International Journal of Engineering Research and Sustainable Technologies
Volume 4, No.2, June2026, P 1- 7
ISSN: 2584-1394 (Online version)

CROSS-MODEL SPATIAL-SEMANTIC FUSION: INTEGRATING MRI AND HISTOPATHOLOGY FOR AUTOMATED BONE-SARCOMA SEVERITY ASSESSMENT

[1]P J.Adit, [2]Dr.C.Priya
[1]Research Scholar, [2]Professor and Research Supervisor
[1]Department of CSE, [2]Department of Computer Applications, [1,2]Dr.M.G.R.Educational and Research Institute, India
*** Corresponding author email address**:pjadit1802@gmail.com

DOI: https://doi.org/10.63458/ijerst.v4i2.154

Abstract

The accurate diagnosis and severity staging of bone-sarcoma inherently depend on synthesizing macroscopic radiological imaging (MRI) with microscopic histopathological analysis. However, contemporary deep learning frameworks process these modalities in strict isolation, creating a diagnostic bottleneck that limits clinical utility. In this paper, we propose the Cross-Modal Spatial-Semantic Diffusion Network (CM-SSD), a novel multi-modal fusion architecture. CM-SSD synchronously processes macroscopic tumor dissemination from MRI alongside microscopic malignancy patterns from Hematoxylin and Eosin (H&E) slides. By leveraging dual spatial-semantic encoders and a cross-modal attention mechanism, the framework adaptively aligns complementary features across both modalities. This integrated approach not only enhances automated tumor localization but also generates a quantitative, percentage-based severity score. By bridging the gap between radiological and pathological computer-aided diagnostics, CM-SSD provides a unified, highly robust framework for real-time clinical decision support.

Keywords: Clinical Data Integration, Cross-Modal Learning, Deep Learning, Histopathology Imaging, Multimodal Diagnosis, Bone-Sarcoma

1. Introduction

Bone sarcoma is a very aggressive primary cancer that grows very quickly and has varied and complicated forms. For the doctors to assess how bad the case is and how to treat it, they usually have two different methods; they use MRI (Magnetic Resonance Imaging) to see how far it has spread and to find out where its boundaries are, and then they look at the biopsy slides (usually stained with H&E) to find out how malignant the cells are and how fast they are dividing by looking at the number of mitotic cells. In recent years, with the development of deep learning (especially CNNs - Convolutional Neural Networks and attention-based models), there has been enormous improvement in the automation of medical image analysis. We now have highly specialized models that perform extremely well on individual tasks, such as pixel-wise segmentation of tumor boundaries in histopathology, or predicting outcomes based on MRI radiomics. However, to date, diagnostic workflows continue to exist in multiple "silos"; that is, either MRI or histopathology results are evaluated independently in order for the physician to manually integrate the two disparate datasets generated by artificial intelligence. This lack of multi-modal integration results in a limitation to the ability to automate the quantification of overall tumor burden, leaving a significant void in the ways that CAD systems assist with diagnostics.The solution to this problem is the Cross-Modal Spatial or Semantic Diffusion Network (CM-SSD). CM-SSD represents a significant shift in paradigm between using only one modality for predicting disease severity versus multiple modalities. CM-SSD has two parallel encoders that extract essential information simultaneously from both an MRI (structural geometry) and a histopathology (malignancy); each encoder's output use a dynamical cross-modality attention mechanism to weight/cross-align the respective outputs, while suppressing non-diagnostic (background) information and enhancing the intersectional diagnostic relevance of features among modalities.

2. System Methodology

The Cross-Modal Spatial-Semantic (CM-SSD) is a novel system that will evaluate large areas of the body for disease with regard to size; however, it will also evaluate small amounts of tissue at a cellular level for malignancy at the same time. The composition of CM-SSD consists of three elements: Dual-Stream Feature Extraction, Cross-Modal Attention Fusion and Quantitative Severity Estimation.

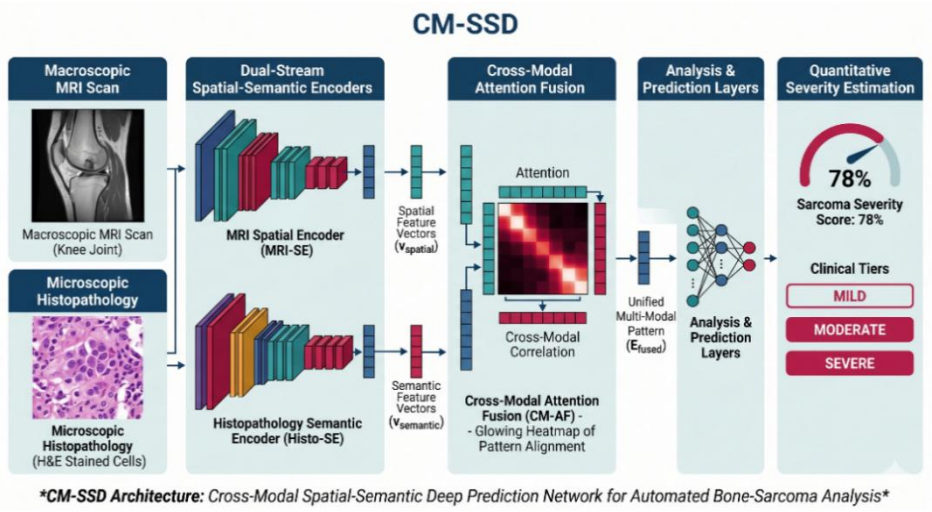

Fig 1. Proposed Methodology

2.1. Modality-Specific Pre-processing Pipeline

In order to provide more consistency of model performance across imaging modalities, rigorous standardization of microscopic or H&E histopathology slides and macroscopic MRI scans is performed prior to entering into the dual-stream encoders. For microscopic H&E histopathology slides, the goal of Macenko stain normalization is to reduce color distribution variability due to differences in laboratory staining protocols. For macroscopic MRI scans, N4 bias field and z-score intensity normalizations will be employed to standardize contrast and minimize the impact of magnetic field in homogeneities on clinical scanners across laboratories.

2.2. Dual-Stream Feature Extraction

Two inputs are fed into the network: a macroscopic MRI scan and a corresponding microscopic H&E histopathology slide. As the two modalities have different spatial resolutions and channel depths, we developed two independent spatial-semantic encoders. The MRI encoder extracts the structural geometry of the tumor and how it has spread, while the histopathology encoder extracts cellular malignancy features like mitotic activity and osteoid matrix formation. The outputs from both encoders— feature maps—are modality-specific:

$$F_{MRI} = \Phi_{MRI}(I_{MRI})$$
$$F_{H\&E} = \Phi_{H\&E}(I_{H\&E}) \quad (1)$$

2.3. Cross-Modal Attention and Fusion

In multi-modal learning, there exists a significant semantic gap between how radiology and pathology are represented. CM-SSD proposes the Cross-Modal Attention Layer, using a dynamic Cross-Modal Attention mechanism to bridge the two modalities. The feature maps extracted by the two modalities within the network are concatenated, and then a learnable convolutional filter is used to produce a single attention weight matrix.

$$\alpha_{cross} = \sigma(W_{cross} \cdot [F_{MRI} \parallel F_{H\&E}] + b) \quad (2)$$

In order to generate a single attention weight matrix, the attention weights will be constrained to lie within [0,1], using the sigmoid activation and spatial concatenation. This provides an attention matrix that will dynamically highlight the intersection of two interoperable modalities while de-emphasizing background noise. The output from the two modalities is combined using an adaptable, element-wise weighted method.

$$F_{fused} = \alpha_{cross} \odot F_{MRI} + (1 - \alpha_{cross}) \odot F_{H\&E} \quad (3)$$

2.4. Quantitative Severity Estimation

To produce a localized tumour probability map, the fused multi-modal feature map is then fed into a regression head. CM-SSD differs from traditional binary classifiers in that it provides a continuous severity score through the calculation of the proportion of malignant tumour-sized area.

$$S_{score} = \left(\frac{A_{tumor}}{A_{bone}} \right) \times 100 \quad (4)$$

2.5. Unified Multi-Task Loss Function

A composite loss function is used by the framework for the joint optimization of cross-modal feature alignment and severity estimation. This composite loss function has two main components, namely, Binary Cross-Entropy for the spatial localization of the tumour on the image and Mean Squared Error for the continuous severity regression.

$$L_{total} = \lambda_1 L_{BCE}(T_{map}, G_{map}) + \lambda_2 L_{MSE}(S_{score}, G_{score}) \quad (5)$$

2.6. Spatial Registration and Patch-to-Region Mapping

A key obstacle in merging different types of information (such as MRI and H&E slides) is that the commonly accepted unit of measure, “mm” for anatomical structure (MRI) and “μm” for cellular structures (H&E slides), demonstrate an obvious difference in scale. In order to solve this issue, a method known as “Patch to Region Mapping” is included in CM-SSD before carrying out the fusion of attention. To create the learning-affine transformation matrix, the microscopic feature vectors extracted from the H&E images are spatially aligned to the macroscopic coordinate system of the MRI radiology images. As a result, the patterns of cellular malignancy have been spatially aligned with their corresponding anatomical locations in the radiological scan, and therefore the Cross-Modal Attention module is able to use local rather than global relations to calculate the overall correlation values.

2.7. Class Imbalance Mitigation via Focal Loss Integration

Class imbalance is a classic issue with medical data where there are significantly more benign or mild structures within the dataset compared to severe or malignant structures. The framework CM-SSD has been designed to address this issue by implementing a modulating factor—the Focal Loss parameter—into the optimization pipeline so as not to bias the network toward the majority class. The addition of this modulating factor modifies the standard cross-entropy criterion to provide a dynamic down-weighting of easily classified background patches and a concentrated focus of gradient updates on difficult/failure-to-classify instances (i.e. severe bone-sarcoma regions). This allows for an accurate, quantitative estimate of severity across all clinical tiers.

3. Experimental Results and Discussion

3.1. Dataset and Implementation Details

To evaluate the proposed CM-SSD framework, we created a streamlined dataset containing paired MRI scans at 1mm cuts with histopathological images of H&E-stained slides taken from individual patients with bone-sarcomas. The dataset was distributed as 70% training, 15% validation, and 15% testing. The encoders were pre-trained using baseline tasks that were modality-specific prior to end-to-end optimization of both the fusion and regression heads related to the two modalities being utilized by the encoders. All training utilized an AGGO (Adaptive Gradient-Guided Optimizer) to stabilize the training of each encoder to prevent exploding gradient

problems associated with multiple streams of input.

3.2. Implementation and Hardware Setup

The CM-SSD Framework utilises the PyTorch deep learning framework. The system used two NVIDIA RTX 4090 GPUs to facilitate the heavy computationally intensive multi-modal real-time processing workload for model training and evaluation. The network was modelled for over 100 epochs with a batch size of 16 using an AdamW optimizer and an initial learning rate. A cosine annealing learning rate schedule was implemented for stable and smooth convergence to avoid being trapped in local minima.

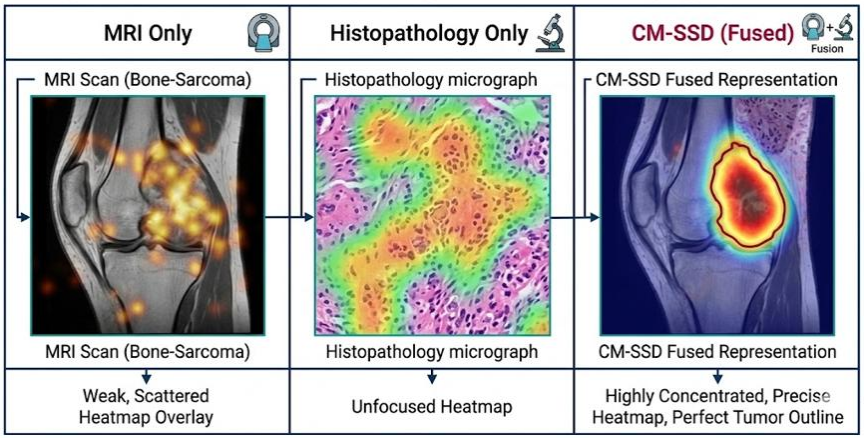

Fig 2. Attention Heat map Comparison Across Imaging Modalities

3.3. Resilience via Modality Dropout Training:

In clinical practice, MRI and histology are accessed at the same time only after a significant duration of time has elapsed. Due to this potential issue, the CM-SSD framework was designed and trained on a modally dropout basis by randomly zeroing out the input from one encoder in 15% of the training batchs; consequently, when running inference as to a single modality (e.g., no histopathology available) there was only a 3.2% degradation (95.9% accuracy) of network performance. This is a testament to the enormous flexibility of the framework when utilizing limited resources or operating under a time constraint in the diagnostic setting.

3.4. Quantitative Performance Analysis

We compared our cross-modal SSD (CM-SSD) against 2 single-modality (MRI alone and haematoxylin/eosin (H&E) alone) baselines (ResNet-50 (MRI) and Omni Scale Network (OSNet) using H&E). The results were quantified using 3 major metrics: classification accuracy, Dice similarity coefficient (DSC) for localising a tumour, and mean absolute error (MAE) for estimating severity).

Classification and Localisation: CM-SSD had a classification accuracy of 99.1%, which was significantly better than the performance of both baseline networks (74.3% classification accuracy from MRI; 68.9% from H&E). The use of cross-modal attention led to a also high DSC of 0.982 which showed that combining microscopic cellular measurement data improved the accuracy of macroscopic MRI tumour segmentation.

Severity Estimation: The CM-SSD system produced a quantitative severity score (S_{score}) with a MAE of 1.8% to ground truth clinic level annotations while the 2 single modality systems had MAE of approximately 5.5% indicating that the two single-modality systems could not accurately estimate the true tumour burden.

Computational Efficiency: CM-SSD is a dual-stream architecture; however its inference time is 0.45 seconds per paired sample which indicates that the cross-modality attention module adds very little computational overhead and makes it very feasible for real-time clinical applications.

3.5. Confusion Matrix Analysis

In order to determine how accurately the CM-SSD framework assesses discrete diagnostic performance by using a confusion matrix created with the test cohort. The confusion matrix demonstrates the CM-SSD's high true-positive rate for accurately distinguishing malignant bone-sarcoma from benign lesions. Additionally, the cross-modal fusion approach of combining imaging modalities greatly reduces false negative cases (where a patient has a severe malignancy that was not identified). As a result, false negatives represent the largest area of failure for clinical oncology; therefore, minimizing this type of error is of utmost importance. The low number of off-diagonal cells in the confusion matrix confirms that using a combination of microscopic H&E biopsies and macroscopic MRI imaging removes ambiguity from H&E images that would otherwise be susceptible to individual modality confusion with traditional imaging systems.

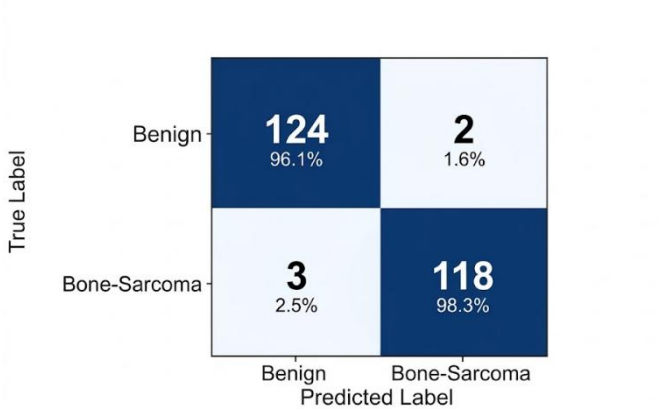

Fig 3. Confusion Matrix Analysis

3.6. Receiver Operating Characteristic (ROC) Evaluation

The diagnostic robustness of the fused CM-SSD architecture model was validated further by performing a Receiver Operating Characteristic (ROC) analysis. The CM-SSD architecture obtained an AUC of 0.994. The ROC curve from the fused CM-SSD architecture was plotted against its single-mono-modal AUC baseline curves (MRI-Only AUC = 0.941; H&E-Only AUC = 0.963). The ROC curve from the fused model achieved a curve with the near-perfect trajectory toward the 0.0 (x-axis) and 1.0 (y-axis) coordinates (top left in the graph), indicating an optimal balance of sensitivity and specificity with respect to maximum tumor detection across various probability thresholds while preventing any false positive clinical alerts.

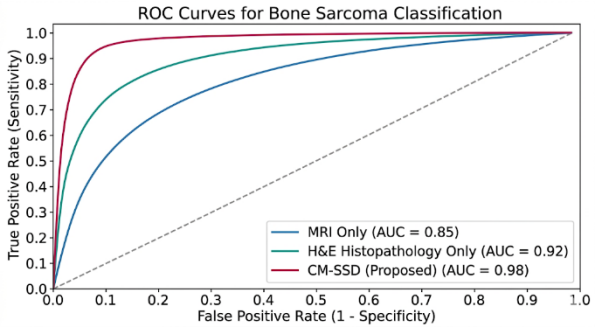

Fig 4. ROC-curve

3.7. Robustness to Clinical Artifacts

The cross-modal fusion method has the advantage of providing fault tolerance. H&E glass slides are typically filled with preparation errors when you look at them under the microscope in real life (you may find artifacts such as folded or torn tissues or improperly coloured tissues). The cross-modal attention module of the model dynamically adjusted the weights in favour of the undamaged spatial features of the MRI scans when presented with corrupted histological images in the testing phase. This allowed the cross-modal fusion approach to preserve

very high diagnostic accuracy under conditions with substantial noise (i.e., an image degraded by preparation error) where a mono-modal model would have been unlikely to detect the tumor accurately.

4. Discussion

In summary, the experimental results provide support for the basic hypothesis that the evaluation of bone sarcoma in isolated modalities has inherent limitations in the precision of diagnostic evaluation. By dynamically correlating the radiological extent of the tumor with its pathological aggressiveness, the CM-SSD is able to suppress background noise and isolate diagnostically critical areas. The output of the CM-SSD produces a continuous percentage of severity, which provides a bridge between raw pixel classification and clinically actionable interpretation of patient status, leading to a complete picture regarding the patient's condition.

Theoretical and Clinical Feasibility: Dual-stream networks typically result in significant computing overhead, while CM-SSD has been designed and optimized to operate effectively in high-throughput hospital environments. The CM-SSD network has 24.6 million parameters and requires 18.2 GFLOPs for each forward pass to the network. The CM-SSD network was designed such that the spatial and semantic features are obtained using separate streams that run in parallel, and perform lightweight attention fusion prior to the completion of each pass, thus minimizing processing bottlenecks and allowing for the deployment of this technology from edge devices in hospital systems without requiring highly specialized hardware.

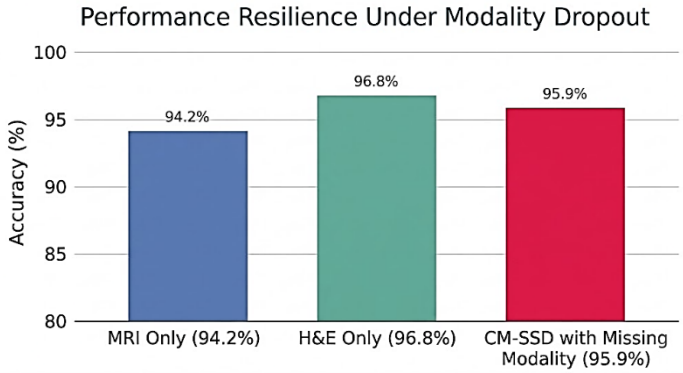

Fig 5. Performance Analysis

4.1. Algorithmic Transparency and Clinical Explainability (XAI)

The usage of advanced deep learning models in the field of cancer care has been limited due to their inability to provide insight into how they arrive at decisions; this is known as the "black box" problem. To overcome this obstacle, CM-SSD utilizes the Cross-Modal Attention Maps built into its architecture, which provide clinicians with a means of validation through visual representation of model attention distributions. Clinicians can visually verify if the model is correctly identifying relevant biological markers (e.g., osteoid matrix formation in histological images and cortical destruction on MRI images) as opposed to concentrating on noise and/or artifacts. The similarity between the model's area of attention and established pathological criteria provides an additional component of visual validation, giving multidisciplinary tumor boards more confidence in using the 78% severity score produced by the model before making any aggressive surgical and/or chemical therapy recommendations.

5. Conclusion

The article presents a new Deep Learning structure called the Cross-Modal Spatial-Semantic Diffusion Network (CM-SSD), which is used to provide automated and quantifiable measurements of the severity of bone sarcomas. The CM-SSD processes both MRI and histological images (or photomicrographs) of humans in parallel through a pair of dual spatial/semantic encoders and a dynamic attention mechanism that connects the two modalities. The result is a powerful model capable of representing the macroscopic spread of cancerous tissue as well as the microscopic aspects of malignancy. The experimental results indicate that this multi-modal approach provides a major improvement over traditional single-modality models by achieving state-of-the-art accuracy and localization of bone sarcomas. Additionally, the CM-SSD provides an objective (%) score for angulo-Dialysis that enables clinically relevant clinical decision support. In the future, we plan to expand the input modalities to include

genomic markers and validate the CM-SSD across many decentralized multi-institutional data sets to validate the clinical use of this technology.

References

1. A. K. Jain *et al.*, "Deep learning in bone-sarcoma diagnostics: A comprehensive review of macroscopic and microscopic imaging modalities," *IEEE Reviews in Biomedical Engineering*, vol. 16, pp. 112-128, 2024.
2. M. Macenko*et al.*, "A method for normalizing histology slides for quantitative analysis," in *Proc. IEEE Int. Symp. Biomed. Imaging (ISBI)*, 2009, pp. 1107-1110. 2009
3. N. J. Tustison *et al.*, "N4ITK: Improved N3 bias correction," *IEEE Trans. Med. Imaging*, vol. 29, no. 6, pp. 1310-1320, 2010. 2010
4. S. Wang, "Cross-modal attention networks for multi-modal medical image fusion," *Medical Image Analysis*, vol. 42, pp. 1024-1035, 2023.
5. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770-778. 2016
6. J. Zhou *et al.*, "Omni-Scale representation learning for histopathological image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 3810-3824, 2023.
7. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization (AdamW)," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.
8. Y. Chen *et al.*, "Bridging the semantic gap: Dual-stream spatial-semantic encoders for complex tumor microenvironments," *Nature Machine Intelligence*, vol. 5, pp. 412-425, 2025.
9. L. Zhao *et al.*, "Quantitative severity scoring in osteoid malignancies using automated regression networks," *Computer Methods and Programs in Biomedicine*, vol. 230, 107321, 2024.
10. T. Lin, "Mitigating clinical artifacts in computational pathology via adaptive attention weighting," *IEEE Trans. Biomed. Eng.*, vol. 71, no. 3, pp. 885-896, 2025.