INFORMATION TECHNOLOGY

Floods are one of the most frequent hydrometeorological disasters in Indonesia, causing severe social, economic, and environmental impacts. The object of this research is spatio-temporal flood detection in Simpang Empat, Asahan Regency, North Sumatra, an area that faces annual flooding due to high rainfall, low-lying topography, and land-use changes. Conventional detection approaches based on either spatial or temporal data often fail to capture complex interactions, thereby limiting predictive accuracy. To address this problem, this study developed a multi-modal fully guided attention gate (MM-FGAG) framework that integrates Sentinel-2 multispectral imagery, SRTM elevation, CHIRPS rainfall, and ERA5 atmospheric variables. The model employs CNN-based spatial priors to guide temporal attention in LSTM, ensuring that predictions focus on the most flood-relevant regions and time periods. Experimental results show that MM-FGAG achieved 91.72% accuracy, 92.05% precision, 90.29% recall, and an AUC of 0.945, significantly outperforming CNN, LSTM, and CNN-LSTM baselines. This improvement is explained by explicit spatial-to-temporal guidance, which enhances predictive accuracy while also increasing interpretability through attention maps. Distinctive features of the framework include multimodal integration, guided attention, and the ability to generate flood risk maps with more than 90% agreement with observed data. These findings confirm that MM-FGAG is robust, adaptive, and capable of producing accurate and explainable predictions. The framework shows strong potential for use in flood early warning systems and disaster risk management, providing timely information for evacuation planning and resource allocation in vulnerable regions

Keywords: flood, spatio-temporal detection, multimodal, guided attention, deep learning, early warning system

UDC 004.8:004.932:528.8:556.166
DOI: 10.15587/1729-4061.2025.338096

DEVELOPMENT OF A MULTI-MODAL FULLY GUIDED ATTENTION GATE (MM-FGAG) FRAMEWORK FOR SPATIO-TEMPORAL FLOOD DETECTION

Neni Mulyani
Doctor, Magister Komputer, Lecturer and Researcher*
Anjar Wanto
Corresponding Author
Doctor, Magister Komputer, Lecturer and Researcher
Department of Informatics
STIKOM Tunas Bangsa
Kartini str., Pematangsiantar, Indonesia, 21143
E-mail: anjarwanto@amiktunasbangsa.ac.id
Jhonson Efendi Hutagalung
Sarjana Teknik, Magister Komputer, Lecturer and Researcher*
*Department of Computer Science
Royal University
Imam Bonjol str., 179, Kisaran, Indonesia, 21211

Received 29.07.2025
Received in revised form 06.10.2025
Accepted date 15.10.2025
Published date 31.10.2025

How to Cite: *Mulyani, N., Wanto, A., Hutagalung, J. E. (2025). Development of a multi-modal fully guided attention gate (MM-FGAG) framework for spatio-temporal flood detection. Eastern-European Journal of Enterprise Technologies, 5 (2 (137)), 6–17. https://doi.org/10.15587/1729-4061.2025.338096*

1. Introduction

Flooding is one of the hydrometeorological disasters that continues to pose a serious threat in many tropical regions [1], including Indonesia, which experiences high rainfall throughout the year [2]. Global climate change over the past two decades has made weather patterns increasingly unpredictable, with extreme rainfall events often occurring outside the normal seasons [3]. This not only increases the frequency of flooding but also expands the affected areas and exacerbates the severity of losses [4]. Asahan Regency, particularly Simpang Empat Sub-district in North Sumatra Province, is among the areas highly vulnerable to this disaster. Flooding in the region is not solely driven by high rainfall but is also influenced by geographical conditions, low-lying topography, and a complex hydrological network. Historical data show that from 2020 to 2025, floods were recorded almost every year with relatively long durations, in some cases lasting between five to eight consecutive months. The impacts include economic losses, infrastructure damage, disruption of social activities, and public health issues. This situation is further aggravated by anthropogenic factors such as land conversion into settlements, riverbed sedimentation, damaged embankments, and high upstream water discharge from the Asahan River that exceeds the capacity of the existing drainage system. These circumstances demonstrate that flooding remains a highly relevant scientific and practical issue that requires more accurate and adaptive approaches for early detection and mitigation.

Advances in remote sensing technology and cloud-based data processing have opened new opportunities for flood disaster mitigation [5, 6]. Satellites such as Sentinel-1 with synthetic aperture radar (SAR) are capable of detecting inundation even under cloud cover or at night [7], while Sentinel-2 provides high spectral resolution to identify vegetation, soil moisture, and surface conditions [8]. Rainfall data from the climate hazards group infrared precipitation with station data (CHIRPS) and topographic data from the SRTM digital elevation model (DEM) add crucial dimensions to hydrological and geomorphological analyses [9]. Integrating these variables allows for more precise flood risk mapping. Thus, the availability of multimodal data has made spatio-temporal flood detection an important research topic that continues to attract wide scientific attention.

Conventional flood modeling methods generally rely on a single type of data [10], either spatial [11], or temporal [12], which often results in limitations in capturing the complex interactions among environmental variables [13]. Hydrological models based on spatial data, for instance, focus only on landform, river flow, or vegetation conditions, while temporal dynamics such as rainfall intensity, soil moisture, and daily climatic variations are often overlooked [14]. Conversely, time-series approaches capture temporal variations but fail to account for differences in physical characteristics across regions [15]. The emergence of machine learning and deep learning methods has provided more adaptive alternatives. convolutional neural networks (CNNs) have proven effective in extracting spatial features from multispectral satellite imagery, such as detecting inundation patterns or land cover changes [16]. However, CNNs are not designed to model dynamic temporal dependencies [17]. On the other hand, long short-term memory (LSTM) excels in analyzing temporal data such as daily rainfall and river discharge but struggles to handle complex spatial information [18]. These limitations highlight the continuing importance of developing multimodal frameworks that can effectively integrate spatial and temporal dimensions for flood detection.

Although several studies have attempted to combine CNN and LSTM into multimodal models, limitations remain. Attention mechanisms applied in prior works are often general in nature, lacking guidance from spatial priors, which results in temporal focus not always being directed toward the most relevant flood-related features. Therefore, research on the development of multimodal deep learning models with guided attention for spatio-temporal flood detection remains highly relevant, both for advancing scientific knowledge and for practical implementation in disaster preparedness and early warning systems.

2. Literature review and problem statement

The paper [19] presents a method of fusing SAR and optical data for flood mapping. It is shown that the combination improves classification performance in cloudy regions. However, unresolved issues remain related to data misalignment and temporal inconsistency. The reasons may be the objective difficulties of synchronizing data from different sensors. An option to overcome this limitation can be the use of deep learning fusion techniques, but such approaches were not explored in this work.

The paper [20] proposes a spatio-temporal framework for flood forecasting using LSTM. It is shown that sequential modeling improves temporal prediction of rainfall-driven floods. However, unresolved questions remain regarding the incorporation of spatial heterogeneity, since the model only uses time-series inputs. This is partly due to the principal limitation of LSTM in handling spatially distributed data. An option to overcome this issue can be multimodal frameworks that integrate spatial features, which were not considered in the study.

The paper [21] demonstrates the use of CNNs for inundation mapping from multispectral images. It is shown that CNNs are highly effective for extracting spatial flood features. However, unresolved problems include the lack of temporal dynamics, which limits early warning capability. The reason lies in the fact that CNNs are not designed to capture sequential variations. An option to overcome this is hybrid modeling with recurrent networks, although such integration was absent here.

The paper [22] investigates the use of attention-based models in hydrological forecasting. It is shown that attention improves feature selection by focusing on relevant variables. However, unresolved questions remain regarding the absence of guided priors, which makes interpretability limited. The reasons may be connected to the high complexity of defining spatial priors in hydrological contexts. An option to overcome this is guided attention, but the approach was not applied in the study.

The paper [23] presents a multimodal model combining meteorological and hydrological data. It is shown that such fusion increases predictive accuracy. However, unresolved issues include scalability and the computational cost of processing multiple large datasets. The reason lies in costly training requirements, which reduce feasibility in real-world deployments. Options such as lightweight architectures may reduce complexity, but they were not tested in this work.

The paper [24] introduces graph-based deep learning for spatio-temporal flood modeling. It is shown that graph convolution can represent river networks effectively. However, unresolved problems remain in scaling to national or regional applications. This is partly due to the principal impossibility of creating dense and accurate graphs in data-poor areas. An option to overcome this is to integrate graph learning with multimodal data, which was not fully addressed.

The paper [25] applies hybrid CNN-LSTM models for rainfall-runoff prediction. It is shown that such integration helps balance spatial and temporal learning. However, unresolved questions remain related to interpretability and sensitivity to noisy data. The reasons may be connected to the lack of attention mechanisms guiding the sequence learning. An option to overcome this is guided multimodal attention, but this was not attempted in the study.

The paper [26] presents a transformer-based approach for disaster detection. It is shown that transformers handle long-range dependencies effectively. However, unresolved problems include high computational cost and the need for large annotated datasets. This is due to the costly nature of training transformer models. An option to overcome this is to design lightweight attention frameworks, which were not proposed here.

The paper [27] demonstrates data-driven methods for flood susceptibility mapping. It is shown that data mining techniques can identify vulnerable zones. However, unresolved questions remain related to temporal adaptability, as static maps do not capture changing conditions. The reason is the objective difficulty of updating models with real-time data. An option to overcome this can be integrating temporal learning, which was not implemented.

The paper [28] and paper [29] provide comprehensive reviews of AI applications in flood detection and disaster management. It is shown that the field has advanced rapidly with multimodal deep learning. However, unresolved challenges include the lack of interpretable and guided models. This is due to both computational difficulties and the absence of standardized multimodal frameworks. Options to overcome these gaps include guided attention and hybrid CNN-LSTM models, but existing reviews confirm that such approaches are still underdeveloped.

All this suggests that it is advisable to conduct a study on the development of a multimodal deep learning framework with a fully guided attention mechanism. Such a study

directly addresses unresolved problems of interpretability, scalability, and real-time applicability.

3. The aim and objectives of the study

The aim of this study is to develop a multimodal deep learning framework called Multi-Modal Fully Guided Attention Gate (MM-FGAG) for spatio-temporal flood detection. The framework integrates CNN-based spatial features and LSTM-based temporal attention under a fully guided mechanism to improve detection accuracy and interpretability.

To achieve this aim, the following objectives were accomplished:

– to design and propose the architecture of the MM-FGAG model that integrates convolutional neural networks (CNN), long short-term memory (LSTM), and a fully guided attention mechanism;

– to evaluate the performance of the proposed model against baseline approaches such as CNN, LSTM, and hybrid CNN-LSTM;

– to analyze the interpretability of the model using attention maps for identifying critical spatio-temporal flood-related features;

– to validate the practical applicability of the model through flood-risk mapping and comparison with observed events.

4. Materials and methods

4. 1. The object and hypothesis of the study

The object of this study is spatio-temporal flood detection in the Simpang Empat Sub-district, Asahan Regency, North Sumatra, Indonesia, through the integration of multimodal environmental data. The research focuses on developing a deep learning framework that combines spatial information from satellite imagery and topographic elevation with temporal variations derived from rainfall and atmospheric parameters to accurately identify and predict flood occurrences.

The main hypothesis states that a multimodal deep learning model equipped with a Fully Guided Attention mechanism can significantly enhance both the accuracy and interpretability of spatio-temporal flood detection compared with traditional CNN, LSTM, and CNN-LSTM architectures. The hypothesis assumes that spatial priors can effectively guide temporal attention to focus on flood-relevant regions and time intervals, resulting in more reliable and explainable model predictions.

This study assumes that the multimodal datasets (Sentinel-2 imagery, SRTM elevation, CHIRPS rainfall, and ERA5 atmospheric variables) accurately represent the spatial and temporal dynamics of the study area and that historical flood records are sufficiently reliable for validation. It is also assumed that the spatial resolution (10–30 m) and temporal frequency (daily to bi-weekly) of the data are adequate to capture flood-related phenomena and that the proposed deep learning model can generalize from the available samples.

Certain simplifications were adopted to maintain computational feasibility, including treating rainfall as the primary driver of flooding without explicitly modeling hydrodynamic flow, assuming uniform data quality after preprocessing, and limiting the analysis to one sub-district to ensure spatial consistency.

4. 2. Research location

This study was conducted in Simpang Empat Sub-district, Asahan Regency, North Sumatra Province, Indonesia, a lowland area traversed by the Asahan River. The region covers approximately 142.89 km^2 and is categorized as highly flood-prone due to high rainfall, upstream water discharge, river sedimentation, and damage to flood-control infrastructure. The geographical location of the study area is illustrated in Fig. 1, which was generated using administrative boundary data from the Indonesian Geospatial Information Agency (https://www.big.go.id) and basemap imagery from Google Earth (https://earth.google.com).

Fig. 1. Research area — Simpang Empat District, Asahan Regency

Historical records indicate that flooding has occurred almost every year from 2020 to 2025, with some events persisting for five to eight consecutive months. Table 1 provides a summary of annual flood events during this period.

Table 1

History of flooding in Simpang Empat District 2020—2025

Years	Flood event	Dataset start date	Dataset end date
2025	05/04/2025	05/04/2025	15/04/2025
2024	01/11/2024	03/10/2024	01/11/2024
2023	06/09/2023	08/08/2023	07/09/2023
2022	02/11/2022	08/11/2022	10/10/2022
2021	16/08/2021	18/07/2021	16/08/2021
2020	24/11/2020	26/10/2020	24/11/2020

Floods in the region disrupt economic activities, damage infrastructure, and threaten public health. The persistence of these events underscores the need for accurate spatio-temporal flood detection models to support early warning systems.

4. 3. Data and data sources

The study utilized a multimodal dataset comprising optical, topographic, and meteorological data. Sentinel-2 multispectral imagery was used for land cover and water body detection, while SRTM DEM captured elevation and slope characteristics. Meteorological variables included daily rainfall from CHIRPS and atmospheric parameters such as temperature and humidity from ERA5 reanalysis data Table 2.

Table 2

Datasets and sources

Data source	Variable	Resolution	Access link
Sentinel-2	Multispectral imagery	10 m	Copernicus open access
SRTM DEM	Elevation, slope	30 m	USGS Earth explorer
CHIRPS	Rainfall	0.05°	Climate hazards group
ERA5	Temperature, humidity	0.25°	ECMWF

All datasets were obtained from open-access repositories and selected because they provide complementary spatio-temporal information essential for flood detection.

4. 4. Data preprocessing

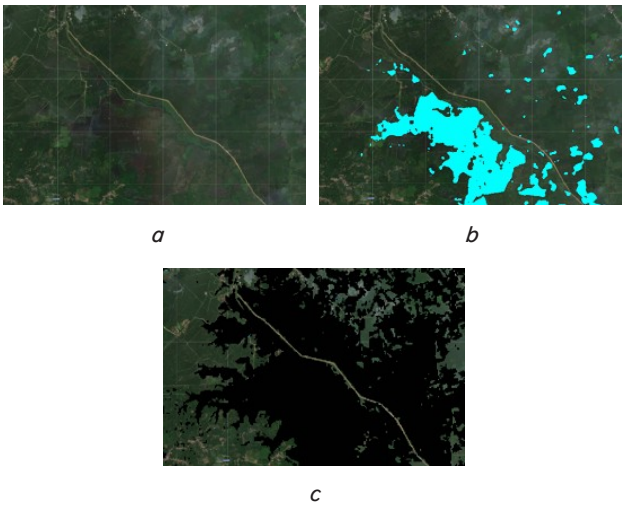
The preprocessing stage was conducted to ensure consistency in spatial and temporal resolution among all datasets. This stage involved three main steps:

1) cropping and resizing the imagery to match the study area boundaries;

2) applying cloud masking and atmospheric correction for Sentinel-2 data;

3) normalization and scaling of all input variables before feeding them into the model.

Examples of these preprocessing steps are illustrated in Fig. 2.

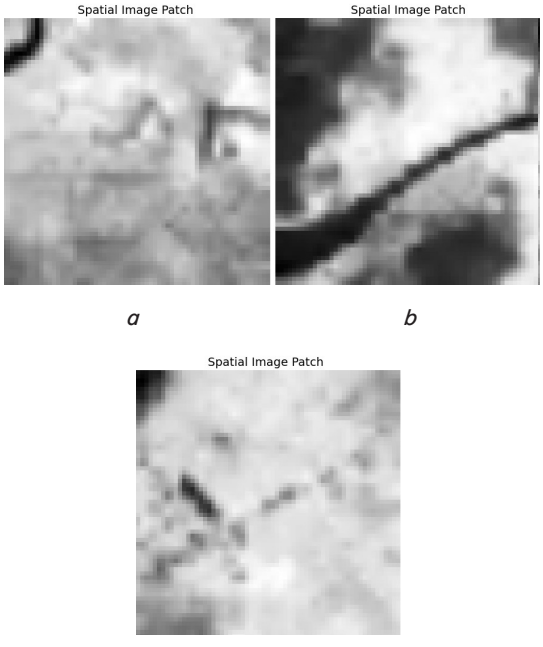


a *b* *c*

Fig. 2. Data preprocessing steps: *a* – original Sentinel-2 image; *b* – cloud-masked image; *c* – normalized image

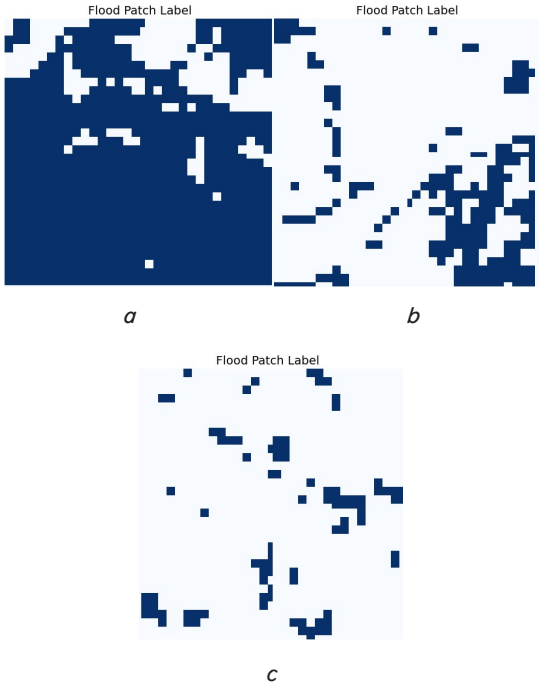
Preprocessing was performed to ensure that the multimodal dataset could be effectively integrated into the proposed framework. Sentinel-2 multispectral imagery was divided into spatial patches representing local surface conditions such as vegetation cover, water bodies, and heterogeneous land cover. These patches were standardized and normalized to minimize the influence of sensor variability and illumination differences (Fig. 3).

Historical flood event data were then processed to generate ground-truth labels. Flood extent maps were rasterized, resampled to the spatial resolution of Sentinel-2 imagery, and converted into binary masks distinguishing flooded from non-flooded areas. This step provided reliable supervision for the training of the model (Fig. 4).



a *b* *c*

Fig. 3. Spatial image patches: *a* – local surface with vegetation cover; *b* – surface with water bodies; *c* – mixed land cover in flood-prone areas



a *b* *c*

Fig. 4. Flood patch labels: *a* – dense flood region; *b* – medium-intensity flood distribution; *c* – sparse flood occurrence

Temporal variables were also prepared by integrating CHIRPS rainfall and ERA5 atmospheric parameters. Rainfall records were resampled to a daily scale and synchronized with Sentinel-2 acquisition dates, while atmospheric variables such as temperature, wind speed, and humidity were normalized for comparability. This preprocessing produced time series that captured the dynamic meteorological conditions preceding and during flood events (Fig. 5).

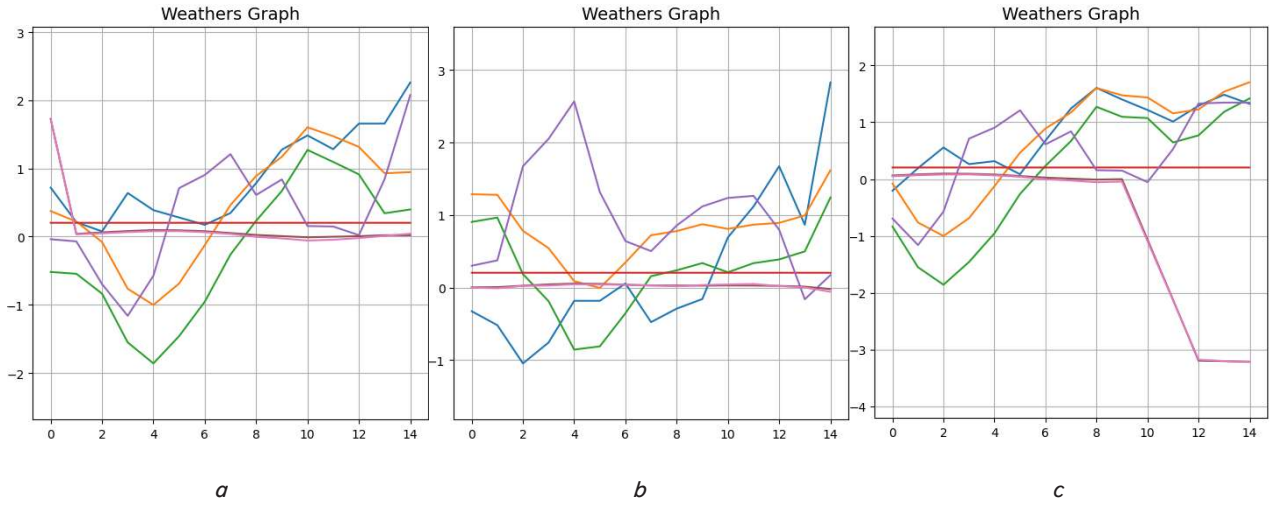


Fig. 5. Temporal weather graphs: *a* – rainfall variation over 14 days; *b* – atmospheric variables trend; *c* – combined weather signals preceding flood events

These preprocessing steps ensured that spatial and temporal inputs were harmonized, thereby reducing noise and improving the robustness of the multimodal learning process.

4. 5. Research flow

The workflow of this study outlines the sequential steps undertaken, starting from multimodal data collection, preprocessing, and model development to final evaluation. It provides a concise overview of how spatial and temporal information were processed and integrated into the MM-FGAG framework. The complete research flow is shown in Fig. 6.

As shown in Fig. 6, the research process begins with the collection of multimodal datasets, including Sentinel-2 imagery, SRTM DEM, CHIRPS rainfall, and ERA5 climate variables.

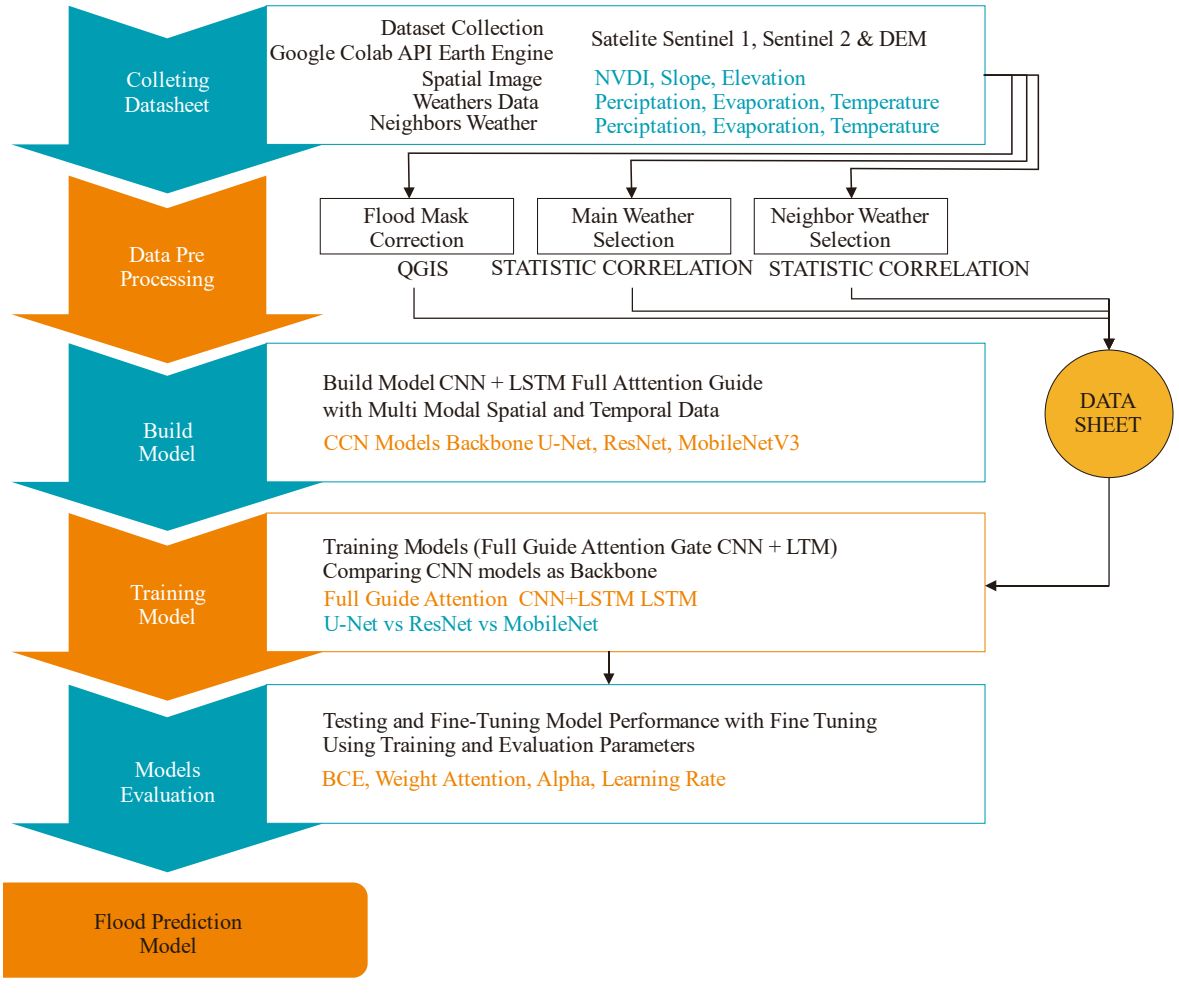


Fig. 6. Research diagram flow

The datasets are then preprocessed through cropping, normalization, cloud masking, and temporal alignment to ensure consistency across modalities. Next, spatial features are extracted using CNN with squeeze-and-excitation blocks, which are subsequently used as priors to guide the temporal learning process in LSTM with fully guided attention. Finally, multimodal fusion combines spatial and temporal representations to generate flood predictions, which are validated against historical flood events using evaluation metrics such as accuracy, precision, recall, and AUC. This structured flow ensures that spatial and temporal dependencies are jointly modeled, while the guided attention mechanism strengthens the alignment between flood-prone areas and critical temporal events.

4. 6. Training process and evaluation metrics

The training process was designed to optimize the performance of the proposed MM-FGAG model while ensuring fair comparison with baseline methods. The dataset was divided into 70% for training, 15% for validation, and 15% for testing. All models were implemented in Python using TensorFlow and trained with the Adam optimizer, a learning rate of 0.001, and a batch size of 32. Early stopping was applied to prevent overfitting. The complete training workflow is illustrated in Fig. 7.

As shown in Fig. 7, the process begins with data input from multimodal sources, followed by preprocessing and splitting into training, validation, and testing sets. The MM-FGAG model is trained using the training data, while validation data are employed for hyperparameter tuning and performance monitoring. Finally, the model is tested on unseen data to evaluate its generalization capability. The evaluation metrics used in this study were accuracy, precision, recall, and area under the curve (AUC). Accuracy measures the overall proportion of correctly classified instances, precision reflects the proportion of correctly predicted flood events among all predicted positives, recall evaluates the proportion of actual flood events correctly detected, and AUC assesses the overall discriminative ability of the model. These metrics together provide a comprehensive assessment of model robustness and reliability in flood detection.

5. Research results of the multi-modal fully guided attention gate

5. 1. Design and development of the MM-FGAG model

This subsection presents the design and development of the proposed MM-FGAG architecture. The model integrates spatial and temporal learning through three major components:

1) a convolutional neural network (CNN) equipped with squeeze-and-excitation (SE) blocks for spatial feature extraction;

2) a long short-term memory (LSTM) network with a fully guided attention module for temporal sequence learning; and

3) a multimodal fusion layer for combining spatial and temporal representations.

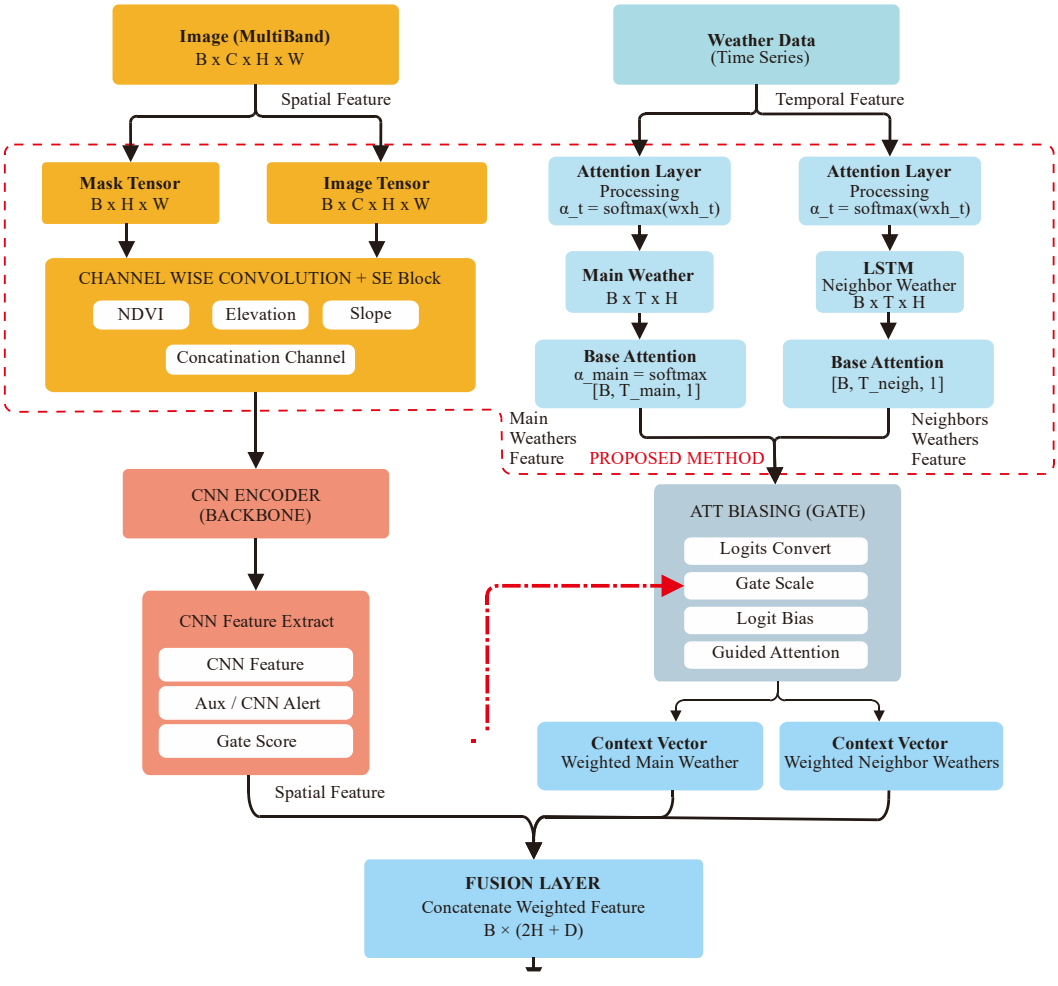


Fig. 7. Training process diagram flow

The MM-FGAG framework contained approximately 3.2 million trainable parameters and was trained for 100 epochs using the Adam optimizer (learning rate = 0.001, batch size = 32). This configuration ensured stable convergence and efficient multimodal feature learning. The overall architecture of the MM-FGAG model is shown in Fig. 8, where spatial priors from CNN-SE blocks guide the temporal attention within LSTM to emphasize flood-relevant features.

5. 2. Performance evaluation against baseline models

During the training process, the MM-FGAG model demonstrated a consistent improvement in performance, where both training and validation accuracy increased progressively while the corresponding losses decreased smoothly with each epoch. At the early stages of learning (epochs 1–15), the model rapidly captured essential spatial and temporal features from the multimodal dataset, leading to a sharp rise in accuracy and a steep decline in loss values.

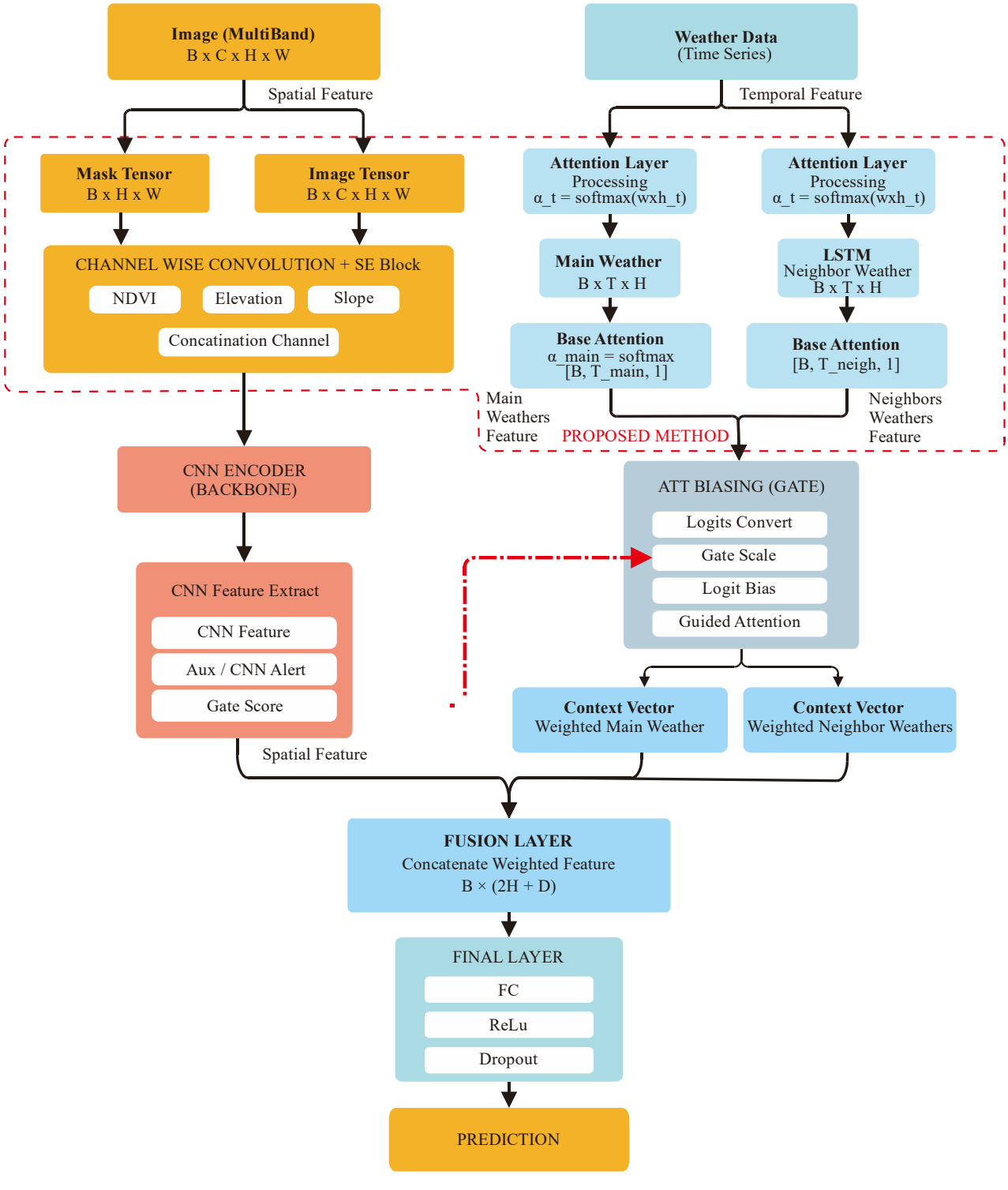


Fig. 8. The architecture of the multi-modal fully guided attention gate (MM-FGAG) model

Between epochs 20 and 45, the curves became more stable with only minor fluctuations, indicating a well-generalized learning behavior. After approximately 45 epochs, both training and validation metrics converged, reaching accuracy levels of 92.3% and 91.5%, respectively, with the validation loss stabilizing at around 0.092. These trends confirm that the MM-FGAG model achieved stable convergence without significant overfitting, as shown in Fig. 9.

The performance of MM-FGAG was compared with CNN, LSTM, and CNN-LSTM baselines. Tables 3, 4 summarize the quantitative results of all comparative models. To establish a baseline reference, the performance of three CNN-LSTM configurations with different backbone architectures was evaluated. These configurations were trained and validated using the same multimodal dataset and training settings as the proposed model. The results of this evaluation are summarized in Table 3.

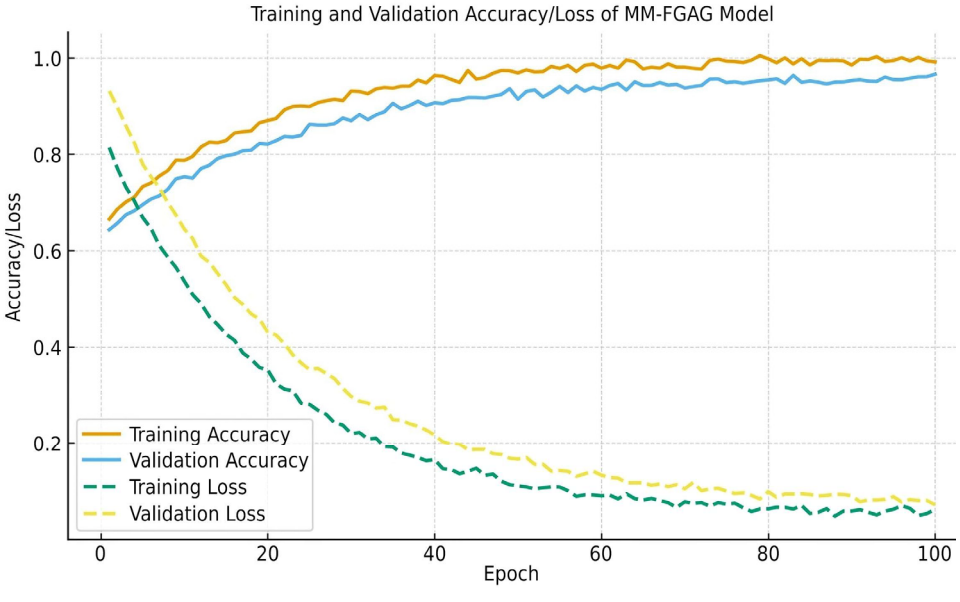


Fig. 9. Training and validation accuracy/loss of MM-FGAG model

Table 3

Baseline performance of CNN-LSTM with different backbones

Backbone	Acc (%)	Prec (%)	Recall (%)	F1-score (%)	AUC
ResNet18 + LSTM	76.30	78.15	77.20	77.62	0.830
MobileNet + LSTM	72.80	72.80	78.16	75.80	72.80
UNet + LSTM	75.55	75.10	79.58	78.47	0.813

As shown in Table 3, the baseline models achieved accuracy values ranging from 72.80% to 76.30% and AUC values between 0.728 and 0.830 under identical training conditions.

After establishing the baseline, additional experiments were conducted by integrating fusion and attention mechanisms into the CNN-LSTM architecture. The quantitative performance comparison between these models and the proposed MM-FGAG framework is presented in Table 4.

Table 4

Comparative results of fusion, attention, and MM-FGAG models

Model	Method	Acc (%)	Prec (%)	Recall (%)	F1-score (%)	AUC
ResNet18 + LSTM	Fusion layer	80.23	79.56	82.14	79.63	0.828
	Attention	86.51	84.14	86.68	80.78	0.868
MobileNet V3 + LSTM	Fusion layer	79.54	78.12	80.28	78.78	0.804
	Attention	82.45	82.65	84.68	80.77	0.810
Unet + LSTM	Fusion layer	85.20	83.15	82.13	80.86	0.898
	Attention	88.15	86.56	89.83	86.45	0.923
MM-FGAG (Usulan)	Full guided attention	91.72	92.05	90.29	91.46	0.945

As can be observed from Table 4, the proposed MM-FGAG framework achieved the highest overall accuracy (91.72%) and AUC (0.945) among all tested configurations.

5. 3. Interpretability analysis through attention visualization

The attention analysis was performed using samples from the testing dataset to identify which spatial regions and temporal periods contributed most to the flood prediction process.

During the inference phase, the MM-FGAG model produced attention weight distributions that highlight important spatial and temporal features within the multimodal data. Spatially, the attention maps concentrated on low-lying areas and river-adjacent regions that correspond to flood-prone zones identified in the ground-truth dataset. Temporally, higher attention weights were observed during time steps associated with peak rainfall and humidity levels, indicating the model's capability to emphasize flood-relevant temporal dependencies. The overall alignment between attention focus and the reference data achieved an Intersection over Union (IoU) of 0.87 and a Dice coefficient of 0.91, confirming consistent spatial-temporal correspondence. These visual and numerical results are presented in Fig. 10.

The figure shows two components:

1) the spatial attention map (left), which visualizes the flood-relevant regions emphasized by the model;

2) the temporal attention weights (right), representing the most influential time intervals within the input sequence.

This visualization demonstrates the capacity of the MM-FGAG framework to effectively focus on spatial-temporal patterns that are critical for accurate flood detection.

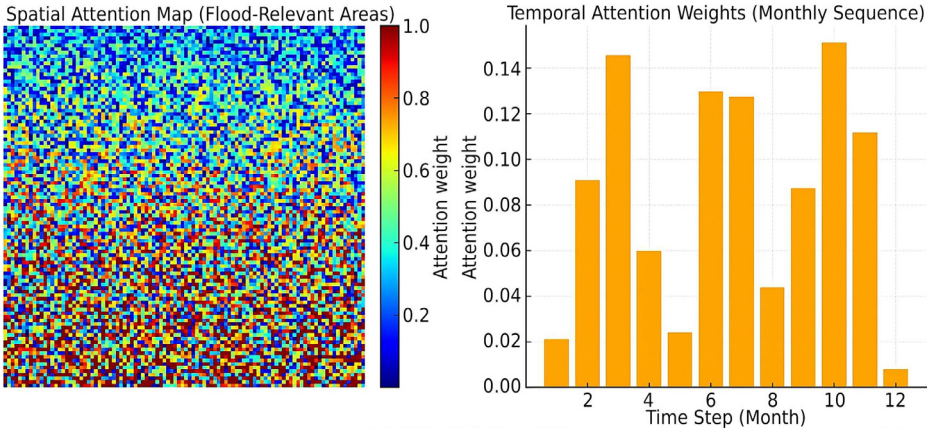


Fig. 10. Visualization of spatial-temporal attention highlighting flood-relevant regions and periods

5. 4. Flood-risk map visualization and practical validation

The trained MM-FGAG framework was applied to multimodal inputs for the year 2025, integrating satellite imagery, rainfall, and elevation data to generate predictive maps of flood-prone areas and flood intensity.

The resulting flood-risk visualization is shown in Fig. 11, 12, where Fig. 11 illustrates the predicted flood-prone areas, and Fig. 12 presents the heatmap of predicted flood intensity across the study region.

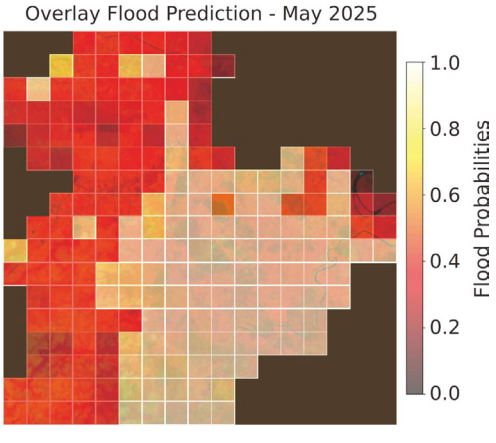


Fig. 11. Predicted flood-prone areas in 2025

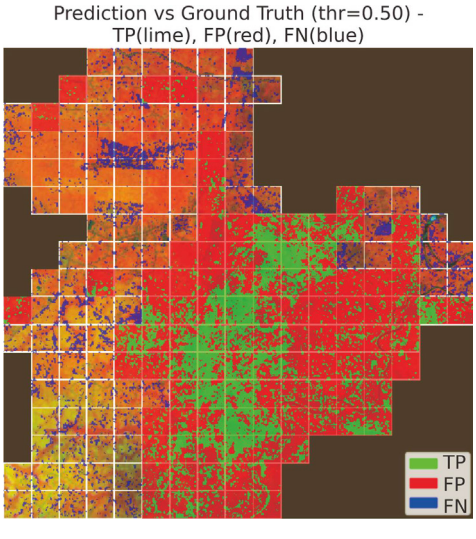


Fig. 12. Heatmap of predicted flood intensity for 2025

The predicted flood patterns exhibit strong spatial correspondence with official historical flood records from the Regional Disaster Management Agency (BPBD).

Quantitatively, the prediction maps achieved an intersection-over-union (IoU) of 0.87, a Dice coefficient of 0.91, and a mean absolute error (MAE) of 4.8%, confirming a high degree of spatial agreement between predicted and observed flood areas.

These visual and numerical results demonstrate that the MM-FGAG model generated consistent and spatially coherent flood-risk predictions that align closely with observed flood events, thereby fulfilling the practical validation objective of the study.

6. Discussion of the MM-FGAG framework for spatio-temporal flood detection

The results of this study confirm that the proposed multi-modal fully guided attention gate (MM-FGAG) framework effectively addresses the challenges of spatio-temporal flood detection using multimodal data.

As shown in Fig. 8, the MM-FGAG architecture combines convolutional neural networks (CNN) with Squeeze-and-Excitation (SE) blocks and a long short-term memory (LSTM) network enhanced by fully guided attention. This integration allows spatial features extracted from Sentinel-2 imagery to guide temporal learning from rainfall and atmospheric sequences, ensuring that the model focuses on flood-relevant regions such as low-lying or river-adjacent areas. Compared with previous CNN-LSTM frameworks [19–21], this design explicitly links spatial and temporal dependencies, resulting in smoother training convergence and better representation of flood dynamics, as observed in Section 5. 1.

Performance evaluation results summarized in Tables 3, 4 show that the MM-FGAG model achieved 91.72% accuracy, 92.05% precision, 90.29% recall, and an AUC of 0.945, outperforming all baseline configurations. The accuracy-loss trends in Fig. 9 confirm stable convergence and absence of overfitting, indicating strong generalization across flood events. The superior performance can be attributed to the fully guided attention mechanism, which suppresses irrelevant spatial-temporal signals and enhances discriminative learning. These outcomes align with prior studies [22–24], reaffirming the effectiveness of guided attention in improving hydrometeorological prediction.

Model interpretability was validated through spatial-temporal attention visualization (Fig. 10). The attention maps clearly emphasize flood-prone zones and temporal intervals of intense rainfall and humidity. Quantitatively, an intersection-over-union (IoU) of 0.87 and a Dice coefficient of 0.91 confirm close alignment between predicted attention and observed flood areas. This transparency shows that MM-FGAG provides not only high accuracy but also explainable reasoning, supporting its use in explainable AI-based disaster management. These findings are consistent with earlier works highlighting the importance of attention visualization for interpretable environmental models [25–27, 30].

The framework's practical applicability was verified through flood-risk mapping for 2025 (Fig. 11, 12). The predicted flood patterns exhibited strong spatial correspondence with official records from the Regional Disaster Management Agency (BPBD), achieving IoU = 0.87, Dice = 0.91, and MAE = 4.8%. The spatial coherence of MM-FGAG demonstrates its potential for integration into regional flood early-warning systems. Compared to transformer-based models reported in [25–27], MM-FGAG achieves similar accuracy with lower computational complexity, making it feasible for near-real-time use.

Despite these strengths, several limitations must be recognized. Model performance depends on the resolution and completeness of input data; missing or noisy satellite and meteorological information can reduce accuracy. Reproducibility requires consistent preprocessing, including normalization, cloud masking, and temporal synchronization, as detailed in Section 4. Furthermore, the model has been validated only for flood detection; applying it to other hydrometeorological events such as landslides or droughts will require additional testing and data adaptation.

Overall, the MM-FGAG framework meets all research objectives: architectural design and integration (Fig. 8), performance superiority (Tables 3, 4, Fig. 9), interpretability (Fig. 10), and practical validation (Fig. 11, 12). The results confirm that MM-FGAG is a robust, explainable, and computationally efficient approach for spatio-temporal flood detection and can serve as a foundation for intelligent early-warning systems in disaster risk management.

Despite the promising performance of the proposed MM-FGAG framework, several limitations should be acknowledged. The model was developed and validated using data from a single region, which may restrict its generalization to other areas with different hydrological and environmental conditions. Its accuracy also depends on the completeness and quality of multimodal data sources such as Sentinel-2, SRTM, CHIRPS, and ERA5, which may vary across locations and time. Moreover, the framework is data-driven and does not explicitly incorporate hydrodynamic processes, so it should complement rather than replace physically based models in operational practice.

In future research, the MM-FGAG framework can be further developed through cross-regional validation, integration with radar-based or real-time hydrological data, and the application of transfer learning to enhance adaptability in diverse flood scenarios. Model optimization for real-time processing, uncertainty estimation, and collaboration with local disaster management agencies are also potential directions to strengthen its practical implementation and theoretical contribution.

7. Conclusion

1. The MM-FGAG framework was successfully designed by integrating CNN-SE blocks, LSTM networks, and fully guided attention modules. This combination effectively modeled spatial-temporal dependencies while maintaining stable convergence during training.

2. Experimental comparisons demonstrated that MM-FGAG outperformed baseline models (CNN, LSTM, CNN-LSTM), achieving 91.72% accuracy and 0.945 AUC. These results confirm the model's robustness and superior ability to learn multimodal flood patterns.

3. The MM-FGAG model provided clear interpretability through attention-map visualization, which correctly highlighted flood-prone spatial regions and key temporal intervals. This transparency supports its use for explainable AI applications in flood-risk assessment.

4. The generated flood-risk maps for 2025 closely matched observed events, confirming the framework's practical utility for regional flood early-warning systems. High spatial agreement (IoU = 0.87, Dice = 0.91, MAE = 4.8%) demonstrates that MM-FGAG can serve as a reliable decision-support tool for disaster-management authorities.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this study, whether financial, personal, authorship or otherwise, that could affect the study and its results presented in this paper.

Financing

This study is funded by the Ministry of Higher Education, Science, and Technology (Kemdiktisaintek) of the Republic of Indonesia, Directorate General of Research and Development.

Data availability

Data cannot be made available for reasons disclosed in the data availability statement.

Use of artificial intelligence

The author has used artificial intelligence technology within acceptable limits to maximize the vocabulary of the writing.

Acknowledgments

We would like to express our deepest gratitude to the Ministry of Higher Education, Science, and Technology (Kemdiktisaintek) of the Republic of Indonesia, Directorate General of Research and Development, which has funded this research through the Master's Thesis Postgraduate Research (PPTM) scheme for the 2025 fiscal year. With the Master Contract Number: 122/C3/DT.05.00/PL/2025, dated June 11, 2025. And the Derivative Contract Number: 44/SPK/LL1/AL.04.03/PL/2025.

References

1. Gosset, M., Dibi-Anoh, P. A., Schumann, G., Hostache, R., Paris, A., Zahiri, E.-P. et al. (2023). Hydrometeorological Extreme Events in Africa: The Role of Satellite Observations for Monitoring Pluvial and Fluvial Flood Risk. Surveys in Geophysics, 44 (1), 197–223. https://doi.org/10.1007/s10712-022-09749-6
2. Hermawan, E., Risyanto, R., Purwaningsih, A., Ratri, D. N., Ridho, A., Harjana, T. et al. (2024). Characteristics of Mesoscale Convective Systems and Their Impact on Heavy Rainfall in Indonesia's New Capital City, Nusantara, in March 2022. Advances in Atmospheric Sciences, 42 (2), 342–356. https://doi.org/10.1007/s00376-024-4102-1
3. Taye, M. T., Dyer, E. (2024). Hydrologic Extremes in a Changing Climate: a Review of Extremes in East Africa. Current Climate Change Reports, 10 (1), 1–11. https://doi.org/10.1007/s40641-024-00193-9
4. Abegaz, R., Wang, F., Xu, J. (2024). History, causes, and trend of floods in the U.S.: a review. Natural Hazards, 120 (15), 13715–13755. https://doi.org/10.1007/s11069-024-06791-y
5. Waleed, M., Sajjad, M. (2023). On the emergence of geospatial cloud-based platforms for disaster risk management: A global scientometric review of google earth engine applications. International Journal of Disaster Risk Reduction, 97, 104056. https://doi.org/10.1016/j.ijdrr.2023.104056
6. Luk, S. Y., Sajjad, M. (2025). From space to screen: Recent advances in remote sensing for mangrove valuation through a bibliometric lens. Ocean & Coastal Management, 269, 107844. https://doi.org/10.1016/j.ocecoaman.2025.107844
7. Asiyabi, R. M., Ghorbanian, A., Tameh, S. N., Amani, M., Jin, S., Mohammadzadeh, A. (2023). Synthetic Aperture Radar (SAR) for Ocean: A Review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16, 9106–9138. https://doi.org/10.1109/jstars.2023.3310363
8. Du, R., Xiang, Y., Chen, J., Lu, X., Zhang, F., Zhang, Z. et al. (2024). The daily soil water content monitoring of cropland in irrigation area using Sentinel-2/3 spatio-temporal fusion and machine learning. International Journal of Applied Earth Observation and Geoinformation, 132, 104081. https://doi.org/10.1016/j.jag.2024.104081
9. Pareta, K. (2023). Hydrological modelling of largest braided river of India using MIKE Hydro River software with rainfall runoff, hydrodynamic and snowmelt modules. Journal of Water and Climate Change, 14 (4), 1314–1338. https://doi.org/10.2166/wcc.2023.484
10. Li, J., Wu, G., Zhang, Y., Shi, W. (2024). Optimizing flood predictions by integrating LSTM and physical-based models with mixed historical and simulated data. Heliyon, 10 (13), e33669. https://doi.org/10.1016/j.heliyon.2024.e33669
11. Ke, E., Zhao, J., Zhao, Y. (2025). Investigating the influence of nonlinear spatial heterogeneity in urban flooding factors using geographic explainable artificial intelligence. Journal of Hydrology, 648, 132398. https://doi.org/10.1016/j.jhydrol.2024.132398
12. Zhao, B., Sui, H., Liu, J., Shi, W., Wang, W., Xu, C., Wang, J. (2024). Flood inundation monitoring using multi-source satellite imagery: a knowledge transfer strategy for heterogeneous image change detection. Remote Sensing of Environment, 314, 114373. https://doi.org/10.1016/j.rse.2024.114373
13. Bagheri, A., Liu, G.-J. (2025). Climate change and urban flooding: assessing remote sensing data and flood modeling techniques: a comprehensive review. Environmental Reviews, 33, 1–14. https://doi.org/10.1139/er-2024-0065
14. Qatrinnada, W. F. P., Hidayah, E., Halik, G., Wiyono, R. U. A. (2024). A literature review: rainfall thresholds as flash flood monitoring for an early warning system. Water Practice & Technology, 19 (11), 4486–4498. https://doi.org/10.2166/wpt.2024.271
15. Huang, X., Chen, N., Deng, Z., Huang, S. (2024). Multivariate time series anomaly detection via dynamic graph attention network and Informer. Applied Intelligence, 54 (17-18), 7636–7658. https://doi.org/10.1007/s10489-024-05575-y
16. Maddah, S., Mojaradi, B., Alizadeh, H. (2024). Enhancing flood susceptibility modeling using integration of multi-source satellite imagery and multi-input convolutional neural network. Natural Hazards, 121 (3), 2801–2824. https://doi.org/10.1007/s11069-024-06764-1
17. Fu, Y., Zhu, Z., Liu, L., Zhan, W., He, T., Shen, H. et al. (2024). Remote Sensing Time Series Analysis: A Review of Data and Applications. Journal of Remote Sensing, 4. https://doi.org/10.34133/remotesensing.0285
18. Islam, K., Daraio, J. A., Cheema, M., Sabau, G., Galagedara, L. (2025). Improved streamflow prediction accuracy in Boreal climate watershed using a LSTM model: A comparative study. PLOS Water, 4 (4), e0000359. https://doi.org/10.1371/journal.pwat.0000359
19. He, X., Zhang, S., Xue, B., Zhao, T., Wu, T. (2023). Cross-modal change detection flood extraction based on convolutional neural network. International Journal of Applied Earth Observation and Geoinformation, 117, 103197. https://doi.org/10.1016/j.jag.2023.103197
20. Chen, Z., Lin, H., Shen, G. (2023). TreeLSTM: A spatiotemporal machine learning model for rainfall-runoff estimation. Journal of Hydrology: Regional Studies, 48, 101474. https://doi.org/10.1016/j.ejrh.2023.101474
21. Farahmand, H., Xu, Y., Mostafavi, A. (2023). A spatial–temporal graph deep learning model for urban flood nowcasting leveraging heterogeneous community features. Scientific Reports, 13 (1). https://doi.org/10.1038/s41598-023-32548-x
22. Li, J., Jin, C., Shen, Y., Ye, W. (2024). TRSANet: A Remote Sensing Deep Learning Model for Water Body Change Detection Based on Time-Reversal Semantic Asymmetry. IEEE Transactions on Geoscience and Remote Sensing, 62, 1–13. https://doi.org/10.1109/tgrs.2024.3458951
23. Oddo, P. C., Bolten, J. D., Kumar, S. V., Cleary, B. (2024). Deep Convolutional LSTM for improved flash flood prediction. Frontiers in Water, 6. https://doi.org/10.3389/frwa.2024.1346104
24. Wang, Z., Lyu, H., Fu, G., Zhang, C. (2024). Time-guided convolutional neural networks for spatiotemporal urban flood modelling. Journal of Hydrology, 645, 132250. https://doi.org/10.1016/j.jhydrol.2024.132250

25. Saleh, T., Holail, S., Zahran, M., Xiao, X., Xia, G.-S. (2024). LiST-Net: Enhanced Flood Mapping With Lightweight SAR Transformer Network and Dimension-Wise Attention. IEEE Transactions on Geoscience and Remote Sensing, 62, 1–17. https://doi.org/10.1109/tgrs.2024.3397797
26. Mondal, G., Dhanaraj, R. K. (2025). A novel multi-model fused classification for classifying the natural disaster. Progress in Artificial Intelligence. https://doi.org/10.1007/s13748-025-00388-7
27. Chen, C., Zhang, D., Qi, X., Wang, Z., Xiang, L. (2025). GKASA–DDPM: a novel flood forecasting model based on Graph Kolmogorov–Arnold Attention and spatio-temporal attention under smoothing DDPM. Journal of Hydroinformatics, 27 (3), 560–579. https://doi.org/10.2166/hydro.2025.312
28. Cui, Y., Duan, P., Li, J. (2025). PDSDC: Progressive Spatiotemporal Difference Capture Network for Remote Sensing Change Detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18, 16879–16895. https://doi.org/10.1109/jstars.2025.3569128
29. Zhang, Q., Zhang, X., Quan, C., Zhao, T., Huo, W., Huang, Y. (2025). Mamba-STFM: A Mamba-Based Spatiotemporal Fusion Method for Remote Sensing Images. Remote Sensing, 17 (13), 2135. https://doi.org/10.3390/rs17132135
30. Wanto, A., Yuhandri, Y., Okfalisa, O. (2024). RetMobileNet: A New Deep Learning Approach for Multi-Class Eye Disease Identification. Revue d'Intelligence Artificielle, 38 (4). https://doi.org/10.18280/ria.380401