intervals of 0.003 nm as features. Additionally, we used Python' s Peak package to calculate R-coordinates and intensity values corresponding to each characteristic peak, incorporating these parameters as additional features into the dataset. For comparison, we also trained neural network models using k-space data and wavelet-transformed data. The wavelet transform employed the cauchy_{wavelet} function from the Larch toolkit with parameters set to kweight = 0, rmax_{out} = 6, transforming k-space data from k_{min} = 3 to k_{max} = 14.

3 Model Construction and Training

To accomplish coordination number prediction tasks, this study constructed a prediction model based on fully-connected neural networks. As shown in Figure 2 [Figure 2: see original paper], the neural network' s input layer contains 326 features corresponding to intensity values of EXAFS R-space spectra at intervals of 0.003-0.0031 nm. The network employs a dual hidden-layer structure with 512 neurons in each layer. Each hidden layer is followed sequentially by a ReLU activation function layer and a Dropout regularization layer: the ReLU activation function enhances the model' s nonlinear expression capability, while the Dropout regularization mechanism mitigates overfitting. The output layer parameter range varies according to the coordination number range corresponding to different elements. During model optimization, the model with optimal validation set performance was selected as the final model through five-fold cross-validation, effectively improving model generalization capability.

To enhance model interpretability and verify neural network prediction results, this study simultaneously constructed ensemble learning models based on decision trees. First, we implemented two ensemble models using the Scikit-learn [36] machine learning framework: decision tree ensemble based on Bagging and Random Forest. Bagging generates sub-training sets equal in size to the original dataset through Bootstrap Sampling, trains independent decision tree models on each subset, and integrates predictions through a voting mechanism. Random Forest introduces feature randomness on top of Bagging—during each node split, only sqrt(d) features (where d is the total number of features) are randomly selected for optimal partitioning, thereby enhancing model diversity. Both models use the Gini Index as the splitting criterion. As shown in Figure 3 [Figure 3: see original paper], each box in the diagram represents a split node. "Freq_r" in the box indicates the intensity corresponding to the abscissa r in R-space of the EXAFS spectrum, while "samples" represents the amount of data contained in that node before splitting.

This study systematically analyzed EXAFS spectra of fourth-period transition metal materials using neural networks and decision tree algorithms. As shown in Figure 4 [Figure 4: see original paper], for each metal, the models from left to right are: neural network model, Bagging model, Random Forest model, Bagging model combined with peak height and position, and Random Forest model combined with peak height and position. The results indicate that both methods achieve prediction accuracies around 70%, with neural network models showing slightly better performance than decision tree models. Notably, the accuracies of both methods show significant correlation, with values tending to increase or decrease together across different elements. In ensemble learning applications, Random Forest models achieve approximately 1-3 percentage points higher prediction accuracy than Bagging models. Additionally, this study examined feature datasets incorporating peak position and height information, finding that decision tree models showed no significant accuracy difference compared to using the original dataset.

Results demonstrate that model prediction accuracy exhibits significant element dependence in coordination number classification. Vanadium (V) shows the best prediction performance in the neural network framework, reaching 81.74% accuracy, while Co exhibits relatively low prediction accuracy at only 67.39%. Further analysis reveals that this difference may be related to the distribution characteristics of coordination numbers for each element in the database. Specifically, V maintains a relatively stable six-coordination structure in most materials, while Co shows more diverse coordination number variations. This may be the main reason for its relatively low prediction accuracy.

As shown in Figure 5 [Figure 5: see original paper], we plotted prediction accuracy against information entropy of coordination number distribution for each element. The scatter plot uses least squares linear regression fitting to reveal the relationship between these variables, with the R-squared value indicating the strength of the fit. The results show that prediction accuracy exhibits certain correlation with coordination number distribution characteristics.

We also compared prediction accuracies based on k-space and q-space (wavelet transform) data. As shown in Figure 6 [Figure 6: see original paper], prediction accuracies vary across the three spaces, but the overall trend is similar, with prediction rates showing common rising or falling patterns across elements. R-space prediction accuracy is superior to k-space and q-space in most cases, leading us to recommend using R-space for neural network learning.

To demonstrate model bias characteristics, we used Zn element as an example to create a confusion matrix comparing true and predicted coordination numbers.

This study further constructed feature importance distribution maps for EXAFS spectra of each element based on the Gini Index from decision tree ensemble models. Figure 7 [Figure 7: see original paper] shows the feature importance and its correspondence with R-space for Co and Mn elements in Bagging and Random Forest models, where (a) represents Bagging Gini Index for Co, (b)

represents Random Forest Gini Index for Co, (c) represents Bagging Gini Index for Mn, and (d) represents Random Forest Gini Index for Mn. Since the R-space dataset contains intensity values at 326 equally spaced coordinate points, the maximum single feature importance value is generally below 0.1. Based on standard EXAFS analysis practices, we only discuss the region where $R < 0.6$ nm. Through comparative analysis of relative importance distribution, we find that feature importance weights are mainly concentrated in the short-range interaction region of $R < 0.3$ nm. This aligns with the empirical understanding in existing EXAFS theory that short-range atomic interactions dominate spectral features.

For distant coordination shells with $R > 0.6$ nm, current models struggle to reliably establish direct correspondence between coordination numbers of high coordination shells and R-space parameters. Because high-shell signal intensities are relatively low, prediction errors increase significantly, failing to meet accuracy requirements. In practice, we also find it difficult to obtain coordination information regarding high coordination shells.

Furthermore, this study observed that Bagging significantly amplifies importance differences among features in feature importance distribution maps, while feature importance distribution in Random Forest models is relatively more balanced. This phenomenon may be related to the overfitting tendency of Bagging models. Specifically, Bagging removes only one feature for optimization in each iteration, causing features with higher importance to be further amplified across multiple iterations. In contrast, the Random Forest model used in this study simultaneously screens 14 features for optimization in each iteration, thereby avoiding excessive enhancement of single feature importance and effectively improving model generalization capability. This finding indicates that Random Forest models can better balance feature importance and reduce overfitting risks when processing high-dimensional data, providing a more robust solution for EXAFS spectral analysis.

Conclusion

This study successfully developed and validated a machine learning-based method for predicting coordination numbers of fourth-period transition metal elements from EXAFS spectra. By collecting large-scale EXAFS data from the Materials Project database and combining neural networks with decision tree algorithms, our models achieved approximately 70% average prediction accuracy in coordination number classification tasks. Notably, neural network models performed exceptionally well on certain elements (such as V), reaching up to 81.74% accuracy, demonstrating their powerful potential for processing complex spectral data.

The results indicate that Random Forest models not only achieve prediction performance comparable to neural networks but also reveal key information in EXAFS spectra through feature importance analysis. Specifically, intensity

information from data points at R < 0.3 nm in Fourier-transformed R-space is crucial for coordination number prediction. This finding is consistent with the perspective in EXAFS theory that short-range atomic interactions dominate spectral features, providing theoretical support and interpretability for the model' s decision-making process.

However, the study also found significant element dependence in model prediction accuracy. For example, Co exhibited lower accuracy (67.39%), possibly related to the diversity of coordination number distributions in the database. Additionally, feature importance analysis revealed that peak position and intensity parameters for certain elements (such as Co) were less important than expected, suggesting that potentially underutilized spectral features may exist. This phenomenon indicates that current feature extraction methods still have room for improvement.

Future research directions should include optimizing feature engineering, exploring more sophisticated spectral feature extraction techniques, and trying other advanced machine learning algorithms to improve prediction accuracy and generalization capabilities across different elements. Meanwhile, hybrid approaches combining physical models with machine learning also warrant further investigation to enhance the efficiency and reliability of EXAFS spectral analysis. The results of this study provide new tools and ideas for rapid analysis of local material structures, holding significant scientific importance and application prospects.

Author Contributions

Haitao Zeng: Experimental design, data collection and computation, manuscript writing; **Longfei Hu:** Proposed experimental framework and participated in manuscript revision; **Tao Yao:** Overall experimental guidance and participated in manuscript revision.

References

[1] Terry J, Lau M L, Sun J, et al. Analysis of Extended X-ray Absorption Fine Structure (EXAFS) Data Using Artificial Intelligence Techniques[J]. Applied Surface Science, 2021, 547: 149059. DOI: 10.1016/j.apsusc.2021.149059.

[2] Tetef S, Govind N, Seidler G T. Unsupervised machine learning for unbiased chemical classification in X-ray absorption spectroscopy and X-ray emission spectroscopy[J]. Physical Chemistry Chemical Physics, 2021, 23(41): 23586-23601. DOI: 10.1039/D1CP02903G.

[3] Timoshenko J, Anspoks A, Cintins A, et al. Neural Network Approach for Characterizing Structural Transformations by X-Ray Absorption Fine Structure Spectroscopy[J]. Physical Review Letters, 2018, 120(22): 225502: 1-6. DOI: 10.1103/PhysRevLett.120.225502.

[4] James G, Witten D, Hastie T, et al. An Introduction to Statistical Learning: with Applications in Python[M]. Springer New York, 2023.

[5] Hoerl A E, Kennard R W. Ridge regression: biased estimation for nonorthogonal problems[J]. Technometrics A Journal of Stats for the Physical Chemical & Engineering Sciences, 2000, 42(1): 80-86. DOI: 10.2307/1271436.

[6] Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2011, 73(3): 267-288. DOI: 10.1111/j.1467-9868.2011.00771.x.

[7] Ouyang R, Curtarolo S, Ahmetcik E, et al. SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates[J]. Phys. Rev. Mater, 2018, 2: 083802. DOI: 10.1103/PhysRevMaterials.2.083802.

[8] Quinlan J R. Induction of Decision Trees[J]. Machine Learning, 1986, 1(1): 81-106. DOI: 10.1007/BF00116251.

[9] Opitz D, Maclin R. Popular Ensemble Methods: An Empirical Study[J]. Journal of Artificial Intelligence Research, 1999, 11: 169-198. DOI: 10.1613/jair.614.

[10] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123–140. DOI: 10.1007/BF00058655.

[11] Ho T K. The random subspace method for constructing decision forests[J]. Transactions on Pattern Analysis & Machine Intelligence, 1998, 20(8): 832-844. DOI: 10.1109/34.709601.

[12] Freund Y, Schapire R. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997, 55(1): 119-139. DOI: 10.1007/3-540-59119-2_{166}.

[13] Torrisi S B, Carbone M R, Rohr B A, et al. Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships[J]. npj Comput Mater, 2020, 6(1): 109. DOI: 10.1038/s41524-020-00376-6.

[14] Sayers D E, Stern E A, Lytle F W. New Technique for Investigating Noncrystalline Structures: Fourier Analysis of the Extended X-Ray—Absorption Fine Structure[J]. Physical Review Letters, 1971, 27(18): 1204-1207. DOI: 10.1103/PhysRevLett.27.1204.

[15] Sun Z, Liu Q, Yao T, et al. X-ray absorption fine structure spectroscopy in nanomaterials[J]. Science China Materials, 2015, 58(4): 313-341. DOI: 10.1007/s40843-015-0043-4.

[16] 杨旺, 林金如, 姚慧等. 针铁矿和赤铁矿吸附 As(V) 的 EXAFS 研究 [J]. 核技术, 2021, 44(12): 1-11. DOI: 10.11889/j.0253-3219.2021.hjs.44.120101.

[17] 陈雨, 王建东, 浦世节等. 用于 X 射线吸收精细结构测量荧光气体探测器的设计与研制 [J]. 核技术, 2024, 47(12): 120401. DOI: 10.11889/j.0253-3219.2024.hjs.47.120401.

CSTR: 32193.14.hjs.CN31-1342/TL.2024.47.120401.

[18] Rehr J J, Kas J J, Vila F D, et al. Parameter-free calculations of X-ray spectra with FEFF9[J]. Physical Chemistry Chemical Physics, 2010, 12(21): 5503-5513. DOI: 10.1039/b926434e.

[19] Joly Y. X-ray absorption near-edge structure calculations beyond the muffin-tin approximation[J]. Physical Review B, 2001, 63(12): 125120. DOI: 10.1103/PhysRevB.63.125120.

[20] Newville M. Larch: An Analysis Package for XAFS and Related Spectroscopies[J]. Journal of Physics Conference Series, 2013, 430: 012007. DOI: 10.1088/1742-6596/430/1/012007.

[21] Pan H, Ganose A M, Horton M, et al. Benchmarking Coordination Number Prediction Algorithms on Inorganic Crystal Structures[J]. Inorganic Chemistry, 2021, 60(3): 1590–1603. DOI: 10.1021/acs.inorgchem.0c02996.

[22] Brunner G O. A definition of coordination and its relevance in the structure types AlB2 and NiAs[J]. Acta Crystallographica, 1977, 33(1): 226-227. DOI: 10.1107/s0567739477000461.

[23] O' Keefe M, Brese N E. Atom sizes and bond lengths in molecules and crystals[J]. Journal of the American Chemical Society, 1991, 113(9): 3226-3229. DOI: 10.1021/ja00009a002.

[24] O' Keeffe M. A proposed rigorous definition of coordination number[J]. Acta Crystallographica Section A, 1979, 35: 772-775. DOI: 10.1107/S0567739479001765.

[25] Voronoi G. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Premier mémoire. Sur quelques propriétés des formes quadratiques positives parfaites[J]. Journal für die reine und angewandte Mathematik, 1908, 133: 97–178. DOI: 10.1515/crll.1908.133.97.

[26] Zimmermann N E R, Jain A. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity[J]. RSC Advances, 2020, 10: 6063–6081. DOI: 10.1039/C9RA07755C.

[27] Ong S P, Richards W D, Jain A, et al. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis[J]. Computational Materials Science, 2013, 68: 314-319. DOI: 10.1016/j.commatsci.2012.10.028.

[28] Agarap A F M. Deep Learning using Rectified Linear Units (ReLU)[J]. arXiv, 2018. DOI: 10.48550/arXiv.1803.08375.

[29] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016, 785–794. DOI: 10.1145/2939672.2939785.

[30] Lee K L K, Gonzales C, Nassar M, et al. MatSciML: A Broad, Multi-Task Benchmark for Solid-State Materials Modeling[J]. arXiv, 2023. DOI: 10.48550/arXiv.2309.05934.

[31] Chanussot L, Das A, Goyal S, et al. Open Catalyst 2020 (OC20) Dataset and Community Challenges[J]. ACS Catalysis, 2021, 11(10): 6059-6072. DOI: 10.1021/acscatal.0c04525.

[32] Jain A, Ong S P, Hautier G, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation[J]. APL Materials, 2013, 1(1): 011002. DOI: 10.1063/1.4812323.

[33] Kirklin S, Saal J E, Meredig B, et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies[J]. npj Computational Materials, 2015, 1: 15010. DOI: 10.1038/npjcompumats.2015.10.

[34] Ghigna P, Muri M D, Spinolo G. Computer simulation approach to reliability and accuracy in EXAFS structural determinations[J]. Journal of Applied Crystallography, 2010, 34(3): 325-329. DOI: 10.1107/S0021889801004745.

[35] Marcella N, Shimogawa R, Xiang Y, et al. First shell EXAFS data analysis of nanocatalysts using neural networks[J]. Journal of Catalysis, 2025, 164145. DOI: 10.1016/j.jcat.2025.116145.

[36] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python[J]. arXiv, 2011. DOI: 10.5555/1953048.2078195.

Note: Figure translations are in progress. See original paper for figures.