Emerging Trends in Digital Transformation

Volume-2, Issue-1, January-March 2026 ISSN: **3107- 4308** *Paper ID:* ***ETDT-V2-I1-60***

FusionMind: Reliability-Aware Adaptive Fusion for Multimodal Sentiment Classification

[*1]**Gundeti Mounika**, *Assistant Professor, School of CSE, Department of CSE*
Malla Reddy (MR) Deemed To Be University, Medchal-Malkajgirii, Telangana, India
E-mail ID: gmounika111@gmail.com

[*2]**Vimula Shwetha,** *Assistant Professor, School of CSE, Department of CSE*
Malla Reddy (MR) Deemed To Be University, Medchal-Malkajgirii, Telangana, India
E-mail ID: shwethavimula9@gmail.com

[*3]**Dr K Sridhar,***Associate Professor, School of CSE, Department of CSE,*
Malla Reddy(MR) Deemed to be University, Medchal,Telangana-500100, India
E-mail ID: sridhark529reddy@gmail.com

Abstract—Multimodal sentiment classification combines textual information, acoustic features, and visual expressions to identify the sentiment conveyed by a speaker. However, the reliability of these systems can decrease when one or more modalities contain noise, are unavailable, or provide conflicting information. To address these challenges, this paper proposes FusionMind, a reliability-aware hybrid deep learning framework for text, audio, and visual sentiment classification. The framework uses separate encoders for each modality, a shared representation space, and an adaptive gating mechanism that adjusts the contribution of each modality according to the input. A classification head generates sentiment predictions, while probability calibration is incorporated to improve the reliability of the predicted confidence scores. Modality dropout is included to examine the model's ability to handle incomplete inputs, and confidence-based selective prediction enables uncertain cases to be referred for further review. The proposed framework will be evaluated primarily on the CMU-MOSEI dataset, with CMU-MOSI used for additional evaluation. Comparisons will be conducted against unimodal classifiers, early-fusion methods, fixed-weight ensembles, and adaptive-fusion approaches. The evaluation will consider accuracy, macro-F1 score, negative log-likelihood, expected calibration error, selective risk, and inference latency. Since training and testing have not yet been conducted, no empirical performance improvements are claimed in this study.

Index Terms—Multimodal learning, sentiment classification, adaptive feature fusion, deep learning, model calibration, uncertainty estimation.

DOI: https://doi.org/10.65713/etdtv2i160

Volume-2, Issue-1, January-March 2026 ISSN: **3107- 4308** *Paper ID:* ***ETDT-V2-I1-60***

I. INTRODUCTION

Human communication conveys meaning through several complementary modalities. Language communicates semantic information, acoustic signals reveal characteristics such as pitch, tone, and intensity, and visual cues provide information through facial expressions and observable behavior. Combining these sources can help identify sentiment more effectively than relying on any single modality.

Multimodal sentiment analysis has become an important research area in natural language processing and affective computing. The CMU-MOSEI dataset provides aligned textual, acoustic, and visual information that supports research on sentiment analysis in video-based communication [1]. Previous studies have explored different approaches to combining these modalities, including low-rank multimodal fusion and recurrent multistage fusion [2], [3].

Despite these developments, three important challenges remain. First, the predictive value of each modality may vary across samples, making fixed fusion strategies less effective when one source contains limited or unreliable information. Second, models trained using complete multimodal inputs may perform poorly when one or more modalities are missing during inference. Third, predicted probabilities may not accurately reflect the model's actual confidence, limiting their usefulness in reliability-sensitive applications. To investigate these issues, FusionMind combines modality-specific encoders, adaptive gated fusion, modality dropout, and explicit reliability assessment within a unified framework.

A. Problem Statement

For each sample, the input is represented as (x_i = (x_i^t, x_i^a, x_i^v)), where the superscripts (t), (a), and (v) represent textual, acoustic, and visual features, respectively. The objective is to learn a model (f_\theta(x_i) \rightarrow \hat{y}_i) that predicts the sentiment associated with each sample while providing useful confidence estimates under both complete and incomplete input conditions.

B. Research Questions

RQ1: Does adaptive fusion achieve better predictive performance than unimodal models and conventional multimodal baselines?

RQ2: Does modality dropout improve robustness when one or more input modalities are unavailable?

RQ3: Does post-hoc probability calibration improve the reliability of the model's confidence estimates?

RQ4: What trade-offs exist between predictive performance, selective risk, and computational cost?

C. Proposed Contributions

This study proposes a modular multimodal learning architecture, a sample-adaptive gated fusion mechanism, an evaluation protocol for missing-modality conditions, and a reliability assessment strategy based on probability calibration and selective prediction. The effectiveness of these contributions will be determined through experiments. Since gated fusion and temperature

DOI: https://doi.org/10.65713/etdtv2i160

Volume-2, Issue-1, January-March 2026 ISSN: **3107- 4308** *Paper ID:* ***ETDT-V2-I1-60***

scaling are established techniques, neither method is presented as an independent methodological innovation.

II. RELATED WORK

A. Multimodal Sentiment Benchmarks

Bagher Zadeh et al. introduced the CMU-MOSEI dataset to support research in multimodal language analysis using textual, acoustic, and visual information, together with annotations for sentiment and emotion-related tasks [1]. The CMU-MOSI dataset provides a smaller benchmark consisting of opinion video segments annotated with sentiment intensity [4]. Both datasets offer useful resources for evaluating multimodal sentiment classification methods under standardized conditions.

B. Fusion Methods

Liu et al. proposed low-rank multimodal fusion, which models interactions between different modalities while reducing the computational cost associated with unrestricted tensor-based fusion [2]. Liang et al. introduced recurrent multistage fusion, a method that progressively combines multimodal representations across multiple stages [3]. These approaches provide important foundations for studying how information from different modalities can be integrated effectively. They also motivate comparisons between conventional feature concatenation, fixed-weight fusion, and learned adaptive fusion mechanisms.

C. Calibration and Research Gap

Guo et al. demonstrated that modern neural networks can produce poorly calibrated probability estimates and investigated temperature scaling as an effective post-processing technique for improving calibration [5]. This finding highlights the importance of evaluating confidence reliability separately from classification performance.

Building on these studies, FusionMind investigates whether combining adaptive fusion, modality dropout, probability calibration, and selective prediction can provide a practical balance between predictive quality, robustness to incomplete inputs, confidence reliability, and computational efficiency. The central research objective is to evaluate these capabilities within a common framework rather than treating each component as an independent solution.

III. PROPOSED FUSIONMIND FRAMEWORK

FusionMind is designed as a modular pipeline consisting of data validation, modality-specific encoders, feature projection and normalization, adaptive gated fusion, a classification head, and a reliability assessment stage. Each component performs a distinct function while contributing to the overall sentiment classification process.

The initial implementation uses pre-extracted, aligned features obtained through the CMU Multimodal SDK. This design allows the experiments to focus primarily on multimodal representation fusion and reliability assessment rather than on extracting features directly from raw audio, video, and text.

DOI: https://doi.org/10.65713/etdtv2i160

Emerging Trends in Digital Transformation
Volume-2, Issue-1, January-March 2026 ISSN: 3107- 4308 Paper ID: ETDT-V2-I1-60

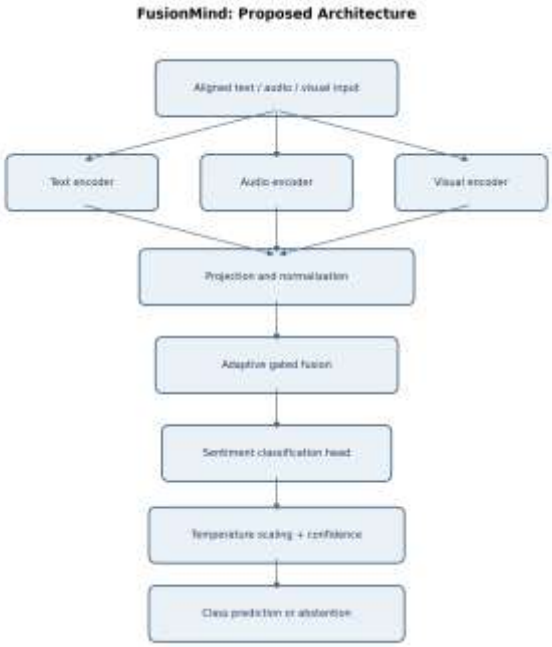


Fig. 1. Proposed FusionMind architecture. The figure illustrates the intended system design and does not represent a completed implementation.

A. Modality-Specific Encoders

Let (x_i^m) denote the input sequence for sample (i) and modality (m \in {t,a,v}), representing text, audio, and visual information, respectively. Each modality is processed by its corresponding encoder to generate a feature representation, expressed as (h_i^m = E_m(x_i^m; \theta_m)), where (E_m) denotes the encoder and (\theta_m) represents its learnable parameters. For sequential inputs, masked mean pooling is applied to aggregate the sequence representations while excluding padded positions from the calculation. The resulting pooled features are then projected into a shared (d)-dimensional representation space, allowing information from different modalities to be combined consistently.

$z_i^m = P_m(\text{Pool}(E_m(x_i^m; \theta_m)))$ (1)

The initial projection dimension is set to (d = 128). This value serves as a preliminary configuration and will be adjusted based on validation results rather than being treated as an experimentally established optimal setting.

B. Adaptive Gated Fusion

A learnable scoring function is applied to each projected modality representation to estimate its relative importance in the fusion process. These scores provide the basis for dynamically adjusting the contribution of textual, acoustic, and visual information for each input sample.

$e_i^m = w_m^T \tanh(W_m z_i^m + b_m) + c_m$ (2)

The modality scores are normalized using the softmax function to obtain sample-dependent weights. These weights determine the relative contribution of the textual, acoustic, and visual representations to the final fused feature vector, allowing the model to adjust the importance of each modality according to the input sample.

DOI: https://doi.org/10.65713/etdtv2i160

$\alpha_i^m = \exp(e_i^m) / \Sigma_r \exp(e_i^r), \quad \Sigma_m \alpha_i^m = 1$ (3)

The fused representation is:

$z_i^f = \Sigma_m \alpha_i^m z_i^m$ (4)

In experiments involving missing modalities, an availability mask (r_i^m \in {0,1}) is incorporated into the gating mechanism to indicate whether each modality is present for a given sample. The mask prevents unavailable modalities from contributing to the fusion process, allowing the model to combine the representations of the remaining available modalities.

$\alpha_i^m = r_i^m \exp(e_i^m) / \Sigma_r r_i^r \exp(e_i^r)$ (5)

At least one modality must be available for each sample to ensure that the fusion process can produce a valid representation. The learned gate weights indicate the relative contributions of different modalities to the fused representation; however, they should not be treated as direct explanations of the model's predictions without further validation.

C. Classification and Calibration

The fused representation is passed to a classification head that generates three output logits (o_i), corresponding to the sentiment classes. The softmax function is then applied to these logits to convert them into class probabilities, representing the model's estimated likelihood for each sentiment category.

$p_i(k) = \exp(o_{ik}) / \Sigma_j \exp(o_{ij}), \quad \hat{y}_i = \arg\max_k p_i(k)$ (6)

Temperature scaling is applied to adjust the confidence distribution of the model's predictions. By introducing a temperature parameter, the method modifies the sharpness of the output probabilities without changing the relative ordering of the class logits. The temperature is optimized using validation data to improve probability calibration and make the predicted confidence levels more consistent with the model's observed accuracy.

$p_i^T(k) = \exp(o_{ik} / T) / \Sigma_j \exp(o_{ij} / T), \quad T > 0$ (7)

The temperature parameter (T) is determined by minimizing the negative log-likelihood on the validation dataset. After calibration, a prediction is accepted when the maximum class probability, (\max_k p_i^T(k)), is greater than or equal to a predefined threshold (\tau). If the confidence score falls below this threshold, the system defers the prediction for further review. The threshold (\tau) is selected exclusively using validation data to prevent information from the test set from influencing the decision policy.

IV. METHODOLOGY AND IMPLEMENTATION

A. Dataset Selection and Labels

CMU-MOSEI is selected as the primary benchmark because it provides aligned textual, acoustic, and visual features for multimodal sentiment analysis. CMU-MOSI serves as a secondary dataset for additional evaluation. The exact number of samples and feature dimensions will be verified from the downloaded CMU Multimodal SDK files, as these values may vary depending on the feature version and preprocessing configuration.

For a sentiment score (s_i), the three-class labeling scheme is defined as follows: (y_i = 0) when (s_i < 0), (y_i = 1) when (s_i = 0), and (y_i = 2) when (s_i > 0). An additional binary

DOI: https://doi.org/10.65713/etdtv2i160

classification task may be conducted by excluding neutral samples and assigning label 0 to negative sentiment and label 1 to positive sentiment. The selected labeling strategy will be documented and applied consistently across all experimental configurations.

B. Preprocessing

The data will be obtained through the official CMU Multimodal SDK. Sample identifiers, sentiment labels, feature arrays, and official dataset partitions will be verified to ensure consistency and reproducibility.

- **Data validation:** Check for missing values, duplicate sample identifiers, variations in sequence lengths, and inconsistencies in modality feature dimensions.
- **Sequence preparation:** Pad or truncate sequences according to a predefined policy while retaining explicit validity masks to distinguish actual observations from padded positions.
- **Feature normalization:** Calculate normalization statistics using only the training data and apply the saved statistics consistently to the validation and test sets.
- **Reproducibility:** Store the processed tensors, label mappings, dataset split identifiers, and preprocessing configuration so that the experimental pipeline can be reproduced reliably.

C. Initial Training Configuration

Parameter	Proposed initial value
Framework	Python, PyTorch
Projected / hidden dimension	128 / 128
Activation / dropout	ReLU / 0.2
Batch size	64
Optimizer	AdamW
Initial learning rate	1×10^{-3}
Weight decay	1×10^{-4}
Maximum epochs / patience	100 / 10
Random seeds	13, 42, 87, 123, 2026
Checkpoint selection	Validation macro-F1
Calibration	Temperature scaling on validation data

The initial hyperparameter values serve as starting configurations rather than experimentally established optimal settings. Parameters such as the learning rate, hidden dimension, dropout rate, and batch size may be adjusted based on validation performance. The test dataset will remain completely separate from model development and will only be used for the final evaluation to ensure an unbiased assessment of the model's performance.

D. Training Objective and Workflow

For a training dataset containing (N) samples and (K = 3) sentiment classes, the model is trained by minimizing the multiclass cross-entropy loss. This objective measures the difference between the predicted class probabilities and the actual sentiment labels, encouraging the model to assign higher probabilities to the correct classes.

$$LCE = -(1/N) \Sigma_{i=1}^{N} \log p_i(y_i) \quad (8)$$

DOI: https://doi.org/10.65713/etdtv2i160

Volume-2, Issue-1, January-March 2026 ISSN: **3107- 4308** *Paper ID:* ***ETDT-V2-I1-60***

If the training dataset exhibits substantial class imbalance, a class-weighted loss function may be evaluated as an additional experimental condition. The class weights will be calculated exclusively from the training labels to prevent information leakage from the validation or test datasets. Modality dropout will also be incorporated during training by randomly masking selected combinations of available modalities, encouraging the model to learn robust representations and maintain predictive capability when one or more input modalities are unavailable.

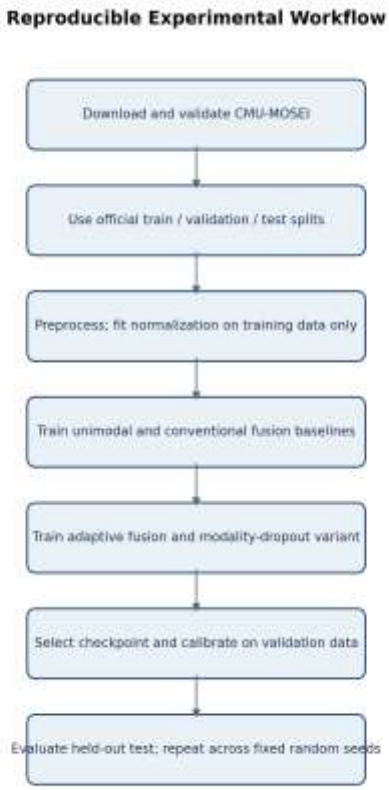


Fig. 2. Reproducible experimental workflow.

E. Reproducibility Controls

For reproducibility, all random seeds, software package versions, configuration files, model checkpoints, training logs, hardware specifications, and training durations will be documented. Random seeds will be set for Python, NumPy, and PyTorch, and deterministic operations will be enabled wherever practical. However, exact reproducibility cannot be guaranteed across different hardware configurations and software environments because certain operations may produce variations in their results [6].

V. EXPERIMENTAL DESIGN

The proposed FusionMind framework will be compared against seven configurations to evaluate the contribution of each modality and fusion strategy.

DOI: https://doi.org/10.65713/etdtv2i160

- **B1: Text-only classifier.** Uses textual features alone to predict sentiment.
- **B2: Audio-only classifier.** Relies exclusively on acoustic features for sentiment classification.
- **B3: Visual-only classifier.** Uses visual features to identify sentiment.
- **B4: Early-fusion model.** Combines the features from all available modalities through concatenation before classification.
- **B5: Fixed-weight probability averaging.** Combines the predicted class probabilities from separate modality-specific classifiers using predefined weights.
- **B6: Adaptive-fusion model without modality dropout.** Uses learned, input-dependent fusion weights but is trained with complete multimodal inputs.
- **B7: FusionMind.** Integrates adaptive fusion, modality dropout, probability calibration, and optional selective prediction to improve robustness and assess prediction reliability.

All models will use the same official dataset partitions, consistent label mappings, and comparable model-selection procedures to ensure a fair evaluation. Parameter counts and inference latency will also be reported alongside predictive metrics to provide a clearer comparison of model performance and computational requirements.

B. Evaluation Metrics

Accuracy and macro-F1 are primary predictive metrics. Accuracy is the fraction of correct predictions:

$$\text{Accuracy} = (1/N)\,\Sigma_i \mathbb{1}(\hat{y}_i = y_i) \quad (9)$$

Macro-F1 is the unweighted mean of the per-class F1 scores:

$$F1_macro = (1/K)\,\Sigma_k F1_k \quad (10)$$

Negative log-likelihood measures the probability assigned to the true label:

$$NLL = -(1/N)\,\Sigma_i \log p_i(y_i) \quad (11)$$

Expected calibration error compares accuracy and mean confidence across B confidence bins:

$$ECE = \Sigma_\beta\,(|S_\beta|/N)\,|acc(S_\beta) - conf(S_\beta)| \quad (12)$$

For confidence $c_i = \max_k p_i^T(k)$, coverage at threshold τ is:

$$\text{Coverage}(\tau) = (1/N)\,\Sigma_i \mathbb{1}(c_i \geq \tau) \quad (13)$$

Selective risk is the error rate among accepted predictions:

$$\text{Risk}(\tau) = \Sigma_i \mathbb{1}(c_i \geq \tau)\mathbb{1}(\hat{y}_i \neq y_i) \,/\, \Sigma_i \mathbb{1}(c_i \geq \tau) \quad (14)$$

ECE depends on binning choices; report the binning protocol and NLL as well. Calibration does not establish robustness under arbitrary distribution shift.

DOI: https://doi.org/10.65713/etdtv2i160

C. Missing-Modality and Ablation Experiments

Evaluate the model using complete multimodal inputs and all six nonempty subsets of the three modalities. Compare the adaptive-fusion model trained on complete inputs with the model trained using modality dropout. Conduct ablation studies by replacing the learned gate with equal weights, substituting concatenation for weighted summation, disabling modality dropout, removing probability calibration, disabling selective prediction, and eliminating one modality at a time.

D. Statistical Analysis

Repeat model training using five predefined random seeds and report the mean and standard deviation of each evaluation metric. Where appropriate, calculate paired bootstrap confidence intervals to assess the reliability of performance differences. Because multiple utterances may originate from the same speaker, speaker-grouped resampling should be preferred whenever feasible. Define the statistical tests and multiple-comparison correction procedures before interpreting the results.

VI. RESULTS AND DISCUSSION

Result status: No training or testing experiments have been conducted yet. All result fields remain pending and must be populated only with verified experimental outputs.

A. Main Benchmark Comparison

Method	Accuracy (%)	Macro-F1 (%)	NLL	ECE
Text-only	Pending	Pending	Pending	Pending
Audio-only	Pending	Pending	Pending	Pending
Visual-only	Pending	Pending	Pending	Pending
Early fusion	Pending	Pending	Pending	Pending
Fixed-weight averaging	Pending	Pending	Pending	Pending
Adaptive fusion	Pending	Pending	Pending	Pending
FusionMind	Pending	Pending	Pending	Pending

Table I. CMU-MOSEI test results template; values require execution.

B. Missing-Modality Results

Available modalities	Adaptive fusion macro-F1	FusionMind macro-F1
Text + audio + visual	Pending	Pending
Text only	Pending	Pending
Audio only	Pending	Pending
Visual only	Pending	Pending
Text + audio	Pending	Pending
Text + visual	Pending	Pending
Audio + visual	Pending	Pending

Table II. Missing-modality evaluation template.

DOI: https://doi.org/10.65713/etdtv2i160

C. Efficiency Measurements

Method	Parameters	Training time	Latency	Peak memory
Early fusion	Pending	Pending	Pending	Pending
Adaptive fusion	Pending	Pending	Pending	Pending
FusionMind	Pending	Pending	Pending	Pending

Table III. Computational measurements template.

D. Interpretation Protocol

If adaptive fusion achieves a higher macro-F1 score than early fusion and fixed-weight probability averaging, this would support the use of sample-dependent fusion for the task. If modality dropout improves performance when one or more modalities are unavailable, it would provide evidence for training with incomplete multimodal inputs. Similarly, if temperature scaling reduces negative log-likelihood (NLL) or expected calibration error (ECE), it would indicate improved calibration under the evaluated data distribution. All conclusions must be drawn from measured experimental results, with uncertainty estimates and computational costs taken into account.

VII. LIMITATIONS AND FUTURE WORK

The initial study relies on pre-extracted, aligned features rather than end-to-end encoders operating directly on raw text, audio, and video signals. Results obtained from CMU-MOSEI and CMU-MOSI alone cannot establish that the framework will generalize to unrelated domains. Furthermore, modality dropout cannot represent every possible form of sensor noise or data corruption, and calibration performance depends on the data used for calibration. Future research could explore explicit modality-quality estimation, ensemble-based uncertainty estimation, corruption-aware training strategies, cross-dataset transfer, and encoders that process raw signals directly.

VIII. CONCLUSION

This paper presents FusionMind, a proposed reliability-aware framework for multimodal sentiment classification. The framework combines textual, acoustic, and visual representations through adaptive gated fusion and incorporates modality dropout, probability calibration, and selective prediction. CMU-MOSEI serves as the primary benchmark, while CMU-MOSI is used as a secondary benchmark. The proposed methodology specifies the label-mapping procedure, training configuration, baseline models, reliability metrics, ablation studies, and reproducibility controls. However, the hypothesis that these components jointly achieve a favorable balance between predictive performance and trustworthiness remains unverified until the planned experiments are conducted and their results are evaluated.

REFERENCES

[1] A. Zadeh et al., “Tensor Fusion Network for Multimodal Sentiment Analysis,” in Proc. EMNLP, 2017, pp. 1103–1114. https://doi.org/10.18653/v1/D17-1115

[2] K. K. Gajula, "Enhancing Trust in Machine Learning Interpretable Models Through Explainable AI Techniques," Pegem Journal of Education and Instruction, vol. 13, no. 4, pp. 909–915, 2023.

[3] A. Zadeh et al., "Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph," in Proc. ACL, 2018, pp. 2236–2246. https://doi.org/10.18653/v1/P18-1208

[4] K. K. Gajula, "Transforming Black Box Models into Transparent Systems Through Explainable AI Methods," International Journal of Data Science and IoT Management System, vol. 4, no. 2, pp. 27–33, 2025.

[5] Y.-H. H. Tsai et al., "Multimodal Transformer for Unaligned Multimodal Language Sequences," in Proc. ACL, 2019, pp. 6558–6569. https://doi.org/10.18653/v1/P19-1656

[6] K. K. Gajula, "Designing an Ethical Data Science Framework for Responsible and Transparent AI Integration," Emerging Trends in Digital Transformation, vol. 1, no. 1, pp. 74–87, 2025.

[7] R. Lin and H. Hu, "MissModal: Increasing Robustness to Missing Modality in Multimodal Sentiment Analysis," Transactions of the Association for Computational Linguistics, vol. 11, pp. 1686–1702, 2023. https://doi.org/10.1162/tacl_a_00628

[8] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in Proc. ICML, vol. 70, 2017, pp. 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

[9] D. Murali and K. K. Gajula, "Deep Learning Image Processing for IT Professionals to Detect Stress," Mukt Shabd Journal, vol. 12, no. 10, pp. 1043–1049, 2023.

[10] "Multimodal Sentiment Analysis Network Based on Distributional Transformation and Gated Cross-Modal Fusion," in Proc. International Conference on Networking and Network Applications (NaNA), 2024. https://doi.org/10.1109/NaNA63151.2024.00088

[11] K. K. Gajula, "Reinforcement Learning with Transparent Policies: An Explainable AI Approach to Adaptive Cyber Security," American Journal of AI Cyber Computing Management, vol. 5, no. 4, pp. 322–328, 2025.

[12] K. K. Gajula, "A Hybrid Interpretable AI Framework for Detecting Financial Fraud in Dynamic Transaction Networks," International Journal of Communication Networks and Information Security, 2022.

[13] "Multimodal Sentiment Analysis—A Comprehensive Survey From a Fusion Methods Perspective," IEEE Access, vol. 13, pp. 64556–64583, 2025. https://doi.org/10.1109/ACCESS.2025.3554665

[14] K. K. Gajula, "AI-Augmented Web Development: A Competency Framework for the Next Generation Workforce," Journal of Science Engineering Technology and Management Science, 2025.

[15] "CorMulT: A Semi-Supervised Modality Correlation-Aware Multimodal Transformer for Sentiment Analysis," IEEE Transactions on Affective Computing, vol. 16, no. 3, pp. 2321–2333, 2025. https://doi.org/10.1109/TAFFC.2025.3559866

[16] K. K. Gajula, "XAI-Driven Green Cloud Intelligence for Transparent and Optimized Resource Allocation in Cloud Computing," International Journal of Research in Engineering and Management Sciences, vol. 2, no. 6, pp. 14–29, 2026.

[17] "Learning Three-Modal Joint Representations With Multimodal Transformer for Multimodal Sentiment Analysis," IEEE Transactions on Consumer Electronics, vol. 71, no. 3, pp. 7559–7569, 2025. https://doi.org/10.1109/TCE.2025.3593333

[18] K. K. Gajula, "Advanced Stress Level Forecasting Using Quantile-Optimized Regression Forest Models," International Journal of Engineering Research and Science Technology, vol. 21, no. 4(1), pp. 1–7, Dec. 2025. https://doi.org/10.62643/ijerst.2025.v21.n4(1).pp1-7

[19] "Hierarchical Signal Calibration and Refinement for Multimodal Sentiment Analysis," IEEE Signal Processing Letters, vol. 32, pp. 3450–3454, 2025. https://doi.org/10.1109/LSP.2025.3603884

[20] K. K. Gajula, Programming With Artificial Intelligence. Hyderabad, India, 2023.

[21] M. Chen and X. Li, "SWAFN: Sentimental Words Aware Fusion Network for Multimodal Sentiment Analysis," in Proc. 28th International Conference on Computational Linguistics (COLING), 2020, pp. 1067–1077. https://doi.org/10.18653/v1/2020.coling-main.93

[22] T. Rajani, K. K. Gajula, M. A. Shaik, Y. Sneha, A. K. Vootla, and K. S. Prasad, "ResUNet-X: Next-Generation Residual U-Net for Accurate Cardiac Imaging," in Proc. International Conference on Signal Processing, Computation, Electronics and Communication Technologies, 2025.

[23] "An Active Learning-Based Alternative Reinforcement Contextual Information Fusion Model for Multimodal Sentiment Analysis," IEEE Transactions on Audio, Speech and Language Processing, vol. 33, pp. 3537–3552, 2025. https://doi.org/10.1109/TASLPRO.2025.3597474

[24] K. K. Gajula, "Intelligent AI Systems and Advanced Machine Learning: Recent Advances and Real-World Applications," Journal of Intelligent Decision Making and Information Science, vol. 3, no. 7s, pp. 1831–1845, 2026.

[25] "Text-Anchored Dual Contrastive Alignment for Multimodal Sentiment Analysis With Temporal Misalignment," IEEE Intelligent Systems, vol. 41, no. 5, pp. 56–64, 2026. https://doi.org/10.1109/MIS.2026.3718547

DATA AND SOFTWARE RESOURCES

CMU-MOSEI: https://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/
CMU-MOSI: https://multicomp.cs.cmu.edu/resources/cmu-mosi-dataset/
CMU Multimodal SDK: https://github.com/CMU-MultiComp-Lab/CMU-MultimodalSDK
PyTorch: https://pytorch.org/