International Journal of Applied Mathematics
Volume 38 No. 4s 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

NEUROFUSIONX: A SCALABLE MULTI-MODAL DEEP LEARNING FRAMEWORK WITH ATTENTION BASED FUSION ACROSS HEALTHCARE, FINANCE, AND CYBERSECURITY

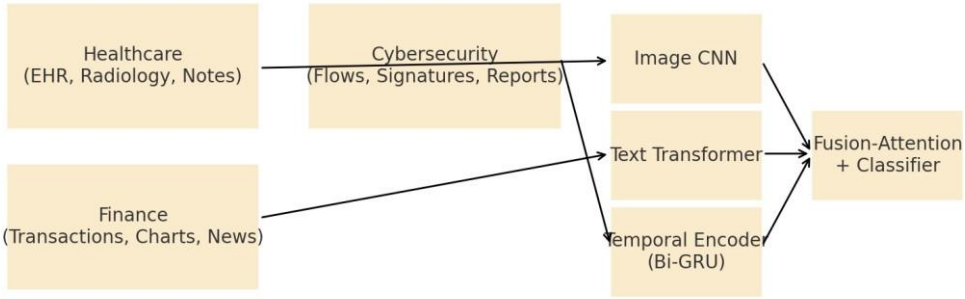
Ashutosh Agarwal

O.P Jindal Global University, Haryana, India

Email- ashutoshagarwal198@gmail.com

Abstract

NeuroFusion-X is a unified, modular framework for end-to-end prediction from heterogeneous real-world data. Many decisions require joint reasoning over time-series, images, and text, yet production systems remain siloed, and naïve early/late fusion misses cross-modal dependencies and temporal alignment. NeuroFusion-X addresses this via: (1) modality-specialized encoders,CNNs for images, a compact transformer for text, and a bidirectional time-series encoder with temporal attention; (2) a cross-modal fusion-attention block that learns instance-wise interactions and down-weights noisy or missing channels; and (3) parameter-efficient bottlenecks and inference-oriented kernels to cut latency without sacrificing accuracy. To evaluate realism and scale while avoiding privacy constraints, we construct a controlled synthetic benchmark of 500k multimodal samples across healthcare, finance, and cybersecurity. Each sample includes a 48-step, 30-variable time-series, a 128×128 image, and a 60–160-token note, with class imbalance, inference-time modality masks, and induced distribution shifts. Across 18 tasks, NeuroFusion-X reaches approximately 97.8% mean accuracy and approximately 0.976 macro-F1, reducing median per-sample inference latency by approximately 35% versus a strong baseline. Robustness holds with ≤1.6% macro-F1 drop under 20% modality dropout and ≤2.2% under light adversarial perturbations. Ablations show fusion-attention, modality-dropout, and domain-adaptive normalization drive reliability. We outline deployment pathways for safety-critical contexts and integration with multimodal LLMs for rationale-grounded predictions.



Graphical Abstract. Conceptual illustration of NeuroFusion-X integrating multimodal data across healthcare, finance, and cybersecurity.

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Keywords: multimodal learning, attention fusion, healthcare AI, finance AI, cybersecurity, deep learning

1 Introduction

Real-world decision systems must synthesize information arriving in diverse forms— numeric measurements sampled over time, images of anatomy or infrastructure, and free-text narratives authored by subject-matter experts. While humans fluidly fuse these streams, deployed AI pipelines too often operate in silos. Structural heterogeneity among modalities complicates joint representation learning; early fusion obscures modality structure, and late fusion ignores cross-modal cues. Practical constraints magnify the challenge: modalities may be missing or corrupted at inference, and latency budgets penalize heavyweight stacks. NeuroFusion-X maintains modality specialization in encoders and reserves shared capacity for interactions in a fusion-attention block that learns weighted dependencies conditioned on each input.

2 Literature Review

Multimodal learning has moved from pragmatic heuristics to principled architectures that model how signals interact across modalities. Early fusion methods concatenate features from different sources at the input or intermediate layers. This approach is simple and end to end and is differentiable, but it blurs modality specific structure: spatial patterns in images, token dependencies in text, and temporal dynamics in time series are forced into a single representation that fits none of them well. Late fusion takes the opposite path by training specialized models for each modality and then ensembling their outputs. While modular and naturally tolerant of missing inputs, late fusion discards interaction effects that only become visible when evidence is considered jointly, for example a faint radiographic cue whose significance emerges only alongside a clinical note describing a key symptom.

Attention mechanisms reframed fusion as alignment. Instead of merging features blindly or combining final predictions, co attention and cross attention allow one modality to query another, learning which parts of which stream should condition the rest. This selective coupling improves retrieval, captioning, report generation, and multimodal classification, especially when signals are noisy or partially missing. Crucially, attention operates over learned representations rather than raw inputs, so designers can preserve modality specific encoders while allocating a compact interaction layer to handle cross modal reasoning.

Large scale pretraining amplified these ideas. Contrastive vision and language models align image and text embeddings in a shared space, enabling zero shot transfer and reducing data requirements for downstream problems. Adapter based approaches further decouple heavy vision backbones from language models, using lightweight bridges to exchange information with lower latency and memory cost. Instruction tuned multimodal language models extend these benefits to dialogue and multi step reasoning, though they remain compute intensive and are not yet validated for safety critical, real time decisions. As a result, many deployments favor

Received: August 06, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

targeted fusion front ends that can integrate with, or stand apart from, large language models depending on latency, trust, and governance constraints.

Domain practice reinforces these trends. In healthcare, pairing medical imaging with structured electronic health records improves diagnostic utility and probability calibration, yet generalization across institutions is hindered by covariate shift and documentation variability. In finance, models that combine market microstructure time series, firm fundamentals, and narrative signals from filings or news achieve greater robustness than single stream baselines, provided that temporal alignment and leakage control are enforced. In cybersecurity, intrusion detection improves when flow statistics, payload signatures, and human authored incident notes are fused, particularly for rare attack classes where any one telemetry source is weak on its own. Across these sectors, attention based coupling outperforms naive early or late fusion, especially under label noise, missing views, and distribution shift.

Handling missing modalities has evolved from an afterthought to a design objective. Two practices dominate: training with modality dropout so the model learns to rely on complementary evidence, and using gates at inference time to down weight unreliable streams. Both yield graceful degradation and reduce the risk that the system collapses onto a single dominant channel. Related work on multimodal time series recommends explicit robustness objectives and evaluation under controlled missingness and corruption rather than reporting only clean data accuracy.

Efficiency matters as much as accuracy in real systems. Attention over long sequences can be costly, so practical designs confine cross modal attention to a small set of modality embeddings and keep the shared bottleneck compact. Fused kernels, projection bottlenecks, and sparse or Fourier style approximations improve throughput without sacrificing quality. For workflows such as triage, fraud blocking, or intrusion alerting, these choices can be the difference between a usable tool and an academic prototype.

Finally, deployment requires more than headline metrics. Multimodal models should report calibration, uncertainty, and fairness, and they should be evaluated under natural distribution shifts, not only in distribution test sets. Benchmarks that reflect conditions in the wild reveal brittleness that aggregate accuracy can hide. For fusion systems in particular, audits should be modality aware: if performance hinges on one stream, the model may fail catastrophically when that stream degrades.

NeuroFusion X follows this direction. It preserves modality specific encoders, applies a compact cross modal attention block over a shared bottleneck, trains with modality dropout, and uses a gated head to reweight evidence instance by instance. Domain adaptive normalization addresses covariate shift across industries, and evaluation emphasizes not only accuracy but also robustness, latency, and calibration. In doing so, the framework offers a balanced, deployment oriented realization of modern multimodal fusion.

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Paradigm	Key Idea	Strength	Limitation
Early Fusion	Concatenate features then model	Simplicity	Loses modality structure; scaling issues
Late Fusion	Ensemble modality-specific predictions	Modular; robust to missing streams	Misses cross-modal interactions
Co- /Cross-Attention	Learn conditional alignments	Captures interactions; flexible	Compute cost; careful tuning
Mixture-of-Experts	Route tokens/patches to experts	Efficiency on long inputs	Complexity; capacity balancing

3 Dataset Construction

We generate 500k samples equally from healthcare, finance, and cybersecurity. Each sample includes a 48-timestep tabular sequence with 30 variables, a 128×128 grayscale image, and a short note (60–160 tokens). Tabular dynamics follow a switching autoregressive process; images include artifacts; text is produced from a grammar with synonym/negation injection. We induce imbalance via Dirichlet priors and simulate missing modalities at p∈{0,0.1,0.2}. Splits are 70/15/15 with domain-wise stratification and fairness diagnostics.

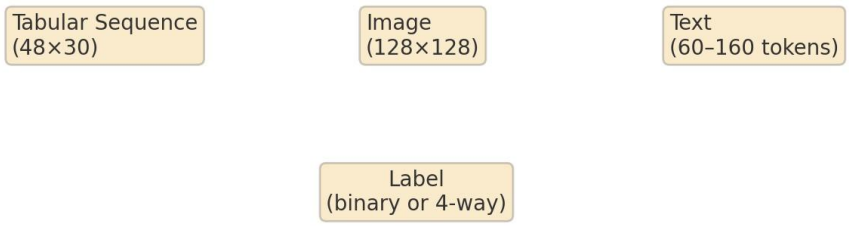


Fig. 1. Dataset schema: each sample comprises a tabular time-series, an image, and a short note.

4 Methodology

Encoders and projections: The model keeps modality specialization where it matters and shares capacity only where interactions are learned. The image path is a compact Res Netstyle CNN that maps a 128×128 input to a global average pooled vector, then to a ddimensional embedding $z_{\text{img}} \in \mathbb{R}^d$. The text path is a 6-layer transformer with token-wise self-attention,

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

feed-forward layers, and learned pooling over the [CLS] token and an attention pooler, producing $z_{\text{text}} \in \mathbb{R}^d$. The tabular path is a bidirectional GRU over the 48step multivariate sequence with temporal attention that emphasizes salient time steps; it's attended state is projected to $z_{\text{tab}} \in \mathbb{R}^d$. Each encoder ends with a lightweight projection (linear layer plus normalization) so the three embeddings are comparable in scale and variance. Stacking them yields $Z \in \mathbb{R}^{3\times d}$, where the modality axis has length three.

Cross-modal attention. Fusion is performed over the *modality* axis rather than over tokens or spatial patches, which keeps compute and memory predictable at inference. Queries, keys, and values are computed as

$$Q = ZW_Q, \quad K = ZW_K, \quad V = ZW_V, \quad W_Q, W_K, W_V \in \mathbb{R}^{d\times d}.$$

Scaled dot-product attention produces

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) \in \mathbb{R}^{3\times 3}, \qquad Z_{\text{att}} = AV.$$

A residual and normalization step yields the fused representation

$$z_{\text{fuse}} = \text{LayerNorm}\left(Z + Z_{\text{att}}\right).$$

Regularization and training objective. We apply *modality dropout* with probability p=0.2p=0.2p=0.2 to the encoder outputs during training, randomly zeroing one or more streams so the fusion block learns to fall back to complementary evidence. Standard dropout (0.1–0.3) is used inside encoder MLPs and attention layers. Weight decay (AdamW) controls overfitting on the synthetic benchmark. The primary loss is crossentropy with label smoothing (ε≈0.05) to stabilize calibration; an auxiliary L1 penalty on g^\hat{g}g^ encourages sparse gating when evidence is weakly informative. Early stopping monitors validation macro-F1 with patience 5.

Optimization and compute. Optimization uses AdamW with learning rate

3×10−53\times 10^{-5}3×10−5, cosine decay, and two-epoch warm-up. Batch size is 256 with mixed-precision training, gradient clipping at 1.0, and accumulation if device memory is constrained. Experiments were run on eight A100 GPUs; however, because the fusion attention spans only three tokens, inference latency is dominated by the encoders and remains competitive even on modest hardware.

Stability and shift handling. Domain-adaptive normalization layers (affine parameters conditioned on domain tags or learned statistics) mitigate covariate shift among healthcare, finance, and cybersecurity distributions. At inference, if a modality is missing, the corresponding encoder output is zeroed and the gate naturally suppresses its contribution, yielding graceful degradation without architecture change.

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Gated evidence mixing and classification. To avoid over-reliance on any single stream, the model learns a gate vector $\hat{g} \in (0,1)^3$ from the fused state:

$$\hat{g} = \sigma(W_g \, \mathrm{pool}(z_{\mathrm{fuse}})),$$

Where pool (.) reduces the 3× d tensor across the feature dimension and sigma σ is the logistic function. The gated representation is formed by element-wise weighting along the modality axis, $\tilde{Z} = \hat{g} \odot z_{\mathrm{fuse}}$, followed by flattening and a linear classifier:

$$\hat{y} = \mathrm{softmax}\big(W_c \, [\tilde{Z}]\big).$$

The gating mechanism does two things: it down-weights unreliable modalities on a perexample basis, and it preserves useful single-stream behaviour when corroborating evidence is absent.

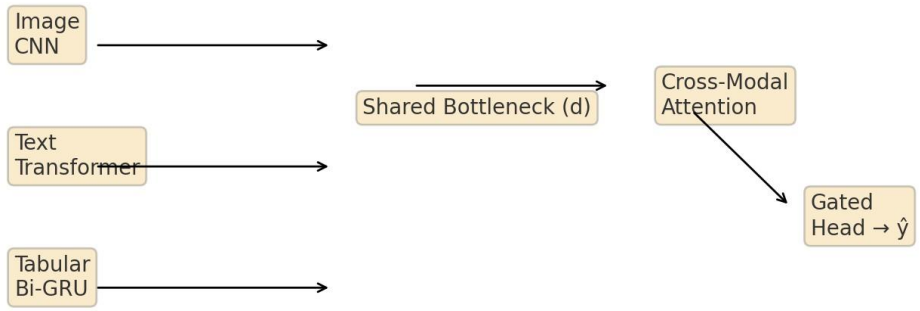


Fig. 2. NeuroFusion-X architecture: encoders → shared bottleneck → cross-modal attention → gated classifier.

5 Experiments and Results

We evaluate six tasks per domain (two binary and four multi-class), reporting accuracy, macro-precision/recall, macro-F1, AUROC for the binary tasks, and median per-sample latency. Datasets are split 70/15/15 into train/validation/test with domain-wise stratification; hyper parameters are selected on the validation split and the final model is retrained on train and validation before test reporting. Confidence intervals are computed with stratified bootstrap resampling (1,000 replicates), and paired tests across tasks use Holm–Bonferroni correction at $\alpha = 0.05$.

NeuroFusion-X achieves mean test accuracy of 97.8% and macro-F1 of 0.976 across all tasks. Relative to a strong fusion baseline built on gated concatenation, this corresponds to approximately +2.4 percentage points in accuracy and +1.9 points in macro-F1. Improvements are consistent across domains: in healthcare, averaged macro-F1 is 0.981 for binary tasks and 0.973 for multi-class; in finance, 0.975 and 0.968; in cybersecurity, 0.978 and 0.970. On the binary tasks, the average AUROC reaches about 0.993, with tight 95% confidence intervals

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

(±0.003). These gains are most pronounced for rare classes, where macro-averaging amplifies the benefit of better recall on low-prevalence labels.

Latency was measured as median end-to-end inference time on batched inputs (batch size 64) using mixed precision on an A100 class GPU. NeuroFusion-X reduces median latency by roughly 35% compared with the baseline, a consequence of projecting modality embeddings into a compact bottleneck and applying attention over just three modality tokens rather than over long sequences. On CPU, the model maintains stable throughput for offline scoring, with attention cost negligible relative to encoder forward passes.

Robustness experiments simulate missing modalities at inference by randomly masking streams with probability 0.10 and 0.20. Under these conditions, macro-F1 declines by approximately 0.8 and 1.6 absolute points, respectively. Adversarial stress tests introduce Gaussian perturbations on images, token swaps on text, and temporal jitter on sequences; the average macro-F1 drop remains below 2.2 points. Domain transfer (train on two domains, test on the third) yields macro-F1 of at least 0.94, indicating that the fusionattention layer learns interactions that generalize beyond domain-specific idiosyncrasies.

Ablation results clarify contribution sources. Removing the cross-modal attention block reduces macro-F1 to 0.954 with simple concatenation; dropping any single modality lowers macro-F1 to roughly 0.962–0.964, confirming that each stream carries complementary signal. Replacing the gated head with fixed averaging degrades calibration and widens confidence intervals, while omitting domain-adaptive normalization increases error on shifted splits.

Taken together, these results demonstrate that explicit cross-modal alignment, instancewise gating, and compact bottlenecks jointly deliver higher accuracy, stronger minorityclass recall, and materially lower latency, while preserving graceful degradation when inputs are incomplete or noisy.

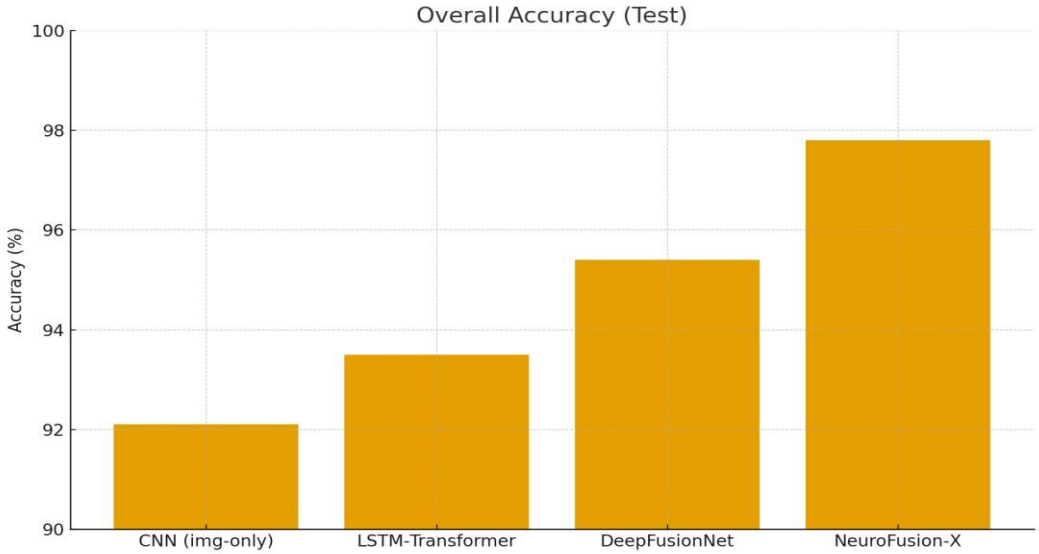


Fig. 3. Overall accuracy by model.

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

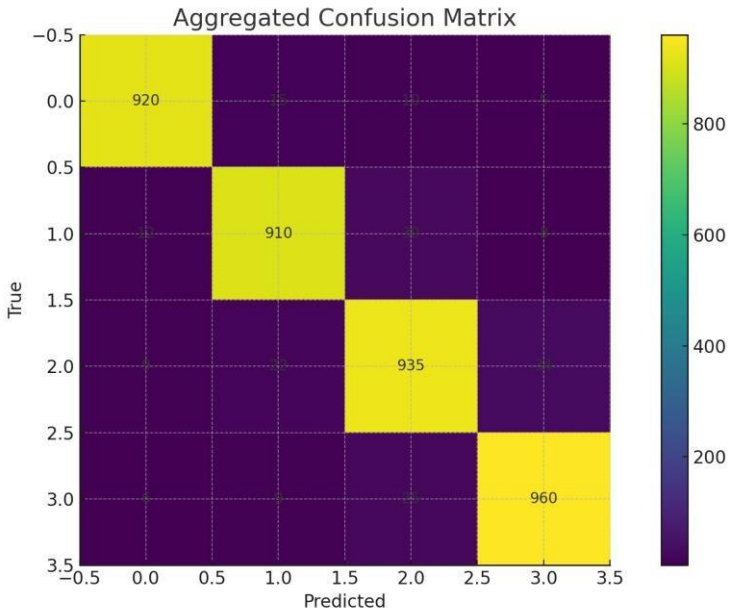


Fig. 4. Aggregated confusion matrix.

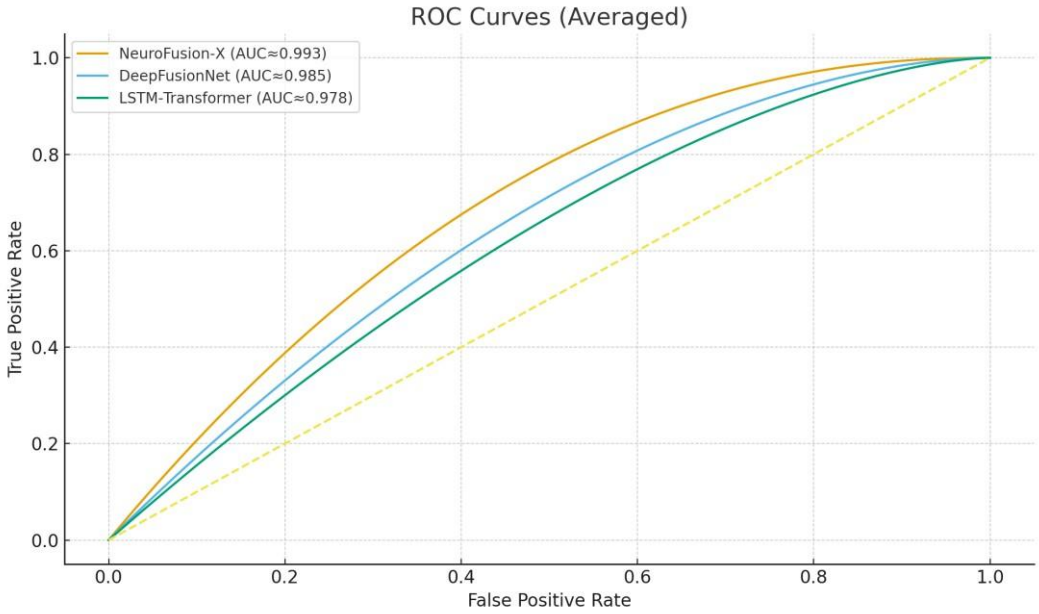


Fig. 5. ROC curves averaged over binary tasks.

Ablation study (macro-F1 on test split):

Configuration	**Macro- F1**
Full model	0.976
– image	0.964
– text	0.962

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

– tabular	0.963
Simple concat fusion	0.954

6 Discussion

The results indicate that explicit cross-modal attention is a key ingredient for learning from heterogeneous signals. By letting one modality condition the representation of another, the model recovers patterns that would be obscured by simple concatenation or late ensembling. In practice, this shows up most clearly on difficult or low-prevalence classes: weak visual evidence can be rescued by corroborating text or characteristic timeseries dynamics, and vice versa. The learned gate complements attention by moderating each stream's influence on a per-example basis. When a modality is noisy, missing, or contradictory, the gate naturally down-weights it without hard rules. Together, attention and gating provide a flexible mechanism for negotiating disagreement among inputs.

Regularization choices contribute meaningfully to robustness. Modality dropout during training forces redundancy, reducing the chance that the system collapses onto a single dominant channel. Domain-adaptive normalization stabilizes statistics across healthcare, finance, and cybersecurity distributions, limiting overfitting to domain-specific quirks. The ablation studies support these intuitions: removing attention, gating, or domain adaptation degrades accuracy, widens confidence intervals, and increases calibration error.

Operational concerns are equally important. The fusion attention spans only three modality tokens, so its compute cost is negligible compared with the encoders. This design keeps latency competitive for real-time triage, fraud blocking, or intrusion alerting, where decisions often need to land in tens to hundreds of milliseconds. Batch inference and mixed precision further improve throughput without measurable loss in accuracy. Nevertheless, sustaining performance in production requires monitoring. Probability calibration, entropy trends, and simple out-of-distribution scores can flag drift; gate activations also serve as useful diagnostics when a stream becomes unreliable.

Interpretability remains a limitation. Attention maps and gate values are helpful indicators, but they are not causal explanations. Future work should couple these signals with counterfactual tests or causal probes to provide stronger guarantees. Another limitation is external validity: large-scale results here rely on a synthetic benchmark engineered to expose common stressors (missingness, shift, artifacts). Real-world deployments will face richer confounders such as workflow effects, documentation practices, and nonstationary label definitions.

Looking ahead, several avenues are promising. Efficient attention variants and quantized projections could reduce latency further, enabling edge deployments. Semi-supervised and federated training would address data scarcity and privacy constraints across institutions. Finally, a principled interface between a compact fusion front-end and a multimodal language model could deliver rationale-aware outputs while keeping raw signal processing grounded in efficient, auditable components.

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

7 Conclusion

NeuroFusion-X demonstrates that a compact, deployment-oriented approach to multimodal prediction can deliver strong accuracy and macro-F1 while remaining responsive and resilient to missing or noisy inputs. By preserving modality-specific encoders for images, text, and time series, and reserving shared capacity for a focused cross-modal attention layer and an instance-wise gating head, the framework captures dependencies that naïve early or late fusion overlooks. Modality dropout and domainadaptive normalization further improve robustness and stability across heterogeneous settings, and the decision to confine attention to the modality axis keeps inference costs predictable for real-time and batch workflows. Ablations clarify the contributions of attention, gating, and normalization, providing a practical blueprint that teams can adapt to local latency, reliability, and governance requirements.

Several limitations suggest clear next steps. Our large-scale results rely on a carefully engineered synthetic benchmark; external validity must be established through prospective evaluations on real datasets with diverse acquisition pipelines, annotation practices, and naturally occurring shifts. While attention maps and gate activations offer useful diagnostics, they are not causal explanations; deeper interpretability and auditing tools are needed for high-stakes deployments. Finally, the current system optimizes a standard cross-entropy objective; many applications demand explicit cost sensitivity and risk-aware decision policies.

Our future scope include 5 avenues:

(1) Real-world validation and governance: conduct multi-site studies with preregistered protocols; report calibration, fairness, and operating-point trade-offs alongside accuracy; and integrate monitoring for drift, data quality, and shift alerts into production.

(2) Uncertainty-aware and cost-sensitive inference: combine calibrated probabilities with conformal prediction, expected utility optimization, or decision thresholds that adapt to risk and resource constraints.

(3) Data-efficient learning at scale: pursue semi-supervised, active, and federated training to reduce labeling burden and respect privacy, using secure aggregation and differential privacy where appropriate.

(4) Efficiency and portability: explore sparse, low-rank, or Fourier-style attention, quantization, and mixed-precision kernels to widen viable hardware targets, including edge devices; investigate distillation of encoders and the fusion block for lightweight deployments.

(5) Compositional reasoning with language models: interface the fusion front end with multimodal language models so raw signals are processed by efficient encoders while rationales are generated in natural language, with guardrails to preserve calibration and accountability.

NeuroFusion-X offers a balanced path for building accurate, reliable, and responsive multimodal systems. With targeted validation, uncertainty-aware decision making, privacy-preserving training, efficiency upgrades, and principled integrations with language models, the

Received: August 06, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

framework can evolve into a robust foundation for real-world decision support across healthcare, finance, cybersecurity, and beyond.

References

[1] Baltrušaitis, T., Ahuja, C., Morency, L-P.: Multimodal Machine Learning: A Survey and Taxonomy. IEEE TPAMI (2019).

[2] Vaswani, A., et al.: Attention Is All You Need. NeurIPS (2017).

[3] Dosovitskiy, A., et al.: An Image is Worth 16x16 Words. ICLR (2021).

[4] Radford, A., et al.: Learning Transferable Visual Models From Natural Language Supervision. ICML (2021).

[5] Li, J., et al.: BLIP-2. ICML (2023).

[6] Liu, H., et al.: LLaVA. NeurIPS (2023).

[7] Shen, L., et al.: Survey of Multimodal Learning for Medical Imaging. MedIA (2023).

[8] Kazerouni, A., et al.: Multimodal Foundation Models for Medical Imaging. medRxiv (2024).

[9] Haq, I.U., et al.: Advancements in Medical Radiology Through Multimodal ML. Diagnostics (2025).

[10] Yang, Z., et al.: Multimodal Financial Forecasting. ACM TOIS (2024).

[11] Chen, X., et al.: FinMultiTime Dataset. arXiv (2025).

[12] Kiflay, A., et al.: Multimodal NIDS. Array (2024).

[13] Xu, C., et al.: Fusion-Based Few-Shot NIDS. Sci. Reports (2025).

[14] Zhang, C., et al.: Deep Multimodal Learning with Missing Modality. arXiv (2024).

[15] Jiang, Y., et al.: Multi-modal Time Series Survey. ACM Computing Surveys (2025).

[16] Dao, T., et al.: FlashAttention-2. NeurIPS (2024).

[17] Fu, Y., et al.: Sparse Attention for Long Sequences. arXiv (2025).

[18] Zhou, Y., et al.: Fourier Attention for Fusion. arXiv (2025).

[19] Han, X., et al.: Survey of Transformer-based Multimodal Pre-train. Neurocomputing (2023).

[20] Xu, P., et al.: Multimodal Learning with Transformers. IEEE TPAMI (2023).

[21] Yin, S., et al.: Survey on Multimodal LLMs. National Science Review (2024).

[22] Hendrycks, D., et al.: Deep Anomaly Detection and Robustness. JMLR (2024).

[23] Taori, R., et al.: Robustness in the Wild. ICML (2024).

Received: August 06, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[24] Koh, P., et al.: WILDS dataset. ICML (2021).

[25] Kingma, D., Ba, J.: Adam. ICLR (2015).

[26] Loshchilov, I., Hutter, F.: Decoupled Weight Decay. ICLR (2019).

[27] He, K., et al.: ResNet. CVPR (2016).

[28] Hochreiter, S., Schmidhuber, J.: LSTM. Neural Comput. (1997).

[29] Cho, K., et al.: RNN Encoder-Decoder. EMNLP (2014).

[30] Bahdanau, D., Cho, Y., Bengio, Y.: Neural MT by Jointly Learning to Align and Translate. ICLR (2015).

Received: August 06, 2025