Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 34(1): 232–250.
DOI: 10.3934/era.2026012
Received: 02 November 2025
Revised: 23 December 2025
Accepted: 25 December 2025
Published: 08 January 2026

Research article

Cross-modal deep fusion based on small samples for early rumor detection

Junqing Yang[1], **Hongzhe Chen**[2], **Yao Zhao**[2], **Wing-Kuen Ling**[3] **and Yang Zhou**[4,*]

[1] School of Photoelectric Engineering, Jiangxi Modern Polytechnic College, Jiangxi 330046, China
[2] School of Automation, Guangdong University of Technology, Guangzhou 510006, China
[3] Center for Integrated Circuits and Artificial Intelligence, Tsientang Institute for Advanced Study, Zhejiang 310024, China
[4] School of Statistics, Beijing Normal University, Beijing 100875, China

* **Correspondence:** Email: bnu_zhouy@bnu.edu.cn.

Abstract: Rumors circulating on social media can have significant adverse impacts on society, highlighting the urgent need for effective rumor detection. However, most existing methods predominantly focus on rumor identification only after widespread dissemination has occurred, when substantial harm has been inflicted. Early-stage rumor detection is a major challenge due to limited reach and small sample sizes, which restrict the use of large datasets and traditional propagation models. To overcome these challenges, a cross-modal deep fusion method based on small samples for early rumor detection is proposed, which includes a multimodal feature extraction network and a cross-modal deep fusion network. The multimodal feature extraction network captures features from multiple modalities, while the multimodal deep information extraction network derives deep representations from these modalities. The cross-modal deep fusion network integrates textual and visual features for rumor classification. Furthermore, an enhanced meta-learning training approach based on model-agnostic meta-learning is proposed to improve the efficiency of rumor detection by employing distinct learning rates both within and between tasks. Experimental results on two publicly available datasets demonstrate that the proposed cross-modal deep fusion method outperforms baseline methods and exhibits promising performance.

Keywords: cross-modal deep fusion network; meta-learning; small samples; early rumor detection

1. Introduction

The rapid advancement of personal computers and mobile phones has greatly facilitated the sharing of news and access to information through online platforms. However, verifying the accuracy of news and information prior to dissemination remains a significant challenge, often leading to the widespread circulation of unverified content, including rumors. Unlike factual news, rumors often possess a sense

of novelty, making them more likely to attract attention and provoke emotional responses, thereby accelerating their spread. Therefore, accurately identifying rumors within the vast volume of shared information is essential.

Several approaches have been proposed to identify rumors, which can be broadly categorized into three main types. The first type of methods centers on analyzing the features of rumors to extract hidden information. Initially, support vector machines (SVM) were used to detect rumors by extracting individual features. The emergence of neural networks has facilitated the fusion and extraction of multiple features, enhancing detection accuracy and driving their widespread adoption in rumor detection. The second type of method focuses on the structure of rumor propagation. Rumors typically spread in a repetitive manner, forming a "cascading" pattern, which makes detection through propagation patterns an effective strategy. Recent studies on propagation dynamics within complex networks—encompassing reaction–diffusion modeling, behavioral spreading mechanisms, and network-driven spatiotemporal pattern evolution—have substantially advanced this line of research and provided more rigorous theoretical insights into the fundamental mechanisms underlying information diffusion. While previous research has made significant progress in rumors detection using machine learning and deep-learning methods, the issue of the imbalanced distribution of rumors samples has not been adequately addressed. To overcome this issue, the third method is to use large models to detect rumors.

In the early stages of rumor propagation, the scope of spread is limited and the number of samples available for extraction is small. Traditional rumor detection methods struggle to accurately identify rumors at this stage. Thus, numerous methods for the early detection of rumors have been proposed in recent years. However, early rumor detection still faces several limitations: 1) Some methods rely on large-scale datasets, overlooking the limited number and narrow dissemination scope of early rumor samples; and 2) Some methods only processes textual information, neglecting the rich multimodal features available on social media platforms.

Therefore, considering that existing methods overlook the deep fusion of multimodal feature information and the limited number of early-stage rumors, this paper proposes a deep cross-modal approach for early rumor detection based on few-shot learning. The principal contributions of this work are delineated as follows.

- A three-layer neural network with cross-modal deep fusion (CMDF) is proposed for feature extraction of text and images, as well as for the interactive fusion of these features. The neural network effectively integrates the features of text and images to enable the early detection of rumors.
- An enhanced meta-learning training method based on model-agnostic meta-learning (MAML) is proposed, which improves the efficiency of rumor detection by adopting different learning rates within and between tasks to update parameters.
- Given the limited number of early rumor samples, a cross-modal deep fusion method for early rumor detection based on small sample sizes is proposed. The performance of the model on two public datasets [1,2] demonstrates its effectiveness and superior performance compared to other baseline methods.

2. Related work

2.1. Rumor detection

The literature on rumor detection has seen the development of diverse methods that can be broadly categorized into three main strands. The first approach primarily examines the distinctive features of rumor-related content for rumor detection. A novel multimodal feature-enhanced attention network (MFAN) is proposed for rumor detection [3], representing the first unified framework that simultaneously integrates textual, visual, and social graph features. An early rumor detection (ERD) model is proposed [4], which combines a dual-attention deep network for tweet-reply feature extraction and a long short-term memory (LSTM)–enhanced deep recurrent Q-learning network (DRQN) for optimized detection timing and accuracy. The work in [5] presents an adaptive rumor detection method that fuses variational autoencoder-extracted word frequency features with test features generated by a convolutional neural network bidirectional long short-term memory network (CNN-BiLSTM), using a credibility-aware mechanism to prevent overfitting.

The second approach employs propagation structure analysis for rumor detection. The work in [6] proposes the path attention graph convolution network (PAGCN), an innovative framework that integrates propagation structure and semantic representation learning to effectively solve the limitations of traditional GCNs, achieving superior performance in rumor detection. To solve the limitations of existing methods, [7] proposes a model named continuous-time dynamic graph networks integrated with knowledge propagation (CGNKP), that captures both spatial structure and continuous-time features of propagation through dynamic graphs. In recent years, research on propagation dynamics based on complex networks and reaction–diffusion systems has further advanced this line of study. The study in [8] develops a reaction-diffusion model using Laplacian matrices, based on secondary transmission mechanisms and network discrete characteristics. The paper in [9] proposes a three-layer network model integrating the microscopic Markov Chain approach and reaction-diffusion systems to analyze green behavior propagation, and applied the model to predict and fit clean energy generation data, while the work in [10] investigates reaction-diffusion systems with advection on discrete networks. These works explore behavioral diffusion mechanisms, spatiotemporal pattern formation, parameter estimation, and diffusion control, thereby providing a more rigorous theoretical foundation for modeling propagation structures. Furthermore, recent approaches [11] have integrated propagation structure analysis with feature-based methods to enhance rumor detection performance.

In recent years, the method of using large language models for rumor detection has gradually become popular. An innovative solution to rumor detection [12] is proposed through a large language model (rumorLLM), fine-tuned for rumor characteristics, with its key contributions including model development and a data augmentation method addressing class imbalance in real-world datasets.

In the early stages of rumor propagation, the scope of spread is limited and the number of samples available for extraction is small. Traditional rumor detection methods struggle to accurately identify rumors at this stage. Thus, numerous methods for the early detection of rumors have been proposed in recent years. A recurrent neural network (RNN)–based deep attention model [13] is proposed which selectively learns temporal hidden representations from consecutive posts to improve early rumor identification. A tweet-vocabulary-user heterogeneous graph is constructed from rumor content and propagation patterns [14], enabling development of a meta-path-based graph attention network that captures global semantic-structural relationships for early rumor detection. A few-shot learning-based

multimodal fusion model [15] is proposed for early rumor detection, incorporating text embedding, feature extraction, and multimodal feature integration. A novel temporal-aware multiview graph attention framework (MTAS) framework [16] is proposed for effective early rumor detection, leveraging generated multiview social data that incorporates semantic, internal, global, and temporal features. However, some methods rely on large-scale datasets or only processes textual information, overlooking the limited number and narrow dissemination scope of early rumor samples and neglecting the rich multimodal features available on social media platforms.

2.2. Few-shot learning

Few-shot learning (FSL) addresses the fundamental challenge of training effective models with limited labeled samples. Recent years have witnessed significant advancements in FSL methodologies. The paper in [17] addresses few-shot camouflaged object detection and segmentation by introducing a new framework which employs instance- and class-level loss functions to effectively distinguish camouflaged instances from backgrounds. By integrating self-knowledge distillation, [18] enhances prototypical networks for FSL and demonstrates competitive performance on standard image classification benchmarks. This investigation [19] proposes an FSL framework augmented with domain-specific attributes to detect unstructured openings using limited labeled data. To address the challenge of limited training samples in rolling bearing fault diagnosis, the study in [20] proposes a novel approach combining Kolmogorov-Arnold networks (KANs) with convolutional neural networks (CNNs).

2.3. Meta-learning

Meta-learning [21], often characterized as "learning to learn", is a machine learning paradigm that focuses on acquiring adaptable learning strategies or optimal parameter initialization through multi-task experience. This approach enables rapid adaptation to novel tasks, effectively addressing FSL challenges. In recent years, many meta-learning algorithms have been proposed. The authors of [22] propose a generic meta-learning framework Meta-SE, a meta-learning framework that employs U-Net as the meta-learner to address few-shot speech enhancement (SE) problem. A reinforcement metric meta-learning approach [23] is proposed which leverages transfer learning to construct an effective feature extraction network. A dynamic meta-learning framework MAML [24] is proposed, featuring a multistep query ensemble loss in the inner loop and adaptive learning rate adjustment through gradient warm-up combined with cosine annealing in the outer loop. The work in [25] analyzes factors contributing to postpartum depression (PPD) and proposes a novel meta-learner model, which integrates multiple base learners to achieve superior predictive accuracy.

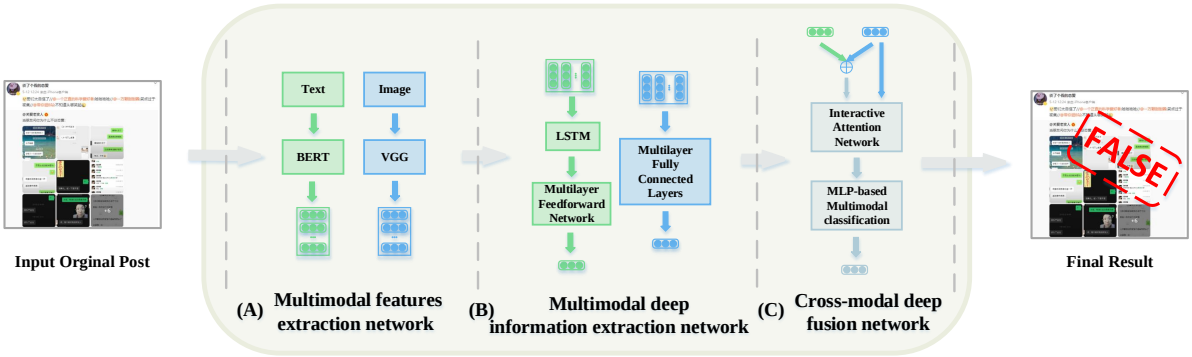

Figure 1. Cross-modal deep fusion neural network based on small samples.

3. The proposed method

This paper proposes a small sample-based cross-modal deep fusion model for early rumor detection. The overall framework of the model is shown in Figure 1, which includes a multimodal feature extraction network, a multimodal deep information extraction network, a cross-modal deep fusion network, and a fusion training network based on MAML.

3.1. Multimodal feature extraction network

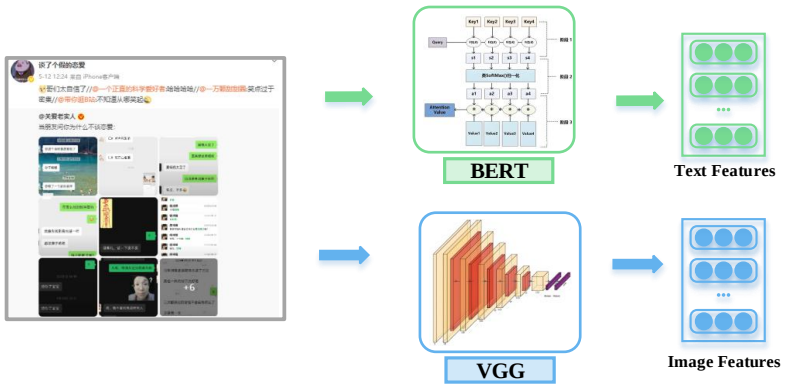

Figure 2. Multimodal feature extraction network.

Social media posts, such as tweets, generally consist of both text and images. The CMDF model capitalizes on this multimodal nature by processing both text and image data to determine whether a tweet is a rumor. Each tweet is represented as $P = T, V_1, V_2, ..., V_M$, where T denotes the textual content and $V_1, V_2, ..., V_M$ refers to the associated images. To extract relevant features, the multimodal feature extraction layer utilizes pretrained models to convert text and image information into feature vectors. As shown in Figure 2, this layer comprises a text feature extraction network and an image feature extraction network.

The text in a social media tweet is represented as a sequence of words, denoted as $T = W_1, W_2, ..., W_M$. Each word is processed individually by the text feature extraction network to generate its corresponding feature vector. Various neural network models for text feature extraction have been proposed, with the bidirectional encoder representations from transformers (BERT) model being the most advanced. In this study, a pretrained BERT model is employed to encode words into vector representations. The BERT model uses a multilayer, bidirectional self-attention mechanism to deeply analyze the text and produce feature representations across multiple hidden layers. Typically, the vector associated with the special token is extracted, serving as the global semantic representation of the entire text. This feature captures the core information of the text and is used in rumor detection tasks to assess the credibility of the content. The calculation for extracting text features from a tweet using the pretrained BERT model is given as.

$$P_T = BERT(W_M), \tag{3.1}$$

where P_T is the vector of the characteristic of the text after the BERT model extracts the text. Since the datasets used in this study are from Weibo and Twitter, the BERT model for both Chinese and English is employed.

In contrast to plain text, tweets often include multiple images alongside the written content. Various image processing techniques are available. The visual geometry group (VGG), a deep convolutional neural network introduced by [26] in 2014, improves the model's representation capacity by increasing the depth with more convolutional layers. Each convolutional layer utilizes 3×3 convolutional filters, with (ReLU) activation functions applied between each layer. The process for extracting image features from tweets is computed as

$$P_V = VGG(V_M), \tag{3.2}$$

where P_V represents the feature vector extracted from images by the pretrained VGG model. This paper employs a pretrained VGG16 model (VGG containing 16 layers) as the image feature extraction network, transforming image data from tweets into vector representations.

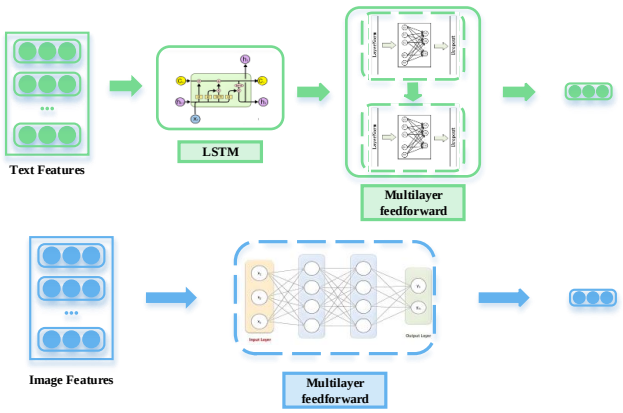

Figure 3. Multimodal deep hidden information extraction network.

3.2. Multimodal deep information extraction network

The multimodal deep information extraction network processes the feature vectors of both text and images, which are extracted from the multimodal feature extraction layer, to reveal deep information within these vectors. The architecture is shown in Figure 3. This layer is composed of two networks: one for text-based deep information extraction and another for image-based deep information extraction.

The network structure for deep text information extraction consists of an input linear transformation, an LSTM network, and a multilayer feedforward network. The text feature vector extracted by the BERT model is first linearly transformed through a fully connected layer to match the input requirements of the LSTM network. To ensure comprehensive information capture, the transformed text features are processed through a double-layer LSTM neural network. The outputs are subsequently concatenated and fed into a multilayer feedforward network to extract deep textual features. The specific mathematical expression is delineated as follows.

Linear transformation through fully connected layers is given as,

$$x_1 = Linear_1(P_T), \tag{3.3}$$

where x_1 is the eigenvector obtained by linear transformation, and $Linear_1$ is the fully connected layer used for linear transformation.

The feature vector obtained by linear transformation is fed into the dual-head LSTM network layer for processing:

$$X_{lstm}, [h_t, c_t] = LSTM(x_1, h_{t-1}, c_{t-1}), \tag{3.4}$$

where X_{lstm} is the hidden state of the sequence after LSTM processing, and h_{t-1} and c_{t-1} are the hidden state and cell state of LSTM at the previous moment.

Normalization standard processing of the sequence hidden state after LSTM processing:

$$X_{norm} = LayerNorm(X_{lstm}), \tag{3.5}$$

where X_{norm} is the eigenvector obtained by normalization standard processing.

The feature vector, obtained through normalization, is fed into the multilayer feedforward network to extract the final deep text information features:

$$\begin{aligned} X_{layer1} &= Dropout(ReLU(X_{norm}W_1 + b_1)), \\ X_{layer2} &= Dropout(ReLU(X_{layer1}W_2 + b_2)), \\ X_{deeptext} &= X_{layer2}W_3 + b_3, \end{aligned} \tag{3.6}$$

where W_1, W_2 and W_3 are the weight matrices of the multilayer feedforward and b_1, b_2 and b_3 are the bias vectors.

In contrast to text feature information, the image deep information extraction network is relatively simple, as the VGG16 model has already extracted high-level image features, such as texture, edges, and shapes. The mathematical expression is given as

$$\begin{aligned} X_{layer1} &= Linear_2(P_V), \\ X_{deepimage} &= Linear_3(X_{layer1}), \end{aligned} \tag{3.7}$$

where $Linear_2$ and $Linear_3$ are the two fully connected layers of the image depth information extraction network, and $X_{deepimage}$ is the image depth information feature extracted by the image depth information extraction network.

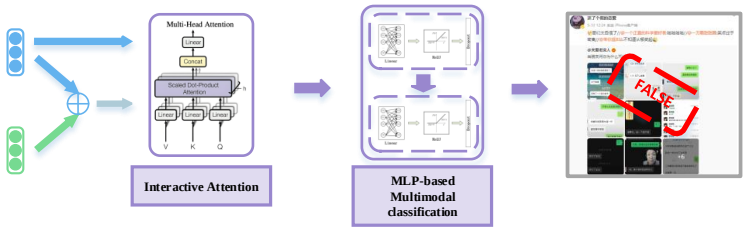

Figure 4. Cross-modal deep fusion neural network.

3.3. Cross-modal deep fusion network

This paper proposes a cross-modal deep fusion network, with text as the primary processing input and images as auxiliary inputs, performing cross-modal fusion between text and images for rumor detection. The architecture is shown in Figure 4. The neural network model includes an interactive attention mechanism and a multilayer perceptron (MLP)-based multimodal classifier. Its mathematical formulation is presented as follows.

To utilize the interactive multihead attention mechanism, it is necessary to generate the Query, Key and Value:

$$\begin{aligned} Q &= X_{deeptext} W_Q, \\ K &= X_{deepimage} W_K, \\ V &= X_{deepimage} W_V, \end{aligned} \tag{3.8}$$

where W_Q, W_K, and W_V are attention weight matrices.

The multihead attention scores are calculated and weighted as follows:

$$\begin{aligned} Attention(Q, K, V) &= softmax(d_K^{-1/2} QK^T), \\ head_i &= Attention(Q_i, K_i, V_i), \\ MutiHead(Q, K, V) &= Concat(head_1, ..., head_n) W_o, \\ T_{Attention} &= MutiHead(X_{deeptext}, X_{deepimage}, X_{deepimage}), \end{aligned} \tag{3.9}$$

where Q, K and V are the Query vector, Key vector and Value vector, respectively. $d_K^{-1/2} QK^T$ is the correlation matrix between Query and Key, i.e. the attention score matrix. $head_i$ represents the attention feature after each attention mechanism has been applied. W_o is the projection matrix of all heads after attention and concatenation to restore the dimensions. $T_{Attention}$ represents the fused feature resulting from the deep processing of the multihead attention mechanism.

After the fused features are concatenated with the original textual deep information features, they are

input into an MLP-based multimodal classification network to produce the rumor classification results:

$$\begin{aligned} X_{tv}^{1} &= Cat(T_{Attention}, X_{deeptext}), \\ X_{tv}^{2} &= Dropout_{0.5}(ReLU(W_4 X_{tv}^{1} + b_4)), \\ X_{tv}^{3} &= Dropout_{0.5}(ReLU(W_5 X_{tv}^{2} + b_5)), \\ X_{classifier} &= W_6 X_{tv}^{3} + b_6, \end{aligned} \tag{3.10}$$

where W_4, W_5 and W_6 are the weight matrices of the hidden layer and the output layer, b_4, b_5 and b_6 are the bias vectors of the hidden layer and the output layer, and $X_{classifier}$ represents the final rumor classification output, which is a two-dimensional vector indicating the probabilities of the object being a rumor or a non-rumor. By comparing these probabilities, the object is classified as either a rumor or a non-rumor.

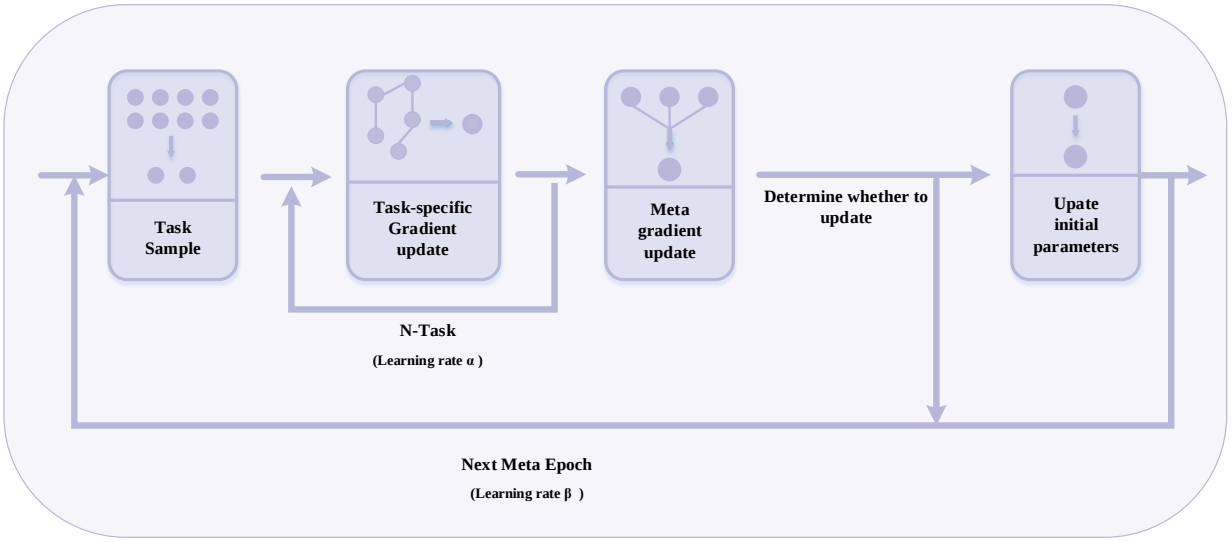

Figure 5. Enhanced meta-learning algorithm based on MAML.

3.4. Enhanced meta-learning algorithm based on MAML

With the continued advancement of deep learning, the complexity of neural networks has increased, but their performance has not shown significant improvement relative to this growing complexity. To address the issue of limited early rumor samples, this paper proposes an enhanced meta-learning training approach based on the model-agnostic meta-learning framework. By employing distinct learning rates both within and between tasks, the method facilitates the efficient updating of target model parameters, effectively processes small-sample data, and improves the training paradigm for early rumor detection. Furthermore, an early stopping mechanism is incorporated to monitor the highest achieved accuracy and compare it with current performance, thereby determining the necessity of subsequent meta-parameter updates. This approach effectively mitigates issues arising from excessive meta-learning updates. The specific steps of the algorithm are shown in Algorithm 1.

MAML is a meta-learning algorithm designed to enable models to quickly adapt to new tasks using minimal data, which is shown in Figure 5. MAML is "model-agnostic", meaning that it can be applied

to any model trained through gradient-based optimization. The fundamental concept of MAML is that, rather than training on a single task, the model is trained across a distribution of tasks, allowing it to learn a robust initialization that generalizes effectively to different tasks. MAML employs a uniform learning rate for both intra-task and inter-task parameter updates during training. This approach, however, is suboptimal for diverse task types, particularly in the context of rumor detection. To address this limitation, an enhanced meta-learning algorithm based on MAML is proposed that utilizes distinct learning rates–denoted as α for intra-task updates and β for inter-task updates, which thereby enhances the model's efficacy in detecting various types of rumors. The specific mathematical expression is as follows.

Task sampling: At the beginning of each meta-learning cycle, both rumor and non-rumor samples are drawn from the dataset to be used as inputs for the meta-learning process. In a *N-way K-shot Q-query* learning task, a mini-dataset is created, consisting of N distinct classes, with K support samples and Q query samples per class. Here, N represents the number of task categories, while the K support and Q query samples are randomly selected from both rumor and non-rumor instances.

Task-specific gradient update: Initially, for each task T_k, the model is initialized with a set of parameters θ, which are the shared parameters across all tasks. Then, these initial parameters can be considered as the starting point for training on any individual task. Lastly, train the model using the support dataset, set an initial learning rate α for each task, and perform a single gradient update:

$$\theta'_k = \theta - \alpha \nabla_\theta \mathcal{L}_k(\theta, D_k^{train}), \tag{3.11}$$

where α is the learning rate for intra-task parameter updates, $\mathcal{L}_k(\theta, D_k^{train})$ is the loss function for task T_k based on the training set D_k^{train}, and $\nabla_\theta \mathcal{L}_k(\theta, D_k^{train})$ is the loss gradient with respect to the model parameters θ. This gradient indicates how much the model parameters need to change to reduce the loss for that task.

Meta-gradient update: After executing task-specific updates, the loss should be evaluated in a separate validation set D_k^{val} for each task in order to assess the performance of the task-specific parameters θ'_k:

$$\mathcal{L}_k^{val}(\theta'_k) = \mathcal{L}_k(\theta'_k, D_k^{val}), \tag{3.12}$$

where θ'_k is the parameter after a gradient update for a specific task, and $\mathcal{L}_k^{val}$ is the loss function obtained using the validation set. In order to minimize the validation loss for all tasks simultaneously, the meta-gradient needs to be calculated, which is the gradient of the loss with respect to the initial model parameters:

$$\nabla_\theta \mathcal{L}^{meta}(\theta) = \frac{1}{K} \sum_{k=1}^{K} \nabla_\theta \mathcal{L}_k^{val}(\theta'_k), \tag{3.13}$$

where ∇_θ represents the gradient, and $\mathcal{L}^{meta}(\theta)$ is the average validation loss over all tasks.

Update the initial parameters: If the validation accuracy exceeds the recorded optimal accuracy, update the meta-knowledge obtained through each task training, set the training learning rate β, and perform gradient descent steps to update the initial model parameters:

$$\theta_{target} = \theta - \beta \nabla_\theta \mathcal{L}^{meta}(\theta), \tag{3.14}$$

where β is the learning rate for the meta-update.

Conventional training methods often fail to extract meaningful patterns from small datasets, limiting their effectiveness in rumor detection. To address this issue, we employ an enhanced meta-learning training algorithm based on MAML, which enhances detection efficiency despite limited training samples. Additionally, meta-learning improves model accuracy while maintaining computational simplicity.

Algorithm 1 Enhanced meta-learning algorithm based on MAML

Input:

Initial target model parameter θ;
Learning rate within the task α;
Learning rate between the task β;
Total sample of rumors P_{Task};
Loss function $\mathcal{L} = -[y * log(p) + (1 - y) * log(1 - p)]$;

Output: Target model parameter θ_{target};

Initialize target parameters θ;
while in Meta epoch **do**
    Obtain N tasks by sampling from the total dataset;
    for all Task **do**
        Perform a single gradient update using α (3.11);
        Evaluate parameter performance within the task (3.12);
    end for
    Simultaneously minimize the validation loss for all tasks (3.13);
    if performance is better than historica **then**
        Using β to update the target parameters (3.14);
    end if
    end while

4. Results

4.1. Experiments set

Table 1. Statistics from two online social media platforms.

Datasets	Weibo	Twitter
Rumor	4749	10,717
Non-rumor	4779	7332
True text	4779	7332
Fake text	4749	10,717
True image	5318	297
Fake image	7954	245

Datasets: To more effectively evaluate the performance of the proposed model, four benchmark models were selected for comparison, and experiments were conducted using datasets from two social media platforms, namely Weibo and Twitter [1, 2], shown in Table 1. The Weibo dataset is a multitopic dataset, including text and image collected from tweets. The dataset is sourced from verified false posts in the official Weibo debunking system from May 2012 to January 2016, as well as authoritative tweets from Chinese news organizations. The Twitter dataset is derived from a multimodal testing task used by

MediaEval in 2016, divided into training and testing sets. The training dataset contains 6225 real posts and 9596 fake posts, while the test dataset contains 1107 real posts and 1121 fake posts.

Implementation Rules: This paper employs a fixed learning rate of 0.00007 and trains the model for 20 epochs on each dataset. To assess the model's overall performance, its accuracy, precision, recall, and F1 score are used as evaluation metrics. Cross-entropy as the loss function and Adam as the optimizer for meta-learning updates are employed. To prevent overfitting and enhance generalization, an early stopping mechanism during training is employed. The training set, validation set, and test set are divided into an 8:1:1 ratio.

Evaluation indicators: To evaluate the performance of the proposed model, four standard metrics were employed: accuracy, precision, recall, and the F1-score. These metrics are calculated from the model's prediction outcomes and can be directly interpreted from the confusion matrix.

A minimal 2×2 confusion matrix for binary classification is shown below:

	Predicted positive	Predicted negative
Actual positive	TP	FN
Actual negative	FP	TN

Accuracy measures the overall proportion of correctly classified samples, reflecting the model's general effectiveness:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.1}$$

Precision evaluates the reliability of the model when predicting the positive class, indicating how many predicted positives are actually correct:

$$Precision = \frac{TP}{TP + FP}. \tag{4.2}$$

Recall (also known as sensitivity) measures the model's ability to detect all actual positive samples, describing how many true positives are successfully retrieved:

$$Recall = \frac{TP}{TP + FN}. \tag{4.3}$$

The F1-score is the harmonic mean of Precision and Recall, balancing both metrics especially in imbalanced classification scenarios:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \tag{4.4}$$

4.2. Comparison baseline

This paper compares the proposed small-sample based CMDF early rumor detection method with four baseline models, demonstrating that the proposed method outperforms these baselines. The baseline models are given as follows.

Table 2. The performance comparisons between small-sample based CMDF early rumor detection method and baseline neural networks on the Weibo dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Prototypical	76.00 (3.57)	38.59 (1.24)	51.71 (1.23)	44.20 (1.12)
Siamese	87.75 (3.89)	60.14 (2.06)	**93.69 (3.02)**	73.26 (1.78)
MFAN	82.75 (3.68)	**90.54 (2.45)**	68.80 (2.54)	78.19 (2.23)
BiGCN-early	86.25 (3.83)	62.77 (1.96)	71.24 (2.68)	66.74 (1.93)
SEQ-CNNs	87.50 (4.23)	64.67 (2.17)	67.43 (2.43)	66.02 (1.90)
CMDF	**92.50 (3.13)**	86.59 (2.35)	76.36 (2.98)	**81.15 (2.41)**

Prototypical networks: This baseline [27] proposes a prototype-based small-sample classification method, where each category is represented by a prototype vector, and classification is performed using Euclidean distance.

Siamese neural networks: This baseline presents a FSL method using Siamese networks to learn a similarity function for comparing feature vectors and determining if inputs belong to the same category [28].

MFAN: This baseline proposes a multimodal feature-enhanced attention network that integrates textual, visual, and social graph features for rumor detection [3].

BiGCN-early: A new graph neural network model, named bidirectional graph convolutional network (BiGCN), is proposed, which integrates bidirectional propagation structure and time series features in social media networks for precise rumor detection [29].

SEQ-CNNs: The sequential convolutional neural networks (SEQ-CNNs) uses a convolutional model to automatically detect and classify false information from text, showcasing its efficiency and scalability on large datasets [30].

Among the benchmark models evaluated, BiGCN-early and SEQ-CNNs utilize convolutional neural networks to extract textual and temporal features for misinformation detection. The Siamese neural network and prototypical network employ distance metrics and similarity comparisons to address few-shot learning challenges. MFAN integrates a multihead interactive attention mechanism to fuse textual and visual features for multimodal misinformation detection. For comparative evaluation, all benchmark models, including CMDF, were trained for 20 epochs with early stopping to prevent overfitting.

4.3. Comparison experiment

Experiments are conducted on the proposed model and baselines using the Twitter and Weibo datasets, with results presented in Tables 2 and 3. From the results in the tables, the conclusions are as follows:

- The Siamese and Prototypical models, which rely on distance metrics, performed poorly on the Weibo and Twitter datasets, except for the Siamese model, which showed limited success on Weibo. This indicates that few-shot learning methods employing Euclidean distance metrics for text and image features comparison are ineffective for early rumor detection.
- The MFAN method employs an attention mechanism for cross-modal processing of text, image, and social graph features, yet demonstrates suboptimal performance. Despite implementing preprocessing for all three modalities and cross-modal fusion techniques, the limited availability

Table 3. The performance comparisons between small-sample based CMDF early rumor detection method and baseline neural networks on the Twitter dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Prototypical	78.50 (3.43)	63.30 (1.23)	49.39 (0.75)	55.49 (1.21)
Siamese	74.25 (3.21)	71.29 (1.54)	51.69 (0.98)	59.93 (1.32)
MFAN	82.50 (3.57)	**91.38 (2.21)**	**81.30 (1.01)**	**86.05 (1.65)**
BiGCN-early	87.25 (3.78)	72.88 (1.48)	65.04 (1.14)	68.74 (1.43)
SEQ-CNNs	80.25 (4.06)	78.30 (1.76)	64.05 (1.11)	70.46 (1.57)
CMDF	**90.50 (3.01)**	83.75 (1.98)	72.48 (1.24)	77.71 (1.61)

of early rumor samples renders this approach ineffective. The method struggles to effectively integrate multimodal features for rumor detection while simultaneously suffering from reduced accuracy due to model overfitting.

- BiGCN-early and SEQ-CNNs, which employ neural networks for feature extraction and processing in rumor detection, demonstrate strong performance. This suggests that simpler network architectures prove more effective for rumor detection given limited early-stage samples. However, these simplified networks fail to comprehensively integrate textual and visual information, potentially overlooking critical features and consequently compromising detection performance.
- The experimental results demonstrate that the proposed CMDF model outperforms others, achieving the highest accuracy and F1-scores of 92.50% and 81.15%, respectively, on the Weibo dataset, and 90.60% and 77.71% on the Twitter dataset. Although the CMDF model achieves the highest accuracy on the Twitter dataset, its F1-score is not the highest among the compared methods. This discrepancy is primarily attributed to the strong class imbalance in the Twitter dataset, where non-rumor samples substantially outnumber rumor samples, causing accuracy to be dominated by majority-class performance while F1-score remains more sensitive to precision and recall on the minority class. The CMDF model employs a pretrained model for text and image feature extraction, constructing a multimodal deep information extraction network to learn comprehensive multimodal representations. Specifically, a bidirectional LSTM architecture prevents information loss during text feature extraction. An interactive attention mechanism then performs cross-modal fusion, with text features serving as the primary key. The model employs an MLP-based classifier for rumor detection output. Finally, the integration of the enhanced meta-learning training algorithm based on MAML improves lightweight cross-modal neural network fusion performance and optimizes meta-learning training methodology.

4.4. Ablation experiment

To validate the innovation of early rumor detection using small samples of cross-modal deep fusion, four simplified neural network models are implemented.

CMDF without Text: In this model, the text depth information extraction network is removed, and only the image depth information is extracted for processing.

CMDF without Image: In this model, the image depth information extraction network is removed, and only extracts the text depth information is extracted for processing.

Table 4. The comparisons of the ablation models and CMDF on the Weibo dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CMDF w/o Text	77.80 (2.31)	67.50 (1.79)	65.42 (1.77)	66.44 (1.54)
CMDF w/o Image	84.65 (2.12)	63.20 (1.63)	62.34 (1.71)	62.77 (1.51)
CMDF w/o Attention	61.34 (1.75)	55.82 (1.58)	48.94 (1.46)	52.15 (1.37)
CMDF w/o MAML	56.75 (1.84)	43.85 (1.54)	34.50 (1.35)	38.62 (1.05)
CMDF	**92.50 (1.96)**	**86.59 (1.94)**	**76.36 (1.89)**	**81.15 (1.71)**

Table 5. The comparisons of the ablation models and CMDF on the Twitter dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CMDF w/o Text	74.53 (2.11)	63.60 (0.97)	66.25 (1.07)	64.90 (1.31)
CMDF w/o Image	81.45 (2.53)	57.89 (0.68)	64.53 (1.11)	61.03 (1.24)
CMDF w/o Attention	65.47 (1.87)	54.82 (0.67)	53.71 (1.21)	54.26 (1.19)
CMDF w/o MAML	64.30 (1.93)	54.87 (0.64)	48.55 (0.87)	51.52 (1.12)
CMDF	**90.50 (2.21)**	**83.75 (0.87)**	**72.48 (1.32)**	**77.71 (1.01)**

CMDF without Attention: In this model, the interactive attention mechanism is replaced with a simple concatenation method to fuse text and image features.

CMDF without MAML: In this model, the MAML algorithm is removed, and the model is trained only using general neural network training methods.

The models are tested on the Twitter and Weibo datasets, and the results are presented in Tables 4 and 5. We summary the conclusions as follows.

- CMDF w/o Text and CMDF w/o Image exhibit significantly inferior performance compared to the CMDF model on both Weibo and Twitter datasets. This result demonstrates that the multimodal hidden information extraction network effectively integrates textual and visual features, thereby improving rumor detection performance.
- On both Weibo and Twitter datasets, the CMDF w/o Image model achieved higher accuracy than the CMDF w/o Text model. This performance difference stems from text features' stronger contextual relationships and richer semantic characteristics compared to image features, enabling neural networks to more effectively detect rumors through textual analysis.
- Experimental results on both Weibo and Twitter datasets demonstrate that the CMDF w/o Attention underperforms not only the standard CMDF model but also the CMDF w/o Text model and the CMDF w/o Image model. This finding indicates that basic unimodal processing of text or images alone is insufficient for effective early-stage rumor detection. The proposed deep cross-modal fusion network addresses this limitation by effectively extracting and integrating cross-modal features, thereby significantly enhancing rumor detection accuracy.
- The CMDF w/o MAML demonstrated the poorest performance on both Weibo and Twitter datasets, showing significantly lower accuracy than the standard CMDF model, CMDF w/o Text model, and CMDF w/o Image. It also performed marginally worse than the CMDF w/o Attention model. These results indicate that the enhanced meta-learning training algorithm based on MAML effectively addresses two key challenges: 1) limited early rumor samples through meta-learning

adaptation, and 2) neural network simplicity by achieving performance comparable to more complex architectures.

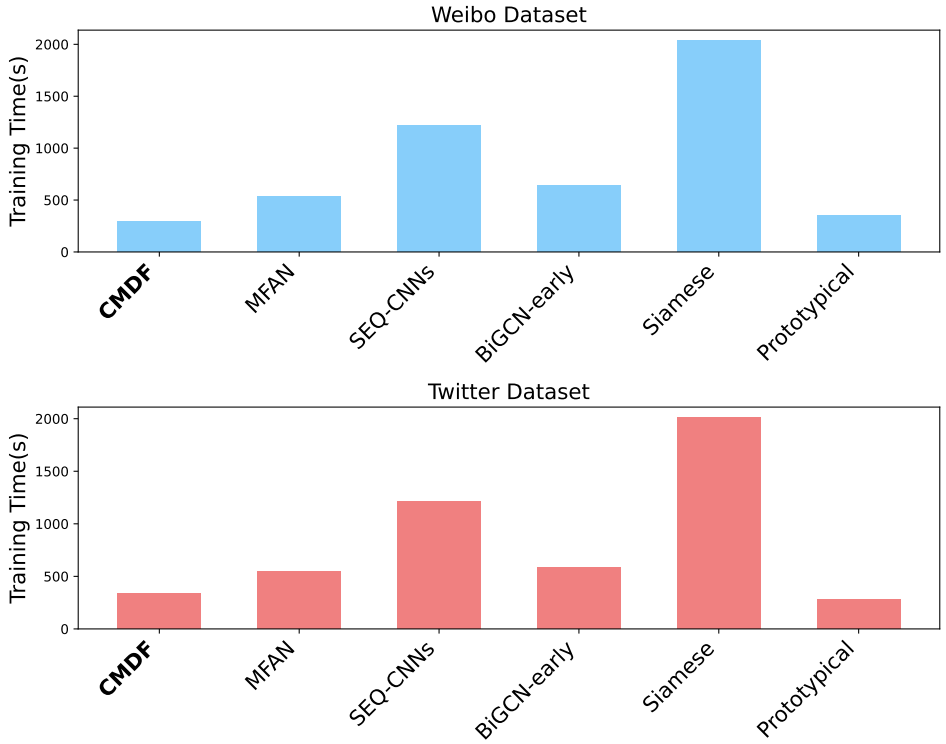

Figure 6. The training time of the CMDF and baselines on Weibo and Twitter dataset.

4.5. Training time

In order to verify the effectiveness of the CMDF method in improving the training method, this section records the training time of the CMDF method and the baseline model. The experimental results are shown in Figure 6. As illustrated in the figure, the CMDF method exhibits a relatively short training time–slightly longer than that of the Prototypical model but shorter than models such as MFAN and Siamese. This indicates that despite incorporating an enhanced MAML algorithm, CMDF does not incur additional training time and, in fact, outperforms typical baseline methods in efficiency. Consequently, the MAML integration not only effectively mitigates the challenge of limited early-stage rumor samples but also maintains a competitive training duration.

5. Conclusions

Given the limited availability of early rumor samples and their constrained propagation scope, this paper proposes a cross-modal deep fusion framework for early rumor detection under small-sample conditions. The framework comprises three key components: a multimodal feature extraction network that processes both textual and visual content, a multimodal deep information extraction network designed to capture richer semantic representations across modalities, and a cross-modal deep fusion

network that enables comprehensive integration of textual and visual features. Furthermore, an enhanced meta-learning training algorithm based on model-agnostic meta-learning is introduced to effectively address the challenge of limited training samples. Experimental evaluation on two benchmark datasets demonstrates the framework's strong performance, with the CMDF model achieving 92.50% accuracy and 81.15% F1-score on the Weibo dataset, and 90.50% accuracy with 77.71% F1-score on the Twitter dataset. These compelling results validate both the effectiveness of the proposed architecture and the efficacy of the multimodal fusion strategy combined with meta-learning in handling sample scarcity. Nevertheless, the method exhibits certain limitations regarding generalization across diverse rumor types and increased computational demands when incorporating additional modalities. Future work will focus on integrating more diverse data modalities and developing more efficient fusion mechanisms to enhance both performance and applicability.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. Z. Jin, J. Cao, H. Guo, Y. Zhang, J. Luo, Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in *Proceedings of the 25th ACM International Conference on Multimedia*, (2017), 795–816. https://doi.org/10.1145/3123266.3123454
2. C. Boididou, S. Middleton, S. Papadopoulos, D. Dang, M. Riegler, G. Boato, et al., The vmu participation@ verifying multimedia use 2016, *Comput. Sci.*, 2016.
3. J. Zheng, X. Zhang, S. Guo, Q. Wang, W. Zang, Y. Zhang, Mfan: Multi-modal feature-enhanced attention networks for rumor detection, in *Proceedings of the 32th International Joint Conference on Artificial Intelligence*, (2022), 2413–2419, https://doi.org/10.24963/ijcai.2022/335.
4. W. Wang, Y. Qiu, S. Xuan, W. Yang, Early rumor detection based on deep recurrent q-learning, *Secur. Commun. Netw.*, **2021** (2021), 5569064.
5. Z. Zhang, Z. Dan, F. Dong, Z. Gao, Y. Zhang, A rumor detection method based on adaptive fusion of statistical features and textual features, *Information*, **13** (2022), 388. https://doi.org/10.3390/info13080388.
6. X. Liu, D. Wang, Pagcn: Structural semantic relationship and attention mechanism for rumor detection, *Appl. Sci.*, **15** (2025), 8984. https://doi.org/10.3390/app15168984.
7. H. Li, L. Jiang, J. Li, Continuous-time dynamic graph networks integrated with knowledge propagation for social media rumor detection, *Mathematics*, **12** (2024), 3453. https://doi.org/10.3390/math12223453.

8. H. Sha, L. Zhu, Dynamic analysis of pattern and optimal control research of rumor propagation model on different networks, *Inf. Process. Manage.*, **62** (2025), 104016. https://doi.org/10.1016/j.ipm.2024.104016
9. L. Zhu, Y. Ding, S. Shen, Green behavior propagation analysis based on statistical theory and intelligent algorithm in data-driven environment, *Math. Biosci.*, **379** (2025), 109340. https://doi.org/10.1016/j.mbs.2024.109340
10. T. Yang, L. Zhu, S. Shen, L. He, Pattern dynamics analysis and parameter identification of spatiotemporal infectious disease models on complex networks, *Math. Biosci.*, **387** (2025), 109502. https://doi.org/10.1016/j.mbs.2025.109502
11. N. Zhong, G. Zhou, W. Ding, J. Zhang, A rumor detection method based on multi-modal feature fusion by a joining aggregation structure, *Electronics*, **11** (2022), 3200. https://doi.org/10.3390/electronics11193200.
12. J. Lai, X. Yang, W. Luo, L. Zhou, L. Li, Y. Wang et al., Rumorllm: A rumor large language model-based fake-news-detection data-augmentation approach, *Appl. Sci.*, **14** (2024), 3532. https://doi.org/10.3390/app14083532.
13. T. Chen, X. Li, H. Yin, J. Zhang, Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection, in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, (2018), 40–52.
14. H. Lu, C. Fan, X. Song, W. Fang, A novel few-shot learning based multi-modality fusion model for covid-19 rumor detection from online social media, *PeerJ Comput. Sci.*, **7** (2021), e688. https://doi.org/10.7717/peerj-cs.688
15. Q. Huang, J. Yu, J. Wu, B. Wang, Heterogeneous graph attention networks for early detection of rumors on twitter, in *2020 International Joint Conference on Neural Networks*, (2020), 1–8.
16. H. T. Le-Pham, T. P. Nguyen, A. D. Tran, K. H. Le, Mtas: A temporal-aware multiview graph attention framework for early rumor detection on social media, *Inf. Fusion*, **126** (2026), 103598. https://doi.org/10.1016/j.inffus.2025.103598
17. T. D. Nguyen, A. K. N. Vu, N. D. Nguyen, V. T. Nguyen, T. D. Ngo, T. T. Do, et al., The art of camouflage: Few-shot learning for animal detection and segmentation, *IEEE Access*, **12** (2024), 103488–103503. https://doi.org/10.1109/ACCESS.2024.3432873
18. M. E. H. Habib, A. Küçükmanisa, O. Urhan, Enhanced protonet with self-knowledge distillation for few-shot learning, *IEEE Access*, **12** (2024), 145331–145340. https://doi.org/10.1109/ACCESS.2024.3472530
19. M. Seo, H. Kim, Irregular openings identification at construction sites based on few-shot learning, *Buildings*, **15** (2025), 1834. https://doi.org/10.3390/buildings15111834.
20. Y. Li, X. Gu, Y. Wei, A deep learning-based method for bearing fault diagnosis with few-shot learning, *Sensors*, **24** (2024), 7516. https://doi.org/10.3390/s24237516.
21. C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in *International Conference on Machine Learning*, (2017), 1126–1135.
22. W. Zhou, M. Lu, R. Ji, Meta-se: A meta-learning framework for few-shot speech enhancement, *IEEE Access*, **9** (2021), 46068–46078. https://doi.org/10.1109/ACCESS.2021.3066609

23. F. Zhou, X. Han, J. Ren, W. Wang, Y. Wang, P. Zhang, et al., A method for few-shot modulation recognition based on reinforcement metric meta-learning, *Computers*, **14** (2025), 346. https://doi.org/10.3390/computers14090346.

24. X. Zhou, H. Bai, Z. Dong, K. Zhou, Y. Liu, Small sample palmprint recognition based on image augmentation and dynamic model-agnostic meta-learning, *Electronics*, **14** (2025), 3236. https://doi.org/10.3390/electronics14163236.

25. S. Nasim, A. S. Al-Shamayleh, N. Thalji, A. Raza, L. Abualigah, A. I. Alzahrani, et al., Novel meta learning approach for detecting postpartum depression disorder using questionnaire data, *IEEE Access*, **12** (2024), 101247–101259. https://doi.org/10.1109/ACCESS.2024.3427685

26. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556

27. J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, *Adv. Neural Inf. Process. Syst.*, **30** (2017), 4080–4090.

28. G. Koch, R. Zemel, R. Salakhutdinov, Siamese neural networks for one-shot image recognition, in *ICML Deep Learning Workshop*, (2015), 1–30.

29. T. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong, et al., Rumor detection on social media with bi-directional graph convolutional networks, in *Proceedings of the AAAI Conference on Artificial Intelligence*, (2020), 549–556. https://doi.org/10.1609/aaai.v34i01.5393

30. F. Yu, Q. Liu, S. Wu, L. Wang, T. Tan, A convolutional approach for misinformation identification, in *IJCAI'17: Proceedings of the 26th International Joint Conference on Artificial Intelligence*, (2017), 3901–3907.

AIMS Press