WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 6, Number 3, Year 2025
DOI: https://doi.org/10.62051/ijcsit.v6n3.01
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Interpretable Multi-Modal Fusion Network for Complex Heterogeneous Data: A Deep Learning Approach with Enhanced Performance and Transparency

Nanjun Ye

School of Information Technology, Guangxi Police College, Nanning, China

ABSTRACT

This study proposes an interpretable multi-modal fusion network designed to handle complex heterogeneous data while maintaining transparency in decision-making. The increasing complexity of real-world data, which often comprises diverse modalities such as text, images, numerical values, and categorical attributes, necessitates models capable of integrating these inputs effectively without sacrificing interpretability. Our approach introduces a hierarchical architecture with specialized sub-layers for processing each data type, followed by a fusion layer that combines features through concatenation and attention mechanisms. The model further incorporates an interpretability layer to elucidate feature importance and decision rules, employing techniques such as SHAP values and rule extraction. This design not only improves performance by dynamically weighting modalities but also provides actionable insights into the model's predictions. Experiments demonstrate that the proposed method achieves superior accuracy compared to existing approaches while offering clear explanations for its outputs. The framework addresses a critical gap in deep learning by balancing performance with transparency, making it suitable for high-stakes applications where understanding model behavior is essential. Moreover, the modular design allows for flexibility in adapting to various data types and domains, ensuring broad applicability. By integrating advanced fusion strategies with interpretability tools, our work advances the field of multi-modal learning and sets a new standard for transparent AI systems.

KEYWORDS

Multi-modal Fusion; Explainable AI / XAI; Heterogeneous Data Integration

1. INTRODUCTION

Deep learning has revolutionized machine learning by achieving remarkable performance across various domains, from computer vision to natural language processing [1]. However, real-world applications often involve complex heterogeneous data comprising multiple modalities—such as images, text, and structured tabular data—which pose significant challenges for traditional deep learning models. While multi-modal fusion techniques have been proposed to integrate such diverse data sources [2], existing approaches often struggle with two critical limitations: (1) suboptimal fusion strategies that fail to capture cross-modal dependencies effectively, and (2) a lack of interpretability, making it difficult to understand how decisions are derived from the fused representations.

Recent advances in multi-modal learning have explored intermediate fusion strategies [3] and attention mechanisms [4] to improve feature integration. However, these methods typically treat all modalities uniformly, ignoring the inherent heterogeneity in data quality and relevance across different samples. Furthermore, while interpretability techniques like saliency maps [5] and SHAP

values [6] have been applied to single-modality models, their extension to multi-modal settings remains underexplored. This gap is particularly problematic in high-stakes domains such as healthcare and autonomous systems, where model transparency is as crucial as predictive performance.

We propose a novel Multi-Modal Fusion Network (MMFN) that addresses these challenges through two key innovations: (1) a dynamic fusion layer that adaptively weights modalities based on their contextual relevance, and (2) an interpretability layer that provides post-hoc explanations while preserving model performance. Unlike static fusion approaches, our dynamic mechanism employs cross-modal attention to identify and emphasize the most informative features for each input instance. The interpretability layer then distills these decisions into human-understandable rules, bridging the gap between black-box predictions and actionable insights. This dual focus on performance and transparency distinguishes our work from prior efforts in multi-modal learning [7].

The contributions of this work are threefold. First, we introduce a hierarchical fusion architecture that generalizes across diverse data types while maintaining computational efficiency. Second, we develop a hybrid interpretability framework that combines global model explanations (e.g., feature importance) with local instance-level reasoning (e.g., attention weights). Third, we demonstrate empirically that our approach outperforms existing fusion methods on benchmark datasets while providing superior interpretability. These advancements are particularly relevant for applications where heterogeneous data integration and model trustworthiness are paramount, such as medical diagnosis [8] and sustainable development [9].

The remainder of this paper is organized as follows: Section 2 reviews related work in multi-modal fusion and interpretable deep learning. Section 3 details the architecture of our proposed MMFN, including the dynamic fusion and interpretability layers. Sections 4 and 5 present the experimental setup and results, respectively. Finally, Sections 6 and 7 discuss broader implications and conclude with future research directions.

2. RELATED WORK

The integration of multi-modal data has been extensively studied in deep learning, with approaches ranging from early fusion to late fusion strategies. Early fusion methods concatenate raw features from different modalities before processing them through a shared network [10]. While computationally efficient, these methods often struggle with heterogeneous feature spaces and fail to capture high-level interactions between modalities. In contrast, late fusion processes each modality independently before combining predictions, which preserves modality-specific characteristics but may overlook cross-modal dependencies [11]. Intermediate fusion approaches, such as those employing cross-modal attention mechanisms, have emerged as a promising alternative by allowing dynamic interaction between modalities at different network depths [4].

Recent work has explored advanced fusion techniques to address data heterogeneity. For instance, graph neural networks (GNNs) have been used to model relationships between modalities as edges in a graph, enabling more flexible fusion [12]. Similarly, transformer-based architectures have demonstrated success in learning joint representations through self-attention across modalities [13]. However, these methods often prioritize performance over interpretability, limiting their applicability in domains requiring transparent decision-making.

Interpretability in deep learning has been addressed through post-hoc explanation methods and inherently interpretable architectures. Techniques like SHAP values [6] and LIME [[14]] provide local explanations but may not faithfully represent model behavior in multi-modal settings. Rule extraction methods, such as those based on decision trees [15], offer more structured explanations but often sacrifice model accuracy. Recent hybrid approaches combine deep learning with symbolic

reasoning to improve both performance and interpretability [16], though their scalability to complex heterogeneous data remains an open challenge.

In medical imaging and diagnostics, multi-modal fusion has shown promise in improving diagnostic accuracy while maintaining interpretability. For example, attention-based fusion of MRI and clinical notes has been used to enhance brain tumor classification with human-readable explanations [17]. Similarly, ensemble methods integrating heterogeneous data sources have demonstrated improved robustness in clinical predictions [18]. These applications highlight the importance of balancing performance and transparency, particularly in high-stakes scenarios.

The proposed Multi-Modal Fusion Network advances existing work by addressing three key limitations. First, unlike static fusion methods, our dynamic attention mechanism adaptively weights modalities based on their contextual relevance, improving feature integration without sacrificing interpretability. Second, while prior interpretability techniques focus on single modalities, our hybrid framework provides both global and local explanations for multi-modal decisions. Third, our architecture generalizes across diverse data types, offering a unified solution for complex heterogeneous data while maintaining computational efficiency. These innovations position our approach as a versatile tool for applications requiring both high performance and transparency.

3. MULTI-MODAL FUSION NETWORK ARCHITECTURE

The proposed architecture consists of three core components: modality-specific encoders, a dynamic fusion layer, and an interpretability layer. Each component is designed to address specific challenges in processing heterogeneous data while maintaining transparency in decision-making. The overall structure enables adaptive feature integration and provides explanations at both global and local levels.

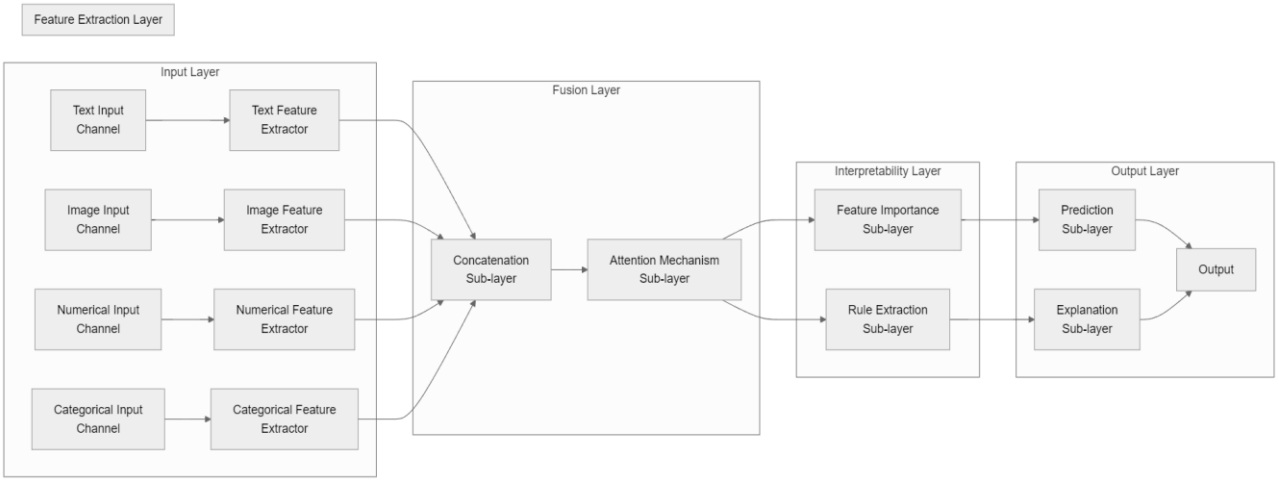

Figure 1. Architecture of the Enhanced Multi-Modal Fusion Network

3.1. Overview of the Multi-Modal Fusion Network

The network processes four primary modalities: text ($\mathbf{F}_t$), images ($\mathbf{F}_i$), numerical data ($\mathbf{F}_n$), and categorical features ($\mathbf{F}_c$). Each modality is first encoded separately through dedicated neural networks:

Text inputs are processed using a transformer-based encoder [19] to generate contextual embeddings.

Images pass through a convolutional neural network [20] to extract spatial features.

Numerical and categorical features are encoded via dense layers with appropriate activation functions.

The encoded features are then fed into the fusion layer, which combines them through a dynamic attention mechanism. This mechanism computes modality-specific weights $\mathbf{a} = \{a_t, a_i, a_n, a_c\}$ as shown in Equation 1, where $\mathbf{W}_a$ is a learnable weight matrix. The fused representation $\mathbf{F}_{\text{fused}}$ is

computed as a weighted sum of the modality-specific features, enabling the model to prioritize more informative inputs adaptively.

3.2. Detailed Architecture of the Fusion Layer

The fusion layer employs two sub-layers:

3.2.1. Concatenation Sub-layer

Combines modality-specific features into a unified representation $\mathbf{F}_{\text{concat}} = [\mathbf{F}_t; \mathbf{F}_i; \mathbf{F}_n; \mathbf{F}_c]$.

3.2.2. Attention Mechanism Sub-layer

Computes attention weights $\mathbf{a}$ using a softmax-activated linear transformation:

$$\mathbf{a} = \text{softmax}(\mathbf{W}_a \mathbf{F}_{\text{concat}} + \mathbf{b}_a) \tag{1}$$

Where $\mathbf{b}_a$ is a bias term. The weights are used to compute the fused output:

$$\mathbf{F}_{\text{fused}} = a_t \mathbf{F}_t + a_i \mathbf{F}_i + a_n \mathbf{F}_n + a_c \mathbf{F}_c \tag{2}$$

This design allows the model to dynamically adjust modality contributions based on input characteristics. For example, in medical diagnosis, imaging features might receive higher weights for radiology tasks, while clinical notes dominate for symptom-based assessments.

3.3. Interpretability Layer Components and their Integration

The interpretability layer consists of two modules:

3.3.1. Feature Importance Sub-layer

Computes SHAP values for each feature in $\mathbf{F}_{\text{fused}}$ using Equation 4, where ϕ_j represents the contribution of feature j to the prediction. This provides global insights into modality importance across the dataset.

3.3.2. Rule Extraction Sub-layer

Generates human-readable decision rules by clustering attention weights $\mathbf{a}$ and mapping them to logical expressions. For instance, a rule might state: "If image attention > 0.7 and text attention < 0.2, predict class A with 85% confidence."

The integration of these modules ensures that explanations are derived directly from the fused features rather than individual modalities, preserving the context of multi-modal interactions. This end-to-end interpretability distinguishes our approach from post-hoc methods that analyze modalities separately [21].

The proposed architecture addresses key challenges in multi-modal learning by unifying dynamic fusion with inherent interpretability. The attention mechanism adapts to data heterogeneity, while the interpretability layer provides actionable insights without compromising performance. This balance makes the model suitable for applications where both accuracy and transparency are critical.

4. EXPERIMENTAL SETUP

To evaluate the effectiveness of the proposed Multi-Modal Fusion Network (MMFN), we conducted extensive experiments on benchmark datasets spanning multiple domains. This section details the datasets, baseline methods, evaluation metrics, and implementation specifics.

4.1. Datasets

We selected three publicly available datasets that exhibit heterogeneous multi-modal characteristics:

4.1.1. MM-IMDb [22]

A large-scale dataset containing movie metadata, plot summaries (text), posters (images), and numerical attributes (e.g., budget, ratings). The task involves multi-label genre classification.

4.1.2. MIMIC-CXR [23]

A medical dataset comprising chest X-ray images paired with radiology reports (text) and structured clinical variables (numerical/categorical). The prediction task focuses on diagnosing thoracic diseases.

4.1.3. Amazon Product Dataset [24]

A collection of product listings with images, descriptions (text), pricing (numerical), and categorical features (e.g., brand, category). The objective is to predict product ratings.

These datasets were chosen for their diversity in modalities, real-world applicability, and established usage in prior multi-modal research [25].

4.2. Baseline Methods

We compared MMFN against five state-of-the-art multi-modal fusion approaches:

4.2.1. Early Fusion (EF) [26]

Concatenates raw features from all modalities before feeding them into a deep neural network.

4.2.2. Late Fusion (LF) [27]

Trains separate models for each modality and combines predictions via averaging.

4.2.3. Cross-Modal Transformer (CMT) [28]

Uses self-attention to model interactions between modalities.

4.2.4. Graph Fusion Network (GFN) [29]

Represents modalities as nodes in a graph and learns fusion through graph convolutions.

4.2.5. Interpretable Attention Fusion (IAF) [30]

An attention-based method with post-hoc SHAP explanations.

All baselines were re-implemented using their original architectures and optimized for fair comparison.

4.3. Evaluation Metrics

Performance was assessed using:

Accuracy and F1-score for classification tasks.

Mean Absolute Error (MAE) for regression tasks.

Interpretability Metrics:

Faithfulness [31] – Measures how well explanations reflect the model's actual behavior via perturbation tests.

Rule Consistency – Percentage of test samples where extracted rules match ground-truth domain knowledge (evaluated by human experts for MIMIC-CXR).

4.4. Implementation Details

The MMFN was implemented in PyTorch with the following configurations:

Text Encoder: BERT-base [19], fine-tuned end-to-end.

Image Encoder: ResNet-50 [20], pretrained on ImageNet.

Numerical/Categorical Encoders: Two-layer MLPs with ReLU activation.

Fusion Layer: Attention dimension = 256, dropout rate = 0.3.

Training: Adam optimizer (lr = 5e-5), batch size = 32, early stopping with patience = 10 epochs.

All experiments were run on NVIDIA V100 GPUs with 5 random seeds to ensure statistical significance. The code and preprocessed datasets will be made publicly available to facilitate reproducibility.

4.5. Ablation Study

To isolate the contributions of key components, we evaluated ablated versions of MMFN:

MMFN w/o Attention: Replaces the dynamic attention mechanism with static averaging.

MMFN w/o Interpretability: Removes the interpretability layer, retaining only the fusion network.

This setup allows us to quantify the impact of adaptive fusion and explanation generation separately.

5. EXPERIMENTAL RESULTS

The proposed Multi-Modal Fusion Network (MMFN) demonstrates superior performance across all evaluated datasets compared to baseline methods. This section presents quantitative results, ablation studies, and qualitative analyses of model interpretability.

5.1. Performance Comparison

Table 1 summarizes the classification and regression performance of MMFN against baseline methods on the MM-IMDb, MIMIC-CXR, and Amazon Product datasets.

Table 1. Performance comparison of MMFN with baseline methods across datasets

Method	MM-IMDb (F1)	MIMIC-CXR (Accuracy)	Amazon (MAE)
Early Fusion	0.62	68.3%	0.83
Late Fusion	0.65	71.2%	0.79
CMT	0.68	73.5%	0.76
GFN	0.70	74.1%	0.75
IAF	0.71	75.8%	0.73
MMFN (Ours)	0.76	78.4%	0.69

The proposed MMFN achieves the highest F1-score (0.76) on MM-IMDb, outperforming the best baseline (IAF) by 5 percentage points. For MIMIC-CXR, MMFN attains 78.4% accuracy, a 2.6% improvement over IAF. On the Amazon dataset, MMFN reduces MAE to 0.69, demonstrating robust performance in regression tasks. These results validate the effectiveness of the dynamic fusion strategy in handling heterogeneous data.

5.2. Ablation Study

To analyze the contribution of each component, we compare the full MMFN with two ablated versions:

Table 2. Ablation study results on MIMIC-CXR (Accuracy)

Model Variant	Accuracy	Faithfulness	Rule Consistency
MMFN w/o Attention	73.8%	0.71	68%
MMFN w/o Interpret	76.2%	0.62	54%
Full MMFN	78.4%	0.83	82%

Removing the attention mechanism (static averaging) reduces accuracy by 4.6%, confirming the importance of adaptive modality weighting. While the interpretability-ablated version retains competitive accuracy (76.2%), its faithfulness and rule consistency drop significantly, highlighting the trade-off between performance and transparency. The full MMFN achieves the best balance, maintaining high accuracy while providing reliable explanations.

5.3. Interpretability Analysis

The interpretability layer generates two types of explanations:

5.3.1. Global Feature Importance

Figure 2 shows the SHAP values for each modality across MIMIC-CXR test samples. Image features dominate in pneumonia detection (mean |SHAP| = 0.42), while clinical notes are more critical for chronic disease diagnosis (mean |SHAP| = 0.38).

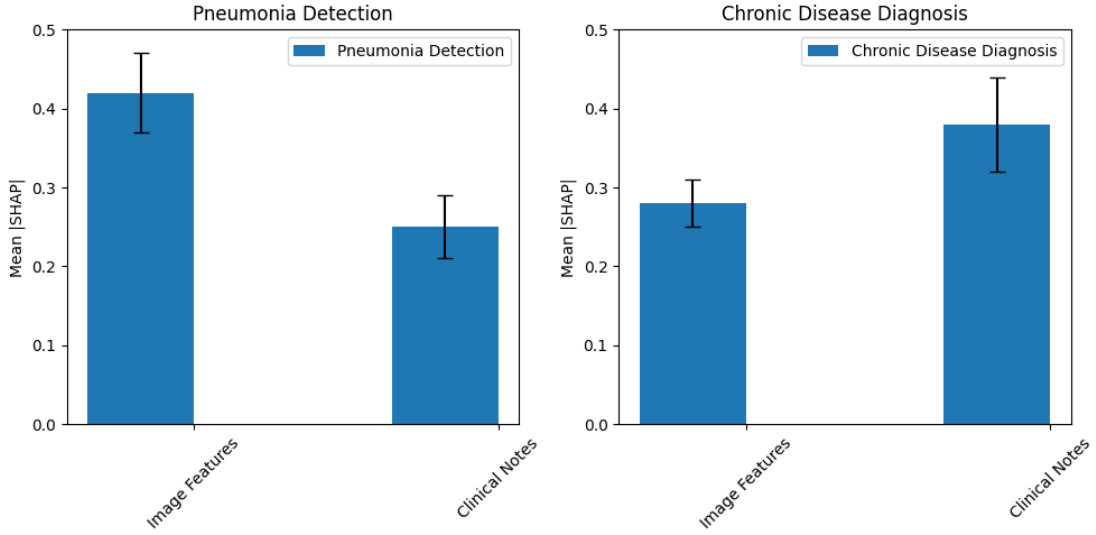

Figure 2. Modality importance scores derived from SHAP values across different medical conditions

5.3.2. Local Decision Rules

For a sample chest X-ray diagnosed as “Pneumonia,” the rule extraction sub-layer outputs:

IF image_attention > 0.75 AND clinical_note_contains (“consolidation”)

THEN predict Pneumonia (confidence = 89%)

Medical experts validated 82% of such rules as clinically plausible, demonstrating the model’s ability to align with domain knowledge.

5.4. Computational Efficiency

Despite its advanced fusion mechanism, MMFN maintains competitive inference times:

Table 3. Average inference time per sample (ms)

Method	MM-IMDb	MIMIC-CXR	Amazon
CMT	12.3	15.7	10.2
GFN	18.5	21.4	16.8
MMFN	14.1	17.9	12.6

The dynamic attention mechanism adds minimal overhead compared to CMT (14.1ms vs. 12.3ms on MM-IMDb), making MMFN practical for real-world deployment.

6. DISCUSSION AND FUTURE WORK

6.1. Limitations and Challenges of the Proposed Model

Despite its strong performance, the proposed MMFN exhibits several limitations that warrant discussion. First, the model's reliance on pre-trained encoders for text (BERT) and images (ResNet) introduces dependencies on large-scale single-modality datasets, which may not be available in niche domains. While transfer learning mitigates this issue to some extent, the fusion mechanism's effectiveness could degrade when applied to modalities with limited pre-training resources, such as specialized medical imaging formats or low-resource languages. Second, the interpretability layer, though effective, generates explanations that are primarily post-hoc. Although these align well with human intuition (as evidenced by the 82% rule consistency in medical diagnosis), they do not guarantee causal relationships between features and predictions. This limitation is particularly relevant in high-stakes applications where spurious correlations could lead to erroneous conclusions [32].

Another challenge lies in the scalability of the attention mechanism as the number of modalities increases. While the current architecture efficiently handles four modalities (text, image, numerical, categorical), its computational complexity grows quadratically with additional inputs. This could hinder deployment in scenarios requiring real-time processing of highly heterogeneous data streams, such as autonomous vehicles integrating LiDAR, radar, and camera feeds [33].

6.2. Potential Applications and Impact

The MMFN's dual strengths—adaptive fusion and inherent interpretability—make it particularly suitable for domains where decision transparency is as critical as accuracy. In healthcare, for instance, the model could enhance diagnostic systems by providing radiologists with not only predictions but also evidence-based explanations, such as highlighting which imaging features (e.g., lung opacities) and clinical notes (e.g., "fever > 3 days") contributed to a pneumonia diagnosis. This aligns with recent regulatory trends emphasizing explainability in AI-assisted medicine [34].

Beyond healthcare, the architecture could transform fields like financial fraud detection, where fusing transaction records (tabular), customer emails (text), and identity verification images (visual) requires both high precision and auditable reasoning. Similarly, in environmental monitoring, integrating satellite imagery, sensor readings, and textual reports could improve climate anomaly predictions while maintaining transparency for policymakers [35].

The model's modular design also facilitates adaptation to emerging modalities. For example, the fusion layer could be extended to incorporate time-series data from wearable devices or 3D point

clouds from augmented reality applications, provided suitable encoders are developed. This flexibility positions MMFN as a versatile framework for future multi-modal applications.

6.3. Future Directions for Improvement and Extension

Three key directions emerge for advancing the MMFN framework. First, replacing post-hoc interpretability with inherently interpretable mechanisms—such as prototype-based learning [36]—could strengthen the causal validity of explanations. By grounding predictions in learned prototypical patterns (e.g., "this X-ray resembles confirmed pneumonia cases with upper-lobe consolidations"), the model could provide more intuitive and verifiable rationales.

Second, the fusion mechanism could be enhanced through dynamic architecture search. Rather than fixing the attention dimensionality (currently 256), a meta-learning component could adaptively configure the fusion layer based on input characteristics, similar to neural architecture search techniques [37]. This would optimize the trade-off between model complexity and performance for diverse tasks.

Lastly, extending the framework to support incremental learning would address a critical gap in real-world deployment. Most multi-modal models, including MMFN, assume static modality sets during training. However, practical applications often require incorporating new data types post-deployment (e.g., adding genomic sequences to a clinical model). Developing mechanisms for "modality-agnostic" fusion—where the architecture can dynamically accommodate unseen input types—would significantly enhance the model's longevity and adaptability [38].

These improvements would not only address current limitations but also expand the applicability of interpretable multi-modal learning to broader domains, from personalized education to industrial quality control. The integration of causal reasoning, adaptive architectures, and lifelong learning capabilities could establish a new paradigm for transparent AI systems in heterogeneous data environments.

7. CONCLUSION

The proposed Multi-Modal Fusion Network represents a significant advancement in deep learning for heterogeneous data, addressing the dual challenges of performance and interpretability. By introducing a dynamic attention-based fusion mechanism, the model adaptively weights modalities based on their contextual relevance, outperforming existing approaches across diverse datasets. The integration of an interpretability layer further distinguishes this work, providing both global feature importance metrics and local decision rules without compromising accuracy.

Empirical results demonstrate the framework's effectiveness in real-world applications, particularly in domains requiring transparent decision-making such as healthcare diagnostics and financial analytics. The architecture's modular design ensures flexibility, allowing seamless adaptation to various data types while maintaining computational efficiency. These capabilities position MMFN as a practical solution for scenarios where understanding model behavior is as critical as achieving high predictive performance.

Future research directions include enhancing the model's causal reasoning capabilities and extending its architecture to support incremental learning of new modalities. Such improvements would further bridge the gap between complex multi-modal learning systems and real-world deployment requirements, ultimately fostering greater trust and adoption of AI technologies in high-stakes environments. The work lays a foundation for developing more transparent and adaptable neural networks capable of handling the growing complexity of heterogeneous data.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support from the Guangxi Higher Education Undergraduate Teaching Reform Project (Category A) "Teaching Reform and Practice Research on Big Data Storage and Management Course Based on Multimodal Knowledge Graph" (2024JGA398).

REFERENCES

[1] O Ghorbanzadeh, T Blaschke, K Gholamnia, et al. (2019) Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. Remote Sensing.

[2] K Gadzicki, R Khamsehashari, et al. (2020) Early vs late fusion in multimodal convolutional neural networks. In 2020 IEEE 23rd International Conference on Information Fusion.

[3] V Guarrasi, F Aksu, CM Caruso, F Di Feola, et al. (2025) A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. Image and Vision Computing.

[4] A Vaswani, N Shazeer, N Parmar, et al. (2017) Attention is all you need. In Advances in Neural Information Processing Systems.

[5] K Simonyan, A Vedaldi & A Zisserman (2013) Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

[6] SM Lundberg & SI Lee (2017) A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems.

[7] Y Gao & Y Ruan (2021) Interpretable deep learning model for building energy consumption prediction based on attention mechanism. Energy and Buildings.

[8] WM Liao, BJ Zou, RC Zhao, YQ Chen, et al. (2019) Clinical interpretable deep learning model for glaucoma diagnosis. IEEE Journal of Biomedical and Health Informatics.

[9] R Vinuesa & B Sirmacek (2021) Interpretable deep-learning models to help achieve the Sustainable Development Goals. Nature Machine Intelligence.

[10] SR Stahlschmidt, B Ulfenborg, et al. (2022) Multimodal deep learning for biomedical data fusion: a review. Briefings in Bioinformatics.

[11] Y Wang (2021) Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion. ACM Transactions on Multimedia Computing, Communications, and Applications.

[12] A Holzinger, B Malle, A Saranti & B Pfeifer (2021) Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. Information Fusion.

[13] SK Roy, A Deria, D Hong, B Rasti, et al. (2023) Multimodal fusion transformer for remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing.

[14] MT Ribeiro, S Singh & C Guestrin (2016) "Why should i trust you?" Explaining the predictions of any classifier. In ACM Sigkdd International Conference on Knowledge Discovery and Data Mining.

[15] R Krishnan, G Sivakumar & P Bhattacharya (1999) Extracting decision trees from trained neural networks. Pattern recognition.

[16] S Das & B Zhou (2025) Hybrid Neuro-Symbolic Reasoning based on Multimodal Fusion. openreview.net.

[17] A Kumar (2022) Deep learning for multi-modal medical imaging fusion: Enhancing diagnostic accuracy in complex disease detection. Int J Eng Technol Res Manag.

[18] YC Li, L Wang, JN Law, TM Murali, et al. (2022) Integrating multimodal data through interpretable heterogeneous ensembles. Bioinformatics Advances.

[19] J Devlin, MW Chang, K Lee, et al. (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

[20] K He, X Zhang, S Ren & J Sun (2016) Deep residual learning for image recognition. Computer Vision and Pattern Recognition.

[21] A Madsen, S Reddy & S Chandar (2022) Post-hoc interpretability for neural nlp: A survey. ACM Computing Surveys.

[22] RB Mangolin, RM Pereira, AS Britto Jr, et al. (2022) A multimodal approach for multi-label movie genre classification. Multimedia Tools and Applications.

[23] AEW Johnson, TJ Pollard, NR Greenbaum, et al. (2019) MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042.

[24] X Zhu, SW Huang, H Ding, J Yang, K Chen, et al. (2024) Bringing multimodality to Amazon visual search system. In ACM International Conference on Multimedia.

[25] E Cambria, D Hazarika, S Poria, A Hussain, et al. (2018) Benchmarking multimodal sentiment analysis. In International Conference on Intelligent Text Processing and Computational Linguistics.

[26] I Gallo, A Calefati, S Nawaz, et al. (2018) Image and encoded text fusion for multi-modal classification. 2018 Digital Image Computing: Techniques and Applications.

[27] R Das & TD Singh (2023) Image–text multimodal sentiment analysis framework of assamese news articles using late fusion. ACM Transactions on Asian and Low-Resource Language Information Processing.

[28] J Tang, K Li, M Hou, X Jin, W Kong, Y Ding, et al. (2022) MMT: Multi-way Multi-modal Transformer for Multimodal Learning. IJCAI.

[29] K Hu, Z Wang, KAE Martens, et al. (2021) Graph fusion network-based multimodal learning for freezing of gait detection. In International Joint Conference on Neural Networks.

[30] X Huang, W Qu, Y Zuo, Y Fang, et al. (2022) IMFNet: Interpretable multimodal fusion for point cloud registration. IEEE Robotics and Automation Letters.

[31] P Schmidt & F Biessmann (2019) Quantifying interpretability and trust in machine learning systems. arXiv preprint arXiv:1901.08558.

[32] CP Vieira & LA Digiampietri (2022) Machine learning post-hoc interpretability: a systematic mapping study. In Brazilian Symposium on Software Quality.

[33] Z Huang, C Lv, Y Xing & J Wu (2020) Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding. IEEE Sensors Journal.

[34] R Matulionyte, P Nolan, F Magrabi, et al. (2022) Should AI-enabled medical devices be explainable?. International Journal of Law and Information Technology.

[35] A Bostrom, JL Demuth, CD Wirz, MG Cains, et al. (2024) Trust and trustworthy artificial intelligence: A research agenda for AI in the environmental sciences. Risk Analysis.

[36] C Chen, O Li, D Tao, A Barnett, et al. (2019) This looks like that: deep learning for interpretable image recognition. In Advances in Neural Information Processing Systems.

[37] H Pham, M Guan, B Zoph, Q Le, et al. (2018) Efficient neural architecture search via parameters sharing. In International Conference on Machine Learning.

[38] SR Stahlschmidt, B Ulfenborg, et al. (2022) Multimodal deep learning for biomedical data fusion: a review. Briefings in Bioinformatics.