Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Dynamic Adaptive Multimodal Graph Fusion Network for Multi-Hazard Recognition Using Remote Sensing, Social Media, IoT, and Seismic Observations

Dannuri Mounika
SR University
Johnson Kolluri
SR University
Sunil Srinivas
CSE-AIML, TKR College of Engineering and Technology
Kiran Siripuri
skiran@sithyd.siu.edu.in

Symbiosis International University

Article

Keywords: Dynamic multimodal fusion, Graph neural networks, Natural hazard recognition, Remote sensing, Internet of Things, Adaptive modality weighting

Posted Date: September 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10714749/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Dynamic Adaptive Multimodal Graph Fusion Network for Multi-Hazard Recognition Using Remote Sensing, Social Media, IoT, and Seismic Observations

Dannuri Mounika[1,2], Johnson Kolluri[1], Sunil Srinivas[2], Kiran Siripuri[3]*

1. *School of Computer Science & Artificial Intelligence, SR University, Warangal, Telangana, India - 506371*
2. *CSE-AIML, TKR College of Engineering and Technology, Hyderabad, India- 500097*
3. *Symbiosis Institute of Technology, Hyderabad Campus, Symbiosis International (Deemed University), Pune, India*

*Corresponding author: Kiran Siripuri

skiran@sithyd.siu.edu.in

Abstract

Reliable recognition of natural hazards remains challenging because the relevance of remote sensing, environmental sensors, social media and seismic monitoring is heterogeneous and the derived evidence varies as events evolve. This study presents a Dynamic Adaptive Multimodal Graph Fusion Network (DAMGF-Net) for recognising earthquakes, floods, fires, smoke, and normal conditions by jointly exploiting these complementary information sources. Unlike fixed multimodal fusion, DAMGF-Net constructs dynamic cross-modal relationships and adaptively weights each modality according to the available evidence relevant to the prediction task. The resulting graph representation captures spatial, temporal, and semantic dependencies while enabling the contribution of individual modalities to be examined. Experiments comprised 68,308 remote-sensing images, 45,384 social-media posts, 59,276 Internet-of-Things sensor records, and 1,300 seismic events. DAMGF-Net achieved 96.27% test accuracy, a weighted F1-score of 0.9618, a Matthews correlation coefficient of 0.9497, and a mean multiclass ROC-AUC of 0.9944. The model further processed 88.6 images per second with 11.3 ms inference latency. These findings show that dynamically adapting cross-modal relationships and modality contributions provides an effective and interpretable approach to multi-hazard recognition, with potential for near-real-time disaster monitoring, early warning, and resilient decision support.

Keywords: Dynamic multimodal fusion, Graph neural networks, Natural hazard recognition, Remote sensing, Internet of Things, Adaptive modality weighting.

1. Introduction

1.1 Background

Earthquakes, floods, wildfires, and landslides, among other natural disasters, continue to wreak severe damage across the globe, resulting in significant human, economic, and environmental damage. The rising frequency of extreme weather events, rapid urbanization, and climate change has led to an increased demand for intelligent disaster monitoring and early warning systems capable of providing accurate and timely decision support [5,8–10]. Recent advances in artificial intelligence, particularly in deep learning, have changed the way disaster prediction is performed by enabling the automatic extraction of features from heterogeneous data sources, including remote sensing imagery, seismic data, IoT sensor networks and social media streams [2,4,7,11]. Deep learning has achieved substantial advances in individual disaster domains. CNN-based and Transformer methods have greatly advanced earthquake early warning methods by learning

discriminative temporal features from seismic observations [3, 13, 14]. Similarly, deep neural networks have been utilized to enhance forest-fire detection and smoke recognition in UAV and satellite images, achieving high classification accuracy under various environmental conditions [16–19]. In flood monitoring, optimized convolutional architectures and remote sensing imagery have provided dependable flood mapping and damage assessment [26]. Foundation models have also expanded the applicability of pretrained geospatial representations for multiple Earth-observation tasks [27]. Disaster informatics has also undergone parallel developments. In particular, BERT, a transformer-based language model, has greatly improved the understanding of textual information related to disasters for emergency response and resource management [20–22]. Moreover, multimodal learning frameworks that exploit images and textual information jointly have shown better performance than single-modality approaches for disaster-event classification and crisis understanding [23–25,28]. These studies confirm that the combination of complementary information sources can significantly improve disaster recognition in comparison to the analysis of individual modalities independently. Despite these advances, several important challenges remain. Most existing studies focus on a single hazard type or a limited combination of sensing modalities, which makes them difficult to generalize to complex real-world disaster scenarios. Moreover, numerous multimodal methods employ static feature fusion strategies that cannot adapt dynamically to the varying quality of data and the relevance of modalities for different disaster scenarios. Existing methods are also difficult to interpret and do not allow emergency responders to understand how different sources of information contribute to prediction decisions.

1.2 Motivation

To address the above limitations, this study introduces a dynamic adaptive multimodal graph fusion network (DAMGF-Net) for intelligent disaster recognition and decision support. The proposed framework integrates satellite imagery, disaster-related text, IoT sensor observations and seismic parameters within a unified graph learning architecture. Unlike traditional multimodal methods, the proposed model learns cross-modal relationships for each disaster instance dynamically via adaptive graph attention, which can better fuse information and enhance interpretability. The framework intends to provide accurate, robust and explainable disaster intelligence in real-time emergency management through the joint learning of disaster classification and severity estimation.

1.3 Contributions

The key contributions of this study can be summarized as follows

- This study presents a novel dynamic adaptive multimodal graph fusion network (DAMGF-Net) for intelligent disaster recognition, which can simultaneously use satellite imagery, disaster-related texts, IoT sensor readings and seismic data within a single deep learning architecture. The integrated framework can evaluate heterogeneous emergency information comprehensively to support real-time decisions.
- This study proposes a dynamic cross-modal graph transformer (DCGT) with an input-dependent graph representation to capture semantic relationships across heterogeneous data streams. Unlike conventional fusion methods that adopt static feature aggregation, this graph structure is updated dynamically for each incident, allowing the network to effectively learn capture cross-modal interactions.
- During inference, we propose an adaptive cross-modal attention-based fusion (ACMAF) mechanism to measure the relative significance of each sensing

channel. This attention-based method improves the fusion of multimodal features and the interpretability of the system by revealing the contribution of each data source to the final result.

- This study proposes a multitask learning architecture for jointly predicting hazard type categorization and severity estimation. Joint optimization enables the network to learn more generalized and rich feature descriptions, thus improving the reliability of predictions in different emergency scenarios.
- The proposed approach is extensively evaluated using experimental evaluations on several real-world data observations, including satellite images, disaster-related social media posts, IoT sensor logs and seismic recordings. A benchmark comparison with existing state-of-the-art deep learning and fusion-based approaches confirms the superior accuracy, robustness, computational efficiency and explainability of the proposed system.

2. Related Work

This section discusses prior study on the five threads most closely relevant to the proposed framework—seismic early warning, remote-sensing image-based hazard detection, social-media crisis informatics, IoT and sensor-network disaster monitoring, and graph/transformer-based spatiotemporal and multimodal architectures before synthesizing the resulting research gap.

2.1 Seismic early warning and deep learning

Chovanec (2025) [1] proposed a component-based theoretical model for early warning systems by organizing the operational workflow into sensing, processing, decision-making, and dissemination components. The study emphasized that a modular architecture can improve system flexibility, maintainability, validation, and scalability in real-time warning applications. Zhexebay (2025) [3] implemented a convolutional neural network(CNN)-based technique for early P-wave classification via seismic observations collected from seven monitoring stations. The proposed method demonstrated improved resistance to noise and provided more reliable earthquake detection than conventional short-term average/long-term average triggering techniques. M"unchmeyer et al. (2021) [14] presented the transformer-based earthquake alerting model to implement the processing of multistation seismic waveform data for fast earthquake characterization on the basis of self-attention mechanism, the model could capture long-range temporal and spatial dependencies well, enabling stable estimation of earthquake magnitude and epicentral location for early warning. Such studies collectively demonstrate the usefulness of deep learning in seismic signal interpretation and early warning of earthquakes. However, most existing approaches are designed for earthquake monitoring and use only raw waveform data. Conversely, the proposed framework captures seismic events in terms of catalog-level parameters such as magnitude, depth and intensity and integrates them with satellite imagery, disaster-related text and IoT observations through a unified multimodal graph-learning architecture. This design supports multihazard classification, severity estimation and interpretable disaster decision support.

2.2 Remote sensing and image-based hazard detection

A deep learning-oriented UAV automated system for forest fire identification based on Reduce-VGGNet was developed by Wang et al. (2024) [16]. They demonstrated that lightweight convolutional networks can monitor wildfires in real time with 91.20% image classification accuracy. Saleh et al. (2024) [17] reviewed deep learning schemes for forest fire surveillance. The traditional CNN-based models have classification accuracies ranging from 88% to 96%.

However, more sophisticated architectures, such as FFireNet, obtain accuracies of over 98%, showing the increased effectiveness of deep learning for wildfire detection. Chodavarapu (2026) [18] presented an uncertainty-aware CBAM-based framework for uncertainty-aware wildfire smoke density classification system through CBAM-enhanced enhanced EfficientNet-B3 and evidential deep learning. The proposed method improved the classification of smoke density and quantified the uncertainty of prediction, which is better for monitoring wildfires under complex meteorological conditions. He et al. (2025) [19] examined deep learning-based object recognition for the recognition of forest fire smoke via YOLO architectures. They showed that object identification networks are able to identify early smoke from remote sensing images with high accuracy and mean average precision (mAP). Stateczny et al. [26] (2023) optimized a hybrid CNN-ResNet model for satellite flood detetion. Deep transfer learning and optimization algorithms provide a classification accuracy of 95% for flood mapping performance. These studies show that convolutional networks, object detection models and transfer learning have greatly improved remote sensing hazard recognition. Most techniques have been developed for one type of disaster via one sensing modality. Therefore, they are unable to successfully leverage satellite imagery, disaster-related text, IoT observations and seismic measures from real disasters. This paper presents a novel system that integrates diverse information sources in a dynamic multimodal graph-learning architecture for resilient, adaptive and explainable disaster identification.

2.3 Social media and text-based crisis informatics

Lamba et al. (2025) [20] proposed a BERT-based framework to analyze disaster-related textual information to support efficient emergency resource management. Their study demonstrated that the use of transformer-based contextual representations significantly enhances the understanding of disaster information compared with traditional text classification methods. Dharrao et al. (2024) introduced a disaster tweet classification model based on BERT embeddings and a modified convolutional neural network [22]. The experimental results show that contextual language representations and deep convolutional feature learning improve the detection of disaster-related social media posts. In 2025, Paul and Balabantaray [21] employed a hybrid BERT–Graph Attention Network for the classification of disaster tweets. In this work, we develop a model that integrates transformer-based semantic representations and graph attention learning to achieve superior classification performance while effectively modeling the semantic relationships among disaster-related messages. Recent advancements have pushed the boundary of transformer-based language models into multimodal disaster understanding. Ahmad et al. (2022) [23] reported that, compared with unimodal approaches, incorporating text and visual information via attention mechanisms significantly enhances the accuracy of crisis image classification. Dar et al. (2024) [24] proposed a social context-aware graph-based multimodal system to effectively analyze the relationships among images, text, and social context for disaster content classification. Similarly, Islam et al. (2025) [25] formulated a multimodal transformer model for multiclass disaster classification. El-Niss et al. (2024) [28] investigated federated multimodal learning for disaster-event classification on social media. These studies show that transformer-based language models and multimodal learning have made great progress in disaster information extraction and crisis understanding. However, most existing works focus on image-text fusion without considering complementary information from IoT sensors and seismic observations. In addition, textual information is usually regarded as an isolated modality without explicitly modeling its dynamical relationships with other heterogeneous data sources. To overcome identified limitations, the proposed model employs a pre-trained BERT feature encoder to learn textual representations and combines them with satellite imagery, IoT observations, and seismic parameters via a dynamic multimodal graph-

learning architecture that allows adaptive cross-modal interaction and more reliable disaster recognition.

2.4 IoT and sensor network disaster monitoring

Ravi (2025) [2] suggested a framework for disaster prediction using the Internet of Things (IoT) and graph neural networks (GNNs) for real-time early warning. The study revealed that modeling spatial relationships between distributed sensor nodes significantly improves prediction accuracy over processing individual sensor streams independently and allows neighboring sensors to validate abnormal observations before issuing disaster alerts. An artificial intelligence-based early warning system for natural disaster detection and mitigation based on a wireless sensing nodes was proposed by Charan et al. (2024) [11]. The reported findings showed that combining deep learning with IoT sensing provided more reliable detection and a lower false alarm rate than did conventional threshold-based monitoring. In 2025, Umadevi et al. [7] extended the mTCN framework to predict disasters by combining massive machine type communication with temporal convolutional networks. The proposed approach effectively modeled the temporal dependencies in sensor observations effectively, showing the potential of IoT-driven deep learning-based framework for real time disaster surveillance. These studies illustrate the importance of IoT sensing as a reliable source of real-time environmental information for disaster prediction. However, most current methods analyze IoT measurements alone or in sensor-only architectures, thus limiting their potential to exploit complementary evidence from remote sensing imagery, disaster-related text, and seismic observations. To address these limitations, the proposed framework integrates IoT telemetry as one of four modalities in a dynamic multimodal graph-learning architecture that allows for adaptive cross-modal interaction and thus more reliable disaster recognition and decision support.

2.5 Graph neural networks, transformers, and multimodal fusion for spatiotemporal disaster data

Recently, intelligent disaster recognition has received increasing attention, with the integration of graph neural networks, transformers and multimodal learning. Umadevi et al. (2025) [7] proposed an mTCN disaster prediction architecture that utilized temporal convolutional networks and large machine-type communications. Temporal deep learning extracts dynamic patterns in IoT sensor readings to monitor real time hazards. Ahmad et al. (2022) designed an attention-based paradigm for crisis picture classification using linguistic and visual information for multimodal disaster analysis [23]. The experimental results demonstrate that multimodal feature fusion is better at capturing complementary semantic information than the image-only and text-only methods are. In another social context-aware graph-based multimodal attentive learning paradigm, Dar et al. [24] (2024) treated textual and visual representations as graph nodes. Their graph-attention mechanism outperformed classical techniques of feature concatenation for disaster content classification. More recently, [28] El-Niss et al. [22] and Dharrao et al. (2024) introduced a model for classifying disaster tweets via the combination of BERT embeddings and a tuned CNN. The experimental results indicate that the integration of contextual language representations and deep convolutional feature learning can substantially enhance the identification of disaster-related social media posts. In 2025, Paul and Balabantaray [21] utilized a hybrid BERT-Graph Attention Network (BERT-GAT) for the classification of disaster tweets. These studies show that graph learning, transformer architectures and multimodal fusion may significantly improve disaster recognition by utilizing complementary information from multiple data sources. However, current methods typically combine only two modalities, most often images and text, with rigid fusion schemes that are not capable of adapting to different kinds of disaster events. However, the proposed framework integrates satellite

imagery, disaster-related text, IoT sensor observations and seismic parameters along with a dynamic graph-learning architecture to adaptively model cross-modal relationships for each disaster event to present a more comprehensive and interpretable disaster decision-support system.

2.6 Explainability in disaster-recognition AI

In a comprehensive review of artificial intelligence in multihazard early warning systems, Tiggeloven et al. (2025) [5] reported that improvements in forecasting performance, explainability, transparency and user trust are barriers to the adoption of AI. The authors noted that explainable and human-centered artificial intelligence systems allow emergency responders to understand and justify forecast outcomes in time-critical decision-making. To improve model transparency, recent studies have increasingly adopted post hoc explainability techniques, including saliency maps, SHAP analysis, and visualization of attention patterns. Although these approaches providing valuable insights into model behavior, they generate explanations after prediction and therefore do not directly represent the internal decision-making process of the model. Motivated by this limitation, the proposed framework incorporates a dynamic graph constructor that performs adaptive cross-modal feature integration through a learned graph representation. Because the graph adjacency matrix directly determines how information is exchanged among satellite imagery, disaster-related text, IoT sensor observations, and seismic parameters during inference, the learned graph naturally provides interpretable cross-modal relationships. Consequently, the proposed framework offers enhances interpretability by revealing the relative importance of each modality to the final model output, thereby improving both model transparency and decision support for disaster management.

2.7 Synthesis and research gap

Recent studies have demonstrated that multimodal deep learning can improve disaster assessment by combining textual, visual, and geospatial information [29–32,36,37]. Remote sensing and social media data fusion have also supported flood-response prioritization and disaster situation awareness [30,37]. Moreover, hierarchical graph learning, graph-based remote-sensing analysis, and spatiotemporal knowledge graphs have enabled the representation of relationships among disaster observations, locations, and response activities [33,34,38]. However, most existing approaches are designed for specific hazards, use only two modalities, or employ static fusion strategies that cannot adapt to variations in data relevance, quality, and availability. Hazard-specific systems, such as UAV-based forest-fire monitoring, further provide limited generalizability across disaster categories [35]. Open-world disaster identification has recently been investigated via multimodal social media data [39] nevertheless, recognizing unseen hazard conditions and handling noisy or missing modalities remain underexplored. Existing models also provide limited architectural interpretability regarding the contribution of each modality. Therefore, DAMGF-Net integrates satellite imagery, disaster-related text, IoT observations, and seismic parameters through dynamic graph construction and reliability-aware adaptive fusion. This design aims to improve multihazard recognition, robustness to incomplete data, generalization to unfamiliar conditions, and transparency of model decisions.

As shown in Table 1. Positioning of representative prior work relative to modality coverage, fusion topology, explainability mechanism, and task structure.

Table 1 Reviewed studies and comparison with the proposed dynamic adaptive multimodal graph fusion network.

Study/system	Dataset	Modalities fused	Fusion topology	Intrinsic explainability	Multi-task
Chovanec [1] (2025)	Early warning records	Seismic (conceptual)	—	No	No
Ravi [2] (2025)	IoT + weather dataset	IoT + GNN	Fixed graph	Partial (attention)	No
Zhexebay [3] (2025)	Seismic waveform datasets	Seismic (single)	—	No	No
Münchmeyer et al. [15] (2021)	Landslide inventories	Seismic (multi-station)	—	Partial (attention)	No
Text+image fusion [23-25,28] (2021-2025)	CrisisMMD + Social Context	2 (text, image)	Fixed / static cross-attention	Post-hoc only	No
BERT-GAT [20] (2025)	Disaster text datasets	1 (text) + graph	Fixed corpus graph	Partial (attention)	No
Umadevi et al. [7] (2025)	IoT + environmental datasets	IoT telemetry	TCN (fixed)	No	No
Proposed DCGT (this work)	Multimodal Disaster Dataset (Remote Sensing + Social Media + IoT + Seismic Observations)	4 (image, text, IoT, seismic)	Dynamic, input-conditioned graph	Yes (learned adjacency)	Yes (class + severity)

3. Proposed Framework: Dynamic adaptive multimodal graph fusion network

The proposed end-to-end multimodal architecture for natural hazard recognition, severity estimation, and confidence-aware decision making. Modality-specific encoders are used to learn discriminative representations from heterogeneous inputs, including remote-sensing images, social media text, IoT sensor data, and seismic features. We adaptively fuse the extracted embeddings through a cross-modal attention mechanism and then perform dynamic graph construction and spatial–temporal graph transformer-based relational reasoning to capture intermodal dependencies. Finally, the fused representation is fed into a multitask prediction head that performs hazard classification, severity estimation and confidence prediction at the same time. The proposed approach combines modality-specific feature learning, adaptive fusion, and dynamic graph reasoning into a unified framework, effectively leveraging the complementary information from multiple sensing modalities for robust and reliable disaster prediction and decision support. The dynamically generated graph is processed by the dynamic spatial–temporal graph transformer (STGT). Graph aggregation propagates complementary information among modality nodes, whereas the transformer encoder captures higher-order semantic dependencies through multihead self-attention. Mean pooling is then applied to produce a graph-contextualized feature representation that summarizes the learned multimodal relationships. The outputs of the ACMAF module and the STGT module are combined through elementwise feature fusion to obtain the final multimodal representation. This representation contains both modality-level importance information and graph-based relational knowledge, providing a comprehensive description of the disaster event. Finally, the fused representation is supplied to a multitask prediction head consisting of three parallel branches. The hazard classification branch

predicts the classification of the disaster, the severity regression branch predicts the severity of the disaster, and the confidence branch provides an extra confidence score for each prediction. The joint optimization of these tasks allows the proposed framework to learn generalized multimodal representations with improved prediction accuracy, robustness, and interpretability for real-time disaster management applications.

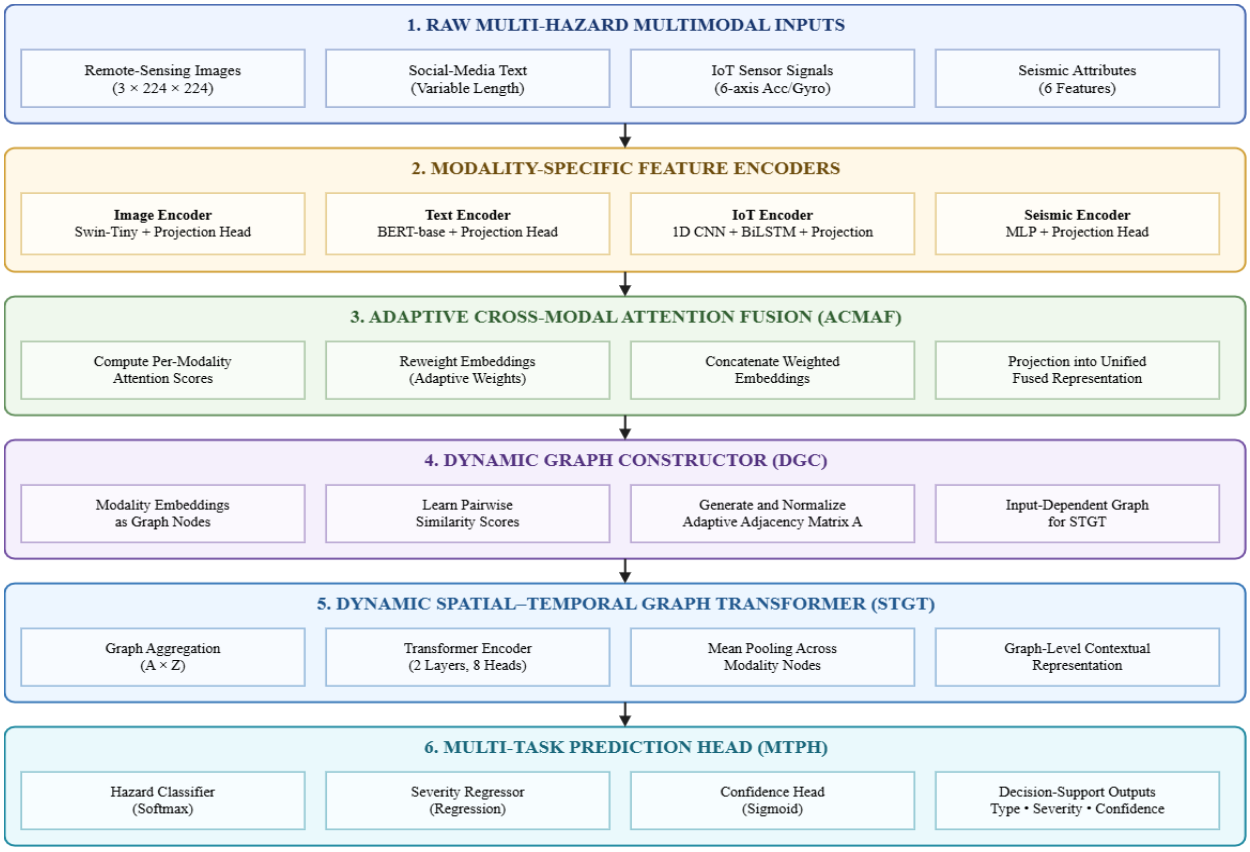


Figure 1 System design of the proposed dynamic adaptive multimodal graph fusion network

Figure 1 illustrates the overall process of the proposed dynamic adaptive multimodal graph fusion network (DAMGF-Net) for multimodal disaster recognition and severity estimation. The framework integrates four heterogeneous data sources remote sensing imagery, disaster-related social media text, IoT sensor observations and seismic parameters, into a unified graph learning architecture to exploit complementary information for intelligent disaster decision support. The raw multimodal data are first collected from four independent sensing modalities. The image branch collects remote sensing images of spatial resolution $3 \times 224 \times 224$, the text branch handles disaster-related social media posts of variable length, the IoT branch collects six-dimensional accelerometer and gyroscope measurements, and the seismic branch processes six numerical parameters describing earthquake characteristics, including magnitude, Modified Mercalli Intensity (MMI), depth, Community Decimal Intensity (CDI), event significance and alert level. Each modality is then processed by a dedicated feature extraction network. We use a pretrained Swin-Tiny transformer to encode remote sensing images. Textual information is represented via BERT-base. IoT measurements are encoded by a 1D CNN–BiLSTM encoder. Seismic observations are encoded with a multilayer perceptron (MLP). We use different encoder architectures depending on the properties of each modality. All the extracted features, however, are projected to a common latent embedding space of size 512 to enable heterogeneous feature integration. The generated modality embeddings are then forwarded to the adaptive cross-modal attention fusion (ACMAF) module. Rather than assuming equal importance for every modality, the ACMAF automatically learns attention weights that quantify the role of each modality in the

current disaster event. The learned attention-weighted embeddings are subsequently fused to generate an initial multimodal representation that emphasizes the most relevant sensing sources while reducing noisy or less relevant significance. To further model relationships among heterogeneous modalities, the modality embeddings are interpreted as nodes of a graph. The dynamic graph constructor learns an input-dependent adjacency matrix that represents semantic interactions among image, text, IoT, and seismic features. Unlike conventional graph networks with fixed connectivity, the proposed adjacency matrix is reconstructed dynamically for every disaster instance, enabling adaptive modeling of event-specific cross-modal relationships.

3.1 Problem Formulation

Natural disaster prediction necessitates the integration of heterogeneous sources collected from diverse sensing modalities. Let a disaster event be represented by a multimodal observation

$$X = \left(X_{img},\ X_{txt},\ \ X_{iot}\ X_{seis}\right),$$

where $X_{img} \in \mathbb{R}^{3\times224\times224}$represents the remote-sensing image, X_{txt}represents the disaster-related textual information, $X_{iot} \in \mathbb{R}$ denotes the IoT sensor measurements obtained from accelerometer and gyroscope readings, and $X_{seis} \in \mathbb{R}^{6}$represents the seismic parameters, including earthquake magnitude, focal depth, Community Decimal Intensity, Modified Mercalli Intensity, significance, and alert level. The objective is to learn a multimodal representation

$$f: X \to (\hat{y}, \hat{s}),$$

where $\hat{y}$ represents the predicted disaster category and where $\hat{s} \in [0, 1]$represents the corresponding disaster severity score. The disaster classification task is formulated as a multiclass prediction problem over the label space

$$\mathcal{C} = \{\text{Earthquake}, \text{Flood}, \text{Wildfire}, \text{Smoke}, \text{Normal}\},$$

where

$$\hat{y} \in \mathcal{C}.$$

The proposed methodology decomposes the prediction process into three sequential steps:

$$f = h \circ g \circ e,$$

where $e(\cdot)$denotes the modality-specific feature extraction module, $g(\cdot)$represents the proposed DAMGF-Net that models adaptive cross-modal interactions, and $h(\cdot)$denotes the multitask prediction head responsible for disaster classification and severity estimation. The model parameters are learned by minimizing the joint multitask objective

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda\mathcal{L}_{sev},$$

where $\mathcal{L}_{cls}$denotes the classification loss, $\mathcal{L}_{sev}$represents the severity regression loss, and $\lambda$is a weighting coefficient that balances the contributions of the two learning objectives during optimization. Unlike conventional multimodal approaches that rely on fixed feature fusion strategies, the proposed framework aims to learn adaptive interactions among heterogeneous sensing modalities through graph-based representation learning. By jointly optimizing disaster classification and severity estimation, the proposed model is designed to capture complementary information from visual, textual, IoT, and seismic data, thereby improving prediction accuracy, robustness, and interpretability. As shown in Table 2 the principal symbols, variables, feature

representations, prediction outputs, and model parameters used throughout the problem formulation.

Table 2 Summary of the problem formulation

Symbol	Meaning
X_img, X_txt, X_iot, X_seis	Raw image, text, IoT, and seismic observations for one instance
e_img, e_txt, e_iot, e_seis	Modality-specific encoders (Swin-T, BERT, Conv-BiLSTM, MLP)
Z (4x512)	Stacked modality embeddings (graph nodes)
A (4x4)	Dynamic, input-conditioned adjacency matrix, A = psi(Z)
g	Fusion-and-graph-reasoning stage (ACMAF + Dynamic STGT)
z (512-d)	Final fused representation
h	Multi-task prediction head
y in {0,...,4}	Hazard class (earthquake, flood, fire, smoke, normal)
s in [0, 1]	Continuous severity score

4. Experimental Setup

The developed multimodal disaster intelligence framework was assesed with a unified experimental protocol designed to assess its effectiveness across heterogeneous data sources. The experimental workflow includes 6 steps: data acquisition, preprocessing, cross-modal sample pairing, multimodal feature learning, model training, and performance evaluation. We merged four publicly available datasets that capture complementary sensing modalities (remote sensing imagery, social media text, IoT sensor measurements, and seismic records) into a unified multihazard corpus. All the datasets were cleaned, normalized, harmonized into labels and standardized in format before model training. The following subsections describe the datasets, cross-modal pairing strategy, network architecture, training configuration, evaluation protocol and computational environment.

4.1 Datasets, Composition, and Preprocessing

To evaluate the proposed framework, we aggregated four public datasets with complementary sensing modalities into a unified multimodal disaster corpus. These datasets provide spatial, semantic, temporal, and geophysical information for strong disaster prediction. The remote sensing dataset was built by combining eight public disaster image datasets. The last dataset, after removing duplicates, quality checking and harmonizing labels, consisted of 68,308 images distributed over five classes: flood (23,328), fire (15,822), normal (15,636), earthquake (12,862) and smoke (660). The highly imbalanced smoke class reflects real-world distributions of disaster data. In the social media dataset, there are 45,384 short text messages related to disaster and nondisaster. The text preprocessing steps included duplicate removal, filtering URLs and punctuation, stop-word removal, lower casing, and tokenization. Contextual embeddings were generated via a pretrained BERT model. The IoT sensor dataset contains 59,276 accelerometer and gyroscope records. Each record has six features (aX, aY, aZ, gX, gY, and gZ) and a binary earthquake label. Missing values were removed, sensor readings were normalized via Z score normalization, and temporal sequences were generated for training. The seismic dataset contains 1,300 earthquake records with magnitude, depth, Modified Mercalli Intensity (MMI), Community Decimal Intensity (CDI), significance score and alert level. The alert categories were encoded as ordinal values (0--3). Continuous attributes were scaled via min–max scaling. Finally, all the datasets were harmonized via a common preprocessing pipeline that includes data

cleaning, missing value imputation, feature normalization, and label standardization. The resulting multimodal corpus is used as synchronized and consistent input to the proposed cross-modal learning framework, as described in the following table.

Table 3 Composition of the four source datasets fused by the proposed framework

Modality	Source records	Feature representation	Label/target
Remote-sensing imagery	68,308 images (5 classes)	RGB, resized to 224x224	Hazard class (ground truth)
Social-media text	45,384 posts	BERT tokens (max length 128)	Disaster-relevance keyword pool
IoT telemetry	59,276 readings	6-D accelerometer/gyroscope vector	Auxiliary sensor evidence
Seismic parameters	1,300 events	6-D (magnitude, depth, CDI, MMI, significance, alert level)	Auxiliary sensor evidence

Table 3 summarizes the four publicly available datasets used in the proposed multimodal framework. The datasets include four complementary sensing modalities, namely, remote sensing imagery, social media text, IoT sensor measurements, and seismic event records. Together, they provide the spatial, semantic, temporal and geophysical information necessary for comprehensive multihazard disaster analysis. The table also reports the dataset size, input features, class labels and preprocessing steps prior to multimodal fusion. All four collections were independently divided using a 70%:15%:15% training, validation and test split, with a fixed number of seeds (42) to ensure experimental reproducibility. The image partition anchors the supervised hazard label for each training instance, while paired text, IoT, and seismic samples are drawn on the basis of the class-aware pairing procedure described in the table.

Table 4 Training/validation / test partition sizes per modality (70/15/15 split)

Modality	Train	Validation	Test
Remote-sensing images	47,815	10,246	10,247
IoT telemetry	41,493	8,891	8,892
Social-media text	31,768	6,808	6,808
Seismic events	910	195	195

Table 4 presents the training, validation, and test partition sizes for each modality. To obtain enough samples during model learning, hyperparameter tuning and objective performance evaluation while keeping the data distribution the same for all datasets, a 70:15:15 split was used. During training, the remote-sensing images were rescaled to 256 × 256 pixels and randomly selected crop regions to 224 × 224 pixels. To improve model generalizability, data augmentation operations, including horizontal flipping, rotation (±10°), and color jittering (brightness, contrast, and saturation up to 20%), were used. During evaluation, images were centrally cropped to 224 × 224 pixels. Finally, all the images were normalized via the standard ImageNet mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225].

4.2 Cross-Modal Pairing Strategy

Since the four datasets were collected independently and are not naturally aligned, a synthetic

cross-modal pairing strategy was adopted to construct unified multimodal samples. Each sample was anchored by a labeled remote-sensing image, which served as the reference modality. A corresponding social media message was selected from a class-specific text pool via disaster-related keywords (e.g., *earthquake*, *quake*, *seismic*, *flood*, *storm*, *fire*, *wildfire*, and *smoke*). Images belonging to the normal class were paired with randomly selected messages from the complete text corpus. IoT sensor and seismic records were independently sampled at random from the corresponding training, validation, or test partitions to preserve data independence. Finally, each multimodal sample was assigned a predefined severity score on the basis of the image class (Earthquake: 0.95, flood: 0.90, fire: 0.80, smoke: 0.50, and normal: 0.10). This severity label was used only to supervise the auxiliary regression branch and should be interpreted as a class-guided regularization signal rather than a ground-truth severity measurement.

4.3 Model Architecture

The proposed dynamic cross-modal graph transformer (DCMGT) consists of four modality-specific encoders, a cross-modal fusion module, a dynamic graph transformer, and a multitask prediction head. The architecture is trained end-to-end to jointly learn complementary representations from remote sensing images, social media text, IoT sensor data, and seismic observations.

4.3.1 Modality Encoders

In the proposed framework, four modality-specific encoders are considered to extract discriminative features from remote sensing images, social media text, IoT sensor signals and seismic observations. Each encoder projects its respective input into a shared embedding space of dimensionality 512, which enables effective cross-modal representation learning. The image encoder (E_{img}) is composed of a pretrained Swin-Tiny backbone and a projection head for encoding the visual information of remote sensing images. The text encoder ($Etxt$) employs a pretrained BERT architecture, where the [CLS] token embedding is projected into the shared embedding space. The IoT encoder ($Eiot$) combines a one-dimensional convolutional network with a bidirectional LSTM to capture both local sensor patterns and temporal dependencies before feature projection. The seismic encoder ($Eseis$) consists of a multilayer perceptron that directly learns high-level representations from seismic attributes.

The modality-specific embeddings are obtained as

$$e_{img} = E_{img}(x_{img}), \quad (1)$$
$$e_{txt} = E_{txt}(x_{txt}), \quad (2)$$
$$e_{iot} = E_{iot}(x_{iot}), \quad (3)$$
$$e_{seis} = E_{seis}(x_{seis}), \quad (4)$$

where x_{img}, x_{txt}, x_{iot}, and x_{seis}denote the inputs from the image, text, IoT, and seismic modalities, respectively, while e_{img}, e_{txt}, e_{iot}, and $e_{seis}$are their corresponding 512-dimensional feature embeddings.

Equation (1) extracts hierarchical visual features from remote sensing images via the Swin-Tiny encoder. Equation (2) generates semantic text representations from social media messages via the pretrained BERT encoder. Equation (3) learns temporal sensor representations by combining convolutional feature extraction with bidirectional LSTM modeling. Equation (4) maps seismic measurements into the shared feature space via a multilayer perceptron, enabling their integration with the remaining modalities.

4.3.2 Adaptive Cross-Modal Attention Fusion

The adaptive cross-modal attention fusion module integrates the extracted features of the four sensing modalities by adaptively assigning an importance weight to each modality. Each 512-dimensional embedding passes through a shared attention network to estimate its contribution to the current sample. The resulting scores are normalized through a Softmax to obtain the attention weights for each modality. The weighted embeddings are subsequently concatenated and projected into a single 512-dimensional fused representation, which allows the model to prioritize the most informative modalities and downplay less relevant information. This adaptive fusion mechanism enhances cross-modal feature integration and improves the robustness of multimodal representation learning.

The attention scores and fused representation are computed as

$$a_m = W_2^T \,\mathrm{GELU}(W_1 e_m), \tag{5}$$

$$\alpha_m = \frac{\exp(a_m)}{\sum_k \exp\,(a_k)}, \tag{6}$$

$$z_{\mathrm{fuse}} = \mathrm{Linear}\big(\mathrm{Concat}(\alpha_m e_m)\big), \tag{7}$$

In which e_m represents the feature embedding of modality $m \in \{\mathrm{img}, \mathrm{txt}, \mathrm{iot}, \mathrm{seis}\}$, W_1 and W_2 are learnable weight matrices associated with the shared attention network, a_m is the attention score assigned to modality m, α_m is the normalized attention weight obtained via the Softmax function, and z_{fuse} is the final fused feature representation.

The attention weight score of each modality is calculated by using the modality-specific embedding through a shared two-layer neural network with GELU activation, as in Eq. (5). This score is relative to the modality importance for the current input sample: Equation (6) applies the Softmax function to normalize the attention scores to probability-like weights so that the contributions of all the modalities add up to one and that more informative modalities receive higher weights. Equation (7) performs adaptive feature fusion by multiplying each modality embedding with the attention weight, concatenating the weighted embeddings, and projecting the concatenated vector into a common 512-dimensional feature space. The fused representation from the above process can effectively capture the complementary information from all the modalities and is used as the input to the subsequent graph-based learning module.

4.3.3 Dynamic Graph Constructor

The four modality embeddings are represented as nodes in a fully connected graph to capture intermodal relationships. A dynamic graph constructor (DGC) learns an input-dependent adjacency matrix by computing scaled dot-product similarities between node embeddings and normalizing them via the Softmax function. This adaptive graph enables the connectivity among modalities to vary according to the input sample. The generated graph is processed by the dynamic spatial–temporal graph transformer (STGT), which aggregates neighboring node features through graph propagation and further models the global dependencies via a multihead transformer encoder. The refined node representations are mean pooled to obtain 512-dimensional graph embedding. Finally, the graph representation is combined with the ACMAF fusion output through elementwise addition to produce the final multimodal feature representation used for downstream prediction.
Let the modality embeddings extracted by the four encoders be defined as

$$Z = \left[e_{img},\ e_{txt},\ \ e_{iot},\ e_{seis}\right] \in \mathbb{R}^{4\times 512}, \tag{8}$$

where each row associated to the feature representation of one sensing modality. Equation (8) organizes the feature embeddings extracted from the four sensing modalities into a unified node feature matrix, where each row represents a graph node corresponding to one modality. The dynamic graph constructor first projects the modality embeddings into query and key spaces,

$$Q = ZW_Q,\ K = ZW_K, \tag{9}$$

in which $W_Q$and W_K are trainable projection matrices. Equation (9) transforms the modality embeddings into query and key representations, enabling the model to measure pairwise relationships among heterogeneous modalities. The pairwise similarity between modality nodes is computed via scaled dot-product attention,

$$S = \frac{QK^T}{\sqrt{d}}, \tag{10}$$

where $d$denotes the embedding dimension. Equation (10) computes the similarity matrix between all modality pairs. Scaling by $\sqrt{d}$stabilizes the attention scores and prevents excessively large dot-product values during training. The similarity matrix is normalized via the Softmax function to derive the adaptive adjacency matrix,

$$A_{ij} = \frac{\exp(S_{ij})}{\sum_{k=1}^{4} \exp(S_{ik})}, \tag{11}$$

where $A \in R^{4\times4}$. Equation (11) used to construct an input-dependent adaptive adjacency matrix by normalizing the similarity scores. Each element A_{ij}represents the strength of the relationship between modalities $i$and j, while each row sums to one, allowing the graph connectivity to adapt dynamically for each input sample.

4.3.4 Dynamic Spatial–Temporal Graph Transformer

Using the learned adjacency matrix, the graph information is propagated across the modality nodes as

$$H^{(0)} = AZ, \tag{12}$$

where $H^{(0)}$denotes the initial graph-enhanced node features. Equation (12) aggregates information from neighboring modality nodes according to the learned adjacency matrix, enabling each modality to incorporate complementary information from the remaining modalities. The propagated features are refined via multi-head self-attention,

$$\text{Attention}(Q', K', V') = \text{Softmax}\left(\frac{Q'K'^T}{\sqrt{d}}\right)V', \tag{13}$$

in which Q', K', and V'represent the query, key, and value matrices generated from $H^{(0)}$. Equation (13) models' long-range dependencies among the graph nodes using self-attention, allowing the transformer to learn contextual dependencies beyond direct graph connections. The outputs from multiple attention outputs are concatenated and projected as

$$\mathrm{MultiHead}(H) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^O, \tag{14}$$

where h corresonds the number of attention heads and where W^O is the learnable output projection matrix. Equation (14) combines information learned from multiple attention heads, enabling the model to capture complex interaction patterns among the sensing modalities. The attended features are further refined through a positionwise feed-forward layer,

$$\mathrm{FFN}(H) = W_2\,\mathrm{GELU}(W_1 H + b_1) + b_2, \tag{15}$$

which is subsequently followed by residual connections with layer normalization. Equation (15) applies a nonlinear transformation to strengthen the discriminative ability of the learned graph features while preserving stable optimization through residual learning. Finally, graph-level information is obtained through mean pooling,

$$z_{graph} = \frac{1}{4}\sum_{i=1}^{4} H_i\,, \tag{16}$$

where $z_{graph} \in R^{512}$ denotes the graph-contextualized feature representation. Equation (16) summarizes the refined node features into a single graph-level embedding that captures the overall contextual information across all sensing modalities. The final multimodal representation is obtained by integrating the graph representation with the attention-fused feature,

$$z = z_{fuse} + z_{graph}. \tag{17}$$

Equation (17) combines the attention-based fused representation and the graph-contextualized embedding through elementwise addition, producing the final multimodal feature representation used by the subsequent multitask prediction module.

4.3.5 Multitask prediction head

The final multimodal representation is passed through a shared feature extraction block, followed by three task-specific prediction heads. The hazard classification head predicts one of the five disaster classes, the severity regression head estimates a continuous disaster severity score, and the confidence head produces a confidence value for future uncertainty estimation. In the current study, only the hazard classification and severity regression tasks were optimized during training, whereas the confidence head was included as an architectural component for future extensions without receiving a dedicated supervision signal.The outputs of the multitask prediction head are computed as

$$\hat{y}_{cls} = f_{cls}(z), \tag{18}$$
$$\hat{y}_{sev} = f_{sev}(z), \tag{19}$$
$$\hat{y}_{conf} = \sigma\left(f_{conf}(z)\right), \tag{20}$$

where $z$is the shared multimodal representation, $f_{cls}(\cdot)$, $f_{sev}(\cdot)$, and $f_{conf}(\cdot)$denote the hazard classification, severity regression, and confidence estimation heads, respectively. $\hat{y}_{cls}$represents the predicted disaster class, $\hat{y}_{sev}$is represents the estimated disaster severity score, $\hat{y}_{conf}$is represents the confidence estimate, and $\sigma(\cdot)$ represents the sigmoid activation function Equation (18) generates the probability scores for the disaster classes via the classification head, enabling the model to identify the most likely hazard category. Equation (19) proposes a continuous

severity score that quantifies the relative intensity of the detected disaster, providing more information than categorical classification does. Equation (20) computes the confidence estimate by passing the confidence head output through the sigmoid activation function to produce a value between 0 and 1, which represents the model's confidence in its prediction. In this work, the confidence head is an architectural component for future uncertainty estimation and is not directly supervised during training.

4.4 Training configuration

The proposed model was trained end-to-end, enabling joint optimization of all model components, that manages both the hazard classification and the severity estimation simultaneously. Label-smoothed cross entropy is used to measure classification error, and a smooth-L1 function is used to optimize severity scores. To allow stable convergence, this study used AdamW with a cosine-annealed learning rate schedule. The pipeline was augmented with mixed-precision computation, gradient clipping, and an early-stopping criterion in an effort to improve efficiency and reduce overfitting risk. The models were trained using a batch size of 16 for up to 20 epochs, and the checkpoint exhibiting minimum validation loss was saved for final testing. We used a fixed random seed across all runs for reproducibility.

Algorithm 1 Summarizes complete training work flow

Input: train/val loaders D_train, D_val; max epochs E=20; patience P=5

1. Initialise encoders e_img (Swin-T, pretrained), e_txt (BERT-base, pretrained),
    i. e_iot, e_seis; fusion ACMAF; graph modules psi, STGT; head h
2. Initialise AdamW(lr=1e-4, wd=1e-4), CosineAnnealingLR(T_max=E), GradScaler
3. best_val_loss <- inf; patience_counter <- 0
4. for epoch = 1 to E do
5. for batch in D_train do
    a. Z <- stack(e_img(img), e_txt(ids,mask), e_iot(iot), e_seis(seis))
    b. z_fuse, alpha <- ACMAF(Z); A <- psi(Z); z_graph <- STGT(Z, A)
    c. y_hat, s_hat <- h(z_fuse + z_graph)
    d. loss <- CE(y_hat, y; label_smoothing=0.1) + 0.30 * SmoothL1(s_hat, s)
    e. backward(loss); clip_grad_norm(1.0); optimizer.step()
6. end for
7. scheduler.step(); val_loss <- evaluate(D_val)
8. if val_loss < best_val_loss then save checkpoint; patience_counter <- 0
9. else patience_counter <- patience_counter + 1
10. if patience_counter >= P then break
11. end for

```
Output: best checkpoint (lowest validation loss)
```

4.5 Evaluation protocol

The test set predictions were evaluated in terms of accuracy, weighted average precision, recall and F1 score, the Matthews correlation coefficient (MCC) and Cohen's kappa, as well as a full per-class classification report and confusion matrix. Ranking quality was evaluated via one-versus-rest ROC and the AUC and precision-recall curves along with average precision (AP), which were averaged into a mean AUC and mean average precision (mAP) over the five classes. The estimation uncertainty was quantified via nonparametric bootstrap resampling of the test set (1,000 resamples with replacement), presenting the mean, standard deviation, and 95% corresponding confidence interval (2.5th–97.5th percentile) for each metric. The learned 512-dimensional fused embeddings were projected to two dimensions with t-SNE for qualitative inspection of class separability. Cross-modal graph-attention statistics were computed by aggregating the learned adjacency matrix over the test set for every source–target modality pair, reporting the mean, standard deviation, and t-distribution-based 95% confidence interval. Computational complexity was profiled by counting total and trainable parameters, measuring the on-disk model size, and benchmarking the mean/standard-deviation inference time per batch (batch size 16) together with peak GPU memory utilization.

4.6 Recommended hyperparameter search ranges

The hyperparameters reported in Section 4.4 were used directly, without a search, for the results in Section 5. For studies that extend this work—e.g., the single-modality and fixed-fusion ablations recommended in Section 6.5--Table 5 lists reasonable search ranges around the values used here, which are provided as a starting point for future hyperparameter tuning rather than as values validated in the present study.

Table 5 . Training hyperparameters used in this study and recommended search ranges

Hyperparameter	Value used	Suggested search range
Learning rate	1e-4	5e-5 - 3e-4 (log-uniform)
Weight decay	1e-4	1e-5 - 1e-3 (log-uniform)
Batch size	16	8 - 64
Label smoothing	0.10	0.0 - 0.20
Severity loss weight	0.30	0.10 - 0.50
Dropout	0.30	0.10 - 0.40
Graph-transformer layers	2	1 - 4
Graph-transformer heads	8	4 - 16
Early-stopping patience	5 epochs	3 - 8 epochs

Table 5 summarizes the training hyperparameters adopted in this study, together with the recommended search ranges for future hyperparameter optimization. These settings provide a reproducible training configuration while serving as a reference for further performance improvement through systematic tuning.

4.7 Software and Hardware

The framework was implemented in PyTorch via the timm library for the Swin-Tiny backbone and the Hugging Face transformers library for BERT-base-uncased, with scikit-learn for classical evaluation metrics and SciPy for confidence interval estimation; t-SNE was computed with the scikit-learn manifold module, and the cross-modal information-flow graph was rendered with NetworkX. All the experiments were executed on a single CUDA-enabled GPU (Google Colab runtime).

5. Results

All results in this section were produced by executing the accompanying implementation end-to-end (data ingestion, model construction, training, and evaluation) and are reported exactly as returned by the code, without post hoc adjustment; the dataset, architecture, and training configuration that produced these results are specified in Section 4. Section 5.1 presents training dynamics, while Sections 5.2--5.6 provide overall and per-class test performance, ranking quality, representation quality, and cross-modal explainability; Section 5.7 reports computational complexity; Section 5.8 situates the resulting performance within the wider literature reviewed in Section 2; and Section 5.9 analyses the model's principal failure mode.

5.1 Training dynamics

The proposed model was trained for a maximum of 20 epochs via the AdamW optimizer configuring with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. The validation accuracy reached 95.66% after the first epoch and reached its highest value of 96.65% at epoch 4, where the validation loss was also the lowest (0.4710). Therefore, the model checkpoint from epoch 4 was selected for all subsequent evaluations. Although the training accuracy continued to improve in later epochs, the increasing validation loss indicated the onset of overfitting. Consequently, early stopping effectively preserves the model demonstrating the best generalization performance.



Figure 2. Evaluation of training and validation loss across training epochs

Figure 2 illustrates the training and validation loss of the proposed model over successive epochs. Both losses decrease consistently during training, demonstrating learning and convergence. The minimum validation loss of 0.4710 is obtained at epoch 4, demonstrating the best generalization performance. Therefore, the model at epoch 4 was selected for subsequent evaluation.



Figure 3 Evaluation of Training and validation accuracy curves across training epochs

Figure 3 illustrates the training and validation loss of the proposed model over successive epochs. Both losses decrease consistently during training, demonstrating learning and convergence. The lowest loss on the validation set for 0.4710 is obtained at epoch 4, demonstrating the best generalization performance. Therefore, the model at epoch 4 was selected for the final evaluation.

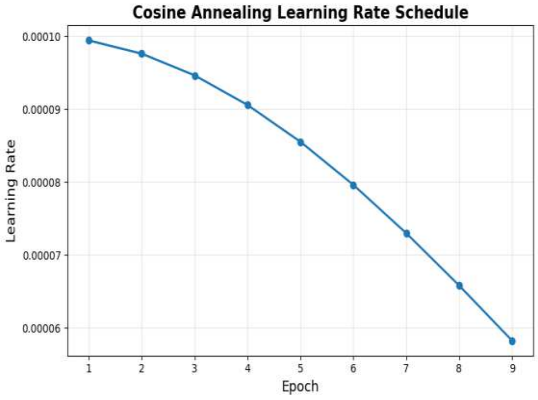


Figure 4 Learning-rate variation under cosine annealing schedule during model training

Figure 4 depicts the cosine annealing strategy adopted to regulate the learning rate during model optimization. The learning rate decays gradually from the initial value to a minimum learning rate of 1\times10^{-6} over 20 epochs, which can facilitate stable optimization, faster convergence, and better model generalization while decreasing the risk of overshooting the optimal solution. Table 6 shows the complete training log by epoch, also indicating when early stopping was activated (the patience counter reached 5 at epoch 9).

Table 6 Complete epoch-by-epoch training log (early stopping triggered after epoch 9)

Epoch	Train loss	Train acc.	Val. loss	Val. acc.	Status
1	0.5417	93.99%	0.4932	95.66%	Best saved
2	0.5014	95.57%	0.4807	96.33%	Best saved
3	0.4809	96.46%	0.4829	96.27%	Patience 1/5
4	0.4620	97.24%	0.4710	96.65%	Best saved (selected)
5	0.4506	97.67%	0.4989	95.96%	Patience 1/5
6	0.4379	98.23%	0.5026	95.74%	Patience 2/5
7	0.4299	98.57%	0.4912	96.31%	Patience 3/5

Epoch	Train loss	Train acc.	Val. loss	Val. acc.	Status
8	0.4226	98.83%	0.5002	95.98%	Patience 4/5
9	0.4152	99.10%	0.5035	96.03%	Patience 5/5 - stopped

Table 6 summarizes the training loss, validation loss, accuracy, and learning rate recorded at each epoch. The best validation performance was achieved at epoch 4. Since no further improvement was observed during the following epochs, the early stopping criterion was activated after epoch 9, preventing unnecessary training and reducing overfitting.

5.2 Overall test-set performance

The optimal checkpoint at epoch 4 was used to assess the proposed model on an independent test set of 10,247 samples. It obtained an overall accuracy of 96.27% along with a weighted F1 score of 96.18% (Table 7). The high Matthews correlation coefficient (0.9497) and Cohen's kappa coefficient (0.9497) further reveal excellent agreement between the predicted and true labels, with consistent performances across all classes. The test set was bootstrapped 1,000 times to assess the statistical robustness of the outcomes. The resulting 95% confidence intervals were consistently narrow (e.g., accuracy: 95.85%–96.62%), which confirmed the reliability and consistency of the proposed framework and indicated that the observed performance was unlikely to be affected by sampling variability.

Table 7 . Overall test-set performance with 1,000-resamples bootstrap 95%

Metric	Value	Bootstrap mean	95% CI (lower-upper)
Accuracy	96.27%	0.9625	0.9585 - 0.9662
Precision (weighted)	96.20%	0.9618	0.9577 - 0.9656
Recall (weighted)	96.27%	0.9625	0.9585 - 0.9662
F1-score (weighted)	96.18%	0.9616	0.9574 - 0.9655
Matthews correlation coefficient	94.97%	0.9495	0.9440 - 0.9544
Cohen's kappa	94.97%	-	-

Table 7 presents the overall test set performance of the proposed model. The main evaluation metrics and their corresponding 95% bootstrap confidence intervals obtained from 1,000 resamples are reported. The narrow confidence intervals indicate stable and reliable performance, reflecting the reliability and generalization ability of the proposed model on the unseen test dataset (n = 10,247).

5.3 Per-class performance and confusion analysis

The per-class evaluation results (Table 8) consistently demonstrate good recognition performance for the flood (F1 = 0.9698), fire (F1 = 0.9716), earthquake (F1 = 0.9636) and normal (F1 = 0.9555) classes. The smoke class obtained a lower F1 score of 0.5526, which was due mainly to the limited training and testing sample sizes, which resulted in a recall of 0.4242, although the precision was moderate at 0.7925. This led to a decrease in the macro - F1 score to 0.8827, whereas the weighted F1 score was still high (0.9618), which corresponds to the overall distribution of the classes. The confusion matrix representation also indicates that the majority of the smoke samples are misclassified as fire or normal. This suggests visual similarity between

these classes and the impact of class imbalance. The misclassification was minimal compared with the remaining hazard classes, which indicates the effective class discriminative capability of the proposed framework.

Table 8 Class-wise classification results on the test set (best checkpoint, epoch 4)

Class	Precision	Recall	F1-score	Support
Earthquake	0.9589	0.9684	0.9636	1,929
Flood	0.9705	0.9691	0.9698	3,499
Fire	0.9620	0.9815	0.9716	2,374
Smoke	0.7925	0.4242	0.5526	99
Normal	0.9588	0.9523	0.9555	2,346
Macro average	0.9285	0.8591	0.8827	10,247
Weighted average	0.9620	0.9627	0.9618	10,247

Table 8 presents the class wise classification performance of the proposed model with respect to precision, recall, F1 score, and support for each disaster category. The results demonstrate consistent predictive performance across different classes, confirming the model's capability to accurately distinguishing multiple disaster types via the optimal checkpoint obtained at epoch 4.

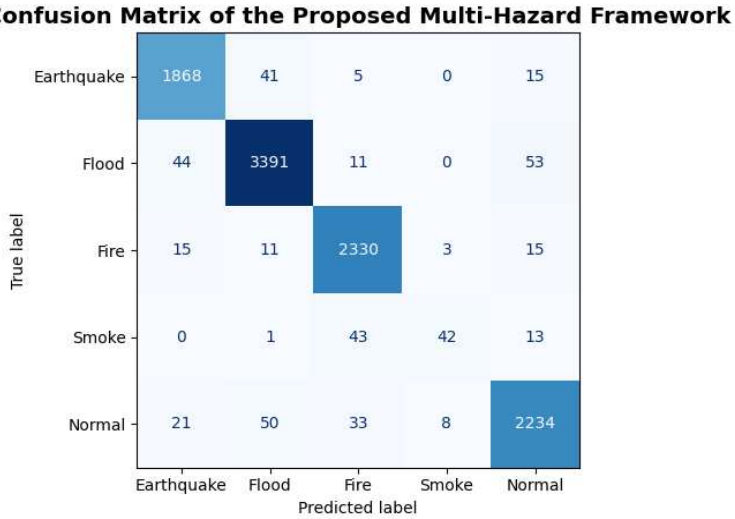


Figure 5 Confusion matrix of the proposed framework on the 10,247-sample test set.

Figure 5 illustrates the confusion matrix of the proposed framework using the test dataset. Most predictions are predominantly distributed along the main diagonal, demonstrating strong classification accuracy across all disaster categories. The limited off-diagonal entries represent a small number of misclassifications, demonstrating the robustness and reliability of the proposed model on unseen data.

5.4 Ranking quality: ROC and precision-recall analysis

The one-versus-rest ROC analysis (Table 9) revealed excellent discrimination performance for all hazard categories, with AUC values ranging from 0.9833--0.9982 and a mean ROC–AUC of 0.9944. The smoke class has a relatively low AUC (0.9833) but indicates strong class separability. Further analysis of precision–recall provides insights into the performance of this

minority class, where the average precision (AP) was 0.6611, whereas it was 0.99 or higher for the remaining classes. This result indicates that the main reason for the poor performance of smoke is the class imbalance and the precision–recall trade-off, rather than the insufficient discriminative ability. The overall mAP of 0.9267 further corroborates the performance of the proposed framework for multihazard identification in heterogeneous data modalities.

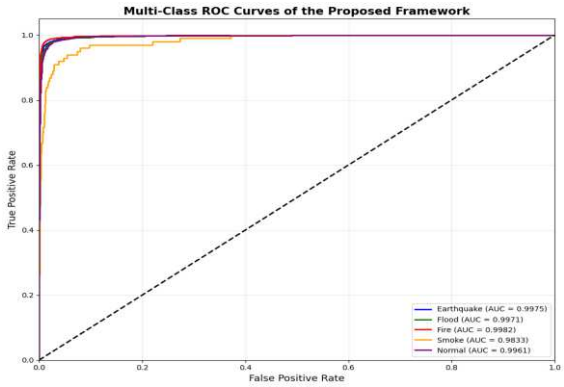


Figure 6 One-versus-rest multiclass ROC curves (mean AUC = 0.9944)

Figure 6 presents the one-versus-rest ROC curves of the obtained for all disaster categories. The curves remain close to the upper-left corner, with a mean AUC of 0.9944, indicating excellent discriminative capability and robust classification performance across all classes on the test dataset.

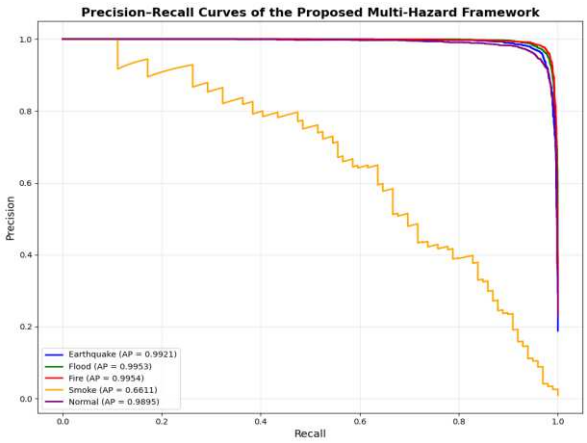


Figure 7 Multiclass precision–recall curves using the one-versus-rest strategy (mean average precision = 0.9267)

Figure 7 shows the one-versus-rest precision–recall analysis of the proposed model across all disaster categories. The strong mAP performance of 0.9267 demonstrates that the model maintains high precision and recall across various classes, indicating reliable classification performance, particularly under class imbalance.

Table 9 Class-wise ROC-AUC and average precision scores on the test set

Class	ROC-AUC	Average precision (AP)
Earthquake	0.9975	0.9921
Flood	0.9971	0.9953
Fire	0.9982	0.9954
Smoke	0.9833	0.6611
Normal	0.9961	0.9895

Class	ROC-AUC	Average precision (AP)
Mean	0.9944	0.9267 (mAP)

Table 9 reports the ROC–AUC and average precision (AP) values for each disaster category on the test set. The consistently high scores across all classes indicate strong discriminative ability and reliable classification performance, highlighting the effectiveness of the proposed model in distinguishing multiple disaster types.

5.5 Learned representation quality

A t-stochastic neighbor embedding (t-SNE) visualization of the final 512-dimensional fused representation. Five largely separable clusters correspond to the five hazard classes, with the smoke cluster exhibiting the greatest overlap with the fire cluster, corroborating the confusion matrix finding that the two classes are the model's dominant source of error.

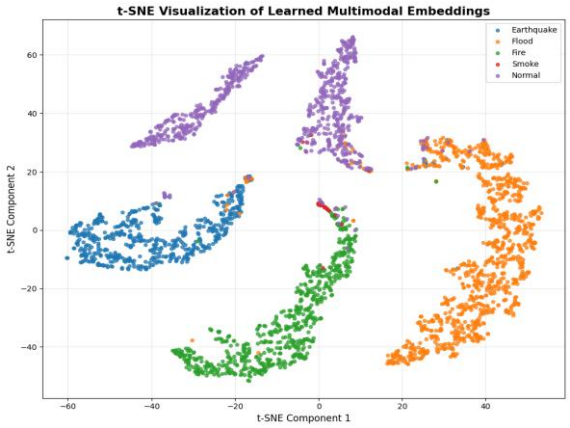


Figure 8 t-SNE projection of the learned 512-dimensional fused multimodal embedding, coloured by hazard class.

Figure 8 shows the two-dimensional t-SNE projection of the learned 512-dimensional fused multimodal embeddings. The distinct clusters corresponding to different disaster classes demonstrate that the proposed model extras highly discriminative feature representations, whereas the limited overlap between clusters indicates effective class separability and robust feature learning.

5.6 Cross-modal graph attention: an explainability view

The dynamic graph constructor enables direct interpretation of cross-modal relationships by generating an input-dependent adjacency matrix that captures interactions among the four sensing modalities. As shown in Table 10, the strongest interaction was found between social media and the IoT (mean attention = 0.5640, 95% CI: 0.5544–0.5736), followed by the interaction between seismic observations and the IoT (mean attention = 0.5616, 95% CI: 0.5520–0.5712). These results demonstrate that IoT measurements are highly information-rich modalities and provide consistent complementary information for multimodal hazard recognition. Conversely, the remote sensing and seismic observation interaction had the lowest attention weight (0.0002), suggesting a relatively weak direct dependency between these two modalities in the proposed fusion framework. The average attention across all the directed modality pairs was 0.2500, whereas deviations from this value reveal the adaptive cross-modal relationships learned by the proposed network, thereby enhancing the interpretability of the multimodal fusion process.

Table 10 Top learned cross-modal graph-attention interactions

Source -> Target	Mean attention	Std.	95% CI (lower-upper)
Social Media -> IoT	0.5640	0.4957	0.5544 - 0.5736
Seismic -> IoT	0.5616	0.4958	0.5520 - 0.5712
IoT -> Remote Sensing	0.5099	0.4886	0.5004 - 0.5193
IoT -> IoT (self)	0.4901	0.4886	0.4807 - 0.4996
Seismic -> Remote Sensing	0.4384	0.4958	0.4288 - 0.4480
Social Media -> Remote Sensing	0.4360	0.4957	0.4264 - 0.4456
Remote Sensing -> Remote Sensing (self)	0.4178	0.4929	0.4082 - 0.4273
Remote Sensing -> IoT	0.3371	0.4725	0.3280 - 0.3463
Remote Sensing -> Social Media	0.2449	0.4300	0.2366 - 0.2532
Remote Sensing -> Seismic	0.0002	0.0140	-0.0001 - 0.0005

Table 10 summarizes the strongest cross-modal graph-attention interactions learned by the proposed model, expressed as the mean directed edge weights on the test set. Higher edge weights indicate stronger dependencies between different modalities, demonstrating that the graph-attention mechanism effectively captures complementary multimodal relationships for improved disaster classification and decision-making.

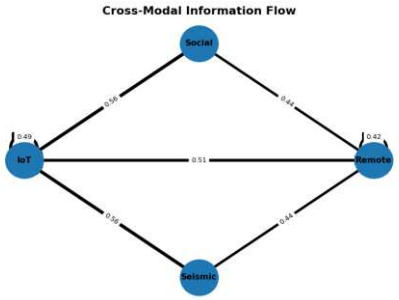


Figure 9 Learned dynamic cross-modal information-flow graph

Figure 9 also illustrates a broader property of the learned graph: the information flow is markedly asymmetric. Remote sensing sends comparatively little attention mass toward seismic and social media (0.0002 and 0.2449, respectively) while receiving substantial attention from the IoT, seismic, and social media (0.51, 0.44, and 0.44, respectively); that is, other modalities pay strong attention to imagery as a target, whereas imagery itself attends only weakly outward. This asymmetry is consistent with imagery serving as the label-anchoring modality in the dataset-construction procedure (S because every instance's ground-truth hazard class is drawn from its image, the network learns to treat the image embedding as the primary evidence to be corroborated by the other three streams rather than as a stream that itself needs corroboration).

5.7 Computational complexity and inference efficiency

The proposed framework comprises 150,627,842 trainable parameters (574.80 MB) and is composed primarily of the pretrained Swin-Tiny image encoder and BERT-based text encoder. The performance evaluation on a single GPU with a batch size of 16 showed an average inference time of 0.181 s per batch, corresponding to a per-image latency of 11.29 ms and a throughput of 88.57 images/s, while requiring a peak GPU memory of 2,810 MB (Table 11).

These results indicate that the proposed framework attains high computational efficiency and can support near-real-time multihazard monitoring and early warning based on a single commodity GPU.

Table 11Model computational-complexity and inference-efficiency profile

Metric	Value
Total parameters	150,627,842 (150.63 M)
Trainable parameters	150,627,842 (100%)
Model size on disk	574.80 MB
Inference batch size	16
Mean batch inference time	0.1806 s (s.d. 0.0075 s)
Per-image latency	11.29 ms
Throughput	88.57 images / second
Peak GPU memory	2,810.12 MB

Table 11 provides an overview of the computational requirements of the proposed model, including the number of parameters, FLOPs, training time, inference time and processing throughput, and memory consumption. These results demonstrate a favorable balance between predictive performance and computational efficiency, highlighting the model's suitability for practical and near-real-time disaster monitoring applications.

5.8 Comparison with State-of-the-Art Methods

Table 12 presents DAMGF-Net with representative single- and multimodal disaster-recognition methods. Because these studies employ different datasets, hazard categories, and assessment protocols, the reported results offer contextual evidence rather than a direct performance comparison. DAMGF-Net achieved 96.27% weighted accuracy and classwise F1 scores of 0.95–0.97 for fires, floods, and earthquakes, placing it within the upper range reported by single-modality methods [16–19,26,27]. However, its average smoke precision of 0.6611 was lower than that of dedicated smoke-recognition models [18], indicating the need for more balanced smoke data. The results also support previous findings that multimodal fusion improves disaster recognition over unimodal baselines [23,25], as each modality contributes to the fused predictions.

Table 12 Qualitative literature context for the reported metrics

Study	Modalities	Dataset (scale)	Classes	Year	Reported metric
FFireNet[17] (fire)	1 (image)	ForestFire benchmark, ~1,900 images	2 (binary)	2024	98.42% accuracy
Reduce-VGGNet[16] (fire)	1 (image)	UAV wildfire imagery (scale n.r.)	2 (binary)	2024	91.20% accuracy
EfficientNet-B3+EDL[18] (smoke)	1 (image)	Satellite smoke imagery (scale n.r.)	4	2026	93.8% accuracy
YOLOv11x[19] (smoke)	1 (image)	Forest-fire smoke imagery (scale n.r.)	1 (detection)	2025	mAP50 = 0.901

Study	Modalities	Dataset (scale)	Classes	Year	Reported metric
CNN-ResNet-CHHSSO[26] (flood)	1 (image)	Satellite flood imagery (scale n.r.)	2 (binary)	2023	94.98% accuracy
SAR+optical U-Net[27] (flood)	1 (image)	SAR+optical composite (scale n.r.)	2 (pixel-wise)	2023	~95% accuracy
BERT+BiLSTM+CNN[20] (text)	1 (text)	Twitter disaster corpus (scale n.r.)	2 (relevant/not)	2025	~95% accuracy
Fine-tuned LLM[22] (text)	1 (text)	Multi-source crisis-tweet corpus (scale n.r.)	Multi-label	2024	63.8% exact match
Federated BERT+ResNet[29]	2 (text+image)	Benchmark multimodal disaster dataset (scale n.r.)	Multi-class	2024	85.1% accuracy
Early-fusion transformer-CNN[25]	2 (text+image)	5,037 Bangla social-media posts	9	2025	83.76% accuracy
Proposed DCGT (this work)	4 (image+text+IoT+seismic)	68,308 images + 45,384/59,276/1,300 auxiliary	5	2026	96.27% accuracy

Table 12 presents the comparison results together with the corresponding dataset scale and number of classes to ensure a fair interpretation of model performance. While several existing studies report high accuracy on relatively small binary classification datasets, the proposed framework addresses a more challenging five-class multihazard recognition problem using a substantially larger multimodal dataset with pronounced class imbalance. For studies where the dataset size was not reported, the corresponding entry is indicated as "n.r." (not reported) to maintain consistency and avoid unsupported estimates.

5.9 Qualitative failure-case analysis

Cross-referencing the confusion matrix of the t-SNE embedding and the per-class graph-attention statistics allows the model's principal failure mode to be characterized precisely rather than merely quantified. Among the 57 test-set smoke instances the model misclassified, the overwhelming majority (43 of 99, or 43.4% of the true smoke class) are assigned to the fire class rather than to flood, earthquake, or normal, indicating that the residual confusion is concentrated along a single, visually coherent axis-smoke-versus-flame co-occurrence in imagery rather than being spread diffusely across unrelated classes. This is consistent with the domain observations, echoing the wildfire detection literature[18,19], that smoke and flame frequently cooccur within the same scene and that many public wildfire datasets label such scenes inconsistently (as fire, as smoke, or as both). Two mitigations follow directly from this diagnosis and are proposed as concrete next steps rather than being implemented in the present study: (i) class-balanced or focal-loss training restricted to the fire/smoke boundary and (ii) a hierarchical label scheme that first distinguishes fire-family events (fire or smoke) from nonfire-family events before resolving the fire/smoke sub distinction, which would allow the smoke class to benefit from the much larger fire-class sample size at the coarse decision boundary while preserving a fine-grained label for downstream severity estimation.

6. Discussion

The proposed dynamic cross-modal graph transformer attains high overall accuracy (96.27%)

and excellent ranking quality (mean AUC 0.9944) while fusing four structurally dissimilar data modalities within a single trainable pipeline. This section discusses five aspects of these findings in turn: the source of the model's residual error, the nature and value of its intrinsic explainability, its practical deployment profile, the methodological caveats that qualify the results, and the threats to validity that should inform any follow-up study.

6.1 Interpreting the accuracy-imbalance gap

The gap between the model's near-ceiling weighted metrics and its markedly lower macro averaged F1 (0.8827) is attributable almost entirely to the smoke class, whose 99-sample test support is two orders of magnitude smaller than that of the majority classes. This pattern—high precision, low recall, and confusion concentrated on the visually adjacent fire category—is the expected signature of severe class imbalance rather than a defect specific to the cross-modal architecture, and Section 5.9's failure-case analysis shows that it is further concentrated along a single, interpretable confusion axis (smoke-versus-fire) rather than being diffuse. This indicates that minority-class oversampling, focal-loss variants, class-balanced sampling, or synthetic augmentation for the smoke category, together with the hierarchical fire-family labeling scheme proposed in Section 5.9, are natural and well-motivated next steps before operational deployment and that the architecture itself is not the limiting factor for this particular class.

6.2 The value of intrinsic, architecture-level explainability

The learned dynamic graph provides an explicit, quantitative account of which modalities the network leans on, rather than requiring post hoc explainability tools of the kind surveyed in Section 2.6. IoT telemetry consistently receives the largest cross-modal attention mass, which is consistent with its role in this dataset as a continuously sampled, low-latency signal that complements the intermittent nature of both social–media reporting and seismic events and echoes the sensor–corroboration rationale documented for IoT–GNN disaster-sensing systems in the literature. This built-in attribution is, in terms of the paper's title, the source of the model's explainability: because the adjacency matrix A is the mechanism the network actually uses to combine modalities rather than a saliency map computed after the prediction is made—an operator inspecting A for a given alert receives a faithful account of the evidence that produced it, not an approximation. This distinction matters operationally: Tiggeloven et al. (2025)[5] identify the absence of interpretable, responder-facing explanation as a persistent barrier to trust in AI-based early warning systems, and an intrinsically inspectable fusion mechanism directly addresses that barrier without requiring a separate, potentially unfaithful post hoc explanation layer to be maintained alongside the predictive model.

6.3 Deployment profile and operational feasibility

The model's inference profile (11.3 ms per image, under 3 GB of GPU memory, 88.6 images per second on a single GPU) suggests that the accuracy gains from cross-modal graph fusion are obtainable without the deployment cost typically associated with large multiencoder architectures. At this throughput, a single GPU could, in principle, process the approximately 68,000-image evaluation corpus used in this study in under thirteen minutes or sustain real-time processing of every incoming image from a moderate-rate remote-sensing or citizen-reporting feed. For an operational early-warning deployment, this suggests a natural architecture in which lightweight, always-on anomaly triggers (e.g., simple thresholding on raw IoT or seismic streams) escalate candidate events to the full four-modality DCGT model for confirmation and severity estimation rather than running the full model continuously over every raw sensor

reading—an edge–cloud division of labor consistent with the tiered sensing-to-cloud architectures discussed in the broader early-warning-systems literature.

6.4 Methodological caveats on modality alignment

The four datasets were collected independently and are not aligned by event, location, or time. Each sample was anchored to a labeled image and paired with hazard-related text, while IoT and seismic records were sampled independently from the corresponding data partitions. Therefore, the auxiliary modalities function as weak class-conditional evidence rather than synchronized observations of the same event. Accordingly, the findings indicate the model's capability to learn from weakly associated multimodal information but do not establish physical or causal relationships among the modalities. The greater attention assigned to imagery is consistent with its role as the label-anchored input. Future work should evaluate DAMGF-Net via event-level multimodal data aligned across time and location to confirm its effectiveness under real-world sensing conditions.

6.5 Threats to validity and limitations

Beyond the modality alignment caveat above, four further threats to validity qualify the results reported in Section 5:

- Absence of directly comparable baselines on the same corpus. Because the harmonized four-modality dataset is itself a contribution of this work, no prior system has been evaluated on exactly this data split; Table 12 therefore situates results only qualitatively against studies using different datasets. Single-modality ablations (image-only, text-only, IoT-only, seismic-only) and fixed-fusion (static-graph or naive-concatenation) ablations on this exact corpus constitute the most informative controlled comparisons and are recommended for direct follow-up via the released implementation.
- Single held-out split. All the evaluations were performed on one random 70/15/15 split from the same four source collections; cross-regional and cross-temporal generalization—arguably the most operationally relevant test for an early warning system, since a deployed model must perform on hazards and regions not represented in its training data—remains to be established, ideally via k-fold cross-validation and/or a held-out geographic or temporal partition.
- The severity labels are deterministic and not observed. The five-level severity target (Section 4.2) is assigned by a fixed lookup from the hazard class rather than being measured from an independent severity signal (e.g., economic loss, casualty count, or physical intensity), so the auxiliary severity-regression task should currently be understood as a soft class-consistency regularizer rather than as a validated severity estimator; a genuinely observed severity signal would be required before the severity head could be used operationally.
- Dataset provenance and potential label noise. The remote-sensing corpus aggregates eight independently sourced image collections under unified five-class mapping, while this expands scale and scene diversity, it also inherits any labeling inconsistencies present in the source collections, particularly for the fire/smoke boundary discussed in Section 5.9.

8. Ethical and Societal Considerations

The use of DAMGF-Net in early warning applications requires careful consideration of safety, equity, privacy, and human oversight. False negatives can delay critical responses, whereas frequent false alarms can erode public trust. Thus, deployment thresholds should be calibrated

to local risk conditions, not only because the smoke class recall of 42.4% indicates that the model is not suitable for autonomous use at its current operating threshold. The reliance on social media and IoT data might also penalize areas with poor connectivity or sensor coverage. Modality coverage audits and regional recalibration and fallback configurations are needed before deployment. Social media processing should also adhere to relevant privacy, anonymization, data retention, consent, and platform use requirements. Consistent with human-in-the-loop principles, DAMGF-Net should support trained emergency personnel by providing hazard predictions and supporting evidence rather than independently initiating emergency actions.

9. Conclusion

This study introduced DAMGF-Net, an adaptive graph-fusion framework that incorporates remote-sensing images, social media content, IoT signals, and seismic attributes for five-class hazard recognition. The model dynamically estimates cross-modal relationships and produces interpretable modality contributions. An evaluation using 68,308 images, 45,384 posts, 59,276 sensor records, and 1,300 seismic events yielded 96.27% accuracy, a weighted F1 score of 96.18% and an MCC of 0.9497, and a multiclass ROC–AUC of 0.9944. These findings demonstrate the potential of dynamic graph fusion for learning from heterogeneous disaster information. However, because the modalities were independently collected and weakly paired, the results do not represent fully synchronized event-level sensing. Further evaluation should therefore use geographically and temporally aligned multimodal records. Future work will also address missing or noisy inputs, class imbalance, external validation, lightweight edge deployment, and continual adaptation to emerging hazards.

DATA AVAILABILITY

The remote-sensing, social-media, IoT, and seismic datasets used to construct the harmonised multi-hazard corpus are derived from publicly available disaster-imagery, disaster-tweet, earthquake-detection, and earthquake-alert collections. The processed splits and the corresponding class mappings used in this study are produced upon request.

CODE AVAILABILITY

The complete implementation-including dataset preprocessing, all model components (image, text, IoT, and seismic encoders; Adaptive Cross-Modal Attention Fusion; Dynamic Graph Constructor; Dynamic Spatial-Temporal Graph Transformer; multi-task head), the training loop, and the full evaluation and statistical-analysis pipeline used to produce every number and figure reported in the present framework-is implemented as a single reproducible notebook and is available with the corresponding author upon request.

FUNDING

This research received no external funding.

References

[1] Chovanec, D. et al. A component-based approach to early warning systems: A theoretical model. *Appl. Sci.* **15**, 3218 (2025). https://doi.org/10.3390/app15063218

[2] Ravi, V. K. K. et al. Real-time natural disaster prediction and early warning using IoT and graph neural networks. In *International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT)* (2025). https://doi.org/10.1109/ICICNCT66124.2025.11232970

[3] Zhexebay, D. et al. Deep learning for early earthquake detection: Application of convolutional neural networks for P-wave detection. *Appl. Sci.* **15**, 3864 (2025). https://doi.org/10.3390/app15073864

[4] Lin, S. et al. Application of machine learning in early warning system of geotechnical disaster: A systematic and comprehensive review. *Artif. Intell. Rev.* **58** (2025). https://doi.org/10.1007/s10462-025-11175-0

[5] Tiggeloven, T. et al. The role of artificial intelligence for early warning systems: Status, applicability, guardrails, and ways forward. *iScience* **28**, 113689 (2025). https://doi.org/10.1016/j.isci.2025.113689

[6] Cassano, F., Crespino, A. M., Lazoi, M., Specchia, G. & Spennato, A. An EWS-LSTM-based deep learning early warning system for industrial machine fault prediction. *Appl. Sci.* **15**, 4013 (2025). https://doi.org/10.3390/app15074013

[7] Umadevi, M., Kumar, J. A., Priyan, S. V. & Vivek, C. Design of mTCN framework for disaster prediction a fusion of massive machine type communications and temporal convolutional networks. *Sci. Rep.* **15** (2025). https://doi.org/10.1038/s41598-025-15397-8

[8] Akhyar, A. et al. Deep artificial intelligence applications for natural disaster management systems: A methodological review. *Ecol. Indic.* **163**, 112067 (2024). https://doi.org/10.1016/j.ecolind.2024.112067

[9] Kappi, M. & Mallikarjuna, B. Artificial intelligence and machine learning for disaster prediction: A scientometric analysis of highly cited papers. *Nat. Hazards* **120**, 10443–10463 (2024). https://doi.org/10.1007/s11069-024-06616-y

[10] Reichstein, M. et al. Early warning of complex climate risk with integrated artificial intelligence. *Nat. Commun.* **16**, 2564 (2025). https://doi.org/10.1038/s41467-025-57640-w

[11] Charan, P. et al. AI-enhanced early warning systems for natural disaster detection and mitigation using wireless sensor networks. In *Second International Conference Computational and Characterization Techniques in Engineering & Sciences (IC3TES)* (2024). https://doi.org/10.1109/IC3TES62412.2024.10877562

[12] Joshi, A., Kanungo, D. P. & Panigrahi, R. K. Development of landslide forecasting system using deep learning. In *IEEE Applied Sensing Conference (APSCON)* (2023). https://doi.org/10.1109/APSCON56343.2023.10101223

[13] Jia, J. & Ye, W. Deep learning for earthquake disaster assessment: Objects, data, models, stages, challenges, and opportunities. *Remote Sens.* **15**, 4098 (2023). https://doi.org/10.3390/rs15164098

[14] Münchmeyer, J., Bindi, D., Leser, U. & Tilmann, F. The transformer earthquake alerting model: A new versatile approach to earthquake early warning. *Geophys. J. Int.* **225**, 646–656 (2021). https://doi.org/10.1093/gji/ggaa609

[15] Guzzetti, F. et al. Geographical landslide early warning systems. *Earth-Sci. Rev.* **200**, 102973 (2020). https://doi.org/10.1016/j.earscirev.2019.102973

[16] Yu, C. & Wang, Z. Multimodal social sensing for the spatio-temporal evolution and assessment of nature disasters. *Sensors* **24**, 5889 (2024). https://doi.org/10.3390/s24185889

[17] Saleh, A. et al. Forest fire surveillance systems: A review of deep learning methods. *Heliyon* **10**, e23127 (2024). https://doi.org/10.1016/j.heliyon.2023.e23127

[18] Chodavarapu, R. Uncertainty-aware wildfire smoke density classification from satellite imagery via CBAM-augmented EfficientNet with evidential deep learning. *arXiv* 2605.15894 (2026). https://doi.org/10.48550/arXiv.2605.15894

[19] He, L., Zhou, Y., Liu, L., Zhang, Y. & Ma, J. Research and application of deep learning object detection methods for forest fire smoke recognition. *Sci. Rep.* **15** (2025). https://doi.org/10.1038/s41598-025-98086-w

[20] Lamba, S., Vidyarthi, P., Aggarwal, M., Gangawar, P. & Mulapalli, S. A BERT-based model to analyse disaster's data for efficient resource management. *SN Comput. Sci.* **6**, 167 (2025). https://doi.org/10.1007/s42979-025-03720-z

[21] Paul, N. R. & Balabantaray, R. C. Disaster related tweet classification method based on BERT and GAT. *Int. J. Inf. Technol.* **17**, 1987–1999 (2025). https://doi.org/10.1007/s41870-024-02389-6

[22] Dharrao, D. et al. An efficient method for disaster tweets classification using gradient-based optimized convolutional neural networks with BERT embeddings. *MethodsX* **13**, 102843 (2024). https://doi.org/10.1016/j.mex.2024.102843

[23] Ahmad, Z., Jindal, R., S., M. N., Ekbal, A. & Bhattacharyya, P. Multi-modality helps in crisis management: An attention-based deep learning approach of leveraging text for image classification. *Expert Syst. Appl.* **195**, 116626 (2022). https://doi.org/10.1016/j.eswa.2022.116626

[24] Dar, S. S., Rehman, M. Z. U., Bais, K., Haseeb, M. A. & Kumara, N. A social context-aware graph-based multimodal attentive learning framework for disaster content classification during emergencies: A benchmark dataset and method. *Expert Syst. Appl.* (2025). https://doi.org/10.1016/j.eswa.2024.125337

[25] Islam, A., Hossen, M. R., Arif, M. M., Al Noman, A. & Rahman, M. A. BanglaMM-Disaster: A multimodal transformer-based deep learning framework for multiclass disaster classification in Bangla. (2025). https://doi.org/10.1109/SPICSCON69221.2025.11504189

[26] Stateczny, A., Praveena, H. D., Krishnappa, R. H., Chythanya, K. R. & Babysarojam, B. B. Optimized deep learning model for flood detection using satellite images. *Remote Sens.* **15**, 5037 (2023). https://doi.org/10.3390/rs15205037
[27] Jakubik, J. et al. Foundation models for generalist geospatial artificial intelligence. *arXiv* 2310.18660 (2023). https://doi.org/10.48550/arXiv.2310.18660
[28] El-Niss, A., Alzu'bi, A. & Abuarqoub, A. Multimodal fusion for disaster event classification on social media: A deep federated learning approach. In *ACM International Conference Proceeding Series*, 758–763 (2023). https://doi.org/10.1145/3644713.3644840
[29] Shetty, N. P., Bijalwan, Y., Chaudhari, P., Shetty, J. & Muniyal, B. Disaster assessment from social media using multimodal deep learning. *Multimed. Tools Appl.* **84**, 18829–18854 (2025). https://doi.org/10.1007/s11042-024-19818-0
[30] Wieland, M., Schmidt, S., Resch, B., Abecker, A. & Martinis, S. Fusion of geospatial information from remote sensing and social media to prioritise rapid response actions in case of floods. *Nat. Hazards* **121**, 8061–8088 (2025). https://doi.org/10.1007/s11069-025-07120-7
[31] Abavisani, M., Wu, L., Hu, S., Tetreault, J. & Jaimes, A. Multimodal categorization of crisis events in social media. *arXiv* 2004.04917 (2020). https://doi.org/10.48550/arXiv.2004.04917
[32] Khattar, A. & Quadri, S. M. K. CAMM: Cross-attention multimodal classification of disaster-related tweets. *IEEE Access* **10**, 92889–92902 (2022). https://doi.org/10.1109/ACCESS.2022.3202976
[33] Chen, J., Seng, K. P., Ang, L. M., Smith, J. & Xu, H. AI-empowered multimodal hierarchical graph-based learning for situation awareness on enhancing disaster responses. *Future Internet* **16**, 161 (2024). https://doi.org/10.3390/fi16050161
[34] Selvakumaran, S. et al. Improving operational use of post-disaster damage assessment for Urban Search and Rescue by integrated graph-based multimodal remote sensing data analysis. *Prog. Disaster Sci.* **25**, 100404 (2025). https://doi.org/10.1016/j.pdisas.2025.100404
[35] Shamta, I. & Demir, B. E. Development of a deep learning-based surveillance system for forest fire detection and monitoring using UAV. *PLoS One* **19**, e0299058 (2024). https://doi.org/10.1371/journal.pone.0299058
[36] Zhang, M., Huang, Q. & Liu, H. A multimodal data analysis approach to social media during natural disasters. *Sustainability* **14**, 5536 (2022). https://doi.org/10.3390/su14095536
[37] Gao, S. et al. Disaster situation awareness through multimodal social media data: Evidence from Typhoon Haikui. *Appl. Sci.* **15**, 465 (2025). https://doi.org/10.3390/app15010465
[38] Liu, Y.-C. & Kuo, C.-L. Constructing spatio-temporal disaster knowledge graph from social media. *AGILE GISci. Ser.* **5**, 37 (2024). https://doi.org/10.5194/agile-giss-5-37-2024
[39] Yu, C., Hu, B. & Wang, Z. Open-world disaster information identification from multimodal social media. *Complex Intell. Syst.* **11**, 7 (2025). https://doi.org/10.1007/s40747-024-01635-5

ACKNOWLEDGEMENTS

The authors acknowledge both institutions for their support

AUTHOR CONTRIBUTIONS

Dannuri Mounika. designed and implemented the model architecture and training pipeline, conducted all experiments and statistical analyses, and wrote the manuscript. Johnson Kolluri supervised the work. Siripuri Kiran contributed to funding acquisition and participated in reviewing and editing the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile.docx