ACS Partner Journal

ENVIRONMENT & HEALTH

Open Access

pubs.acs.org/EnvHealth

Article

Metallomic Classification of Pulmonary Nodules Using Blood by Deep-Learning-Boosted Synchrotron Radiation X-ray Fluorescence

Published as part of Environment & Health special issue "Artificial Intelligence and Machine Learning for Environmental Health".

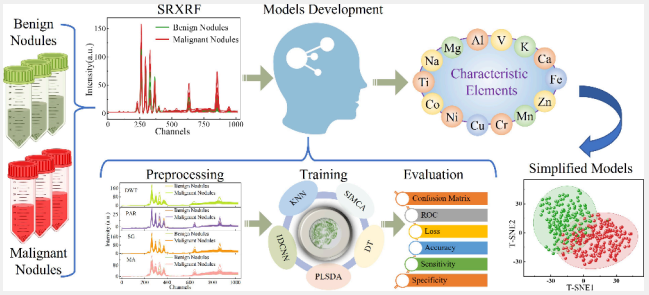
Chaojie Wei,[#] Chao Li,[#] Hongxin Xie,[#] Wei Wang,* Xin Wang,* Dongliang Chen, Bai Li, and Yu-Feng Li*

Cite This: *Environ. Health* 2025, 3, 40−47

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Ambient air pollution is an important contributor to increasing cases of lung cancer, which is a malignant cancer with the highest mortality among all cancers. It primarily manifests in the form of pulmonary nodules, but not all will develop into lung cancer. Therefore, it is highly desired to distinguish between benign and malignant pulmonary nodules for the early prevention and treatment of lung cancer. Currently, histopathological examination is the gold standard for classifying pulmonary nodules, which is invasive, time-consuming, and labor-intensive. This study proposes a metallomics approach through synchrotron radiation X-ray fluorescence (SRXRF) with a simplified one-dimensional convolutional neural network (1DCNN) to distinguish pulmonary nodules by using serum samples. SRXRF spectra of serum samples were obtained and preliminarily analyzed using principal component analysis (PCA). Subsequently, machine learning algorithms (MLs) and 1DCNN were applied to develop classification models. Both MLs and 1DCNN based on full-channel spectra could distinguish patients with benign and malignant pulmonary nodules, but the highest accuracy rate of 96.7% was achieved when using 1DCNN. In addition, it was found that characteristic elements in serum from patients with malignant nodules were different from those in benign nodules, which can serve as the fingerprint metallome profile. The simplified model based on characteristic elements resulted in good performance of sensitivity and F1-score > 91.30%, G-mean, MCC and Kappa > 85.59%, and accuracy = 94.34%. In summary, metallomic classification of benign and malignant pulmonary nodules using serum samples can be achieved through 1DCNN-boosted SRXRF, which is easy to handle and much less invasive compared to histopathological examination.



KEYWORDS: *pulmonary nodules, metallomics, synchrotron radiation X-ray fluorescence, one-dimensional convolutional neural network, serum*

1. INTRODUCTION

Ambient air pollution exerts a significant impact on both the deterioration of ecosystems and the heightened burden of respiratory ailments, which has become a major public health problem.[1] As the primary organ directly exposed to ambient air pollutants, the lungs bear the brunt of the adverse health impacts.[2] Lung cancer, a particularly devastating respiratory disease, is a significant global health threat, with approximately 2.5 million new cases and 1.8 million deaths reported worldwide in 2022.[3] Early diagnosis and treatment can improve prognosis and reduce medical costs for lung cancer patients.[4] Pulmonary nodules, characterized as round or irregular lesions with a diameter of 30 mm or less in the lung, can be benign or malignant.[5] Untreated malignant pulmonary nodules often develop to lung cancer, significantly impacting patient survival.[6] On the other hand, misdiagnosing benign nodules as malignant ones can lead to unnecessary and potentially harmful treatments such as surgery, chemotherapy, or radiation therapy. Currently, commonly used diagnostic methods for pulmonary nodules include endoscopy examinations, medical imaging diagnosis, and histopathological examinations. Endoscopy examinations may not locate all of the nodules. Medical imaging can detect pulmonary nodules but cannot distinguish between benign and malignant ones.[7] Histopathological examination is the gold standard for the diagnosis of malignant nodules, but it is invasive, time-consuming, and labor-intensive.[8] The development of more

Received: June 27, 2024
Revised: August 22, 2024
Accepted: August 26, 2024
Published: September 4, 2024

https://doi.org/10.1021/envhealth.4c00124
Environ. Health 2025, 3, 40−47

effective and less destructive methods for the diagnosis of malignant pulmonary nodules remains a pressing clinical challenge.[9]

The advantages of body fluids, especially serum, in cancer diagnosis lie in their less-invasive, simple, and rapid nature.[10] The discrimination between benign and malignant pulmonary nodules with serum has been achieved through metabolomics and proteomics.[11−14] Metallomics is the systematic investigation of the concentration, speciation, distribution, structure, and function of metals and metalloids in biological systems.[15,16] By studying the metallome difference, it has been applied the screening of neurodevelopmental disorders,[17] pediatric inflammatory bowel disease,[18] cardiovascular diseases,[19] and cancer.[20,21]

Conventional techniques for quantification of the metallome, such as Atomic Fluorescence Spectroscopy (AFS) and Inductively Coupled Plasma Mass Spectrometry (ICP-MS), offer high accuracy and low detection limits, but these methods often require destruction of samples and are time-consuming.[18] On the other hand, X-ray fluorescence, especially synchrotron radiation-based X-ray fluorescence (SRXRF) spectroscopy, allows for simultaneous detection of multiple elements directly in one run with a relatively low detection limit, which is an ideal tool for metallomics study. We recently developed a nontargeted metallomics method based on SRXRF spectra with machine learning algorithms (MLs) to screen cancer patients, which is rapid and accurate.[22]

Data mining is essential in metallomics.[23] Principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) can achieve data dimension reduction and clustering.[24,25] PCA is suitable for global feature extraction and dimension reduction, while t-SNE is mainly used for visualization. MLs including support vector machine (SVM), k-nearest neighbor (KNN), and Bayesian Network (BN) have been applied in the identification and screening of breast cancer,[26] lung cancer,[27] and gastric cancer.[28] However, these methods rely heavily on high-quality data, effective preprocessing, and feature extraction.[29] Deep learning exhibited automatic feature extraction[30] and better classification performance[31] to address these shortcomings. The one-dimensional convolutional neural network (1DCNN) with low computational complexity effectively captures features for spectra.[32] Therefore, in comparison with MLs, the 1DCNN was applied with SRXRF to explore its potential for end-to-end data mining and discrimination of pulmonary nodules.

The aim of this study was to develop an easy-to-handle and much less invasive method compared to histopathological examination for the identification of malignant pulmonary nodules using serum samples with SRXRF data mining as shown in Figure 1. Blood from patients with pulmonary nodules was collected, and then serum samples were separated from the blood. SRXRF spectra of serum were obtained and preprocessed to establish machine learning and deep learning classification models. The characteristic elements that distinguish pulmonary nodules were analyzed, and the model results were evaluated.

2. EXPERIMENTAL SECTION

2.1. Patients Recruitment and Sample Preparation

This study was approved by the Ethics Committee of the Second Affiliated Hospital of Anhui Medical University (No. YX2023−193), and informed consent was obtained from each participant before blood collection. A total of 60 patients participated and were recruited

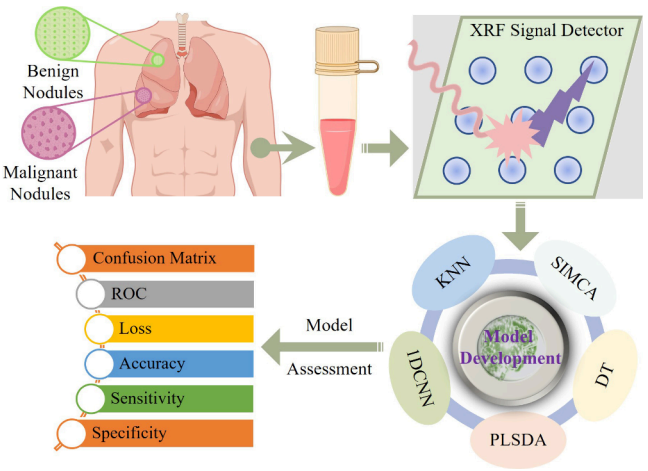


Figure 1. Scheme for classification of benign and malignant pulmonary nodules using serum samples.

in this study and were divided into the two groups: benign pulmonary nodules group (21 cases) and malignant pulmonary nodules group (39 cases). The criteria below were followed:

(1) Participants were detected with pulmonary nodules from CT, and their clinical data were complete; (2) participants did not have a lung cancer history, pulmonary tuberculosis, or other malignant tumors; (3) all the nodules were determined by pathological results, while some benign pulmonary nodules were clinically diagnosed, with size remaining stable after 3−5 years of follow-up; and (4) individuals who had both types of nodules were not recruited in this study.

There was no significant difference in the gender and age between the two groups. Serum samples were obtained from fasting venous blood by centrifugation at 3000 rpm for 10 min.

2.2. SRXRF Spectra Collection

A 60 μL portion of serum was dropped onto filter paper, air-dried, and placed on a precision translation stage for SRXRF data acquisition. SRXRF spectra were collected at the 4W1B beamline of the Beijing Synchrotron Radiation Facility (BSRF), which operates at 2.5GeV with a current of 250 mA. A polychromatic beam (pink beam) with energy of 10−18 keV was used as the incident X-ray. A four-element Hitachi Vortex-ME4 silicon drift detector coupled to a Quantum Detectors Xspress3 multichannel analyzer system was used to collect XRF spectra. The serum was excited by the incident X-ray excitation source with a spot size of 50 μm, and 121 spectra were collected for each sample by manipulating the translation stage.

2.3. SRXRF Spectral Preprocessing and Dimension Reduction

The fluorescence characteristic peaks of almost all elements were in the channel range between 1 and 1024, which were selected from original SRXRF spectra of 4096 channels for further analysis. Prior to the development of classification models, various preprocessing methods, such as discrete wavelet transform (DWT), Pareto scaling (PAR), Savitzky-Golay smoothing (SG), and moving average filtering (MA), were applied to the spectral data. DWT separates interference components, such as system errors and baseline drift. PAR scales the range and maintain the original structure of the data, which was calculated according to eq 1, where x_{ij} and $\tilde{x}_{ij}$ represent the data before and after PAR, and $\bar{x}_i$ and s_i refer to the mean value and standard deviation of the i-th spectra, respectively. SG reduces noise and extracts trends through polynomial fitting and sliding windows. In this study, a cubic polynomial and a 15-point window were utilized for the smoothing process. MA is instrumental in reducing the noise and smoothing data. A window size of 15 points was used to calculate the average value, which was substituted for the central data point. To reduce noise and maintain peak shape, SG and MA were applied to the SRXRF spectra.

https://doi.org/10.1021/envhealth.4c00124
Environ. Health 2025, 3, 40−47

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}} \quad (1)$$

PCA reveals the intrinsic elemental differences of SRXRF in serum and performs dimensionality reduction and clustering. This study applied PCA to the averaged spectra to preliminarily explore the distribution between benign and malignant pulmonary nodules in low dimensional space. Additionally, t-SNE reduces distances between points with high similarity and increases distances between points with low similarity, visualizing high-dimensional data in a low dimensional space.

2.4. Model Development

Partial least-squares discriminant analysis (PLSDA) and soft independent modeling of class analogy (SIMCA) are supervised learning models for handling high-dimensional correlated data. PLSDA transforms the data into a new space while maintaining strong correlation and minimizing the squared error, which improves the performance of linear discrimination and probability calculation. In contrast, SIMCA constructs separate principal component subspaces for each class and determines the class label of a test sample by calculating the squared distance between the projected sample and the centroid of each class subspace. On the other hand, decision tree (DT) adopts a bottom-up hierarchical structure, using features to recursively partition the data, ultimately yielding the classification results. To avoid overfitting, pruning techniques were employed, and Gini coefficients were used as the splitting criteria. In contrast to the feature-based PLSDA, SIMCA and DT, KNN is an instance-based learning algorithm. It finds the K nearest labeled samples to the unknown sample and determines the class label of the unknown sample based on majority voting. In this study, the Euclidean distance was used as the similarity metric for KNN.

1DCNN has demonstrated efficacy in capturing and extracting features from SRXRF data,[23] which is characterized by high dimensionality and intricate correlations. Compared to previous studies, the model employed an average pooling layer (AvgPool) in the last layer, allowing for the input of spectra of varying sizes. Additionally, the model structure discarded the maximum pooling layer between convolutional layers for retaining more features. The basic structure, as depicted in Figure 2, consisted of three stacked

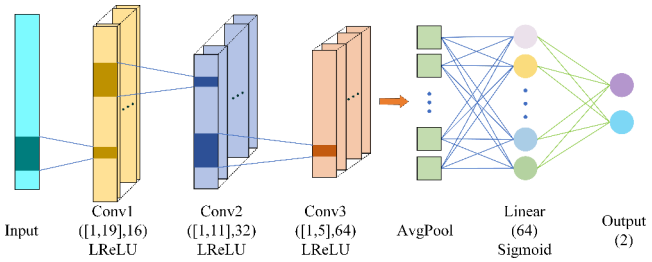


Figure 2. Architecture of the 1DCNN model.

convolution layers (labeled as Conv1, Conv2, and Conv3), one AvgPool, and one fully connected layer (labeled as Linear). The three convolution kernel sizes were 19, 11, 5, with strides of 6, 4, and 3, respectively. Normalization layers and nonlinear activation functions were applied after each convolutional layer to enhance the model's performance. Specifically, the Leaky Rectified Linear Unit (LReLU) was adopted as the activation function, as its calculation formula (eq 2) allows for enhanced nonlinear response. To convert the model output into probabilities corresponding to benign and malignant nodules, the sigmoid function was used. Additionally, the labels of the nodules were converted into a one-hot encoding, which stores the n-bit state in the form of 0 and 1. This transformation facilitated the model's learning and classification of the nodule types.

$$f(x) = \begin{cases} \alpha x, & x < 0 \\ x, & x \geq 0 \end{cases} \quad (2)$$

where α refers to the Leaky coefficient with a value of 0.01.

During the development of the 1DCNN model, the batch size was set to 256, the initial learning rate to 1×10^3, and the epoch of iterations to 600. The model utilized Binary Cross Entropy (BCE) loss as loss function. The Adaptive Moment Estimation (Adam) optimizer was employed. In order to achieve better convergence during the training process, the weight attenuation coefficient β_1 of the Adam momentum term was set to 0.9, and the attenuation coefficient of the learning rate was set to 0.988. A total of 660 spectra were randomly partitioned into calibration set with 462 spectra and validation set with 198 spectra. The quantity ratio in the calibration set and validation set was consistent with the sample proportion of benign and malignant nodules. Cross validation is used to select the optimal hyperparameters and evaluate its performance on different data subsets.

2.5. Model Performance Assessment

Accuracy (eq 3) is calculated as the ratio of correctly classified samples to the total number of samples in each data set. Sensitivity (eq 4) represents the proportion of correctly classified patients with malignant and benign nodules. For imbalanced samples, comprehensive evaluation indicators including Matthews correlation coefficient (MCC), Geometric-mean (G-mean), F1-score, and Kappa coefficient have been introduced into for model evaluation, and their calculation formulas are shown in eqs 5−8.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \times 100\% \quad (3)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \quad (4)$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (5)$$

$$\text{G-mean} = \sqrt{\frac{\text{TP}}{\text{TP} + \text{FN}} \times \frac{\text{TN}}{\text{TN} + \text{FP}}} \quad (6)$$

$$\text{F1-score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \quad (7)$$

$$\text{Kappa} = \frac{p_o - p_e}{1 - p_e} \quad (8)$$

where TP represents malignant nodules correctly identified as positive, TN represents benign nodules correctly identified as negative, FN represents malignant nodules misclassified as negative, and FP represents benign nodules misclassified as positive. In eq 8, p_0 is the accuracy rate of correctly predicted samples, and p_e is the ratio of the sum of product of the actual and predicted values of benign and malignant nodules to the square of the total sample number.

The receiver operating characteristic (ROC) curve plotted with true positive rate on the vertical axis and false positive rate on the horizontal axis was achieved by varying different segmentation thresholds. The curve allows for a direct comparison of the generalization performance of the model. The closer the curve is to the upper left corner, the higher the accuracy of the model. Aera Under the Curve (AUC) is the area enclosed by the ROC curve and the coordinate axis. A higher AUC indicates greater authenticity in the detection method. When AUC approaches 1.0, the method's authenticity is higher, while values close to 0.5 suggest limited practical value.

3. RESULTS AND DISCUSSION

3.1. Comparative Analysis of SRXRF Spectra

Figure 3 shows the total raw and averaged SRXRF spectra of serum samples from patients with benign or malignant pulmonary nodules. The spectral patterns from both groups

https://doi.org/10.1021/envhealth.4c00124
Environ. Health 2025, 3, 40−47

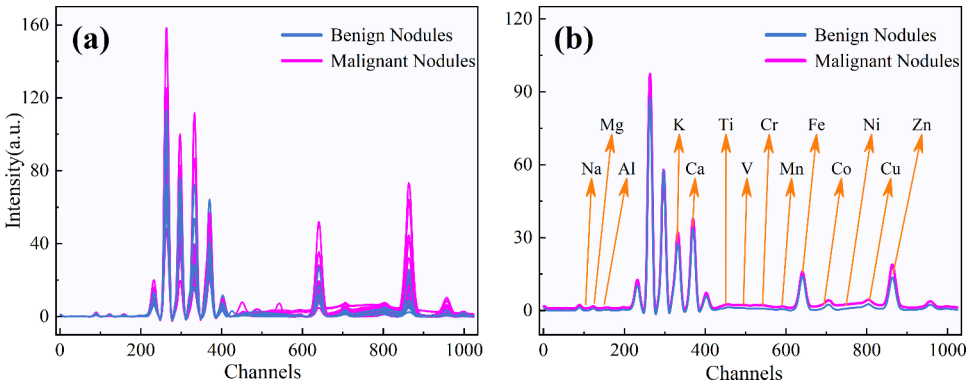


Figure 3. SRXRF spectra of benign and malignant pulmonary nodules: (a) total raw and (b) averaged spectra.

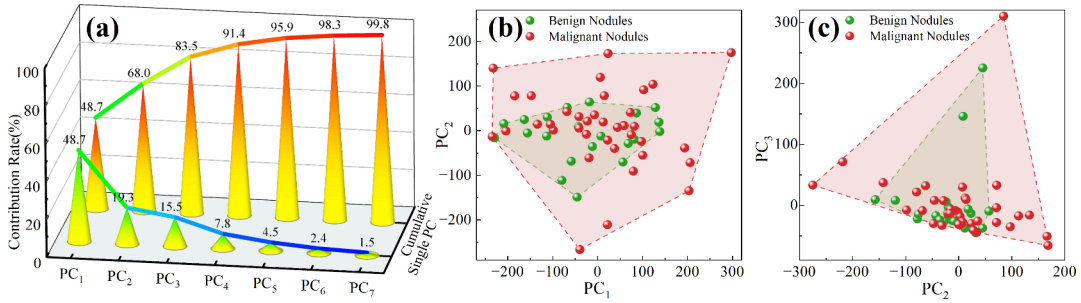


Figure 4. PCA for SRXRF spectra with different nodules: (a) histogram of PC contribution rate, (b) PCA score plots of PC_1 vs PC_2, and (c) PC_2 vs PC_3.

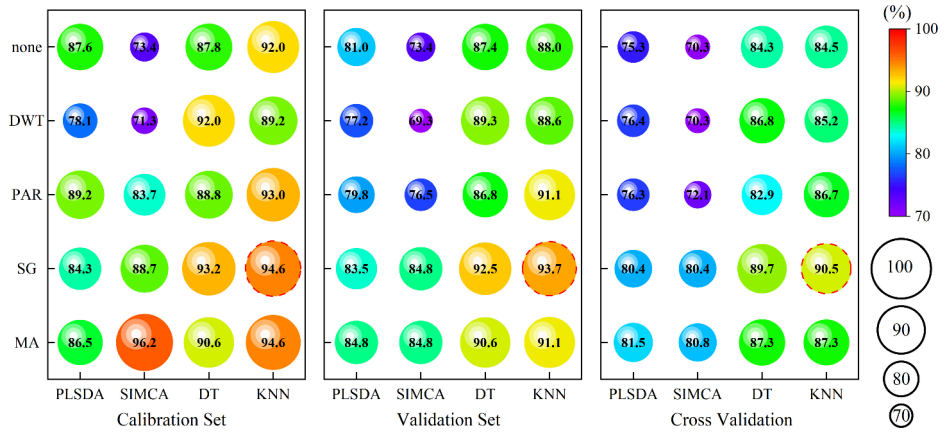


Figure 5. Accuracies of machine learning models based on preprocessed spectra.

are basically consistent, but the peak areas vary for 14 elements, indicating disparities in metal element concentrations within the serum between malignant and benign nodules. Unmarked peaks observed between Al and K may be attributed to Rayleigh Scatter and Compton Scatter.[33]

3.2. Explanatory Classification Analysis by PCA

Figure 4 illustrates the contribution rates of PCs and the PCA score plots of 60 samples. As depicted in Figure 4a, the first 7 PCs accounted for 99.8% variance in the original data. Especially, the conversion of PC_1 to PC_3 explained a significant portion of the original information, reaching up to 83.5%. Consequently, the clustering information on the first three PCs was analyzed in detail. From the score plot of PC_1 vs PC_2 shown in Figure 4b, it can be observed that patients with malignant nodules exhibited a higher dispersion, while benign nodules tended to cluster together. Interestingly, the benign nodules were situated within the scatter of the malignant nodules. Figure 4c displays the score plot of PC_2 vs PC_3, and the clustering pattern of nodules in the two categories aligns with the score plot of PC_1 vs PC_2, with benign nodules surrounded by the clusters of malignant nodules. This phenomenon could potentially be attributed to the presence of cancerous cells spreading into the blood of patients with malignant nodules, leading to elemental abnormalities.[34] Consequently, it can be deduced that PCA holds the potential to distinguish partially malignant nodules from benign ones, although additional methods are still required for further refinement.

3.3. Full-Channel Classification Models

Four MLs (SIMCA, PLSDA, DT, and KNN) and the 1DCNN model were used to classify benign and malignant pulmonary nodules based on full-channels of SRXRF spectra.

https://doi.org/10.1021/envhealth.4c00124
Environ. Health 2025, 3, 40−47

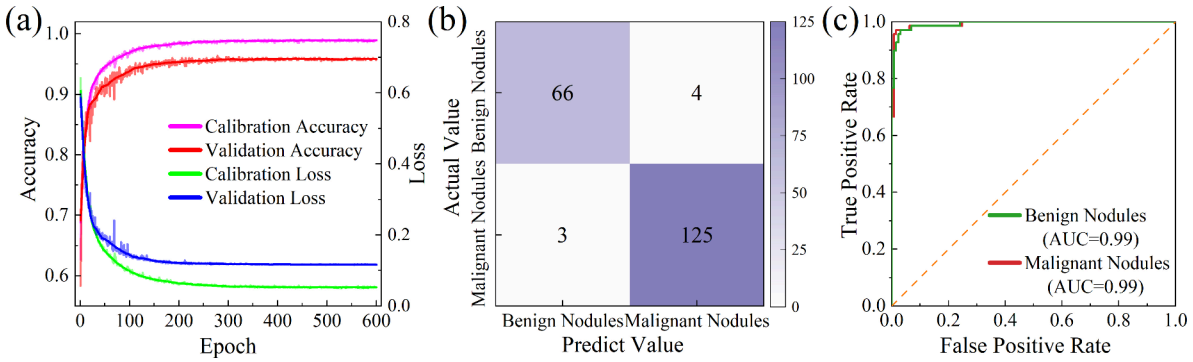


Figure 6. Performance of 1DCNN model. (a) Accuracy and loss curves, (b) confusion matrix, and (c) ROC curve.

3.3.1. Classification Models with ML. The preprocessed spectra are shown in Figure S1. The weak peaks in the spectra after DWT disappear, and the relative peak intensity after PAR changes compared to the spectra without preprocessing. The spectra after SG and MA maintain the same shape and achieve denoising. Figure 5 shows the accuracy of MLs with both raw and preprocessed full-channels spectra. In terms of pretreatment methods, the accuracies of DWT and PAR are comparable to or even lower than those without pretreatment in certain cases. These unexpected results could be attributed to the loss of crucial information caused by improper preprocessing. On the other hand, SG and MA preprocessing methods exhibited improved accuracy suggesting that noise present in the spectra can negatively impact the model's accuracy, and smoothing operations appear to be more suitable preprocessing methods.

In this study, four MLs were employed to classify benign and malignant pulmonary nodules. The PLSDA model achieved the highest accuracy of 84.8% when the MA was applied. Similarly, the SIMCA model also attained the same accuracy with either SG or MA preprocessing. Notably, the DT model exhibited the best performance with an accuracy of 92.5% when SG preprocessing was utilized, showcasing an improvement compared to the previous two models. Furthermore, the KNN model achieved the highest accuracy of 93.7% with the SG, representing the most favorable result among the MLs.

The ROC curves (Figure S2a) and confusion matrices (Figure S2b) were obtained with the optimal preprocessing method. The ROC curves in Figure S2a depict the performance of the PLSDA, SIMCA, DT, and KNN models. The AUC values for benign nodules were 0.92, 0.93, 0.97, and 0.97, respectively, while the AUC values for malignant nodules were 0.93, 0.91, 0.96, and 0.97, respectively. Figure S2b presents the confusion matrices of MLs. The SIMCA model correctly detected 116 malignant nodules, but incorrectly classified 12 of them as benign nodules, indicating relatively poorer performance in detecting malignant nodules compared to other models. Similarly, the PLSDA model misdiagnosed 12 out of 128 malignant nodules as benign nodules, suggesting it performed poorly in detecting benign nodules. Fortunately, the KNN model demonstrated the highest effectiveness among the models, with only 13 misclassifications out of 198 spectra.

3.3.2. Classification Model with 1DCNN. Figure 6 illustrates the performance of the improved 1DCNN model. The accuracy and loss curves of the 1DCNN model with no preprocessing are shown in Figure 6a. The accuracy of the network exhibits a rapid initial increase, followed by a gradual slowdown until it eventually stabilized. After 600 iterations of network training, the validation accuracy reached 96.73%. The loss value demonstrates a sharp decline during the initial stages of network training, gradually decreasing until convergence. The loss of the validation set stabilized at 0.09. Remarkably, in comparison to MLs, the accuracy of 1DCNN model shows a significant improvement, with a 4.0% increase for the calibration set and a 3.6% increase for the validation set compared to the KNN model. The confusion matrix is shown in Figure 6b, where 4 benign nodules were misdiagnosed as malignant, which is an improvement compared to the MLs. The ROC curves are shown in Figure 6c, with AUC values of both nodules reaching 0.99.

3.4. The Simplified Models through Characteristic Elements

Previous studies have attempted to assess the concentrations of metal elements in the serum samples of lung cancer patients. It was found that there were lowered concentrations of Zn, Mg, Fe, Co, Na, K, Ca, Mo, and Se and elevated concentrations of Al, Cu, Pb, Mn, Co, Fe, Cr, Mg, Cd, Hg, As, and Ni in sera of lung cancer patients compared to the control group. However, there are also conclusions that contradict the above situation.[35−39] Besides direct comparing of element concentration, ratios between elements have been identified important biomarkers for lung cancer disease in serum (V/Mn, V/Pb, V/Zn, Cr/Pb).[40] All of the above reflect the dysregulation of metal elements in serum from lung cancer patients.

The SRXRF spectra, known for their low detection limit, allow for the simultaneous excitation of multiple elements from Sodium (Na) to Uranium (U) in the periodic table. In this study, 14 channels corresponding to 104 (Na), 125 (Mg), 149 (Al), 331 (K), 369 (Ca), 451 (Ti), 495 (V), 542 (Cr), 590 (Mn), 641 (Fe), 693 (Co), 748 (Ni), 805 (Cu), and 864 (Zn) are the characteristic channels in SRXRF spectra. Figure 7 presents the relative intensity distributions of the selected characteristic elements. The violin plot's curve represents the degree of adherence to a normal distribution, with the center point indicating the mean value and the length of the box representing the standard deviation. The data distribution adhered to the standard normal distribution, suggesting that the samples being studied were representative. The relative average intensity of each element in serum of benign nodules was lower to that of malignant nodules, indicating that malignant nodules led to a general increase or decrease in the concentration of elements compared to patients with benign nodules. Among the 14 elements, Na, Al, Co, Ni, and Ca showed relatively small differences, whereas the relative levels of the remaining elements were greater in malignant nodules compared to benign nodules. This may be due to the disorder

https://doi.org/10.1021/envhealth.4c00124
Environ. Health 2025, 3, 40−47

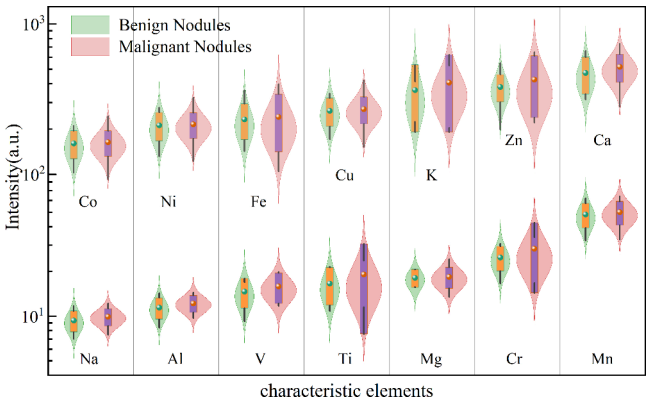


Figure 7. Intensity distribution diagram of characteristic elements.

of metal element markers in the serum caused by malignant nodules, while benign nodules remain stable within a certain area. This observation was consistent with the PCA score map, which shows the benign nodules surrounded by the clusters of malignant nodules.

To eliminate irrelevant information and improve the computation speed for data analysis, simplified models were established based on characteristic channels. Both KNN, an optimal machine learning algorithm with full-channels, and 1DCNN, a representative deep learning algorithm, were established as simplified models to evaluate and compare for classifying nodules, the performance of which is shown in Table 1. The 1DCNN model obtained better results with

Table 1. Performance for the Classification Models Based on Characteristic Channels with KNN and 1DCNN

	Predicted Class			
	KNN		1DCNN	
Actual Class	Benign Nodules	Malignant Nodules	Benign Nodules	Malignant Nodules
Benign Nodules	56	13	63	6
Malignant Nodules	9	120	5	124
Sensitivity (%)	81.16	93.02	91.30	96.12
F1-score (%)	83.58	91.60	91.97	95.75
G-mean (%)	75.48		87.75	
MCC (%)	75.27		87.73	
Kappa (%)	71.08		85.59	
Accuracy (%)	88.79		94.34	

accuracies of 94.34% for validation set. The sensitivity and F1-score of 1DCNN were both greater than 90%, while G-mean, MCC, and Kappa are all greater than 85%, showing a significant improvement compared to the simplified KNN. The simplified model with acceptable performance demonstrated that the characteristic elements play an important role in distinguishing pulmonary nodules. The input parameters were reduced from 1024 to 14, reducing the complexity of the model by 98.63%, which improve the prediction speed. Simplified model reduces risk of overfitting, improves generalization ability, and is crucial in practical applications and resource limited fields. Therefore, the simplified 1DCNN model based on characteristic channels was used to classify benign and malignant pulmonary nodules, although the accuracy decreased 2.39% compared with full-channels model.

T-SNE is a powerful visualization tool that is suitable for exploring the structure and patterns of high-dimensional data. Full-channel spectra, characteristic channel spectra, and features extracted from the convolutional layers of 1DCNN were applied to visualize the data distribution in 2D spaces. As shown in Figure S3, the scatter points of the two types of pulmonary nodules almost completely overlap, and the scatter points of the full-channel spectra (Figure S3a) are more chaotic, while the scatter points of the characteristic channels (Figure S3b) show a trend of separation. The features (Figure S3c) extracted by 1DCNN exhibit significant separability in the 2D space. This phenomenon demonstrates a series of dynamic changes from the original spectra to the features extracted by 1DCNN, as well as the effectiveness of 1DCNN in pulmonary nodule classification.

This study proposes a metallomics approach through SRXRF with simplified 1DCNN to distinguish pulmonary nodules using serum samples. Because 1DCNN is a data-hungry approach, the sample size is further expanded for training and validation to improve the model performance. Meanwhile, further research is needed to determine whether this method is applicable for distinguishing other types of lung diseases

4. CONCLUSIONS

The study highlights the feasibility of metallomics, coupled with SRXRF and 1DCNN, for distinguishing between benign and malignant pulmonary nodules. This approach enables accurate classification without the need for quantifying of metal element concentrations in serum. The 1DCNN and MLs model were developed based on full-channels with optimal accuracy of 96.73%. The simplified model based on characteristic elements resulted in good performance of sensitivity and F1-score > 91.30%, G-mean, MCC, and Kappa > 85.59%, and accuracy = 94.34% for eliminate redundancy. Compared with metabolomics or proteomics study on the classification of benign and malignant pulmonary nodules, the obtaining of metallomics profiling with SRXRF is easy-handle and less-destroy to serum samples.

Throughout the model development, 14 serum metal elements were identified as significantly impacting nodule classification, which paves the way for the potential use of metal elements in serum as biomarkers for diagnosis and prognosis. These elements may serve as the fingerprint metallome profile as diagnostic and treatment indicators for patients with nodules and provide a basis for the development of rapid detection methods for nodules classification. Considering the low availability of synchrotron radiation based XRF, further attempts with commercially available XRF machines are under investigation.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/envhealth.4c00124.

Preprocessed SRXRF spectra, performance of machine learning with full-channels, and T-SNE analysis of features (PDF)

https://doi.org/10.1021/envhealth.4c00124
Environ. Health 2025, 3, 40−47

AUTHOR INFORMATION

Corresponding Authors

Wei Wang − *College of Engineering, China Agricultural University, Beijing 100083, China*; Email: playerwxw@cau.edu.cn

Xin Wang − *School of Basic Medical Sciences, Anhui Medical University, Hefei 230032 Anhui, China*; Email: wxchem@ahmu.edu.cn

Yu-Feng Li − *CAS-HKU Joint Laboratory of Metallomics on Health and Environment, & CAS Key Laboratory for Biomedical Effects of Nanomaterials and Nanosafety, & Beijing Metallomics Facility, & National Consortium for Excellence in Metallomics, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China*; orcid.org/0000-0002-5013-5849; Email: liyf@ihep.ac.cn

Authors

Chaojie Wei − *College of Engineering, China Agricultural University, Beijing 100083, China*

Chao Li − *Department of Oncology, The Second Affiliated Hospital, Anhui Medical University, Hefei 230601 Anhui, China*

Hongxin Xie − *CAS-HKU Joint Laboratory of Metallomics on Health and Environment, & CAS Key Laboratory for Biomedical Effects of Nanomaterials and Nanosafety, & Beijing Metallomics Facility, & National Consortium for Excellence in Metallomics, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China*

Dongliang Chen − *Beijing Synchrotron Radiation Facility, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China*

Bai Li − *CAS-HKU Joint Laboratory of Metallomics on Health and Environment, & CAS Key Laboratory for Biomedical Effects of Nanomaterials and Nanosafety, & Beijing Metallomics Facility, & National Consortium for Excellence in Metallomics, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/envhealth.4c00124

Author Contributions

#C.W., C.L., and H.X. contributed equally to this work. Chaojie Wei: Investigation, Data curation, Visualization, Software, Writing- original draft. Chao Li: Methodology, Formal analysis, Writing- original draft. Hongxin Xie: Data curation, Software, Writing- original draft. Wei Wang: Supervision, Visualization, Software, Project administration, Funding acquisition, Writing- Reviewing and Editing. Xin Wang: Methodology, Project administration, Writing- Reviewing and Editing. Dongliang Chen: Data curation, Software, Bai Li: Data curation, Software. Yu-Feng Li: Resources, Software, Writing- Reviewing and Editing.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China [32272410]; the National Key Research and Development Program of China [2022YFF0607900, 2022YFA1207300]; the Natural Science Foundation of Anhui Province [2108085MA26]; and Research Fund of Anhui Institute of Translational Medicine [2022zhyx-C77]. We gratefully acknowledged the staffs from BL4W1B at Beijing Synchrotron Radiation Facility for beam time allocation and assistance during data collection.

DEDICATION

Dedicated to Ms. Yuanxiu Ding and Mr. Jiadong Liu on the occasion of their 80th birthday.

REFERENCES

(1) Loomis, D.; Grosse, Y.; Lauby-Secretan, B.; El Ghissassi, F.; Bouvard, V.; Benbrahim-Tallaa, L.; Guha, N.; Baan, R.; Mattock, H.; Straif, K.; et al. The Carcinogenicity of Outdoor Air Pollution. *Lancet Oncol.* **2013**, *14* (13), 1262−1263.

(2) Pei, Z.; Wu, M.; Zhu, W.; Pang, Y.; Niu, Y.; Zhang, R.; Zhang, H. Associations of Long-Term Exposure to Air Pollution with Prevalence of Pulmonary Nodules: A Cross-Sectional Study in Shijiazhuang, China. *Ecotoxicol. Environ. Saf.* **2023**, *262*, 115311.

(3) WHO. *Global Cancer Burden Growing, Amidst Mounting Need for Services*. https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services (accessed 2024−02−01).

(4) Wu, Y.-J.; Wu, F.-Z.; Yang, S.-C.; Tang, E.-K.; Liang, C.-H. Radiomics in Early Lung Cancer Diagnosis: From Diagnosis to Clinical Decision Support and Education. *Diagnostics.* **2022**, *12* (5), 1064.

(5) Mazzone, P. J.; Lam, L. Evaluating the Patient with a Pulmonary Nodule a Review. *JAMA* **2022**, *327* (3), 264−273.

(6) Huang, S.; Yang, J.; Shen, N.; Xu, Q.; Zhao, Q. Artificial Intelligence in Lung Cancer Diagnosis and Prognosis: Current Application and Future Perspective. *Semin. Cancer Biol.* **2023**, *89*, 30−37.

(7) Qiu, S.; Li, B.; Zhou, T.; Li, F.; Liang, T. Multi-View Auxiliary Diagnosis Algorithm for Lung Nodules. *CMC-Comput. Mater. Contin.* **2022**, *72* (3), 4897−4910.

(8) In Jae, L.; Hyoung June, I.; Miyeon, Y.; Kwanseop, L.; Yul, L.; Sang Hoon, B. Percutaneous Core Needle Biopsy for Small Lung Nodules with Diameters ≤ 10 mm: Accurate Diagnosis and Complication Rates. *Diagn. Interv. Radiol.* **2012**, *18*, 527−530.

(9) Li, Y.; Jiang, G.; Wu, W.; Yang, H.; Jin, Y.; Wu, M.; Liu, W.; Yang, A.; Chervova, O.; Zhang, S.; et al. Multi-Omics Integrated Circulating Cell-Free DNA Genomic Signatures Enhanced the Diagnostic Performance of Early-Stage Lung Cancer and Post-operative Minimal Residual Disease. *EBioMedicine.* **2023**, *91*, 104553.

(10) Zhang, Y.; Liu, Y.; Liu, H.; Tang, W. H. Exosomes: Biogenesis, Biologic Function and Clinical Potential. *Cell Biosci.* **2019**, *9* (1), 9−28.

(11) Yao, Y.; Wang, X.; Guan, J.; Xie, C.; Zhang, H.; Yang, J.; Luo, Y.; Chen, L.; Zhao, M.; Huo, B.; et al. Metabolomic Differentiation of Benign Vs Malignant Pulmonary Nodules with High Specificity Via High-Resolution Mass Spectrometry Analysis of Patient Sera. *Nat. Commun.* **2023**, *14* (1), 2339.

(12) Wang, M.; Dai, X.; Yang, X.; Jin, B.; Xie, Y.; Xu, C.; liu, Q.; Wang, L.; Ying, L.; Lu, W.; et al. Serum Protein Fishing for Machine Learning-Boosted Diagnostic Classification of Small Nodules of Lung. *ACS Nano* **2024**, *18* (5), 4038−4055.

(13) Li, H.; Wang, J.; Li, X.; Zhu, X.; Guo, S.; Wang, H.; Yu, J.; Ye, X.; He, F. Comparison of Serum from Lung Cancer Patients and from Patients with Benign Lung Nodule Using Ftir Spectroscopy. *Spectroc. Acta Pt. A-Mol. Biomol. Spectr.* **2024**, *306*, 123596.

(14) Luo, H.; Zu, R.; Li, L.; Deng, Y.; He, S.; Yin, X.; Zhang, K.; He, Q.; Yin, Y.; Yin, G.; et al. Serum Laser Raman Spectroscopy as a Potential Diagnostic Tool to Discriminate the Benignancy or Malignancy of Pulmonary Nodules. *iScience.* **2023**, *26* (5), 106693.

(15) Li, Y.-F.; Sun, H. *Applied Metallomics: From Life Sciences to Environmental Sciences*; Wiley, 2024.

(16) Li, Y.-F. Metallomics in Multidisciplinary Research and the Analytical Advances. *At. Spectrosc.* **2021**, *42* (5), 227−230.

(17) Shayganfard, M. Are Essential Trace Elements Effective in Modulation of Mental Disorders? Update and Perspectives. *Biol. Trace Elem. Res.* **2022**, *200* (3), 1032−1059.
(18) Stochel-Gaudyn, A.; Fyderek, K.; Kościelniak, P. Serum Trace Elements Profile in the Pediatric Inflammatory Bowel Disease Progress Evaluation. *J. Trace Elem. Med. Biol.* **2019**, *55*, 121−126.
(19) Tan, C.; Chen, H.; Xia, C. The Prediction of Cardiovascular Disease Based on Trace Element Contents in Hair and a Classifier of Boosting Decision Stumps. *Biol. Trace Elem. Res.* **2009**, *129* (1−3), 9−19.
(20) Wei, C.; Li, C.; Xie, H.; Wang, X.; Li, Y.-F.; Li, Y.; Liu, Y.; Wang, W. Research on Cancer Screening Method Based on Synchrotron Radiation X-Ray Fluorescence Spectroscopy and One-Dimensional Convolutional Neural Network. *Chin. J. Inorg. Anal. Chem.* **2024**, *14* (1), 104−111.
(21) Sheng, D.; Wang, Y.; Jianjun, Y.; Zhang, Y.; Jiang, X.; Li, Y.-F.; Wang, X.; Li, C. Application of Metallomics in the Early Diagnosis of Acute Myocardial Infarction. *Chin. J. Inorg. Anal. Chem.* **2024**, *14* (01), 112−116.
(22) He, L.; Lu, Y.; Li, C.; Xie, H.; Zhao, J.; Wang, Y.; Wang, L.; Wang, X.; Wang, W.; Chen, D.; et al. Non-Targeted Metallomics through Synchrotron Radiation X-Ray Fluorescence with Machine Learning for Cancer Screening Using Blood Samples. *Talanta.* **2022**, *245*, 123486.
(23) Xie, H.; Wei, C.; Wang, W.; Chen, R.; Cui, L.; Wang, L.; Chen, D.; Yu, Y.-L.; Li, B.; Li, Y.-F. Screening the Phytotoxicity of Micro/Nanoplastics through Non-Targeted Metallomics with Synchrotron Radiation X-Ray Fluorescence and Deep Learning: Taking Micro/Nano Polyethylene Terephthalate as an Example. *J. Hazard. Mater.* **2024**, *463*, 132886.
(24) Attallah, O.; Aslan, M. F.; Sabanci, K. A Framework for Lung and Colon Cancer Diagnosis Via Lightweight Deep Learning Models and Transformation Methods. *Diagnostics.* **2022**, *12* (12), 2926.
(25) Pradhan, K. S.; Chawla, P.; Tiwari, R. Hrdel: High Ranking Deep Ensemble Learning-Based Lung Cancer Diagnosis Model. *Expert Syst. Appl.* **2023**, *213*, 118956.
(26) Houssein, E. H.; Emam, M. M.; Ali, A. A.; Suganthan, P. N. Deep and Machine Learning Techniques for Medical Imaging-Based Breast Cancer: A Comprehensive Review. *Expert Syst. Appl.* **2021**, *167*, 114161.
(27) Shin, H.; Oh, S.; Hong, S.; Kang, M.; Kang, D.; Ji, Y.-g.; Choi, B. H.; Kang, K.-W.; Jeong, H.; Park, Y.; et al. Early-Stage Lung Cancer Diagnosis by Deep Learning-Based Spectroscopic Analysis of Circulating Exosomes. *ACS Nano* **2020**, *14* (5), 5435−5444.
(28) Cheong, J.-H.; Wang, S. C.; Park, S.; Porembka, M. R.; Christie, A. L.; Kim, H.; Kim, H. S.; Zhu, H.; Hyung, W. J.; Noh, S. H.; et al. Development and Validation of a Prognostic and Predictive 32-Gene Signature for Gastric Cancer. *Nat. Commun.* **2022**, *13* (1), 774.
(29) Greener, J. G.; Kandathil, S. M.; Moffat, L.; Jones, D. T. A Guide to Machine Learning for Biologists. *Nat. Rev. Mol. Cell Biol.* **2022**, *23* (1), 40−55.
(30) Paladini, E.; Vantaggiato, E.; Bougourzi, F.; Distante, C.; Hadid, A.; Taleb-Ahmed, A. Two Ensemble-Cnn Approaches for Colorectal Cancer Tissue Type Classification. *J. Imaging.* **2021**, *7* (3), 51.
(31) Acquarelli, J.; van Laarhoven, T.; Gerretzen, J.; Tran, T. N.; Buydens, L. M. C.; Marchiori, E. Convolutional Neural Networks for Vibrational Spectroscopic Data Analysis. *Anal. Chim. Acta* **2017**, *954*, 22−31.
(32) Kiranyaz, S.; Avci, O.; Abdeljaber, O.; Ince, T.; Gabbouj, M.; Inman, D. J. 1d Convolutional Neural Networks and Applications: A Survey. *Mech. Syst. Signal Proc.* **2021**, *151*, 107398.
(33) Okonda, J. J.; Angeyo, H. K.; Dehayem-Kamadjeu, A.; Rogena, A. E. Feasibility for Early Cancer Diagnostics by Machine Learning Enabled Synchrotron Radiation Based Micro X-Ray Fluorescence Imaging of Trace Biometals as Cancer Biomarkers. *Spectroc. Acta Pt. B-Atom. Spectr.* **2023**, *204*, 106671.
(34) Mazzone, P. J.; Lam, L. Evaluating the Patient with a Pulmonary Nodule: A Review. *JAMA* **2022**, *327* (3), 264−273.
(35) Zabłocka-Słowińska, K.; Płaczkowska, S.; Prescha, A.; Pawełczyk, K.; Porębska, I.; Kosacka, M.; Pawlik-Sobecka, L.; Grajeta, H. Serum and Whole Blood Zn, Cu and Mn Profiles and Their Relation to Redox Status in Lung Cancer Patients. *J. Trace Elem. Med. Biol.* **2018**, *45*, 78−84.
(36) Cobanoglu, U.; Demir, H.; Sayir, F.; Duran, M.; Mergan, D. Some Mineral, Trace Element and Heavy Metal Concentrations in Lung Cancer. *Asian Pac. J. Cancer Prev.* **2010**, *11* (5), 1383−1388.
(37) Qayyum, M. A.; Farooq, Z.; Yaseen, M.; Mahmood, M. H. R.; Irfan, A.; Zafar, M. N.; Khawaja, M.; Naeem, K.; Kisa, D. Statistical Assessment of Toxic and Essential Metals in the Serum of Female Patients with Lung Carcinoma from Pakistan. *Biol. Trace Elem. Res.* **2020**, *197* (2), 367−383.
(38) Zhang, K.; Zhu, T.; Quan, X.; Qian, Y.; Liu, Y.; Zhang, J.; Zhang, H.; Li, H.; Qian, B. Association between Blood Heavy Metals and Lung Cancer Risk: A Case-Control Study in China. *Chemosphere.* **2023**, *343*, 140200.
(39) Callejón-Leblic, B.; Sánchez Espirilla, S.; Gotera-Rivera, C.; Santana, R.; Díaz-Olivares, I.; Marín, J. M.; Macario, C. C.; Cosio, B. G.; Fuster, A.; García, I. S.; et al. Metallomic Signatures of Lung Cancer and Chronic Obstructive Pulmonary Disease. *Int. J. Mol. Sci.* **2023**, *24* (18), 14250.
(40) Callejón-Leblic, B.; Gómez-Ariza, J. L.; Pereira-Vega, A.; García-Barrera, T. Metal Dyshomeostasis Based Biomarkers of Lung Cancer Using Human Biofluids. *Metallomics.* **2018**, *10* (10), 1444−1451.

https://doi.org/10.1021/envhealth.4c00124
Environ. Health 2025, 3, 40−47