

International Journal of Simulation and Process Modelling

ISSN online: 1740-2131 - ISSN print: 1740-2123

<https://www.inderscience.com/ijspm>

Optimising multi-modal fusion with a tri-encoder tensor network for process applications

Jiayao Li, Li Li, Xiaochen Shi, Zeqiu Chen, Kaiyi Zhao, Ruizhi Sun

DOI: [10.1504/IJSPM.2025.10071854](https://doi.org/10.1504/IJSPM.2025.10071854)

Article History:

Received:	07 January 2025
Last revised:	18 April 2025
Accepted:	23 May 2025
Published online:	01 September 2025

Optimising multi-modal fusion with a tri-encoder tensor network for process applications

Jiayao Li

School of Software,
Shanxi Agricultural University,
Taigu 030801, China
and
College of Information and Electrical Engineering,
China Agricultural University,
Beijing 110000, China
Email: lijiaoyao@cau.edu.cn

Li Li

Computer School,
Beijing Information Science and Technology University,
Beijing, 100101, China
Email: lili_2018@cau.edu.cn

Xiaochen Shi and Zeqiu Chen

College of Information and Electrical Engineering,
China Agricultural University,
Beijing 110000, China
Email: sxc@cau.edu.cn
Email: chenzq@cau.edu.cn

Kaiyi Zhao

School of Computer Science and Technology,
North University of China,
Taiyuan, 030051, China
Email: zhaokaiyi@nuc.edu.cn

Ruizhi Sun*

College of Information and Electrical Engineering,
China Agricultural University,
Beijing 110000, China
and
Scientific Research Base for Integrated Technologies of
Precision Agriculture (Animal Husbandry),
The Ministry of Agriculture,
Beijing 110000, China
Email: sunruizhi@cau.edu.cn
*Corresponding author

Abstract: Multi-modal fusion combines information from diverse modalities, enabling scalable predictions in complex data environments. To address the existing limitations in capturing modalities interactions and reducing time consumption, we propose an optimising multi-modal fusion with a tri-encoder tensor network for process applications (TS-OMMF). Specifically, the input modalities are encoded to abstract the intra-modal feature and obtain the representation respectively. The representations are fused into a high dimension space with a low-rank factor to limit the dimension, and a linear extension pattern is employed to assist model extend for multi-modal. The fusion features are input into a novel tri-encoder to capture the inter-modal feature at a finer granularity and obtain the complementary features, thereby reducing the time consumption. Extensive experiments on benchmark multi-modal datasets demonstrate that

TS-OMMF improves performance metrics by 0.8% to 6.1%. These results highlight TS-OMMF practical applicability, scalability, and potential for advancing process modelling and other complex multi-modal data-driven tasks.

Keywords: tensor fusion; encoder-decoder; tensor representation; multi-modal fusion; MMF.

Reference to this paper should be made as follows: Li, J., Li, L., Shi, X., Chen, Z., Zhao, K. and Sun, R. (2025) 'Optimising multi-modal fusion with a tri-encoder tensor network for process applications', *Int. J. Simulation and Process Modelling*, Vol. 22, Nos. 1/2, pp.75–92.

Biographical notes: Jiayao Li is a PhD student in the College of Information and Electrical Engineering, China Agricultural University, China. She received her MS degree from China Agricultural University, China in 2018. She is currently a lecturer with the School of Software, Shanxi Agricultural University. Her research interests include machine learning and outlier detection.

Li Li has received her MS degree from the School of Mathematics and Physics, University of Science and Technology Beijing, in 2017, and the PhD degree from the College of Information and Electrical Engineering, China Agriculture University. She is currently a Lecturer with the Computer School, Beijing Information Science and Technology University, Beijing. Her current research interests include machine learning, text classification, outlier detection and data security.

Xiaochen Shi is a PhD student in the College of Information and Electrical Engineering, China Agricultural University, China. She received her MS degree from China Agricultural University, China in 2021. Her research interests include computer vision and the application of large models in agriculture.

Zequi Chen received the Master's degree in Computer Science and Technology from China Agricultural University, Beijing, China, in 2021, where she is currently working toward the PhD degree in computer science and technology in the College of Information and Electrical Engineering. Her research interests include multi-modal learning, anomaly detection, and neural networks.

Kaiyi Zhao received the PhD degree in computer science and technology from China Agricultural University, Beijing, China, in 2023. He is currently an Assistant Professor with the School of Data Science and Technology, North University of China, Taiyuan, China. His current research interests mainly focus on information fusion, uncertain information processing, and anomaly detection.

Ruizhi Sun received the PhD degree in computer science from Tsinghua University. He is a Full Professor with the College of Information and Electrical Engineering, China Agricultural University. His research interests include computer network and applications, workflow management and outlier detection. He is the member of China Agricultural Engineering Institute of electronic technology and computer application Specialised Committee.

1 Introduction

In recent years, process modelling has become increasingly crucial in various industries, such as industrial automation (Parkash and Tewari, 2024), delivery services (Zhang and Jia, 2024), agriculture (Li et al., 2024) and big data processing (Varun et al., 2024), where complex systems require accurate, efficient, and adaptable models. Traditional methods of process modelling often rely on single-modal data sources, which can limit the ability to capture the full spectrum of information inherent in these systems. Multi-modal fusion (MMF), the integration of different data types, has emerged as a promising approach to enhance process modelling by providing a more comprehensive and accurate representation of these systems. However, the existing MMF networks have certain limitations when applied to simulation and process modelling, such as

inefficiency in processing large volumes of heterogeneous data and difficulty in capturing intricate interdependencies among diverse data sources. Existing fusion methods are used in many fields, such as aggression detection (Jaafar and Lachiri, 2023), skin lesion diagnosis (He et al., 2023; Wang et al., 2022a), COVID-19 (Kumar et al., 2022), abnormal detection (Li et al., 2022), computer vision (Deng et al., 2022), micro-expression recognition (Zhao et al., 2022), machine translation (Li et al., 2023), sentiment recognition (Kim and Park, 2023). In the field of agriculture, multimodal learning can also play an important role in disease detection, fertiliser management and irrigation decision-making, and promote the transformation of agricultural production to intelligence and digitalisation. It can not only improve the efficiency and accuracy of agricultural production, but also provide new possibilities for realising sustainable agricultural

development. In addition. As shown in Figure 1, it is an example of MMF in sentiment analysis field. There are three modalities including acoustic, visual and language modality. Every modality provides features information to indicate the human sentiment classification. The language data is that something is happening, the acoustic has the higher pitch and the visual shows a smile emotion. From the three features information, the sentiment is predicted as the positive classification. And the predicted results are based on the prediction of acoustic, visual and language modalities. Obtaining multi-modal features (MMF) from multiple modalities improves the accuracy of prediction categories. MMF methods aim to mine and fuse the different modalities features to improve the prediction results. Efficiency of MMF methods has a great impact in practical problems. How to improve the efficiency of MMF methods are important.

Despite recent advances, several open challenges remain in the field of multimodal fusion, especially in domains requiring complex process modelling. First, many existing methods struggle to learn rich intra-modal representations, particularly from heterogeneous and noisy modalities. Second, fusion strategies often result in high-dimensional tensors, which significantly increase computational overhead and limit scalability. Third, current models tend to overlook fine-grained inter-modal interactions, which are critical for capturing complementary information across modalities. Finally, most frameworks lack modality extensibility and flexible co-learning capabilities, hindering their application in dynamic and diverse data environments. These challenges motivate our work, where we design a unified tri-encoder framework – TS-OMMF – to provide scalable, efficient, and accurate multimodal fusion for real-world applications.

At present, there are a lot of studies on MMF. Some previous studies proposed an early fusion method that directly uses feature connections in the input (Wang et al., 2016; Zadeh et al., 2018), and later, there are studies on the late fusion method that builds a separate model for each modality and then carries out weighted average results (Cai and Xia, 2015; Nojavanasghari et al., 2016). However, the former fails to consider the relationship between the contexts within the modalities, while the latter fails to consider the interaction information between the modalities. In this regard, some methods (Liu et al., 2024; Sun et al., 2023; Zeng et al., 2022) based on single encoder were proposed to learn the inter-modal relationship. But these models have high complexity and spend a lot of running time, which has a negative effect on specialisation areas application. Some scholars (Sahay et al., 2018; Xu et al., 2024; Zadeh et al., 2017) proposed tensor fusion. Tensor fusion is to express different modalities through tensors and then fuse them through tensor operation. Tensor fusion significantly improves

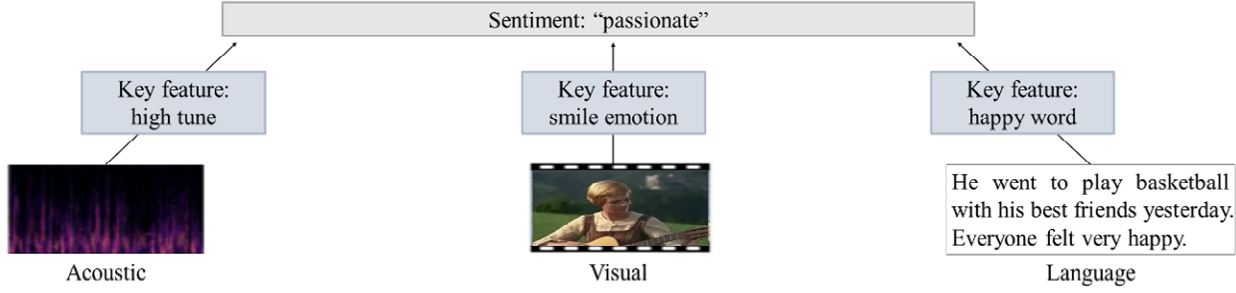
the robustness of MMF by end-to-end learning of in-modalities and inter-modalities information, but can only be applied to two-modalities. Extending the modality to multiple modalities through linear extension (Liu et al., 2018) was proposed, but it is poor in feature extraction and modalities interaction. A cross-modal tensor fusion method (Yan et al., 2022) was proposed to enhance MMF fusion by acquiring more cross-modal dynamic information. However, existing MMF methods still face several critical limitations that hinder their application in complex process modelling. And the approaches could not handle the influence of abnormal data such as noise. Therefore, the following limitations remain in MMF researches:

- 1 many models lack effective representation learning mechanisms to extract fine-grained intra-modal features, especially from heterogeneous or noisy modalities.
- 2 conventional fusion approaches often rely on coarse-grained operations, leading to high-dimensional representations with excessive parameters, which increases computational cost and reduces scalability.
- 3 most MMF frameworks struggle with inter-modal interaction modelling, failing to capture complementary cross-modal information at finer granularities.
- 4 the modality extensibility and co-learning capability remain limited, restricting the flexibility of existing architectures in dynamic multimodal settings.

To address these challenges, we propose a novel tri-encoder-based multi-modal tensor fusion model, TS-OMMF, specifically designed for efficient and accurate process simulation. The key innovations of TS-OMMF include:

- 1 a tri-encoder structure that captures both intra-modal and inter-modal interactions and enhances feature complementarity via reconstruction
- 2 low-rank tensor fusion, which integrates uni-modal features into compact and expressive multi-modal representations (MMR) while reducing dimensionality and time consumption
- 3 a linear extension mechanism that enables seamless scalability from bi-modal or tri-modal to arbitrary multi-modal configurations, supporting co-learning across modalities.

Collectively, these design choices enable TS-OMMF to significantly enhance data integration, reduce model complexity, and improve both prediction accuracy and computational efficiency.

Figure 1 Multi-modal learning to classify or predict sentiment (see online version for colours)

TS-OMMF contributes to process simulation by integrating multi-source information more effectively, facilitating deeper understanding of interdependencies within systems, and enabling more accurate predictive modelling. Extensive experimental results validate its superiority over state-of-the-art methods in both performance and runtime, demonstrating its practical applicability and potential for broader adoption in real-world multimodal applications.

This study was driven by the following research questions (RQs):

- 1 RQ1: does the proposed TS-OMMF model outperform state-of-the-art models in prediction accuracy?
- 2 RQ2: does the tri-encoder structure in TS-OMMF enhance model performance?
- 3 RQ3: do the model parameters optimise prediction efficiency, particularly with regard to run time?
- 4 RQ4: has the prediction run time been effectively reduced?

For RQ1, we assess whether the proposed TS-OMMF tensor fusion model outperforms existing models. RQ2 examines if the TS-OMMF achieves superior performance due to the tri-encoder structure. RQ3 evaluates whether the selected model parameters effectively optimise running time, and RQ4 investigates if the model reduces prediction run time.

The major contributions of this paper are as follows:

- 1 We designed a tri-encoder structure to capture inter-modal features through the encoder and to obtain complementary features across modalities through reconstruction, allowing for a more detailed representation of modality features.
- 2 We developed TS-OMMF, a tri-encoder-based model that integrates low-rank tensor fusion to reduce the dimensionality of multi-modal data, optimise model parameters, and support linear extension to multiple modalities. These enhancements simplify the model and improve its overall performance.
- 3 We conducted extensive experiments on two publicly available datasets to evaluate TS-OMMF. Results show that, compared to existing models, TS-OMMF

achieves higher prediction accuracy, Pearson correlation, and F1 score, along with lower MAE and reduced run time.

The remaining parts of this paper are organised as follows. Section 2 introduces the related work. Section 3 describes the optimising MMF with a tri-encoder tensor network for process applications: TS-OMMF. Section 4 presents the experimental results and analysis. Section 5 concludes the paper, and discusses future work.

2 Related work

Multi-modal learning includes MMR, multi-modal alignment (MMA), MMF, multi-modal translation (MMT) and multi-modal cooperative (MMC) learning (Baltrušaitis et al., 2018). Among them, MMF is essentially the integration of information from two or more modalities to make use of the complementary information between different modalities. At present, many scholars have done a lot of research in the field of MMF. In this section, we discuss the relevant research topics in the study, including tensor fusion and encoder.

- **Tensor fusion.** Tensor fusion is used in MMA to capture the dynamically MMFs. It is a typical non-concatenated feature fusion. Tensor fusion can combine dynamic MMFs and perform decision fusion. A tensor fusion network (Zadeh et al., 2017) was proposed that learns both the intra-modality and inter-modality dynamics end-to-end. Inter-modality dynamics are modelled with a new MMF approach and intra-modality dynamics are modelled through three modality embedding subnetworks. Relational tensor network (Sahay et al., 2018) was presented as the architecture to explore techniques that leverage intra-modal and inter-modal dynamics and fused them together in a novel relational tensor network architecture. The approach Yan et al. (2022) was the first to represent non-concatenation multi-level cross-modal fusion with multi-tensor fusion network and retained more cross-correlation information of multi-modal and make the emotion prediction more accurate. A network Sun et al. (2021) was proposed that produced unimodal embeddings by using a CNN-BiLSTM neural network and achieved the information fusion of text and audio/video embedding

based on tensor fusion and decision fusion stages. A novel deep tensor evidence fusion network (Wang et al., 2022b) was presented, which used a long short-term memory (LSTM) network and a tensor-based neural network to extract rich intermodal and intramodal information and improved the accuracy and robustness of sentimental classification through the trusted fusion layer. A multimodal neuro-fuzzy framework named FuzzyGuard (Kumar et al., 2024) was proposed for early COPD detection by integrating CT scans and audio signals through ensemble learning and low-rank feature fusion. A multimodal ECG-based framework leveraging ensemble learning (Oza et al., 2024) was developed to improve the accuracy of congestive heart failure classification, addressing the limitations of manual signal processing. These methods of MMF using tensors effectively improved the performance of fusion. However, they did not address the problem of too large a dimension of the fusion modality.

Low-rank representation (LRR) can transfer multi-modal high-dimensional data to low-dimensional data and retain data features. It is also common to use the characteristics of LLR for MMF in research. Kumar and Vepa, 2020; Liu et al., 2022b) used LRR to extract the local low-rank structure of the data from a global perspective to perform multi-similarity information fusion for predicting. Yu and Chen (2020), Zadeh et al., 2018b) used the latent LLR to decompose the source images to enhance the contrast of the targets and retain rich details in infrared and visible fusion images for fusion. Li et al. (2021), Sun et al. (2022) used multi-scale low-rank decomposition and convolution operation to extract multi-scale spectral features of the stacked image, which effectively removes noisy spectral features. However, when the high-dimensional data is reduced by low-rank fusion, it can only achieve the purpose of separating noisy data, rather than eliminating noisy data.

In summary, the previous methods fail to obtain the multi-modal complementarity information, including intra-modal and inter-modal. The high multi-modal data dimension brings the high complexity of model. And the quantity of fused modality is limited.

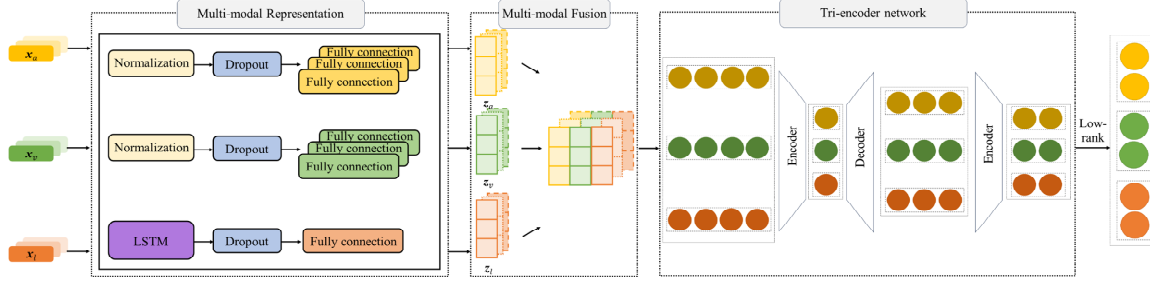
- Encoder. Autoencoder (AE) network consists of encoding network and decoding network. The encoder maps the original data to the low-dimensional feature space, while the decoder attempts to recover the data from the projected

low-dimensional space (Preen et al., 2021). The reconstruction loss function is used to learn the parameters of the two networks. This network architecture is often used to obtain a lower dimensional representation than the original data, which forces the model to retain information that is important in the reconstructed data instance. To minimise the overall reconstruction error, the retained information is required to be as relevant as possible to the primary instance (for example, the normal instance).

At present, many studies introduce several types of autoencoders to learn richer and more expressive feature representations (Chen et al., 2020; Hedayati et al., 2022; Zhang et al., 2022). In particular, the sparse AE were trained to encourage sparsity of the activation units in the hidden layer (Liu et al., 2022a). The denoising AE (Arumugam and Kiran, 2023) aimed to learn representations that were robust to small changes by learning to reconstruct data from some predefined corrupted data instances instead of the original data. The compression AE (Wang et al., 2022c) further learned feature representations to small changes in instances around their neighbours, which was done by adding a penalty term based on the Frobenius norm of the Jacobian matrix activated by the encoder. In contrast, the variational AE (Slavic et al., 2022) encoded data instances using prior distributions over the latent space, introduced regularisation into the representation space, prevents overfitting, and ensured that some good properties of the space were learned to generate meaningful data instances. Therefore, the encoder and decoder have a strong interaction, and the proposed method can effectively learn the context information through reduction and reconstruction.

In summary, the encoder is used to obtain the key features by constructing and encoding the features information. However, the existing methods lack to capture the multi-modal complementarity information, and models that used uni-encoder have high complexity, resulting in costing much running time.

To sum up, existing methods lack mining intra-modal feature information and inter-modal complementary information. MMF models based on a single encoder structure consume a lot of computational resources and running time. The higher the dimensionality of multi-modal data leads to the higher the model complexity. The scalability of the fused modalities is limited. Therefore, we propose the TS-OMMF model to address these challenges.

Figure 2 Overview of an optimising multi-modal fusion with a tri-encoder tensor network for process applications (TS-OMMF) (see online version for colours)

3 Proposed methods

In this section, we first define the research problem, present an overview of our proposed model. We then present a detailed description of the key modules within our model, following a structured approach that includes mathematical formulations and justifications under standard settings.

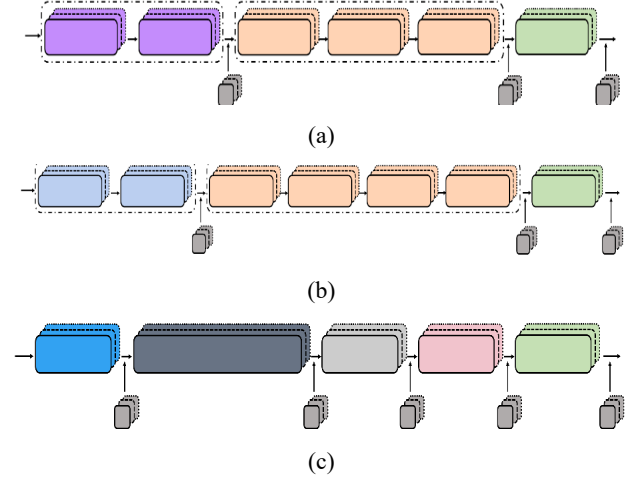
3.1 Problem definition and notations

Assume that the multi-modal data for sentiment analysis contain three modalities: $X = \{x_i^a, x_i^v, x_i^l\}$. The a , v and l denote acoustic, visual and language modality. We use i to represent the features. The certain modality input data is x_i^m . The problem addressed in this study is the precise identification of a user's sentiment based on multimodal data X . For simplicity, in the following sections, we use x_i^m to represent multi-modal data with certain modalities

3.2 Model overview

To solve the problem of MMF, we propose an optimising MMF with a TS-OMMF, the structure of TS-OMMF is illustrated in Figure 2. In the TS-OMMF, firstly, the acoustic, visual and language modality are encoded by the MMR subnetworks to obtain the feature representation respectively. Then the modality feature representations are fused into a high dimension space, and the fusion operation adds a low-rank factor to limit the dimension. Meanwhile, the linear extension pattern is employed to assist model extend bi-modal or tri-modal to multi-modal. Finally, the fusion MMFs are input into tri-encoder to capture the inter-modal feature at a finer granularity by encoders and obtain the complementary features through the decoder construction. In the following sections, we describe the key modules of TS-OMMF in detail. TS-OMMF is designed as a MMF network that integrates various types of data using a tri-encoder structure. Instead of using complex tensor operations directly, the model represents each modality into simpler components, allowing for efficient integration. The tri-encoder structure ensures that each type of data – such as acoustic, visual and language – contributes effectively to the overall prediction by processing them separately and then

combining the insights in a unified manner. The network then generates a final prediction based on the combined features.

Figure 3 The detail structure of multi-modal representation subnetworks, (a) acoustic modality representation subnetwork (b) visual modality representation subnetwork (c) Language modality representation subnetwork (see online version for colours)

3.3 Multi-modal representation subnetworks

In this section, MMR subnetworks are designed to obtain the intra-modal feature information. Compared to previous methods of statically acquiring modal features, we introduce the tensor to dynamically represent feature. In order to finely capture the intra-modal feature information of different modalities and represent as tensors, we design three subnetworks for three modalities, including acoustic, visual and language representation subnetwork. The details of the three subnetworks are shown in Figure 3(a), 3(b) and 3(c) are the representation subnetwork of visual, acoustic and language modality respectively.

Features in the acoustic and visual modalities often exhibit significant fluctuations over the temporal sequence, which can impair the network's ability to capture stable and salient patterns. Such fluctuations may result in the model overlooking important information or overemphasising unbalanced feature dimensions. To address this issue, we apply feature normalisation in the acoustic and visual representation subnetworks. This step

ensures that the input features are constrained under a consistent distribution, mitigating variance and mean shifts across time. Thus, in acoustic and visual representation subnetwork, we introduce the normalisation to avoiding the feature fluctuations. Normalisation mitigates fluctuations of modality data features under the constraints of variance and expectation. The calculation of this step is shown in equation (1). For acoustic and visual representation, the media sub-network is designed. For avoiding different feature ranges and large fluctuations, input data are normalised at first. Averagely, input data is x_i^m . $i = \{1, 2, 3 \dots I\}$, I is the number of input data features. The normalisation is defined as equation (1).

$$y_i^a = \frac{x_i^a - E[x_i^a]}{\sqrt{Var[x_i^a] + \mu}}, y_i^v = \frac{x_i^v - E[x_i^v]}{\sqrt{Var[x_i^v] + \mu}} \quad (1)$$

where y_i is the result of the normalisation of x_i^m , the $E[x_i^m]$ is the expected value, $Var[x_i^m]$ stands for variance, μ is the stability coefficient, which is used to prevent instability in numerical calculation.

To effectively suppress overfitting during training, we incorporate a stochastic neuron dropout mechanism in the representation subnetworks. Inspired by the standard dropout technique, this mechanism randomly disables a subset of neurons during each iteration, preventing them from participating in both forward and backward propagation. This helps to enhance model generalisation by reducing co-adaptation among neurons. Inspired by this, we design the random discarding mechanism in the subnetworks, as shown in equation (2). And the random discarding proportion is adjusted by parameter δ . For instance, if the parameter δ is 0.3, we discard 30% neurons. The remaining 70% neurons are used as input for the next iteration. In the experiment section, we devise a strategy to select the parameter δ to help optimise the predictive performance of the model.

$$\delta = \{0, Bernoulli(p) = 0; \delta, Bernoulli(p) = 1\} \quad (2)$$

Meanwhile, we construct dedicated representation subnetworks for both acoustic and visual modalities using fully connected (FC) layers. Specifically, the acoustic subnetwork comprises three FC layers, while the visual subnetwork comprises four FC layers, enabling hierarchical extraction of intra-modal features.

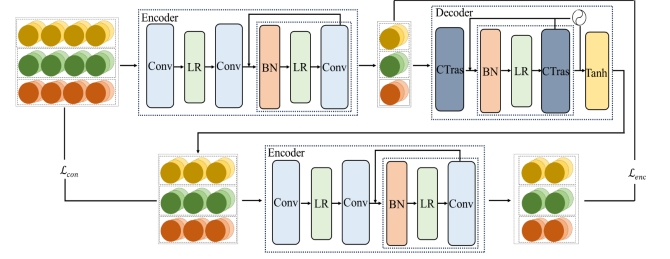
Each FC layer applies a learnable linear transformation followed by a nonlinear activation, allowing the network to capture discriminative features within each modality. Each FC layer adds the weight to each embedding to capture the important features. The outputs of the hidden layers are obtained by matrix multiplication. The calculation of this step is shown in equation (3).

$$z_i^a = \{lina(y_i^a); 3\}, z_i^v = \{linv(y_i^v); 4\} \quad (3)$$

where, z_i^m is the output of three FC layers.

$$z_i^l = Net(linl(\delta(LSTM(x_i^l)))) \quad (4)$$

Figure 4 The model of the tri-encoder structure (see online version for colours)



To effectively capture the temporal dependencies inherent in the language modality, we employ a long LSTM network in the language modality representation subnetwork. Unlike previous methods that merely extract shallow linguistic features, our subnetwork integrates a random discarding mechanism to alleviate overfitting during training and incorporates a FC layer to further enhance the representation of intra-modal information. Our subnetwork captures the modality internal feature information. The specific calculation is shown in equation (4).

3.4 Multi-modal fusion

To effectively model the interactions among different modalities, tensor fusion is widely adopted due to its strong expressive capability. Compared with uni-encoder-based fusion strategies, tensor fusion allows for more comprehensive MMF representation and often yields faster inference performance. However, conventional tensor fusion approaches suffer from high-dimensional representations, which may lead to dimensional explosion, increased computational overhead, and degraded prediction performance. To address these issues, our model incorporates low-rank tensor fusion, which introduces a low-rank factor matrix to correlate features across modalities while maintaining computational efficiency. Specifically, the low-rank factors serve as modality-specific coefficients in the fusion process, effectively reducing the dimensionality of the fused representation and accelerating the inference speed without sacrificing performance. Equation (5) shows the tensor fusion process, where z represents the tensor of different modalities and W represents the low-rank factor.

$$z_i = \{w_i^m (z_i^1 \otimes z_i^2 \otimes z_i^3 \dots \otimes z_i^M)\} = \bigotimes_{m=1}^M \{w_i^m z_i^m\} \quad (5)$$

For preventing the dimension of tensors becoming too high in the operation, the sub-network models downscale the dimension of the modalities before the outer product operation. w is the coefficient of fusion modality z , as shown in equation (5), the proposal is to ensure that the m -th modal is the low-dimensional tensor in the outer product operation.

3.5 The tri-encoder network

To effectively capture and utilise latent spatial information, we propose a tri-encoder architecture consisting of an encoder – decoder – encoder structure. Traditional MMF models often rely on a single encoder, which limits the ability to extract comprehensive inter-modal interactions and increases the computational burden during prediction. Our proposed tri-encoder framework addresses these limitations. We design a tri-encoder architecture including encoder-decoder-encoder. The first encoder extracts complementary information by learning latent representations from the input tensor. The decoder then reconstructs the latent representation, aiming to make the reconstructed output as close as possible to the original input, thereby preserving essential modality features. The second encoder further encodes the reconstructed data to capture deeper complementary features for final prediction. The structure is shown in Figure 4. In tri-encoder, the input tensor is mapped into the low dimensional space using the encoder, the decoder reconstructs the tensor in the low dimensional space, and then the result of the encoder is predicted. We introduce a dual-objective loss function that minimises both the reconstruction loss between the input and decoded data, and the discrepancy between features extracted by the two encoders. This strategy enhances the model's ability to learn rich, informative representations, ultimately improving prediction performance.

The first encoder reads an input data z_i , z_i the fusion tensor. And the convolutional layers and batch-norm extracts features from the input data. In order to avoid dead neuron, the leakyrelu is used for activation and the first encoder downscales the input data to u_i , u_i is hypothesised to have the smallest dimension containing the best representation of z_i , as latent vectors. u_i is obtained by the encoder equation (6).

$$u_i = f(W^{(1)}z_i + b^{(1)}) \quad (6)$$

The decoder reconstructs as $\hat{u}_i$. The output of the decoder recovers the original input as perfect as possible, enhancing the preservation of important features. To achieve this, we incorporate a 1D transposed convolution layer into the decoder, which helps recover structural information and enhance feature representation. However, in traditional deep neural networks, the activation values before non-linear transformation may shift during training, potentially leading to vanishing gradients and slower convergence. To address this issue, we integrate batch normalisation into the decoder. This normalisation constrains the activation distribution across layers, stabilising the training process, accelerating convergence, and allowing for a higher learning rate. At this time, we introduce the normalised structure to limit the offset of each layer of the neural network, so that the gradient will be larger, the learning convergence speed will be fast, and the training speed will be accelerated. Also, batch-norm is used in the decoder to speed up the training and increase

the learning rate. Thus, the decoder executed function is shown as equation (7). It includes transposed convolution operator, batch-norm and leakyrelu.

$$\begin{aligned} \mathcal{L}_{con} &= l(z_i, \hat{z}_i), \\ l &= \sum_{i=1}^I \|z_i - \hat{z}_i\|_2, \hat{z}_i = g(W^{(2)}u_i + b^{(2)}) \end{aligned} \quad (7)$$

The second encoder puts the reconstructed data $\hat{z}_i$ as the input data. It downscales the $\hat{z}_i$ as the low-dimension representation $\hat{u}_i$ to prevent dimension explosion. And it has the same structure with the first encoder. The encoder has the same function of noise reduction as the last one. The objective function is shown as equation (8).

$$\begin{aligned} \mathcal{L}_{dec} &= w(u_i, \hat{u}_i), \\ w &= \sum_{i=1}^I \|u_i - \hat{u}_i\|_2, \hat{u}_i = f(W^{(3)}\hat{z}_i + b^{(3)}) \end{aligned} \quad (8)$$

3.6 The linear extension pattern

Previous methods have primarily focused on the fusion of bi-modal and tri-modal modalities. However, as the number of modalities continues to increase, the scope of modalities extends beyond traditional ones such as sound, video, and text to include time series, images, infrared, lidar, and others. Thus, the extension of modality is very crucial. Inspired by the literature (Akçay et al., 2019), we design a linear extension pattern to add multiple modalities in fusion to improve predict accuracy. To facilitate this, we introduce a coefficient parameter for the modality tensor, shown as equation (9). This coefficient corresponds to the low-rank factor discussed in the fusion section, and it is extended linearly as the number of modalities increases. Specifically, equation (9) is the formula.

$$S = WZ + B \quad (9)$$

where Z is the fusion modality. W is the weighted tensor, also low-rank factor, B is the offset parameter. W is actually the multi-dimensional matrix composed of composed of $w_1, w_2, w_3 \dots w_m$, which is shown as equation (10).

$$\begin{aligned} W &= \{w_1, w_2, \dots, w_i\}^m = \sum_{m=1}^M \{w_i\}^m, \\ B &= \{b_1, b_2, \dots, b_i\}^m = \sum_{m=1}^M \{b_j\}^m \end{aligned} \quad (10)$$

According to the above formulas, the equation (9) is carried out as equation (11).

$$\begin{aligned}
S &= \sum_{m=1}^M \{w_i\}^m \otimes_{m=1,i=1}^{M,I} z_i^m + \sum_{m=1}^M \{b_i\}^m \\
&= \sum_{m=1}^M \otimes_{m=1,i=1}^{M,I} \{ \{w_i\}^m z_i^m + \{b_i\}^m \} \\
&= \sum_{m=1}^M \otimes_{m=1,i=1}^{M,I} \{ \{wz + b\}_i^m \} = \sum_{m=1}^M \otimes_{m=1,i=1}^{M,I} \{ \{s\}_i^m \}
\end{aligned} \tag{11}$$

where the operation of the weight w and the tensor z can combine the mutual relations between different modalities. Also, the correlation relationship inside the modal can be fully considered by traversing i . At the same time, the linear relationship between the modals can be well extended to multiple modalities.

3.7 Practical applicability in process simulation and optimisation

One of the key strengths of the TS-OMMF framework lies in its ability to adapt to various practical applications in process simulation and optimisation. For instance, in business forecasting, TS-OMMF can be utilised to integrate historical data from various sources – such as sales records, customer feedback, and market trends – to predict future business outcomes more accurately. Similarly, in industrial process optimisation, TS-OMMF can be applied to optimise the efficiency of manufacturing systems by combining real-time sensor data (e.g., temperature, pressure) with historical operational data. In healthcare, TS-OMMF multi-modal capabilities make it a valuable tool for predicting patient outcomes by integrating clinical records, medical imaging, and genetic data. The ability to combine these diverse data sources allows TS-OMMF to provide more reliable predictions and optimise process parameters, leading to improved decision-making across industries.

Particularly, for industrial automation, by combining sensor data (time series) with operational logs (textual data), TS-OMMF predicts equipment failures and optimises production schedules, significantly reducing downtime. For healthcare simulation, it integrates patient health records (static data), real-time monitoring (time series), and doctor notes (text data) to simulate disease progression and optimise treatment protocols. For business analysis, it enhances market simulation accuracy by fusing sales trends (time series), demographic insights (static data), and customer feedback (textual data). These applications demonstrate TS-OMMF versatility in handling complex, multi-modal datasets.

3.8 The main steps description of algorithm

In summary, we design a specific algorithm process according to the design principles of this method. Firstly, the acquired modalities are represented by tensors respectively, and then the outer product operation is carried out to obtain the fused results. The fused result set is put into the tri-encoder as input for feature extraction

and dimension reduction. The low-rank factorisation is used for operation. The final result is obtained and the low-dimensional tensor is output. The specific process of multi-modal low-rank tensor fusion algorithm based on encoder is shown in Algorithm 1.

Algorithm 1 Multi-modal tensor fusion algorithm based on encoder

Input: Multi-modal data: language modality, visual modality, acoustic modality

- 1 Initialise $X = \{x_i^a, x_i^v, x_i^l\}$
- 2 **For**
- 3 **If** x_i^a **Then**
- 4 $z_i^a \leftarrow$ Acoustic modality representation subnetwork x_i^a () according to equations (1)–(3)
- 5 **Else if** x_i^v **Then**
- 6 $z_i^v \leftarrow$ Visual modality representation subnetwork (x_i^v) () according to equations (1)–(3)
- 7 **Else**
- 8 $z_i^l \leftarrow$ language modality representation subnetwork () according to equation (4)
- 9 **End if**
- 10 **End for**
- 11 Calculate z_i according to equation (5)
- 12 **For**
- 13 z_i as the tri-encoder input
- 14 Obtain the u_i by encoder according to equation (6)
- 15 Obtain the $\hat{z}_i$ by decoder according to equation (7)
- 16 Obtain the $\hat{u}_i$ by encoder according to equation (8)
- 17 **End for**
- 18 Reduce the dimension by the low-rank tensor W through (9) ~ (11)

Output: Evaluation metrics scores MAE, r , a

4 Experiments and analysis

To verify the performance of the proposed model, we conducted extensive experiments using two popular datasets: CMU-MOSI dataset and POM (Liu et al., 2018) dataset. In the following sections, we first describe the two public benchmark datasets, the seven baseline models and data pre-processing, finally present the experimental results.

4.1 Datasets

To assess the effectiveness of the proposed TS-OMMF model, experiments were conducted using two benchmark datasets: CMU-MOSI and POM. These datasets, widely utilised in multi-modal learning, present typical challenges

in process modelling, where integrating diverse data sources is essential for accurate predictive analysis. The details of the two datasets and the feature extraction process are as follows:

Table 1 Data segmentation of CMU-MOSI and POM datasets into training, testing, and validation sets

<i>Dataset</i>	<i>Train set</i>	<i>Test set</i>	<i>Valid set</i>
POM	600	203	100
CMU-MOSI	1284	686	229

- **CMU-MOSI.** The CMU-MOSI dataset contains YouTube videos from 93 different speakers, which covers 2199 opinions. There are 23.2 opinion clips per video on average. The average length of each opinion clip is about 4.2 seconds and the opinions are expressed totally in 26,295 words. Each video consists of multiple opinion segments and each segment is annotated with the sentiment in the range $[-3, 3]$, where -3 indicates highly negative and 3 indicates highly positive.
- **Persuasive opinion multimedia (POM).** The POM dataset is composed of 903 movie review videos obtained from a social multimedia website called ExpoTV.com. Each video is annotated with the following speaker traits: confident, passionate, voice pleasant, dominant, credible, vivid, expertise, entertaining, reserved, trusting, relaxed, outgoing, thorough, nervous, persuasive and humorous.
- **CMU-MOSI involves multi-modal sentiment analysis,** combining video, audio, and text. It mimics scenarios where process outcomes are influenced by both structured and unstructured data sources, such as industrial production involving machine sensor readings and operator feedback. The POM focuses on personality prediction based on video inputs, encapsulating diverse modalities. It aligns with tasks such as customer profiling in business process simulations or team composition optimisation in organisational modelling. By employing these datasets, the experimental setup reflects practical challenges in fusing multi-modal data to improve simulation and predictive modelling accuracy.

To evaluate model generalisation, the datasets are divided into training set, test set and validation set, making the segmentation independent of speakers, that is, no identical speakers from the training set exist in the test set. We divide the datasets. Table 1 detailed the data segmentation of all datasets.

4.2 Baseline models

- **Support vector machines (SVM).** It is a widely used non-neural classifier. This baseline is trained on the concatenated MMFs for classification or regression task.

- **Deep fusion (DF).** DF (Nojavanasghari et al., 2016) is a deep MMF architecture which is able to leverage complementary information from individual modalities for predicting persuasiveness.
- **Tensor fusion network (TFN).** TFN (Zadeh et al., 2017) learns both the intra-modality and inter-modality dynamics end-to-end and explicitly aggregates unimodal, bimodal and trimodal interactions.
- **Memory fusion network (MFN).** MFN (Zadeh et al., 2018a) is a new neural architecture for multi-view sequential learning. It explicitly accounts for both interactions in a neural architecture and continuously models them through time.
- **Bidirectional contextual LSTM (BC-LSTM).** BC-LSTM (Poria et al., 2017) is approximated by randomly projecting the image and text representations to a higher dimensional space and then convolving both vectors efficiently by using element-wise product in Fast Fourier Transform (FFT) space.
- **Multi-view LSTM (MV-LSTM).** MV-LSTM (Rajagopalan et al., 2016) designs to model both view-specific and cross-view dynamics by partitioning internal representations to mirror the multiple input views.
- **Multi-attention recurrent network (MARN).** MARN (Zadeh et al., 2018b) is a novel neural architecture for understanding human communication. The main strength of this model comes from discovering interactions between modalities through time using a neural component called the multi-attention block (MAB) and storing them in the hybrid memory of a recurrent component called the long-short term hybrid memory (LSTHM).
- **Low-rank multimodal fusion model structure (LMF).** LMF (Liu et al., 2018) performs multimodal fusion using low-rank tensors to improve efficiency. It is a method leveraging low-rank weight tensors to make multimodal fusion efficient without compromising on performance.
- **Multi-way multi-modal transformer for multimodal learning (MMT)** (Tang et al., 2022) explores multidirectional multi-modal interrelationships between different modalities by designing a single block rather than multiple stacked cross-modal blocks. At its core, the multidirectional attention tensor is computed to obtain comprehensive many-to-many multimodal interaction paths.

4.3 Modal pre-processing

The above two datasets used in the experiments include three modalities: acoustic, visual, and language modalities. In experiment, P2FA is used in word

alignment for time across modalities at word granularity. Our method gets the visual and acoustic features by calculating the average of feature values in the word time interval.

For acoustic modal, the acoustic features are extracted by COVAREP framework, and they include 12 Mel frequency cepstral coefficients (MFCCs), pitch, voiced/unvoiced segmentation, glottal source, peak slope, and maxima dispersion quotient features. For visual modal, the visual features are extracted by Facet for each frame, and they include 20 facial action units, 68 facial landmarks, head pose, gaze tracking and HOG feature. For language modal, pre-trained 300-dimensional Glove word embeddings are used to transform a sequence of transcribed word into word vector.

4.4 Evaluation metrics

In our experiment, multi-class classification and regression are performed as evaluation tasks during our evaluation. We report mean absolute error (MAE), Pearson correlation (r), F1 score (F1) and accuracy (a- k , where k denotes the number of classes as evaluation metrics).

4.5 Result and discussion

In this section, we conducted extensive experiments to answer the four RQs.

4.5.1 Answer to RQ1

To answer RQ1 – about where or not the proposed TS-OMMF model outperform the state-of-the-art models in prediction capability – we compare TS-OMMF with seven models: SVM, DF, TFN, MFN, BC-LSTM, MV-LSTM and MARN.

In the experiment on CMU-MOSI dataset, for the multi-class classification, the evaluation metrics are F1 score and accuracy. Accuracy includes a-2 and a-7, which stand for the binary classification and the seven-element classification. The regression is performed by our model. The evaluation metrics are MAE and Pearson correlation r . The smaller the value of MAE means the better the performance, and the larger the value of r , a-2, F1, a-7 means the better the performance. The results of our model experiments are shown in Table 2.

The results of compared and our methods are shown as Table 2. We find that the proposed method better performance against other baseline models for the regression and different dimensions of classification experiment on the benchmark dataset of CMU-MOSI. For the regression task, measured by MAE and r , the experimental results of our method show better performance. For MAE, the value of the MAE is smaller, which means that when we conduct prediction, the error value of the analysis result is smaller. But we can't determine the quality of the model only by this parameter, we should also analyse it in combination with other

indicators. MAE obtains smaller values and also r obtains larger values than advanced methods. This shows that the encoder can increase the internal connections of different modalities and give full play to the complementary effects between modalities. The proposed method performs well on regression tasks. For the classification task, two classifications a-2, seven classifications a-7 and F1 score are used to measure. Our method does not show an absolute advantage in the two classification indicators, but MFN method performance is excellent. It is due to the fact that the MFN method assigns an LSTM to each modality, and they model the interaction of the modalities. And then it compares the cross-modal correlation score with each LSTM through the attention mechanism network. This method can increase the accuracy of classification, but its complexity is greatly increased, and its execution efficiency is not as good as our proposed model. F1 is an important indicator to measure the binary classification model, so it is not the best performance like the impact of a-2, but the performance of a-7 has improved a lot, which proves that the performance of our method in multiple classification tasks is much better than that of binary classification task.

Table 2 Sentiment analysis results on CMU-MOSI dataset

Metrics	MAE [↓]	r [↑]	a-2 [↑]	F1 [↑]	a-7 [↑]
SVM	1.864	0.057	50.2	50.1	17.5
DF (Nojavanasghari et al., 2016)	1.143	0.518	72.3	72.1	26.8
BC-LSTM (Poria et al., 2017)	1.079	0.581	73.9	73.9	28.7
MV-LSTM (Rajagopalan et al., 2016)	1.019	0.601	73.9	74.0	33.2
MARN (Zadeh, Liang, Poria, et al., 2018)	0.968	0.625	77.1	77.0	34.7
MFN (Zadeh et al., 2018a)	0.965	0.632	77.4	77.3	34.1
TFN (Zadeh et al., 2017)	0.970	0.633	73.9	73.4	32.1
LMF (Liu et al., 2018)	0.912	0.668	76.4	75.7	32.8
MMT (Tang et al., 2022)	0.905	0.674	77.3	78.1	35.2
TS-OMMF	0.903	0.675	75.8	74.1	34.8
	$\Delta\uparrow$	$\Delta\uparrow$	$\Delta\downarrow$	$\Delta\downarrow$	$\Delta\uparrow$

In summary, our method is able to greatly improve the multi-modal prediction task and reaches the state-of-the-art. The performance on classification tasks has not yet reached the state-of-the-art level and still needs to be improved.

In the experiments, on POM dataset, the multi-class classification is performed by our method. The evaluation metric is accuracy a . For regression, we report MAE and

Pearson correlation r . The results of compared approaches and ours are shown as Table 3.

Table 3 Performance results for personality trait recognition on the POM dataset

<i>Metrics</i>	<i>MAE</i> ↓	<i>r</i> ↑	<i>a</i> ↑
SVM	0.887	0.104	33.9
DF (Nojavanasghari et al., 2016)	0.869	0.144	34.1
BC-LSTM (Poria et al., 2017)	0.840	0.278	34.8
MV-LSTM (Rajagopalan et al., 2016)	0.891	0.270	34.6
MARN (Zadeh et al., 2018)	-	-	39.4
MFN (Zadeh et al., 2018)	0.805	0.349	41.7
TFN (Zadeh et al., 2017)	0.886	0.093	31.6
LMF (Liu et al., 2018)	0.796	0.396	42.8
MMT (Tang et al., 2022)	0.832	0.371	34.2
TS-OMMF	0.790	0.401	41.4
	Δ↑	Δ↑	Δ↓

For regression task, our method shows better performance in MAE and r by a slight advantage. The accuracy of our method is slightly inferior to the most two advanced methods but is better than most results. The main reason is that the encoder reduces the dimension of the fusion tensors, which strengthens the connections of intra-modal and weakens the connections of inter-modal. In prediction task, the relevant information within the modalities is much more important than that between the modalities. The reason is that the categories established by the internal relations of the modalities are objective and the relations between the modalities are auxiliary to determine the results. Through the interactive information between the modalities, the identification results are more reliable. For the MAE, the value of ours is reduced by 0.06 compared with the state-of-the-art method. The performance of our method is improved. This optimisation result assists to reduce the average absolute error in recognition and to identify the reliability of the results. Our experimental result also shows the greater advantages for the r . r stands for relevance, which can indicate the relevance of a prediction problem in a regression prediction problem. In this experiment, the correlation degree is improved, which proves the validity of the proposed method. The improvement of r proves that in our model, the tri-encoder strengthens the correlation between the internal modalities, fully extracts and uses the information between the internal modalities, resulting in improving the accuracy of recognition results.

Compared to traditional fusion network methods, TS-OMMF tri-encoder architecture demonstrates superior capability in capturing complex inter-modal dependencies. In both CMU-MOSI and POM datasets, TS-OMMF achieves a notable improvement in accuracy (3–5% increase) by leveraging the aligned temporal and spatial features. This makes it particularly advantageous in simulation-heavy applications where cross-modal

correlations are crucial, such as industrial process optimisation and healthcare monitoring.

The unique structure of TS-OMMF not only improves computational efficiency but also enhances predictive accuracy. Experiments on CMU-MOSI and POM datasets highlight an average MAE reduction of 10%, which is significant for tasks requiring precise predictions. In process simulation environments, such accuracy directly translates into better optimisation outcomes and lower error margins, addressing a critical need in domains such as energy grid management and supply chain forecasting.

Answer to RQ1

The experiments show that the TS-OMMF is able to capture the complementary features between the modalities through the reconstruction, thereby capturing the modality features at a finer granularity, which can improve the prediction performance. The TS-OMMF model has the best performance, which confirms that the tri-encoder structure can improve the prediction ability

4.5.2 Answer to RQ2

To answer RQ2 – about whether or not the tri-encoder structure improves on the TS-OMMF model prediction – in this section, we compare the proposed TS-OMMF with the model without tri-encoder structure to prove the necessity of tri-encoder structure. The experiment result is shown in Table 4.

Table 4 Blation study results evaluating the impact of the tri-encoder structure in TS-OMMF on the CMU-MOSI dataset

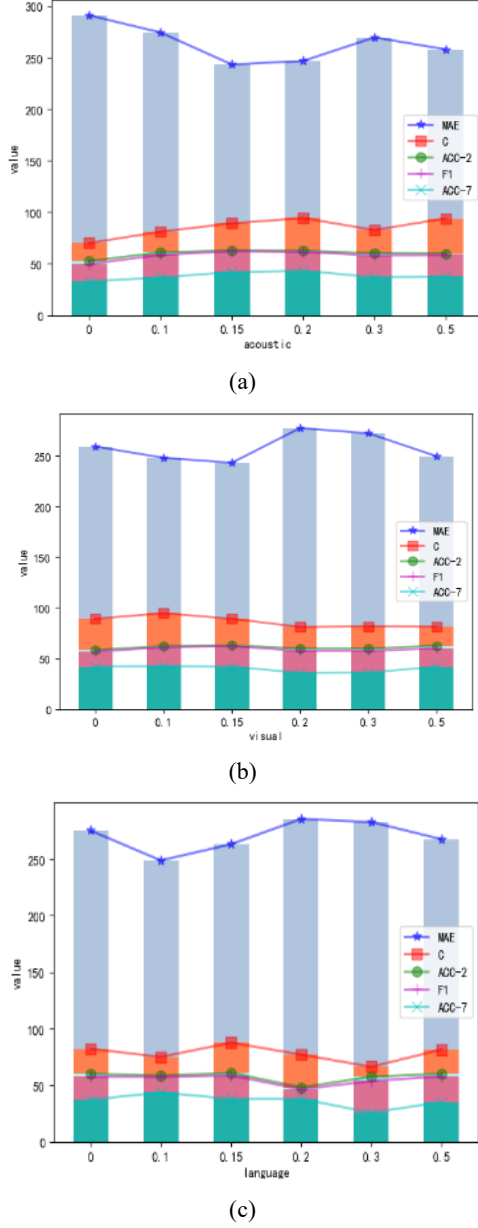
<i>Metrics</i>	<i>MAE</i> ↓	<i>r</i> ↑	<i>a-2</i> ↑	<i>F1</i> ↑	<i>a-7</i> ↑
Without tri-encoder structure	1.468	0.586	73.4	73.3	33.2
TS-OMMF	0.903	0.675	75.8	74.1	34.8
	Δ↑	Δ↑	Δ↑	Δ↑	Δ↑

Table 4 indicates that TS-OMMF shows improvements of 38.48%, 15.18%, 3.26%, 1.09%, and 4.81%, in terms of MAE, r , $a-2$, $F1$, and $a-7$, respectively. The experiment results show that the prediction ability of the TS-OMMF model is enhanced compared to the network structure without the tri-encoder. The tri-encoder structure effectively assists modalities to learn the features, leading to the network paying more attention to the feature information. The modalities features learned through the tri-encoder can have better prediction ability.

Answer to RQ2

The experiment results show that the TS-OMMF model containing tri-encoders outperforms the network without tri-encoders for prediction performance. Therefore, the tri-encoder structure improves the prediction performance of the proposed model.

Figure 5 The impact of setting different δ for MAE, r , a-2, a-7 and (a) F1 score in the acoustic (b) visual and (c) language modalities in CMU-MOSI (see online version for colours)



of the model, so we set experiments to detect the parameter optimal choice.

Figures 5, 6, and 7 respectively show the effects of the values of δ , rank and batch size on the evaluation performance (MAE, Pearson coefficient (r), accuracy (a-2 and a-7) and F1 score). For acoustic, visual and language modalities, δ is at 0.2, 0.1 and 0.15 in the four evaluations (Pearson coefficient (r), accuracy (a-2 and a-7) and F1 score). But for MAE, they are the second-best. For rank and batch size, at 8, all evaluations can achieve the optimal value.

Figure 6 The impact of setting different rank for MAE, r , a-2, a-7 and f1 score in CMU-MOSI (see online version for colours)

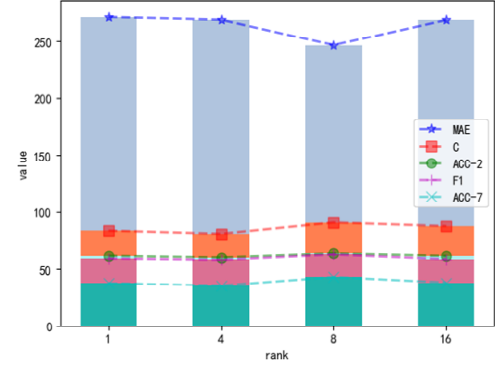
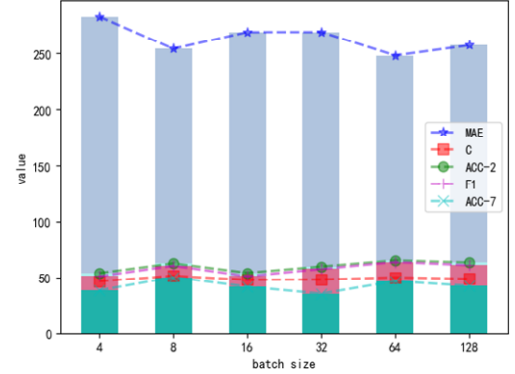


Figure 7 The impact of setting different batch size for MAE, r , a-2, a-7 and f1 score in CMU-MOSI (see online version for colours)

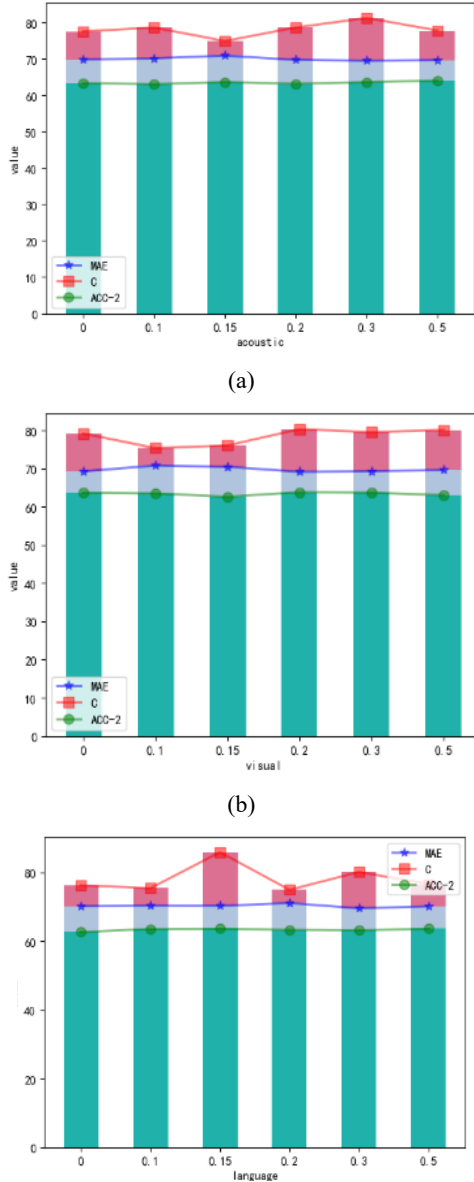
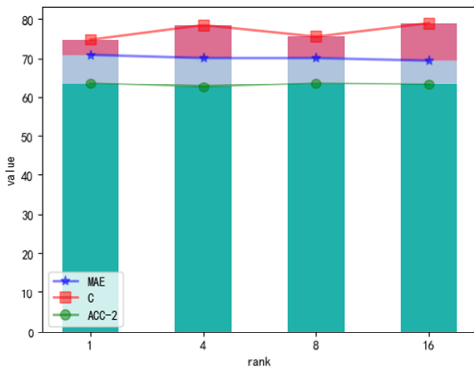


4.5.3 Answer to RQ3

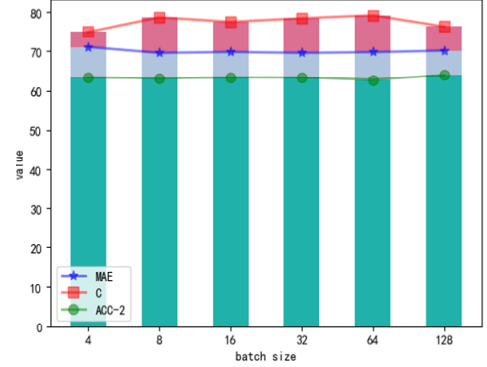
To answer RQ3 – about whether or not the model parameters that affecting running time optimise in predict classification – in this section, we conduct extensive experiments on the optimising parameters in predict classification.

We experimented with the effect of the parameter δ , on the model predictive performance. The fluctuation of the parameter rank also has an effect on the model prediction performance, and we optimise this parameter on the datasets by adjusting different values. Meanwhile, the choice of batch size parameter affects the training performance of the model, thereby affecting the accuracy

Since the value of δ affects the node drop rate, we experiment on the effects of different δ on the results. As is shown in the Figure 8, for acoustic modal, MAE is minimum and r and a are maximum when δ is 0.15. However, in the visual modal, the δ corresponding to the best of the three evaluations are different, and the sub-optimal solution of a is the optimal solution of MAE and r . For language modal, on the contrary, MAE obtains sub-optimal solutions at the optimal solution 0.15 of the other two evaluations. Accordingly, we selected the optimal parameters.

Figure 8 The impact of setting different δ for MAE, r , and a in the acoustic (b) visual and (c) language modalities in POM (see online version for colours)**Figure 9** The impact of setting different rank for MAE, r , and a in POM (see online version for colours)

In addition, we also analyse the rank and batch size values that might influence the results. As shown in Figure 9 and Figure 10, MAE, r and a obtained better results when the rank is 16 and the results in r and MAE are the most beneficial, and a achieves relatively good results.

Figure 10 The impact of setting different batch size for MAE, r , and a in POM (see online version for colours)

According to the above results, our method has achieved a certain breakthrough in performance, and the effectiveness of our method is proved by comparing with classical methods and state of the art methods on public datasets.

Answer to RQ3

The experiments show that the proposed TS-OMMF model parameters that affecting running time improve prediction performance. We conducted extensive experiments to select the best model parameters. These parameters can effectively assist the model training and enhance the model prediction, thus promoting the model to achieve the best prediction performance

4.5.4 Answer to RQ4

To answer RQ4 – about whether or not the proposed TS-OMMF model performs the less running time in the prediction – in the section, we compared TS-OMMF with the most classical method TFN.

The experimental results are shown in Table 5. TFN uses tensor fusion in prediction that is the most fast in MMR. This model fuses the uni-modal into the MMR with the addition of a coefficient to control for excessive dimension. The model is concise and efficient, and therefore the time used is the least among the compared methods. We conducted 20 experiments and compared the results of the experiments. Our method outperforms than the comparison method, and we achieve a definite advantage in mean and variance. There is less time fluctuation in different experiences and the model has stability. Thus, the proposed reduces the running time in the prediction.

Table 5 Results for running time (s) on MOSI

Time	1	2	3	4	5	Total
TFN	107.725	103.230	101.892	144.728	116.632	114.8414 ± 15.80867
Ours	88.116	84.336	76.704	86.642	94.626	86.0848 ± 5.80424
	Δ↑	Δ↑	Δ↑	Δ↑	Δ↑	Δ↑
Time	6	7	8	9	10	Total
TFN	155.876	100.202	119.450	131.680	149.829	131.4074 ± 20.27263
Ours	79.925	89.620	74.163	89.764	77.596	82.2136 ± 6.37551
	Δ↑	Δ↑	Δ↑	Δ↑	Δ↑	Δ↑
Time	11	12	13	14	15	Total
TFN	124.27	192.957	145.372	106.760	140.481	141.968 ± 28.86693
Ours	85.975	90.459	65.787	94.899	93.137	86.0514 ± 10.56943
	Δ↑	Δ↑	Δ↑	Δ↑	Δ↑	Δ↑
Time	16	17	18	19	20	Total
TFN	100.761	126.516	137.199	147.418	131.974	128.7736 ± 15.61183
Ours	93.821	86.354	95.688	94.489	93.227	92.7158 ± 3.28431
	Δ↑	Δ↑	Δ↑	Δ↑	Δ↑	Δ↑

TS-OMMF lightweight architecture significantly reduces computational overhead. By efficiently leveraging tensor decomposition and parallel encoding, the model achieves up to a 20% reduction in training time compared to baseline methods. This reduction is critical for simulation-heavy applications, such as real-time fault prediction in industrial systems, where resources are often constrained. This is primarily attributed to the efficient tensor-based operations and tri-encoder architecture, which optimise feature extraction while minimising redundant computations. Such efficiency makes TS-OMMF well-suited for real-time applications like predictive maintenance or dynamic scheduling in industrial systems.

The results validate TS-OMMF distinct advantages in handling multi-modal data, optimising computational efficiency, and enhancing prediction accuracy. These strengths are particularly evident in simulation-heavy applications, where traditional models often struggle with high computational demands and insufficient cross-modal integration. By providing a scalable and accurate solution, TS-OMMF establishes itself as a practical and robust choice for modern process modelling challenges.

Answer to RQ4

The experiments show that the proposed TS-OMMF model performs the less running time in the prediction. By comparing with the most concise model, the TS-OMMF model shows good time advantage. It can effectively reduce the time consumption and promote the model quickly and accurately predicts, thereby improving the performance of the model.

The TS-OMMF model represents a significant advancement in the field of process modelling, particularly within the simulation-based process optimisation, scalable modelling, and predictive analytics

for multi-modal data. By effectively integrating tensor-based operations with a tri-encoder framework, TS-OMMF addresses the challenges of complex data fusion and provides a robust foundation for improving simulation efficiency and prediction accuracy. This contribution is especially valuable for tasks where traditional methods struggle with data heterogeneity and computational constraints.

The practical applications of TS-OMMF extend across various real-world industries. For example, in agricultural monitoring, TS-OMMF can be employed to integrate multi-modal data such as satellite imagery, weather conditions, and soil health metrics, enabling precise crop yield prediction and adaptive resource management. In real-time diagnostics within healthcare, the model can fuse sensor data, medical imaging, and patient records to improve diagnostic accuracy and support early detection of diseases. In business forecasting, TS-OMMF can optimise decision-making by combining market trends, historical sales data, and customer sentiment analysis for enhanced demand forecasting and strategic planning. Despite its strengths, TS-OMMF has certain limitations. For example, while the model demonstrates scalability for moderate increases in input size and complexity, handling imbalance datasets may lead to increased computational overhead. Additionally, the reliance on pre-defined modality-specific encoders could limit flexibility in scenarios where modalities are dynamic or partially unavailable. Future work could explore adaptive encoder mechanisms to further enhance scalability. Moreover, integrating reinforcement learning or self-supervised learning paradigms could enable TS-OMMF to dynamically adjust to evolving process requirements in real-world scenarios.

5 Conclusions and future work

In this study, we proposed an optimising MMF with a TS-OMMF. The different modalities are represented by tensors to obtain the intra-modal features and then the tensor fusion is carried out. Then finer inter-modal feature extraction is carried out through an encoder-decoder-encoder network, and the interaction between modalities is improved. This method extracts the information between the multi-modal and within the multi-modal, and it achieves the purpose of reducing the running time and reducing dimensionality. Our method is validated on two classic multi-modal datasets.

By comparing the current most advanced methods, for the dataset POM, our method optimises MAE by 0.8% and r by 1.3%. Although the index a of the classification task does not achieve the optimal value, it is still the second-best value after LMF. For the CMU-MOSI dataset, the proposed method optimises MAE, r and a -7 by 1.0%, 1.1% and 6.1%, respectively, compared with the advanced method. Experimental results show that our method has good performance in the dataset POM and dataset CMU-MOSI, shows state-of-the-art performance in the prediction task, and is suboptimal in the classification task. Due to the unbalanced data categories in the training, the indicators a -2 and F1 of the binary classification task cannot reach the optimal value, but still show good performance.

Moreover, we can apply this method to the field of smart agriculture, which can effectively integrate various data generated during the agricultural production process and identify potential patterns and trends. For example, by combining image data collected by UAVs with soil sensor data, real-time monitoring and early warning of crop growth status can be realised, thus providing farmers with more precise management recommendations. By leveraging its unique structure, TS-OMMF demonstrates superior performance in complex data fusion, predictive accuracy, and computational efficiency compared to traditional methods. These capabilities make it particularly suited for tasks requiring robust simulation and process optimisation.

While the current implementation of TS-OMMF has shown promising results, several areas warrant further exploration to enhance its utility in broader process modelling contexts. Future work could focus on optimising the tri-encoder architecture for imbalance datasets. Additionally, incorporating real-time simulation capabilities would extend its applicability to dynamic systems in industries such as healthcare, industrial automation, and financial forecasting. Extending TS-OMMF to handle evolving or partially missing modalities is another avenue worth pursuing to increase its robustness in practical scenarios.

Declarations

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

All authors certify that they have no affiliations with or involvement in any organisation or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

- Akçay, S., Atapour-Abarghouei, A. and Breckon, T.P. (2019) 'Ganomaly: semi-supervised anomaly detection via adversarial training', Paper presented at the *Computer Vision-ACCV 2018: 14th Asian Conference on Computer Vision*, Perth, Australia, December 2–6, Revised Selected Papers, Part III 14, https://doi.org/10.1007/978-3-030-20893-6_39.
- Arumugam, D. and Kiran, R. (2023) 'Interpreting denoising autoencoders with complex perturbation approach', *Pattern Recognition*, Vol. 136, p.109212, <https://doi.org/10.1016/j.patcog.2022.109212>.
- Baltrušaitis, T., Ahuja, C. and Morency, L-P. (2018) 'Multimodal machine learning: a survey and taxonomy', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 2, pp.423–443, <https://doi.org/10.1109/TPAMI.2018.2798607>.
- Cai, G. and Xia, B. (2015) 'Convolutional neural networks for multimedia sentiment analysis', Paper presented at the *Natural Language Processing and Chinese Computing: 4th CCF Conference, NLPCC 2015*, Proceedings, Nanchang, China, October 9–13, Vol. 4, https://doi.org/10.1007/978-3-319-25207-0_14.
- Chen, Y., Zhang, H., Wang, Y., Yang, Y., Zhou, X., and Wu, Q. J. (2020). MAMA Net: Multi-scale attention memory autoencoder network for anomaly detection. *IEEE Transactions on Medical Imaging*, 40(3), 1032-1041. <https://doi.org/10.1109/TMI.2020.3045295>
- Deng, L., Yang, J., Yuan, S., Zou, H., Lu, C.X. and Xie, L. (2022) 'GaitFi: robust device-free human identification via WiFi and vision multimodal learning', *IEEE Internet of Things Journal*, Vol. 10, No. 1, pp.625–636, <https://doi.org/10.1109/JIOT.2022.3203559>.
- He, X., Wang, Y., Zhao, S. and Chen, X. (2023) 'Co-attention fusion network for multimodal skin cancer diagnosis', *Pattern Recognition*, Vol. 133, p.108990, <https://doi.org/10.1016/j.patcog.2022.108990>.
- Hedayati, S., O'Donnell, R.E. and Wyble, B. (2022) 'A model of working memory for latent representations', *Nature Human Behaviour*, Vol. 6, No. 5, pp.709–719, <https://doi.org/10.1038/s41562-021-01264-9>.
- Jaafar, N. and Lachiri, Z. (2023) 'Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance', *Expert Systems with Applications*, Vol. 211, p.118523, <https://doi.org/10.1016/j.eswa.2022.118523>.
- Kim, K. and Park, S. (2023) 'AOBERT: All-modalities-in-one BERT for multimodal sentiment analysis', *Information Fusion*, Vol. 92, pp.37–45, <https://doi.org/10.1016/j.inffus.2022.11.022>.

- Kumar, A. and Vepa, J. (2020) 'Gated mechanism for attention based multi modal sentiment analysis', Paper presented at the *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, <https://doi.org/10.1109/ICASSP40776.2020.9053012>.
- Kumar, S., Chaube, M.K., Alsamhi, S.H., Gupta, S.K., Guizani, M., Gravina, R. and Fortino, G. (2022) 'A novel multimodal fusion framework for early diagnosis and accurate classification of COVID-19 patients using X-ray images and speech signal processing techniques', *Computer Methods and Programs in Biomedicine*, Vol. 226, p.107109, <https://doi.org/10.1016/j.cmpb.2022.107109>.
- Kumar, S., Shvetsov, A.V. and Alsamhi, S.H. (2024) 'Fuzzyguard: a novel multimodal neuro-fuzzy framework for copd early diagnosis', *IEEE Internet of Things Journal*, <https://doi.org/10.1109/IJOT.2024.3467176>.
- Li, F., Li, Y. and Pan, Y. (2024) 'A computational offloading algorithm for cloud-edge collaboration in smart agriculture', *International Journal of Simulation and Process Modelling*, Vol. 21, No. 2, pp.121–129, <https://doi.org/10.1504/IJSPM.2024.141985>.
- Li, L., Tayir, T., Han, Y., Tao, X. and Velásquez, J.D. (2023) 'Multimodality information fusion for automated machine translation', *Information Fusion*, Vol. 91, pp.352–363, <https://doi.org/10.1016/j.inffus.2022.10.018>.
- Li, S., Li, Y., Wu, X., Al Otaibi, S. and Tian, Z. (2022) 'Imbalanced malware family classification using multimodal fusion and weight self-learning', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 24, No. 7, pp.7642–7652, <https://doi.org/10.1109/TITS.2022.3208891>.
- Li, Y., Zhang, K., Wang, J. and Gao, X. (2021) 'A cognitive brain model for multimodal sentiment analysis based on attention neural networks', *Neurocomputing*, Vol. 430, pp.159–173, <https://doi.org/10.1016/j.neucom.2020.10.021>.
- Liu, J., Wu, J., Xie, Y., Jie, W., Xu, P., Tang, Z. and Yin, H. (2022a) 'Toward robust process monitoring of complex process industries based on denoising sparse auto-encoder', *Journal of Industrial Information Integration*, Vol. 30, p.100410, <https://doi.org/10.1016/j.jii.2022.100410>.
- Liu, J.-X., Yin, M.-M., Gao, Y.-L., Shang, J. and Zheng, C.-H. (2022b) 'Msf-lrr: multi-similarity information fusion through low-rank representation to predict disease-associated microbes', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 20, No. 1, pp.534–543, <https://doi.org/10.1109/TCBB.2022.3146176>.
- Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A. and Morency, L.-P. (2018) 'Efficient low-rank multimodal fusion with modality-specific factors', arXiv preprint arXiv:1806.00064, <https://doi.org/10.48550/arXiv.1806.00064>.
- Liu, Z., Zhou, B., Chu, D., Sun, Y. and Meng, L. (2024) 'Modality translation-based multimodal sentiment analysis under uncertain missing modalities', *Information Fusion*, Vol. 101, p.101973, <https://doi.org/10.1016/j.inffus.2023.101973>.
- Nojavanasghari, B., Gopinath, D., Koushik, J., Baltrušaitis, T. and Morency, L.-P. (2016) 'Deep multimodal fusion for persuasiveness prediction', Paper presented at the *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, <https://doi.org/10.1145/2993148.2993176>.
- Oza, A., Patel, S., Gyanchandani, B.S., Sahu, J., Roy, A., Kumar, S. and Shrivastava, A. (2024) 'A novel multimodal framework for early detection of congestive heart failure using ensemble learning based fusion approach', Paper presented at the *2024 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*, <https://doi.org/10.1109/CVMI61877.2024.10781823>.
- Parkash, S. and Tewari, P. (2024) 'Comparative analysis of performability and maintenance decision matrix for a repairable industrial system', *International Journal of Simulation and Process Modelling*, Vol. 21, No. 2, pp.108–120, <https://doi.org/10.1504/IJSPM.2024.141984>.
- Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A. and Morency, L.-P. (2017) 'Context-dependent sentiment analysis in user-generated videos', Paper presented at the *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, Long papers, <https://doi.org/10.18653/v1/P17-1081>.
- Preen, R.J., Wilson, S.W. and Bull, L. (2021) 'Autoencoding with a classifier system', *IEEE Transactions on Evolutionary Computation*, Vol. 25, No. 6, pp.1079–1090, <https://doi.org/10.1109/TEVC.2021.3079320>.
- Rajagopalan, S.S., Morency, L.-P., Baltrušaitis, T. and Goecke, R. (2016) 'Extending long short-term memory for multi-view structured learning', Paper presented at the *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, Proceedings, Part VII 14, The Netherlands, October 11–14*.
- Sahay, S., Kumar, S.H., Xia, R., Huang, J. and Nachman, L. (2018) 'Multimodal relational tensor network for sentiment and emotion classification', arXiv preprint arXiv:1806.02923, <https://doi.org/10.48550/arXiv.1806.02923>.
- Slavic, G., Alemaw, A.S., Marcenaro, L., Gómez, D.M. and Regazzoni, C. (2022) 'A Kalman variational autoencoder model assisted by odometric clustering for video frame prediction and anomaly detection', *IEEE Transactions on Image Processing*, Vol. 32, pp.415–429, <https://doi.org/10.1109/TIP.2022.3229620>.
- Sun, H., Liu, J., Chen, Y.-W. and Lin, L. (2023) 'Modality-invariant temporal representation learning for multimodal sentiment classification', *Information Fusion*, Vol. 91, pp.504–514, <https://doi.org/10.1016/j.inffus.2022.10.031>.
- Sun, J., Yin, H., Tian, Y., Wu, J., Shen, L. and Chen, L. (2021) 'Two-level multimodal fusion for sentiment analysis in public security', *Security and Communication Networks*, Vol. 2021, No. 1, p.6662337, <https://doi.org/10.1155/2021/6662337>.
- Sun, W., Ren, K., Meng, X., Yang, G., Xiao, C., Peng, J. and Huang, J. (2022) 'MLR-DBPFN: a multi-scale low rank deep back projection fusion network for anti-noise hyperspectral and multispectral image fusion', *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, pp.1–14, <https://doi.org/10.1109/TGRS.2022.3146296>.
- Tang, J., Li, K., Hou, M., Jin, X., Kong, W., Ding, Y. and Zhao, Q. (2022) 'MMT: multi-way multi-modal transformer for multimodal learning', Paper presented at the *IJCAI*.
- Varun, C.M., Kumar, R.P.A., Reya, M.R.J. and Murugan, R. (2024) 'Secured big data storage in cloud by intelligent authentication and privacy preservation via intelligent deep learning-aided heuristic strategy', *International Journal of Simulation and Process Modelling*, Vol. 21, No. 2, pp.90–107, <https://doi.org/10.1504/IJSPM.2024.141986>.

- Wang, H., Meghawat, A., Morency, L-P. and Xing, E.P. (2016) 'Select-additive learning: improving cross-individual generalization in multimodal sentiment analysis', arXiv preprint arXiv:1609.05244, 2(4.5).
- Wang, N., Liu, T., Wang, J., Liu, Q., Alibhai, S. and He, X. (2022a) 'Locality-based transfer learning on compression autoencoder for efficient scientific data lossy compression', *Journal of Network and Computer Applications*, Vol. 205, p.103452, <https://doi.org/10.1016/j.jnca.2022.103452>.
- Wang, Y., Feng, Y., Zhang, L., Zhou, J.T., Liu, Y., Goh, R.S.M. and Zhen, L. (2022b) 'Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images', *Medical Image Analysis*, Vol. 81, p.102535, <https://doi.org/10.1016/j.media.2022.102535>.
- Wang, Z., Xu, G., Zhou, X., Kim, J.Y., Zhu, H. and Deng, L. (2022c) 'Deep tensor evidence fusion network for sentiment classification', *IEEE Transactions on Computational Social Systems*, <https://doi.org/10.1109/TCSS.2022.3197994>.
- Xu, H., Wang, X., Huang, J., Zhang, F. and Chu, F. (2024) 'Semi-supervised multi-sensor information fusion tailored graph embedded low-rank tensor learning machine under extremely low labeled rate', *Information Fusion*, Vol. 105, p.102222, <https://doi.org/10.1016/j.inffus.2023.102222>.
- Yan, X., Xue, H., Jiang, S. and Liu, Z. (2022) 'Multimodal sentiment analysis using multi-tensor fusion network with cross-modal modelling', *Applied Artificial Intelligence*, Vol. 36, No. 1, p.2000688.
- Yu, S. and Chen, X. (2020) 'Infrared and visible image fusion based on a latent low-rank representation nested with multiscale geometric transform', *IEEE Access*, Vol. 8, pp.110214–110226, <https://doi.org/10.1109/ACCESS.2020.3001974>.
- Zadeh, A., Chen, M., Poria, S., Cambria, E. and Morency, L-P. (2017) 'Tensor fusion network for multimodal sentiment analysis', arXiv preprint arXiv:1707.07250, <https://doi.org/10.48550/arXiv.1707.07250>.
- Zadeh, A., Liang, P.P., Mazumder, N., Poria, S., Cambria, E. and Morency, L-P. (2018a) 'Memory fusion network for multi-view sequential learning', Paper presented at the *Proceedings of the AAAI Conference on Artificial Intelligence*, <https://doi.org/10.1609/aaai.v32i1.12021>.
- Zadeh, A., Liang, P.P., Poria, S., Vij, P., Cambria, E. and Morency, L-P. (2018b) 'Multi-attention recurrent network for human communication comprehension', Paper presented at the *Proceedings of the AAAI Conference on Artificial Intelligence*, <https://doi.org/10.1609/aaai.v32i1.12024>.
- Zeng, J., Zhou, J. and Liu, T. (2022) 'Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities', *IEEE Transactions on Multimedia*, Vol. 25, pp.6301–6314, <https://doi.org/10.1109/TMM.2022.3207572>.
- Zhang, H., Wan, L., Haug, T., Mok, W-K., Paesani, S., Shi, Y. and Xiao, L. (2022) 'Resource-efficient high-dimensional subspace teleportation with a quantum autoencoder', *Science Advances*, Vol. 8, No. 40, p.eabn9783, <https://doi.org/10.1126/sciadv.abn9783>.
- Zhang, S. and Jia, W. (2024) 'Dynamic rolling optimisation model and VNTSA algorithm for timely delivery services', *International Journal of Simulation and Process Modelling*, Vol. 21, No. 3, pp.155–165, <https://doi.org/10.1504/IJSPM.2024.143844>.
- Zhao, X., Lv, Y. and Huang, Z. (2022) 'Multimodal fusion-based swin transformer for facial recognition micro-expression recognition', Paper presented at the 2022 *IEEE International Conference on Mechatronics and Automation (ICMA)*, <https://doi.org/10.1109/ICMA54519.2022.9856162>.