frontiers in Medicine

ORIGINAL RESEARCH
published: 05 January 2022
doi: 10.3389/fmed.2021.775587

MPMR: Multi-Scale Feature and Probability Map for Melanoma Recognition

Dong Zhang[1,2†], Hongcheng Han[1,3†], Shaoyi Du[1†], Longfei Zhu[4], Jing Yang[3], Xijing Wang[1], Lin Wang[5] and Meifeng Xu[4*]*

[1] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China, [2] School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an, China, [3] School of Software Engineering, Xi'an Jiaotong University, Xi'an, China, [4] Dermatology Department, Second Affiliated Hospital of Xi'an Jiaotong University (Xibei Hospital), Xi'an, China, [5] School of Information Science and Technology, Northwest University, Xi'an, China

Malignant melanoma (MM) recognition in whole-slide images (WSIs) is challenging due to the huge image size of billions of pixels and complex visual characteristics. We propose a novel automatic melanoma recognition method based on the multi-scale features and probability map, named MPMR. First, we introduce the idea of breaking up the WSI into patches to overcome the difficult-to-calculate problem of WSIs with huge sizes. Second, to obtain and visualize the recognition result of MM tissues in WSIs, a probability mapping method is proposed to generate the mask based on predicted categories, confidence probabilities, and location information of patches. Third, considering that the pathological features related to melanoma are at different scales, such as tissue, cell, and nucleus, and to enhance the representation of multi-scale features is important for melanoma recognition, we construct a multi-scale feature fusion architecture by additional branch paths and shortcut connections, which extracts the enriched lesion features from low-level features containing more detail information and high-level features containing more semantic information. Fourth, to improve the extraction feature of the irregular-shaped lesion and focus on essential features, we reconstructed the residual blocks by a deformable convolution and channel attention mechanism, which further reduces information redundancy and noisy features. The experimental results demonstrate that the proposed method outperforms the compared algorithms, and it has a potential for practical applications in clinical diagnosis.

Keywords: malignant melanoma, whole slide image, multi-scale feature, probability map, neural networks

OPEN ACCESS

Edited by:
Jun Feng,
Northwest University, China

Reviewed by:
Qiang Yan,
Taiyuan University of Technology, China
Jialin Peng,
Huaqiao University, China

****Correspondence:***
Lin Wang
wanglin@nwu.edu.cn
Meifeng Xu
xumf96@163.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Precision Medicine, a section of the journal Frontiers in Medicine

Received: *14 September 2021*
Accepted: *08 December 2021*
Published: *05 January 2022*

Citation:
Zhang D, Han H, Du S, Zhu L, Yang J, Wang X, Wang L and Xu M (2022) MPMR: Multi-Scale Feature and Probability Map for Melanoma Recognition. Front. Med. 8:775587. doi: 10.3389/fmed.2021.775587

1. INTRODUCTION

Malignant melanoma (MM) is a highly aggressive form of skin cancer whose incidence continues to increase at a great rate worldwide (1). It is characterized by an extraordinary metastasis capacity and chemotherapy resistance, and the difficulty of effective treatment increases with its continually developing aggression. Therefore, early diagnosis is essential to improve the survival rate of MM patients. Pathological examination is the gold standard for the diagnosis of MM (2), which enables the most reliable diagnosis based on pathological features at the cell level compared to other methods. Tissue cut from the lesion on the skin is made into pathological slices and scanned by a Digital Pathology Microscope Slide Scanner to get a whole-slide image (WSI). Through the WSI,

the pathologist finds out the property of the tissue and marks the MM region, if it exists, to measure related pathological indicators, such as lesion size, invasion depth, etc., which provide an important reference for treatment planning and surgical prognosis (3).

Analyzing WSIs is a challenging task (4). Even an experienced pathologist spends an average of 10-20 min recognizing the region of MM in a WSI, of which identifying the MM region takes up much time. First, a WSI has billions of pixels, and the physician needs to perform a scanned screening of the pathology images in a zoomed-in window. Second, the complex visual characteristics of the skin lesions, such as irregular-shaped texture, fuzzy boundaries, etc., increase the difficulty of recognition. Some MM tissues are hard to distinguish from some benign tissues (5), which is a challenge for MM recognition. These problems aggravate the work burden of pathologists, affecting the efficiency of pathological examination. Third, the difficulty in training and scarcity of pathologists, as well as the uneven distribution of medical resources, make it difficult to obtain a timely and accurate diagnosis for every melanoma patient. Therefore, there is an urgent need for an effective method for automatic MM recognition in WSIs.

MM region screening in WSIs is an image recognition task that utilizes computer vision. Since convolutional neural networks (CNNs) have provided state-of-the-art image classification and segmentation performance, medical image analysis methods based on CNNs have been developed. The U-Net proposed by Ronneberger et al. (6) and its derivative improved networks (7–10) have achieved considerable success in medical image segmentation in recent years. However, pixel-wise image segmentation methods have limitations in MM region recognition in WSIs. The huge size of WSIs poses problems to the computation of the network. Some MM recognition methods based on deep learning are proposed. For example, Hekler et al. (11) trained a CNN based on ResNet-50 (12) to realize the classification of histopathological images of melanomas and nevi with an accuracy of 81%. The limited feature extraction capabilities of ResNet make it challenging to achieve higher accuracy. Wang et al. (13) used a deep CNN to establish a diagnosis model through the patch of eyelid melanoma histopathological slides and obtained good results. Yu et al. (14) proposed a method for melanoma recognition by leveraging very deep CNNs and constructed a fully convolutional residual network for accurate MM segmentation. However, it applies only to dermoscopy images analysis, which is easier to realize but not as reliable and detailed as pathological analysis.

However, these methods only work for the region of interest marked by pathologists. They cannot achieve good results in WSIs. The huge number of pixels makes network training difficult or impossible. Resizing images by down-sampling will lead to the loss of detailed information, which is unacceptable for MM diagnosis focusing on pathological features at the cellular level. Furthermore, due to the characteristics of WSIs and the limited feature extraction capability of related networks, the existing methods are difficult to adapt for WSIs-based MM recognition.

Based on the above considerations, we proposed a novel MM recognition method based on a multi-scale feature representation and probability map to recognize the MM tissue region in WSIs, as shown in **Figure 1**. The following contributions are made to our work.

- Aiming at the difficult and inaccurate problems of recognizing the enormous size of WSIs, the breaking up the whole into parts idea is introduced to recognize melanoma. Furthermore, using predicted results and probabilities generates the mosaic-style mask and lesion region.
- To take both global and local features, we propose an efficient multi-scale network for improving melanoma recognition, combining high-level features with more semantic information and low-level features with more detail information. A multi-scale sliding cropping operation is used to obtain patch and sub-patch images.
- To enhance the feature representation of irregular-shaped lesions, highlight the critical features, and reduce the impact of information redundancy and data noise, we reconstruct the residual block by deformable convolution and channel attention.

The paper is arranged as follows. Section 2 details the proposed method, including the description of our method's framework, the realization principles, and the equations of each module. Section 3 shows the experimental results of our method, compares algorithms on an available WSI dataset, and further provides the ablation analysis to prove the effectiveness and rationality of the proposed method. Section 4 provides a further discussion on the feature representation capability of the proposed multi-scale network. And section 5 provides a brief summary and the conclusions of this work.

2. METHODOLOGY

2.1. Framework of Our Method

On super-large WSI images, patch-based recognition is necessary and feasible. Melanoma pathological analysis mainly focuses on cell-scale characteristics. We set patch size according to pathologists' professional advice, which ensures that cell morphology and local distribution are well represented in the patches. On the boundary of the patch, some cells may be torn, but the overlapping sampling method can effectively avoid the loss of information caused by incomplete splitting. For lesion areas without clear boundaries, mixed cell tissue limits feature extraction by conventional rectangular convolution. Therefore, the proposed method reconstructed the residual block by deformable convolution and channel attention to overcome the irregular-shaped lesion and focus on important features. Furthermore, to overcome the influence of cell-scale differences, we built multi-scale feature fusion layers to enhance feature information and improve identification accuracy. The framework of the proposed method shown in **Figure 2** consists of the following seven components: patch processing, feature extraction, feature fusion, feature selection, predictive classification, mask generation, and loss function in training.

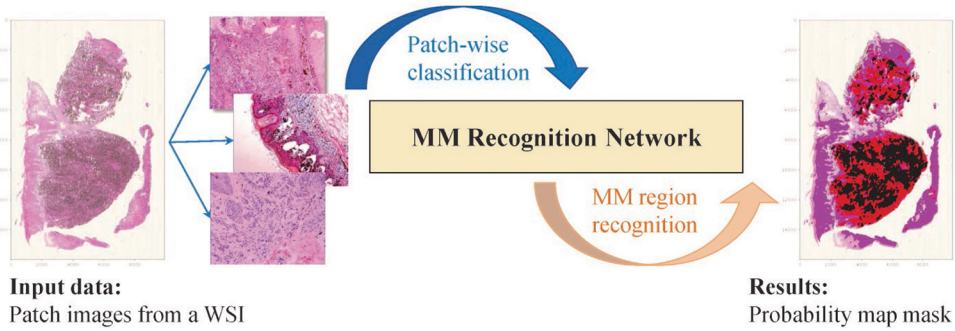

FIGURE 1 | The idea of the proposed method. The WSI is broken up into patch images. The classification results are aggregated and generate the mosaic-type mask based on the probability map, and the MM regions are segmented based on a settable probability threshold.

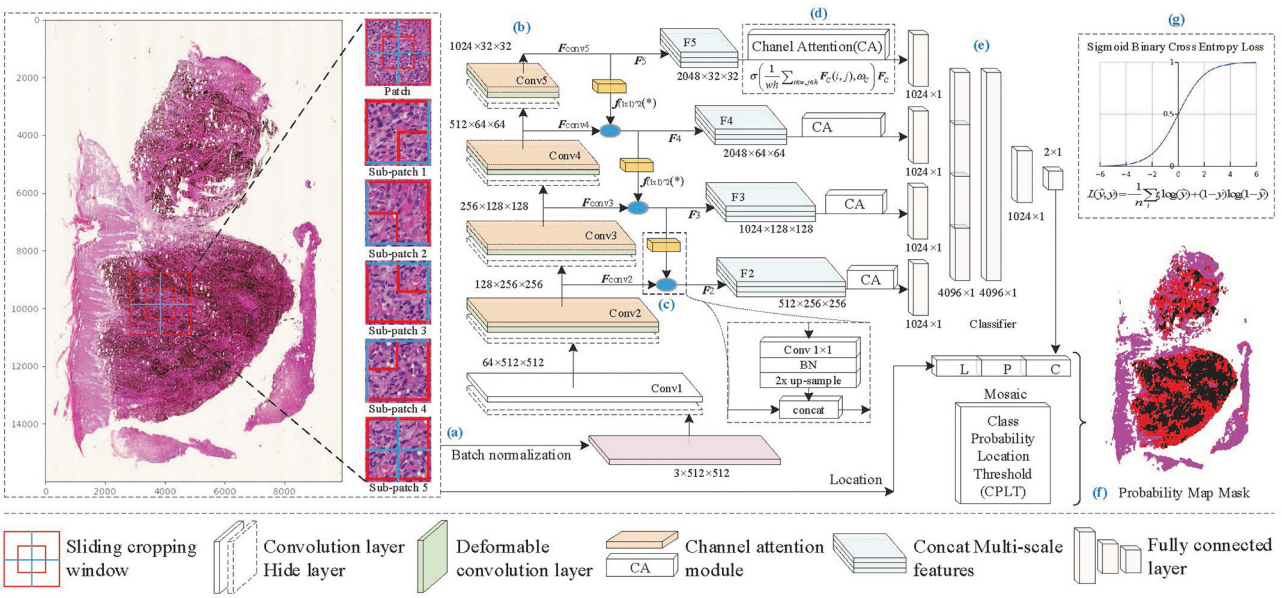

FIGURE 2 | The framework of the proposed method. **(A)** Patch processing, **(B)** feature extraction, **(C)** feature fusion, **(D)** feature selection, **(E)** predictive classification, **(F)** probability map generation, and **(G)** loss function of training model.

- **Patch processing**: A WSI is broken up into N patches through sliding cropping (N depends on the sliding window size and sliding stride), from which tissue-contained patches are picked out by a color analysis method. Each tissue-contained patch is broken up into sub-patches. And then, patch images and sub-patch images are normalized to a uniform size.
- **Feature extraction**: The lesion features are extracted by backbone Conv1 to Conv5. Considering the irregular-shaped cells, and focusing on essential features, deformable convolution (DC) and channel attention (CA) operations are embedded in the Conv2 to Conv5 layers to enhance the feature extraction capability of the network. Then extracted features $\boldsymbol{F}_{convi}(i = 2, 3, 4, 5)$ are produced separately from Conv2 to Conv5.
- **Feature fusion**: As the network is gradually deepened, the resolution of the feature map decreases, and the semantic properties of the features are enhanced. The features of the next layer, which contains richer semantic information, are concatenated with those of the current layer, which contains richer detailed texture information, to enhance the lesion feature representation capability of the network.
- **Feature selection**: After the fused features $\boldsymbol{F}_i(i = 2, 3, 4, 5)$, the channel attention mechanism is separately used to select the critical features and to enhance the correlation between high-level semantic features and low-level detailed features.
- **Predictive classification**: The output features from each branch are flattened into a vector, respectively, and then they are concatenated. Fully connected layers are constructed to obtain the predictive classification results of patch images.

- **Probability map generation**: The prediction results of patches (containing prediction labels and confidence probabilities) are combined with the location information to generate the probability map of malignant tissues. And a mosaic-type mask of MM regions is obtained through a confidence probability threshold.
- **Loss function**: Sigmoid binary cross-entropy loss function is used in training for parameter optimization.

2.2. Multi-Scale Features

In the pathological examination, it is necessary to carry out comprehensive analysis according to various characteristics of lesions, such as tissue morphology and cell distribution, which are reflected on a large scale, and cell morphology and nuclear size, which are reflected on a smaller scale (15). Therefore, computational analysis of WSIs at different scales is beneficial to represent pathological features at different scales. A multi-scale sliding cropping method is embedded in the proposed algorithm. For each patch of a WSI, besides the whole patch image, cropped sub-patch images from the patch are normalized to a uniform size and input into the network together for the classification of the patch. The size and quantity of the sub-patch depend on the cropping method, for example, **Figure 2A** shows five sub-patches cropped from a patch image.

Furthermore, the idea of multi-scale is also reflected in the construction of the feature extraction network. As the network deepens, feature resolution decreases and channels increase, low-level detail information is being transformed into higher-level semantic information. However, factors such as data noise and chain derivative attenuate or lose the information in the forward and back propagation, which becomes more and more apparent with increasing network depth. The fusion of shallow features and deeper features, which are with different scales, to supplement the semantic information of high-level features is beneficial to improve the feature representation capability of the network. Based on the above considerations, a network with enhanced multi-scale feature extraction capability is constructed. As **Figure 2** shows, additional branches are added to the backbone network for feature fusion. In each branch, the feature of $(i+1)$-th level $\boldsymbol{F}_{\text{Conv}i+1}(i=2,3,4)$ is concatenated with the feature of i-th level $\boldsymbol{F}_{\text{Conv}i}(i=2,3,4)$ after 1×1 convolution and up-sampling, as shown in **Figure 2C**, and then $\boldsymbol{F}_i(i=2,3,4)$ is obtained, as shown in Equation (1).

$$\boldsymbol{F}_i = f_{(1\times1)*2}\left(\boldsymbol{F}_{i+1}\right) \oplus \boldsymbol{F}_{\text{Conv}i} \tag{1}$$

where $f_{(1\times1)*2}(*)$ indicates that the $\boldsymbol{F}_i$ is obtained by the convolution of 1×1 and double up-sampling of features $\boldsymbol{F}_{i+1}$ and has the same shape maps as the i-th conv output features $\boldsymbol{F}_{\text{Conv}i}$. $\bigoplus$ denotes the concatenation of the normalized features of the two groups.

The concatenated features $\boldsymbol{F}_2 \sim \boldsymbol{F}_4$, together with $\boldsymbol{F}_5$, which is obtained from $\boldsymbol{F}_{\text{Conv5}}$, are transferred to feature vectors and input into the fully connected layer via shortcut connections for classification.

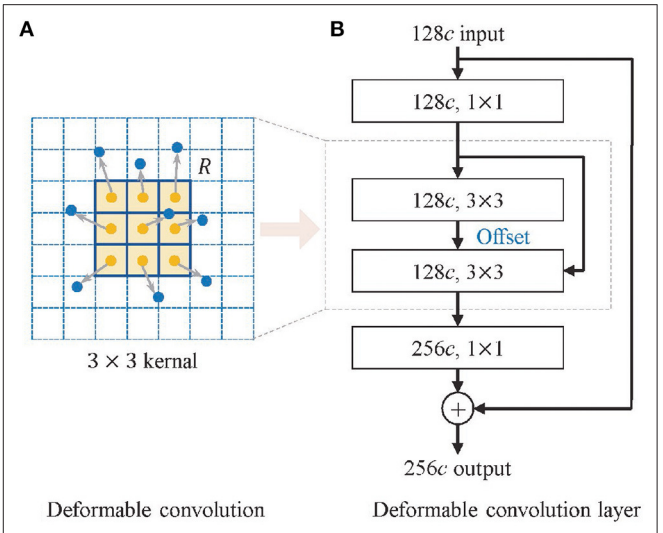

FIGURE 3 | Structure of deformable convolution module. **(A)** The deformable convolution and **(B)** the deformable convolution layer.

2.3. Deformable Convolution

MM tissues in histopathological images mainly show as interstitial or heterogeneous tumor cells, which are mainly enlarged and darkly stained nuclei with varying shapes (16). This irregularity leads to the inadequate learning of melanoma feature information by traditional convolution for its fixed rectangular receptive field of the kernel. Inspired by Dai et al. (17) and Zhu et al. (18), we introduce offsets in the traditional convolution to make the geometry of the kernel more flexible, as shown in **Figure 3**, which improves the representation of irregular-shaped features. The deformable convolution format for each position p in the input feature map is shown in Equation (2).

$$\boldsymbol{y}(p) = \sum_{p_k \in \boldsymbol{R}} \boldsymbol{w}\left(p_k\right) \cdot \boldsymbol{x}\left(p + p_k + \Delta p_k\right) \tag{2}$$

where $\boldsymbol{y}(p)$ indicates the feature obtained by the convolution on one sampling point p of the feature map. $\boldsymbol{R}$ is the receptive field size of the regular kernel. p_k denotes the difference between the sampling points and $\boldsymbol{y}(p)$, $k=1,2,3\ldots N, N=|\boldsymbol{R}|$, Δp_k is the learned offset, and $\boldsymbol{w}$ is the kernel parameter. The offset of the deformable convolution has a dilated value, which determines the maximum distance for resampling and is set to 2.

2.4. Channel Attention

First, multi-scale feature fusion enriches the extracted feature information of the network, but while enhancing feature representation capability, it also brings some redundant features, which are unrelated to melanoma recognition, and interferes with model learning. It is particularly obvious in low-level features with higher resolution, and this effect becomes more prominent when low-level features are fused with high-level features through additional branch paths. Second, the deformable convolution helps in the feature extraction of irregular lesions and enhances lesion feature representations, but also generates

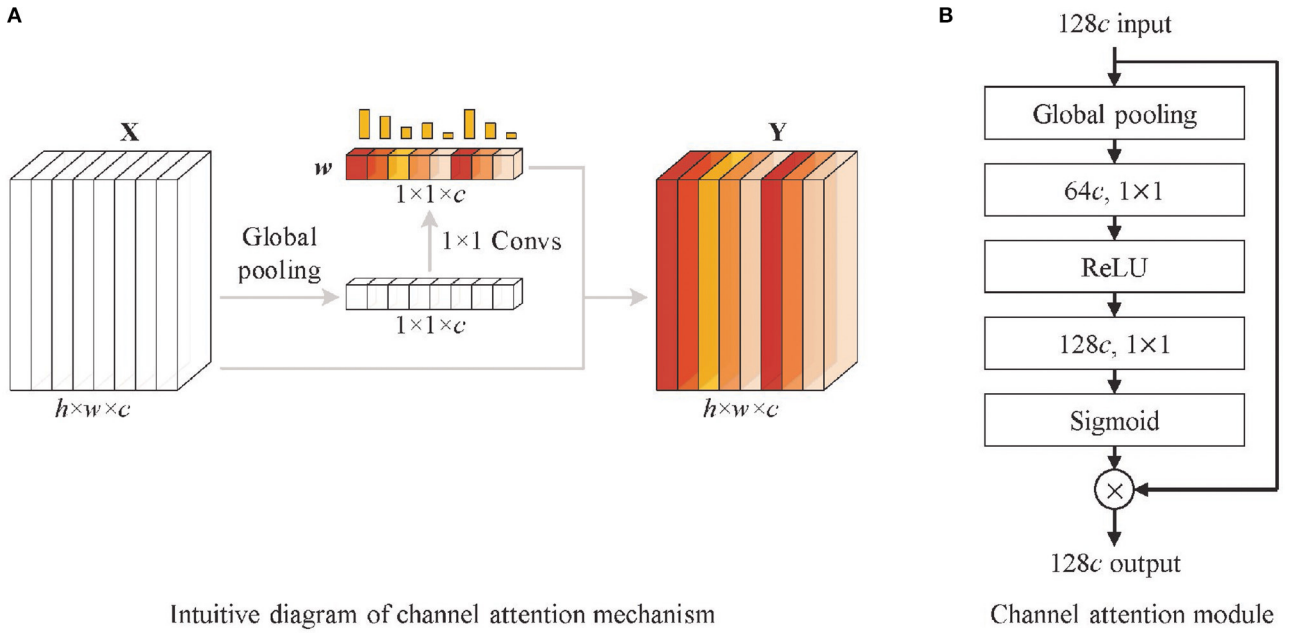

FIGURE 4 | Structure of channel attention module. **(A)** The intuitive diagram of the channel attention mechanism and **(B)** channel attention module. It firstly performs global pooling to the input feature map (128c denotes 128 channels) to obtain the overall representations. Then a weight vector is learned by 1 × 1 convolution operations. The learned weight vector allocates weight coefficients for each channel of the original feature map, and the weighted feature map is output.

some noisy features influenced by non-lesion tissue. Therefore, to extract more valuable information and suppress the impact of redundant information and noises features, we need a mechanism that focuses on essential features and filters the irrelevant features.

Based on the above considerations, and inspired by the work of Hu et al. (19), a channel attention mechanism is used in the shortcut connection between features $\boldsymbol{F}_i(i = 2, 3, 4, 5)$ and fully connected layers, as shown in **Figure 2D**. It highlights high-value feature maps by a series of weights learned by the channel attention (CA) module. The filtering of channels is actually the weighting of different types of features. Although the convolution operation itself also correlates each channel of the feature map with each other, it is difficult to accurately assign appropriate weights to each channel due to the influence of the w and h dimensional feature distributions. To address this problem, the channel attention mechanism obtains a global representation of each channel by global pooling, and the weights of each channel are calculated by 1 × 1 convolution based on the resulting feature vectors. In **Figure 4**, the CA module firstly performs global pooling to the input feature map to obtain the overall representation of it. Then a weight vector is learned by 1 × 1 convolution. The learned weight vector allocates weight coefficients for each channel of the original feature map, and the weighted feature map is output. The mathematic description of the channel attention module is formatted as Equation (3).

$$\boldsymbol{Y} = \boldsymbol{\sigma}\left(\boldsymbol{W}_{Conv2}\boldsymbol{\delta}\left(\boldsymbol{W}_{Conv1}\frac{1}{hw}\sum_{i\in h, j\in w}\boldsymbol{X}\left(i,j\right)\right)\right)\otimes\boldsymbol{X} \quad (3)$$

where $\boldsymbol{X}$ means the input feature map, and $\boldsymbol{Y}$ denotes the output feature map of the channel attention module, h and w are the height and width in the input feature maps. $\boldsymbol{W}_{Conv1}$ and $\boldsymbol{W}_{Conv2}$ indicate the parameters of two 1 × 1 convolution operations, which are equivalent operations to fully connected layers. $\boldsymbol{\delta}$ is the ReLU activation. $\boldsymbol{\sigma}$ is the sigmoid function, and ⊗ means the weighting calculation of the learned weight vector and the input feature map.

In addition to calculating the channel weights of feature weights, the channel attention mechanism also strengthens the correlation between channels through global pooling and 1 x 1 convolution; making up for the defect of the weak correlation between channels in the convolution module is conducive to the enhancement of feature expression ability. Therefore, the channel attention modules are also embedded into the backbone network, as shown in **Figure 2B**.

3. EXPERIMENTS

3.1. Experimental Setup

MM WSIs labeled by pathologists are rare and valuable data. The dataset is collected from the Second Affiliated Hospital of Xi'an Jiaotong University (Xibei Hospital), containing 30 WSIs labeled by experienced pathologists. Sliding window size is set to 1, 024 × 1, 024, sliding stride is set to 1024, 18,698 tissue-included patches are obtained, containing 7,369 malignant tissue patches and 11,329 benign tissue patches. They are divided into training, validation, and test datasets by a ratio of 6:2:2. Five sub-patches are cropped from each patch, as shown in **Figure 2A**, and

TABLE 1 | The experimental results of the proposed algorithms and comparison algorithms, the higher the values of precision, recall, accuracy, and *F*1, the better the recognition performance.

Algorithms	Layer	Precision	Recall	Accuracy	*F*1
Inception V3 (2016)	50	0.8919	0.8876	0.8326	0.8897
ResNeXt (2017)	50	0.8974	0.9241	0.8618	0.9106
SENet (2018)	50	0.8915	0.9472	0.8721	0.9185
SENet (2018)	101	0.9120	0.9477	0.8812	0.9295
ResNeSt (2020)	50	0.9314	0.9327	0.8877	0.9321
ResNeSt (2020)	101	0.9526	0.9601	0.9355	0.9513
MPMR (ours)	50	0.9683	0.9709	0.9498	0.9696
MPMR (ours)	101	**0.9740**	**0.9861**	**0.9553**	**0.9749**

Bold indicate maximum values.

all images are resized to 512 × 512 when input to the network. Considering that melanoma tissue features are non-directional and non-chiral, we introduce data augmentation operations by the mirror and random rotation in the range of $(-90°, +90°)$.

The proposed method is developed by Python 3.6 on Ubuntu18.04, and the hardware is RTX2080-12G with CUDA-10.1. The development libraries include MXNet-1.5, Gluoncv-0.5, Numpy-1.17, OpenCV-4.2, etc. The models iterate 30 epochs, and the batch size is 32. Gradient descent with momentum (20) is used for optimization. We set the momentum to 0.9. The learning rate is 0.001, and the decay rate is 0.99. Both recall (R) and precision (P) for MM recognition are considered in diagnosis, so the evaluation criterion $F1$ score is used to comprehensively measure the performance of the proposed method, which is calculated as Equation (4).

$$F1 = \frac{2 \times P \times R}{P + R} \tag{4}$$

3.2. Results of Patch Classification

In order to verify the effectiveness and recognition performance of the algorithm, the proposed method is compared with some popular algorithms in recent years, including Inception V3 (21), ResNeXt (22), SENet (19), and ResNeSt (23). The experimental results are shown in **Table 1**, the higher the values of $F1$, recall, precision, and accuracy, the better the recognition performance. The $F1$ values of all algorithms exceeded 90%, except Inception V3, and the scores of the proposed method also achieved the best results. SENet and ResNeSt, containing the channel attention module, outperform other comparison algorithms, indicating that the channel attention mechanism improves performance.

The proposed method outperforms all the comparison algorithms for the same number of layers, mainly benefiting from the deformable convolution, the channel attention, and the multi-scale feature fusion. In particular, the learning capability of multi-scale features in the proposed method effectively adapts the different scale samples. It sufficiently learns the feature information of melanoma in the training and validation datasets and has better robustness on the testing dataset. Therefore, the proposed algorithm outperforms other algorithms on the WSI test dataset.

3.3. Results of the Probability Map

The prediction results of patch images containing prediction labels and confidence probabilities are combined with the location information to generate the probability map. The visualization results of a WSI containing malignant melanoma tissues are shown in **Figure 5**. The probability of being predicted as MM tissues is visualized as different colors, from 0 to 1. The threshold of malignant tissues and benign tissues is set to 0.5; red regions display the recognized MM tissues. The prediction results of some difficult samples of different algorithms are compared, and the proposed method provides the most correctly recognized patches, marked by green boxes, while other comparison algorithms provide some incorrect recognition results, marked by red boxes. The results indicate that the proposed method can obtain more accurate recognition results in WSIs.

3.4. Ablation Analyses

To analyze the contributions of multi-scale feature fusion, deformable convolution, and channel attention in the proposed method, ablation analyses are performed for these impacts. The results of ablation analyses are shown in **Tables 2**–**4**, the higher the values of precision, recall, accuracy, and $F1$, the better the recognition performance.

3.4.1. Multi-Scale Features

The proposed method realizes multi-scale feature fusion by constructing additional branch paths and adopting shortcut connections between fused features and the fully connected layers. Low-level features containing more detail information are expected to supplement the semantic features of high-level features for enhancing the classification capability of the model. The experimental results of the networks with different numbers of branch paths are shown in **Table 2**. The more branch paths added, the higher the values of $F1$, recall, precision, and accuracy obtained. It indicates that the prediction method based on multi-scale features helps the proposed method to recognize melanoma. In the results of F5/F4/F3/F2, the accuracy of the proposed method decreases compared to F5/F4/F3, and the other evaluation indicators show weak increases. It indicates that multi-scale feature fusion should be carried out in an

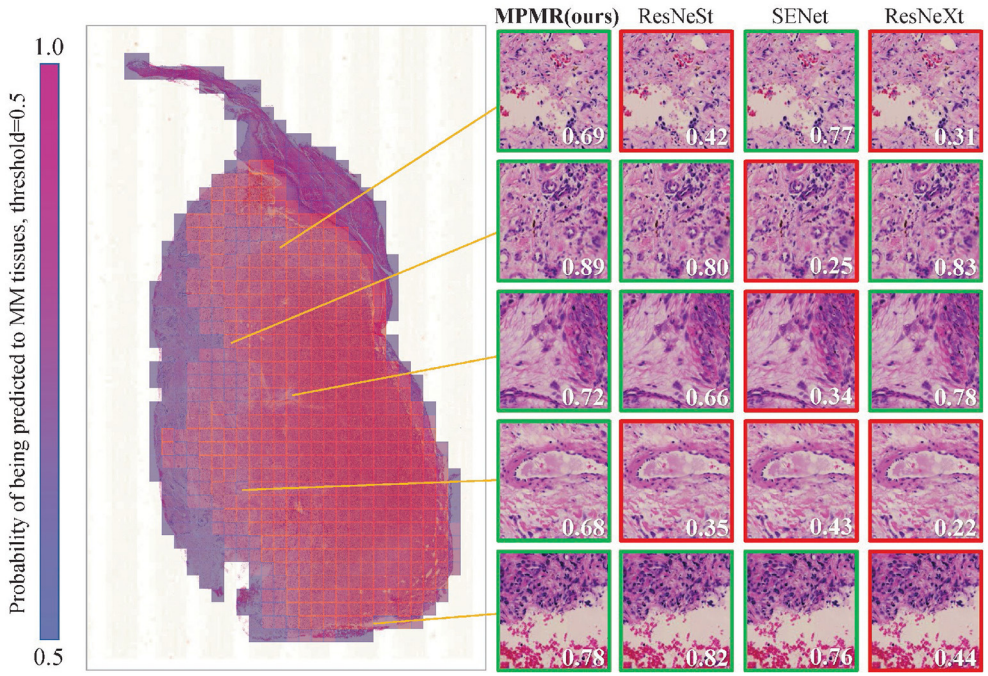

FIGURE 5 | The mosaic-style mask result of a WSI in the test dataset is generated through the probability map obtained through patch image classification. The recognition results of some patches, which are difficult to recognize, through different algorithms are compared. A green box represents a correct prediction, and a red box represents a misclassification. The probabilities of being predicted as MM tissues are marked in each patch image, and the threshold is 0.5.

TABLE 2 | The experimental results of the ablation analysis of multi-scale features, the higher the values of precision, recall, accuracy, and *F*1, the better the recognition performance.

Multi-scale features	Layer	Precision	Recall	Accuracy	*F*1
F5	50	0.9400	0.9457	0.9367	0.9428
F5/F4	50	0.9448	0.9633	0.9493	0.9540
F5/F4/F3	50	0.9590	0.9675	**0.9567**	0.9632
F5/F4/F3/F2	50	**0.9683**	**0.9709**	0.9498	**0.9696**

Bold indicate maximum values.

TABLE 3 | The experimental results of the ablation analysis of deformable convolution, the higher the values of precision, recall, accuracy, and *F*1, the better the recognition performance.

Deformable convolution	Layer	Precision	Recall	Accuracy	*F*1
None	50	0.9380	0.9509	0.9387	0.9444
DConv5	50	0.9396	0.9522	0.9402	0.9458
DConv5/4	50	0.9464	0.9563	0.9462	0.9513
DConv5/4/3	50	0.9604	0.9619	**0.9569**	0.9611
DConv5/4/3/2	50	**0.9683**	**0.9709**	0.9498	**0.9696**

Bold indicate maximum values.

appropriate range, and an excess of fusions will cause information redundancy, which is not conducive to feature representation.

3.4.2. Deformable Convolution

The melanoma characteristics in the pathological images are mainly enlarged and darkly stained nuclei with varying shapes. This irregularity leads to the inadequate learning of melanoma feature information by traditional convolution. Deformable convolution is embedded into the convolution layers of the proposed network to enhance irregular-shaped feature representation ability. The experimental results of the networks with different numbers of deformable convolution layers are shown in **Table 3**, indicating that the more deformable convolution layers embedded, the better the recognition

TABLE 4 | The experimental results of the ablation analysis of channel attention, the higher the values of precision, recall, accuracy, and *F*1, the better the recognition performance.

Channel attention	Layer	Precision	Recall	Accuracy	*F*1
None	50	0.9320	0.9216	0.9222	0.9242
Backbone (B)	50	0.9340	0.9321	0.9256	0.9331
Shortcut (S)	50	0.9472	0.9552	0.9460	0.9512
Both B and S	50	**0.9683**	**0.9709**	**0.9498**	**0.9696**

Bold indicate maximum values.

performance of the proposed method. Accuracy decrease occurs in the results of DConv5/4/3/2. It indicates that too many deformable convolution layers may amplify the impact of noise on learning and influence feature representation.

3.4.3. Channel Attention

Channel attention in the proposed method selectively enhances information-rich features, allowing subsequent processing of the networks to take full advantage of these features and suppress noisy features. The experimental results of channel attention with different numbers of layers are shown in **Table 4**, where the B-case indicates that channel attention modules are only embedded in the backbone network for feature extraction, and the S-case indicates that channel attention modules are used in the shortcut connection between fused features and fully connected layers. The recognition performance of the model is significantly improved after using channel attention modules. However, both the embedded B-case and S-case can obtain the best performance of the proposed method. This further demonstrates that the embedding of channel attention can facilitate positive network learning.

4. DISCUSSION

The pathological features related to melanoma are at different scales, such as tissue, cell, and nucleus, and enhancing the representation of multi-scale features is important for melanoma recognition. From the experimental results, it can be concluded that the residual block based on deformable convolution and multi-scale feature fusion brings considerable performance improvement in the patch-wise classification of WSIs.

The visual features of malignant melanoma and benign nevi tissues are very similar, and the shape of the features is irregular and has uneven distribution, which further increases the difficulty of recognition. When learning the feature space of a histopathological skin image, the traditional convolutional network is limited by its fixed spatial geometric structure, as shown in **Figure 3**, which is not suitable for the irregular shape of lesions and the uneven distribution of melanoma cells. However, the deformable convolution layers effectively avoid the rectangular limitation of traditional convolution sampling. The experimental results in **Table 3** demonstrate that dynamic convolution can better extract the features of tissue images. The performance improvement of deformable convolution on the model grows as the number of layers increases, which introduces extra computational consumption but can be neglected for some tasks with low real-time requirements.

Shortcut connections, also known as skip connections, show considerable advantages in residual networks and U-shaped networks. The residual connections ensemble the feature at different layers through sum operation, and (24) put forward similar views. The connections between the encoder and the decoder in U-shaped networks, through deconvolution and concatenation, realize the fusion of features at different scales. In addition, extra information flows, brought by shortcut connections, provide shorter paths for the transmission of parameters in the forward- and back-propagation, reducing information attenuation. These ideas are embodied in the construction of the proposed networks. The additional branch paths in the proposed network realize multi-scale feature fusion through 1×1 convolution, $2\times$ up-sampling, and concatenation. Another fusion of several fused features is performed through the shortcut connections between the fused features and the fully connected layers. The above operations are expected to enhance feature representation and make contributions to improve MM recognition precision.

In order to further analyze the influence of multi-scale fusion on the quality of features extracted from the network, t-distributed stochastic neighbor embedding (t-SNE) (25), a manifold learning dimensionality reduction method, is used to visualize the features extracted from the network with the different number of feature fusion branch paths. The feature vectors in the second-to-last layer of the full connection layers are transferred from 1,024 dimensions to 2 dimensions and visualized as shown in **Figure 6**. The higher the linear separability of benign features and malignant features after dimensionality reduction, the more beneficial the features extracted from the network are for classification. The dimension reduction results of F5, which represents the features extracted by the network without multi-scale feature fusion, are shown in **Figure 6A**, and some of the benign features are interspersed with the malignant features. The results of F5/F4 shown in **Figure 6B**, which represents the features extracted by the network with one branch path for multi-scale feature fusion, show considerable improvement. **Figures 6C,D** show more improvements, which indicates that the additional branch paths for multi-scale feature fusion improve the quality of the features extracted by the network, enhancing the feature representation for MM recognition, and finally provide more accurate MM recognition results. This is consistent with the experimental results in **Table 2**.

5. CONCLUSIONS

This work proposes a novel automatic MM recognition method in WSI based on multi-scale features and the probability map.

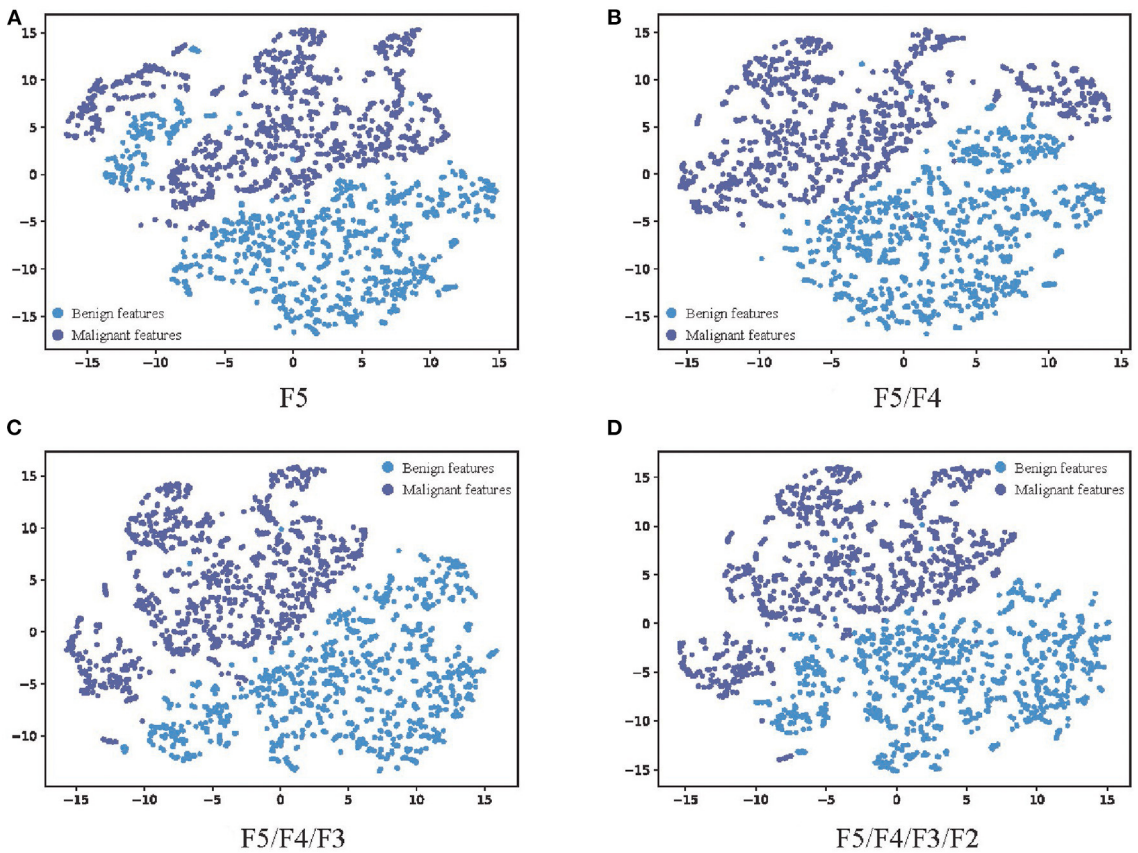

FIGURE 6 | The visualization results of the features extracted by networks with different numbers of feature fusion branch paths. **(A)** F5, **(B)** F5/F4, **(C)** F5/F4/F3, and **(D)** F5/F4/F3/F2. The feature vector in the second-to-last layer of the full connection layers are transferred from 1,024 dimensions to 2 dimensions through t-SNE. The blue scatter plots represent malignant features, and the purple scatter plots represent malignant features. The higher the linear separability of benign features and malignant features after dimensionality reduction, the more beneficial the features extracted from the network are for classification.

The idea that breaking up a WSI into patches and sub-patches through multi-scale sliding cropping solves the difficult-to-calculate problem of WSIs with huge sizes, and the probability map is generated based on the predicted class and confidence probabilities and location information of patch images to visualize the recognition result of MM tissues in WSIs. Additional branch paths and shortcut connections are established for multi-scale feature fusion, which realizes the information supplement of low-level features containing more detail information to deep features containing more semantic information. Deformable convolution operations are embedded into the backbone network to enhance the representation capability of irregular-shaped features in tissues. Channel attention modules are used in the shortcut connection between fused features and fully connected layers, and also the backbone network to highlight the high-value features and reduce the negative impacts of information redundancy caused by additional branch paths.

The results of comparison experiments indicate that the proposed method outperforms Inception V3, ResNeXt, SENet, and ResNeSt. The results of ablation analyses prove the effectiveness of multi-scale feature fusion, deformable convolution, and channel attention modules. Through the proposed method, MM regions in WSIs can be recognized accurately and efficiently, which is a great help to pathological examination and the diagnosis of MM.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

DZ and HH were in charge of experiments and manuscript writing. LZ and MX were responsible for medical analysis and annotation of pathological data. SD and JY provided guidance for method formulation. LW and XW checked the experimental results. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Key Research and Development Program of China under (grant no. 2017YFA0700800), the National Natural Science Foundation of China under (grant no. 61971343), the Shaanxi Provincial Social Science Fund under (grant no. 2021K014), and the Key Research and Development Program of Shaanxi Province of China under (grant no. 2020GXLH-Y-008).

REFERENCES

1. Palacios-Ferrer JL, García-Ortega MB, Gallardo-Gómez M, García MÁ, Díaz C, Boulaiz H, et al. Metabolomic profile of cancer stem cell-derived exosomes from patients with malignant melanoma. *Mol Oncol.* (2021) 15:407–28. doi: 10.1002/1878-0261.12823
2. Shoo BA, Sagebiel RW, Kashani-Sabet M. Discordance in the histopathologic diagnosis of melanoma at a melanoma referral center. *J Amer Acad Dermatol.* (2010) 62:751–6. doi: 10.1016/j.jaad.2009.09.043
3. Scolyer RA, Rawson RV, Gershenwald JE, Ferguson PM, Prieto VG. Melanoma pathology reporting and staging. *Mod Pathol.* (2020) 33:15–24. doi: 10.1038/s41379-019-0402-x
4. Lu C, Mandal M. Automated analysis and diagnosis of skin melanoma on whole slide histopathological images. *Pattern Recognit.* (2015) 48:2738–50. doi: 10.1016/j.patcog.2015.02.023
5. Khan MQ, Hussain A, Rehman SU, Khan U, Maqsood M, Mehmood K, et al. Classification of melanoma and nevus in digital images for diagnosis of skin cancer. *IEEE Access* (2019) 7:90132–44. doi: 10.1109/ACCESS.2019.2926837
6. Ronneberger O, Fischer P, Brox T. "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. (Springer) (2015). p. 234–41.
7. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. "Unet++: a nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer (2018). p. 3–11.
8. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention u-net: learning where to look for the pancreas. *arXiv preprint* arXiv:180403999. (2018).
9. Li C, Tan Y, Chen W, Luo X, Gao Y, Jia X, et al. "Attention Unet++: a nested attention-aware U-Net for liver CT image segmentation," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE (2020). p. 345–9.
10. Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, et al. "Unet 3+: a full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE (2020). p. 1055–9.
11. Hekler A, Utikal JS, Enk AH, Berking C, Klode J, Schadendorf D, et al. Pathologist-level classification of histopathological melanoma images with deep neural networks. *Eur J Cancer* (2019) 115:79–83. doi: 10.1016/j.ejca.2019.04.021
12. He K, Zhang X, Ren S, Sun J. "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2016). p. 770–8.
13. Wang L, Ding L, Liu Z, Sun L, Chen L, Jia R, et al. Automated identification of malignancy in whole-slide pathological images: identification of eyelid malignant melanoma in gigapixel pathological slides using deep learning. *Brit J Ophthalmol.* (2020) 104:318–23. doi: 10.1136/bjophthalmol-2018-313706
14. Yu Z, Jiang X, Zhou F, Qin J, Ni D, Chen S, et al. Melanoma recognition in dermoscopy images via aggregated deep convolutional features. *IEEE Trans Biomed Eng.* (2018) 66:1006–16. doi: 10.1109/TBME.2018.2866166
15. Elmore JG, Barnhill RL, Elder DE, Longton GM, Pepe MS, Reisch LM, et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. *BMJ* (2017) 357:j2813. doi: 10.1136/bmj.j2813
16. Chopra A, Sharma R, Rao UN. Pathology of melanoma. *Surgical Clin.* (2020) 100:43–59. doi: 10.1016/j.suc.2019.09.004
17. Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, et al. "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*. (2017). p. 764–73.
18. Zhu X, Hu H, Lin S, Dai J. "Deformable convnets v2: more deformable, better results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2019). p. 9308–16.
19. Hu J, Shen L, Sun G. "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2018). p. 7132–41.
20. Qian N. On the momentum term in gradient descent learning algorithms. *Neural Netw.* (1999) 1:145–51. doi: 10.1016/S0893-6080(98)00116-6
21. Xia X, Xu C, Nan B. "Inception-v3 for flower classification," in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*. (IEEE) (2017). p. 783–7.
22. Xie S, Girshick R, Dollár P, Tu Z, He K. "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2017). p. 1492–500.
23. Zhang H, Wu C, Zhang Z, Zhu Y, Lin H, Zhang Z, et al. Resnest: split-attention networks. *arXiv preprint* arXiv:200408955. (2020).
24. Veit A, Wilber MJ, Belongie S. Residual networks behave like ensembles of relatively shallow networks. *Adv Neural Inf Process Syst.* (2016) 29:550–8.
25. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res.* (2008) 9:2579–605.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.