scientific reports

Check for updates

OPEN

# Multimodal multi-instance evidence fusion neural networks for cancer survival prediction

Hui Luo[1,2], Jiashuang Huang[3], Hengrong Ju[3], Tianyi Zhou[3] & Weiping Ding[1,3]✉

**Accurate cancer survival prediction plays a crucial role in assisting clinicians in formulating treatment plans. Multimodal data, such as histopathological images, genomic data, and clinical information, provide complementary and comprehensive information, significantly enhancing the accuracy of this task. However, existing methods, despite achieving some promising results, still exhibit two significant limitations: they fail to effectively utilize global context and overlook the uncertainty of different modalities, which may lead to unreliable predictions. In this study, we propose a multimodal multi-instance evidence fusion neural network for cancer survival prediction, called M2EF-NNs. Specifically, to better capture global information from images, we employ a pre-trained vision transformer model to extract patch feature embeddings from histopathological images. Additionally, we are the first to apply the Dempster–Shafer evidence theory to the cancer survival prediction task and introduce subjective logic to estimate the uncertainty of different modalities. We then dynamically adjust the weights of the class probability distribution after multimodal fusion based on the estimated evidence from the fused multimodal data to achieve trusted survival prediction. Finally, the experimental results on three cancer datasets demonstrate that our method significantly improves cancer survival prediction regarding overall C-index and AUC, thereby validating the model's reliability.**

**Keywords** Survival prediction, Multimodal fusion, Vision transformer, Dempster–Shafer evidence theory

Cancer has become a significant global public health issue due to factors such as population aging, environmental pollution, unhealthy lifestyle habits, and changes in dietary patterns[1]. Histopathological images are considered the gold standard for cancer diagnosis, and their diagnostic results directly affect the choice of patient treatment and the prediction of prognosis[2–4]. Survival prediction is an essential aspect of cancer prognosis research. However, relying solely on histopathological changes cannot objectively and accurately reflect the occurrence and development of the disease due to individual variations and cancer heterogeneity. With the advancements in molecular biology, there is now substantial research evidence indicating that many molecular pathological changes occur in the early stages of cancer, even before apparent morphological changes in cancer tissue. In addition, even among cancer patients with the same histopathological type, there can be completely different molecular alterations within the cancer, leading to variations in drug treatment response and prognosis. Therefore, cancer experts often combine histopathology information with genomics information to predict patient survival outcomes[5]. However, it is important to note that histopathological images and genomic data are heterogeneous, originating from different sources. Therefore, it is crucial to consider and address the heterogeneity among various modalities, as well as the unique information inherent in each modality. Fortunately, the emergence of whole slide imaging (WSI) and breakthroughs in deep learning methods have provided exciting possibilities for the effective integration of histopathology and genomic information[6–9]. This integration allows for better quantification of the tumor microenvironment for prognostic prediction.

Recently, researchers have introduced a variety of deep learning-based methods aimed at enhancing the performance of cancer survival prediction through the integration of data from different modalities[10,11]. For instance, Wang et al.[12] proposed the GPDBN, a fusion method that aims to enhance cancer survival prediction performance by taking into account the relationships within and between different modalities. But, it fails to fully explore the dynamic relationships within each modality. Additionally, Chen et al.[13] presented a novel pathology fusion method for integrating histopathology and genomic data. The method utilized CNNs and graph convolutional networks (GCNs) to train morphological features from histopathological images and self-normalizing neural networks (SNNs) to train genomic features. They then simplified the output representation

[1]Faculty of Data Science, City University of Macau, Macau 999078, China. [2]School of Information and Management, Guangxi Medical University, Nanning 530021, China. [3]School of Artificial Intelligence and Computer Science, Nantong University, Nantong 226019, China. ✉email: dwp9988@163.com

with a gate-based attention mechanism to reduce noise. Finally, they employed the Kronecker product to combine the deep features from histopathology and genomics, enabling the prediction of patient survival outcomes. The experiments yielded promising results. However, this approach does not effectively model cross-modal interactions. Subsequently, Chen et al.[14] put forward a novel architecture named MCAT, which combines multiple instance learning and the Transformer model, to investigate intra-modality interactions. This model employs attention mechanisms to learn the interaction between images and genomics, thereby visually demonstrating the interpretability of multi-modality interactions.

The current methods, despite yielding promising results, suffer from two notable limitations: they do not effectively utilize global context and disregard modal uncertainty. Specifically, traditional methods preserve limited global information and utilize patch features obtained from pre-trained ResNet models on ImageNet. In comparison to vision transformer (ViT) models, CNN models lack spatial positional information and retain less global context. These structural differences result in variations in out-of-distribution generalization capabilities[15,16]. However, survival prediction is a challenging ordinal regression task aimed at predicting the relative risk of cancer mortality, requiring complex interactions across instances and between instances throughout the entire WSI. Also, traditional approaches primarily focus on improving accuracy by leveraging the complementarity of different modalities while neglecting uncertainty, which may lead to unreliable predictions. The quality of data often varies across samples due to the presence of varying amounts of information and potential noise in histopathological images and genomic data[17,18]. Existing methods either simply treat each modality as equally important or adjust the weights of different modalities to fixed values, integrating them into a shared representation for subsequent tasks. Although these methods are effective, they often overlook uncertainty and are insufficient to capture the dynamic noise in multimodal data, potentially resulting in unreliable predictions. Therefore, ensuring the reliability of multimodal integration and final decision-making is also crucial.

To address these challenges, we propose a multimodal multi-instance evidence fusion neural networks (M2EF-NNs) framework for cancer survival prediction. Our proposed method is capable of capturing comprehensive global information while considering the uncertainties associated with individual modalities. Specifically, we employ a pre-trained ViT model on ImageNet to extract feature embeddings from histopathology images. Then, using genomic embeddings as queries, we learn the co-attention mapping between the genomic features and histopathological images to achieve early interaction and fusion of multimodal information. Subsequently, we parameterize the distribution of class probabilities using the processed multimodal features and introduce subjective logic to estimate the uncertainty of different modalities. By integrating the Dempster–Shafer theory (DST), we can dynamically adjust the weights of the class probabilities after multimodal fusion based on the estimated evidence from the fused multimodal data, thereby achieving reliable survival predictions. The primary contributions of this study are summarized as follows:

- To effectively capture the global information in the histopathological images, we employ a pre-trained ViT model trained on the ImageNet dataset. Specifically, we sliced the histopathological images at a magnification scale of 20x and extracted features using the pre-trained model.
- We introduce a multimodal attention module to fuse the processed histopathology images and genomic data. By learning the co-attention mapping between the genomic features and histopathology images, we achieve early interaction and fusion of multimodal information, enabling us to better capture their correlations.
- We consider the uncertainty of each modality in the cancer survival prediction by using the DST. We parameterize the distribution of class probabilities using the processed multimodal features and introduce subjective logic to estimate the uncertainty associated with different modalities. By combining with the Dempster–Shafer theory, we can dynamically adjust the weights of class probabilities after multimodal fusion to achieve trusted survival prediction. Based on the information available to us, our work is the first to explore the application of the Dempster–Shafer theory in cancer survival prediction.
- Experimental validation on three different cancer datasets confirms the significant improvements achieved by our proposed method in cancer survival prediction and enhances the reliability of the model.

## Related work

### Weakly supervised learning in WSIs

In pathology, the heterogeneity and high resolution nature of WSI makes it challenging to directly label the entire image and use it for training. Moreover, only a small portion of the WSI is usually relevant to the disease. Therefore, many methods have started to adopt weakly supervised learning approaches to deal with pathology images. Currently, these weakly supervised methods based on multiple instance learning (MIL) and other set-based deep learning techniques have achieved remarkable progress in the field of WSI applications. Researchers such as Edwards, Storkey, and Zaheer were among the pioneers in proposing neural network architectures based on set data structures[19,20]. Subsequently, Ilse et al.[21] further extended the set-based architecture by introducing attention mechanisms and applying them to WSIs. To further enhance predictive performance, Yao et al.[22] proposed an attention-guided deep MIL network, which combines multiple instance learning and attention mechanisms for survival analysis in cancer patients, offering good interpretability. Recently, Chen et al.[14] introduced MCAT, a novel structure that combines multiple instance learning and transformer models for early fusion interaction of modalities, further exploring the interactions within modalities and achieving good predictive performance. In summary, while MIL-based pathology analysis methods may not fully capture the complex interactions between instances, they have been able to effectively address the needle-in-a-haystack problem in pathology. These methods provide powerful tools for pathology researchers to tackle the challenges of whole-slide digital pathology images and offer more accurate disease analysis and prediction.

### Vision transformer

Google introduced the Transformer architecture in 2017 as a deep learning framework. Initially designed to tackle machine translation tasks, it utilizes a self-attention mechanism and incorporates a multi-head attention mechanism[23]. In 2020, Dosovitskiy et al.[24] introduced the architecture to image classification for the first time, proposing the ViT. This work demonstrated that image classification is not necessarily dependent on traditional CNNs and becoming a benchmark method for many computer vision tasks. Recent studies on ViT have primarily focused on robustness[25], the impact of self-attention on the models[26], model improvements[27], and comparisons with CNN models[15,16]. In the medical domain, histopathology images exhibit complex composition, where some abnormal images consist mostly of abnormal patches, while others contain only a small fraction of abnormal patches. Therefore, models used for histopathology image feature extraction tasks must possess strong capabilities to extract global information. The ViT architecture introduces positional encoding, improving accessibility to global information and exhibiting excellent out-of-distribution generalization and prominent feature representation capabilities. As a result, ViT has demonstrated outstanding performance in histopathology images segmentation and classification tasks[28,29]. For example, Gao et al.[30] introduced i-ViT, a method designed to learn robust representations of histopathological images for tasks involving subtyping papillary renal cell carcinoma. This approach focuses on extracting detailed features from patches to enhance the quality of the learned representations. Chen et al.[31] presented GasHisTransformer, a model that combines the strengths of ViT and CNN architectures to achieve automatic global detection of gastric cancer images. Huang et al.[32] proposed a novel cross-scale fusion Transformer model, designed to efficiently integrate patch embeddings from diverse fields of view and learn cross-scale contextual relationships. However, these methods based on vision transformers face the challenge of high computational costs. Some approaches utilize random sampling[33] or patch clustering[22] for downsampling to reduce computational complexity, which may lead to information loss. In this context, the Swin Transformer, as a variant of ViT, employs a local window self-attention mechanism that reduces computational complexity, making it more suitable for processing high-resolution images. Therefore, we adopt a pre-trained Swin Transformer model for feature extraction from pathological images.

### Dempster–Shafer evidence theory

DST, initially proposed by Dempster, is a theory that deals with belief functions[34]. It serves as an extension of Bayesian theory by incorporating subjective probabilities[35]. It has been developed as a general framework for modeling uncertainty in cognition[36]. Unlike Bayesian neural networks that indirectly model uncertainty through parameter sampling, DST directly models uncertainty. It enables the combination of beliefs from multiple sources using different fusion operators, resulting in new beliefs that incorporate all available evidence[37]. In data classification, the belief functions and Dempster's combination rule of evidence theory have been combined with various classification algorithms to handle uncertainty[38–40]. Denoeux et al.[41] proposed a multi-layer neural network that combines the Dempster rule for adaptive pattern classification. Sensoy et al.[42] applied evidence theory to quantify uncertainty in deep convolutional neural network classification and constructed an evidence-deep neural network. Furthermore, evidence-based deep neural networks have been used to build algorithms for uncertain data classification[43–45]. Recently, Han et al. introduced a trusted multimodal classification method that combines the DS evidence theory and subjective logic. It parameterizes the distribution parameters of class probabilities as Dirichlet distributions using evidence from different views and estimates the uncertainty of different viewpoints[46]. In order to ensure the reliability of multimodal integration and the final prediction of histopathology and genomics, we are the first to introduce the Dempster–Shafer evidence theory for trusted multimodal survival prediction.

## Method

In this section, we will introduce our overall framework, as shown in Fig. 1. The M2EF-NNs proposed in this paper contain three main modules, including the multimodal multiple instance feature extraction, the multimodal feature fusion, and the DST-based trusted survival prediction.

### Multimodal multiple instance feature extraction

In multiple instance learning, training samples are organized in the form of instance bags, where each bag consists of multiple instances and is assigned a label indicating its class. In our study, we constructed corresponding instance bags for each patient's histopathology images and genomic data. Each instance bag contains a varying number of instances. The label for each instance bag represents the total survival time of the patient in months. Additionally, we used a label to indicate the survival status, where 0 represents death and 1 represents survival.

*Feature extraction for histopathology images*: To handle large-scale WSIs, we adopted a two-stage MIL approach for feature extraction. First, we needed to crop the complete WSI into multiple equal-sized patches and extract features from each patch to obtain instance-level features for subsequent tasks. Considering the high resolution of WSIs, processing the entire image at once in memory is challenging. Therefore, we opted for a staged approach. Before model training, we first used a pre-trained model for feature extraction, saving the extracted features for subsequent model training. The entire feature extraction process is illustrated in Fig. 2. We selected WSIs at a magnification scale of 20x for analysis. First, for each WSI, we performed tissue segmentation employing the method described in Lu et al.[47]. The segmented tissue regions were then cropped into $256 \times 256$ patches, ensuring no spatial overlap between the patches. For each patch, we extracted a 1536-dimensional feature embedding $h \in \mathbb{R}^{1536\times 1}$ using the pre-trained Swin-L model[27] trained on ImageNet. Because among the different variants of the Swin Transformer, Swin-L has the most parameters and a more complex network structure, it is capable of capturing richer features and more intricate patterns. Additionally, when handling high-resolution images and complex tasks, Swin-L typically offers higher accuracy and better performance. We then map the feature embeddings to a 256-dimensional space using a fully connected (FC) layer. Finally, we

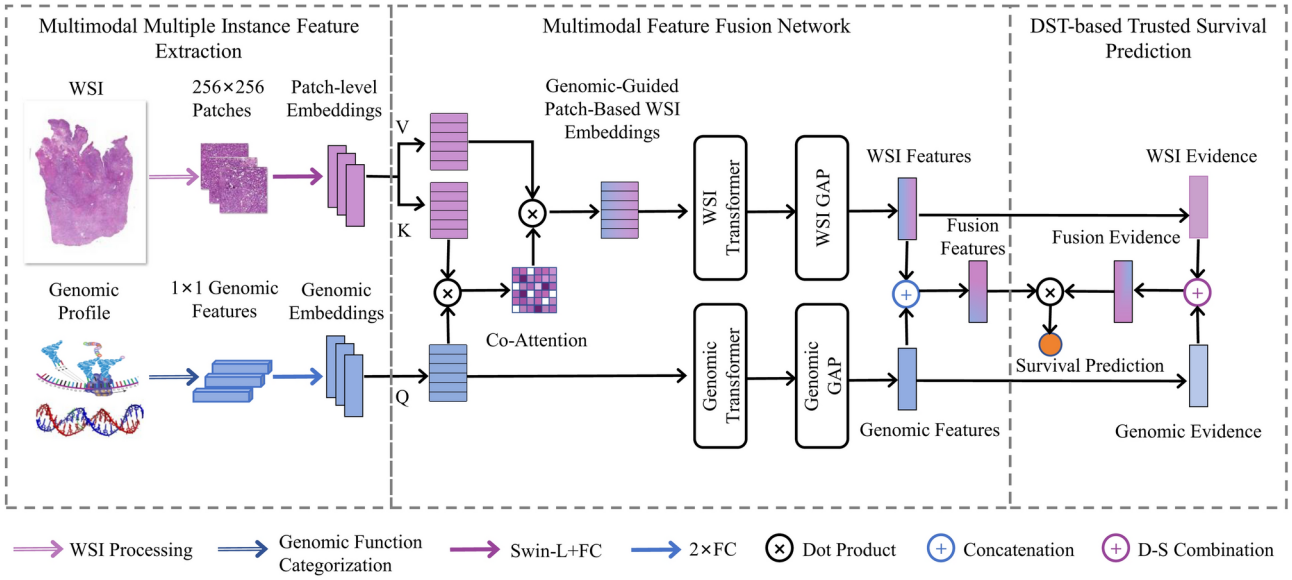


**Fig. 1**. Overview of M2EF-NNs. It consists of three parts: (1) Multimodal multiple instance feature extraction: The part is to extract features from multimodal data. Specifically, patches are sampled from WSI with 20x magnification and then fed them into a pre-trained Swin-L model to extract patch features at the instance level. Second, based on the biological function, genomic profiles are classified, and instance-level genomic features are extracted. (2) Multimodal feature fusion: In this part, multiple strategies are employed to fuse features from different modalities. First, we learn the co-attention mapping between the genomic features and histopathology images to achieve early interaction and fusion of multimodal information. Second, multi-instance features are aggregated using global attention pooling, transforming instance-level features into image-level features. Finally, the features from different modalities are fused using a concatenation method to obtain the final fused feature vector. (3) DST-based trusted survival prediction: In this stage, the Dempster–Shafer theory is utilized to integrate multimodal information at the evidence level and dynamically adjust the weights of fused class probabilities after multimodal fusion for trusted survival prediction.

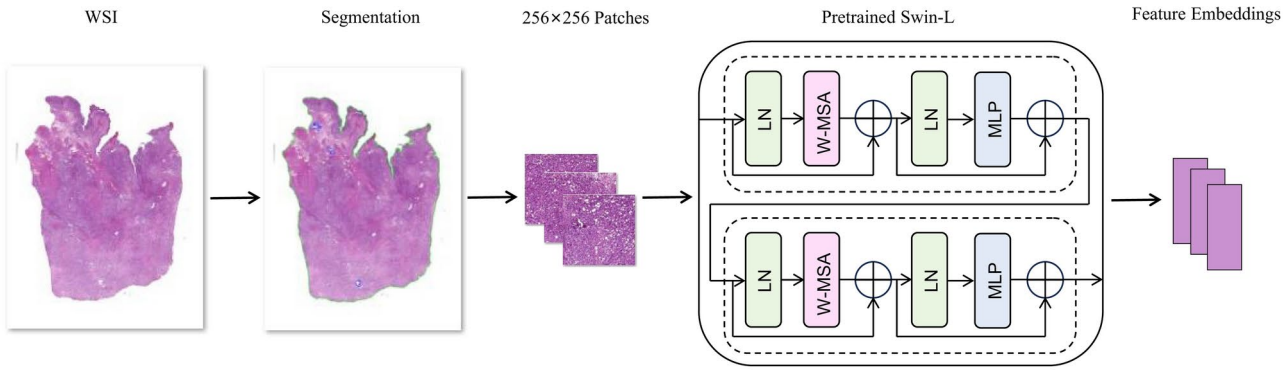


**Fig. 2**. The specific framework of WSI feature extraction.

packaged each patch-level embedding into an instance bag $H_{bag} \in \mathbb{R}^{M\times 256}$, where $M$ is the instance size of the bag and the instance size of each WSI is not unique.

*Feature extraction for genomics*: Genomic data primarily consists of gene mutation status, copy number variations, and gene expression data, which are typically represented as $1 \times 1$ attributes. In our study, we utilized the semantic information within genes and grouped genes with similar biological functions into one instance for feature extraction. We divided the gene sets using six functional category labels obtained from[48,49]. These sets are tumor suppressor, tumor occurrence, protein kinase, cell differentiation, transcription, and cytokine and growth. The size of each gene set may vary. To process these genomic data, we employed two fully connected layers (FC layers). The first FC layer was used to partition all gene attributes and assign them to their corresponding gene sets. The second FC layer mapped each gene set to a 256-dimensional feature embedding $\{g_n \in \mathbb{R}^{256\times 1}\}_{n=1}^{N}$. These feature embeddings were then packaged into a genomic bag $G_{bag} \in \mathbb{R}^{N\times 256}$. Here, $N$ was set to 6, indicating that each genomic bag contained six instances corresponding to six functional category label sets.

## Multimodal feature fusion network

In multimodal feature fusion, our goal is to integrate features from different modalities to achieve a more comprehensive and robust feature representation than a single modality alone. In this stage, our primary emphasis is on the multimodal fusion of histopathology images and genomic features. The objective is to construct a unified feature representation for subsequent survival prediction. This process involves three steps: cross-modal early fusion, multiple instance aggregation, and multimodal late fusion.

*Cross-modal early fusion*: First, inspired by the work of Chen et al.[14] and Vaswani et al.[23], we found that we can use genomic embeddings $G_{bag} \in \mathbb{R}^{N\times 256}$ as a substitute for the query $Q$ in the Transformer attention mechanism. We then multiply the query with the histopathology image embeddings $H_{bag} \in \mathbb{R}^{M\times 256}$ to obtain a co-attention matrix $A_{coat}$ between the two modalities, achieving early interaction fusion. Here, our goal is to map histopathology image embeddings $H_{bag}$ to a set of genomic-guided histopathology image embeddings $\hat{H}_{bag}$. This mapping can be represented as:

$$CoAt_{G\to H}(G,H) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) = softmax\left(\frac{W_q G H^T W_k^T}{\sqrt{256}}\right) W_v H \to A_{coat} W_v H \to \widehat{H} \quad (1)$$

where $W_q$, $W_k$, and $W_v$ are trainable weight matrices that are multiplied with the query $G_{bag}$ and key-value pairs ($H_{bag}$,$H_{bag}$), respectively, and $A_{coat} \in \mathbb{R}^{N\times M}$ is the co-attention matrix used to compute the weighted average of $H_{bag}$. By employing this approach, we achieve an early fusion of histopathology image features and genomic features, allowing them to influence and interact with each other, thereby extracting richer multimodal features. Furthermore, in the experiments, the batch size of Q is 6, while the batch sizes of $K$ and $V$ are generally larger than $Q$. As a result, aggregating $\hat{H}_{bag} \in \mathbb{R}^{N\times 256}$ is computationally more efficient than aggregating $H_{bag} \in \mathbb{R}^{M\times 256}$, significantly reducing the computational cost. The obtained $\hat{H}_{bag}$ is essentially a WSI feature embedding that is enhanced with the attention of genomic information.

*Multiple instance aggregation*: Then, we pass the obtained genomic-guided histopathology image embeddings $\hat{H}_{bag}$ and the genomic feature matrix $G_{bag}$ through two Transformer encoder layers and a global attention-based pooling layer for multiple instance aggregation. In the instance aggregation, we first compute their attention scores as follows:

$$A_{ih} = \frac{\exp\left\{W_{\rho h}\left[\tanh\left(V_{\rho h} h_i^T\right) \odot sigmod\left(U_{\rho h} h_i^T\right)\right]\right\}}{\sum_{ih=1}^{N} \exp\left\{W_{\rho h}\left[\tanh\left(V_{\rho h} h_i^T\right) \odot sigmod\left(U_{\rho h} h_i^T\right)\right]\right\}}, \quad (2)$$

$$A_{ig} = \frac{\exp\left\{W_{\rho g}\left[\tanh(V_{\rho g} g_i^T) \odot sigmod(U_{\rho h} g_i^T)\right]\right\}}{\sum_{ig=1}^{N} \exp\left\{W_{\rho g}\left[\tanh(V_{\rho g} g_i^T) \odot sigmod(U_{\rho h} g_i^T)\right]\right\}}, \quad (3)$$

where $W_{\rho h}$, $V_{\rho h}$, $U_{\rho h}$, $W_{\rho g}$, $V_{\rho g}$, and $U_{\rho g}$ are all trainable parameters. Then, we concatenate all the instance vectors $h$ and $g$ within the bags $H_{bag}$ and $G_{bag}$ to form the feature vector as shown below:

$$R_h = ReLU\left(W_{\zeta h} \sum_{ih-1}^{N} A_{ih} h_i\right), \quad (4)$$

$$R_g = ReLU\left(W_{\zeta g} \sum_{ig=1}^{N} A_{ig} g_i\right), \quad (5)$$

where $W_{\zeta h}$ and $W_{\zeta g}$ are also trainable parameters.

*Multimodal late fusion*: Finally, we pass the aggregated genomic-guided histopathology image feature vector and the genomic feature vector to an MLP (Multi-Layer Perceptron) for late fusion. This MLP comprises two layers with ReLU activation functions. Ultimately, we obtain a 256-dimensional feature vector as follows:

$$R_{fusion} = MLP(R_h \oplus R_g), \quad (6)$$

where $R_{fusion}$ includes both histopathological images and genomic information.

## DST-based trusted survival prediction

In the survival prediction task, we adopted the same model as Chen et al.[14] and modeled the survival time using discrete time intervals. We divided the survival time into four non-overlapping intervals based on its quantiles. Therefore, the discrete survival time corresponds to four labeled categories, then $K$ is equal to 4. Next, we input the unified feature representation constructed through late fusion into a fully connected layer (FC layer). In this FC layer, for each time interval, we can calculate a survival risk score that represents the patient's survival risk within that interval. We use the sigmoid activation function to fit the patient's survival risk score as follows:

$$s_{risk} = sigmod(FC(R_{fusion})), \quad (7)$$

where $s_{risk}$ is a four-dimensional vector, representing the survival risk for each category.

Inspired by the research of Han et al.[46], to ensure the reliability of multimodal integration and the final prediction of histopathology and genomics, as shown in Fig. 3, we used evidence from different modes to

parameterize the distribution of category probability into Dirichlet distribution and introduced subjective logic to estimate the uncertainty of different modes, combined with Dempster–Shafer theory, the weights of category probabilities after multimodal fusion are dynamically adjusted to make trusted multimodal survival prediction. Specifically, firstly, we can get non-negative evidence vectors $e_h$ and $e_g$ from the feature embeddings aggregated by histopathology and genomics through a full connection layer and a softplus activation layer respectively as follows:

$$e_h = Softplus(FC(R_h)), \quad e_g = Softplus(FC(R_g)). \tag{8}$$

Subjective logic links the evidence with the parameters of Dirichlet distribution, for $\alpha = e + 1$. Accordingly, the parameters of Dirichlet distribution of the two modes $\alpha_h$ and $\alpha_g$ can be obtained as follows:

$$\alpha_h = e_h + 1, \quad \alpha_g = e_g + 1. \tag{9}$$

$S_h$ and $S_g$ are Dirichlet strengths which are calculated as follows:

$$S_h = \sum_{i=1}^{K}(e_{kh} + 1) = \sum_{i=1}^{K}\alpha_h, \quad S_g = \sum_{i=1}^{K}(e_{kg} + 1) = \sum_{i=1}^{K}\alpha_g. \tag{10}$$

Subsequently, we calculate the confidence mass and overall uncertainty for each modality separately as follows:

$$b_{kh} = \frac{e_{kh}}{S_h} = \frac{\alpha_{kh} - 1}{S_h}, \quad b_{kg} = \frac{e_{kg}}{S_g} = \frac{\alpha_{kg} - 1}{S_g}, \tag{11}$$

$$u_h = \frac{K}{S_h}, \quad u_g = \frac{K}{S_g}, \tag{12}$$

where $u_h, u_g, b_h, b_g \geq 0$, and meet $u_h + \sum_{k=1}^{K} b_{kh} = 1$ ,$u_g + \sum_{k=1}^{K} b_{kg} = 1$. After obtaining the evidence and uncertainties from individual modalities, we utilized the DST to combine evidence from various sources and generate a degree of belief that incorporates all available evidence. Through the utilization of the Dempster–Shafer combination rule, we can derive the final confidence quality for each category and assess the overall uncertainty as follows:

$$b_k = \frac{1}{1 - C}\left(b_{kh}b_{kg} + b_{kh}u_g + b_{kg}u_h\right), \quad u - \frac{1}{1 - C}u_h u_g. \tag{13}$$

Thereby, we can estimate the fused evidence from multiple modalities and determine the corresponding parameters of the Dirichlet distribution as follows:

$$S = \frac{K}{u}, \quad e_k = b_k \times S, \quad \alpha_k = e_k + 1. \tag{14}$$

Then, the final evidence for each category, obtained through evidence fusion, is normalized using the sigmoid function. This normalization process allows us to obtain the weight parameters for the risk scores of each category, derived from the multimodal fusion network. The calculation of the final risk scores is as follows:

$$o_{risk} = sigmod(e) \cdot s_{risk}. \tag{15}$$

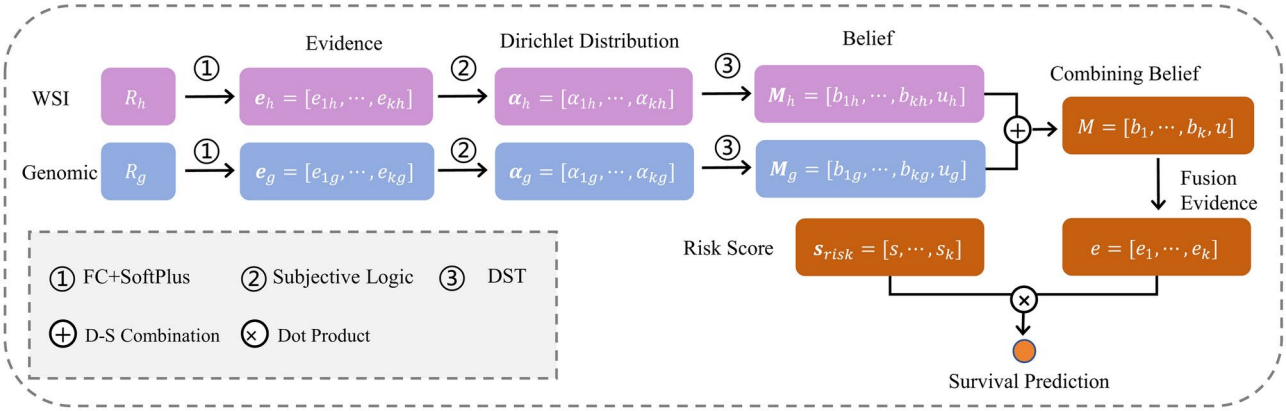


**Fig. 3**. The process of DST-based trusted survival prediction.

In the final risk score $o_{risk}$, the uncertainty of each modality is taken into account. The sizes of the final evidence for each class, obtained through evidence fusion, are used to dynamically adjust the risk scores obtained from the modality fusion network. The greater the evidence accumulated, the higher the weight assigned, leading to an increased final risk score for that particular class. Conversely, the less evidence there is, the smaller the weight assigned to the corresponding risk score, leading to a lower final risk score for the class. Compared to other methods, our approach reduces decision-making risks and improves the reliability of the model. When updating the parameters of the survival model, we followed the approach of Chen et al.[14] by considering the log-likelihood[50] that takes into account the patient's survival status.

## Experiments and results

### Dataset descriptions

To validate our proposed method, we collected data on Bladder Urothelial Carcinoma (BLCA), Glioblastoma & Lower Grade Glioma (GBMLGG), and Breast Invasive Carcinoma (BRCA) from The Cancer Genome Atlas (TCGA). We matched the diagnostic WSIs of each patient with their genomic information and clinical information, including survival time labels and review status. The statistical information for the three constructed datasets is shown in Table 1. For the BLCA dataset, we had 373 patients, 437 WSIs, and 3392 genes. For the GBMLGG dataset, we had 567 patients, 1014 WSIs, and 2723 genes. For the BRCA dataset, we had 955 patients, 1020 WSIs, and 2630 genes.

### Implementation details

We trained our model using an NVIDIA GTX 4090 GPU. Firstly, we sampled slices at a magnification scale of 20x and extracted 1536-dimensional image feature embeddings using the Swin-L model pretrained on ImageNet. Then, we applied two fully connected layers to transform the genomic data into genomic feature embeddings. Next, we fed the extracted histopathology feature embeddings and patient genomic feature embeddings into the multimodal fusion network. We first learned a co-attention mapping between the genomic data and WSIs. Then, we processed the WSI and genomic features separately using two visual transformers and an attention pooling layer. The preprocessed features were subsequently combined through concatenation and fed into a fully connected layers with sigmoid activations to obtain the concatenated risk scores. Simultaneously, the evidence fusion of different modalities was performed using the final evidence from the modal fusion network as the weight parameters for the risk scores of each class obtained from the modal fusion network. During the training phase, we employed the Adam optimizer with a learning rate of $2 \times 10^{-4}$, weight decay of $1 \times 10^{-5}$, batch size of 1, and a total of 20 training epochs.

### Evaluation metrics

In our experiments, we utilized the concordance index (C-index) as an evaluation metric. The concordance index quantifies the correlation between risk scores and survival times and is a commonly used evaluation metric for survival prediction. The calculation formula is as follows:

$$C - index = \frac{1}{n} \sum_{i \in \{1,2,\cdots,N\}} \sum_{t_j > t_i} I[o_i > o_j]. \tag{16}$$

where $n$ is the number of comparable pairs, $s$ are the observed values, $o$ is the survival risk calculated by the model, and $I[\cdot]$ is the indicator function. As for the AUC, it quantifies the ranking quality at the event time level, providing a complementary quantitative measure to the c-Index. It is calculated as follows:

$$AUC = \frac{1}{m} \sum_{t \in T} \sum_{t_i < t} \sum_{t_j > t} I[o_i > o_j]. \tag{17}$$

where $m$ represents the cumulative number of comparable pairs computed over all event times, $t$ is the set of all distinct event times in the dataset. And both the C-index and AUC quantify the ranking quality, with values ranging from 0 to 1. A higher value indicates better predictive performance, while a lower value indicates poorer performance.

In addition to the concordance index, We also employed the log-rank test[51] to evaluate the statistical significance of differences between the two survival curves. The log-rank test allows for the comparison of survival curve differences between different groups, such as the predicted high-risk and low-risk groups. This test helps determine whether the risk stratification predicted by the model has statistical significance. Furthermore, we employed the Kaplan–Meier estimation and the predicted risk distribution to visualize patient stratification[52]. The Kaplan–Meier estimation is a non-parametric approach utilized to estimate survival curves based on information regarding a patient's survival time and events, such as death or recurrence. By visualizing

| Dataset | BLCA | GBMLGG | BRCA |
|---|---|---|---|
| Patient number | 373 | 567 | 955 |
| WSI number | 437 | 1014 | 1020 |
| Gene number | 3392 | 2723 | 2630 |

**Table 1**. Data statistics of datasets.

the predicted risk distribution and survival curves, we can intuitively demonstrate how the model stratifies patients.

## Methods for comparison

In comparisons with state-of-the-art methods on two cancer datasets, we first compare the proposed M2EF-NNs method with a single-modality genomic-based approach and three single-modality MIL methods based on WSIs, including (1) self-normalizing neural network method, (2) sum pooling-based method, (3) global attention pooling-based method, and (4) cluster-based method. Subsequently, we further compare M2EF-NNs with four state-of-the-art multimodal MIL methods, including (1) sum pooling-based method, (2) global attention pooling-based method, (3) cluster-based method, and (4) co-attention-based method. Now we briefly summarize these competing methods as follows.

(1) SNN[53]: SNN is a self-normalizing neural network architecture used for genomic single-line comparison. It utilizes self-normalizing activation functions and weight initialization methods to enhance the training effectiveness and robustness of the network.

(2) Deep sets[20]: Deep sets is one of the earliest set-based neural network architecture. It is a method that performs feature aggregation and pooling at the instance level. This approach can handle input sets with an uncertain number of instances and models the set using a pooling operation with a variable order.

(3) Attention MIL[21]: Attention MIL is a multiple instance learning architecture based on attention mechanisms. It extends the concept of set-based learning and uses global attention pooling instead of sum pooling to aggregate instances in a deep set. This method can adaptively learn the importance weights of different instances to better capture key instances within the set.

(4) DeepAttnMISL[22]: DeepAttnMISL is a survival analysis method based on multiple-instance learning and attention mechanisms. It first clusters instance features into K clusters using the K-Means method and then extracts features from each cluster using convolutional neural networks. Finally, the cluster instance features are aggregated using global attention pooling. This method demonstrates good performance and interpretability in explaining survival-related pathology images.

(5) MCAT[14]: MCAT is an advanced interpretable and weakly supervised multimodal survival prediction architecture. It is based on multi-instance learning and the Transformer model, to explore intra-modality interactions. This model learns the interaction between images and genes using attention mechanisms and visually demonstrates the interpretability of multi-modality interactions.

There are two major strategies in M2EF-NNs, i.e., (1) using a pre-trained Swin-L model to extract pathological image features, and (2) performing trusted survival prediction based on the DST. To investigate the effectiveness of these strategies, we further compare M2EF-NNs with its three variants, including (1) uses the pre-trained Swin-L model for extracting pathological image features but does not employ the survival predictor based on the D–S evidence theory (denoted as MCAT (Swin-L)), (2) does not use the pre-trained Swin-L model for extracting pathological image features but employs the survival predictor based on the D–S evidence theory (denoted as M2EF-NNs (Resnet 50)), and (3) uses the pre-trained Swin-L model for extracting pathological image features and employs the survival predictor based on the D–S evidence theory (denoted as M2EF-NNs (Swin-L)). Additionally, for the purpose of multimodal comparison with M2EF-NNs, we only employed concatenation fusion as a late fusion mechanism to integrate histopathological and genomic features. For all methods, we used the same 5-fold cross-validation splits, training hyperparameters, and loss functions on the three cancer datasets.

## Experimental results

Tables 2 and 3 show the experimental results on three cancer datasets using different methods. From the table, we can find that M2EF-NNs outperform DeepSets, attention MIL, and DeepAttnMISL in terms of single-modality data training methods. The overall C-index performance improved by 20.80%, 13.70%, and 15.59%, while the overall AUC improved by 14.64%, 17.95%, and 16.09%, respectively. This indicates that M2EF-NNs are more accurate in survival prediction tasks compared to these methods. Relative to the genomic baseline, M2EF-NNs achieved a performance improvement of 8.90%. This suggests that M2EF-NNs exhibit better survival prediction capability on genomic data. M2EF-NNs show improvements in all benchmark tests compared to their corresponding single-modality tasks, which is consistent with similar work that utilizes multimodal fusion to enhance supervised learning tasks. This further validates the effectiveness of multimodal data fusion in

| Methods | Model | BLCA | GBMLGG | BRCA | Overall |
|---|---|---|---|---|---|
| Unimodal | SNN (Genomic only) | $0.583 \pm 0.029$ | $0.807 \pm 0.014$ | $0.531 \pm 0.072$ | 0.640 |
| | DeepSets (WSI only) | $0.497 \pm 0.013$ | $0.744 \pm 0.014$ | $0.491 \pm 0.060$ | 0.577 |
| | Attention MIL (WSI only) | $0.541 \pm 0.075$ | $0.771 \pm 0.032$ | $0.527 \pm 0.039$ | 0.613 |
| | DeepAttnMISL (WSI only) | $0.537 \pm 0.024$ | $0.754 \pm 0.025$ | $0.519 \pm 0.061$ | 0.603 |
| Multimodal | DeepSets | $0.597 \pm 0.036$ | $0.801 \pm 0.046$ | $0.546 \pm 0.099$ | 0.648 |
| | Attention MIL | $0.601 \pm 0.024$ | $0.801 \pm 0.028$ | $0.554 \pm 0.085$ | 0.652 |
| | DeepAttnMISL | $0.606 \pm 0.025$ | $0.809 \pm 0.014$ | $0.554 \pm 0.071$ | 0.656 |
| | MCAT | $0.624 \pm 0.023$ | $0.817 \pm 0.024$ | $0.591 \pm 0.085$ | 0.677 |
| | M2EF-NNs | $0.651 \pm 0.021$ | $0.821 \pm 0.034$ | $0.618 \pm 0.030$ | 0.697 |

**Table 2**. The results of comparison experimental by different methods using C-index values.

| Methods | Model | BLCA | GBMLGG | BRCA | Overall |
|---|---|---|---|---|---|
| Unimodal | SNN (Genomic only) | $0.599 \pm 0.037$ | $0.824 \pm 0.021$ | $0.610 \pm 0.077$ | 0.678 |
| | DeepSets (WSI only) | $0.560 \pm 0.046$ | $0.764 \pm 0.059$ | $0.601 \pm 0.039$ | 0.642 |
| | Attention MIL (WSI only) | $0.545 \pm 0.092$ | $0.789 \pm 0.053$ | $0.538 \pm 0.063$ | 0.624 |
| | DeepAttnMISL (WSI only) | $0.547 \pm 0.043$ | $0.778 \pm 0.033$ | $0.576 \pm 0.061$ | 0.634 |
| Multimodal | DeepSets | $0.613 \pm 0.054$ | $0.832 \pm 0.067$ | $0.594 \pm 0.085$ | 0.680 |
| | Attention MIL | $0.613 \pm 0.034$ | $0.837 \pm 0.038$ | $0.650 \pm 0.065$ | 0.700 |
| | DeepAttnMISL | $0.614 \pm 0.043$ | $0.841 \pm 0.026$ | $0.651 \pm 0.098$ | 0.702 |
| | MCAT | $0.652 \pm 0.033$ | $0.854 \pm 0.031$ | $0.667 \pm 0.031$ | 0.724 |
| | M2EF-NNs | $0.675 \pm 0.042$ | $0.862 \pm 0.043$ | $0.671 \pm 0.019$ | 0.736 |

**Table 3**. The results of comparison experimental by different methods using AUC values.

improving survival prediction performance. When compared to the multimodal methods of DeepSets, Attention MIL, DeepAttnMISL, and MCAT, M2EF-NNs achieved increases in overall C-index of 7.56%, 6.90%, 6.25%, and 2.95%, respectively, and improvements in overall AUC of 8.24%, 5.14%, 4.84%, and 1.66%. This indicates that M2EF-NNs have improved overall performance in multimodal survival prediction tasks.

Additionally, stratifying patients into low-risk and high-risk categories is clinically significant, providing prognostic information that guides treatment decisions, patient communication, and resource allocation. For low-risk patients, who typically have a better prognosis, doctors may choose a more conservative treatment approach and maintain an optimistic outlook on treatment outcomes, thereby reducing the frequency of follow-ups and diagnostic checks. In this study, we utilized the 50th percentile as the risk index in the Kaplan–Meier survival analysis to categorize patients into low-risk and high-risk groups. Figure 4 presents the results of the analysis. The Kaplan–Meier survival curves clearly demonstrate that the M2EF-NNs method outperforms other multimodal methods in both datasets, indicating its superior performance. In the BLCA dataset, the log-rank test *P*-values were 1.904e-03 for DeepSets, 1.053e-02 for Attention MIL, 2.939e-03 for DeepAttnMISL, and 3.045e-04 for MCAT, while M2EF-NNs achieved the most significant *P*-value of 2.130e-06. In the GBMLGG dataset, the *P*-values were 4.586e-25 for DeepSets, 1.731e-23 for Attention MIL, 1.120e-24 for DeepAttnMISL, and 3.789e-29 for MCAT, with M2EF-NNs showing the most significant *P*-value of 3.590e-31. For the BRCA dataset, the *P*-values were 2.705e-01 for DeepSets, 1.519e-04 for Attention MIL, 1.673e-02 for DeepAttnMISL, and 2.753e-02 for MCAT, with M2EF-NNs again achieving the most significant *P*-value of 4.002e-06. Therefore, these experiments indicate that the proposed M2EF-NNs method can enhance the performance of cancer survival prediction.

## Ablation studies

To further evaluate the effectiveness of the pre-trained Swin-L model for patch feature extraction and the trusted survival predictor based on the DST, we conducted ablation experiments. Tables 4 and 5 present the results of the ablation experiments. The results demonstrate that both using the Swin-L model for patch feature extraction and the multimodal evidence fusion-based trusted survival predictor contributes to the overall C-index. Specifically, comparing M2EF-NNs (Swin-L) with M2EF-NNs (ResNet 50), we found that the model using the Swin-L model for patch feature extraction outperformed the method using the ResNet 50 model, with an overall average C-index of 0.697 and an AUC of 0.736 across all three datasets. These findings demonstrate that the Swin-L model effectively captures representative features from medical images. Furthermore, the performance difference between M2EF-NNs (Swin-L) and MCAT (Swin-L) validates the effectiveness of the trusted survival predictor based on the DST.

Furthermore, upon examining the Kaplan–Meier survival curves in Fig. 5 and the corresponding log-rank test *p*-values, it becomes evident that our proposed M2EF-NNs method, which leverages the pre-trained Swin-L model for patch feature extraction and the trusted survival predictor based on the DST, significantly improves its performance. In the BLCA dataset, M2EF-NNs(Swin-L) achieved a log-rank test *P*-value of 2.130e-06, which was the most significant compared to M2EF-NNs(ResNet 50) and MCAT(Swin-L). In the GBMLGG dataset, M2EF-NNs(Swin-L) recorded a log-rank test *P*-value of 3.590e-31, again the most significant compared to the other methods. Similarly, in the BRCA dataset, M2EF-NNs(Swin-L) had a log-rank test *P*-value of 4.002e-06, again demonstrating superior significance compared to M2EF-NNs(ResNet 50) and MCAT(Swin-L).

## Multimodal interpretability

In addition to the improved c-Index and significant patient stratification, our M2EF-NNs method is highly interpretable. We utilize genomic embeddings to guide the computation of shared attention weights for histopathological patches. These shared attention weights are overlaid onto the corresponding spatial locations in the original whole-slide images (WSI) to construct genomic-guided visual concept-attention heatmaps. As shown in Fig. 6, in both low-risk and high-risk cases of TCGA-BLCA, the genomic-guided heatmaps reflect several known genotype-phenotype relationships in cancer pathology. Specifically, in BLCA, the genomic-guided visual concept-attention heatmaps often reflect normal stroma, glands, and adipocytes, which are associated with tumor suppression, protein kinases, and cellular differentiation. Across all cases, the cellular differentiation embeddings predominantly focus on tumor-associated stroma, while tumor suppression and protein kinase embeddings primarily concentrate on adipocyte-adjacent stroma and glandular structures. In the

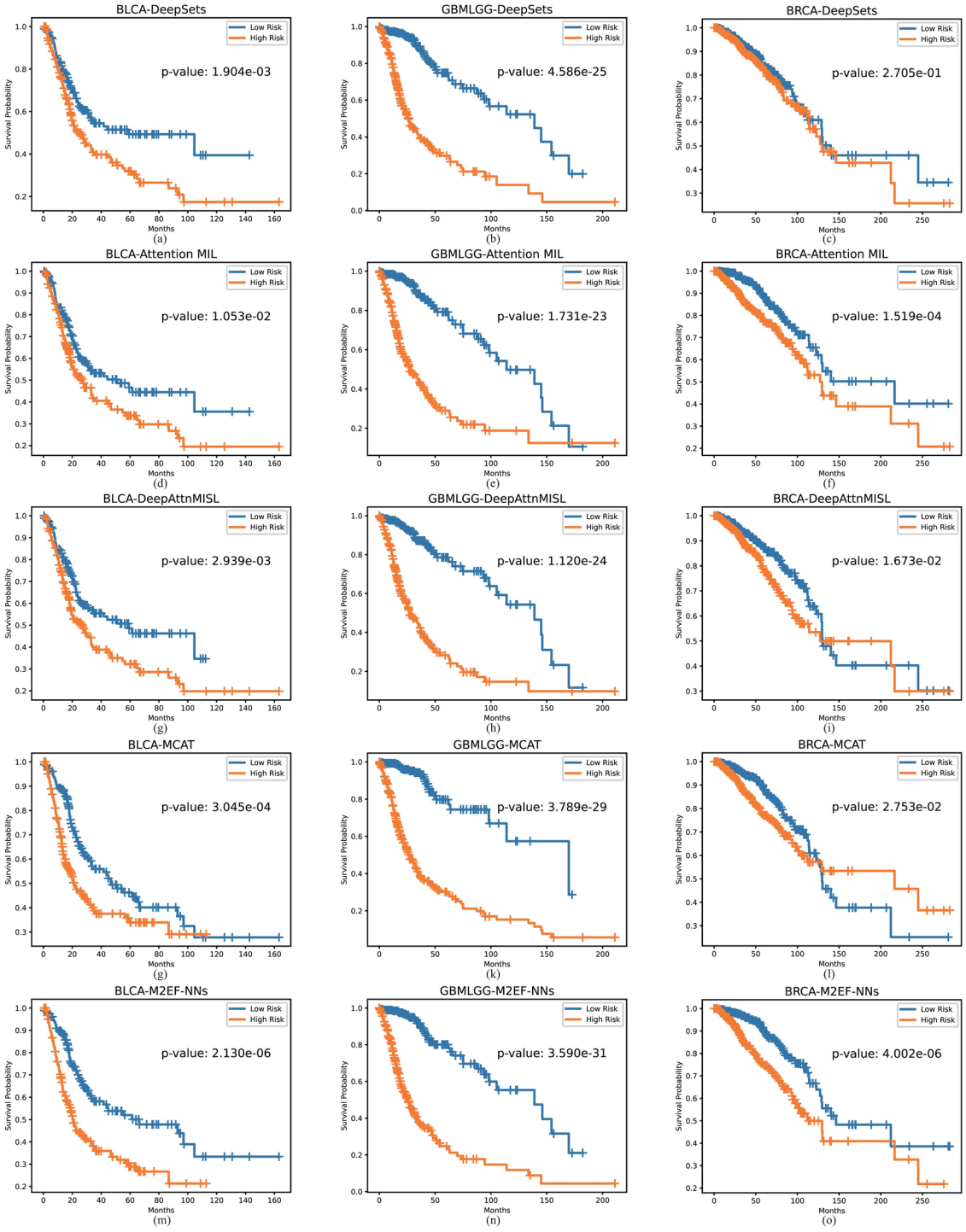


**Fig. 4**. Kaplan–Meier survival curves of the multimodal methods..

| Methods | BLCA | GBMLGG | BRCA | Overall |
|---|---|---|---|---|
| MCAT (Swin-L) | $0.633 \pm 0.030$ | $0.817 \pm 0.030$ | $0.600 \pm 0.065$ | 0.683 |
| M2EF-NNs (Resnet 50) | $0.641 \pm 0.016$ | $0.818 \pm 0.022$ | $0.608 \pm 0.044$ | 0.689 |
| M2EF-NNs (Swin-L) | $0.651 \pm 0.021$ | $0.821 \pm 0.034$ | $0.618 \pm 0.030$ | 0.697 |

**Table 4**. The results of ablation studies using C-index values.

| Methods | BLCA | GBMLGG | BRCA | Overall |
|---|---|---|---|---|
| MCAT (Swin-L) | $0.656 \pm 0.045$ | $0.858 \pm 0.044$ | $0.640 \pm 0.054$ | 0.718 |
| M2EF-NNs (Resnet 50) | $0.665 \pm 0.027$ | $0.858 \pm 0.025$ | $0.667 \pm 0.033$ | 0.730 |
| M2EF-NNs (Swin-L) | $0.675 \pm 0.042$ | $0.862 \pm 0.043$ | $0.671 \pm 0.019$ | 0.736 |

**Table 5**. The results of ablation studies using AUC values.

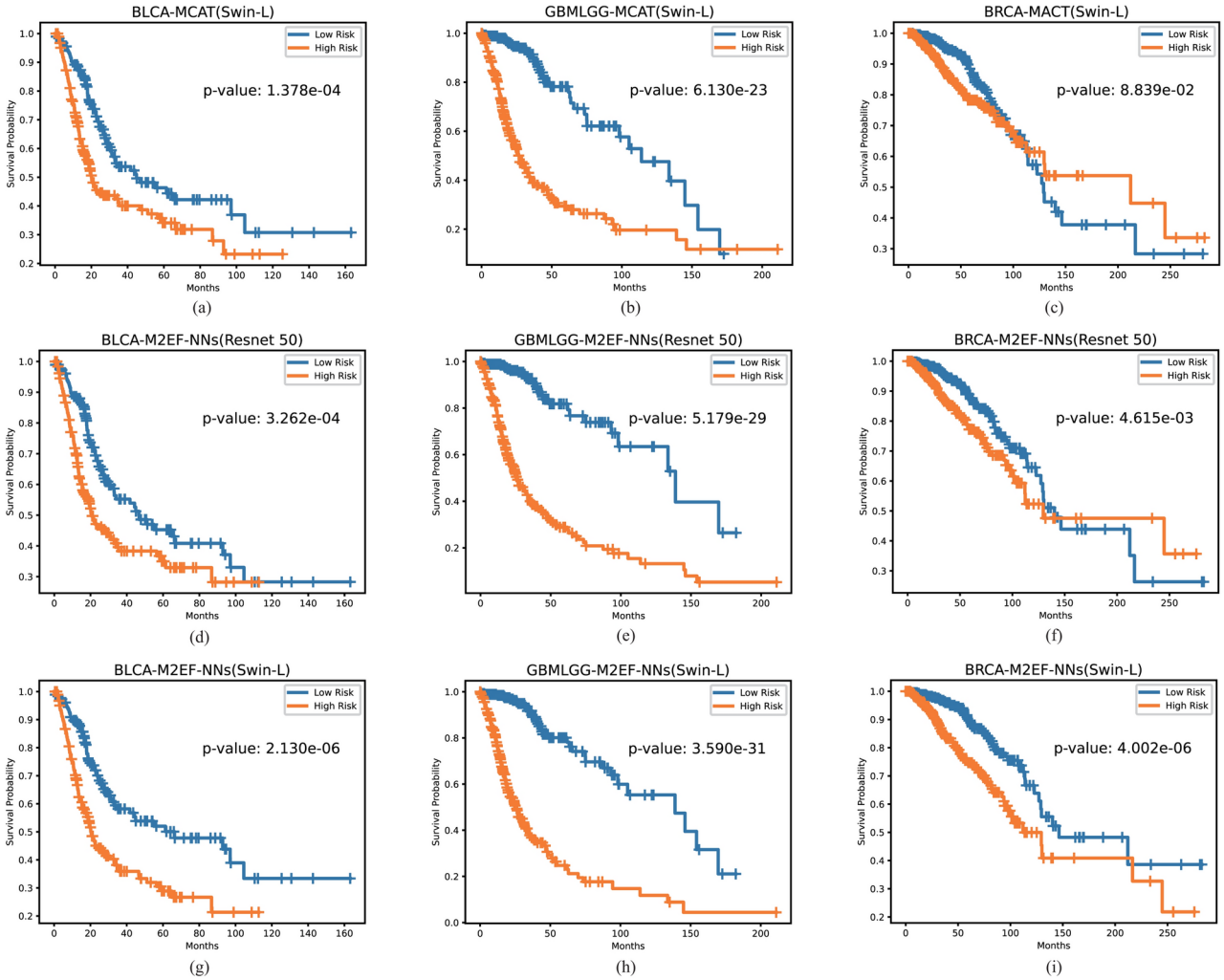


**Fig. 5**. Kaplan–Meier survival curves of ablation study..

tumor and transcriptomic embeddings, regions with high attention weights are localized to aggressive, higher-grade tumor morphologies, such as dense tumor cells and tumor-infiltrated stroma. In the cytokine embeddings, high attention regions are centered around immune cells and tumor cells infiltrating the normal stroma.

These observations indicate that through genomics-guided shared attention heatmaps, the M2EF-NNs method is able to capture some known genotype-phenotype relationships in cancer pathology, aiding in the understanding of cancer development and pathological features. This highly interpretable approach can provide medical researchers and clinicians with deeper insights to support cancer diagnosis, treatment, and prognosis assessment.

## Discussion

In the diagnosis and prognosis assessment of cancer, it is typically necessary to obtain multiple biomarkers to accurately evaluate the disease status and progression stage. Among these biomarkers, histopathological images and genomic data are two primary types. While there have been some studies proposing multimodal survival prediction methods that integrate histopathology images and genomics data, achieving good results, there are two significant limitations: they cannot effectively utilize global context and they disregard the uncertainty

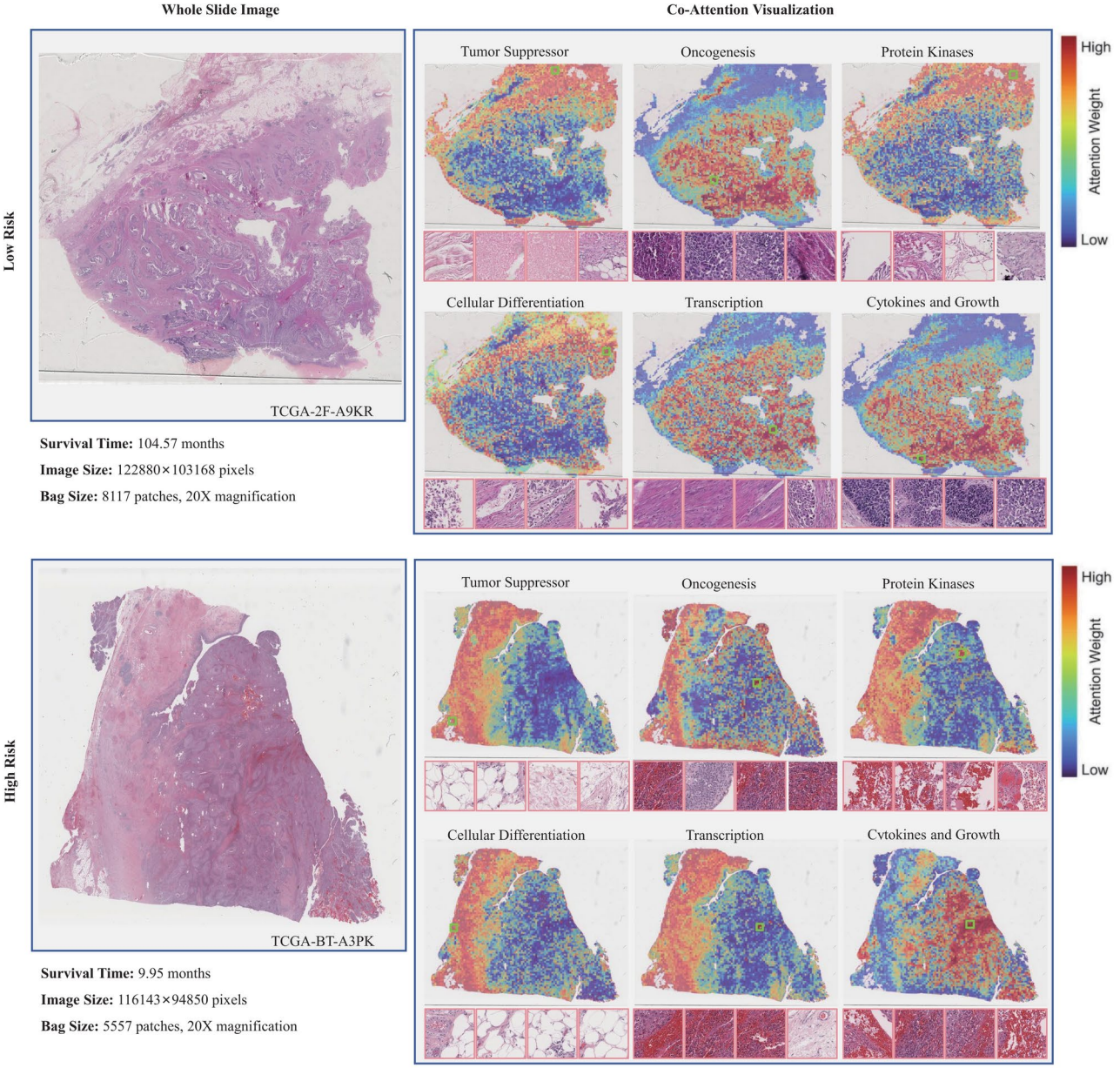


**Fig. 6**. Attention heatmaps with corresponding high attention patches.

of modalities. To extract effective predictive features from these multimodal data for multimodal integration and ensure the reliability of the final decision, we propose a multimodal multi-instance evidence fusion neural networks (M2EF-NNs) framework for cancer survival prediction. Our proposed method is capable of capturing comprehensive global information while considering the uncertainties associated with individual modalities. Specifically, we employ a pre-trained Swin-L model to extract features from histopathology images. Then, using genomic embeddings as queries, we learn the co-attention mapping between the genomic features and histopathology images to achieve early interaction and fusion of multimodal information. Subsequently, we integrate the multimodal information at the evidence level using Dempster–Shafer theory and dynamically adjust the weights of the class probability distribution after multimodal fusion to achieve trusted survival prediction.

In line with similar studies that employ multimodal fusion to enhance supervised learning tasks, M2EF-NNs demonstrates improvements over its respective single-modal counterparts in all benchmark evaluations. This further validates the effectiveness of multimodal data fusion in improving survival prediction performance. Compared to the multimodal methods of DeepSets, Attention MIL, DeepAttnMISL, and MCAT, the M2EF-NNs method exhibits overall performance improvement for multimodal survival prediction. This provides a new computational tool for cancer survival prediction. Additionally, ablation experiments further validate the effectiveness of the pre-trained Swin-L model in capturing representative patch features and the trusted survival predictor based on the DST.

Our study has two limitations. Firstly, it only analyzes pathological images at a single scale to reduce training time costs. WSIs are typically organized in a pyramid structure, with data scanned at different magnification levels stored in different pyramid levels. Pathologists continuously review and analyze pathological images at different magnifications while moving across the microscope or terminal screen, rather than focusing on local or fixed magnifications. Analyzing slices of pathological images at a single scale may not adequately represent the overall heterogeneous microenvironment of the tumor. Secondly, the separation of the WSI feature extraction process from the final prediction stage in M2EF-NNs does not allow for end-to-end training. In future research, we plan to combine data of different resolutions and explore the use of multi-granularity computational methods, such as fuzzy sets[54], clustering[55,56] and feature weighting[57], to perform multiscale feature fusion from coarse to fine granularity in order to capture more comprehensive features. We will also consider using techniques like transfer learning or self-supervised learning to achieve end-to-end learning from input to output, thereby better utilizing existing data.

## Conclusion

In this work, we propose a multimodal multi-instance evidence fusion neural networks (M2EF-NNs) framework for cancer survival prediction. Our proposed method is capable of capturing comprehensive global information while considering the uncertainties associated with individual modalities. Specifically, we employ a pre-trained Swin-L model to extract features from histopathology images. Then, using genomic embeddings as queries, we learn the co-attention mapping between the genomic features and histopathology images to achieve early interaction and fusion of multimodal information. Subsequently, we integrate the multimodal information at the evidence level using Dempster–Shafer theory and dynamically adjust the weights of the class probability distribution after multimodal fusion to achieve trusted survival prediction. The experimental results unequivocally demonstrate a substantial enhancement in cancer survival prediction performance with the proposed method, thereby significantly boosting the model's reliability. Moreover, our approach not only provides powerful computational tools for cancer survival prediction but also advances research in this field. It aims to offer clinicians and patients more accurate and reliable cancer survival predictions, as well as support for personalized treatment decision-making.

## Data availibility

The data used in this study are publicly available through the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/).

Received: 21 November 2024; Accepted: 10 March 2025
Published online: 26 March 2025

## References

1. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer statistics, 2022. *CA Cancer J. Clin.* **72**(1), 7–33 (2022).
2. Liu, P., Fu, B., Ye, F., Yang, R. & Ji, L. DSCA: A dual-stream network with cross-attention on whole-slide image pyramids for cancer prognosis. *Expert Syst. Appl.* **227**, 120280 (2023).
3. Shao, W. et al. Weakly supervised deep ordinal cox model for survival prediction from whole-slide pathological images. *IEEE Trans. Med. Imaging* **40**(12), 3739–3747 (2021).
4. Zhang, J., Guo, X., Wang, B. & Cui, W. Automatic detection of invasive ductal carcinoma based on the fusion of multi-scale residual convolutional neural network and SVM. *IEEE Access* **9**, 40308–40317 (2021).
5. Gallego, O. Nonsurgical treatment of recurrent glioblastoma. *Curr. Oncol.* **22**(4), 273–281 (2015).
6. Shao, W. et al. Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis. *IEEE Trans. Med. Imaging* **39**(1), 99–110 (2019).
7. Li, R., Wu, X., Li, A. & Wang, M. HFBSurv: Hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. *Bioinformatics* **38**(9), 2587–2594 (2022).
8. Wu, X., Shi, Y., Wang, M. & Li, A. CAMR: Cross-aligned multimodal representation learning for cancer survival prediction. *Bioinformatics* **39**(1), btad025 (2023).
9. Shi, Y. et al. MIF: Multi-shot interactive fusion model for cancer survival prediction using pathological image and genomic data. *IEEE J. Biomed. Health Inform.* 1–12 (2024).
10. Lv, Z., Lin, Y., Yan, R., Wang, Y. & Zhang, F. TransSurv: Transformer-based survival analysis model integrating histopathological images and genomic data for colorectal cancer. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **20**(6), 3411–3420 (2022).
11. Xie, Y., Niu, G., Da, Q., Dai, W. & Yang, Y. Survival prediction for gastric cancer via multimodal learning of whole slide images and gene expression. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 1311–1316 (IEEE, 2022).
12. Wang, Z., Li, R., Wang, M. & Li, A. GPDBN: Deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction. *Bioinformatics* **37**(18), 2963–2970 (2021).
13. Chen, R. J. et al. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans. Med. Imaging* **41**(4), 757–770 (2020).
14. Chen, R. J., Lu, M. Y., Weng, W.-H., Chen, T. Y., Williamson, D. F., Manz, T., Shady, M. & Mahmood, F. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF international conference on computer vision* 4015–4025 (2021).
15. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C. & Dosovitskiy, A. Do vision transformers see like convolutional neural networks?. *Adv. Neural Inf. Process. Syst.* **34**, 12116–12128 (2021).
16. Zhang, C., Zhang, M., Zhang, S., Jin, D., Zhou, Q., Cai, Z., Zhao, H., Liu, X. & Liu, Z. Delving deep into the generalization of vision transformers under distribution shifts. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition* 7277–7286 (2022).
17. Yagi, Y. Color standardization and optimization in whole slide imaging. In *Diagnostic pathology* Vol. 6 1–12 (Springer, 2011).
18. Huang, Z. et al. Salmon: Survival analysis learning with multi-omics neural networks on breast cancer. *Front. Genet.* **10**, 166 (2019).
19. Edwards, H. & Storkey, A. Towards a neural statistician. Preprint at arXiv:1606.02185 (2016).
20. Zaheer, M. et al. Deep sets. *Adv. Neural Inf. Process. Syst.* **30**, 3391–3401 (2017).
21. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In *International conference on machine learning* 2127–2136 (PMLR, 2018).

22. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N. & Huang, J. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Med. Image Anal.* **65**, 101789 (2020).
23. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
24. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. Preprint at arXiv:2010.11929 (2020).
25. Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T. & Veit, A. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision* 10231–10241 (2021).
26. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. & Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* 9650–9660 (2021).
27. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* 10012–10022 (2021).
28. Luo, J. et al. DCA-DAFFNet: An end-to-end network with deformable fusion attention and deep adaptive feature fusion for laryngeal tumor grading from histopathology images. *IEEE Trans. Instrum. Meas.* **72**, 1–15 (2023).
29. Wang, J. et al. Nuclei instance segmentation using a transformer-based graph convolutional network and contextual information augmentation. *Comput. Biol. Med.* **167**, 107622 (2023).
30. Gao, Z. et al. in *24th International Conference, Strasbourg, France, Proceedings, Part VIII 24* 299–308 (Springer, 2021).
31. Chen, H. et al. GasHis-transformer: A multi-scale visual transformer approach for gastric histopathological image detection. *Pattern Recogn.* **130**, 108827 (2022).
32. Huang, S.-K., Yu, Y.-T., Huang, C.-R. & Cheng, H.-C. Cross-scale fusion transformer for histopathological image classification. *IEEE J. Biomed. Health Inform.* **28**(1), 297–308 (2023).
33. Li, C., Zhu, X., Yao, J. & Huang, J. Hierarchical transformer for survival prediction using multimodality whole slide images and genomics. In *26th international conference on pattern recognition (ICPR)* 4256–4262 (IEEE, 2022).
34. Dempster, A. P. Upper and lower probabilities induced by a multivalued mapping. In *Classic works of the Dempster–Shafer theory of belief functions* 57–72 (Springer, 2008).
35. Dempster, A. P. A generalization of Bayesian inference. *J. R. Stat. Soc.: Ser. B (Methodol.)* **30**(2), 205–232 (1968).
36. Shafer, G. *A mathematical theory of evidence* (Princeton University Press, 1976).
37. Jøsang, A. & Hankin, R. Interpretation and fusion of hyper opinions in subjective logic. In *2012 15th International Conference on Information Fusion* 1225–1232 (IEEE, 2012).
38. Zhu, C., Qin, B., Xiao, F., Cao, Z. & Pandey, H. M. A fuzzy preference-based Dempster–Shafer evidence theory for decision fusion. *Inf. Sci.* **570**, 306–322 (2021).
39. Peñafiel, S., Baloian, N., Sanson, H. & Pino, J. A. Applying Dempster–Shafer theory for developing a flexible, accurate and interpretable classifier. *Expert Syst. Appl.* **148**, 113262 (2020).
40. Zhang, L., Xiao, F. & Cao, Z. Multi-channel EEG signals classification via CNN and multi-head self-attention on evidence theory. *Inf. Sci.* **642**, 119107 (2023).
41. Denoeux, T. A neural network classifier based on Dempster–Shafer theory. *IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum.* **30**(2), 131–150 (2000).
42. Sensoy, M., Kaplan, L. & Kandemir, M. Evidential deep learning to quantify classification uncertainty. *Adv. Neural Inf. Process. Syst.* **31**, 3183–3193 (2018).
43. Gao, B., Zhou, Q. & Deng, Y. BIM-AFA: Belief information measure-based attribute fusion approach in improving the quality of uncertain data. *Inf. Sci.* **608**, 950–969 (2022).
44. Ghesu, F. C. et al. in *22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22* 676–684 (Springer, 2019).
45. Tardy, M., Scheffer, B. & Mateus, D. Uncertainty measurements for the reliable classification of mammograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 495–503 (Springer, 2019).
46. Han, Z., Zhang, C., Fu, H. & Zhou, J. T. Trusted multi-view classification with dynamic evidential fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(2), 2551–2566 (2022).
47. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**(6), 555–570 (2021).
48. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**(6), 417–425 (2015).
49. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. *et al.*, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profile. In *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, 15545–15550 (2005).
50. Zadeh, S. G. & Schmid, M. Bias in cross-entropy-based training of deep survival networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(9), 3126–3137 (2020).
51. Bland, J. M. & Altman, D. G. The logrank test. *BMJ* **328**(7447), 1073 (2004).
52. Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**(282), 457–481 (1958).
53. Klambauer, G., Unterthiner, T., Mayr, A. & Hochreiter, S. Self-normalizing neural networks. *Adv. Neural Inf. Process. Syst.* **30**, 120 (2017).
54. Ding, W. et al. FMDNN: A fuzzy-guided multi-granular deep neural network for histopathological image classification. *IEEE Trans. Fuzzy Syst.* **32**, 4709–4723 (2024).
55. Oskouei, A. G., Abdolmaleki, N., Bouyer, A., Arasteh, B. & Shirini, K. Efficient superpixel-based brain MRI segmentation using multi-scale morphological gradient reconstruction and quantum clustering. *Biomed. Signal Process. Control* **100**, 107063 (2025).
56. Oskouei, A. G., Hashemzadeh, M., Asheghi, B. & Balafar, M. A. CGFFCM: Cluster-weight and group-local feature-weight learning in fuzzy c-means clustering algorithm for color image segmentation. *Appl. Soft Comput.* **113**, 108005 (2021).
57. Oskouei, A. G., Balafar, M. A. & Akan, T. A brain MRI segmentation method using feature weighting and a combination of efficient visual features. In *Applied computer vision and soft computing with interpretable AI* 15–34 (Chapman and Hall/CRC, 2024).


## Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under Grant 61976120, Grant 62006128, Grant 62371261, Grant 62102199 and in part by the Youth Science Foundation of Guangxi Medical University under Grant GXMUYSF202432.

## Author contributions

All authors contributed to the study conception and design. H.L. developed and implemented the model and performed the data analyses. T.Z implemented the Python package and visualization. J.H. and H.J. supervised experimental design. W.D. supervised the project and offered funding. The first draft of the manuscript was written by H.L and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. All authors reviewed the manuscript.

## Declarations

### Competing interests
The authors declare no competing interests.

### Additional information
**Correspondence** and requests for materials should be addressed to W.D.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.