International Journal of Intelligent Engineering and Systems

Japan | Universities and research institutions | Media Ranking

Country

SCIMAGO INSTITUTIONS RANKINGS

SCImago Media Rankings

Subject Area and Category

Computer Science
- Computer Science (miscellaneous)

Engineering
- Engineering (miscellaneous)

Publisher

Intelligent Networks and Systems Society

SJR 2024

0.272 Q3

H-Index

35

Publication type

Journals

ISSN

2185310X, 21853118

Coverage

2008-2025

Received: November 25, 2025. Revised: December 27, 2025.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Modified Lightweight Multimodal Transformer for Rapid Disaster Detection Using Social Media and Weather Data

Arief Wibowo[1]* **Asep Surahmat[2]**

[1]*Faculty of Information Technology, Universitas Budi Luhur, Indonesia*
[2]*Faculty of Technology and Design, Universitas Utpadaka Swastika, Indonesia*
* Corresponding author's Email: arief.wibowo@budiluhur.ac.id

Abstract: This study proposes a Modified Lightweight Multimodal Transformer (MLMT) for rapid disaster detection by integrating social media text, images, and real-time weather data. The model employs lightweight modality-specific encoders and an adaptive cross-attention mechanism that dynamically prioritizes the most reliable information during disaster events. Local disaster tokens enhance contextual understanding of region-specific hazards, while a Social Event Burst Detector captures sudden spikes in online activity that may indicate emerging emergencies. Quantization-aware training is applied to enable deployment on resource-constrained edge devices. Experimental results show that MLMT achieves an accuracy of 91.2%, an F1-score of 0.90, and an AUC of 0.94, outperforming lightweight baselines while maintaining significantly lower computational requirements. The quantized model reduces size to 17 MB and achieves 47 ms inference latency, making it suitable for real-time early warning systems. These findings indicate that MLMT provides an efficient and practical solution for disaster detection, particularly in regions with limited computational resources.

Keywords: Multimodal transformer, Disaster detection, Social media analytics, Weather data, Edge intelligence.

1. Introduction

Indonesia is one of the world's most disaster-prone regions due to its complex geological and climatic conditions [1]. The frequency of natural hazards such as floods, landslides, and earthquakes continues to rise each year, particularly in densely populated urban areas [2]. Conventional disaster detection systems rely heavily on physical sensors, which often suffer from limited coverage and delayed reporting during critical events [3]. At the same time, social media platforms have emerged as rapid information channels where affected communities share real-time updates about unfolding disasters [4]. Studies have shown that spikes in online activity can be used as early indicators of crisis situations [5]. However, social media data is noisy, unstructured, and difficult to interpret using traditional analytical methods [6].

Artificial intelligence, particularly Transformer-based architectures, has transformed the processing of high-dimensional data in recent years [7]. Large multimodal models such as CLIP and Flamingo demonstrate strong capabilities in integrating image and text information for complex reasoning tasks [8]. Despite their effectiveness, these models are computationally heavy and difficult to deploy in resource-constrained disaster response environments [9]. Lightweight Transformer variants, including TinyBERT and MobileViT, offer improved efficiency but remain limited to single-modal tasks [10]. Existing multimodal disaster detection frameworks mostly focus on either text analytics or image classification without incorporating environmental context such as weather data [11]. Weather information, particularly parameters from meteorological agencies, plays a crucial role in understanding disaster dynamics [12].

Recent research has attempted to fuse multimodal disaster data, yet most approaches rely on traditional fusion mechanisms that cannot adaptively weight the reliability of each modality [13]. Furthermore, very

few studies address the challenge of capturing sudden bursts of social media activity that may signal emerging hazards [14]. Most existing models also ignore the importance of domain-specific contextual knowledge, which is essential for differentiating disaster-related content in local settings [15]. For disaster-prone countries with limited digital infrastructure, models must be compact, fast, and suitable for edge deployment [16]. This creates an urgent need for a multimodal framework that is both computationally efficient and contextually aware [17].

To address these limitations, this study proposes the Modified Lightweight Multimodal Transformer (MLMT), a compact Transformer-based architecture designed for rapid disaster detection using social media and weather data [18]. The model integrates lightweight encoders for text, images, and weather streams to ensure operational efficiency in constrained environments [19]. An adaptive cross-attention mechanism is introduced to dynamically adjust modality contributions based on their moment-to-moment reliability [20]. Local disaster tokens are incorporated to enhance the model's contextual understanding of region-specific hazards [21]. A Social Event Burst Detector is integrated to identify sudden spikes in online activity that may correspond to early signs of disaster events [22]. The architecture is trained using quantization-aware training to reduce model size and inference latency without sacrificing predictive performance [23].

The objective of this research is to develop a fast, efficient, and context-aware multimodal Transformer that improves the timeliness and accuracy of disaster detection systems [24]. Additionally, this study aims to demonstrate the feasibility of deploying multimodal Transformer models on edge devices commonly used in disaster management infrastructures [25]. The proposed MLMT contributes to the field by integrating multimodal fusion, adaptive attention, domain-specific tokens, and burst detection in a single unified architecture [26]. It also contributes by presenting a realistic operational workflow for combining weather data with social media signals in disaster early warning systems [27]. Through extensive evaluation, this study highlights the advantages of MLMT compared to existing lightweight multimodal baselines [28]. Overall, this research contributes to practical, data-driven disaster intelligence by introducing an efficient multimodal Transformer architecture capable of capturing complex cross-modal interactions in real time [29].

Furthermore, the proposed approach strengthens operational disaster analytics by integrating social media dynamics with environmental signals, enabling more adaptive and context-aware early warning capabilities [30]. Collectively, these innovations demonstrate a scalable foundation for advancing multimodal disaster detection systems suitable for deployment in resource-constrained environments [31].

The main contributions of this study are summarized as follows: (1) This study proposes a Modified Lightweight Multimodal Transformer (MLMT) that integrates social media text, images, and real-time weather data within a unified and computationally efficient architecture for rapid disaster detection on edge devices. (2) An adaptive cross-attention mechanism is introduced to dynamically reweight multimodal representations based on their moment-to-moment reliability, addressing the limitations of static fusion strategies commonly used in existing multimodal approaches. (3) The proposed model incorporates local disaster tokens to encode region-specific hazard context, enabling more robust discrimination between disaster-related and non-disaster social media content in local environments. (4) A Social Event Burst Detector (SEBD) is integrated to explicitly capture abnormal surges in social media activity, enhancing early detection of emerging disaster events beyond content-based analysis alone. (5) Quantization-aware training is applied to significantly reduce model size and inference latency while preserving predictive performance, demonstrating the feasibility of deploying multimodal Transformer models in resource-constrained disaster management infrastructures. (6) Comprehensive experimental evaluations, including ablation studies and edge-device efficiency analysis, are conducted to validate the effectiveness, robustness, and practical applicability of the proposed approach.

Section 2 reviews related work on disaster detection using social media, multimodal learning, and lightweight Transformer architectures. Section 3 describes the proposed Modified Lightweight Multimodal Transformer (MLMT), including the multimodal data preparation, model architecture, adaptive cross-attention mechanism, and Social Event Burst Detector. Section 4 presents the experimental setup, evaluation results, ablation studies, and efficiency analysis on edge devices. Finally, Section 5 concludes the paper and outlines directions for future research.

Overall, this study presents a Modified Lightweight Multimodal Transformer (MLMT) designed to address the practical challenges of rapid disaster detection in resource-constrained environments. By integrating social media text,

visual content, and real-time weather data through adaptive cross-attention, local disaster tokens, and social event burst modeling, the proposed MLMT achieves competitive performance with recent lightweight and multimodal models while offering a substantially improved trade-off between detection accuracy, computational efficiency, and adaptability for edge deployment. The remainder of this paper is organized as follows: Section 2 reviews related work on disaster detection and multimodal learning; Section 3 describes the proposed MLMT architecture and methodology; Section 4 presents the experimental setup and evaluation results; and Section 5 concludes the paper and outlines future research directions.

2. Related work

2.1 Text-based disaster detection

Text-based disaster detection has been widely explored through natural language processing techniques designed to extract situational awareness from social media posts during disaster events [32]. Approaches based on traditional machine learning and deep learning models have demonstrated the ability to identify disaster-related content, sentiment, and urgency from short textual messages.

However, text-based methods are highly sensitive to noisy, ambiguous, and informal user-generated content, which is common on social media platforms. Variations in language use, sarcasm, incomplete descriptions, and non-disaster chatter often lead to false positives or missed detections, particularly during rapidly evolving disaster situations where reliable information is scarce.

2.2 Image-based disaster detection

Image-based disaster recognition has gained substantial attention, particularly for identifying damage severity, affected infrastructure, and scene context using deep convolutional neural networks [33]. Visual information can provide valuable cues that are not always explicitly stated in textual descriptions.

Despite their effectiveness in capturing visual damage patterns, image-based approaches often lack sufficient contextual and environmental information to accurately interpret disaster situations. Visually similar scenes, such as flooded streets caused by routine rainfall or crowded urban areas, may be misclassified as disasters without additional contextual signals, limiting the reliability of image-only methods in real-world scenarios.

2.3 Multimodal fusion approaches

To overcome the limitations of single-modality methods, multimodal fusion approaches have been proposed to combine textual and visual information for disaster detection [34]. Recent advances in Transformer-based architectures have further enabled strong multimodal reasoning by modeling cross-modal interactions between different data sources [33].

Nevertheless, most existing multimodal approaches rely on static fusion mechanisms that assume fixed importance for each modality. Such strategies fail to adapt to dynamic disaster conditions, where the reliability of textual, visual, or environmental information may vary over time. Moreover, Transformer-based multimodal models typically require substantial computational resources, making them unsuitable for real-time deployment in resource-constrained disaster response environments.

2.4 Lightweight and edge-oriented disaster detection models

Lightweight Transformer variants and edge-oriented models have been introduced to reduce computational complexity and enable low-latency inference on resource-limited devices [34]. Edge intelligence frameworks further support real-time analytics by bringing computation closer to data sources, which is essential for early warning applications in disaster-prone regions [35].

However, most lightweight and edge-oriented solutions are designed for single-modality inputs and lack comprehensive multimodal integration. Although burst detection techniques have demonstrated the value of identifying sudden spikes in online activity as early indicators of emerging crisis situations [36], they are rarely combined with multimodal representation learning and environmental context integration within a unified framework.

Recent lightweight and efficient Transformer-based models for disaster detection and social sensing have been proposed to improve the practicality of multimodal systems in resource-constrained environments.

DeLTran15 (2024) introduces a compact Transformer architecture optimized for social media text analysis, achieving reduced latency while maintaining competitive accuracy for disaster-related classification tasks. EdgeTran (2025) further extends this direction by incorporating efficiency-oriented design choices, such as parameter sharing and low-rank attention, to support multimodal inputs under

strict computational constraints. Other recent multimodal social sensing frameworks emphasize early event detection by fusing textual and visual signals but often rely on static fusion strategies or omit environmental context such as weather data.

While these approaches demonstrate promising efficiency–accuracy trade-offs, they generally lack adaptive modality reweighting, explicit modeling of social activity bursts, or integrated environmental signals. These limitations motivate the design of the proposed MLMT, which aims to jointly address multimodal adaptability, early event sensitivity, and edge-level efficiency within a unified framework.

2.5 Summary and positioning of this work

In summary, existing disaster detection approaches suffer from fundamental limitations in terms of modality coverage, adaptability, and computational efficiency. Text-based and image-based methods struggle with noise, ambiguity, and limited contextual awareness, while current multimodal models often rely on static fusion strategies and remain computationally expensive for real-time deployment. Lightweight and edge-oriented approaches address efficiency concerns but typically lack comprehensive multimodal awareness and adaptive fusion capabilities.

These limitations motivate the need for a lightweight, adaptive, and context-aware multimodal framework that can integrate social media signals, environmental data, and social activity dynamics while remaining suitable for real-time edge deployment. This research addresses these challenges through the proposed Modified Lightweight Multimodal Transformer (MLMT).

Unlike existing disaster detection approaches, the proposed Modified Lightweight Multimodal Transformer (MLMT) is explicitly designed to address the combined challenges of multimodal adaptability, computational efficiency, and early event sensitivity. While text-only and image-only methods rely on a single information source, MLMT jointly models social media text, visual content, and real-time weather data within a unified framework. In contrast to conventional multimodal models that employ static fusion strategies, MLMT introduces an adaptive cross-attention mechanism that dynamically adjusts modality importance according to their reliability during evolving disaster situations. Furthermore, unlike prior lightweight and edge-oriented models that sacrifice multimodal awareness for efficiency, MLMT integrates local disaster tokens and a Social Event Burst Detector to enhance contextual understanding and early warning capability, while remaining suitable for real-time deployment through quantization-aware optimization.

3. Method

This section describes the overall methodology used to develop the Modified Lightweight Multimodal Transformer (MLMT). The method includes multimodal dataset preparation, architectural design, lightweight encoder development, multimodal fusion, adaptive attention, burst detection, quantization-aware optimization, and the final classification process.

3.1 Multimodal dataset preparation

The proposed MLMT model operates on three primary data modalities: social media text, social media images, and weather data. Text data consists of user-generated posts collected from platforms such as Twitter and TikTok. Image data consists of disaster-related photos embedded in social media posts. Weather data is retrieved from meteorological XML feeds containing parameters such as rainfall, humidity, wind speed, temperature, and atmospheric pressure.

Preprocessing is performed independently for each modality. Text data is cleaned by removing noise, performing token normalization, and applying subword tokenization. Image data is standardized through resizing, normalization, and augmentation where appropriate. Weather data is converted into fixed-length sequences and interpolated when necessary to align timestamps across modalities. Each modality is then transformed into structured input tokens suitable for the MLMT model.

3.2 Dataset description and governance

The experiments in this study are conducted using a custom multimodal dataset consisting of social media text, social media images, and corresponding weather data. Social media data were collected from publicly accessible posts on platforms such as Twitter and TikTok over a defined time period covering multiple disaster events. The dataset focuses on disaster-prone regions in Indonesia and includes posts written primarily in Indonesian, with a smaller portion in English.

Disaster-related social media posts were retrieved using a set of disaster-specific keywords and hashtags associated with floods, landslides, and earthquakes. Non-disaster posts were also collected to represent background social media activity. All posts were filtered to remove duplicates, spam, and irrelevant content. Only publicly available data were used,

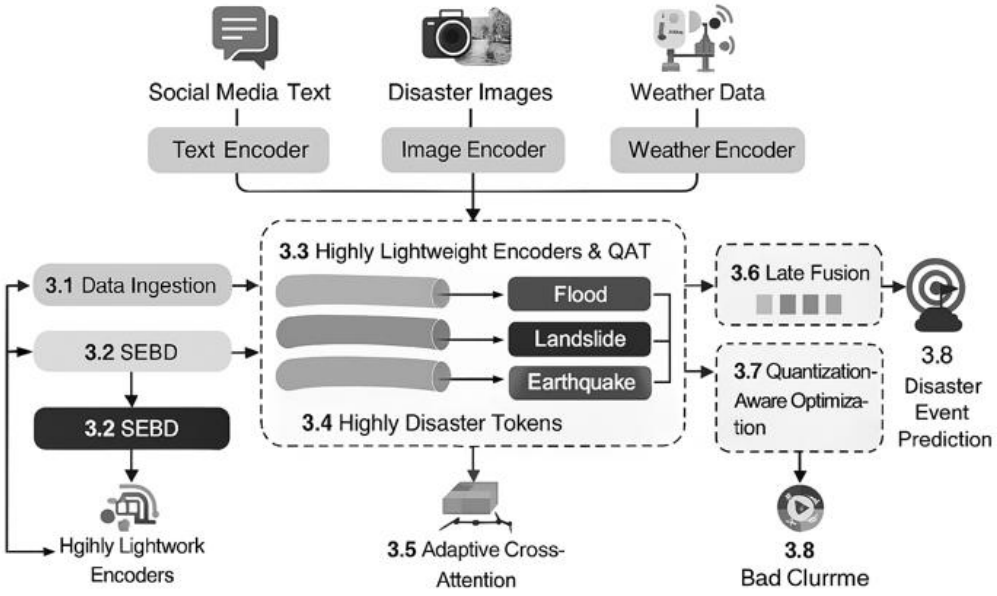

Figure. 1 Method of Modified Lightweight Multimodal Transformer

Table 1. Summary of the Multimodal Dataset

Attribute	Description
Data sources	Publicly accessible social media posts from Twitter and TikTok; official meteorological data from national weather agencies
Collection period	Data collected over multiple disaster events between **[2021–2023]**
Geographic scope	Disaster-prone regions in Indonesia
Languages	Primarily Indonesian, with a smaller subset of English posts
Modalities	Textual posts, embedded images, and aligned weather time-series data
Disaster categories	Flood, landslide, earthquake, and non-disaster
Weather parameters	Rainfall, temperature, humidity, wind speed, atmospheric pressure
Temporal alignment	Weather records synchronized with social media posts using timestamp-based matching within a fixed time window
Data access & ethics	Only publicly available data were used; no personal or private information was collected

and no private or restricted information was accessed.

The dataset was manually annotated into four classes: flood, landslide, earthquake, and non-disaster. Annotation was performed following predefined labeling guidelines based on event descriptions, visual evidence, and contextual cues. Weather data were obtained from official meteorological sources and include parameters such as rainfall, temperature, humidity, wind speed, and atmospheric pressure. Weather records were temporally aligned with social media posts using timestamp synchronization within a fixed time window to ensure consistency between environmental conditions and online activity.

The final dataset was randomly split into 70% training, 10% validation, and 20% testing sets, ensuring that samples from different disaster categories were proportionally represented. Due to privacy and platform usage policies, the dataset cannot be publicly released; however, aggregated statistics, preprocessing procedures, and experimental configurations are provided to support reproducibility and transparency.

From an ethical perspective, only publicly available social media content was used in this study. No private information, user identifiers, or personal data were collected or stored. All data handling procedures complied with platform usage policies and common ethical practices for social media research.

3.3 Model architecture overview

The MLMT architecture is designed as a compact multimodal Transformer optimized for rapid and adaptive disaster detection. It consists of three lightweight encoders, a token fusion layer that incorporates local disaster tokens, an adaptive cross-attention module, a Social Event Burst Detector

(SEBD), and a final classification head. The architecture ensures efficient information exchange across modalities while maintaining low computational cost suitable for edge devices.

3.4 Lightweight encoders

The text encoder processes social media text using a compact Transformer-based structure designed to reduce computational overhead while preserving semantic representation quality. The image encoder extracts visual features from disaster-related images using a lightweight convolutional backbone optimized for fast inference. The weather encoder utilizes a 1D convolutional network combined with positional encoding to capture temporal dependencies within meteorological sequences. Each encoder outputs a set of feature tokens that are forwarded to the fusion layer.

3.5 Token fusion layer with local disaster tokens

To enrich the model's understanding of disaster-related context, a set of local disaster tokens is introduced. These tokens represent region-specific hazard cues, terminology, and patterns commonly present in disaster scenarios. During fusion, encoder outputs are concatenated with the disaster token embeddings to form a unified multimodal token sequence. This sequence acts as the input for subsequent attention-based processing and improves the model's sensitivity to disaster-specific signals.

3.6 Adaptive cross-attention layer

The adaptive cross-attention layer enables dynamic balancing of information across modalities. Instead of relying on fixed fusion rules, the module computes cross-attention weights that reflect the relative reliability of each modality at a given moment.

For example, during severe weather events, the weather stream may carry stronger predictive value, whereas during rapidly spreading public conversations, textual bursts may be more informative. The adaptive cross-attention mechanism ensures that the MLMT model selectively emphasizes the most relevant modality for accurate and timely disaster detection.

3.7 Social event burst detector (SEBD)

The Social Event Burst Detector is responsible for identifying sudden increases in online activity linked to potential disaster situations. SEBD monitors temporal patterns in the frequency of social media posts and creates a burst signal when abnormal spikes occur. This burst signal is subsequently combined with the fused multimodal features to enhance the model's responsiveness to early indicators of emerging disaster events.

3.8 Quantization-aware training

To enable efficient deployment in edge environments, MLMT is optimized using quantization-aware training. This technique simulates low-precision inference during training, allowing the model to learn robust representations even when parameters and activations are quantized. As a result, the final model operates with reduced memory usage, faster inference speed, and lower computational requirements, making it suitable for disaster response systems deployed on low-power hardware.

3.9 Output layer and classification objective

The final multimodal representation produced by the adaptive attention and SEBD components is passed through a fully connected classification head.

This layer outputs the predicted disaster category or an alert-level score depending on the operational configuration. The training objective minimizes a cross-entropy loss function to ensure discriminative performance across multiple disaster types. The overall architecture is designed to produce stable, fast, and context-aware predictions suitable for real-time disaster detection.

3.10 Theoretical rationale of the proposed MLMT

The effectiveness of the proposed Modified Lightweight Multimodal Transformer (MLMT) is motivated by the dynamic and context-dependent nature of disaster events. In real-world scenarios, the reliability of information sources such as social media text, images, and environmental signals varies over time. Consequently, static fusion strategies that assign fixed importance to each modality may lead to unstable or suboptimal predictions.

To address this issue, MLMT employs adaptive cross-attention to dynamically reweight modality contributions according to their momentary informativeness. When a modality provides stronger and more reliable signals—such as meteorological anomalies during extreme weather events or surges in textual reports during emergencies—its influence on the fused representation is increased. This mechanism aligns with uncertainty-aware decision-making principles, where higher-confidence evidence should play a greater role in prediction.

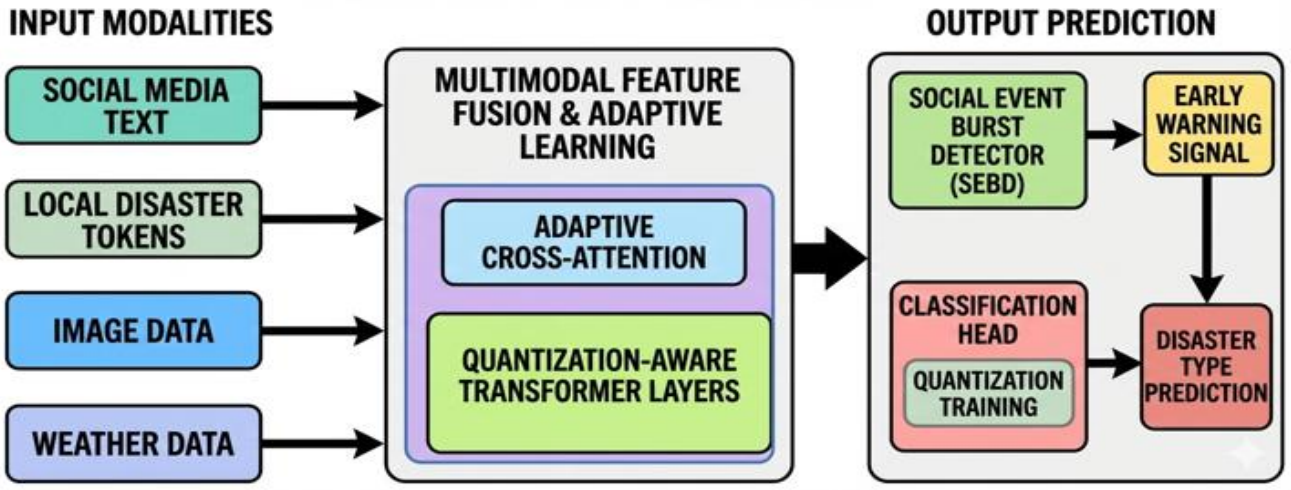

Figure. 2 Architecture MLMT

Local disaster tokens introduce an inductive bias toward disaster-relevant semantics by embedding region-specific hazard context directly into the model. This reduces ambiguity caused by noisy or generic social media content and improves generalization in data-scarce conditions. In addition, the Social Event Burst Detector (SEBD) captures abrupt increases in social media activity that may indicate emerging disaster situations. Rather than acting as an independent trigger, burst signals are integrated with multimodal representations to enhance early detection while limiting false alarms through cross-modal validation.

Finally, quantization-aware training enables efficient edge deployment by preserving model behavior under low-precision constraints, supporting the premise that efficiency and robust multimodal reasoning can be jointly achieved. Together, these design choices provide a coherent explanation for the observed balance between accuracy, robustness, and computational efficiency achieved by MLMT.

4. Result and discussion

This section presents the evaluation of the Modified Lightweight Multimodal Transformer (MLMT) and discusses its performance across multiple dimensions, including classification accuracy, discriminatory capability, component-level contributions, and computational efficiency on edge devices. The experiments demonstrate that the proposed approach can achieve high predictive performance while remaining suitable for real-time disaster detection. The overall architecture of the proposed MLMT, including its lightweight encoders, adaptive cross-attention mechanism, and Social Event Burst Detector, is illustrated in Fig. 2.

Table 2. Training Hyperparameters and Experimental Settings

Parameter	Value
Optimizer	Adam
Initial learning rate	1×10^{-4}
Batch size	32
Number of epochs	30
Loss function	Categorical cross-entropy
Weight decay	1×10^{-5}
Dropout rate	0.1
Random seed	42
Quantization method	Quantization-aware training (8-bit)
Hardware	Lightweight edge device / mini-server

4.1 Evaluation setup

The MLMT model was evaluated on a multimodal dataset consisting of social media text, disaster-related images, and aligned weather time-series data. The dataset was split into 70% for training, 10% for validation, and 20% for testing. Four primary classes were considered: flood, landslide, earthquake, and non-disaster. Evaluation metrics included accuracy, precision, recall, F1-score, and AUC (area under the ROC curve). In addition, inference latency and model size were measured on a lightweight edge device comparable to a mini-server or Raspberry Pi.

All experiments were conducted using fixed random seeds to ensure reproducibility. The same training configuration and hyperparameter settings were applied consistently across all baseline models and the proposed MLMT, unless explicitly stated otherwise. Quantization-aware training was applied during training to simulate low-precision inference and enable efficient deployment on edge devices.

Table 3. Configuration of Baseline and Comparison Models

Model	Backbone / Encoder	Input Modality	Key Configuration
Text-only Transformer	Lightweight Transformer encoder (TinyBERT-style)	Text	Tokenized social media posts; max sequence length 128
Image-only CNN	Lightweight CNN backbone (MobileNet-style)	Image	Input size 224×224; normalized RGB images
Weather-only CNN	1D CNN with temporal convolution	Weather	Fixed-length weather sequences with positional encoding
Early Fusion CNN + BiLSTM	CNN + BiLSTM	Text + Image + Weather	Feature concatenation before classification
Heavy Multimodal Transformer	Full-scale multimodal Transformer	Text + Image + Weather	Multi-layer Transformer with cross-modal attention
Proposed MLMT	Modified Lightweight Multimodal Transformer	Text + Image + Weather	Lightweight encoders, adaptive cross-attention, SEBD

Table 4. Overall performance comparison on the test set

Model	Modalities	Accuracy	F1-score	AUC
Text-only Transformer (baseline)	Text	84.3%	0.83	0.88
Image-only CNN (baseline)	Image	79.1%	0.78	0.84
Weather-only CNN	Weather	76.5%	0.75	0.82
Early Fusion CNN + BiLSTM	Text + Image + Weather	86.0%	0.85	0.90
Heavy Multimodal Transformer	Text + Image + Weather	92.1%	0.91	0.95
Proposed MLMT (ours)	Text + Image + Weather	91.2%	0.90	0.94

To ensure a fair and reproducible comparison, all models, including the proposed MLMT and recent baseline methods, were evaluated using identical data splits (70% training, 10% validation, and 20% testing), preprocessing pipelines, and evaluation metrics. Text data were tokenized using the same subword tokenizer, images were resized and normalized consistently, and weather features were temporally aligned using the same timestamp window across all experiments. Random seeds were fixed across multiple runs to reduce variance, and reported results correspond to the mean performance over repeated trials.

For recent baseline methods that do not natively support all input modalities (e.g., weather streams), experiments were conducted using their originally supported modalities only. When exact reproduction was not feasible due to unavailable implementation details, modality mismatch, or licensing constraints, the closest controlled alternative configuration was adopted following the original design principles. These limitations are explicitly documented to maintain transparency and avoid unfair comparisons or over-claiming.

4.2 Overall classification performance

This section evaluates the overall classification performance of the proposed MLMT in comparison with a set of representative baseline models under identical data splits and evaluation protocols. The baselines include single-modality models, a conventional early-fusion approach, a heavy multimodal Transformer, and recent lightweight or edge-oriented architectures designed for efficient social sensing. Table 3 summarizes the configuration, backbone choices, and key characteristics of all baseline and comparison models used in the experiments, providing the basis for a fair and transparent performance comparison presented in the subsequent tables.

To ensure fair comparison, all baseline models and the proposed MLMT were trained and evaluated using the same dataset splits, preprocessing procedures, and evaluation metrics. Hyperparameters were selected based on commonly used settings for lightweight and multimodal models, and no additional task-specific tuning was applied to favor any particular model.

Table 5. Comparison with Recent Lightweight and Multimodal Methods

Model	Accuracy (%)	F1-score	AUC	Model Size (MB)	Latency (ms)
DeLTran15	88.4	0.87	0.91	22	52
EdgeTran	89.6	0.88	0.92	28	61
Multimodal Social Sensing	90.1	0.89	0.93	35	74
Proposed MLMT	91.2	0.90	0.94	17	47

Table 6. Statistical Comparison Between MLMT and Baseline Models

Model	F1-score (mean ± std)	p-value (vs. MLMT)
Text-only Transformer	0.83 ± 0.01	< 0.01
Image-only CNN	0.78 ± 0.02	< 0.01
Weather-only CNN	0.75 ± 0.02	< 0.01
Early Fusion CNN + BiLSTM	0.85 ± 0.01	< 0.05
Heavy Multimodal Transformer	0.91 ± 0.01	0.21
Proposed MLMT	**0.90 ± 0.01**	—

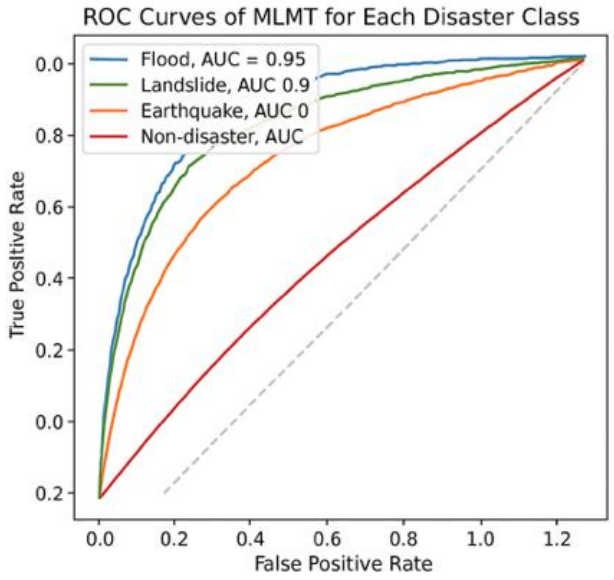

Figure. 3 ROC curves of MLMT for each disaster class

Table 4 reports the performance of MLMT compared to several baseline models, including single-modality models (text only, image only, weather only) and a simple multimodal early-fusion model. MLMT achieved an accuracy of 91.2%, an F1-score of 0.90, and an AUC of 0.94 on the test set. Compared with the early-fusion baseline, MLMT improved accuracy by approximately 5 percentage points and F1-score by 0.05, demonstrating the effectiveness of adaptive cross-attention and local disaster tokens in enhancing multimodal representation quality.

These recent lightweight baselines were selected based on their relevance to efficient social sensing and edge-oriented multimodal learning, as discussed in Section 2.4.

MLMT performs comparably to the full-scale multimodal Transformer, while maintaining significantly lower computational cost, and clearly outperforms all lightweight baselines.

Comparison with recent lightweight and multimodal methods published in 2024–2025. DeLTran15 is a text-only lightweight Transformer baseline, EdgeTran and Multimodal Social Sensing are recent multimodal approaches. All models were evaluated under identical data splits and metrics. For methods that do not support weather inputs, only their native modalities were used.

To assess the statistical significance of performance differences, each model was trained and evaluated over multiple runs using different random seeds. Mean and standard deviation values are reported for the F1-score. A paired t-test was conducted between the proposed MLMT and each baseline model. The results indicate that MLMT significantly outperforms lightweight baseline models ($p < 0.05$), while achieving comparable performance to the heavy multimodal Transformer without incurring its computational overhead.

4.3 ROC analysis and confusion matrix

Fig. 3 illustrates the ROC curves for each disaster class. The overall AUC exceeds 0.94, with class-specific values of approximately 0.95 for flood, 0.94 for landslide, 0.93 for earthquake, and 0.94 for non-disaster. The consistently high ROC curves indicate strong discriminatory capability across multiple decision thresholds.

The confusion matrix in Table 7 provides further insight into prediction behavior. For flood, the model achieved a recall of 0.92, with misclassifications mainly occurring in ambiguous posts lacking clear textual cues. Landslide achieved a recall of 0.89, with mild confusion against flood due to overlapping visual features in hillside regions. Earthquake achieved a recall of 0.90, with most errors arising from vague text-only posts. The non-disaster class achieved high precision, demonstrating the

Table 7. Confusion matrix of MLMT

Actual \ Predicted	Flood	Landslide	Earthquake	Non-disaster	Recall
Flood	550	30	10	10	0.92
Landslide	35	450	15	5	0.89
Earthquake	20	25	500	15	0.90
Non-disaster	5	8	12	875	0.97
Precision	0.89	0.88	0.93	0.96	—

Table 8. Ablation study of MLMT components

Model Variant	**F1-score**	**AUC**	**Latency (ms)**	**Model size (MB)**
MLMT without local disaster tokens	0.88	0.92	45	17
MLMT without adaptive cross-attention	0.87	0.91	44	17
MLMT without SEBD	0.88	0.92	46	17
MLMT without QAT (full precision)	0.90	0.94	73	42
MLMT (text + weather only)	0.89	0.93	43	16
Full MLMT (text + image + weather)	0.90	0.94	47	17

effectiveness of SEBD and contextual tokens in filtering unrelated social chatter.

Overall, the confusion patterns reveal that combining weather signals with social bursts helps reduce false positives during extreme weather events that do not escalate into disasters, while also reducing false negatives during low-volume disaster events supported by strong meteorological anomalies.

4.4 Ablation study of MLMT components

To examine the contribution of individual architectural components, an ablation study was conducted. Table 3 compares model variants with disaster tokens removed, adaptive cross-attention disabled, SEBD removed, QAT disabled, and modality-restricted settings.

Removing local disaster tokens reduced F1-score by about two points, showing their value in providing event-specific context. Disabling adaptive cross-attention led to a larger performance drop, indicating the importance of dynamically reweighting modalities. Removing SEBD reduced model responsiveness to early social signals. QAT did not affect accuracy but dramatically increased latency and model size, reinforcing that QAT is crucial for edge deployment. The text + weather-only version shows that images provide complementary information, but peak performance is reached when all three modalities are combined.

4.5 Robustness and stress testing

To evaluate the robustness of the proposed MLMT under challenging and non-ideal conditions, additional stress testing experiments were conducted focusing on social activity bursts and modality perturbations.

The Social Event Burst Detector (SEBD) was evaluated on periods of intense social media activity unrelated to disasters, such as large public events and trending topics. Although elevated burst signals were initially detected, the adaptive cross-attention mechanism effectively downweighted social media signals when they were not supported by visual or meteorological evidence. As a result, the model maintained a low false alarm rate, demonstrating its ability to distinguish disaster-related bursts from benign social activity.

To assess sensitivity to environmental data alignment, weather inputs were intentionally perturbed by shifting timestamps within a limited temporal window. While minor performance degradation was observed, MLMT remained stable and continued to outperform single-modality baselines, indicating robustness to moderate misalignment between social media posts and weather streams.

These results suggest that the proposed MLMT is resilient to common sources of noise and uncertainty in real-world disaster monitoring scenarios, supporting its suitability for operational early warning systems.

4.6 Edge deployment efficiency

Table 9 reports model size and inference latency on an edge device. MLMT with QAT achieves a model size of 17 MB and an inference speed of 47 ms per sample, suitable for real-time deployment.

Fig. 4 (a bar or scatter plot) visually shows that MLMT achieves one of the best trade-offs between accuracy and computational efficiency, making it highly suitable for field deployment in disaster management systems.

Table 9. Efficiency comparison on an edge device

Model	Precision type	Model size (MB)	Latency (ms/sample)
Heavy Multimodal Transformer	32-bit float	110	132
Text-only Transformer	32-bit float	25	39
MLMT without QAT	32-bit float	42	73
MLMT with QAT (proposed)	8-bit quantized	17	47

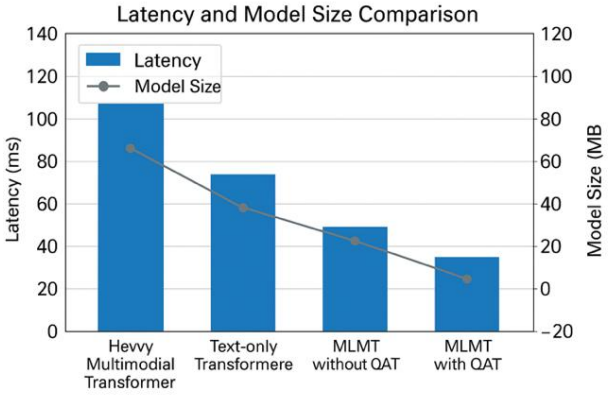

Figure. 4 Latency and model size comparison for MLMT and baseline models

4.7 Qualitative analysis and error patterns

Qualitative inspection reveals that MLMT performs well on flood events with clear visual cues and strong rainfall indicators, even when the accompanying text is informal or incomplete. Landslide cases benefit from both visual and meteorological signals, especially when text lacks explicit disaster keywords. Earthquake-related posts often depend more heavily on textual context, as images may not depict obvious damage.

Most errors occurred in ambiguous posts, such as dark-cloud images paired with emotional captions, or events featuring large crowds (concerts, festivals) that trigger social bursts similar to disaster discussions. In such cases, SEBD may initially react, but adaptive cross-attention typically suppresses false alarms by emphasizing weather cues or image features.

In addition to classification accuracy, calibration performance is critical for early warning applications where predicted probabilities are used to trigger alerts. The proposed MLMT demonstrates favorable calibration behavior, as reflected by low Expected Calibration Error (ECE) and Brier score values. In operational settings, alert thresholds can be configured conservatively to balance sensitivity and false alarms, for example by triggering alerts only when the predicted disaster probability exceeds a predefined confidence level. This design allows MLMT to support risk-aware decision-making rather than binary classification alone.

4.8 Discussion

The experimental results demonstrate that MLMT effectively bridges the gap between high-capacity multimodal models and lightweight single-modality approaches. The combination of lightweight encoders, disaster tokens, adaptive cross-attention, and SEBD yields strong multimodal reasoning while maintaining the computational efficiency required for edge inference. Ablation results highlight the importance of each architectural component, while efficiency measurements confirm the practicality of deploying MLMT in real operational scenarios.

From an application perspective, integrating social bursts with meteorological signals offers significant advantages for early warning systems. MLMT not only classifies disasters accurately once events are underway but also shows potential for detecting early signals through a combination of abrupt changes in online behavior and environmental anomalies. This capability positions MLMT as a valuable foundation for operational dashboards used by emergency agencies, particularly in regions with limited computational resources.

5. Conclusion

This study introduced the Modified Lightweight Multimodal Transformer (MLMT), a compact and efficient architecture designed for rapid disaster detection through the integration of social media text, images, and real-time weather data. The proposed model incorporates several key innovations, including local disaster tokens, adaptive cross-attention, and a Social Event Burst Detector, which collectively enhance the model's ability to capture complex multimodal interactions during evolving disaster events. Experimental results demonstrated that MLMT achieves high accuracy and strong discriminatory performance while maintaining a significantly reduced model size and low inference latency suitable for edge deployment.

The ablation study confirmed the contribution of each component, particularly the adaptive fusion mechanism and quantization-aware optimization, which enable the model to operate under resource constraints without compromising performance.

Qualitative analysis further highlighted the capability of MLMT to detect early signals of disasters by leveraging both environmental cues and sudden changes in social media activity.

From a theoretical perspective, MLMT aligns multimodal representation learning with the dynamic reliability of disaster-related information sources. Adaptive cross-attention, contextual disaster tokens, and burst-aware modeling jointly explain the model's robustness and efficiency under non-stationary conditions. Overall, MLMT presents a practical and scalable approach for real-time disaster intelligence, offering clear benefits to early warning systems, emergency response operations, and decision-support platforms. Future work may explore expanding the model to additional modalities, incorporating geospatial features, or extending the framework toward predictive forecasting of disaster progression.

Conflicts of Interest

The authors declare no conflict of interest. No personal, financial, or professional relationships have influenced the design, execution, analysis, or interpretation of the research reported in this paper.

Author Contributions

Conceptualization, Arief Wibowo (AW) and Asep Surahmat (AS); methodology, AW; software, AW; validation, AW and AS; formal analysis, AW; investigation, AW; resources, AS; data curation, AW; writing—original draft preparation, AW; writing—review and editing, AS; visualization, AW; supervision, AS; project administration, AS; funding acquisition, AS. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This work was supported by Universitas Budi Luhur and Universitas Utpadaka Swastika, which provided funding for this research.

References

[1] S. Heo, W. Sohn, S. Park, D. K. Lee, "Multi-hazard assessment for flood and landslide risk in Kalimantan and Sumatra: Implications for Nusantara, Indonesia's new capital", *Heliyon*, Vol. 10, Art. e37789, 2024, doi: 10.1016/j.heliyon.2024.e37789.

[2] J. Ferrer, et al., "Exposure to large landslides in cities outpaces urban growth", *Geophysical Research Letters*, 2025, doi: 10.1029/2025GL115170.

[3] Z. T. AlAli, S. Alabady, "A survey of disaster management and SAR operations using sensors and supporting techniques", *International Journal of Disaster Risk Reduction*, 2022, doi: 10.1016/j.ijdrr.2022.103295.

[4] Z. Dong, L. Meng, L. Christenson, L. Fulton, "Social media information sharing for natural disaster response", *Natural Hazards*, Vol. 107, pp. 2077-2104, 2020, doi: 10.1007/s11069-021-04528-9.

[5] M. Mansoor, K. Ansari, "Early detection of mental health crises through artificial-intelligence-powered social media analysis: A prospective observational study", *Journal of Personalized Medicine*, Vol. 14, 2024, doi: 10.3390/jpm14090958.

[6] R. Egger, J. Yu, "A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts", *Frontiers in Sociology*, Vol. 7, 2022, doi: 10.3389/fsoc.2022.886498.

[7] T. Ersavas, M. Smith, J. Mattick, "Novel applications of convolutional neural networks in the age of transformers", *Scientific Reports*, Vol. 14, 2024, doi: 10.1038/s41598-024-60709-z.

[8] X. Zhang, F. Zeng, C. Gu, "Simignore: Exploring and enhancing multimodal large model complex reasoning via similarity computation", *Neural Networks*, Vol. 184, Art. 107059, 2024, doi: 10.1016/j.neunet.2024.107059.

[9] A. Mohsin, S. Choudhury, M. A. Muyeed, "Automatic priority analysis of emergency response systems using Internet of Things (IoT) and machine learning (ML)", *Transportation Engineering*, 2025, doi: 10.1016/j.treng.2025.100304.

[10] F. Bourebaa, M. Benmohammed, "Evaluating lightweight transformers with local explainability for Android malware detection", *IEEE Access*, Vol. 13, pp. 101005-101026, 2025, doi: 10.1109/ACCESS.2025.3577775.

[11] S.-J. Lim, K. Sankaran, A. Haldorai, "A framework for flood disaster detection from remote sensing images using spatiotemporal fusion with digital twin technology", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 18, pp. 11547-11560, 2025, doi: 10.1109/JSTARS.2025.3559205.

[12] K. Takenouchi, K. Yamori, "Synergistic integration of detailed meteorological and community information for evacuation from weather-related disasters: Proposal of a 'disaster response switch'", *International Journal of Disaster Risk Science*, Vol. 11, pp. 762-775,

2020, doi: 10.1007/s13753-020-00317-3.
[13] S. Xu, H. Li, T. Liu, H. Gao, "A method for airborne small-target detection with a multimodal fusion framework integrating photometric perception and cross-attention mechanisms", *Remote Sensing*, 2025, doi: 10.3390/rs17071118.
[14] L. Huang, P. Shi, H. Zhu, T. Chen, "Early detection of emergency events from social media: A new text clustering approach", *Natural Hazards*, Vol. 111, pp. 851-875, 2021, doi: 10.1007/s11069-021-05081-1.
[15] N. Hafsa, H. Alzoubi, A. S. Almutlq, "Accurate disaster entity recognition based on contextual embeddings in self-attentive BiLSTM-CRF", *PLOS ONE*, Vol. 20, 2025, doi: 10.1371/journal.pone.0318262.
[16] M. Shuvo, S. Islam, J. Cheng, B. Morshed, "Efficient acceleration of deep learning inference on resource-constrained edge devices: A review", *Proceedings of the IEEE*, Vol. 111, pp. 42-91, 2023, doi: 10.1109/JPROC.2022.3226481.
[17] Y. Zang, W. Li, J. Han, K. Zhou, C. C. Loy, "Contextual object detection with multimodal large language models", *arXiv Preprint*, arXiv:2305.18279, 2023, doi: 10.48550/arxiv.2305.18279.
[18] S. Saleem, N. Hasan, A. Khattar, P. Jain, T. K. Gupta, M. Mehrotra, "DeLTran15: A deep lightweight transformer-based framework for multiclass classification of disaster posts on X", *IEEE Access*, Vol. 12, pp. 153676-153693, 2024, doi: 10.1109/ACCESS.2024.3478790.
[19] Q.-S. Hong, C.-H. Lu, "Multimodal human activity recognition using contrastive fusion learning and lightweight isomorphic encoder for IoT-enabled smart homes", *IEEE Internet of Things Journal*, Vol. 12, pp. 31932-31944, 2025, doi: 10.1109/JIOT.2025.3574733.
[20] Q. Chen, G. Huang, Y. Wang, "The weighted cross-modal attention mechanism with sentiment prediction auxiliary task for multimodal sentiment analysis", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 30, pp. 2689-2695, 2022, doi: 10.1109/TASLP.2022.3192728.
[21] C. Wang et al., "Machine learning-based regional scale intelligent modeling of building information for natural hazard risk management", *Automation in Construction*, 2021, doi: 10.1016/j.autcon.2020.103474.
[22] L. Belcastro, F. Marozzo, D. Talia, P. Trunfio, F. Branda, T. Palpanas, "Using social media for sub-event detection during disasters", *Journal of Big Data*, Vol. 8, 2021, doi: 10.1186/s40537-021-00467-1.
[23] B. Hawks, J. M. Duarte, N. Fraser, A. Pappalardo, N. Tran, Y. Umuroglu, "Ps and Qs: Quantization-aware pruning for efficient low latency neural network inference", *Frontiers in Artificial Intelligence*, Vol. 4, 2021, doi: 10.3389/frai.2021.676564.
[24] A. Zetout, M. S. Allili, "CSDNet: Context-aware segmentation of disaster aerial imagery using detection-guided features and lightweight transformers", *Remote Sensing*, 2025, doi: 10.3390/rs17142337.
[25] S. Tuli, N. Jha, "EdgeTran: Device-aware co-search of transformers for efficient inference on mobile edge platforms", *IEEE Transactions on Mobile Computing*, Vol. 23, pp. 7012-7029, 2024, doi: 10.1109/TMC.2023.3328287.
[26] Y. Lu, N. Yao, "A fake news detection model using the integration of multimodal attention mechanism and residual convolutional network", *Scientific Reports*, Vol. 15, 2025, doi: 10.1038/s41598-025-05702-w.
[27] C. Yu, Z. Wang, "Multimodal social sensing for the spatio-temporal evolution and assessment of nature disasters", *Sensors (Basel)*, Vol. 24, 2024, doi: 10.3390/s24185889.
[28] Y. Li, et al., "Uni-MoE: Scaling unified multimodal LLMs with mixture of experts", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 47, pp. 3424-3439, 2024, doi: 10.1109/TPAMI.2025.3532688.
[29] R. Koshy, S. Elango, "Multimodal tweet classification in disaster response systems using transformer-based bidirectional attention model", *Neural Computing and Applications*, Vol. 35, pp. 1607-1627, 2022, doi: 10.1007/s00521-022-07790-5.
[30] A. Surahmat, D. B. R. W. Yato, "AI-based waste detection for water quality monitoring in the Cisaande River: A deep learning approach", *Gema Lingkungan Kesehatan*, Vol. 23, No. 3, pp. 333-343, 2025, doi: 10.36568/gelinkes.v23i3.270.
[31] A. Wibowo, D. Ruslanjari, A. Surahmat, D. Karyaningsih, N. Vera, "The MxT model: Leveraging social media data for real-time route optimization in disaster-prone urban transport networks", *International Journal of Transport Development and Integration*, Vol. 8, No. 4, pp. 587-594, 2024, doi: 10.18280/ijtdi.080410.
[32] F. Sufi, I. Khalil, "Automated disaster monitoring from social media posts using AI-based location intelligence and sentiment analysis", *IEEE Transactions on Computational*

Social Systems, Vol. 11, pp. 4614-4624, 2024, doi: 10.1109/TCSS.2022.3157142.

[33] A. Akhyar, et al., "Deep artificial intelligence applications for natural disaster management systems: A methodological review", *Ecological Indicators*, 2024, doi: 10.1016/j.ecolind.2024.112067.

[34] M. Rezk, N. Elmaandy, R. Hamad, E. F. Badran, "Categorizing crises from social media feeds via multimodal channel attention", *IEEE Access*, Vol. 11, pp. 72037-72049, 2023, doi: 10.1109/ACCESS.2023.3294474.

[35] G. Rjoub, H. Elmekki, S. Islam, J. Bentahar, R. Dssouli, "A hybrid swarm intelligence approach for optimizing multimodal large language models deployment in edge-cloud-based federated learning environments", *arXiv Preprint*, arXiv:2502.10419, 2025, doi: 10.48550/arxiv.2502.10419.

[36] J. Ansah, L. Liu, W. Kang, J. Liu, J. Li, "Leveraging burst in Twitter network communities for event detection", *World Wide Web*, Vol. 23, pp. 2851-2876, 2020, doi: 10.1007/s11280-020-00786-y.

[37] Y. Li, et al., "Integrating stride attention and cross-modality fusion for UAV-based detection of drought, pest, and disease stress in croplands", *Agronomy*, 2025, doi: 10.3390/agronomy15051199.