



OPEN ACCESS

EDITED BY

Rui Zhang,
University of Jinan, China

REVIEWED BY

Hongmei Deng,
Guangzhou University, China
Muhammad Zafar Irshad Khan,
Zhejiang University, China
Vinod Chandra S. S.,
University of Kerala, India

*CORRESPONDENCE

Jianbo Jia,
✉ jbjiaqu@163.com
Xiaolong Xu,
✉ xuxl@wyu.edu.cn

RECEIVED 30 October 2025

REVISED 15 January 2026

ACCEPTED 29 January 2026

PUBLISHED 16 February 2026

CITATION

Zhao L, Li A, Zhang Y, Jiang B, Liu C, Xu X and Jia J (2026) Predicting environmental pollutant concentrations via cell image-derived damage features using a hybrid model.
Front. Environ. Sci. 14:1729913.
doi: 10.3389/fenvs.2026.1729913

COPYRIGHT

© 2026 Zhao, Li, Zhang, Jiang, Liu, Xu and Jia. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Predicting environmental pollutant concentrations via cell image-derived damage features using a hybrid model

Linying Zhao¹, Anqi Li¹, Yanyan Zhang¹, Beier Jiang¹,
Changyu Liu^{1,2}, Xiaolong Xu^{1,2,3*} and Jianbo Jia^{1,2*}

¹Jiangmen Key Laboratory of Synthetic Chemistry and Cleaner Production, School of Environmental and Chemical Engineering; Carbon Neutrality Innovation Center, Wuyi University, Jiangmen, China, ²Guangdong Provincial Laboratory of Chemistry and Fine Chemical Industry Jieyang Center, Jieyang, China, ³School of Emergency Technology and Management; Wuyi University, Jiangmen, China

Introduction: Effective detection of environmental pollution relies on reliable methods for determining pollutant concentrations, with cellular damage reflecting pollutant toxicity as a vital detection tool. This study presents a novel quantitative method for predicting environmental pollutant concentrations using cell images and a hybrid model.

Methods: The approach processes conventional optical microscope images by extracting grayscale statistical features and constructing a hybrid predictive framework that integrates stepwise regression for feature selection and multilayer perceptron for nonlinear modeling, enabling accurate mapping from image-based damage features to pollutant concentrations.

Results: Experiments show that the model performs consistently well across five cell types: HeLa, A549, HUVEC, PC12, and HaCaT. For example, it achieves an R^2 of 0.9911 on the HeLa test set, demonstrating strong generalization ability and robustness.

Discussion: The method does not require expensive equipment or complex sample preparation, offering an innovative, rapid, and low-cost solution for monitoring environmental pollutant concentrations.

KEYWORDS

cell image, environmental toxicology, grayscale statistical features, hybrid model, pollutant concentration prediction

1 Introduction

Cell damage is a direct reflection of the effects of environmental pollutants on living organisms (Sheller-Miller et al., 2020; Yang et al., 2024), and is gradually becoming an effective way to detect pollutant concentrations. Pollutants enter organisms through water, atmosphere or soil, and then cause physiological damage to cells (Yang et al., 2024; Suzuki et al., 2020; Pereira et al., 2021), and the extent of this damage is closely related to the pollutant concentration. The interference of environmental pollutants on cell functions is usually shown to be concentration dependent; that is, the greater the concentration of pollutants, the more significant the damage or functional interference to cells (Chukhlovin et al., 2001; Xin et al., 2015). Therefore, the degree of cellular damage can be used as a sensitive response to changes in pollutant concentrations, providing a scientific basis for the quantitative detection of pollutants. This method not only has important application value

in the concentration detection of environmental pollutants, but also can provide new technical means for environmental toxicology research.

In the fields of environmental toxicology, drug discovery and disease diagnosis, cellular damage usually reflects the effects of external environment or drug treatment on the physiological functions of cells, and common methods of assessing cellular damage mainly include cellular activity assay, determination of cellular oxidative stress markers, and observation of cellular morphology. Cell viability detection methods (Kamiloglu et al., 2020) can be roughly divided into: the dye exclusion methods, the colorimetric analysis, fluorescent dye methods, the bioluminescence assays (such as Real-time viability assay (Hiers et al., 2024)) and the flow cytometry assays. Measurement of cellular oxidative stress markers is an important tool for assessing the extent of oxidative damage to cells, such as ROS levels by fluorescent probe assay. In principle, cells are stimulated with morphological changes (e.g., smaller cell size, rounded cells, reduced cell apposition, etc.) (Kerr et al., 1972), which can be used as an important means to analyze cell damage. Traditional cell damage assessment methods usually rely on physicochemical indicators for qualitative or semi-quantitative analysis, which is a complex and time-consuming process and relies on specific equipment. In contrast, cell images provide rich structural and optical information, and analyzing the multidimensional features of the images through machine learning is expected to provide a more accurate and convenient way for contaminant toxicity assessment.

However, morphological analysis typically relies on complex image segmentation and parameter calculations, making it susceptible to interference from imaging conditions, cell density, and segmentation errors, affecting feature consistency and reproducibility (Chen et al., 2012). In contrast, grayscale distribution features directly reflect changes in optical density within cellular regions and are more sensitive to alterations in intracellular composition, density, and translucency caused by damage. This feature is computationally simple, highly stable, and less dependent on segmentation accuracy, making it more suitable as a robust indicator for high-throughput, automated analysis. Therefore, this study chooses grayscale distribution as the basic feature to establish a reliable and reproducible quantitative relationship between cell damage and contaminant concentration.

Previously, we developed a method based on grayscale frequency analysis to quantify the degree of cellular damage using image brightness, which achieved good results (Li et al., 2025). However, this method, which relies only on grayscale frequency analysis by calculating the ratio of the number of maximum pixel points in abnormal regions to the total number of maximum pixel points within normal and abnormal regions, may not adequately capture the complex information in cell images. We propose the use of hybrid models to capture complex information in cell images. Based on our experimental research observations, we found that the gray-scale distribution data of cell images can be used as an effective indicator of cell damage for quantitative analysis of environmental pollutant concentrations. However, direct processing of all grayscale distribution data may lead to redundant information and increased computational cost, so feature selection methods are needed to extract the most representative information to improve

model efficiency and performance. Feature selection methods have helped us simplify the data, but in order to mine the information in the data more comprehensively, we introduce a hybrid model based on feature selection. Hybrid modeling significantly improves model performance by integrating the strengths of multiple methods and compensating for the weaknesses of a single method. Many hybrid methods have been reported and shown excellent results in several application areas. For example, Improved Principal Component Analysis-Convolutional Neural Network-Long and Short-Term Memory Network (improved PCA-CNN-LSTM) (Zhu et al., 2023), Principal Component Analysis- Back Propagation-Artificial Neural Network (PCA-BP-ANN) (Sadra and Khorrami, 2023), Principal Component Analysis-Ridge Regression (PCA-Ridge Regression) (Wang W. Y. et al., 2021), and so on.

In summary, cellular images provide an intuitive means for assessing pollutant toxicity, yet constructing a high-precision, low-cost quantitative prediction system still faces dual challenges. First, the grayscale frequency analysis method we previously proposed has significant limitations (Li et al., 2025). It relies solely on a single statistical measure, failing to fully leverage the rich grayscale distribution information within cellular images, resulting in insufficient prediction accuracy and robustness. More critically, mainstream detection technologies for environmental pollutant concentrations, such as electrochemical sensors, infrared spectroscopy, and mass spectrometry, while highly accurate, generally require expensive, specialized instruments and involve complex sample pretreatment (Kumunda et al., 2021; Meher and Zarouri, 2025). This makes achieving low-cost, high-throughput rapid screening difficult. Addressing these dual bottlenecks of “insufficient feature utilization” and “excessive equipment dependency,” this study aims to explore a novel “one-stop” prediction paradigm: starting directly from cell images obtained via conventional optical microscopy, extracting their full-band grayscale statistical features, and constructing intelligent prediction models to ultimately achieve precise quantitative determination of pollutant concentrations.

To this end, we propose the stepwise regression-multilayer perceptron (SR-MLP) hybrid model and compare it with two linear hybrid variants: stepwise regression-linear regression (SR-LR) and stepwise regression-ridge regression (SR-RR). The dataset is divided into training and testing sets in an 8:2 ratio for all three models. SR-LR combines stepwise regression with linear regression, offering simplicity but a tendency to overfit. SR-RR integrates stepwise regression (SR) with ridge regression (RR), which mitigates multicollinearity through L2 regularization, yet remains limited to linear mappings. In contrast, SR-MLP retains SR for feature selection and further employs a multilayer perceptron with L2 regularization to model nonlinear interactions, effectively capturing complex relationships between features and pollutant concentrations. Experimental results demonstrate that while SR-LR and SR-RR perform adequately under linear assumptions, SR-MLP achieves superior fitting and generalization in scenarios involving nonlinear dependencies. Moreover, SR-MLP consistently surpasses baseline models without feature selection (linear regression (LR), RR, and multilayer perceptron (MLP)) across all evaluation metrics, confirming its enhanced accuracy and robustness. Thus, SR-MLP provides a more accurate and

flexible approach for quantifying environmental pollutants and elucidating their dose-dependent impact on cellular damage.

The proposed SR-MLP hybrid framework in this study innovates by establishing a new paradigm for quantitative mapping between “image features and pollutant concentrations” that does not rely on complex model architectures or high computational resources. Current mainstream approaches in bioimage analysis typically rely on deep learning models such as convolutional neural networks (CNNs) or Transformers to extract multi-level, structured visual features (Shobayo and Saatchi, 2025). While these methods possess strong feature representation capabilities, they often suffer from limitations including complex model structures, high training costs, strong dependence on large-scale labeled data, and relatively weak interpretability (Atabansi et al., 2023). In contrast, this study demonstrates for the first time that one-dimensional grayscale statistical features extracted solely from cell images can establish a highly accurate quantitative relationship with pollutant concentrations through feature selection (SR) and nonlinear modeling (MLP). This approach significantly reduces model complexity and computational overhead while maintaining predictive performance. It enhances transparency in feature selection and interpretability of results, offering a lighter, more deployable solution for rapid environmental pollutant detection. Therefore, the core contributions of this study can be summarized as follows:

1. Proposing the SR-MLP hybrid model to achieve high-precision pollutant concentration prediction across cell types.
2. Validating the “one-stop” quantitative prediction paradigm based on image damage features for its cross-cell universality and robustness.
3. Establishing a new low-cost, non-invasive, and easily deployable imaging-based detection scheme for environmental pollutants.

2 Experiments and data

2.1 Data acquisition and image processing

First, well-conditioned HeLa cells were spread in six-well plates and cultured in adherence at an inoculum density of 1×10^6 cells/mL per well for 24 h. Then, cells were stimulated with Diniconazole (DIN) at concentrations of 0, 2.5, 5, 10, and 20 $\mu\text{mol/L}$ for 24 h. The culture environment was maintained at 37 °C and 5% CO_2 . After stimulation, bright-field imaging was performed using a 10 \times objective of an optical microscope (OLYMPUS CKX53), and conditions such as exposure time and light intensity were kept consistent to ensure uniform imaging conditions for all sample images. One hundred images were acquired for each concentration, and the resolution of the images was 3072 \times 2048 pixels, totaling 500 samples. For HUVEC, and HaCaT cells, stimulation was performed using DIN at concentrations of 0, 2.5, 5, 10, and 20 $\mu\text{mol/L}$, and a total of 150 samples were collected. For A549 cells, stimulation was performed using DIN at concentrations of 0, 12.5, 25, and 50 $\mu\text{mol/L}$, and a total of 120 samples were collected. For PC12 cells, stimulation was performed using DIN at

concentrations of 0, 5, 10, 20 and 40 $\mu\text{mol/L}$, and a total of 150 samples were collected. It should be noted that this study adopted a phased image acquisition strategy. Initially, HeLa cells were used as the model system for method development. To ensure statistical robustness and provide a solid data foundation for feature analysis, 100 images were collected at each concentration. After successful method validation, to assess its cross-cell line generalization ability and feasibility in scenarios with limited resources, 30 images were collected per concentration in the subsequent experiments with A549, HUVEC, PC12, and HaCaT cells.

In order to further analyze the grayscale characteristics of the cellular regions, we performed batch automatic segmentation of the cellular images by calling Cellpose’s API and exported the corresponding masks (Stringer et al., 2021). Cell segmentation was performed using the Cellpose pre-trained model (cyto3). The parameters were set as follows: automatic diameter estimation, grayscale image channels, a flow threshold of 0.4, and a minimum cell size of 15 pixels. Tiled processing and network averaging were enabled to balance the processing of large images with segmentation accuracy. All segmentation tasks were accelerated using GPU processing. To visually demonstrate the image segmentation results, we overlaid the mask generated by Cellpose onto the original image during the image processing stage in Figure 1. This clearly shows the extracted cell regions (as illustrated by the grayscale example in the image processing module), thereby providing an intuitive basis for quality verification in subsequent grayscale feature extraction. Then, we batch-calculated the grayscale values of cell regions and their corresponding pixel numbers based on mask, and adjusted the data with normalization based on the average occupancy of all samples to ensure comparability among different images. Ultimately, we constructed five datasets containing the cell grayscale distribution and DIN concentration for subsequent analysis and modeling.

2.2 Data preprocessing

We performed a visual analysis of the samples at each concentration, using the sample stimulated with 10 $\mu\text{mol/L}$ DIN in HeLa cells as an example. Figure 2 shows the grayscale distribution of 100 representative samples, where a small number of outliers were observed. To address this, we applied the interquartile range (IQR) method to remove outliers (Swami et al., 2023). When the scaling factor $k = 3$, it effectively eliminated apparent outlier curves while retaining most of the samples, thereby enhancing data reliability and preventing outliers from negatively affecting the model. Observing the gray scale distribution curve after removing the outliers in Figure 3, we found that the data still has a certain degree of dispersion. Therefore, we standardized the data to make the data obey the standard normal distribution, which in turn improves the training effect of the model (Muhammad and Faraj, 2014).

3 Methods

We observed an increase in cell dehiscence and a tendency towards a rounded morphology with increasing DIN concentration, which may be attributed to DIN-induced apoptosis, leading to

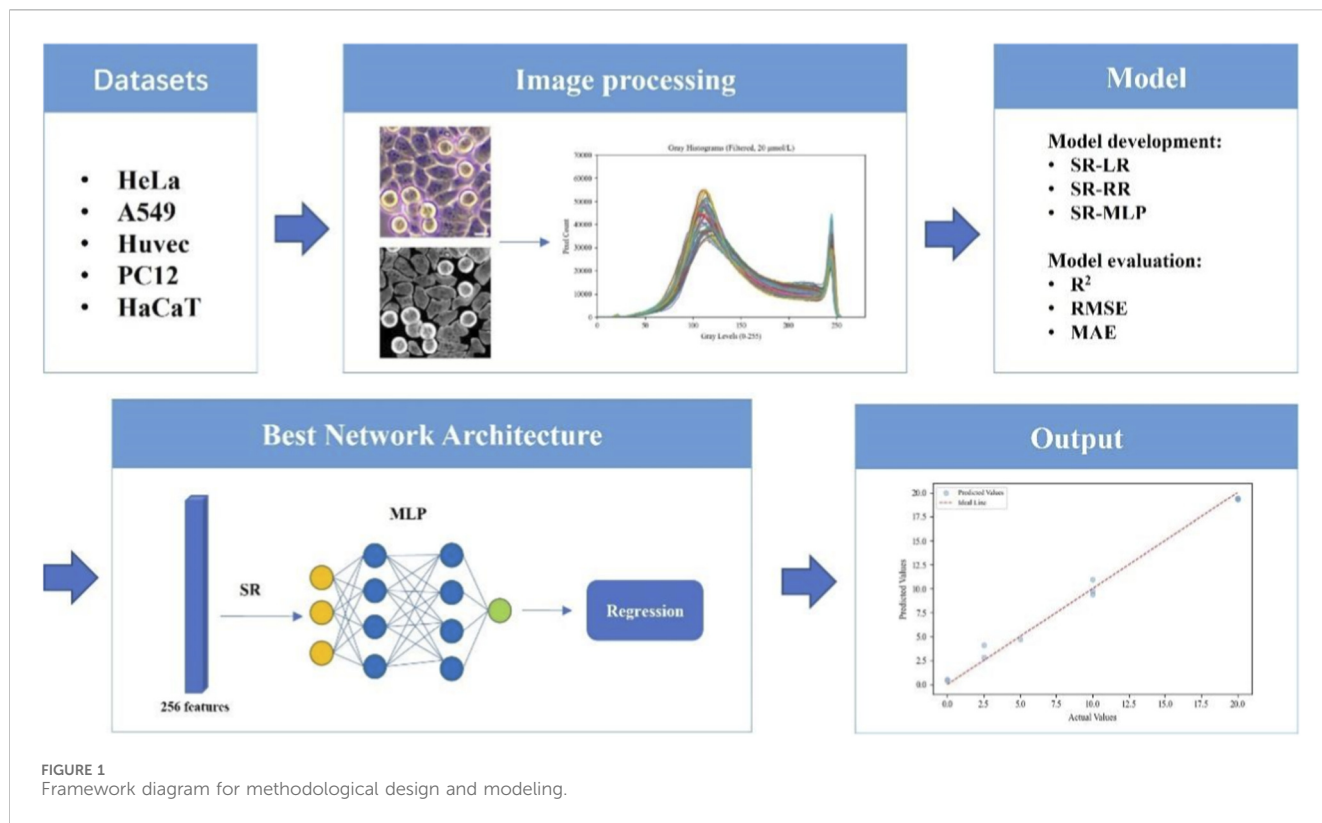


FIGURE 1 Framework diagram for methodological design and modeling.

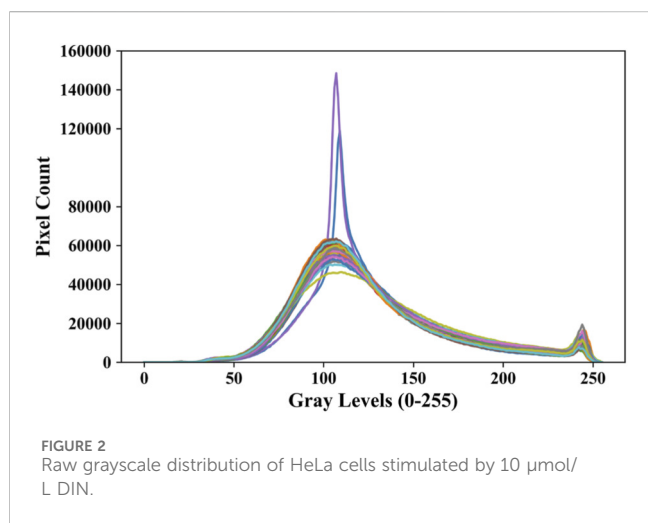


FIGURE 2 Raw grayscale distribution of HeLa cells stimulated by 10 $\mu\text{mol/L}$ DIN.

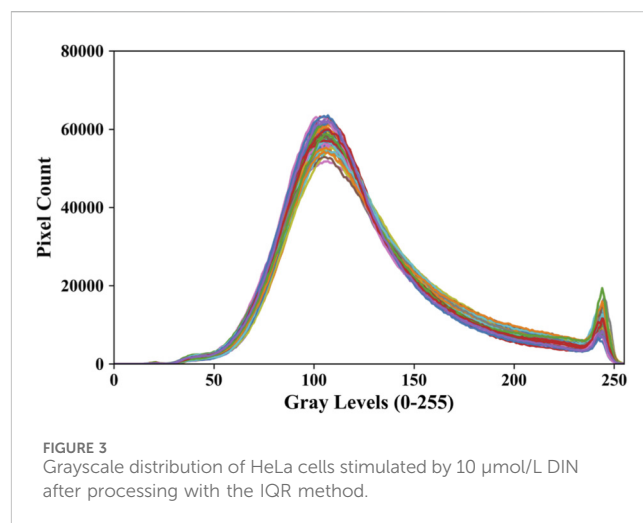


FIGURE 3 Grayscale distribution of HeLa cells stimulated by 10 $\mu\text{mol/L}$ DIN after processing with the IQR method.

coalescence of nuclei and cytoplasm, gradual contraction of the cells and loss of adhesion, and the eventual formation of membrane-encapsulated apoptotic vesicles, which resulted in a rounded morphology of the cells (Kerr et al., 1972). Intuitively, the surrounding brightness of round cells is high, and the grayscale values of damaged cells are mainly concentrated in the 200-255 interval (e.g., Figure 4), and the number of pixels in this interval tends to increase with increasing DIN concentration (e.g., Figure 5). We inferred that there is a certain quantitative relationship between the gray scale distribution of cell images and pollutant concentration. To further explore this relationship, we chose three hybrid modeling approaches: SR-LR, SR-RR, and SR-

MLP. The feature selection methods, modeling approaches, optimization strategies, and model evaluation metrics of these three models are briefly described below. The specific modeling framework of the methodology is shown in Figure 1.

3.1 SR

SR is a feature selection method based on statistical significance (P-value). It iteratively filters out the most relevant features to the target variable, thus reducing redundant features and improving the efficiency and predictive power of the model (Zhou and Jiang, 2016).

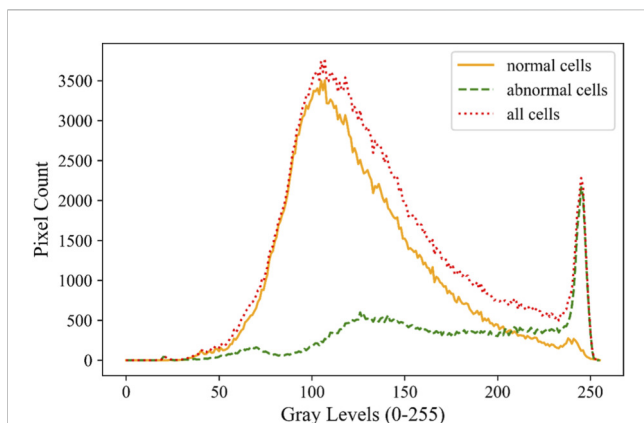


FIGURE 4
Grayscale distribution of normal cells and abnormal cells.

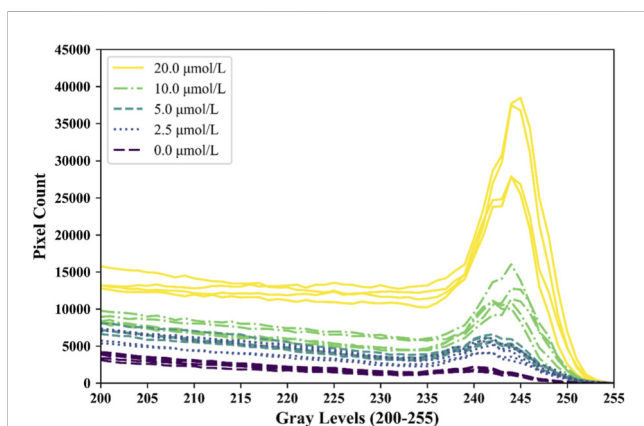


FIGURE 5
Pixel distribution of abnormal cells in high grayscale range (200-255) at different DIN concentrations.

3.2 SR-LR and SR-RR

The SR-LR model is a combination of SR and LR, which is a simple and interpretable method of linear modeling through screened features (James et al., 2023). The SR-RR model introduces L2 regularization on this basis, i.e., it is a combination of SR and RR. RR reduces multicollinearity through L2 regularization, limits model weights, and reduces the risk of overfitting, thus improving the stability and predictive ability of the model (Wang X. et al., 2021). The advantage of RR is that it can better handle the situation of correlation between features and control the model complexity while maintaining the linear modeling framework.

3.3 SR-MLP

In the SR-MLP model, we employ the MLP for nonlinear modeling. The MLP is a feedforward neural network consisting of multiple hidden layers, which captures complex nonlinear relationships in data through weighted summation and activation functions (Liu et al., 2022). The structure of the MLP—such as the

number of hidden layers, the number of neurons, and the activation functions—can be flexibly adjusted according to the specific task requirements. In this study, we constructed an MLP network with three hidden layers, containing 72, 48, and 24 neurons, respectively. The hidden layers use the Rectified Linear Unit (ReLU) activation function to enhance the modeling capability for complex nonlinear relationships (Alkhouly et al., 2021; Fu et al., 2020), while the output layer employs a linear transformation (Linear) to directly output continuous predictions for the regression task (Kılıçarslan et al., 2021).

To prevent overfitting and enhance the model's generalization capability, we employed the following regularization strategies during training: First, L2 regularization was introduced by adding a weight penalty term to the loss function, thereby constraining the magnitude of weights and suppressing overfitting (Xie et al., 2022). Second, the Dropout method was applied to randomly drop a portion of hidden layer neurons during training, reducing model complexity and further improving its generalization ability (Salehin and Kang, 2023). We monitored the trends of training loss and validation loss to guide hyperparameter optimization. Model convergence was considered satisfactory when both losses decreased smoothly, remained closely aligned, and eventually stabilized. A scenario where training loss continued to decline while validation loss plateaued or increased prematurely indicated potential overfitting, whereas persistently high losses with slow reduction suggested underfitting (Wang et al., 2022). Accordingly, we systematically adjusted hyperparameters including the number of training epochs, batch size, regularization strength, learning rate, and Dropout rate to achieve optimal convergence performance.

3.4 Evaluation metrics

To fully evaluate the performance of the hybrid model, we use the following commonly used metrics:

1. Coefficient of determination (R^2): a measure of the goodness of fit of the model, the closer the value is to 1 the better the model fits the data and the better the prediction (Chicco et al., 2021).
2. Root Mean Square Error (RMSE): a measure of the difference between the predicted value and the true value, emphasizing the effect of larger errors. Smaller values indicate higher model prediction accuracy (Chicco et al., 2021).
3. Mean Absolute Error (MAE): Calculates the average of the absolute error between the predicted value and the true value, reflecting the overall error level of the model. The smaller the value indicates the higher the prediction accuracy of the model (chicco et al., 2021).

4 Results and discussion

4.1 Analysis of models performance

We evaluated the performance of the three hybrid models (SR-LR, SR-RR, and SR-MLP) and the baseline models (LR, RR, and

TABLE 1 List of statistical parameters.

Dataset	Method	R ²	RMSE	MAE
HeLa	LR	0.8306	2.7705	1.5626
	RR	0.9386	1.6684	1.0388
	MLP	0.9770	1.1481	0.7993
	SR-LR	0.9707	1.1518	0.8827
	SR-RR	0.9708	1.1504	0.8837
	SR-MLP	0.9911	0.6364	0.3730
A549	LR	0.8453	6.5973	5.6149
	RR	0.8457	6.5899	5.6121
	MLP	0.8786	5.3887	4.5710
	SR-LR	0.8809	5.7900	4.5223
	SR-RR	0.8805	5.8001	4.5419
	SR-MLP	0.9716	2.8269	1.5795
HUVEC	LR	0.7287	3.9980	3.1179
	RR	0.7382	3.9277	3.0682
	MLP	0.8398	2.7994	2.2135
	SR-LR	0.8533	2.9403	2.3650
	SR-RR	0.8507	2.9659	2.3633
	SR-MLP	0.9753	1.2055	0.8668
PC12	LR	0.9340	3.8624	2.8333
	RR	0.9340	3.8628	2.8382
	MLP	0.8942	3.9298	3.3227
	SR-LR	0.9375	3.7580	3.2913
	SR-RR	0.9384	3.7308	3.2686
	SR-MLP	0.9821	2.0093	1.3450
HaCaT	LR	0.9018	2.3391	1.7062
	RR	0.9024	2.3316	1.7084
	MLP	0.8633	2.6931	2.1108
	SR-LR	0.9174	2.1452	1.5500
	SR-RR	0.9140	2.1898	1.5842
	SR-MLP	0.9222	2.0822	1.2773

Datasets: HeLa, A549, Huvec, PC12, and HaCaT represent the five different cell types used in the study. Model development: SR-LR, SR-RR, and SR-MLP stand for the hybrid models “stepwise regression–linear regression,” “stepwise regression–ridge regression,” and “stepwise regression–multilayer perceptron,” respectively. Model evaluation: R², RMSE, and MAE denote the coefficient of determination, root mean square error, and mean absolute error, respectively, which are used to quantify model performance.

MLP) using R², RMSE, and MAE metrics, as summarized in Table 1. The results indicate that all hybrid models incorporating SR outperformed their non-SR counterparts across most datasets, underscoring the efficacy of SR in feature selection for enhancing model performance. And among these three hybrid models, the R² values of SR-MLP on all cellular datasets were significantly better than those of SR-LR and SR-RR and higher than 0.9, indicating a

better fitting ability. In the case of the HeLa cell dataset, for example, the R² of SR-MLP was as high as 0.9919, while SR-LR and SR-RR were 0.9707 and 0.9708, respectively. Moreover, the RMSE and MAE of SR-MLP are lower than those of SR-LR and SR-RR on all datasets, which indicates that its prediction error is smaller and its stability is higher. In contrast, the R² of SR-LR and SR-RR was higher than 0.9 on the HeLa, PC12 and HaCaT cell datasets, but lower than 0.9 on the A549 and HUVEC cell datasets, which indicated a poorer fit and limited generalization ability.

To further assess the predictive power of SR-MLP, we compared the predicted and true values of SR-RR (e.g., Table 2) with those of SR-MLP (e.g., Table 3) models. Taking the HeLa cell dataset as an example, the predicted values of SR-MLP and SR-RR were 19.44 μmol/L and 19.07 μmol/L, respectively, at a high concentration of 20 μmol/L, whereas the predicted values of SR-MLP and SR-RR were 2.65 μmol/L and 3.54 μmol/L, respectively, at a low concentration of 2.5 μmol/L. It can be seen that the predicted values of SR-MLP are closer to the true values, and the errors are smaller, indicating that the model has the ability to predict the true values at different concentrations. MLP’s prediction value is closer to the real value with smaller error, indicating that the model has strong prediction accuracy under different concentration conditions.

4.2 Analysis of feature generalisation of cross-cell

In order to reduce data dimensionality and enhance model stability, we employed the SR method to screen gray-scale features based on statistical significance. During the screening process, differentiated p-value inclusion thresholds (ranging from <0.001 to <0.05) were set according to the data characteristics of different cell datasets and the need to control model complexity. Specifically, for the HeLa dataset, which exhibited clearer feature-concentration relationships and higher data quality, a stricter threshold (p < 0.001) was applied to construct a highly concise initial feature set. For the other datasets, relatively lenient thresholds (p < 0.05 or p < 0.01) were adopted to retain more potentially relevant feature information for the subsequent MLP to perform nonlinear learning. The setting of all thresholds followed the statistical principle of controlling Type I errors while also considering the connection with the subsequent nonlinear modeling stage (Aguinis et al., 2021).

Despite the differences, the screening results also indicate some similarities among key features across different cell datasets, as shown in Table 4. For instance, feature 244 was selected in the HeLa, A549, and HUVEC cell datasets, while key feature 204 for PC12 cells and key feature 248 for HaCaT cells, along with feature 244, all fell within the gray-scale range of 200–255. This suggests that these features may possess strong generalizability across different cell types and can stably reflect the effects of pollutants on cells. The gray-scale range of 200–255 typically corresponds to abnormal cells, and features within this range exhibit more significant variations. Moreover, the features screened in this range demonstrated better p-values, further validating the importance of these features in the cellular damage response. Therefore, during the feature selection process, we focused on the key features of cell damage and

TABLE 2 Comparison of predicted and true values of the SR-RR model.

SR-RR									
Predicted (μmol/L)					Actual (μmol/L)				
HeLa	A549	HUVEC	PC12	HaCaT	HeLa	A549	HUVEC	PC12	HaCaT
0.24	2.65	3.11	5.71	0.23	0.00	0.00	0.00	0.00	0.00
3.54	13.60	2.34	4.28	3.54	2.50	12.50	2.50	5.00	2.50
5.57	30.99	6.97	13.57	0.20	5.00	25.00	5.00	10.00	5.00
10.33	47.94	8.40	14.84	9.24	10.00	50.00	10.00	20.00	10.00
19.07		23.65	40.54	17.82	20.00		20.00	40.00	20.00

TABLE 3 Comparison of predicted and true values of the SR-MLP model.

SR-MLP									
Predicted (μmol/L)					Actual (μmol/L)				
HeLa	A549	HUVEC	PC12	HaCaT	HeLa	A549	HUVEC	PC12	HaCaT
0.11	0.37	0.20	0.26	0.73	0.00	0.00	0.00	0.00	0.00
2.65	12.93	3.35	5.29	2.44	2.50	12.50	2.50	5.00	2.50
4.94	27.29	5.05	10.80	2.92	5.00	25.00	5.00	10.00	5.00
9.92	49.85	8.98	18.95	10.96	10.00	50.00	10.00	20.00	10.00
19.44		19.75	42.48	20.11	20.00		20.00	40.00	20.00

TABLE 4 Features of five cell types screened by SR method.

Dataset	P-value	Filtered features (grayscale values)
HeLa	<0.001	27, 49, 77, 81, 115, 133, 136, 183, 231, 238, 242, 244, 245, 250
A549	<0.05	32, 37, 39, 40, 54, 97, 215, 218, 221, 244
HUVEC	<0.05	16, 18, 20, 41, 47, 48, 52, 54, 117, 159, 180, 186, 191, 198, 220, 223, 244, 247
PC12	<0.01	16, 18, 20, 54, 55, 76, 83, 90, 118, 193, 194, 197, 200, 204
HaCaT	<0.005	12, 37, 52, 56, 64, 66, 67, 77, 122, 124, 127, 159, 172, 192, 248

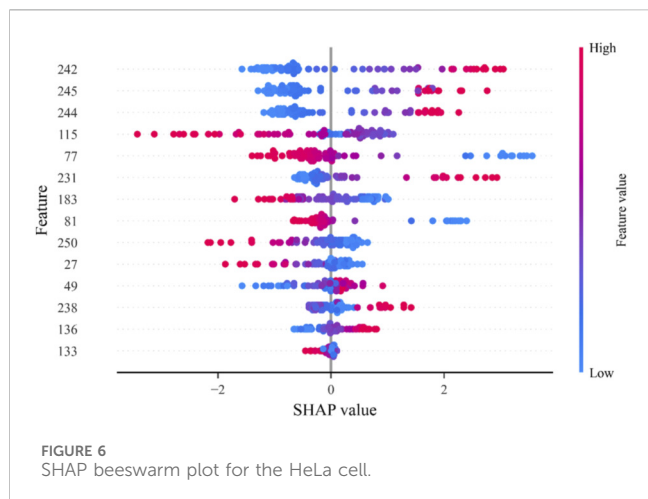
controlled the number of features within 20 to optimize the computational efficiency and generalization capability of the model.

4.3 SHAP-based analysis of the importance of high-grayscale features

To elucidate the contribution of key features in the SR-MLP model to pollutant concentration prediction, we applied the SHapley Additive exPlanations (SHAP) method. Derived from game-theoretic Shapley values, SHAP provides a consistent and interpretable quantification of feature contributions. Using SHAP’s Explainer, we computed SHAP values for key features on the test set and visualized the results with beeswarm plots. In these plots, the X-axis represents SHAP values (indicating directional contribution), the Y-axis lists features in descending order of mean absolute SHAP value (importance), and

point color reflects the original feature magnitude (red for high, blue for low), thus enabling an intuitive assessment of how feature values relate to their predictive influence (Ekanayake et al., 2022).

Analysis results are highly consistent with the key features selected based on statistical significance, strongly confirming the experimental finding that “features in the high grayscale range (200–255) are the core drivers for predicting pollutant concentration.” The SHAP importance ranking reveals that across all cell types, features located within this high-brightness interval dominate (The SHAP beeswarm plot for HeLa cells is shown in Figure 6, while the corresponding results for the remaining cell types (A549, HUVEC, PC12, HaCaT) are provided in the Supporting Information as Supplementary Figures S1–S4). Specifically, in HeLa cells, the top three features by importance are 242, 245, and 244; in A549 cells, they are 244, 221, and 218; in Huvec cells, they are 244, 198, and 48; in PC12 cells, they are 204, 76, and 200; and in HaCaT cells, they are 192, 37, and 172. It is particularly



noteworthy that feature 244 ranks highly in HeLa, A549, and HUVEC cells, indicating its stable predictive ability across different cell types. The results for HaCaT cells show certain particularities: the top-ranked feature in SHAP importance is 192, which, although close to, does not fully fall within the 200–255 high-brightness range. This may suggest that under DIN stimulation, HaCaT cells exhibit differences in brightness response thresholds or grayscale distribution patterns related to morphological damage compared to other cell lines—for example, a smaller increase in brightness during apoptosis or interference from optical properties of other subcellular structures. This phenomenon warrants further investigation in future studies, considering cell type specificity.

In Figure 6, key high-grayscale features (such as 242, 245, and 244) exhibit a distribution where blue points are concentrated on the left side of the X-axis, indicating that their lower feature values are associated with negative SHAP values. This suggests that the model has learned not a simple linear positive correlation but a more complex nonlinear response relationship: it is likely that only when these feature values exceed a certain threshold do they contribute positively and clearly to the prediction of high concentrations. This precise capture of nonlinear dependencies and feature interactions is precisely the intrinsic key to the SR-MLP model’s ability to achieve high-precision predictions.

Overall, although individual features exhibit complex nonlinear influences in their SHAP values, from the perspective of overall feature importance ranking and dominant patterns, these observations align highly with morphological findings: after DIN stimulation, cells contract, become rounded, and gradually detach, leading to a significant increase in brightness in the cell regions, with grayscale values concentrated in the 200–255 range (as shown in Figure 4). Therefore, SHAP analysis confirms, at the level of model decision-making mechanisms, that the model ultimately relies on and integrates the information carried by these high-grayscale features, thereby successfully mapping the biological causal relationship between “the degree of cellular damage (manifested as an increase in high-brightness pixels) and the rise in pollutant concentration.”

4.4 Analysis of model training and convergence

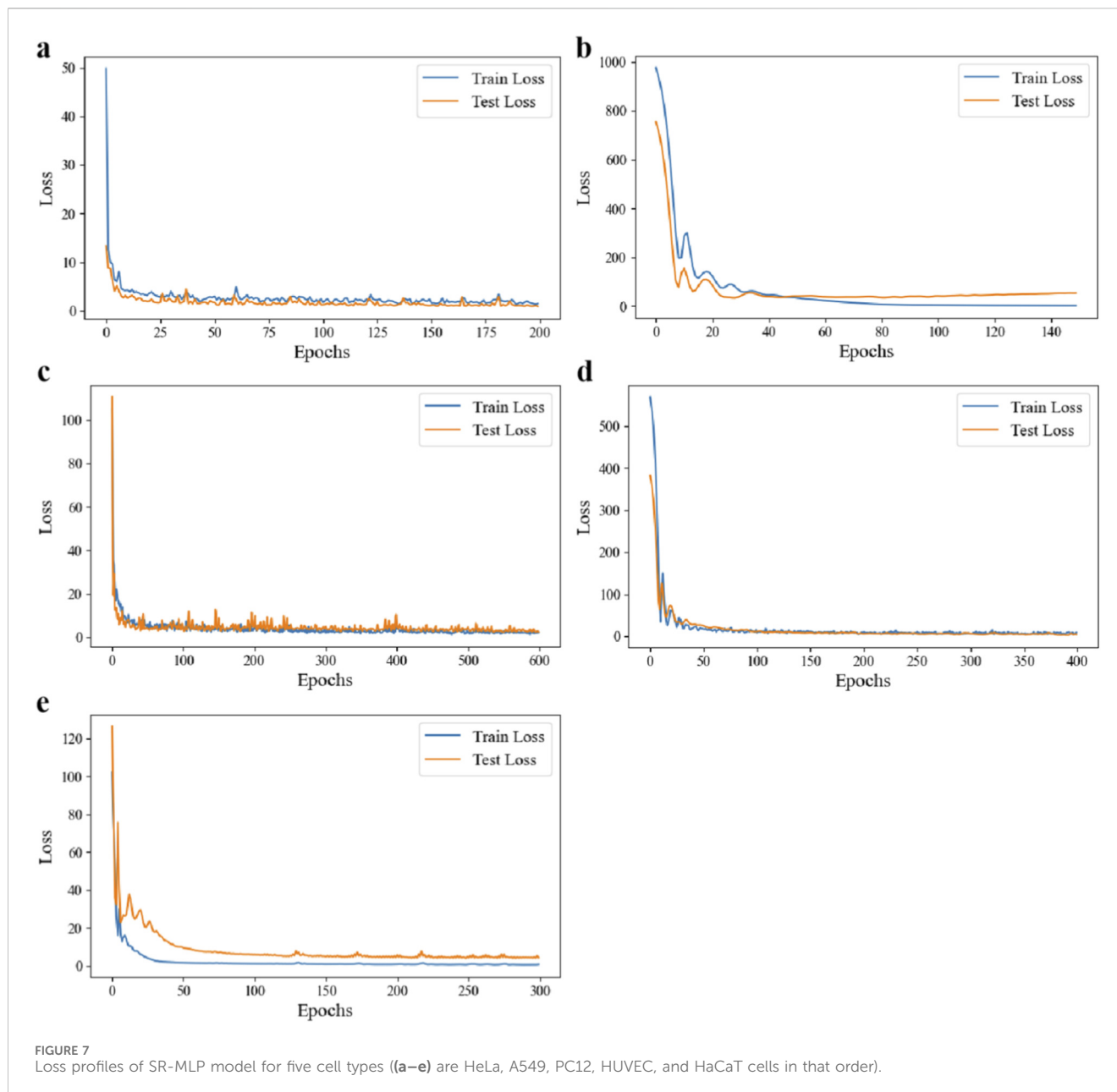
By monitoring the trends of training loss and validation loss, we systematically optimized the hyperparameters of the SR-MLP model. The optimized hyperparameter configurations are presented in Table 5. As shown in Figure 7, after tuning, both the training loss and validation loss exhibited a smooth downward trend and eventually stabilized, maintaining a small gap between them without evident divergence or premature convergence. These results indicate that the model did not suffer from overfitting or underfitting during training, demonstrating good convergence behavior and generalization ability (Wang et al., 2022). This validates the effectiveness of the adopted regularization strategies and hyperparameter configurations.

4.5 Analysis of the superiority of nonlinear modeling methods

Based on the comparison of model performance, we conducted an in-depth analysis of the fundamental reasons

TABLE 5 Hyperparameter list with Mult. (multiplicity coefficient in IQR), Reg. (L2 regularization coefficient), Dropout 1 (first layer Dropout Ratio), Dropout 2 (second layer Dropout Ratio), Dropout 3 (third layer Dropout Ratio), LR (Learning Rate), Epochs (Training rounds) and Batch Size.

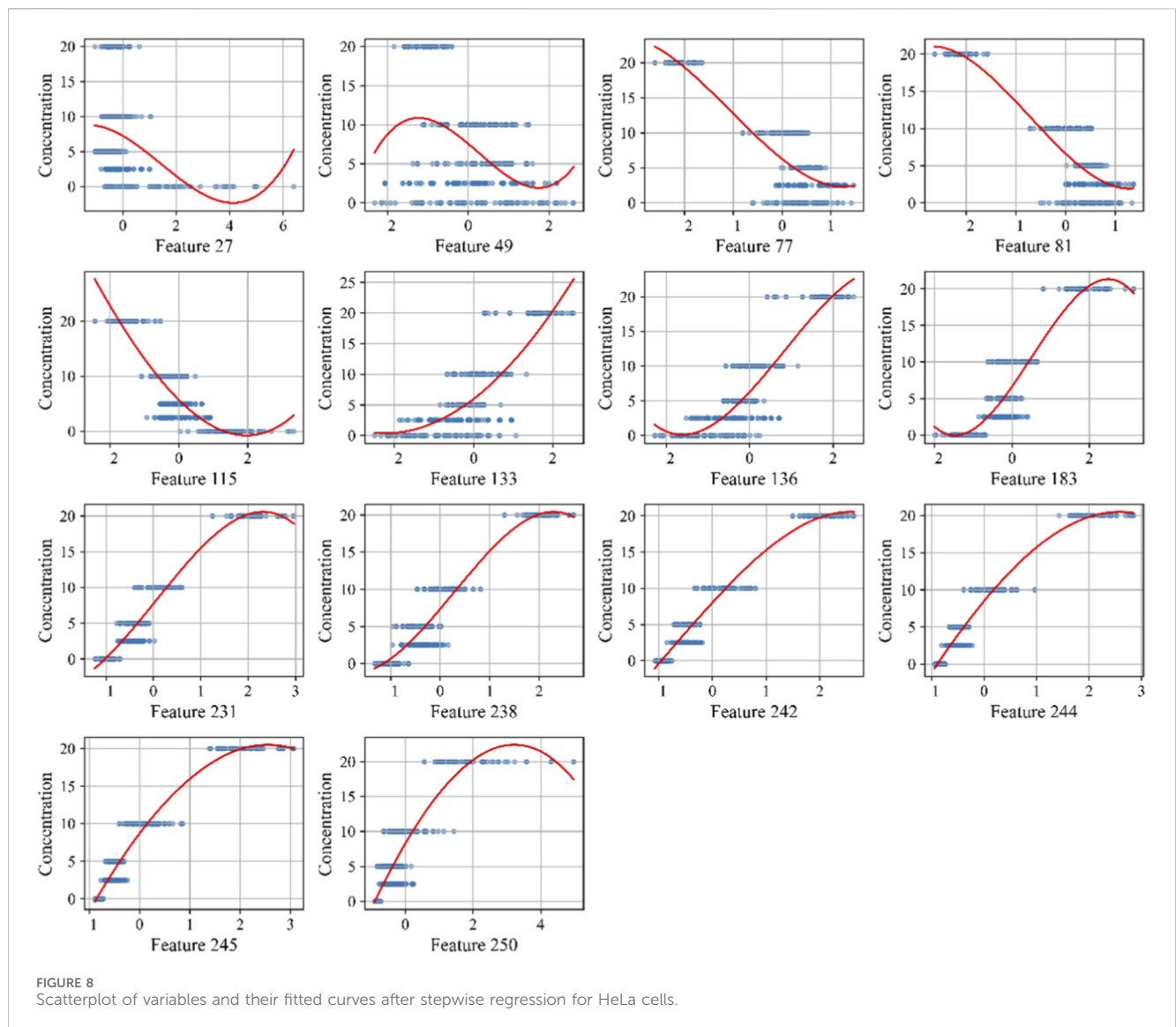
Dataset	Method	Mult	Reg	Dropout_1	Dropout_2	Dropout_3	LR	Epochs	Batch size
HeLa	SR-RR	3	0.01	—	—	—	—	—	—
	SR-MLP	3	0.01	0.1	0.1	0.1	0.01	200	32
A549	SR-RR	2	0.01	—	—	—	—	—	—
	SR-MLP	2	0.01	0	0	0	0.01	150	64
HUVEC	SR-RR	3	0.01	—	—	—	—	—	—
	SR-MLP	3	0.01	0.1	0.1	0.1	0.01	200	32
PC12	SR-RR	2	0.01	—	—	—	—	—	—
	SR-MLP	2	0.01	0.1	0.1	0	0.01	400	64
HaCaT	SR-RR	3	0.01	—	—	—	—	—	—
	SR-MLP	3	0.01	0	0	0	0.01	300	64



behind the superior performance of the SR-MLP model. Analysis reveals that the core reason for the superior predictive performance of SR-MLP over SR-LR and SR-RR is its ability to model nonlinear relationships. The SR-LR and SR-RR model are based on linear regression, which assumes a simple linear relationship between the input features and pollutant concentrations, whereas SR-MLP is able to capture more complex patterns through a multilayered neural network and nonlinear activation functions. To validate the nonlinear relationship between the variables, we visualized the relationship between features and concentrations and used polynomial fit curves to observe data trends. Common manifestations of nonlinearity: the fitted curve is not a straight line, has inflection points, and exhibits significant curvature or fluctuation (e.g., U-shape (Yang et al., 2020), inverted U-shape

(Yang et al., 2020; Northoff and Tumati, 2019), and S-shape (Xiao et al., 2009), etc.). By observing the variable fitting curves of HeLa cells as in Figure 8, we found that the fitting curves of most of the features showed obvious nonlinearities, such as Feature 115 and Feature 133 showed U-shape variations. The results indicate that SR-MLP overcomes the limitations of the SR-LR and SR-RR linearity assumption by modeling nonlinear relationships, and is able to more accurately describe the complex mapping between the gray-scale features of cell damage and pollutant concentrations, improving the fitting ability and prediction accuracy of the model.

Having established the core advantage of the SR-MLP model in handling nonlinear relationships, we further applied it to quantitatively predict environmental pollutant concentrations based on cellular image features. This approach establishes a



direct, data-driven correlation between cellular morphological damage severity and pollutant concentration, offering a novel research perspective for environmental toxicity assessment. The efficacy of the SR-MLP framework in environmental concentration prediction is corroborated by related studies. For instance, Liu et al. successfully calibrated air quality monitoring data using a similar hybrid model (Liu et al., 2021). By transferring and validating this effective modeling strategy at the cellular level, this study proposes a novel, efficient, and reliable computational method for predicting pollutant concentrations based on cellular damage features. This expands the application scope of the hybrid model to microscopic biological response analysis.

4.6 Analysis of method applicability

To validate the applicability of the method in different cell types, we applied the SR-MLP model to four other cell types (A549, HUVEC, PC12, HaCaT) and evaluated the performance of the model on different cell datasets. As can be seen from the data in

Table 4, the SR-MLP model showed relatively small errors between the predicted and true values at each concentration level in the HeLa, A549, HUVEC, and PC12 cell datasets, whereas the predictions of certain concentrations in the HaCaT cell dataset were slightly biased, e.g., the predicted value of 2.92 $\mu\text{mol/L}$ at 5 $\mu\text{mol/L}$ true value, but the predictions of other concentrations were still accurate. This indicates that the model still has good adaptability and generalization ability across different cellular datasets. We found that despite the differences in biological properties and data distribution among cell types, the overall trends were highly similar, and the SR-MLP model was able to accurately predict contaminant concentrations and compare the degree of cellular damage in different cell models.

4.7 Methodological limitations and future directions

To ensure that changes in image grayscale features primarily reflect the toxic effects of pollutants, this study implemented

systematic experimental design and data processing procedures to control for potential confounding factors. Specific measures included: during the cell culture phase, all cell lines were cultured simultaneously in a standard incubator under constant temperature and humidity, using the same batch of culture medium and serum, while controlling passage number and inoculation density to maintain approximate consistency; during the image acquisition phase, all samples were imaged using the same optical microscope and a 10× objective, with fixed exposure time, light intensity, and white balance parameters, and each imaging session was preceded by background correction and uniform field calibration; during the image analysis phase, an automated segmentation pipeline based on a pre-trained model (Cellpose cyto3) was employed, with unified flow threshold and minimum cell size settings to eliminate morphological and grayscale biases due to segmentation variability; during data processing, rigorous outlier removal and data normalization procedures were applied.

Nevertheless, we recognize that certain potential confounding factors are difficult to fully eliminate, such as intrinsic cellular heterogeneity (e.g., cell cycle distribution, metabolic state fluctuations) (Kondo et al., 2021), spontaneous morphological changes unrelated to toxicity that may occur during apoptosis or stress, and potential edge illumination attenuation in microscopic imaging fields despite calibration. These unmeasured or incompletely controlled variables may lead the model to partially rely on non-toxicity-related imaging artifacts or biological noise. Therefore, the quantitative “image-concentration” relationship established in this study should be interpreted as a strong statistical association captured under highly controlled experimental conditions. While this association provides a robust computational foundation and scalable framework for image-based toxicity prediction, the underlying biological causal chain warrants further systematic validation in the future through approaches such as intervention experiments (e.g., co-imaging with oxidative stress probes), multimodal data integration (e.g., combining fluorescence labeling with transcriptomics), and causal inference modeling.

5 Conclusion

This study successfully establishes a novel quantitative “image-to-concentration” prediction paradigm based on cellular grayscale statistical features and an SR-MLP hybrid model. The approach enables high-accuracy, low-cost quantification of environmental pollutants using only conventional optical microscopy images, while demonstrating the broad applicability of high-grayscale features as cross-cellular damage markers. In advancing environmental toxicology, this work contributes methodologically by introducing a data-driven framework that transforms subjective morphological observations into objective concentration predictions, thereby enhancing the standardization and throughput of toxicity assessment. Furthermore, through interpretable analysis, it reveals nonlinear relationships between high-grayscale features and pollutant concentrations, opening new pathways for linking morphological changes to toxic

mechanisms such as apoptosis and oxidative stress. Future research should expand the range of pollutants and toxicity endpoints to build a comprehensive “image-feature-mechanism” database. Developing dynamic monitoring models based on live-cell imaging will advance the method from static detection to real-time tracking and early warning. Additionally, efforts should focus on clarifying the biological significance of image features and establishing cross-platform standardization protocols to promote wider adoption in environmental monitoring, toxicity screening, and risk assessment.

Data availability statement

The dataset for this study and the analysis code supporting the findings are located in the following Google Drive folder: https://drive.google.com/drive/folders/1H9F113bZ4DlOXx1_jkmN_BTLeoY1BktX?usp=sharing.

Author contributions

LZ: Conceptualization, Formal Analysis, Data curation, Investigation, Writing – original draft. AL: Investigation, Writing – original draft, Methodology, Validation. YZ: Investigation, Data curation, Writing – original draft. BJ: Data curation, Formal Analysis, Software, Writing – original draft. CL: Funding acquisition, Project administration, Writing – review and editing. XX: Funding acquisition, Project administration, Writing – review and editing, Conceptualization, Supervision. JJ: Conceptualization, Funding acquisition, Writing – review and editing, Formal Analysis.

Funding

The author(s) declared that financial support was received for this work and/or its publication. We acknowledge financial support from the National Natural Science Foundation of China (Grant Nos. 22536007 and 22274117), Education Department of Guangdong Province (Grant No. 2022ZDJS027).

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the

authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be

evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fenvs.2026.1729913/full#supplementary-material>

References

- Aguinis, H., Vassar, M., and Wayant, C. (2021). On reporting and interpreting statistical significance and p values in medical research. *BMJ Evid-Based Med.* 26 (2), 39–42. doi:10.1136/bmjebm-2019-111264
- Alkhouly, A. A., Mohammed, A., and Hefny, H. A. (2021). Improving the performance of deep neural networks using two proposed activation functions. *IEEE Access* 9, 82249–82271. doi:10.1109/access.2021.3085855
- Atabansi, C. C., Nie, J., Liu, H., Song, Q., Yan, L., and Zhou, X. (2023). A survey of transformer applications for histopathological image analysis: new developments and future directions. *Biomed. Eng. Online* 22 (1), 96. doi:10.1186/s12938-023-01157-0
- Chen, S., Zhao, M., Wu, G., Yao, C., and Zhang, J. (2012). Recent advances in morphological cell image analysis. *Comput. Math. Methods Med.* 2012 (1), 101536. doi:10.1155/2012/101536
- Chicco, D., Warrens, M. J., and Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* 7, e623. doi:10.7717/peerj-cs.623
- Chukhlovin, A. B., Tokalov, S. V., Yagunov, A. S., Westendorf, J., Reincke, H., and Karbe, L. (2001). *In vitro* suppression of thymocyte apoptosis by metal-rich complex environmental mixtures: potential role of zinc and cadmium excess. *Sci. Total Environ.* 281 (1–3), 153–163. doi:10.1016/s0048-9697(01)00843-9
- Ekanayake, I., Meddage, D., and Rathnayake, U. (2022). A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using shapley additive explanations (SHAP). *Case Stud. Constr. Mat.* 16, e01059. doi:10.1016/j.cscm.2022.e01059
- Fu, D. D., Wang, Q. H., and Gao, S. (2020). Analysis of S-Ovalbumin content of different varieties of eggs during storage and its nondestructive testing model by visible-near infrared spectroscopy. *Chin. J. Anal. Chem.* 48 (2), 289–297.
- Hiers, R. D., Khajotia, S. S., Merritt, J., and Esteban Florez, F. L. (2024). Optimization of an ultra-bright real-time high-throughput renilla luciferase assay for antibacterial assessment of *Streptococcus mutans* biofilms. *Dent. Mater* 40 (9), 1313–1321. doi:10.1016/j.dental.2024.06.012
- James, G., Witten, D., and Hastie, T. (2023). *Linear regression [M]. An introduction to statistical learning: with applications in python*. Springer, 69–134.
- Kamiloglu, S., Sari, G., Ozdal, T., and Capanoglu, E. (2020). Guidelines for cell viability assays. *Food Front.* 1 (3), 332–349. doi:10.1002/fft2.44
- Kerr, J. F., Wyllie, A. H., and Currie, A. R. (1972). Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *Br. J. Cancer* 26 (4), 239–257. doi:10.1038/bjc.1972.33
- Kılıçarslan, S., Adem, K., and Çelik, M. (2021). An overview of the activation functions used in deep learning algorithms. *J. New Results Sci.* 10 (3), 75–88. doi:10.54187/jnrs.1011739
- Kondo, H., Ratcliffe, C. D., Hooper, S., Ellis, J., MacRae, J. I., Hennequart, M., et al. (2021). Single-cell resolved imaging reveals intra-tumor heterogeneity in glycolysis, transitions between metabolic states, and their regulatory mechanisms. *Cell Rep.* 34 (7), 108750. doi:10.1016/j.celrep.2021.108750
- Kumunda, C., Adekunle, A. S., Mamba, B. B., Hlongwa, N. W., and Nkambule, T. T. I. (2021). Electrochemical detection of environmental pollutants based on graphene derivatives: a review. *Front. Mater* 7, 616787. doi:10.3389/fmats.2020.616787
- Li, A., Zhao, L., Liu, C., Xu, X., and Jia, J. (2025). Gray frequency-based methodology for assessing cell damage. *ACS Omega* 10 (14), 14084–14093. doi:10.1021/acsomega.4c11226
- Liu, B., Zhao, Q. B., Jin, Y. Q., and Shen, J. (2021). Application of combined model of stepwise regression analysis and artificial neural network in data calibration of miniature air quality detector. *Sci. Rep.* 11 (1), 3247. doi:10.1038/s41598-021-82871-4
- Liu, Y., Zhao, C., and Huang, Y. (2022). A combined model for multivariate time series forecasting based on MLP-Feedforward attention-LSTM. *IEEE Access* 10, 88644–88654. doi:10.1109/access.2022.3192430
- Meher, A. K., and Zarouri, A. (2025). Environmental applications of mass spectrometry for emerging contaminants. *Molecules* 30 (2), 364. doi:10.3390/molecules30020364
- Muhammad, A. L. I. P., and Faraj, R. (2014). Data normalization and standardization: a technical report.
- Northoff, G., and Tumati, S. (2019). Average is good, extremes are bad – non-linear inverted U-shaped relationship between neural mechanisms and functionality of mental features. *Neurosci. Biobehav. Rev.* 104, 11–25. doi:10.1016/j.neubiorev.2019.06.030
- Pereira, S. P., Santos, S. M. A., Fernandes, M. A. S., Deus, C. M., Martins, J. D., Pedroso de Lima, M. C., et al. (2021). Improving pollutants environmental risk assessment using a multi model toxicity determination with *in vitro*, bacterial, animal and plant model systems: the case of the herbicide alachlor. *Environ. Pollut.* 286, 117239. doi:10.1016/j.envpol.2021.117239
- Sadrara, M., and Khorrami, M. K. (2023). Principal component analysis-multivariate adaptive regression splines (PCA-MARS) and back propagation-artificial neural network (BP-ANN) methods for predicting the efficiency of oxidative desulfurization systems using ATR-FTIR spectroscopy. *Spectrochim. Acta A*, 300. doi:10.1016/j.saa.2023.122944
- Salehin, I., and Kang, D.-K. (2023). A review on dropout regularization approaches for deep neural networks within the scholarly domain. *Electronics* 12 (14), 3106. doi:10.3390/electronics12143106
- Sheller-Miller, S., Radnaa, E., Arita, Y., Getahun, D., Jones, R. J., Peltier, M. R., et al. (2020). Environmental pollutant induced cellular injury is reflected in exosomes from placental explants. *Placenta* 89, 42–49. doi:10.1016/j.placenta.2019.10.008
- Shobayo, O., and Saatchi, R. (2025). Developments in deep learning artificial neural network techniques for medical image analysis and interpretation. *Diagnostics* 15 (9), 1072. doi:10.3390/diagnostics15091072
- Stringer, C., Wang, T., Michaelos, M., and Pachitariu, M. (2021). Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* 18 (1), 100–106. doi:10.1038/s41592-020-01018-x
- Suzuki, T., Hidaka, T., Kumagai, Y., and Yamamoto, M. (2020). Environmental pollutants and the immune response. *Nat. Immunol.* 21 (12), 1486–1495. doi:10.1038/s41590-020-0802-6
- Swami, R., Dave, M., and Ranga, V. (2023). IQR-based approach for DDoS detection and mitigation in SDN. *Def. Technol.* 25, 76–87. doi:10.1016/j.dt.2022.10.006
- Wang, W. Y., Zhang, X. F., Lu, K., Dai, B., Zhang, J., Chen, P., et al. (2021). Prediction method of core dead stock column temperature based on PCA and ridge regression. *ISIJ Int.* 61 (11), 2785–2791. doi:10.2355/isijinternational.isijint-2020-497
- Wang, X., Wang, X., Ma, B., Li, Q., and Shi, Y. Q. (2021). High precision error prediction algorithm based on ridge regression predictor for reversible data hiding. *IEEE Signal Proc. Lett.* 28, 1125–1129. doi:10.1109/lsp.2021.3080181
- Wang, Q., Ma, Y., Zhao, K., and Tian, Y. (2022). A comprehensive survey of loss functions in machine learning. *Ann. Data Sci.* 9 (2), 187–212. doi:10.1007/s40745-020-00253-5
- Xiao, J.-Q., Ding, D.-X., Xu, G., and Jiang, F. L. (2009). Inverted S-shaped model for nonlinear fatigue damage of rock. *Int. J. Rock Mech. Min.* 46 (3), 643–648. doi:10.1016/j.ijrmms.2008.11.002
- Xie, X., Xie, M., Moshayedi, A. J., and Noori Skandari, M. H. (2022). A hybrid improved neural networks algorithm based on L2 and dropout regularization. *Math. Probl. Eng.* 2022 (1), 8220453. doi:10.1155/2022/8220453

Xin, L., Wang, J., Wu, Y., and Guo, S. (2015). The development of GADD45 α luciferase reporter assays in human cells for assessing the genotoxicity of environmental pollutants. *Toxicol. Mech. Methods* 25 (2), 136–142. doi:10.3109/15376516.2014.1003357

Yang, G., Zha, D., Wang, X., and Chen, Q. (2020). Exploring the nonlinear association between environmental regulation and carbon intensity in China: the mediating effect of green technology. *Ecol. Indic.* 114, 106309. doi:10.1016/j.ecolind.2020.106309

Yang, L., Cai, X., and Li, R. (2024). Ferroptosis induced by pollutants: an emerging mechanism in environmental toxicology. *Environ. Sci. Technol.* 58 (5), 2166–2184. doi:10.1021/acs.est.3c06127

Zhou, X., and Jiang, T. (2016). Metamodel selection based on stepwise regression. *Struct. Multidiscip. Optim.* 54 (3), 641–657. doi:10.1007/s00158-016-1442-1

Zhu, M., Guan, X. R., Li, Z., He, L., Wang, Z., and Cai, K. (2023). sEMG-Based lower limb motion prediction using CNN-LSTM with improved PCA optimization algorithm. *J. Bionic Eng.* 20 (2), 612–627. doi:10.1007/s42235-022-00280-3