Journal of Zhejiang University-SCIENCE A in press
www.jzus.zju.edu.cn; www.springer.com/journal/11582
E-mail: jzus_a@zju.edu.cn

Research Article
https://doi.org/10.1631/jzus.A2500545

A novel robust cross-modal integration fusion model for rapid moisture content detection in concrete sand

Zhijian CAI, Jun ZHANG✉, Xiaoling WANG, Jiajun WANG, Kehao ZHAO, Guohua WU

State Key Laboratory of Hydraulic Engineering Intelligent Construction and Operation, Tianjin University, Tianjin 300350, China

Abstract: The rapid and accurate detection of concrete sand moisture content (MC) is crucial for ensuring concrete quality. However, existing unimodal detection methods are constrained by limited representative features and lack robustness. Multimodal operations often involve simple concatenation of features from different modalities, lacking potential interactivity among features. To address this issue, a novel robust cross-modal integration fusion model, which uses five branches to extract the features of images, near-infrared spectrum, and dielectric constant and a multilevel cross-modal integration fusion network to fuse these features, is proposed for the rapid detection of MC in concrete sand. Specifically, the multilevel cross-modal integration fusion network comprises a feature attention module, a cross-modal self-attention fusion module, and an integrated output module. The feature attention module enhances the feature representation from each modality, reducing the interference from redundant features and noise. The cross-modal self-attention fusion module employs a residual self-attention mechanism to deeply mine and fuse interactions between modalities while retaining low-level features, improving model accuracy and stability. The integrated output module is utilized to obtain more robust prediction results. The results show that the proposed model outperformed unimodal, traditional multimodal, and cross-modal methods on our concrete sand dataset, achieving excellent and robust prediction results for both machine-made sand (RMSE = 0.458, R^2 = 0.983, RPD = 7.9) and natural sand (RMSE = 0.705, R^2 = 0.984, RPD = 7.931). The detection time was within 71 seconds, significantly enhancing the detection frequency and efficiency, which provides a reliable solution for the rapid detection of MC in concrete sand.

Key words: Concrete sand; Rapid moisture content detection; Cross-modal integration fusion; Robust prediction

1 Introduction

Fine aggregate such as sand is one of the most common building materials in the construction industry and has been widely used in concrete production. Fine aggregate stockpiled in open air exhibits high fluctuation in moisture content (MC), particularly during the rainy season. The high water content fluctuation directly affects the water–cement ratio of concrete and influences its physical and mechanical properties. It has been shown that slight changes in the water–cement ratio can significantly affect the strength of concrete (Gavela et al., 2018). Amlashi et al. (2019) found that the amount of added water had the greatest effect on concrete collapse. In concrete mixing plants, cementitious materials are sealed in storage tanks, and under the premise of ensuring accurate measurements, the direct weighing of these materials has less impact on the water–cement ratio. However, fine aggregates are stacked bare during engineering construction. According to engineering statistics, fine aggregate has a small particle size and high water absorption, and its MC can fluctuate within the range of 1 wt.-% to 15 wt.-%. Therefore, the MC of sand is one of the most important factors affecting the water–cement ratio. According to the standard for technical requirements and test method of sand and crushed stone (or gravel) for ordinary concrete (JGJ52−2006), most of the concrete mixing plants currently use the traditional bin sampling and drying method to detect the MC of fine aggregate. However, this detection method is time consuming (8 hours for the drying

✉ Jun ZHANG, zhangdajun@tju.edu.cn
Jun ZHANG, https://orcid.org/0000-0002-8429-3693
Zhijian CAI, https://orcid.org/0009-0004-8841-751X

Received Oct. 24, 2025; Revision accepted Apr. 17, 2026; Crosschecked

method), inefficient, and has poor sample representation. Therefore, the development of a rapid and accurate method for measuring the MC is of great practical significance and application value.

MC detection methods based on data acquired from physical sensing devices have become prevalent in current applications. Recently, many studies based on unimodal data have been applied to the detection of MC. Near-infrared (NIR) spectroscopy offers several advantages, including its nondestructive, rapid, and simple (Watanabe et al., 2021) nature. This has led to its widespread adoption in evaluating the MC of tea, construction and demolition waste characteristics, fruit quality and soil characteristics (Cao et al., 2024; Huang et al., 2021; Ng et al., 2019; Radica et al., 2024). The principle lies in the overtones and vibrational combinations of molecular bonds such as O−H, C−H, and N−H, and the hydrogen-containing chemical components have unique absorption features in the NIR spectra (Miao et al., 2023). Many chemometric methods have been proposed for classification and regression problems based on NIR spectral data, including partial least squares (Guo et al., 2023), support vector regression (SVR) (Zhang et al., 2021), random forests (Nawar and Mouazen, 2019), and artificial neural networks (ANNs) (Wang et al., 2021). It should be noted that full NIR spectral data contain a large amount of noise and irrelevant variables, which not only increase the complexity of data processing but also reduce the robustness of the model (Miao et al., 2023). Therefore, some NIR data preprocessing methods have been proposed to solve the above problems, for example, standard normal variable transformation, multivariate scattering correction, multiorder derivatives, and wavelet transform (Abasi et al., 2019), to eliminate large amounts of noise and extraneous variables, thereby improving the accuracy of the model. However, specific preprocessing methods are only applicable to a limited number of samples. As data samples are continuously collected, factors such as equipment aging, fluctuations in ambient temperature and humidity, and shifts in equipment calibration may introduce more random errors over time, and conventional preprocessing methods are no longer suitable, leading to reduced generalization ability and robustness of the model (Ni et al., 2019). At the same time, preprocessing increases the modeling difficulty.

In recent years, the advent of deep learning has revolutionized the field of chemometrics, with its powerful linear and nonlinear feature extraction capabilities outperforming traditional chemometric methods. As an end-to-end method, deep learning can automatically extract complex nonlinear features without data preprocessing, simplifying the difficulty of modeling and enhancing the generalization ability of the model. Therefore, convolutional neural networks (CNNs) with strong local abstract feature extraction and recurrent neural networks (RNNs) with powerful positional feature extraction have gradually been applied to spectral analysis (Wang et al., 2021; Zhang et al., 2019). Yang et al. (2020) combined a CNN and gated recurrent unit (GRU) framework for soil property prediction. The combined model achieved a better performance than the baseline model. Wang et al. (2021) developed an end-to-end deep learning network based on an inception network for full-waveband spectral feature extraction and soil nitrogen content (STN) prediction. Song et al. (2025) proposed a novel deep learning network (DiSENet), which showed superiority over traditional calibration methods for green tea MC prediction. Yuan et al. (2022) introduced a novel network architecture that combines a one-dimensional CNN and attention-based bigated RNN (AT-BiGRU) for NIR abstract and positional feature extraction and sand gravel MC detection, which achieved better prediction results than other deep learning and traditional chemometrics methods. The frequency domain reflection (FDR) technique for measuring the dielectric constant (DC) of substances is another rapid method for detecting soil MC. It allows for the rapid measurement of MC through the inversion of the DC of the soil and application of an alternating current (Fragkos et al., 2024). This is because the DC of freshwater-amended wet soil is primarily dependent on its MC. Mouazen et al. (2018) combined DC and NIR spectroscopy to detect soil volumetric MC, which showed the feasibility of this approach for achieving relatively high-precision prediction results within a certain MC range. Krzeminska et al. (2022) used the DC measured by FDR to inversely derive the soil MC. Their results effectively reflected the point information of the soil MC. Han et al. (2022) used the DC measured by an FDR sensor to monitor temporal variations in soil MC. Therefore, using the dielectric

properties of soil for MC detection is regarded as an effective, rapid, simple, and reliable method (Dean et al., 1987). Moreover, with the continuous advancement of imaging technology, researchers have conducted extensive studies on the detection of soil moisture, soil salinity, and related parameters utilizing technologies such as smartphones, unmanned aerial vehicles (UAVs), and remote sensing, revealing the potential applications of image-based techniques in the field of soil property assessment (Somkunwar et al., 2024; Tobiszewski and Vakh, 2023; Zhu et al., 2022). Khalkho et al. (2024) estimated soil moisture based on composite optical band images consisting of blue, green, and red bands collected by UAV in conjunction with the normalized difference moisture index. Nijaguna et al. (2023) derived vegetation indices (VI) from satellite imagery and combined them with an improved water cloud model and a deep learning model to detect soil moisture, achieving high detection accuracy. Rahman et al. (2025) developed a deep neural network termed MoistNetMax, which extracts wood chip image features for MC detection and demonstrated superior prediction performance compared to prevailing state-of-the-art deep learning models.

Most of the above studies applied single-modality data for modeling and have achieved some competitive results in MC detection. However, the limited representative features hinder further improvement in the detection accuracy of related tasks. Moreover, in the field of soil MC detection, near-infrared spectroscopy is susceptible to the influence of soil particle size, which in turn affects the measurement of absorbance (Wu et al., 2021). The soil DC is not only influenced by soil organic matter but also varies with changes in temperature, dry density, and salinity. Image data can be affected by environmental lighting, acquisition equipment, and the distribution of sand particles. Furthermore, due to external factors such as equipment aging, prolonged operation, and calibration errors, the measurement process inevitably introduces photon, thermal, electronic, and quantization noise, which approximately follow a Gaussian distribution. These factors contribute to the lack of robustness in MC measurements based on single-modal data. Therefore, it is essential to integrate multiple modalities of data to achieve more competitive detection results.

Multimodal data fusion provides a feasible approach to addressing the aforementioned issues, which integrates heterogeneous data from different modalities to exploit the complementarity between data and provide better predictive performance (Jaafar and Lachiri, 2023). Recently, moisture detection models based on multimodal data, including NIR, multispectral, thermal, and image data, have progressed in the field of agricultural and soil moisture detection (Liu et al., 2022; Song et al., 2021; Zuo et al., 2023). Cheng et al. (2022) investigated the impact of fusing RGB image, multispectral, and thermal sensor data on the accuracy of soil MC estimation. The results indicated that multimodal data achieved better estimation accuracy than single-modal data. Wu et al. (2024) extracted VI metrics from multimodal UAV remote sensing data, including RGB image, thermal infrared, and multispectral data, and constructed a hybrid CNN-LSTM model to achieve high-accuracy soil moisture estimation. Li et al. (2024) utilized characteristic wavelengths, RGB images, and thermal imaging data, incorporating two feature extraction branches to separately extract intermediate features from images and wavelengths, and then fused these features to achieve simultaneous detection of soil organic matter and MC. However, most multimodal fusion operations often involve data-level fusion or simple concatenation of features from different modalities, which ignores the heterogeneity and interaction between different modalities. Cross-modal feature fusion methods, which can consider the interaction between different modalities, offer new insights into the above problems (Hua et al., 2022; Shu et al., 2023; Wei et al., 2025). Yuan et al. (2024) proposed a Deep Multimodal Fusion (DMF) model that currently represents the state-of-the-art in multimodal fusion approaches for MC detection in sand gravel to integrate multimodal data—including NIR, RGB image, and DC—for MC detection, achieving satisfactory measurement accuracy. He et al. (2023) developed a coattention fusion network that achieved robust prediction results by using two branches based on the attention mechanism to extract features from dermoscopy and clinical images and a hyperbranch network for refining and fusing these features from the two branches. Li et al. (2023) constructed a residual cross-modal fusion attention

module for speech and text multimodal information fusion, effectively extracting intermodal interactions and achieving state-of-the-art results. Gan et al. (2024) proposed a novel multimodal emotion analysis model for videos, which demonstrated excellent performance under both word-aligned and nonaligned settings. Tong et al. (2025) proposed two cross-modal feature fusion modules designed to integrate mid-level and high-level features from multiple modalities, thereby enhancing the precision of robotic object grasping. The previously summarized cross-modal fusion methods offer new perspectives for rapid MC detection. However, most of the research has primarily concentrated on cross-modal interactions at the middle or final stages of different modalities, overlooking the combined contribution of cross-modal features at different hierarchical levels to the final MC output and the corresponding interpretability analysis.

To address these issues, based on the multimodal data (NIR, images, and DC) mentioned earlier for MC detection, a novel robust cross-modal integration fusion model for rapid detection of MC in concrete sand is proposed in this study. The major contributions of this study are as follows:

(1) A novel robust cross-modal integration fusion model is proposed for the rapid and accurate MC detection of concrete sand, which can adapt to the influence of complex engineering environmental changes on the concrete sand MC and help to improve the management level of concrete production quality in the construction stage.

(2) A multilevel cross-modal integration fusion network is proposed to comprehensively consider both intramodal and intermodal feature interactions. Specifically, the multilevel cross-modal integration fusion network comprises a feature attention module, a cross-modal self-attention fusion module, and an integrated output module. The feature attention module enhances the feature representation from each modality, reducing the interference from redundant features and noise. The cross-modal self-attention fusion module employs a residual self-attention mechanism to deeply mine and fuse interactions between modalities while retaining low-level features, improving model accuracy and stability. The integration output module consolidates the results from different levels of fused features to obtain more robust MC detection results.

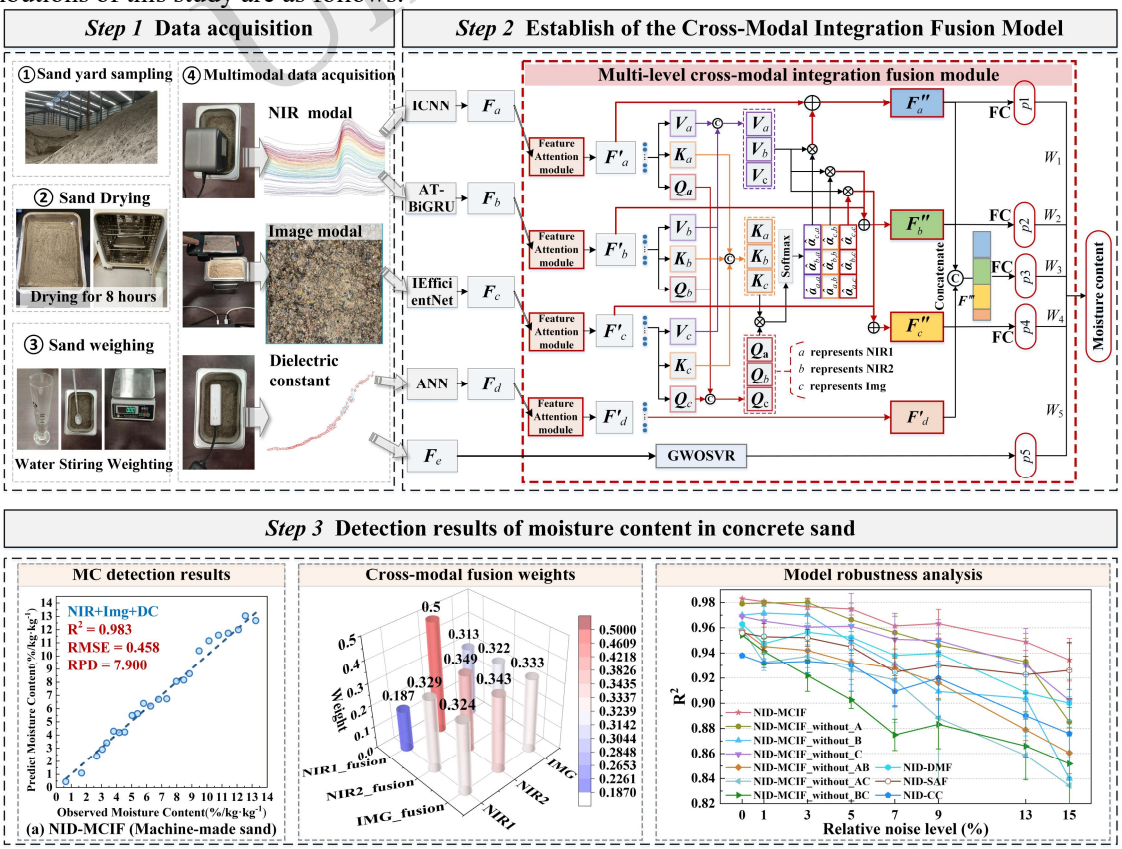


Fig. 1 Research framework

2 Research framework

Fig. 1 presents the research framework of this study, which consists of three parts: (1) Data acquisition. The onsite sand yard of a hydropower station in southwest China was selected as the research object. In accordance with the on-site MC testing workflow for fine aggregates, this study conducted multimodal data-based MC detection via on-site sampling and subsequent testing in a field laboratory. Specifically, various moisture states were simulated by adding different amounts of water to the dry sand materials. Multimodal data were collected using a near-infrared spectrometer, a smartphone, and a DC measurement device. The data were then divided into training, validation, and test datasets. (2) Establish a cross-modal integration fusion model. Based on the acquired training and validation datasets, a robust cross-modal feature fusion model for sand MC detection was developed by combining five unimodal branch models with the proposed multilevel cross-modal integration fusion network. (3) Based on the proposed cross-modal integrated fusion model, the MC of the test dataset was detected, and the superiority of the proposed model over the comparative models was validated. Additionally, the robustness of the proposed model was discussed under different types and levels of noise in the test dataset.

3 Data acquisition

The concrete sand used in this study was obtained from a concrete sand yard of a large-scale water conservancy project in southwest China. Based on engineering data, the fineness moduli of natural sand and machine-made sand are 1.77 and 2.87, respectively. Three gradation tests were conducted on the sampled sand, and the results are displayed in Fig. S1 of the electronic supplementary materials (ESM). The fineness modulus of the sampled sand is consistent with that of the sand in the field, indicating good representation. Following the JGJ 52−2006 standard, the collected sand samples were weighed in a preweighed container and dried in an electric oven at 105 °C. The fine aggregate was stirred thoroughly every 30 minutes of drying. After 6 hours of drying, the samples were cooled to room temperature in an experimental cabinet and weighed. This weighing process was repeated every 30 minutes until a constant mass was achieved (defined as a mass difference of < 0.1% between successive weighings). After drying and natural cooling processes, the natural and machine-made sands were evenly divided into three groups. Each group was then placed into a dry aluminum box and sealed with an impermeable membrane. Each set of samples was weighed on an electronic scale, and the masses of the corresponding aluminum box, mixing spoon, and dry sand were recorded. It should be noted that the ambient environment was regulated via an air conditioning system to maintain a temperature of 25 °C and a relative humidity of approximately 60%. By limiting the natural cooling time to within 30 minutes, the influence of ambient humidity on the dried fine aggregates was effectively minimized. Subsequently, a random amount of tap water—consistent with the water source used in the actual engineering project—was added to each set of samples and stirred thoroughly to establish their initial MC. It should be noted that the mixing spoon was included in the subsequent weighing to avoid mass loss caused by small amounts of sand and water adhering to the spoon. The formula for calculating the MC of the concrete sand is given by Eq. (1).

$$w = \frac{m_{\mathrm{w}} - m_{\mathrm{d}}}{m_{\mathrm{d}} - m_0} \tag{1}$$

where w denotes the MC of the sand (wt.%); m_{w} represents the total mass of the wet sand, aluminum box, and spoon (kg); m_{d} is the total mass of the dry sand, aluminum box, and spoon (kg); and m_0 is the total mass of the aluminum box and spoon (kg).

After initializing the MC of all sand samples, multimodal data acquisition was carried out. For detailed procedures, please refer to Section S3 of the ESM. Finally, 147 machine-made and 183 natural sand samples were obtained. The statistical characteristics of the NIR spectra, DC, and MC of the measured sand are listed in Table S1 of the ESM.

4 Methodology

As described in Section 3, this study acquired three modalities of data for the sand samples: NIR spectra, images, and DC. The NIR spectra cover a wavelength range of 1350 nm to 2150 nm, containing absorbance features across 160 continuous bands. As demonstrated by (Yuan et al., 2022), NIR spectral curves exhibit distinct local characteristic peaks and sequential patterns that vary with wavelength, reflecting significant abstract and sequential positional properties. Consequently, these abstract and sequential positional characteristics are comprehensively considered during NIR spectral modeling in this study. Furthermore, the experimental results indicate that the surface color, texture, and particle distribution states of the sand vary with MC. Therefore, when processing the image modality data, the physical input consists of the entire camera field of view, which is resized into a complete 224 × 224 pixel RGB three-channel image. The modeling of these images aims to extract high-dimensional feature representations of the darkening color and altered surface texture associated with varying moisture levels, thereby capturing their nonlinear mapping relationship with the MC. The DC is unidimensional data, serving as a single scalar value that reflects the MC. During modeling, its fusion with the high-dimensional NIR spectra and image features, along with its specific contribution to decision-level fusion, is fully considered. The proposed theory is detailed below from two main aspects: unimodal feature extraction and multimodal feature fusion.

4.1 Unimodal feature extraction method

4.1.1 NIR feature extraction under the ICNN and AT-BiGRU frameworks

Inspired by Yuan et al. (2022), this study designed a dual-branch architecture to separately extract abstract and positional features from NIR data. CNNs can extract local and abstract features from spectral data via multilayer convolution and pooling, as described in detail in (Yang et al., 2020). Currently, the most popular CNNs stack the convolutional layers in increasingly deeper configurations to achieve better results (Inthiyaz et al., 2023); however, this leads to degradation in the network performance. The Inception model (Szegedy et al., 2015), which improves the network performance by increasing both the depth and width of the network, can solve the aforementioned problem. In addition, depthwise separable convolution (DWConv) (Shaheed et al., 2022) can balance model accuracy and operational complexity, significantly reducing the number of parameters. Therefore, DWConv is employed as a replacement for ordinary convolution to extract the NIR features in this study. Recently, the channel spatial attention mechanism (CBAM) has been widely used to perform adaptive weighting on the channel and spatial dimensions of feature maps to enhance key information (You et al., 2025). Therefore, to effectively extract key features from NIR spectra, this study combines the strengths of DWConv, the Inception architecture, and CBAM to propose an improved CNN (ICNN) for abstract spectral feature extraction, as illustrated in Fig. S2 of the ESM.

The GRU model, which can effectively alleviate the gradient vanishing problem in RNNs and shorten the training time, was first proposed by Chung et al. (2014) for sequence modeling. The GRU uses the so-called update gate and reset gate (Liu et al., 2020). In the GRU, the states are always output from forward to reverse, ignoring the dependency between spectral sequence positions. A BiGRU consists of forward and reverse GRUs that can effectively extract the interdependencies between the positions of spectral sequences. However, fully connected networks cannot consider the importance of different sequences when processing long sequence samples, which adversely affects the performance of the model. The multihead self-attention mechanism, which can efficiently identify correlations between different parts of an input, has been widely used in natural language processing (Vaswani et al., 2017) and fault detection of wind turbine gearboxes (Yu et al., 2024). Therefore, an improved GRU model (AT-BiGRU) (Yuan et al., 2022) model that integrates a BiGRU with a multihead self-attention mechanism was proposed to extract the potential positional features of spectral sequences. The network structure of the AT-BiGRU is shown in Fig. S3 of the ESM.

4.1.2 Image feature extraction under improved EfficientNet

EfficientNet is an architectural and scaling technique for convolutional neural networks that uses

a compound coefficient to scale all depth, breadth, and resolution dimensions evenly (Tan and Le, 2019). EfficientNet B0 is composed of 2D Depthwise convolution blocks, which have been proven to be incredibly efficient in terms of both processing time and cost. Therefore, EfficientNet-B0 was selected as the image feature extractor. To further enhance the model's robustness against measurement noise, an improved EfficientNet was constructed. Specifically, the CBAM was embedded into the final convolutional layer to improve the extraction of key features, and the network architecture is illustrated in Fig. S4 of the ESM. To adapt to the task of sand MC detection, we applied transfer learning by utilizing the pretrained weights of EfficientNet_B0 available in the Torchvision open-source library. The layer after the global average pooling layer in the original EfficientNet_B0 was replaced with our fully connected layer, and the model was fine-tuned specifically for sand MC detection. In addition, image flipping and standardization were employed to enhance the generalization ability of the model.

4.1.3 DC feature extraction under ANN and GWOSVR

The DC features exhibit significant nonlinear characteristics. SVR, based on the principle of structural risk minimization (SRM) in statistical learning theory, can effectively address highly nonlinear regression and classification problems (Fan et al., 2023). Therefore, this study leverages the advantages of SVR in nonlinear modeling to construct a regression model between DC features and MC. The SVR regression problem is detailed in Section S4.1.3 of the ESM. The selection of SVR hyperparameters determines the quality of the model's fitting capability. Therefore, the gray wolf optimizer (GWO), characterized by its simple structure and ease of adjustment, was used to optimize the hyperparameters, and an accurate MC detection model GWOSVR was obtained. The specific principles are detailed in (Fan et al., 2023). The establishment of the DC feature-based GWOSVR model provides a reliable MC prediction branch for the multilevel fusion strategy within the final MC detection framework under small-sample conditions. Furthermore, to address the challenge of integrating the unidimensional DC feature with high-dimensional NIR spectral and image features, an ANN was used for nonlinear mapping to obtain high-dimensional nonlinear features. The structure of the ANN is presented in Table 1.

Table 1 The structure of the ANN model

Input	Hidden1	Hidden2	Output
1	10	5	1

4.2 Multilevel cross-modal integration fusion model

Most existing MC detection models have lacked consideration of the potential interactions among multimodal features. To effectively integrate features at different levels and enhance interactions between both intramodal and intermodal modalities and further leverage the contributions of features at various levels to the final output, a multilevel cross-modal integration fusion (MCIF) model was proposed in this study, and the structure of the proposed model is illustrated in Fig. 2. The MCIF comprises a feature attention module (A), a cross-modal self-attention fusion module (B), and an integrated output module (C). In module A, considering the impact of noise and redundant features within multimodal features, two learnable parameter matrices, $\boldsymbol{U} \in \mathbb{R}^{1\times n}$ and $\boldsymbol{W} \in \mathbb{R}^{n}$, were introduced for calculating the intramodal feature attention scores. Subsequently, feature weights were obtained and multiplied with the multimodal features to enhance the representation of intramodal features while reducing the impact of noise and redundant data. This principle is expressed in Eqs. (2)–(4).

$$\boldsymbol{F}_k^i = \boldsymbol{U}\tanh\left(\boldsymbol{W}\boldsymbol{F}_k^i + b\right), i = 1,2,\cdots,n \tag{2}$$

$$\boldsymbol{a}_k^i = \frac{\exp\left(\boldsymbol{F}_k^i\right)}{\sum_{i=0}^{n}\sum_{i=0}^{n}\exp\left(\boldsymbol{F}_k^i\right)}, i = 1,2,\cdots,n \tag{3}$$

$$\boldsymbol{F}_k^{i'} = \boldsymbol{a}_k^i \times \boldsymbol{F}_k^i, i = 1,2,\cdots,n \tag{4}$$

where k represents one of the following modalities: the abstract feature modality of NIR data (a), the positional feature modality of NIR data (b), or the image modality (c), n denotes the number of intramodal features, and $\boldsymbol{F}_k^i$ denotes the i-th feature of the k-th modality. The weight matrix $\boldsymbol{W}$ and bias term b are used to map $\boldsymbol{F}_k^i$ to the attention space, which is then multiplied by the weight matrix $\boldsymbol{U}$ to

obtain the attention score, $\boldsymbol{F}_k^i$, which is the attention score computed for the i-th feature of the k-th modality. $\boldsymbol{F}_k^i$ is fed into softmax to generate the feature attention weights $\boldsymbol{a}_k^i$, which are multiplied by $\boldsymbol{F}_k^i$ to obtain the enhanced unimodal feature $\boldsymbol{F}_k^{i'}$.

Module B was proposed for intermodal relation extraction and deep feature fusion. Specifically, the intramodal enhancement feature $\boldsymbol{F}_k' \in \mathbb{R}^{1\times n}$ was used as the input, and three weight matrices, namely, $\boldsymbol{Q}_k \in \mathbb{R}^n$, $\boldsymbol{K}_k \in \mathbb{R}^n$, and $\boldsymbol{V}_k \in \mathbb{R}^n$, were obtained by feature mapping through three weight vectors: $\boldsymbol{W}_k^1 \in \mathbb{R}^{1\times n}$, $\boldsymbol{W}_k^2 \in \mathbb{R}^{1\times n}$, and $\boldsymbol{W}_k^3 \in \mathbb{R}^{1\times n}$. $\boldsymbol{Q}_k \in \mathbb{R}^n$ and $\boldsymbol{K}_k \in \mathbb{R}^n$ were dot multiplied to compute the similarity between the sequences and obtain the weights, which were then divided by $\sqrt{d_k}$ for scaling to prevent the result from falling on a very small gradient. Subsequently, the weights were normalized using the softmax function. Finally, the weighted values of the weights with the corresponding weight matrices $\boldsymbol{V}_k \in \mathbb{R}^n$ were utilized to obtain fusion features $\boldsymbol{F}_k'' \in \mathbb{R}^{1\times n}$ that consider intermodal correlations. In addition, to minimize the loss of modal information and avoid gradient vanishing and explosion during the training process, we further fused the intramodal enhancement features with the intermodal fusion features by adding residual connections. This concept is expressed in Eqs. (5)–(11)

$$\boldsymbol{Q}_k = \boldsymbol{W}_k^1 \times \boldsymbol{F}_k', \boldsymbol{K}_k = \boldsymbol{W}_k^2 \times \boldsymbol{F}_k', \boldsymbol{V}_k = \boldsymbol{W}_k^3 \times \boldsymbol{F}_k' \tag{5}$$

$$\boldsymbol{a}_{m,k} = \frac{\boldsymbol{Q}_m \bullet \boldsymbol{K}_k^{\mathrm{T}}}{\sqrt{d_k}} \tag{6}$$

$$\hat{\boldsymbol{a}}_{m,k} = \mathrm{softmax}\left(\boldsymbol{a}_{m,k}\right) \tag{7}$$

$$\boldsymbol{F}_{\mathrm{a}}'' = \hat{\boldsymbol{\alpha}}_{\mathrm{a,a}}\boldsymbol{V}_{\mathrm{a}} + \hat{\boldsymbol{\alpha}}_{\mathrm{b,a}}\boldsymbol{V}_{\mathrm{b}} + \hat{\boldsymbol{\alpha}}_{\mathrm{c,a}}\boldsymbol{V}_{\mathrm{c}} + \boldsymbol{F}_{\mathrm{a}}' \tag{8}$$

$$\boldsymbol{F}_{\mathrm{b}}'' = \hat{\boldsymbol{\alpha}}_{\mathrm{a,b}}\boldsymbol{V}_{\mathrm{a}} + \hat{\boldsymbol{\alpha}}_{\mathrm{b,b}}\boldsymbol{V}_{\mathrm{b}} + \hat{\boldsymbol{\alpha}}_{\mathrm{c,b}}\boldsymbol{V}_{\mathrm{c}} + \boldsymbol{F}_{\mathrm{b}}' \tag{9}$$

$$\boldsymbol{F}_{\mathrm{c}}'' = \hat{\boldsymbol{\alpha}}_{\mathrm{a,c}}\boldsymbol{V}_{\mathrm{a}} + \hat{\boldsymbol{\alpha}}_{\mathrm{b,c}}\boldsymbol{V}_{\mathrm{b}} + \hat{\boldsymbol{\alpha}}_{\mathrm{c,c}}\boldsymbol{V}_{\mathrm{c}} + \boldsymbol{F}_{\mathrm{c}}' \tag{10}$$

$$\boldsymbol{F}''' = \mathrm{Concat}\left(\boldsymbol{F}_{\mathrm{a}}'', \boldsymbol{F}_{\mathrm{b}}'', \boldsymbol{F}_{\mathrm{c}}''\right) \tag{11}$$

where $m,k \in \{\mathrm{a}, \mathrm{b}, \mathrm{c}\}$, $\boldsymbol{a}_{m,k}$ and $\hat{\boldsymbol{a}}_{m,k}$ represent the feature score and feature weight of multimodality, respectively. $F_{\mathrm{a}}'' \in \mathbb{R}^{N\times n}$, $F_{\mathrm{b}}'' \in \mathbb{R}^{N\times n}$, and $F_{\mathrm{c}}'' \in \mathbb{R}^{N\times n}$ are the NIR abstract fusion features, NIR positional fusion features, and image fusion features considering the correlation between modalities, respectively. For the DC feature, the ANN described in Section 4.1.3 is employed to perform nonlinear transformation, generating 10 nonlinear features for multilevel feature fusion. After obtaining the DC features and fusion features for each modality, the cross-modal fusion features $\boldsymbol{F}''' \in \mathbb{R}^{N\times(3n+10)}$ were obtained by concatenation.

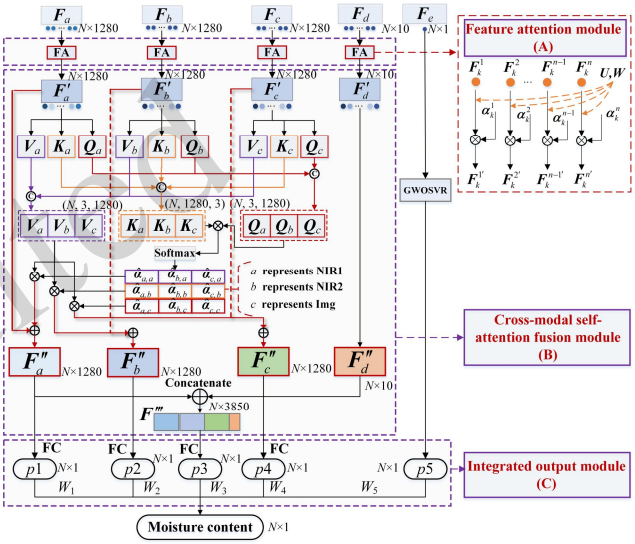


Fig. 2 Flowchart of the multilevel cross-modal integration fusion model

Module C was proposed to integrate the output results of features from different levels, aiming to achieve more robust MC detection results. The DC features are modeled using the GWOSVR described in Section 4.1.3, and the output results are utilized for decision-level fusion in module C. Finally, the unimodal fusion feature prediction results ($p1, p2, p4$), cross-modal fusion feature prediction results ($p3$), and DC prediction results ($p5$) were weighted summation using trainable weights to obtain the final MC prediction results, as shown in Eq. (12).

$$w = W_1 p1 + W_2 p2 + W_3 p3 + W_4 p4 + W_5 p5 \tag{12}$$

4.2 Model setting and evaluation

As described above, the flowchart of our proposed improved cross-modal integration fusion method for the rapid detection of MC in concrete sand is displayed in Fig. 1. The input features of the model

were standardized using Z score normalization. The loss function of the model is the RMSE between the predicted and measured values. In the training stage, the Adam (Kingma and Ba, 2014) optimizer was used to adaptively tune the model parameters. The batch size of the model was set to 20. The initial learning rate was set to 0.0005 and was decreased by a factor of 0.9 every 30 epochs. The maximum number of iterations was set to 500. To avoid overfitting, the dropout layer was set before the fully connected layer, and the dropping rate was set to 0.1. Training was stopped when the absolute error of the loss value between 10 consecutive epochs in the verification dataset was less than 0.00001 or when the maximum epoch was reached. Finally, the model corresponding to the minimum loss value of the verification dataset was selected as the calibration model for the final MC detection. The model was implemented in Python using PyTorch with a single NVIDIA GeForce GTX 1060 GPU. The model performance was evaluated using the RMSE, R^2, and residual predictive deviation (RPD). The detailed principles are elaborately described in (Yuan et al., 2022). Additionally, the improvement rate (IR) of the test set RMSE was employed to evaluate the comparative performance of the models, defined as follows.

$$R_{\mathrm{I}} = \frac{\left|E_B - E_A\right|}{E_B} \tag{13}$$

where R_{I} is the IR, E_A is the RMSE of the superior model A on the test set, and E_B is the RMSE of the inferior model B on the test set.

5 Results and analysis

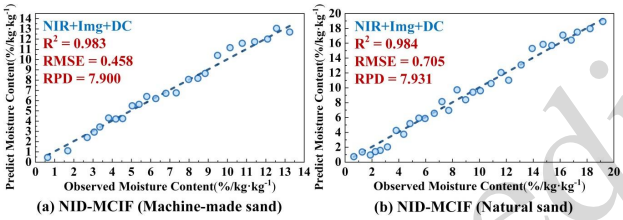
Using the data acquisition method described in Section 3, we obtained 147 multimodal samples of machine-made sand and 183 multimodal samples of natural sand. To ensure that the trained model could achieve a better generalization performance, we sorted all data according to the MC, from large to small. Then, a group of six samples was divided into training, validation, and test sets at a ratio of 4:1:1. The relevant statistical properties are presented in Table S1 of the ESM. For convenience, the following definitions are provided: N refers to NIR modal information, I represents image modal information, and D denotes DC modal information. The naming conventions and their semantic interpretations for the corresponding models are systematically detailed in Table 2.

Table 2 Model nomenclature and semantic conventions

Model	Description
NID-MCIF	The MCIF-based trimodal fusion prediction model, with all subsequent multimodal combinations adhering to this nomenclature convention.
NID-MCIF_ without_A	The MCIF-based trimodal fusion prediction model with module A removed.
NID-MCIF_ without_B	The MCIF-based trimodal fusion prediction model with module B removed.
NID-MCIF_ without_C	The MCIF-based trimodal fusion prediction model with module C removed.
NID-MCIF_ without_AB	The MCIF-based trimodal fusion prediction model with modules A and B removed.
NID-MCIF_ without_AC	The MCIF-based trimodal fusion prediction model with modules A and C removed.
NID-MCIF_ without_BC	The MCIF-based trimodal fusion prediction model with modules B and C removed.
NID-CC	The simple feature-concatenation-based trimodal fusion prediction model, with all subsequent multimodal combinations adhering to this nomenclature convention.
NID-DMF	The trimodal fusion prediction model based on the deep multimodal fusion model (DMF) proposed in (Yuan et al., 2024), with all subsequent multimodal combinations adhering to this nomenclature convention.
NID-SAF	The Self-Attention based Fusion (SAF) trimodal fusion prediction model, with all subsequent multimodal combinations adhering to this nomenclature convention.

Fig. 3 depicts the performance of the proposed model on both machine-made and natural sand multimodal test sets. It can be observed that the proposed model achieves high accuracy for the MC detection of both sands. This result highlights the model's excellent performance in predicting the MC of sand and its potential applicability for different sand materials. In terms of detection time, the multimodal data acquisition process requires approximately 70 s, and the model inference takes approximately 0.62 s, resulting in an overall MC detection time of approximately 71 s. Compared to the traditional oven-drying method (8 h) and the frying method (30 min), this approach yields a significant improvement in detection efficiency. The testing frequency can be increased by a factor of

approximately 24, thereby substantially enhancing the ability to capture fluctuations in fine aggregate MC. According to engineering records, the batching plant evaluated in this study operates with a standard concrete production capacity of 131 m³/h, corresponding to an hourly output of approximately 50 batches. Given that the single-batch detection time is approximately 72 s, the proposed method clearly demonstrates the capability for real-time detection. Furthermore, this study analyzed the generalization performance of the proposed model and the weights of multimodal fusion and conducted ablation experiments on the proposed model to verify the superiority of the NID-MCIF architectural design. For details, please refer to Sections S5.1, S5.2 and S5.3 of the ESM.



(a) NID-MCIF (Machine-made sand) (b) NID-MCIF (Natural sand)

Fig. 3 Prediction results of the proposed model for the MC of concrete sand

6 Discussion

6.1 Comparison with the unimodal model

To verify the superiority of the proposed approach over the unimodal method, models were built for the NIR, DC, and image modal data, including ICNN, AT-BiGRU, GWOSVR, ANN and EfficientNet. Furthermore, to investigate their effectiveness in fusing different types of features within the NIR modality and to compare them with a simple feature concatenation model (CNN-GRU), this study applied the proposed MCIF model, the DMF model from (Yuan et al., 2024), and the traditional self-attention model (SAF) to fuse spectral abstract and positional features. This resulted in the construction of the following NIR fusion models: the MCIF-based NIR fusion model (CNN-GRU-MCIF), the DMF-based NIR fusion model (CNN-GRU-DMF), and the SAF-based NIR fusion model (CNN-GRU-SAF). The model parameters described above are consistent with those in Section 4.1. For NIR data modeling, ANN and GWOSVR were also utilized for comparative analysis. The ANN consists of an input layer (160 neurons), two hidden layers (256 and 128 neurons), and an output layer (1 neuron). The kernel and penalty parameters of the SVR are optimized iteratively by the GWO algorithm between 0.01 and 10, with the population size defined as 30 and the maximum number of iterations set as 50. The model training strategy is consistent with that described in Section 4.3. Fig. 4 illustrates the IRs achieved by the proposed model compared to the comparison models in MC detection for two types of sand materials, and Fig. 5 displays the prediction results of each unimodal model for the MC test set of machine-made sand. The results demonstrate that GWOSVR (R^2=0.952, RMSE=0.774, RPD=4.679) employed for modeling DC features achieved the highest accuracy in MC detection. This can be attributed to the complete insertion of the DC detection device's probe into the sand material, which minimizes interference from external environmental fluctuations and consequently enhances the model's detection precision. For NIR data, CNN-GRU-MCIF, CNN-GRU-DMF, and CNN-GRU-SAF all exhibit significantly superior performance relative to CNN-GRU, demonstrating that modeling interactions between NIR abstract and positional features enhances predictive capability. Notably, the CNN-GRU model exhibits inferior performance compared to the ICNN and AT-BiGRU models, which contrasts with the findings of (Yuan et al., 2022). This can be attributed to the simpler two-branch architecture employed by (Yuan et al., 2022), which features lower feature dimensions and relatively lower heterogeneity between the two feature types. Consequently, their dual-branch model, based on a straightforward concatenation fusion strategy, can still achieve improved prediction accuracy. In contrast, the two-branch model in this study extracts a larger volume and higher capacity of features, leading to increased feature heterogeneity. As a result, the dual-branch model relying on a simple concatenation fusion strategy yields suboptimal performance. Simultaneously, compared to the ICNN and AT-BiGRU models, the IR of the CNN-GRU-MCIF model for MC detection in machine-made and natural sand are {14.44%, 4.84%} and {11.40%, 13.68%}, respectively, further validating the efficacy of utilizing both the dual-branch architecture and the MCIF fusion strategy to enhance overall model

performance.

Moreover, these three cross-modal fusion models exhibit minimal performance variation among themselves, attributable to the constrained representational capacity of single-modal features limiting their information mining potential. Furthermore, IEfficientNet (R^2=0.943, RMSE=0.843, RPD=4.293) exhibited relatively strong performance on image data. These results indicate that different models exhibit significant variations across different data types. Compared to the optimal unimodal MC detection model (GWOSVR (DC)), the proposed NID-MCIF significantly enhances MC detection accuracy by effectively integrating key information from multimodal features. Table S4 of the ESM presents the performance of the proposed cross-modal fusion model alongside various unimodal models in MC detection of natural sand. It is evident that the proposed model maintains significant superiority across different types of sand materials, thereby validating its applicability.

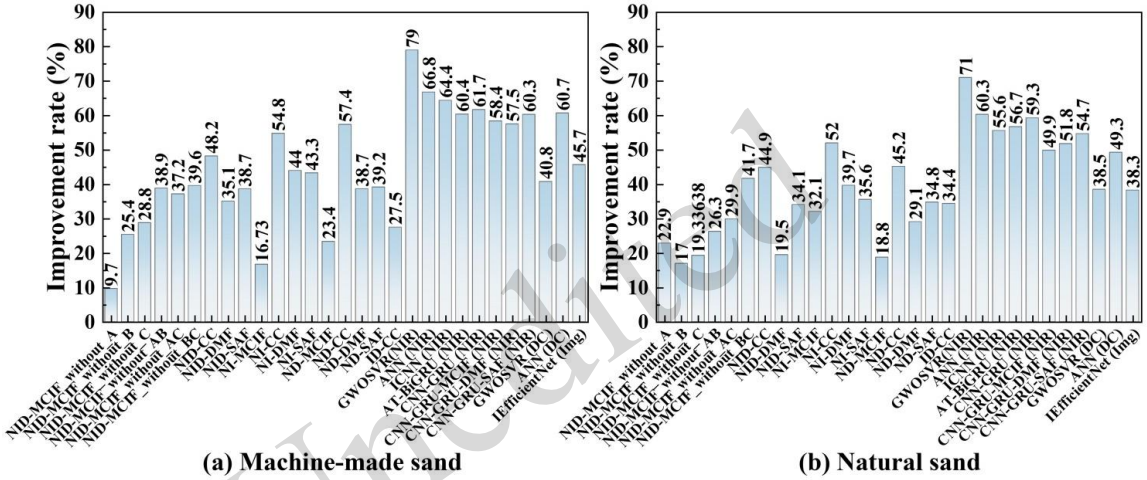


Fig. 4 IRs of the proposed model compared with those of other models

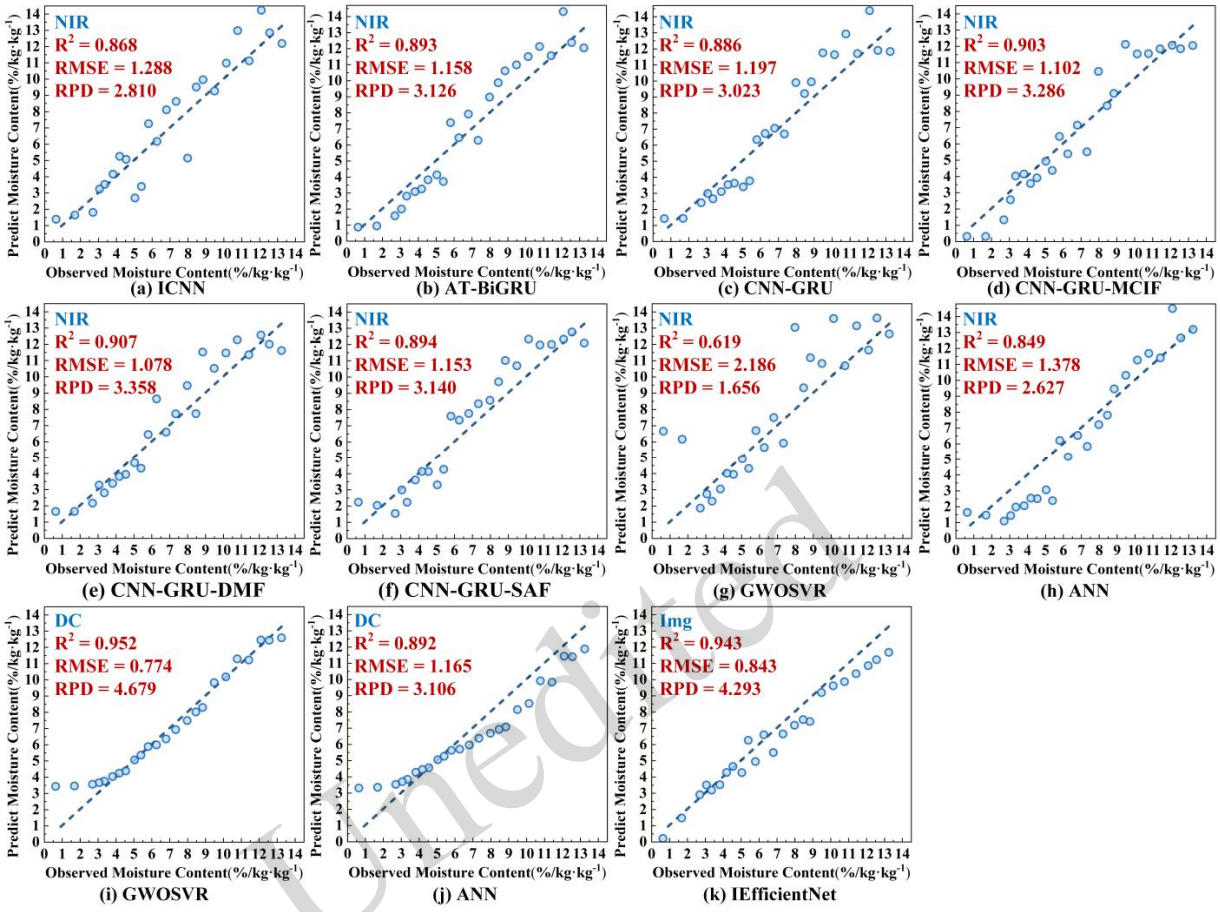


Fig. 5 Scatter diagram of the MC prediction results of machine-made sand using unimodal models

6.2 Comparison with the multimodal model

The ablation experiments presented in Section S5.3 of the ESM demonstrate the effectiveness of the proposed MCIF network in cross-modal feature fusion across three modalities. To further explore the performance of different modality combinations and fusion strategies in MC detection for sand materials, this study designed multiple comparative experiments encompassing MCIF, DMF, and SAF models alongside feature concatenation baselines, all implemented within both tri-modal and bimodal frameworks. The results are shown in Fig. 6 and Table S5 of the ESM. As shown in Fig. 4, the proposed NID-MCIF model demonstrates superior performance in MC detection for two types of sand. Compared to other multimodal fusion models, the proposed model achieves a minimum IR of 16.73% for machine-made sand and 18.8% for natural sand and a maximum IR of 57.4% for machine-made sand and 52% for natural sand, highlighting its remarkable advantage. The NI-MCIF model (machine-made sand: R^2=0.976, RMSE =0.550, RPD=6.590; natural sand: R^2=0.964, RMSE=1.038, RPD=5.388) and the ND-MCIF model (machine-made sand: R^2=0.971, RMSE=0.599, RPD=6.043; natural sand: R^2=0.975, RMSE=0.868, RPD=6.443) demonstrate significant improvements in RMSE compared to the NI-CC model (machine-made sand: R^2=0.918, RMSE=1.013, RPD=3.572; natural sand: R^2=0.928, RMSE=1.469, RPD=3.807) and the ND-CC model (machine-made sand: R^2=0.908, RMSE=1.075, RPD=3.366; natural sand: R^2=0.945, RMSE=1.286, RPD=4.351). Specifically, the IRs in RMSE are 45.71% (machine-made sand) and 29.34% (natural sand) for NI-MCIF and 44.28% (machine-made sand) and 32.50% (natural sand) for ND-MCIF. Furthermore, within the same modal framework, the MCIF model outperformed the DMF model and the SAF model on both the machine-made and natural sand test sets. The results above demonstrate that the proposed MCIF remains effective in extracting

intermodal interactions and multilevel feature contributions within bimodal data and exhibits significant advantages over the state-of-the-art multimodal fusion model for MC detection and simple concatenation-based fusion models. The comparative analysis of model FLOPs indicates that, compared to the NID-DMF, NID-SAF, and NID-CC models, the proposed NID-MCIF model introduces a marginal increase in FLOPs (0.65%) while achieving a substantial enhancement in performance (19.5%–48.2%). This minimal increase in computational overhead is highly disproportionate to the significant performance gains, indicating that the improvements primarily stem from the refined fusion strategy. The comparative analysis of model parameters indicates that the NID-MCIF model contains 4.83, 1.38, and 4.35 times the parameters of the NID-DMF, NID-SAF, and NID-CC models, respectively, while achieving performance IRs of 35.1%, 38.7%, and 48.2% (using the machine-made sand test set as an example). Furthermore, the parameter volumes of the NID-DMF and NID-SAF models are 0.9 and 3.15 times that of the NID-CC model, yielding IRs of 20.14% and 15.50%, respectively. The relative performance of the NID-SAF and NID-DMF models compared to the NID-CC model demonstrates that a mere increase in parameter volume does not guarantee a significant enhancement in model performance. Notably, although the NID-MCIF and NID-SAF models possess comparable parameter volumes, the NID-MCIF model exhibits a substantial performance improvement over both the NID-SAF and NID-CC models. This effectively validates the superiority and effectiveness of the proposed MCIF fusion strategy. Furthermore, this study further investigates different modality combinations under the same fusion strategy. For details, please refer to Section S6.2 of the ESM.

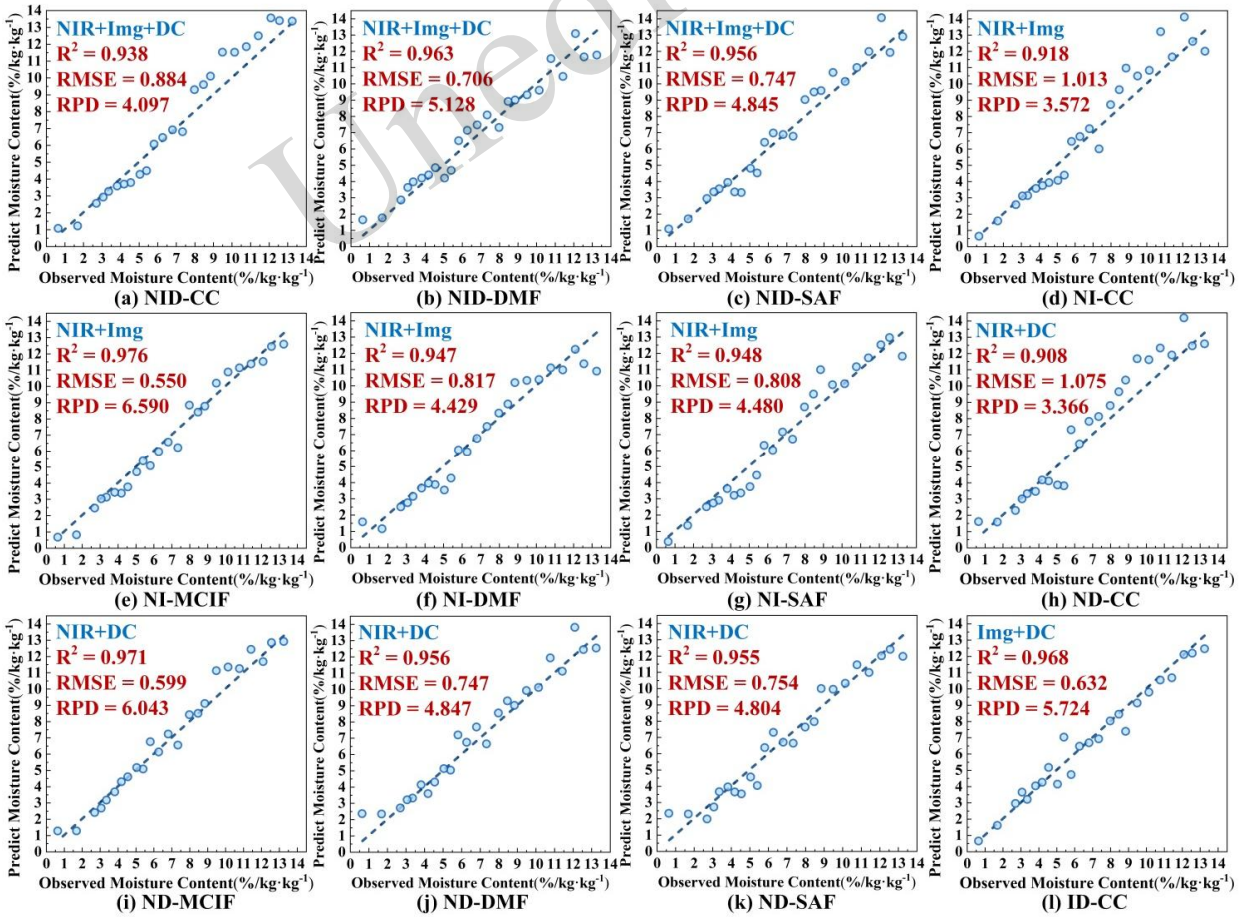


Fig. 6 Scatter diagram of the MC prediction results of machine-made sand based on multimodal models

6.3 Robustness analysis

As previously discussed, during the operation of sensing equipment, the multimodal data acquisition process is inevitably affected by factors such as equipment aging, operational runtime, and calibration errors. Consequently, various types of noise—including photon noise, thermal noise, and electronic and quantization noise—are introduced. During NIR spectral measurements, the random arrival of photons at the detector follows a Poisson distribution, which can be approximated as a Gaussian distribution when the photon count is sufficiently large. Furthermore, devices such as the NIR spectrometer, mobile phones, and DC detection equipment are prone to heating during measurements, thereby generating thermal noise. This noise originates primarily from the thermal motion of electrons within the detectors and circuitry; theoretically, it acts as white noise with an amplitude conforming to a Gaussian distribution. Therefore, according to the central limit theorem, the superposition of these multiple independent noise sources results in a total noise profile that tends to follow a Gaussian distribution. Consequently, it is necessary to further investigate the model's robustness under varying signal-to-noise ratios. Building upon the foundational work of Zhong et al. (2019) and Yuan et al. (2024), this study introduced Gaussian noise at eight distinct relative noise levels—specifically 1%, 3%, 5%, 7%, 9%, 13%, and 15%—into each modality of the test set. For each noise level, six repeated experiments were conducted to examine the model's stability when exposed to identical noise conditions. For the analysis of the impact of unimodal noise on model performance, please refer to Section S6.3 of the ESM.

Fig. 7 illustrates the variation in R^2 values for the two sand datasets under simultaneous image noise, NIR noise, and DC noise at different levels. In the machine-made sand dataset, when the relative noise level is less than or equal to 5%, the proposed NID-MCIF model exhibits a relatively slow decline in R^2, maintaining values above 97% and demonstrating good accuracy. However, as the relative noise level increases further, the model's performance declines more rapidly, indicating reduced robustness. In contrast, the NI-MCIF model shows better robustness. The ND-MCIF model also demonstrates strong robustness when the relative noise level is less than or equal to 5%, but its performance decreases as the noise level increases. In the natural sand dataset, the proposed model performs better than in the machine-made sand dataset, with a slower decline in R^2 and consistently superior performance compared to other models. The ND-MCIF model similarly demonstrates better robustness than other models. The NI-MCIF model exhibits no significant decline in R^2 when the relative noise level is less than or equal to 5%, showing good robustness. However, as the noise level increases further, the model's performance declines significantly. Additionally, under various modal combinations and noise levels, the MCIF model significantly outperforms both the DMF and SAF models. Models based on simple concatenation fusion strategies, such as NID-CC, NI-CC, and ND-CC, show clear downward trends with large fluctuations across repeated experiments, indicating poor robustness. The ID-CC model exhibits relatively better robustness due to the stability of image features across noise levels and the lack of interaction with DC features. These results indicate that the proposed model, through the MCIF network, effectively extracts multilevel cross-modal fusion features, demonstrating strong robustness at low to medium relative noise levels and achieving superior detection performance compared to other models. MCIF-based dual-modality models also show good robustness at low to medium relative noise levels and perform significantly better overall than their counterparts based on simple concatenation fusion. The ID-CC model benefits from the relatively stable detection accuracy of image features across noise levels and its lack of interaction with DC features, leading to relatively robust detection results. However, this also limits its ability to improve detection accuracy at low to medium noise levels. In practical engineering applications, multimodal data sampling is typically conducted in indoor environments where environmental factors remain relatively stable. Consequently, the NID-MCIF model, which demonstrates high accuracy and strong robustness under low to moderate noise conditions, is well suited to meet the practical requirements of engineering fields.

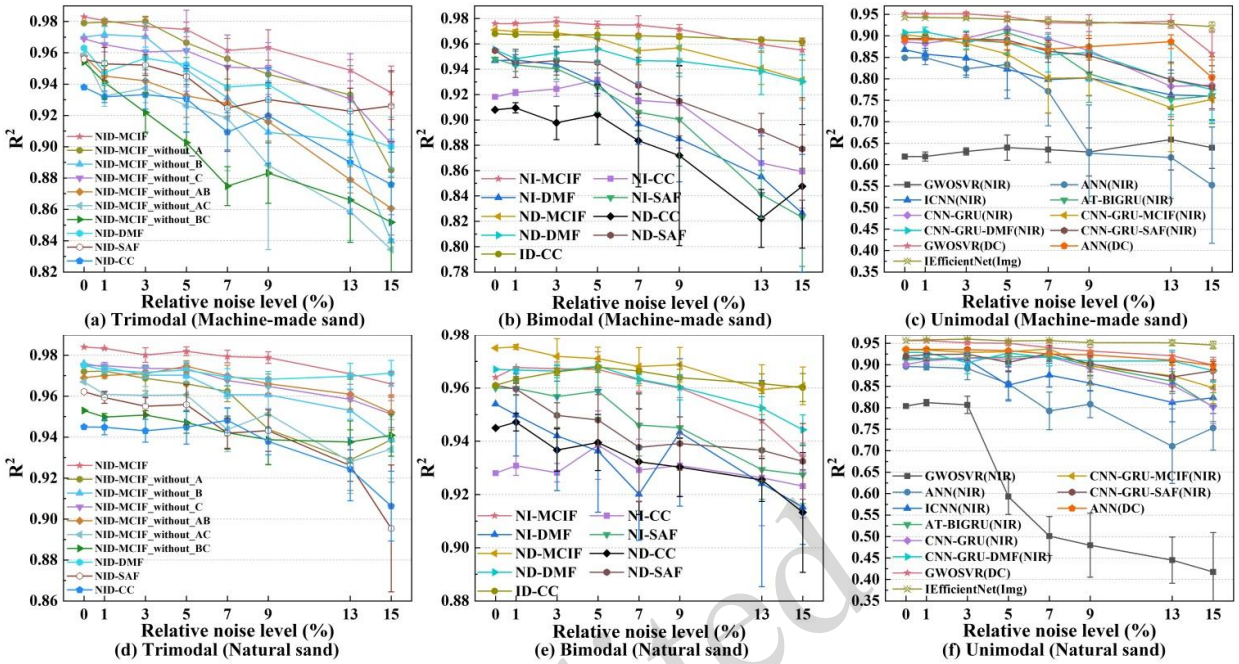


Fig. 7 Model performance under varying levels of image, NIR, and DC noise intensity

7 Conclusions

This study proposes a novel and robust cross-modal integration fusion model that considers multilevel features for the rapid detection of MC in concrete sand materials, utilizing multimodal measurement data such as images, NIR, and DC. The detection accuracy and robustness of the proposed model were validated based on two types of sand materials from engineering sites. The conclusions are as follows:

(1) The proposed NID-MCIF model (machine-made sand: R^2=0.983, RMSE=0.458, RPD=7.900; natural sand: R^2=0.984, RMSE=0.705, RPD=7.931) achieves the highest detection accuracy and significantly outperforms other comparative models. Furthermore, the proposed MCIF fusion strategy also demonstrates competitive detection accuracy when applied to bimodal data. Models based on the MCIF fusion strategy significantly outperform models employing simple concatenation fusion strategies under corresponding modalities, as well as unimodal models. In addition, ablation experiments confirm that the proposed MCIF fusion strategy effectively extracts key features and enhances model performance by considering interactions between multilevel features.

(2) By introducing Gaussian noise at different levels into the test datasets of the two types of sand materials, the noise resistance of the model under noisy environments was evaluated. The results indicate that the proposed NID-MCIF exhibits strong overall robustness across various noisy scenarios, including scenarios with noise added only to image data, only to spectral data, only to DC data, and simultaneously to image, NIR, and DC data. This performance is significantly better than that of models based on simple concatenation fusion strategies. Similarly, the bimodal models NI-MCIF and ND-MCIF, which are based on the MCIF fusion strategy, also demonstrate competitive performance in terms of detection accuracy and robustness, significantly outperforming their counterparts that use simple concatenation fusion strategies under corresponding modalities.

Case studies demonstrate that the proposed model achieves a detection time of less than 71 seconds, offering a novel and robust approach for the rapid determination of MC in concrete sand. Consequently, it possesses the capability for real-time fine aggregate MC detection in small- to medium-capacity batching plants. However, this study still presents certain limitations, such as the insufficient automation level in the data acquisition process. Additionally, the current research did not account for the impact of varying ambient light intensities, relative humidity, different water sources,

and fine aggregate types with distinct mineral compositions on the MC detection results. For instance, fine aggregates with varying mineral compositions exhibit distinct absorption characteristics across different wavelengths of the NIR spectrum; variations in particle morphology lead to differences in the diffuse reflectance and scattering of NIR light; and differing salinity levels significantly impact DC measurements. Furthermore, to meet the requirements of higher-intensity concrete production, the detection efficiency of the proposed theory needs to be further enhanced. In future studies, the diversity and volume of the dataset will be expanded. We also intend to enhance the data collection of potential influencing factors—such as ambient light, environmental humidity, mineral composition, and particle morphology—and deeply analyze their underlying mechanisms affecting MC. Furthermore, the proposed method will be integrated with intelligent concrete mixing systems to construct a more universal and competitive framework for fine aggregate MC detection.

Acknowledgments

This research was supported by the National Natural Science Foundation of China Joint Fund Key Project (Grand no. U24B20111) and the Basic Theory and Damming Technology for low-hot Cement or Concrete of the National Natural Science Foundation of China (Grand no. 51839007). In addition, we are grateful for the valuable suggestions of the editor and reviewers.

Author contributions

Zhijian CAI: Conceptualization, Methodology, Investigation, Data curation, Software, Writing – original draft; Jun ZHANG and Xiaoling WANG: Writing – review & editing, Supervision; Xiaoling WANG: Funding acquisition; Jiajun WANG: Writing – review & editing, Methodology; Kehao ZHAO: Formal analysis, Data curation; Guohua WU: Visualization.

Conflict of interest

Zhijian CAI, Jun ZHANG, Xiaoling WANG, Jiajun WANG, Kehao ZHAO, and Guohua WU declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Abasi S, Minaei S, Jamshidi B, et al., 2019. Rapid measurement of apple quality parameters using wavelet de-noising transform with Vis/NIR analysis. *Scientia Horticulturae*, 252, 7–13. https://doi.org/10.1016/j.scienta.2019.02.085

Amlashi AT, Abdollahi SM, Goodarzi S, et al., 2019. Soft computing based formulations for slump, compressive strength, and elastic modulus of bentonite plastic concrete. *Journal of Cleaner Production*, 230, 1197–1216. https://doi.org/10.1016/j.jclepro.2019.05.168

Cao YY, Yang W, Li H, et al., 2024. Development of a vehicle-mounted soil organic matter detection system based on near-infrared spectroscopy and image information fusion. *Measurement Science and Technology*, 35(4). https://doi.org/10.1088/1361-6501/ad179f

Cheng MH, Jiao XY, Liu YD, et al., 2022. Estimation of soil moisture content under high maize canopy coverage from UAV multimodal data and machine learning. *Agricultural Water Management*, 264, 107530. https://doi.org/10.1016/j.agwat.2022.107530

Chung JY, Gulcehre C, Cho K, et al., 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. 1–9. https://doi.org/10.48550/arXiv.1412.3555

Dean TJ, Bell JP, Baty AJB, 1987. Soil moisture measurement by an improved capacitance technique, Part I. Sensor design and performance. *Journal of Hydrology*, 93(1–2), 67–78. https://doi.org/10.1016/0022-1694(87)90194-6

Fan CC, Zheng YX, Wang SQ, et al., 2023. Prediction of bond strength of reinforced concrete structures based on feature selection and GWO-SVR model. *Construction and Building Materials*, 400, 132602. https://doi.org/10.1016/j.conbuildmat.2023.132602

Gan CQ, Tang Y, Fu X, et al., 2024. Video multimodal sentiment analysis using cross-modal feature translation and dynamical propagation. *Knowledge-Based Systems*, 299, 111982. https://doi.org/10.1016/j.knosys.2024.111982

Gavela S, Nikoloutsopoulos N, Papadakos G, et al., 2018. Multifactorial experimental analysis of concrete compressive strength as a function of time and water-to-cement ratio. *Procedia Structural Integrity*, 10, 135–140. https://doi.org/10.1016/j.prostr.2018.09.020

Guo Z, Zhang J, Ma CY, et al., 2023. Application of visible-near-infrared hyperspectral imaging technology coupled with wavelength selection algorithm for rapid determination of moisture content of soybean seeds. *Journal of Food Composition and Analysis*, 116, 105048. https://doi.org/10.1016/j.jfca.2022.105048

Han ZJ, Cao DF, Zhu HH, et al., 2022. A field test to investigate spatiotemporal distribution of soil moisture under different cropland covers in the semiarid Loess

Plateau of China. *Paddy and Water Environment*, 20(3), 339–353.
https://doi.org/10.1007/s10333-022-00896-5

He XY, Wang Y, Zhao S, et al., 2023. Co-Attention Fusion Network for Multimodal Skin Cancer Diagnosis. *Pattern Recognition*, 133, 108990.
https://doi.org/10.1016/j.patcog.2022.108990

Hua CJ, Sun MC, Zhu Y, et al., 2022. Pedestrian detection network with multi-modal cross-guided learning. *Digital Signal Processing*, 122, 103370.
https://doi.org/10.1016/j.dsp.2021.103370

Huang ZX, Sanaeifar A, Tian Y, et al., 2021. Improved generalization of spectral models associated with Vis-NIR spectroscopy for determining the moisture content of different tea leaves. *Journal of Food Engineering*, 293, 110374.
https://doi.org/10.1016/j.jfoodeng.2020.110374

Inthiyaz S, Altahan BR, Ahammad SH, et al., 2023. Skin disease detection using deep learning. *Advances in Engineering Software*, 175, 103361.
https://doi.org/10.1016/j.advengsoft.2022.103361

Jaafar N, Lachiri Z, 2023. Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance. *Expert Systems with Applications*, 211, 118523.
https://doi.org/10.1016/j.eswa.2022.118523

Khalkho D, Thakur S, Tripathi MP, 2024. Soil Moisture Determination by Normalized Difference Index Based on Drone Images Analysis. *Journal of the Indian Society of Remote Sensing*, 52(7), 1623–1632.
https://doi.org/10.1007/s12524-024-01885-3

Kingma DP, Ba JL, 2014. Adam: A method for stochastic optimization. ArXiv Preprint ArXiv:1412.6980, 1–15.
https://doi.org/10.48550/arXiv.1412.6980

Krzeminska D, Bloem E, Starkloff T, et al., 2022. Combining FDR and ERT for monitoring soil moisture and temperature patterns in undulating terrain in south-eastern Norway. *Catena*, 212, 106100.
https://doi.org/10.1016/j.catena.2022.106100

Li F, Luo JS, Wang LL, et al., 2023. GCF2-Net: global-aware cross-modal feature fusion network for speech emotion recognition. *Frontiers in Neuroscience*, 17.
https://doi.org/10.3389/fnins.2023.1183132

Li H, Song YM, Wang ZY, et al., 2024. Development of an online prediction system for soil organic matter and soil moisture content based on multi-modal fusion. *Computers and Electronics in Agriculture*, 227, 109514.
https://doi.org/10.1016/j.compag.2024.109514

Liu ZY, Zhang RT, Yang CS, et al., 2022. Research on moisture content detection method during green tea processing based on machine vision and near-infrared spectroscopy technology. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, 271, 120921.
https://doi.org/10.1016/j.saa.2022.120921

Liu ZQ, Guo D, Lacasse S, et al., 2020. Algorithms for intelligent prediction of landslide displacements. *Journal of Zhejiang University: Science A*, 21(6), 412–429.
https://doi.org/10.1631/jzus.A2000005

Fragkos A, Loukatos D, Kargas G, et al., 2024. Response of the TEROS 12 soil moisture sensor under different soils and variable electrical conductivity. *Sensors*, 24(7): 2206.
https://doi.org/10.3390/s24072206

Miao XX, Miao Y, Liu Y, et al., 2023. Measurement of nitrogen content in rice plant using near infrared spectroscopy combined with different PLS algorithms. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, 284, 121733.
https://doi.org/10.1016/j.saa.2022.121733

Tan MX, Le Q, 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. International conference on machine learning, PMLR, p. 6105-6114.

Mouazen AM, Al-Asadi RA, 2018. Influence of soil moisture content on assessment of bulk density with combined frequency domain reflectometry and visible and near infrared spectroscopy under semi field conditions. *Soil and Tillage Research*, 176, 95–103.
https://doi.org/10.1016/j.still.2017.11.002

Nawar S, Mouazen AM, 2019. On-line vis-NIR spectroscopy prediction of soil organic carbon using machine learning. *Soil and Tillage Research*, 190, 120–127.
https://doi.org/10.1016/j.still.2019.03.006

Ng W, Minasny B, Montazerolghaem M, et al., 2019. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma*, 352, 251–267.
https://doi.org/10.1016/j.geoderma.2019.06.016

Ni C, Wang DY, Tao Y, 2019. Variable weighted convolutional neural network for the nitrogen content quantization of Masson pine seedling leaves with near-infrared spectroscopy. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, 209, 32–39.
https://doi.org/10.1016/j.saa.2018.10.028

Nijaguna GS, Manjunath DR, Abouhawwash M, et al., 2023. Deep Learning-Based Improved WCM Technique for Soil Moisture Retrieval with Satellite Images. *Remote Sensing*, 15(8).
https://doi.org/10.3390/rs15082005

Radica F, Iezzi G, Trotta O, et al., 2024. Characterization of CDW types by NIR spectroscopy: Towards an automatic selection of recycled aggregates. *Journal of Building Engineering*, 88, 109005.
https://doi.org/10.1016/j.jobe.2024.109005

Rahman A, Street J, Wooten J, et al., 2025. MoistNet: Machine vision-based deep learning models for wood chip moisture content measurement. *Expert Systems with Applications*, 259, 125363.
https://doi.org/10.1016/j.eswa.2024.125363

Shaheed K, Mao A, Qureshi I, et al., 2022. DS-CNN: A pre-trained Xception model based on depth-wise separable convolutional neural network for finger vein recognition. *Expert Systems with Applications*, 191,

116288.
https://doi.org/10.1016/j.eswa.2021.116288

Shu T, Wang XK, Wang RT, et al., 2023. Mutilmodal Feature Extraction and Attention-based Fusion for Emotion Estimation in Videos. ArXiv Preprint ArXiv:2303.10421. http://arxiv.org/abs/2303.10421

Somkunwar R, Gupta AK, Anand A, et al., 2024. CNN-based Soil Image Analysis for Enhanced Crop Prediction in Smart Agriculture. *2024 MIT Art, Design and Technology School of Computing International Conference* (MITADTSoCiCon), 1–5.
https://doi.org/10.1109/MITADTSoCiCon60330.2024.10575651

Song Y, Wang XZ, Xie HL, et al., 2021. Quality evaluation of Keemun black tea by fusing data obtained from near-infrared reflectance spectroscopy and computer vision sensors. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, 252, 119522.
https://doi.org/10.1016/j.saa.2021.119522

Song Y, Yi WQ, Liu Y, et al., 2025. A robust deep learning model for predicting green tea moisture content during fixation using near-infrared spectroscopy: Integration of multi-scale feature fusion and attention mechanisms. *Food Research International*, 203(130), 115874.
https://doi.org/10.1016/j.foodres.2025.115874

Szegedy C, Liu W, Jia YQ, et al., 2015. Going Deeper with Convolutions Christian. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p.1–9.
https://doi.org/10.4324/9781410605337-29

Tobiszewski M, Vakh C, 2023. Analytical applications of smartphones for agricultural soil analysis. *Analytical and Bioanalytical Chemistry*, 415(18), 3703–3715.
https://doi.org/10.1007/s00216-023-04558-1

Tong L, Qian K, 2025. HFNet: High-precision robotic grasp detection in unstructured environments using hierarchical RGB-D feature fusion and fine-grained pose alignment. Measurement, 253, 117775.
https://doi.org/10.1016/j.measurement.2025.117775

Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Wang QQ, Zhang H, Li FD, et al., 2021. Assessment of calibration methods for nitrogen estimation in wet and dry soil samples with different wavelength ranges using near-infrared spectroscopy. *Computers and Electronics in Agriculture*, 186, 106181.
https://doi.org/10.1016/j.compag.2021.106181

Wang YT, Li MZ, Ji RH, et al., 2021. Construction of complex features for predicting soil total nitrogen content based on convolution operations. *Soil and Tillage Research*, 213, 105109.
https://doi.org/10.1016/j.still.2021.105109

Wang Y, Li M, Ji R, et al., 2021. A deep learning-based method for screening soil total nitrogen characteristic wavelengths. *Computers and Electronics in Agriculture*, 187, 106228.
https://doi.org/10.1016/j.compag.2021.106228

Watanabe A, Tokuda S, Mizuta Y, et al., 2021. Toward automated non-destructive diagnosis of chloride attack on concrete structures by near infrared spectroscopy. *Construction and Building Materials*, 305, 124796.
https://doi.org/10.1016/j.conbuildmat.2021.124796

Wei YY, Cao L, Dong YL, et al., 2025. MCNN-CMCA: A multiscale convolutional neural networks with cross-modal channel attention for physiological signal-based mental state recognition. *Digital Signal Processing*, 156, 104856.
https://doi.org/10.1016/j.dsp.2024.104856

Wu CW, Zheng Y, Yang H, et al., 2021. Effects of different particle sizes on the spectral prediction of soil organic matter. *Catena*, 196, 104933.
https://doi.org/10.1016/j.catena.2020.104933

Wu ZJ, Cui NB, Zhang WJ, et al., 2024. Estimation of soil moisture in drip-irrigated citrus orchards using multi-modal UAV remote sensing. *Agricultural Water Management*, 302, 108972.
https://doi.org/10.1016/j.agwat.2024.108972

Yang JC, Wang XL, Wang RH, et al., 2020. Combination of Convolutional Neural Networks and Recurrent Neural Networks for predicting soil properties using Vis–NIR spectroscopy. *Geoderma*, 380, 114616.
https://doi.org/10.1016/j.geoderma.2020.114616

You ZY, Wang X, Xu JW, et al., 2025. Signal generation for bolt loosening detection with unbalanced datasets based on the CBAM-VAE. *Measurement*, 240, 115589.
https://doi.org/10.1016/j.measurement.2024.115589

Yu XX, Zhang ZG, Tang BP, et al., 2024. A multi-head self-attention autoencoder network for fault detection of wind turbine gearboxes under random loads. *Measurement Science and Technology*, 35(8).
https://doi.org/10.1088/1361-6501/ad4dd4

Yuan Q, Wang JJ, Wu BP, et al., 2024. Deep multimodal fusion model for moisture content measurement of sand gravel using images, NIR spectra, and dielectric data. *Measurement*, 227, 114270.
https://doi.org/10.1016/j.measurement.2024.114270

Yuan Q, Wang JJ, Zheng MW, et al., 2022. Hybrid 1D-CNN and attention-based Bi-GRU neural networks for predicting moisture content of sand gravel using NIR spectroscopy. *Construction and Building Materials*, 350, 128799.
https://doi.org/10.1016/j.conbuildmat.2022.128799

Zhang XL, Lin T, Xu JF, et al., 2019. DeepSpectra: An end-to-end deep learning approach for quantitative spectral analysis. *Analytica Chimica Acta*, 1058, 48–57.
https://doi.org/10.1016/j.aca.2019.01.002

Zhang XL, Yang J, Lin T, et al., 2021. Food and agro-product quality evaluation based on spectroscopy and deep learning: A review. *Trends in Food Science and Technology*, 112, 431–441.
https://doi.org/10.1016/j.tifs.2021.04.008

Zhong H, Wang JJ, Jia HJ, et al., 2019. Vector field-based support vector regression for building energy consumption prediction. *Applied Energy*, 242, 403–414. https://doi.org/10.1016/j.apenergy.2019.03.078

Zhu CM, Ding JL, Zhang ZP, et al., 2022. Exploring the potential of UAV hyperspectral image for estimating soil salinity: Effects of optimal band combination algorithm and random forest. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, 279, 121416. https://doi.org/10.1016/j.saa.2022.121416

Zuo ZY, Mu JD, Li WJ, et al., 2023. Study on the detection of water status of tomato (Solanum lycopersicum L.) by multimodal deep learning. *Frontiers in Plant Science*, 14, 1–10. https://doi.org/10.3389/fpls.2023.1094142

Electronic Supplementary Materials

Section S3–S6, Figs. S1–S9, Tables S1–S5

中文概要

题　目： 一种用于混凝土砂含水率快速检测的新型鲁棒跨模态集成融合模型

作　者： 蔡志坚，张君，王晓玲，王佳俊，赵科皓，吴国华

机　构： 天津大学，水利工程智能建设与运维全国重点实验室，中国天津，300350

目　的： 快速且准确地检测混凝土砂的含水量（MC）对于确保混凝土质量至关重要。本文旨在提出一种快速、准确且鲁棒的跨模态含水率检测模型，以提高混凝土砂料含水率检测的准确性和稳定性。

创新点： 1. 提出基于 ICNN、AT-BiGRU、IEfficientNet、ANN 和 GWOSVR 的多分支单模态特征提取方法，以获取多模态深度特征；2. 提出一种多层级跨模态集成融合模型，以全面考虑模态内和模态间的特征交互；3. 通过实验验证所提模型的优越性，并进行可解释性与鲁棒性分析，体现所提方法的应用潜力。

方　法： 1. 建立基于 ICNN、AT-BiGRU、IEfficientNet、ANN 和 GWOSVR 的多分支单模态特征提取方法；2. 建立基于 MCIF 的跨模态特征提取方法，实现跨模态特征深度融合；3. 通过计算机模拟，深入剖析所提模型改进策略的有效性，基于对比实验，验证所提方法的有效性和可靠性。

结　论： 1. 所提 NID-MCIF 模型在两种混凝土砂料含水率检测中均取得了高精度的准确率，R^2>0.98；2. 基于 MCIF 融合策略的模型在相应的模态下，准确性和鲁棒性显著优于采用简单拼接融合策略的模型以及单模态模型，为快速测定混凝土砂料含水率提供了一种新颖且可靠的解决方案。

关键词： 混凝土砂料；快速含水率检测；跨模态集成融合；鲁棒性预测