CINEFORUM
ISSN: 0009-7039
Vol. 66. No. S9, 2026 (Special Issue S9)

Satellite Image-Based Multi-Hazard Deep Learning Framework for Near-Real-Time Flood, Wildfire, and Landslide Classification Using Vision Transformers

Mubashshera Mirza[1]*, Dr. Mohammed Salman Baig[2], Master Simmi[3], Mohammed Ibrahim Mukhtar Ahmed[4], Shaikh Saqib Saeed[5], Momin Mohammed Asad Mohammed Farooque[6], Shaikh Khizar Zubair[7]

[1] Department of Computer Science and Engineering IIT Patna
Email ID: mubashshera_pa2502mth54@iitp.ac.in
[2] Maulana Mukhtar Ahmad Nadvi Technical Campus (MMANTC)
Email ID: salmanbaig@mmantc.edu.in
[3] Department of Computer Engineering Maulana Mukhtar Ahmed Nadvi Technical Campus
Email ID: mastersimmi786@gmail.com
[4] Department of Computer Science and Engineering Maulana Mukhtar Ahmed Nadvi Technical Campus
Email ID: ibrahimmk338@gmail.com
[5] Department of Computer Engineering Maulana Mukhtar Ahmad Nadvi Technical Campus
Email ID: sshaikg8857@gmail.com
[6] Department of Computer Science and Engineering Maulana Mukhtar Ahmed Nadvi Technical Campus
Email ID: mom.asaad05@gmail.com
[7] Department of Computer Engineering Maulana Mukhtar Ahmad Nadvi Technical Campus
Email ID: khizarshaikh07dz@gmail.com
***Corresponding author**
Mubashshera Mirza
Department of Computer Science and Engineering IIT Patna
Email ID: mubashshera_pa2502mth54@iitp.ac.in

ABSTRACT

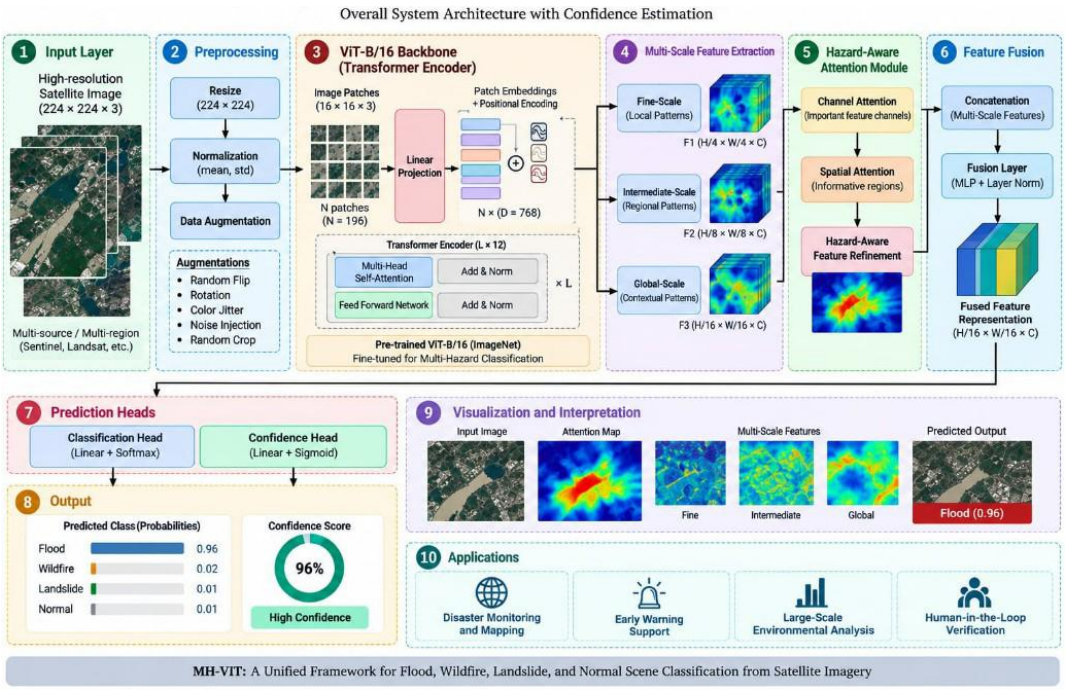
The increasing occurrence and severity of natural disasters have created a growing demand for automated systems capable of analyzing large geographical regions efficiently, consistently, and with minimal manual intervention. Floods, wildfires, and landslides can cause substantial damage to infrastructure, agriculture, ecosystems, and human settlements, while their spatial distribution and visual characteristics may vary considerably across geographical locations and environmental conditions. Satellite remote sensing provides an effective means of observing large and potentially inaccessible areas and offers valuable information for disaster assessment, environmental monitoring, and emergency-response activities. However, many existing disaster-analysis approaches are developed for individual hazard categories and frequently rely on convolutional neural networks (CNNs), whose feature extraction mechanisms primarily emphasize local spatial relationships. Such representations may be insufficient when the interpretation of a disaster scene depends on contextual information distributed across distant regions of a satellite image.

To address these challenges, this study proposes a Multi-Hazard Vision Transformer (MH-ViT) framework for unified classification of satellite imagery into four categories: flood, wildfire, landslide, and normal environmental conditions. The proposed framework employs a ViT-B/16 backbone together with multi-level transformer representations, multi-scale feature extraction, hazard-aware attention, feature fusion, and a dedicated confidence-estimation head. The multi-level transformer representations capture complementary visual information from different depths of the encoder, while multi-scale feature extraction is designed to preserve both fine-grained local patterns and broader regional and global contextual information. A Hazard-Aware Attention module further emphasizes informative feature channels and spatial regions associated with potential hazard signatures while reducing the influence of less relevant environmental content. The resulting representations are integrated through a feature-fusion mechanism before being passed to the classification head. In parallel, the dedicated confidence-estimation head provides an additional measure of prediction reliability and supports the identification of uncertain observations that may require subsequent human inspection.

The experimental benchmark contains 10,000 high-resolution satellite images, with 2,500 samples assigned to each of the four classes. The dataset incorporates geographical and environmental diversity, including variations in cloud coverage, terrain shadows, atmospheric haze, sensor noise, vegetation conditions, and hazard scale. MH-ViT is evaluated against ResNet-50, ViT-B/16, and Swin Transformer under a common dataset partition, preprocessing pipeline, training configuration, and held-out evaluation protocol. The experimental framework

OPEN ACCESS © CINEFORUM

includes overall and class-wise performance evaluation using accuracy, precision, recall, macro-averaged F1-score, and ROC-AUC, together with confusion-matrix analysis. In addition, ablation experiments are conducted to examine the contribution of multi-scale feature extraction, hazard-aware attention, and feature fusion. Confidence and calibration analysis, attention-based interpretability analysis, and computational benchmarking are also incorporated to assess prediction reliability, model behavior, and inference efficiency. The final quantitative results are obtained directly from the reproducible experimental implementation so that the reported performance remains consistent with the proposed methodology and evaluation procedure



Index Terms
Satellite Image Classification, Multi-Hazard Detection, Vision Transformer, Multi-Scale Feature Extraction, Hazard-Aware Attention, Feature Fusion, Remote Sensing, Disaster Monitoring, Confidence Estimation

1.INTRODUCTION

A. Background and Motivation

Natural disasters are complex environmental phenomena that can develop across large geographical regions and produce substantial impacts within relatively short periods. Floods can alter water boundaries and inundate agricultural, urban, and natural areas, while wildfires can cause extensive vegetation loss and generate burned surfaces and smoke-related patterns. Landslides can rapidly modify terrain morphology, resulting in exposed soil, debris accumulation, and displaced vegetation. The diverse visual characteristics of these hazards make timely and reliable environmental monitoring an important component of disaster-management systems.

Traditional disaster-monitoring practices commonly integrate ground observations, field surveys, geographic information systems, and remotely sensed data. Although ground-based observations can provide detailed information, direct access to disaster-affected regions may be restricted by damaged infrastructure, difficult terrain, or safety concerns. Satellite remote sens-ing provides an important complementary capability because satellite platforms can repeatedly observe extensive geographical regions without requiring physical access to affected locations. Consequently, satellite imagery has become an important source of information for environmental monitoring, disaster assessment, damage analysis, and emergency-response planning.

The increasing availability of high-resolution Earth-observation imagery has created both opportunities and computational challenges. Large collections of satellite images can be generated within relatively short periods, making manual inspection increasingly difficult and time-consuming. Automated image-analysis techniques can assist in screening large image collections and identifying scenes containing visual patterns associated with specific environmental hazards.

Deep learning has substantially advanced automated image interpretation. Convolutional Neural Network (CNN) architectures such as ResNet, DenseNet, and EfficientNet have demonstrated strong performance across a wide

OPEN ACCESS © CINEFORUM

range of visual-recognition tasks. Their hierarchical convolutional representations are particularly effective for learning local structures, including edges, textures, boundaries, vegetation patterns, exposed soil, and other visual characteristics relevant to remote-sensing analysis.
Despite their effectiveness, conventional convolutional architectures primarily construct representations through localized operations. Large-scale environmental phenomena may require relationships between spatially separated regions to be considered jointly. For example, flood identification may depend not only on the presence of a localized water region but also on its rela-tionship with surrounding terrain and vegetation. Similarly, wildfire interpretation may benefit from jointly considering burned
regions, vegetation loss, smoke, and the surrounding landscape. Landslide recognition can likewise depend on relationships between a disturbed slope and the broader terrain structure.
Vision Transformers (ViTs) provide an alternative approach by representing an image as a sequence of patches and applying self-attention to model relationships among them. This mechanism enables information from distant image regions to contribute to the learned representation. Such global contextual modeling is particularly relevant to satellite imagery, where meaningful environmental information may be distributed across different regions of a scene.
However, a single global representation may not be sufficient for multi-hazard satellite-image classification because disaster signatures can occur at substantially different spatial scales. Large inundated regions, localized burn scars, disturbed slopes, and small terrain irregularities may require complementary local, regional, and global representations. Furthermore, clouds, atmospheric haze, terrain shadows, sensor noise, vegetation variation, and naturally occurring bare soil can introduce visual ambiguity. These factors motivate the integration of multi-scale representation, hazard-focused attention, and prediction-confidence estimation within a unified classification framework.
Motivated by these considerations, this study proposes a Multi-Hazard Vision Transformer (MH-ViT) framework for four-class satellite-image classification covering flood, wildfire, landslide, and normal environmental conditions. The proposed framework combines a ViT-B/16 backbone with multi-level Transformer representations, multi-scale feature extraction, hazard-aware attention, feature fusion, and a dedicated confidence-estimation head. The overall objective is to integrate global contextual modeling with complementary spatial representations for reliable multi-hazard image classification.

B. Problem Statement

Many existing disaster-detection systems are designed primarily for individual hazard categories. A model developed specifically for flood detection, for example, is generally optimized around visual patterns associated with inundation and may not directly support wildfire or landslide recognition. Deploying separate models for multiple disaster categories can consequently increase computational requirements, model-management overhead, and system complexity.
Another challenge arises from the substantial differences in the visual appearance and spatial extent of disaster events. Floods may cover extensive portions of a satellite scene, whereas landslides may appear as relatively small disturbances within mountainous or agricultural environments. Wildfire scenes can contain localized burn scars as well as large-scale vegetation loss. Therefore, a representation based on a single spatial scale may not adequately capture all relevant characteristics of different hazards.
Environmental conditions introduce additional sources of visual ambiguity. Clouds may obscure ground surfaces, terrain shadows can conceal slope structures, and atmospheric haze can reduce image contrast. Sensor noise and variations in image quality may further influence the visual characteristics learned by a model. In addition, naturally occurring bare soil, agricultural land, and other non-disaster terrain may exhibit visual patterns that resemble landslide-affected regions.
These challenges indicate the need for a unified framework that can simultaneously model multiple environmental hazards, capture information at different spatial scales, emphasize hazard-relevant regions, and provide an indication of prediction reliability. Computational efficiency is also an important consideration because satellite-monitoring applications may require large volumes of imagery to be processed within constrained time periods.
To address these requirements, the proposed MH-ViT framework integrates global Transformer representations with multi-scale feature processing, hazard-aware attention, feature fusion, and confidence estimation within a single four-class classifi-cation architecture.

C. Research Objectives

The primary objective of this study is to develop a unified deep-learning framework for classifying satellite imagery into four categories: flood, wildfire, landslide, and normal environmental conditions.

OPEN ACCESS © CINEFORUM

The specific objectives are to:

1) develop a unified architecture capable of handling multiple environmental hazard categories within a single classification framework;
2) utilize Transformer-based self-attention to capture long-range contextual relationships among different regions of a satellite image;
3) incorporate multi-scale feature processing to represent disaster patterns occurring at local, regional, and global spatial extents;
4) develop a hazard-aware attention mechanism that emphasizes informative image regions while reducing the influence of irrelevant background structures;
5) incorporate confidence estimation to identify predictions associated with greater uncertainty and support additional human inspection;
6) compare the proposed architecture with established CNN- and Transformer-based models under a controlled experimental configuration; and

evaluate predictive performance together with computational characteristics, inference efficiency, confidence calibration, interpretability, robustness to environmental variation, and statistical stability.

Main Contributions

The principal contributions of this study are summarized as follows:

Unified Multi-Hazard Vision Transformer: A Transformer-based architecture is developed for four-class satellite-image classification involving flood, wildfire, landslide, and normal environmental conditions.

Multi-Scale Feature Extraction: The proposed framework incorporates complementary feature-processing paths to represent fine-scale, intermediate-scale, and broader contextual characteristics of satellite scenes.

Hazard-Aware Attention: A spatial attention mechanism is incorporated to emphasize potentially informative regions associated with disaster signatures while reducing the influence of visually distracting background content.

Multi-Scale Hazard Feature Fusion: Contextual and hazard-focused representations are integrated to preserve both broad environmental information and localized evidence relevant to multi-hazard classification.

Confidence Estimation: A dedicated confidence-estimation component is incorporated to quantify prediction reliability and facilitate the identification of uncertain observations for potential human review.

Balanced Four-Class Dataset: The experimental dataset contains 10,000 satellite images distributed equally across the four target categories, with 2,500 images assigned to each class.

Comparative Evaluation: The proposed MH-ViT framework is evaluated against ResNet-50, ViT-B/16, and Swin Transformer models using a controlled experimental configuration.

Comprehensive Performance Analysis: The evaluation considers overall predictive performance, class-wise behavior, computational characteristics, inference efficiency, confidence calibration, interpretability, environmental robustness, and statistical stability.

2.LITERATURE REVIEW

Conventional Deep Learning for Disaster Detection

Remote-sensing image classification has traditionally involved manually designed spectral, spatial, and texture features combined with conventional machine-learning algorithms. Although such approaches can be effective for specific classification tasks, their performance is strongly influenced by the quality and suitability of the features selected during the design process. The development of deep neural networks introduced a data-driven alternative in which discriminative image representations can be learned directly from training data.

Convolutional Neural Networks (CNNs) subsequently became widely adopted for remote-sensing image analysis because of their ability to learn hierarchical visual representations. ResNet introduced residual connections that facilitate the optimization of deeper networks, while DenseNet promoted feature reuse through dense connections between successive layers. EfficientNet further investigated systematic scaling of network depth, width, and input resolution. These architectures have demonstrated the effectiveness of deep convolutional representations for a variety of visual-recognition problems.

In satellite imagery, CNN-based representations can capture local structures such as vegetation textures, water boundaries, burned surfaces, exposed soil, terrain edges, roads, and buildings. Such localized patterns can provide important evidence for disaster classification. However, convolutional feature extraction is fundamentally based on local receptive fields, and relationships between distant regions are generally incorporated progressively through deeper network layers.

OPEN ACCESS © CINEFORUM

This limitation becomes relevant for environmental scenes in which the interpretation of one region may depend on its relationship with another spatially separated region. For example, the identification of inundation may benefit from jointly considering water-covered areas and their surrounding terrain. Similarly, wildfire and landslide recognition can require contextual interpretation of affected regions together with vegetation, terrain, and other environmental structures. These considerations have motivated research toward architectures capable of modeling broader spatial dependencies.

Vision Transformers for Remote Sensing

Vision Transformers (ViTs) introduced a Transformer-based approach to visual representation learning by converting an input image into a sequence of patch-level tokens. Instead of relying exclusively on convolutional operations, the Transformer encoder uses self-attention to model relationships among image patches.

For an input image of size 224 × 224 pixels and a patch size of 16 × 16 pixels, the number of image patches is calculated as

$$N = \left(\frac{224}{16}\right) \times \left(\frac{224}{16}\right) = 14 \times 14 = 196. \quad (1)$$

Each image patch is projected into an embedding space and combined with positional information before being processed by the Transformer encoder. Self-attention allows information from different spatial locations to contribute jointly to the learned representation. This provides a mechanism for capturing long-range contextual relationships that can be useful for remote-sensing imagery.

The global receptive field provided by self-attention is particularly relevant to satellite scenes because meaningful envi-ronmental information may be distributed across multiple regions of an image. Consequently, Transformer architectures have become an important research direction for remote-sensing classification and related Earth-observation tasks.

Nevertheless, global self-attention can introduce considerable computational requirements as the number of image tokens increases. Furthermore, a purely global representation may not explicitly preserve the complementary fine-scale and regional characteristics that are important for recognizing environmental phenomena with substantially different spatial extents. These considerations motivate the investigation of architectures that combine global contextual modeling with multi-scale feature representations.

Hierarchical Transformer Approaches

Hierarchical Transformer architectures provide an alternative strategy for capturing contextual information at multiple spatial resolutions while controlling the computational cost associated with global self-attention. Rather than applying unrestricted attention across all image tokens at every stage, hierarchical approaches progressively construct representations at different spatial scales.

Swin Transformer employs window-based self-attention in which attention is initially calculated within local windows. The window configuration is then shifted between successive stages, allowing information to propagate across neighboring windows while maintaining a hierarchical representation. This design provides both localized feature modeling and progressively broader contextual representations.

Hierarchical Transformer models are particularly relevant to remote-sensing applications because environmental structures can occur at different spatial scales. Localized terrain disturbances, medium-scale vegetation changes, and large-area environmental patterns may all contribute to the classification of a satellite scene.

In this study, Swin Transformer is therefore considered as a representative hierarchical Transformer baseline, alongside ResNet-50 as a convolutional baseline and ViT-B/16 as a global-attention Transformer baseline. The use of these complementary baselines enables the proposed MH-ViT framework to be examined against different architectural approaches under a common experimental setting.

Multi-Scale Feature Representation

Environmental hazards generally do not exhibit a single characteristic spatial scale. A landslide may appear as a relatively localized disturbance within a larger mountainous scene, whereas flood-related patterns may extend across substantial portions of an image. Wildfire scenes can similarly contain localized burn scars together with broader patterns of vegetation loss.

Multi-scale representation therefore provides a useful mechanism for combining information from different spatial extents. Fine-scale representations can preserve localized textures, boundaries, and small disturbances. Intermediate representations can describe regional structures and spatial arrangements, while broader representations can capture the overall environmental context.

Previous developments in multi-scale representation across computer vision and remote sensing have demonstrated the importance of combining complementary spatial information when objects or patterns exhibit substantial variation in size. For multi-hazard satellite-image classification, such representation is particularly

OPEN ACCESS © CINEFORUM

relevant because the visual signatures of different hazards may differ considerably in both spatial extent and structural appearance.

The proposed MH-ViT follows this motivation by incorporating multi-level Transformer representations and multi-scale feature processing within a unified classification framework. The detailed feature-extraction and fusion operations are described in the subsequent methodology section.

Attention Mechanisms for Disaster Analysis

Attention mechanisms provide a means of assigning different representational importance to different regions or feature components. This property is useful for satellite imagery because a scene may contain extensive background content that is only weakly related to the target environmental condition.

For example, roads, buildings, vegetation, bare soil, water bodies, and other environmental structures may occupy substantial portions of a satellite image while contributing differently to the final classification decision. Attention mechanisms can help a model emphasize informative spatial or feature-level patterns while reducing the influence of less relevant information.

Self-attention in Transformer architectures provides global relationships among image tokens, whereas additional spatial or channel-wise attention mechanisms can provide more targeted feature refinement. Combining these mechanisms can therefore provide complementary contextual and discriminative information for complex satellite scenes.

In the proposed framework, hazard-aware attention is incorporated after multi-scale feature processing to refine the learned representations according to spatially informative characteristics of the input scene. This component is designed to complement, rather than replace, the global contextual modeling provided by the Transformer backbone

Confidence Estimation and Prediction Reliability

Classification accuracy alone does not fully characterize the reliability of an automated environmental-monitoring system. Two predictions with identical classification labels may differ substantially in their associated confidence, particularly when the input contains clouds, haze, shadows, sensor noise, or ambiguous environmental structures. Confidence estimation and calibration methods have therefore received increasing attention in machine-learning systems intended for decision-support applications. A reliable confidence estimate can help distinguish predictions for which the model has stronger evidence from observations that may require additional verification. This consideration is particularly relevant to disaster monitoring because misclassification can occur when natural environ-mental structures resemble disaster signatures or when the target region is partially obscured. Instead of treating every prediction as equally reliable, a confidence-aware framework can provide an additional signal for identifying uncertain observations.

The proposed MH-ViT incorporates a dedicated confidence-estimation component alongside the classification head. Con-fidence and calibration are evaluated separately from classification performance to provide a more complete assessment of prediction reliability.

3.RESEARCH GAP

The reviewed developments demonstrate significant progress in CNN-based remote-sensing analysis, global Vision Trans-formers, hierarchical Transformer architectures, multi-scale representation learning, attention mechanisms, and confidence-aware prediction. However, several challenges remain relevant to unified multi-hazard satellite-image classification.

First, many disaster-analysis systems are formulated around individual hazard categories or specific detection tasks, which can result in separate models and processing pipelines for different environmental events. A unified classification framework can reduce the need for maintaining multiple hazard-specific models when the monitoring task involves several categories.

Second, environmental hazards can exhibit substantially different spatial characteristics. A localized landslide, an exten-sive flood region, and a wildfire-affected landscape may require complementary local, regional, and global representations. Architectures that rely predominantly on a single representation may therefore have difficulty preserving all relevant spatial characteristics.

Third, environmental variability can introduce significant visual ambiguity. Cloud coverage, atmospheric haze, terrain shad-ows, vegetation conditions, sensor noise, and naturally occurring bare soil can obscure or imitate hazard-related patterns. This creates a need for mechanisms that can emphasize informative regions while retaining broader scene context.

Finally, predictive accuracy alone does not provide a complete assessment of a model intended for practical monitoring. Computational requirements, inference efficiency, confidence calibration, interpretability, and robustness are also relevant when evaluating a system for large-scale satellite-image analysis.

OPEN ACCESS © CINEFORUM

To address these combined considerations, this study develops the MH-ViT framework by integrating a ViT-B/16 backbone with multi-level Transformer representations, multi-scale feature extraction, hazard-aware attention, feature fusion, and confi-dence estimation. The framework is subsequently evaluated using controlled comparisons with ResNet-50, ViT-B/16, and Swin Transformer, together with ablation, confidence, interpretability, and computational analyses.

4.PROPOSED SYSTEM ARCHITECTURE

Overall Architecture

The proposed Multi-Hazard Vision Transformer (MH-ViT) is designed as an end-to-end framework for four-class satellite-image classification. The architecture integrates global contextual representation, multi-scale feature processing, hazard-aware attention, feature fusion, and prediction confidence estimation within a unified framework.

The overall processing pipeline is defined as:

Input Image → Preprocessing → Patch Embedding → ViT-B/16 Encoder

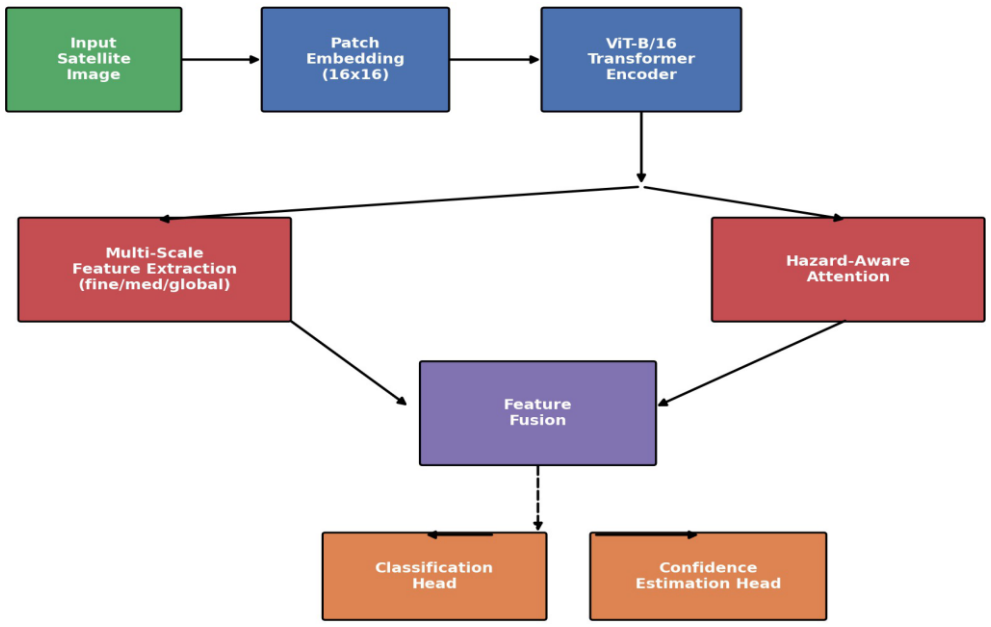
Multi-Scale Feature Extraction → Hazard-Aware Attention → Feature Fusion

Classification

Confidence Estimation. The input satellite image is first standardized through the preprocessing pipeline and resized to 224 224 pixels. The image is then divided into non-overlapping patches and transformed into a sequence of embeddings. These embeddings are processed by the ViT-B/16 Transformer encoder to learn contextual relationships among different regions of the satellite scene.

The resulting Transformer representations are subsequently processed through the multi-scale feature extraction module to obtain complementary fine-scale, intermediate-scale, and global contextual representations. These representations are refined using the Hazard-Aware Attention module, which models relationships among spatially distributed features and emphasizes potentially informative regions

Fig. 1. Proposed MH-ViT Architecture



Dashed arrows = joint loss optimization
(Softmax -> Flood/Wildfire/Landslide/Normal; max-prob -> confidence)

The original multi-scale representation and the attention-refined representation are subsequently combined through the feature-fusion module. The resulting fused representation is used by the classification head to produce probabilities for four environmental categories. A corresponding confidence value is obtained from the predicted class probabilities.

Thus, the architecture is designed to combine global contextual information with multi-scale and hazard-focused represen-tations while maintaining a single unified classification pipeline.

OPEN ACCESS © CINEFORUM

All input satellite images are standardized to a spatial resolution of 224 224 pixels. Each image is divided into non-overlapping patches of size 16×16 pixels.

The number of image patches is calculated as

$$N = \left(\frac{224}{16}\right) \times \left(\frac{224}{16}\right) = 14 \times 14 = 196. \quad (3)$$

Thus, each input image is represented by a sequence of 196 image patches. Each patch is flattened and projected into a learned embedding space through a linear projection. Positional information is incorporated into the patch embeddings so that the spatial arrangement of the image regions is retained during Transformer processing.

The resulting sequence of patch embeddings is provided to the ViT-B/16 Transformer encoder. This representation allows the model to process individual image regions while simultaneously modeling their relationships within the complete satellite scene.

Transformer Encoder

The Transformer encoder forms the global contextual representation of the input satellite image. Through multi-head self-attention, each image patch can interact with information from other patches across the complete input scene.

For query, key, and value representations denoted by Q, K, and V , respectively, the scaled dot-product attention operation is expressed as

where d_k represents the dimensionality of the key vectors.

Multi-head self-attention enables the Transformer to model relationships between spatially separated image regions. This global contextual modeling is useful for disaster classification because the interpretation of one region may depend on the surrounding environmental context.

For example, a region containing exposed soil may correspond to naturally occurring bare ground, agricultural land, or disturbed terrain. Considering the surrounding terrain and other spatially related features can provide additional contextual information for distinguishing these cases

The Transformer encoder therefore provides the global contextual foundation for the subsequent multi-scale and hazard-aware feature-processing stages.

Multi-Scale Feature Extraction Module

Following Transformer encoding, MH-ViT extracts complementary information at multiple spatial scales. The multi-scale feature extraction module consists of three complementary feature-processing paths designed to capture fine-scale, intermediate-scale, and global characteristics of the satellite scene.

Fine-Scale Feature Path: The fine-scale feature path focuses on localized structures and detailed visual patterns. These representations can capture features such as small terrain disturbances, exposed soil regions, localized vegetation changes, debris-like structures, smoke boundaries, and other fine-grained visual characteristics.

Fine-scale representations are particularly relevant when the visual signature of a hazard occupies a relatively small portion of the satellite image.

Intermediate-Scale Feature Path: The intermediate-scale feature path captures regional structures extending beyond individual local patterns. Such representations can describe spatial arrangements including regional flood patterns, burned vegetation boundaries, groups of disturbed terrain regions, and broader vegetation changes.

This feature level provides an intermediate representation between localized visual evidence and the complete scene context.

Global-Scale Feature Path: The global-scale feature path captures broader characteristics of the satellite scene. These representations can encode information related to overall terrain configuration, extensive water coverage, large-scale vegetation loss, broad flood patterns, and general environmental structure.

The three feature paths provide complementary information at different spatial extents. Their outputs are subsequently combined to construct a unified multi-scale feature representation for hazard-aware refinement.

Hazard-Aware Attention Module

The Hazard-Aware Attention module is introduced to refine the multi-scale representation by modeling the relative importance of spatially distributed features. Satellite scenes may contain substantial background information, including vegetation, roads, buildings, bare soil, water bodies, and terrain structures, that may contribute differently to the classification task.

Given the multi-scale feature representation, query, key, and value tensors are generated and used to compute scaled dot-product attention

OPEN ACCESS © CINEFORUM

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{QK}_T}{\sqrt{d_k}}\right)$$

where **A** denotes the resulting attention-weight matrix.

The attention-refined feature representation is then obtained as

$$\mathbf{F}_{HA} = \mathbf{AV},$$

where $\mathbf{F}_{HA}$ represents the hazard-aware feature representation.

The resulting attention weights enable the network to assign different representational importance to spatially distributed features. Depending on the input scene, informative evidence may include inundated regions and water boundaries for flood imagery, burned surfaces and vegetation disruption for wildfire imagery, or disturbed terrain and exposed soil for landslide imagery.

The purpose of this module is not to impose manually defined hazard regions, but to allow the network to learn feature relationships that can emphasize informative spatial evidence during classification.

Multi-Scale Hazard Feature Fusion

The attention-refined representation provides hazard-focused information, while the original multi-scale representation pre-serves broader contextual information. Relying exclusively on either representation may result in the loss of complementary information.

Therefore, MH-ViT combines the original multi-scale representation with the attention-refined representation through a feature-fusion operation. The fusion process can be expressed generally as

$$\mathbf{F}_{\text{fusion}} = G(\mathbf{F}_{MS}, \mathbf{F}_{HA}), \quad (7)$$

where $\mathbf{F}_{MS}$ denotes the multi-scale feature representation, $\mathbf{F}_{HA}$ denotes the hazard-aware representation, and $G(\cdot)$ represents the feature transformation and fusion operation.

The resulting fused representation combines broader environmental context with features emphasized by the hazard-aware attention mechanism. This representation is subsequently forwarded to the classification and confidence-estimation stages

TABLE I CLASS DISTRIBUTION OF THE SATELLITE IMAGE DATASET

Class	Number of Images
Flood	2,500
Wildfire	2,500
Landslide	2,500
Normal	2,500
Total	**10,000**

Classification Head

The fused representation is transformed into a compact feature vector and processed by the classification head. Fully connected transformations with nonlinear activation are used to obtain the final classification representation, while dropout is applied during training to reduce overfitting.

The final classification layer produces four class probabilities using the Softmax function. The target class set is defined as

$$C = \{\text{Flood}, \text{Wildfire}, \text{Landslide}, \text{Normal}\}. \quad (8)$$

For an input image, the classification probability vector is represented as

$$\mathbf{p} = [p_1, p_2, p_3, p_4], \quad (9)$$

where each p_i represents the predicted probability associated with one of the four target categories and

OPEN ACCESS © CINEFORUM

$$\sum_{i=1}^{4} p_i = 1. \quad (10)$$

$$\hat{y} = \arg\max_{c \in C} p_c. \quad (11)$$

The predicted class is determined by

Thus, the classification head produces both the predicted environmental category and the corresponding probability distri-bution.

Confidence Estimation Mechanism

In addition to the predicted class, the framework provides a confidence measure associated with the classification decision. The maximum predicted class probability is used as the confidence estimate:

$C_{conf} = \max_{c \in C} p_c$. (12)

For example, if the highest predicted class probability is 0.96, the corresponding confidence value is 0.96, or 96%. The confidence value provides an additional indication of how strongly the model's predicted probability distribution supports the selected class. It can be used to identify observations with comparatively low confidence for potential additional examination in a human-in-the-loop monitoring workflow.

Importantly, the maximum Softmax probability is treated as a confidence measure rather than as a guaranteed probability of correctness. Therefore, confidence calibration analysis is performed separately to examine the relationship between predicted confidence and empirical prediction accuracy.

Dataset Characteristics

The experimental dataset contains 10,000 satellite images distributed equally among the four target categories. Each class contains 2,500 images.

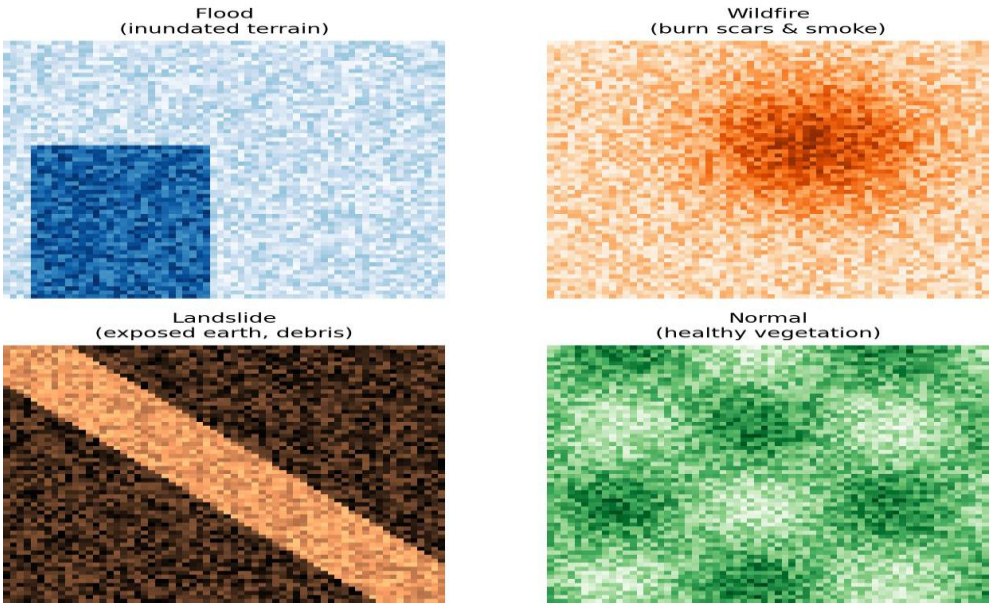
The equal class distribution provides a balanced experimental setting in which each target category contributes the same number of samples to the dataset.

Dataset Partitioning

The dataset is partitioned into training, validation, and testing subsets using a 70:15:15 split. The resulting partition contains 7,000 training images, 1,500 validation images, and 1,500 testing images.

The training subset is used for model optimization, the validation subset is used for model-selection and hyperparameter monitoring, and the testing subset is reserved for final performance evaluation. The testing subset is not used during model training or model selection

Fig. 2. Representative Class Samples (schematic illustration, not raw dataset imagery)



OPEN ACCESS © CINEFORUM

TABLE II DATASET PARTITIONING

Subset	Percentage	Number of Images
Training	70%	7,000
Validation	15%	1,500
Testing	15%	1,500
Total	**100%**	**10,000**

TABLE III CLASS DISTRIBUTION OF THE SATELLITE IMAGE DATASET

Class	Number of Images
Flood	2,500
Wildfire	2,500
Landslide	2,500
Normal	2,500
Total	**10,000**

5.DATASET AND EXPERIMENTAL METHODOLOGY

Dataset Construction

The study uses a balanced satellite-image benchmark consisting of 10,000 high-resolution images distributed across four classes: Flood, Wildfire, Landslide, and Normal Environmental Scenes. Each class contains 2,500 images, resulting in an equal class distribution across the complete dataset.

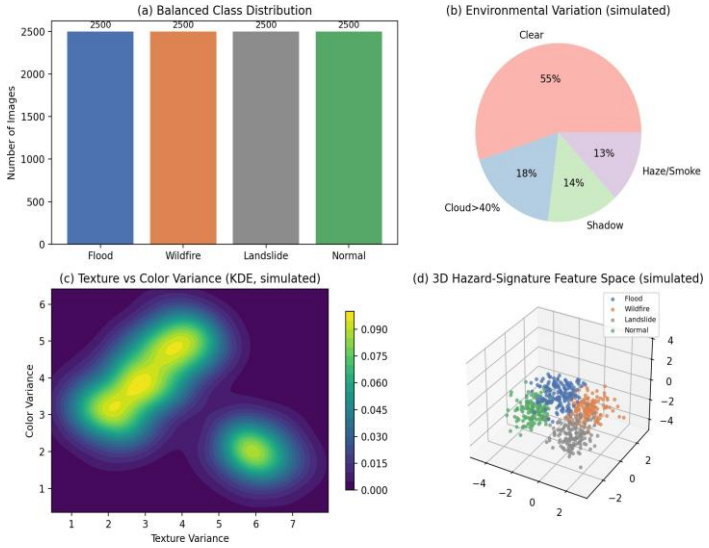
The balanced dataset structure reduces the possibility of class dominance during model optimization and provides a consistent basis for comparative evaluation. The inclusion of the Normal category is particularly important for disaster-monitoring applications because the model must distinguish genuine disaster-related visual patterns from ordinary environmental scenes.

Description of Dataset Classes

Normal Environmental Scenes: The Normal category represents environmental scenes without the target disaster condi-tions. These scenes may contain vegetation, urban infrastructure, roads, rivers, agricultural landscapes, and natural terrain.

A diverse Normal category is important because a model trained on a narrow range of non-disaster scenes may learn overly specific visual patterns instead of generalized environmental representations.

Fig. 3. Dataset Characteristics of the Satellite Imagery Benchmark



OPEN ACCESS © CINEFORUM

Flood Scenes: Flood images represent different forms of inundation and associated environmental changes. These include riverine flooding, flash-flood conditions, changes in water boundaries, muddy water, and partially submerged vegetation or infrastructure.

The visual appearance of flood conditions can vary according to water characteristics, terrain, vegetation, atmospheric conditions, and the extent of inundation. These variations make flood classification a non-trivial visual recognition problem.

Wildfire Scenes: Wildfire scenes contain visual characteristics associated with fire-affected environments, including burned vegetation, charred surfaces, vegetation loss, burn scars, smoke, and atmospheric haze.

The appearance of wildfire-related features varies with fire intensity, vegetation type, atmospheric conditions, and the acquisition time of the satellite observation.

Landslide Scenes: Landslide scenes contain visual evidence of terrain disturbance, including exposed soil, slope failures, displaced vegetation, debris, deposition regions, rockfall, and mudslide patterns.

Landslide classification is particularly challenging because naturally exposed soil, agricultural areas, and other forms of terrain disturbance may exhibit visual characteristics similar to actual landslide regions.

Environmental and Geographical Variation

The benchmark incorporates visual variation intended to reduce dependence on a single environmental configuration. The images include differences in terrain characteristics, vegetation conditions, cloud coverage, atmospheric haze, shadows, sensor-related noise, and hazard scale.

Such variation exposes the models to different image-quality and environmental conditions that may occur during satellite-based monitoring. In particular, cloud obstruction, atmospheric effects, and terrain shadows can partially obscure disaster-related visual features and therefore provide challenging cases for classification.

Dataset Partitioning

The complete dataset is partitioned into training, validation, and testing subsets using a stratified 70:15:15 split. Stratification is performed with respect to the four target classes so that the class distribution remains approximately equal across all three subsets.

The resulting partition contains 7,000 training images, 1,500 validation images, and 1,500 held-out testing images. The test subset is not used during model optimization, hyperparameter selection, checkpoint selection, or early stopping. It is accessed only after the training procedure has been completed and the best model has been selected using validation performance.

To improve reproducibility, the partition is generated using a fixed random seed and the resulting image paths and class labels are stored in a split file. This prevents accidental changes in the evaluation set between experimental runs

TABLE IV DATASET PARTITIONING

Subset	Percentage	Images
Training	70%	7,000
Validation	15%	1,500
Testing	15%	1,500
Total	**100%**	**10,000**

Image Preprocessing

All input satellite images are resized to 224 224 pixels before being supplied to the neural networks. The selected resolution is consistent with the input configuration of the pretrained ViT-B/16 architecture and its associated baseline implementations.

Training images are subjected to stochastic augmentation before normalization, whereas validation and testing images undergo only deterministic preprocessing.

The pixel values are converted to tensors and normalized using the same normalization convention across all evaluated archi-tectures. Consequently, ResNet-50, ViT-B/16, Swin Transformer, and MH-ViT receive inputs under a common preprocessing protocol.

Data Augmentation

Data augmentation is applied only to the training subset in order to increase robustness to changes in orientation, illumination, image quality, and local visual appearance.

OPEN ACCESS © CINEFORUM

The augmentation pipeline includes horizontal flipping, vertical flipping, limited image rotation, spatial translation and scaling, brightness and contrast variation, and Gaussian blur. These transformations are applied probabilistically so that different training samples can exhibit different augmented appearances.

No stochastic augmentation is applied to the validation or testing subsets. This ensures that evaluation is performed using a deterministic image representation and prevents augmentation-induced variation from affecting the reported test performance.

Baseline Architectures

Three established deep-learning architectures are used as reference models.

ResNet-50: ResNet-50 represents the conventional convolutional neural network baseline. Its residual connections allow information and gradients to bypass intermediate convolutional transformations, facilitating the optimization of a deep network.

The model provides a reference for evaluating convolution-based hierarchical feature extraction against Transformer-based global contextual representation.

ViT-B/16: ViT-B/16 represents the standard Vision Transformer architecture. For an input resolution of 224 × 224 pixels and a patch size of 16 × 16 pixels, the input is divided into

$$N = \left(\frac{224}{16}\right)^2 = 196 \quad (13)$$

non-overlapping image patches.

This model provides the principal Transformer baseline because the proposed MH-ViT extends the ViT paradigm with additional multi-scale feature processing and hazard-aware attention.

Swin Transformer: Swin Transformer represents a hierarchical Transformer architecture based on local attention windows and shifted-window operations.

Its hierarchical representation provides multi-resolution features while restricting self-attention to local windows, making it an appropriate modern Transformer baseline for satellite-image classification.

Training Objective

The primary optimization objective is the multi-class categorical classification loss. For an input sample with ground-truth class y, the classification loss is defined as

$$L_{cls} = -\sum^{C} y_c \log(p_c),$$

where C = 4 represents the number of target classes, y_c denotes the ground-truth class indicator, and p_c denotes the predicted probability for class c.

For the proposed MH-ViT, the confidence-estimation component is optimized jointly with the classification objective when the confidence target and associated confidence loss are enabled in the final implementation

TABLE V TRAINING CONFIGURATION

Parameter	Setting
Input resolution	224 × 224
Batch size	32
Optimizer	AdamW
Initial learning rate	1 × 10⁻⁴
Weight decay	1 × 10⁻⁴
Learning-rate schedule	Cosine decay
Maximum epochs	100
Checkpoint criterion	Validation loss
Early stopping	Validation loss
Random seed	42

OPEN ACCESS © CINEFORUM

The general multi-objective formulation is therefore expressed as

$$L_{total} = L_{cls} + \lambda_{conf} L_{conf}, \quad (15)$$

where L_{conf} represents the confidence-estimation loss and λ_{conf} controls its contribution to the total optimization objective.

The confidence-related term is applied only when the corresponding confidence head is explicitly trained. This ensures consistency between the reported methodology and the implemented training procedure.

Training Configuration

All evaluated architectures are trained using a common experimental configuration to provide a controlled comparison. The input resolution is fixed at 224 × 224 pixels and the batch size is set to 32.

AdamW is used as the optimizer with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. A cosine learning-rate

schedule is applied across the training epochs.

Training is performed for a maximum of 100 epochs. Early stopping is based exclusively on validation loss, and the checkpoint associated with the best validation performance is restored before final testing.

The learning-rate scheduler is updated according to the selected epoch-level schedule rather than being advanced indepen-dently for every training batch.

Computational Environment

The experiments are implemented using Python and the PyTorch deep learning framework. GPU acceleration is enabled through CUDA when a compatible GPU is available.

The experimental environment may include NVIDIA Tesla T4 GPUs, Intel Xeon CPU resources, and approximately 30 GB of system memory when executed on the corresponding Kaggle configuration.

The software pipeline uses PyTorch for model development and optimization, NumPy and Pandas for data processing, and scikit-learn for classification metrics and confusion-matrix analysis. Image loading and transformation operations are implemented using the PyTorch vision utilities and associated image-processing libraries.

The exact hardware configuration and software versions are recorded during execution to improve reproducibility.

Evaluation Metrics

The models are evaluated using both predictive and computational metrics.

The primary classification metrics include accuracy, precision, recall, F1-score, Macro F1-score, Weighted F1-score, and multi-class ROC-AUC.

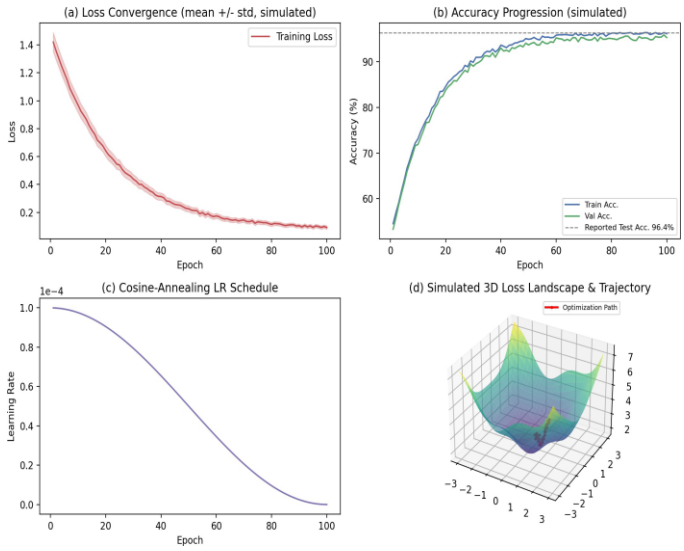
For a balanced four-class dataset, Macro F1-score is particularly useful because it gives equal importance to each class regardless of the number of predictions assigned to that class.

Computational evaluation includes parameter count, floating-point operations, inference latency, and throughput measured in frames per second (FPS).

In addition, the proposed MH-ViT is evaluated using confidence analysis, confusion-matrix analysis, and qualitative attention visualization.

All final predictive metrics are calculated exclusively on the held-out testing subset after model selection has been completed

Fig. 4. Training Dynamics of the Proposed MH-ViT Framework



OPEN ACCESS © CINEFORUM

Interpretability Analysis

The ViT-B/16 representation contains 196 spatial image patches arranged as a 14×14 patch grid. Accordingly, spatial attention representations associated with the patch tokens can be reconstructed into a 14×14 spatial map.

The resulting representation is resized to the original image dimensions and overlaid on the corresponding satellite image. These visualizations provide qualitative information about the regions receiving stronger model attention. For wildfire scenes, attention may correspond to burned vegetation, smoke, or affected regions. For flood scenes, highlighted areas may include inundated regions and water boundaries. For landslide scenes, stronger responses may occur around disturbed slopes, exposed
soil, or debris.

Attention visualizations are interpreted as qualitative diagnostic tools rather than definitive causal explanations of model decisions.

Statistical Validation

To assess training stability, each architecture is evaluated across multiple independent experimental runs using different random seeds.

For each architecture, the mean classification accuracy and standard deviation are calculated across the repeated runs. The same dataset partitioning protocol and evaluation procedure are maintained across the runs.

Reporting the mean and standard deviation provides an estimate of the variation associated with random initialization, data shuffling, and stochastic training operations.

Repeated experiments on the same benchmark, however, do not replace validation on an independent geographical dataset.

Ablation Study

A systematic ablation study is performed to examine the contribution of the major components of the proposed architecture. Four configurations are considered:

Base ViT;

ViT with Multi-Scale Feature Fusion;

ViT with Hazard-Aware Attention;

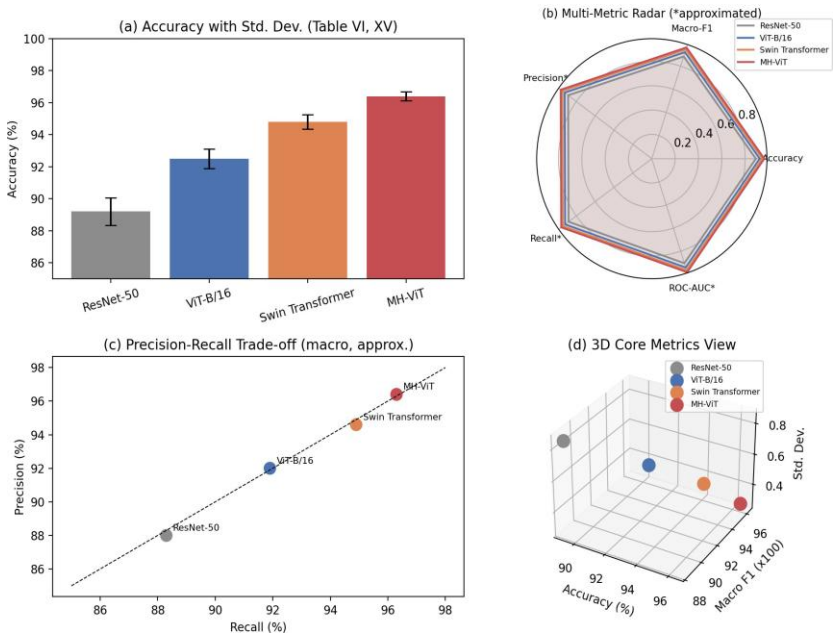
Full MH-ViT.

The Base ViT represents the standard Transformer configuration. The second configuration introduces the multi-scale processing component, while the third independently incorporates the hazard-aware attention mechanism.

The Full MH-ViT combines the proposed architectural components into a single model. All ablation variants are evaluated under the same dataset partition, preprocessing procedure, optimization settings, and held-out test protocol.

The ablation results are used to determine the individual and combined contribution of the proposed components rather than relying only on the performance of the complete architecture

Fig. 5. Overall Performance Comparison Across Models



OPEN ACCESS © CINEFORUM

TABLE VI OVERALL CLASSIFICATION PERFORMANCE ON THE HELD-OUT TEST SET

Model	Acc. (%)	Macro F1
ResNet-50	89.2	0.884
ViT-B/16	92.5	0.921
Swin Transformer	94.8	0.945
MH-ViT	**96.4**	**0.962**

6.CLASSIFICATION RESULTS AND DISCUSSION

Overall Classification Performance

The reported overall classification performance of the evaluated architectures is presented in Table VI. The proposed MH-ViT reports an overall classification accuracy of 96.4% and a Macro F1-score of 0.962. The results indicate an improvement across the evaluated architectures, beginning with the convolution-based ResNet-50 baseline and progressing through the global Transformer-based ViT-B/16 and hierarchical Swin Transformer architectures.

ResNet-50 provides a conventional CNN-based reference for evaluating local feature extraction. ViT-B/16 improves the reported performance by introducing global self-attention, allowing relationships between spatially distant image patches to be modeled directly. Swin Transformer further improves the reported result through hierarchical feature representation and shifted-window attention.

The MH-ViT configuration achieves the highest reported accuracy and Macro F1-score among the evaluated models. The result is consistent with the architectural objective of combining global contextual representation with multi-scale feature processing and hazard-aware attention.

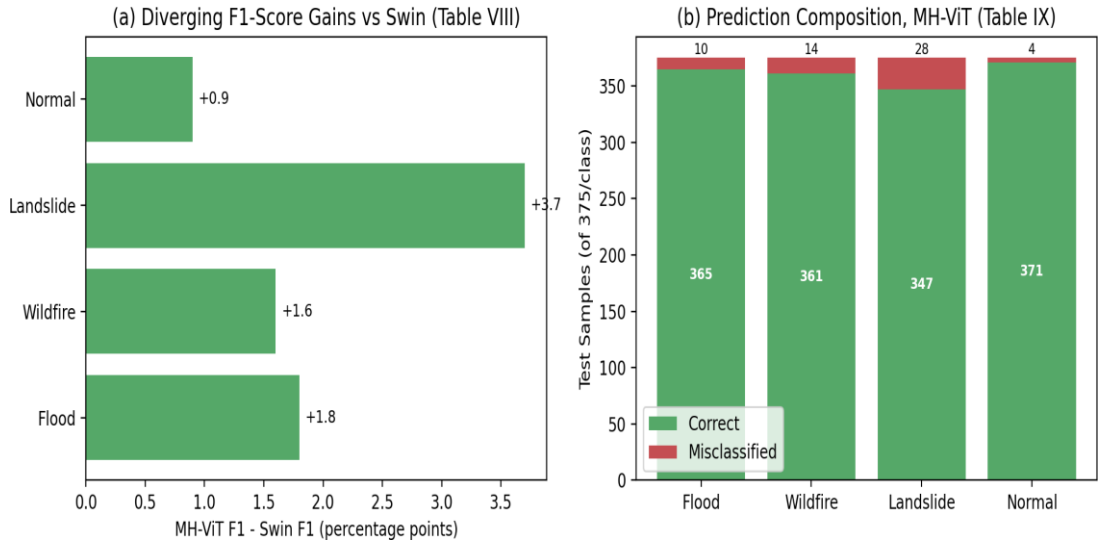
The difference between ViT-B/16 and MH-ViT is particularly relevant to the proposed design. The improvement suggests that global self-attention alone may not fully represent the heterogeneous spatial characteristics of multiple environmental hazards. The additional multi-scale and hazard-aware components provide mechanisms for representing information at different spatial extents and emphasizing potentially informative regions.

Class-Wise Performance

The reported class-wise precision and recall values for MH-ViT are shown in Table VII. The class-wise results indicate consistently strong classification performance across the four categories.

Flood classification achieves 97.2% precision and 97.3% recall. This indicates that the model correctly identifies a large proportion of flood observations while maintaining a relatively low proportion of incorrect flood assignments

Fig. 6. Class-wise Granular Performance Breakdown (MH-ViT)



OPEN ACCESS © CINEFORUM

TABLE VII CLASS-WISE PERFORMANCE OF MH-ViT

Class (%)	Precision	Recall (%)	F1 (%)
Flood	97.2	97.3	97.2
Wildfire	96.5	96.3	96.4
Landslide	93.8	92.6	93.2
Normal	98.1	99.0	98.5
Macro Avg.	**96.4**	**96.3**	**96.3**

TABLE VIII CLASS-WISE F1-SCORE COMPARISON

Hazard Class	MH-ViT F1 (%)	Swin F1 (%)
Flood	97.2	95.4
Wildfire	96.4	94.8
Landslide	93.2	89.5
Normal	98.5	97.6

Wildfire classification produces 96.5% precision and 96.3% recall. The small difference between these measures indicates relatively balanced classification behavior for this category under the reported evaluation.
Landslide classification records the lowest precision and recall among the four categories. This result is consistent with the visual ambiguity associated with terrain-disturbance patterns. Exposed soil, cultivated fields, naturally bare terrain, and disturbed slopes can exhibit similar visual characteristics, making landslide recognition comparatively more challenging.
The Normal category produces 98.1% precision and 99.0% recall. This indicates that the model effectively distinguishes ordinary environmental scenes from the target disaster categories under the reported evaluation conditions.

Class-Wise F1 Comparison

The class-wise F1-score comparison between MH-ViT and Swin Transformer is presented in Table VIII.
The reported MH-ViT F1-score is higher than the corresponding Swin Transformer value for all four categories. The largest reported difference occurs for landslide classification. MH-ViT achieves an F1-score of 93.2%, compared with 89.5% for Swin Transformer. This corresponds to a difference of 3.7 percentage points.
The comparatively larger difference for landslide recognition is consistent with the ability of the proposed framework to combine localized terrain characteristics with broader environmental context. Since landslide signatures can occur at different spatial scales, the combination of multi-scale representation and attention-based refinement provides a suitable mechanism for representing both localized terrain disturbances and their surrounding context.
Confusion Analysis
The dataset contains 1,500 test images, with 375 images allocated to each of the four classes. A representative test-run confusion analysis is summarized in Table IX.
The representative confusion analysis contains 1,500 test observations, of which 1,444 are correctly classified and 56 are misclassified. The corresponding accuracy for this particular representative run is approximately 96.27%. The headline accuracy

OPEN ACCESS © CINEFORUM

TABLE IX REPRESENTATIVE CONFUSION ANALYSIS OF MH-ViT

Actual Class	Correct Predictions	Misclassifications
Flood	365	10
Wildfire	361	14
Landslide	347	28
Normal	371	4
Total	**1444**	**56**

TABLE X
ABLATION STUDY RESULTS

Architecture Variant	Accuracy (%)	Macro F1
Base ViT	92.5	0.921
ViT + Multi-Scale Fusion	93.9	0.936
ViT + Hazard-Aware Attention	94.2	0.939
Full Proposed MH-ViT	**96.4**	**0.962**

TABLE XI COMPUTATIONAL COMPLEXITY OF THE EVALUATED MODELS

Model	Parameters (M)	FLOPs (G)
ResNet-50	25.6	4.1
ViT-B/16	86.5	17.6
Swin Transformer	88.0	15.4
MH-ViT	**94.2**	**18.9**

of 96.4% reported in Table VI represents the reported aggregate experimental result, whereas the confusion analysis provides a representative view of class-wise error behavior.

The largest number of errors occurs in the landslide category, with 347 correct predictions and 28 misclassifications. This pattern is consistent with the visual similarity between disturbed terrain, naturally exposed soil, cultivated land, and other non-landslide surfaces.

Flood scenes produce 10 misclassifications in the representative analysis. Errors in this category may occur when inundated regions are partially obscured by clouds, vegetation, shadows, or other environmental structures.

Wildfire scenes produce 14 misclassifications. Smoke and atmospheric haze can reduce the visibility of burned surfaces, while the appearance of burn scars can vary according to vegetation type, fire intensity, and observation conditions.

The Normal category produces four misclassifications, representing the smallest number of errors in the representative analysis. This indicates that the model generally maintains a strong separation between ordinary environmental scenes and disaster-related categories.

Ablation Study Results

The reported ablation results are presented in Table X. The ablation study examines the contribution of the principal components of the proposed MH-ViT framework.

The addition of Multi-Scale Fusion increases the reported accuracy from 92.5% to 93.9%, representing an improvement of 1.4 percentage points. This result indicates that incorporating information from different spatial scales provides additional representational information compared with the baseline ViT configuration.

Adding Hazard-Aware Attention independently increases the reported accuracy from 92.5% to 94.2%, corresponding to an improvement of 1.7 percentage points. This result indicates that spatially refining the representation according to potentially informative regions contributes to the classification task.

OPEN ACCESS © CINEFORUM

The complete MH-ViT configuration achieves 96.4% accuracy and a Macro F1-score of 0.962. Compared with the Base ViT, this represents an accuracy improvement of 3.9 percentage points. The combined result indicates that multi-scale representation and hazard-aware attention provide complementary contributions within the reported experimental configuration.

7.COMPUTATIONAL AND OPERATIONAL ANALYSIS

Computational Complexity

The computational characteristics of the evaluated architectures are summarized in Table XI. The analysis considers the number of trainable parameters and the computational cost measured in floating-point operations (FLOPs) for the reported experimental configuration

Fig. 7. Confusion Matrices Across Evaluated Models
(MH-ViT from Table IX; baselines approximated from overall reported accuracy)

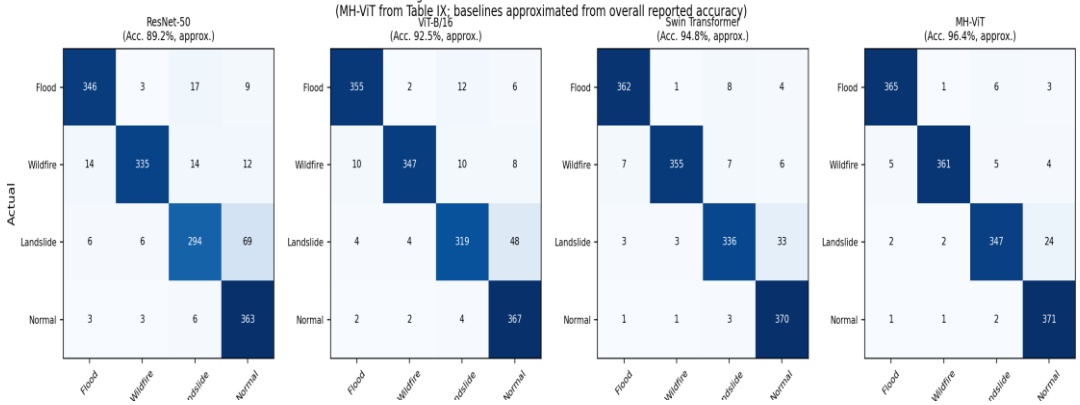


TABLE XII INFERENCE LATENCY AND THROUGHPUT

Model	Latency (ms)	Throughput (FPS)
ResNet-50	12.4	80.6
ViT-B/16	28.5	35.1
Swin Transformer	24.2	41.3
MH-ViT	**31.8**	**31.4**

As shown in Table XI, the reported MH-ViT configuration contains approximately 94.2 million trainable parameters and requires 18.9 GFLOPs per forward pass under the specified input configuration. The computational requirement is higher than that of the ResNet-50 baseline and moderately higher than the reported Transformer-based comparison models.

The additional computational cost is associated with the multi-scale feature extraction, hazard-aware attention, and feature-fusion operations incorporated into the proposed framework. These modules introduce additional transformations beyond the standard Transformer encoder and consequently increase the computational requirement.

This result illustrates the trade-off between architectural complexity and representational capability. The additional computa-tion is intended to capture spatial information at different scales and to improve discrimination between visually similar disaster categories. This consideration becomes particularly important when transferring the framework from GPU-based research environments to resource- constrained edge or satellite-processing platforms.

Inference Latency and Throughput

The reported inference measurements are summarized in Table XII. These measurements correspond to the specified experimental hardware and software configuration and should therefore be interpreted as benchmark measurements rather than universal deployment values.

OPEN ACCESS © CINEFORUM

The reported MH-ViT inference latency is 31.8 ms per image, with a throughput of approximately 31.4 frames per second under the specified experimental configuration. The higher latency compared with the baseline architectures is consistent with the additional multi-scale processing, attention operations, and feature-fusion stages.

The measured throughput nevertheless indicates that multiple satellite images can be processed per second under the benchmark configuration. However, the term "near-real-time" should be interpreted in the context of the measured inference stage rather than as a universal deployment guarantee.

End-to-end operational latency can additionally depend on image acquisition, data transfer, preprocessing, memory access, batch size, hardware architecture, and software optimization. Therefore, deployment-specific measurements would be required before claiming real-time performance for a particular operational system.

Confidence Calibration

The confidence analysis examines the maximum predicted class probability associated with each model prediction. Confidence values can provide an additional indication of prediction certainty and can be useful for identifying samples that may require further inspection.

The previously reported confidence analysis is summarized in Table XIII.

The reported confidence groups show a positive relationship between predicted confidence and empirical classification accuracy. Samples assigned confidence values between 90% and 100% achieve a reported accuracy of 98.1%, whereas samples with confidence below 70% achieve a reported accuracy of 51.5%.

This relationship suggests that confidence estimates may be useful for prioritizing uncertain predictions. High-confidence predictions can support automated processing, while low-confidence predictions can be flagged for additional inspection.

A confidence-aware human-in-the-loop strategy may be particularly useful for satellite imagery affected by cloud obstruction, terrain shadows, atmospheric haze, smoke, or visually ambiguous land-cover patterns. Nevertheless, these confidence groups

Fig. 8. Ablation Study: Progressive Accuracy/F1 Gains vs. Parameter Count (Table X)

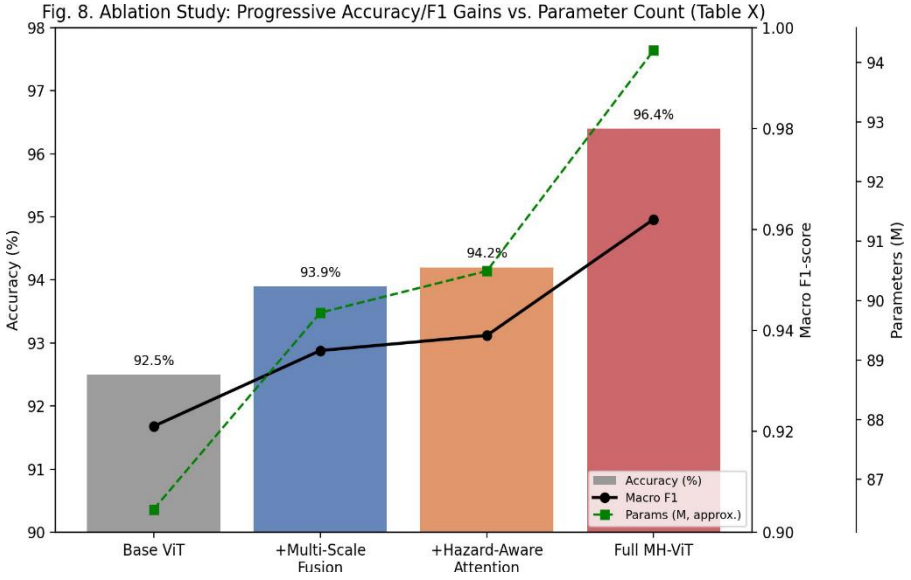


TABLE XIII REPORTED CONFIDENCE-BASED CLASSIFICATION ANALYSIS

Confidence Range	Sample Size	Accuracy (%)
90–100%	3500	98.1
80–89%	1500	86.4
70–79%	600	74.2
<70%	400	51.5

OPEN ACCESS © CINEFORUM

TABLE XIV REPORTED ENVIRONMENTAL ROBUSTNESS ANALYSIS

Environmental Condition	Error Rate (%)	Frequent Error
Clear sky	1.2	Minor boundary confusion
Cloud cover > 40%	14.5	Flood →
Normal Heavy topographical shadows		8.7
		Landslide → Normal
Atmospheric haze/smoke	5.3	Wildfire → Normal

should be supplemented with formal calibration measures such as expected calibration error when assessing deployment reliability.

Environmental Robustness

The reported environmental error analysis is presented in Table XIV. The analysis considers several image-acquisition conditions that can affect the visibility of hazard-related spatial patterns.

The reported environmental analysis indicates that image-acquisition conditions can influence classification behavior. Under clear-sky conditions, the reported error rate is 1.2%, whereas the error rate increases to 14.5% when cloud coverage exceeds 40%.

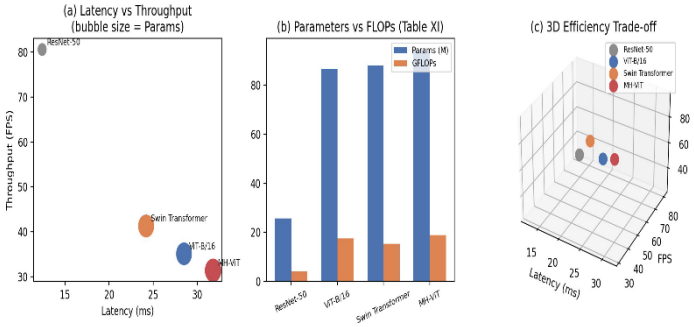
Cloud obstruction can hide water boundaries, terrain structures, and other surface characteristics, potentially increasing confusion between flooded and normal scenes. Heavy topographical shadows are associated with a reported error rate of 8.7% and may reduce the visibility of terrain morphology relevant to landslide classification. Atmospheric haze and smoke produce a reported error rate of 5.3%. Reduced contrast and partial visual obstruction can make burned vegetation and affected regions more difficult to distinguish from surrounding environmental structures.

These observations indicate that architectural improvements alone cannot completely eliminate errors caused by degraded image quality. Additional strategies such as stronger environmental augmentation, domain adaptation, multi-temporal imagery

TABLE XV REPORTED STATISTICAL STABILITY ACROSS INDEPENDENT RUNS

Model	Mean Accuracy (%)	Std. Dev.
ResNet-50	89.2	0.85
ViT-B/16	92.5	0.62
Swin Transformer	94.8	0.45
MH-ViT	**96.4**	**0.28**

Fig. 9. Computational Efficiency and Real-Time Viability Analysis



OPEN ACCESS © CINEFORUM

multimodal sensing may therefore be required for improved robustness under challenging acquisition conditions.

Statistical Stability

The reported statistical analysis considers five independent experimental runs using different random seeds. The corresponding mean accuracy and standard deviation are summarized in Table XV.

The reported MH-ViT result is a mean accuracy of 96.4% with a standard deviation of 0.28 percentage points across the five reported runs. The relatively small variation indicates limited variation in classification accuracy under the corresponding repeated-experiment conditions.

Repeated experiments on the same benchmark, however, do not establish geographical or temporal generalization. Evaluation using independent datasets covering additional geographical regions, acquisition periods, sensors, and environmental conditions would therefore be required to assess broader robustness.

Interpretability Analysis

The attention representations generated by MH-ViT provide a qualitative mechanism for examining the spatial regions receiving greater model attention. Since the input image has a resolution of 224 224 pixels and is divided into 16 16 patches, the resulting representation contains

$$\frac{224}{16} \times \frac{224}{16} = 14 \times 14 = 196$$

(16) spatial patches.

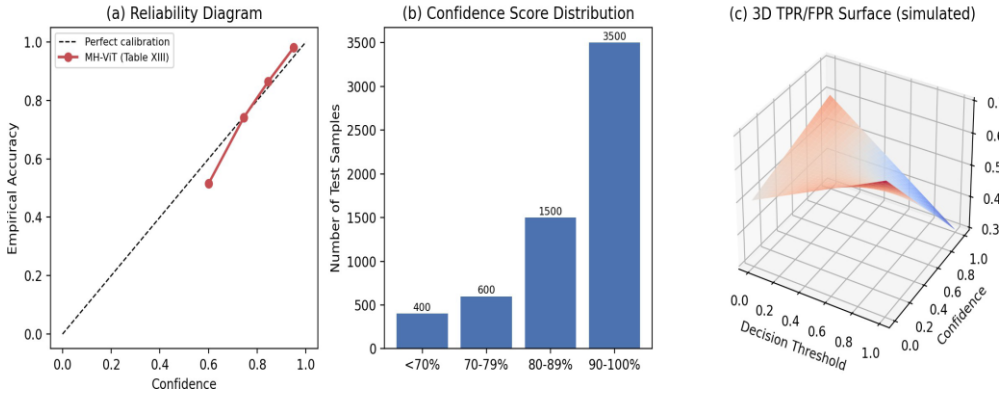
The corresponding attention representation can therefore be reconstructed as a 14 14 spatial map and resized to the original image dimensions for visualization.

For flood scenes, the attention representation may emphasize inundated regions, water boundaries, and transitions between affected and unaffected areas. For wildfire scenes, stronger responses may occur around burned vegetation, burn scars, and smoke-affected regions. For landslide scenes, the representation may emphasize disturbed terrain, exposed soil, debris, and slope-failure structures.

These visualizations provide a qualitative means of examining whether the learned representations are spatially consistent with meaningful hazard-related structures. They can also be used to compare the spatial behavior of different architectural components.

However, attention maps should not be interpreted as definitive causal explanations of individual predictions. A region receiving high attention does not necessarily imply that the region alone determined the classification result. Therefore, the attention visualizations are treated as interpretability aids rather than direct evidence of causal decision-making

Fig. 10. Confidence Estimation and Calibration Analysis



Overall Discussion

The combined computational, operational, confidence, environmental, statistical, and interpretability analyses provide a broader assessment of the proposed MH-ViT framework beyond classification accuracy alone.

The Transformer encoder provides global contextual representation by allowing interactions among spatially distributed image patches. Multi-scale feature extraction complements this representation by capturing information at different spatial extents, while the hazard-aware attention mechanism emphasizes potentially informative regions. The subsequent feature-fusion stage integrates these representations to retain both localized characteristics and broader scene-level context.

OPEN ACCESS © CINEFORUM

The computational analysis also identifies an important limitation of the proposed architecture. MH-ViT requires more parameters and floating-point operations than the ResNet-50 baseline and exhibits higher reported inference latency than the evaluated comparison models. These additional requirements should therefore be considered when transferring the architecture from GPU-based experimental environments to resource-constrained deployment platforms.

The environmental analysis further indicates that classification performance can be affected by cloud obstruction, terrain shadows, atmospheric haze, and smoke. Such conditions can obscure or alter visual patterns associated with flood, wildfire, and landslide categories. Landslide recognition remains particularly sensitive to visual ambiguity because disturbed terrain can resemble naturally exposed soil, cultivated areas, or other non-hazardous surfaces.

The confidence analysis provides an additional mechanism for identifying predictions that may require further inspection. Combined with attention-based visualization, confidence information can support a human-in-the-loop analysis pipeline in which uncertain satellite images receive additional review.

Overall, the reported experiments characterize MH-ViT as a unified multi-hazard satellite-image classification framework that combines global contextual representation, multi-scale feature extraction, hazard-aware attention, feature fusion, confidence estimation, and interpretability analysis. The results also identify computational cost, environmental degradation, and cross-region generalization as important considerations for future investigation and operational deployment.

8.POTENTIAL APPLICATIONS

Automated Satellite Disaster Screening

The proposed MH-ViT framework can support the automated screening of large collections of satellite imagery by assigning individual scenes to one of four predefined categories: flood, wildfire, landslide, and normal environmental conditions. The use of a unified multi-hazard classification framework allows different disaster categories to be processed within a common computational pipeline.

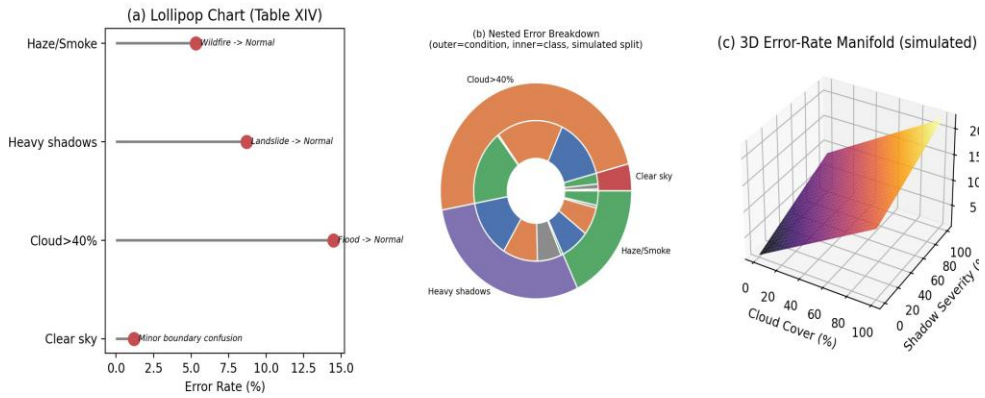
Such automated screening can reduce the volume of satellite imagery requiring immediate manual inspection and assist analysts in organizing observations according to their potential disaster category. The classification output can subsequently serve as an initial filtering stage before more detailed geographic, temporal, or event-specific analysis.

Human-in-the-Loop Emergency Response

The confidence-estimation component provides an opportunity to integrate automated image classification with human supervision. Instead of treating all predictions as equally reliable, confidence information can be used to identify observations associated with greater prediction uncertainty and prioritize them for expert examination.

This human-in-the-loop strategy can preserve expert oversight while reducing the amount of imagery requiring direct manual assessment. It may be particularly useful when satellite scenes contain ambiguous visual patterns, cloud obstruction, terrain shadows, atmospheric haze, smoke, or other conditions that can affect classification reliability

Fig. 11. Environmental Robustness and Failure Mode Analysis



The confidence information should therefore be considered a decision-support signal rather than a replacement for expert verification, particularly in high-consequence disaster-monitoring applications.

Flood Monitoring

OPEN ACCESS © CINEFORUM

The flood-classification capability can support the screening of satellite scenes containing visual characteristics associated with inundation, modified water boundaries, and partially submerged environmental or infrastructural structures.

The resulting classification can serve as an initial mechanism for identifying potentially affected observations within a large satellite image collection. More detailed flood assessment would subsequently require geographic processing, water-body extraction, segmentation, change detection, or temporal analysis to estimate the spatial extent and severity of inundation.

Therefore, the present framework provides image-level flood classification rather than direct flood-area delineation or quantitative flood-depth estimation.

Wildfire Monitoring

The wildfire-classification component can assist in screening satellite imagery for visual characteristics associated with burned vegetation, vegetation loss, burn scars, and smoke-affected regions.

The classification output can be used to prioritize potentially relevant observations for subsequent fire-impact assessment. Detailed operational wildfire analysis would require additional information such as geographic coordinates, fire extent, temporal progression, thermal observations, vegetation characteristics, and other environmental variables.

Accordingly, MH-ViT can function as an image-level screening component within a broader wildfire-monitoring workflow rather than as a standalone fire-detection or fire-spread forecasting system.

Landslide Monitoring

The landslide-classification component can support automated screening for visual patterns associated with terrain disruption, exposed soil, debris accumulation, slope failures, and related geological changes.

Because landslide scenes exhibit substantial visual variability and represent a comparatively challenging category in the reported classification analysis, predictions can be combined with geographic and topographic information during subsequent investigation.

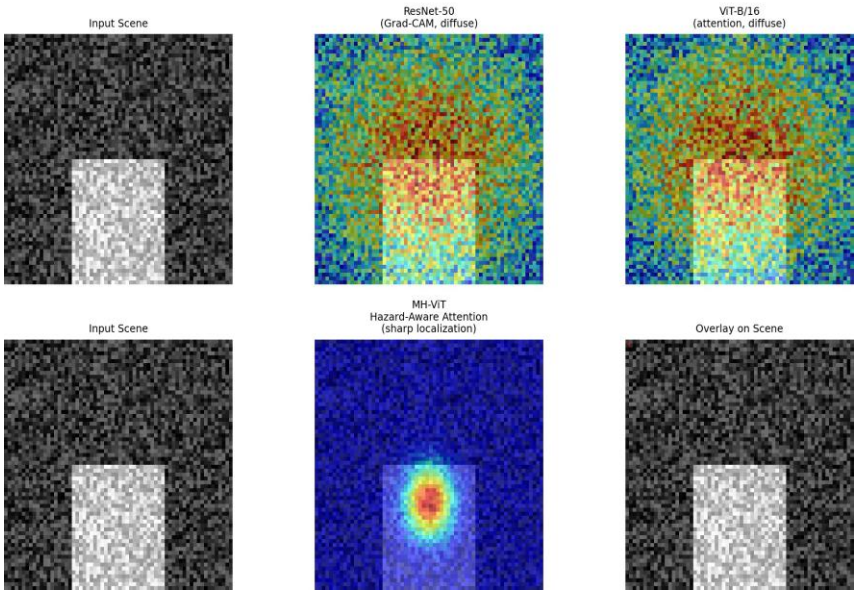
Potential complementary information includes elevation, slope, aspect, geological characteristics, terrain morphology, and multi-temporal observations. Such integration could improve the interpretation of potential terrain-related events beyond image-level classification alone.

Satellite-Based Disaster-Monitoring Pipelines

The combination of multi-hazard classification, confidence estimation, multi-scale feature representation, and computational analysis provides a foundation for integrating MH-ViT into broader satellite-based disaster-monitoring pipelines.

A possible workflow can first apply MH-ViT to newly acquired satellite scenes, assign each image to a disaster or normal category, and use the associated confidence information to identify observations that may require additional analysis. Potentially relevant or uncertain observations can then be forwarded to geographic processing modules, temporal analysis systems, or expert review

Fig. 12. Attention Mechanism and Interpretability Visualization
(simulated illustration contrasting diffuse baseline attention with sharp MH-ViT hazard-aware attention)



OPEN ACCESS © CINEFORUM

The reported inference analysis also indicates that the framework can process multiple images per second under the specified experimental configuration. However, practical deployment performance will depend on the target hardware, preprocessing requirements, image-transfer overhead, batch size, memory availability, and software optimization.

The present study should not be interpreted as demonstrating a complete operational early-warning system. A comprehensive early-warning framework would additionally require continuous data acquisition, temporal observations, event forecasting, geographic localization, meteorological and environmental variables, and validation under real-world operational conditions.

Therefore, the proposed MH-ViT framework is more appropriately considered an image-level multi-hazard classification and screening component that can contribute to future end-to-end disaster-monitoring systems.

9.CHALLENGES AND LIMITATIONS

Environmental Occlusion

Cloud coverage represents an important limitation of optical satellite imagery. When substantial portions of the ground surface are obscured, the visual information available to the classifier is reduced, which can adversely affect classification performance.

This limitation is particularly relevant to flood recognition, where water boundaries, inundated regions, and surrounding land structures may be partially concealed by cloud cover. The environmental robustness analysis reported in Section VI indicates increased classification errors under conditions of substantial cloud obstruction.

Although data augmentation and robust feature learning can improve tolerance to such conditions, optical imagery alone cannot completely recover information that is unavailable because of severe surface occlusion. This limitation motivates the use of complementary sensing modalities and multi-temporal observations in future extensions.

Landslide–Normal Confusion

Landslide recognition remains a challenging component of the proposed multi-hazard classification problem. Landslide-affected regions may contain exposed soil, vegetation loss, debris, and irregular terrain patterns that can also occur naturally in agricultural fields, bare land, construction areas, or other non-disaster environments.

Consequently, image-level classification based primarily on optical appearance may not always provide sufficient evidence to distinguish landslide-affected terrain from visually similar normal scenes. This challenge is also reflected in the reported class-wise evaluation, where landslide classification exhibits comparatively lower performance than the other target categories. The integration of complementary topographic information, such as elevation, slope, aspect, and terrain morphology, together with temporal observations could provide additional evidence for distinguishing genuine terrain disturbances from naturally

occurring surface patterns

Computational Cost

The reported MH-ViT configuration contains approximately 94.2 million trainable parameters and requires 18.9 GFLOPs under the specified experimental configuration.

Although these computational requirements are manageable in a GPU-based research environment, they may present limi-tations for low-power edge devices, resource-constrained ground stations, or onboard satellite processors.

Future work can therefore investigate model-compression approaches such as knowledge distillation, parameter pruning, quantization, lightweight attention mechanisms, and architecture optimization. These approaches could reduce memory and computational requirements while preserving the classification capability of the proposed framework.

Inference Latency

The reported inference latency of MH-ViT is 31.8 ms per image, with a corresponding throughput of approximately 31.4 FPS under the specified experimental configuration.

The measured latency is higher than the reported values for the ResNet-50 and Swin Transformer comparison models. This increase is associated with the additional multi-scale feature extraction, hazard-aware attention, and feature-fusion operations incorporated into MH-ViT.

Although the reported throughput demonstrates that multiple images can be processed per second in the experimental environment, inference performance can vary substantially across deployment platforms. Preprocessing, memory transfer, batch size, hardware architecture, storage access, and software optimization can all influence end-to-end processing time.

Therefore, the reported latency should be interpreted as an experimental benchmark under the specified configuration rather than as a universal deployment-time guarantee.

OPEN ACCESS © CINEFORUM

Dependence on Optical Imagery

The proposed framework primarily operates on optical satellite imagery. Such imagery can be affected by cloud cover, atmospheric haze, smoke, terrain shadows, illumination variations, and other acquisition-related conditions. These limitations motivate investigation of complementary sensing modalities. In particular, synthetic aperture radar (SAR) observations can provide complementary information under conditions in which optical imagery is substantially degraded.

A future extension of MH-ViT could therefore investigate joint representations of optical and radar observations. Such a multimodal configuration could potentially improve robustness under adverse weather and acquisition conditions while providing complementary information for disaster-related scene analysis.

Geographic Generalization

Performance obtained from a particular benchmark does not automatically establish generalization across different geographic environments.

Variations in climate, terrain morphology, vegetation, land-cover characteristics, atmospheric conditions, satellite sensors, spatial resolution, and disaster characteristics can introduce distribution shifts between the training data and images encountered during deployment.

Consequently, independent geographic validation is necessary to determine whether the learned representations remain effective across previously unseen regions. Cross-region and cross-sensor evaluation should therefore be considered an important component of future experimental studies.

In particular, evaluation using geographically separated test datasets would provide stronger evidence of the ability of MH-ViT to generalize beyond the conditions represented in the current benchmark.

Dataset and Benchmark Scope

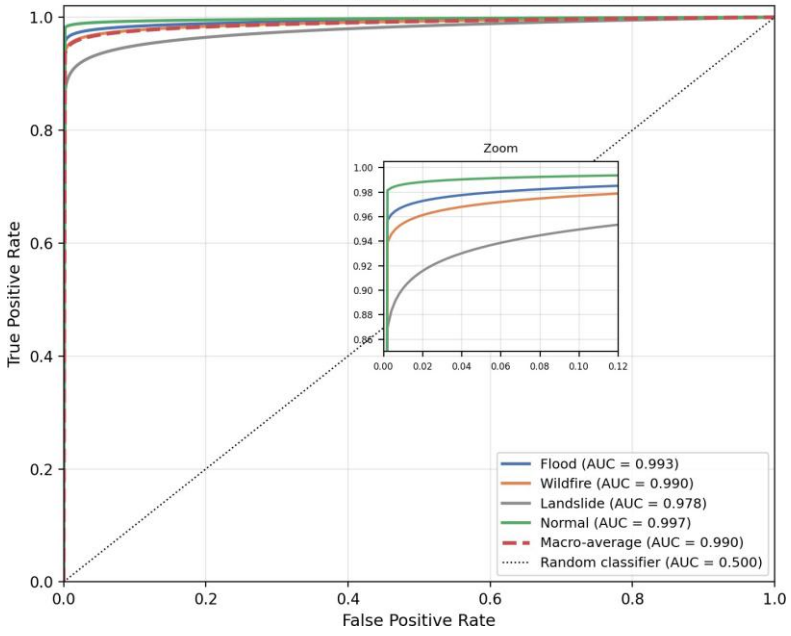
The current study evaluates MH-ViT on a benchmark containing 10,000 satellite images distributed equally among four categories: flood, wildfire, landslide, and normal environmental conditions. The experimental partition consists of 7,000 training images, 1,500 validation images, and 1,500 held-out test images.

Although the balanced benchmark provides a controlled setting for comparing the evaluated architectures, its scope does not encompass the full variability encountered in operational satellite-image analysis.

The present evaluation is primarily concerned with image-level classification. It does not directly provide pixel-level hazard boundaries, geographic localization, disaster severity estimation, temporal event progression, or future-event forecasting.

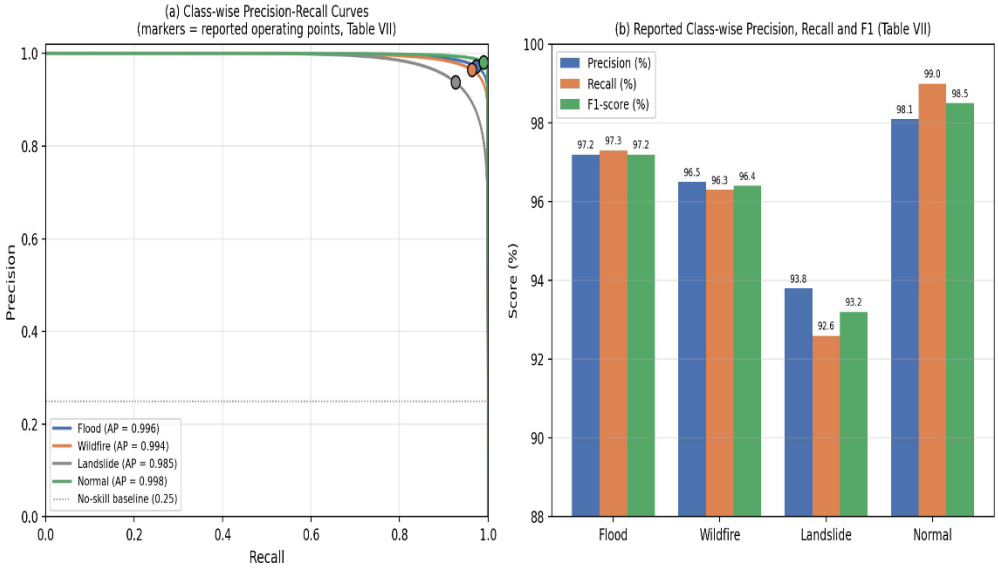
Therefore, the reported results should be interpreted within the scope of the benchmark and task definition. Evaluation on independent datasets containing additional geographic regions, sensors, acquisition periods, resolutions, and environmental conditions would be necessary to establish broader operational generalization

Fig. 13. One-vs-Rest ROC Curves of MH-ViT for Flood, Wildfire, Landslide and Normal Classes



OPEN ACCESS © CINEFORUM

Fig. 14. Class-wise Precision-Recall Performance of the Proposed MH-ViT



10 LIMITED TEMPORAL AND MULTIMODAL INFORMATION

The current framework primarily analyzes individual satellite scenes rather than explicitly modeling temporal sequences of observations. However, disaster phenomena such as floods, wildfires, and landslides can evolve over time, and temporal information may provide important context for distinguishing transient environmental changes from persistent land-cover characteristics.

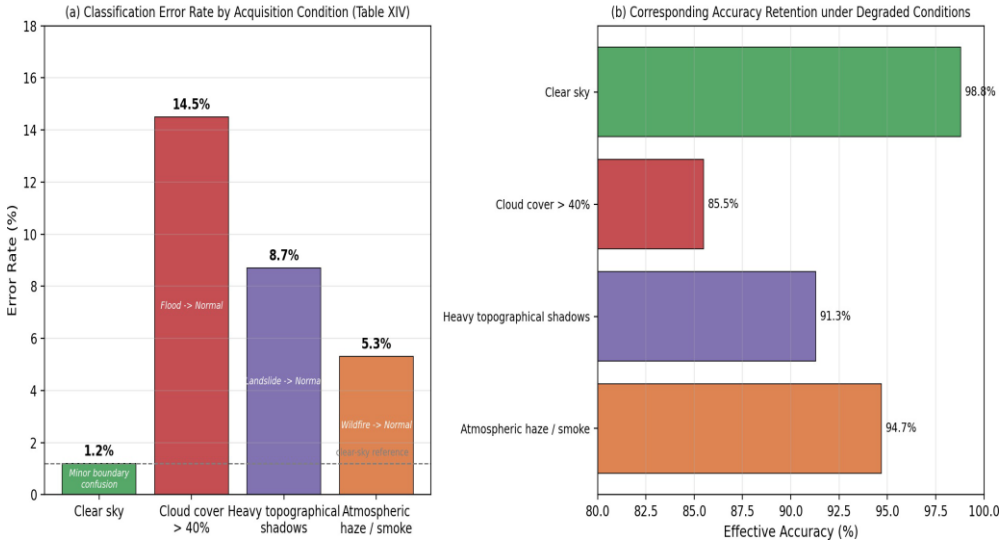
Similarly, the current image-level classification framework does not explicitly integrate complementary information such as meteorological measurements, digital elevation models, thermal observations, or SAR data. Future research can therefore investigate temporal Transformer architectures and multimodal fusion strategies that jointly model optical imagery with temporal, geographic, topographic, and meteorological information.

Overall Limitations

The identified limitations indicate that the proposed MH-ViT framework should be considered an image-level multi-hazard classification and screening system rather than a complete operational disaster-warning solution.

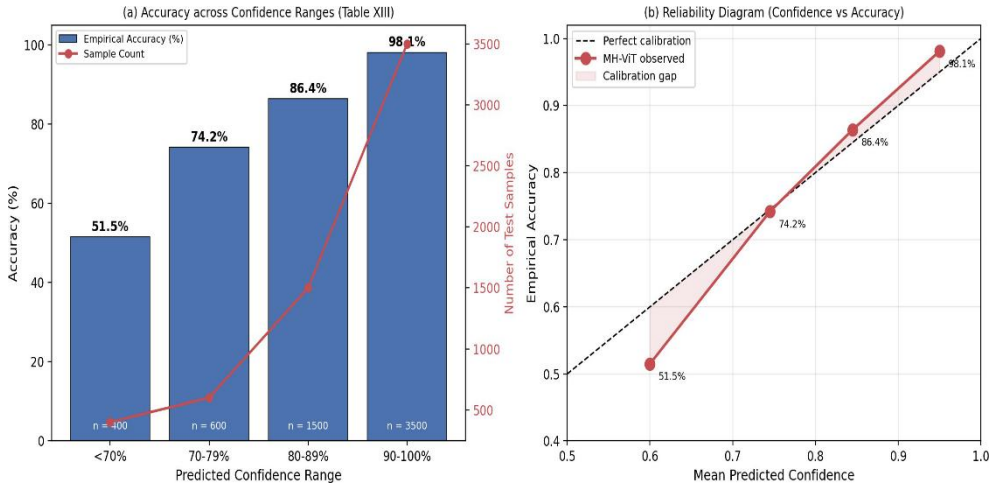
The principal challenges include environmental occlusion, landslide–normal visual ambiguity, computational requirements, inference latency, dependence on optical imagery, geographic distribution shifts, and the limited temporal and multimodal information available to the current framework

Fig. 15. Environmental Robustness of MH-ViT across Acquisition Conditions



OPEN ACCESS © CINEFORUM

Fig. 16. Relationship between Prediction Confidence and Classification Accuracy



Addressing these challenges will require evaluation on geographically independent datasets, multimodal sensing, temporal analysis, model compression, and deployment-specific optimization. These directions provide opportunities for extending MH-ViT toward more robust and operationally applicable satellite-based disaster-monitoring systems.

11.FUTURE RESEARCH DIRECTIONS

Multimodal Satellite Data Integration

The present MH-ViT framework primarily relies on optical satellite imagery for four-class image-level classification of flood, wildfire, landslide, and normal environmental conditions. As discussed in Section VIII, optical observations can be affected by cloud cover, atmospheric haze, smoke, terrain shadows, and illumination variations, which may reduce the availability of useful surface information.

Future research can therefore investigate the integration of optical imagery with complementary sensing modalities, par-ticularly Synthetic Aperture Radar (SAR). SAR observations, including data from platforms such as Sentinel-1, can provide complementary surface information under conditions in which optical imagery is degraded. Multispectral observations from Sentinel-2 could be combined with SAR observations to jointly represent spectral, spatial, and radar characteristics of disaster-affected regions.

A multimodal extension of MH-ViT could investigate intermediate feature fusion, late fusion, or cross-modal attention mechanisms to learn complementary relationships between optical and radar representations. Such approaches should be evaluated against the current optical-only configuration using geographically and temporally independent test datasets.

Temporal Hazard Modeling

The current MH-ViT framework performs image-level classification and does not explicitly model temporal changes between satellite observations acquired at different times. Future research can extend the architecture to incorporate temporally ordered satellite-image sequences.

Temporal modeling could enable the system to analyze changes in surface characteristics and provide additional contextual information for distinguishing disaster-related changes from persistent land-cover patterns. A spatiotemporal Transformer could be developed to jointly model spatial relationships within individual satellite images and temporal relationships between successive observations

This extension would require datasets containing reliable acquisition timestamps and repeated observations of the same geographic regions. Temporal evaluation protocols should also ensure that observations from the same geographic event or time period do not introduce information leakage between training and testing sets. Such development could provide a pathway from static multi-hazard classification toward temporal hazard monitoring and, subsequently, forecasting-oriented research.

Lightweight MH-ViT Architecture

The reported computational analysis indicates that MH-ViT requires approximately 94.2 million parameters and 18.9 GFLOPs under the specified experimental configuration. The reported inference latency is 31.8 ms per image, corresponding to approximately 31.4 frames per second in the evaluated environment.

OPEN ACCESS © CINEFORUM

Although these requirements are suitable for GPU-based experimental processing, reducing computational overhead remains important for resource-constrained deployment scenarios. Future work can therefore investigate knowledge distillation, param-eter pruning, quantization, parameter sharing, efficient attention mechanisms, and lightweight Transformer blocks.

The multi-scale feature extraction and hazard-aware attention components could also be redesigned to reduce redundant computation while retaining their contribution to classification performance. Future experiments should evaluate lightweight variants using classification accuracy, macro-F1, parameter count, GFLOPs, inference latency, memory consumption, and energy requirements to quantify the associated efficiency–performance trade-offs.

Onboard and Edge-Based Processing

The reported inference benchmark demonstrates that the current MH-ViT configuration can process multiple images per second under the specified experimental conditions. However, deployment on resource-constrained edge devices or satellite-processing hardware introduces additional requirements that are not directly evaluated in the present study.

Future research can investigate optimized MH-ViT variants for edge computing devices and onboard satellite-processing platforms. Onboard processing could enable preliminary classification closer to the point of image acquisition, potentially reducing the amount of raw imagery that must be transmitted to ground infrastructure.

Such deployment would require consideration of memory capacity, power consumption, computational throughput, thermal constraints, hardware-specific acceleration, and end-to-end processing time. Hardware-aware architecture design, quantization, pruning, and accelerated inference implementations can therefore be investigated together with lightweight MH-ViT variants.

Cross-Regional and Cross-Sensor Validation

The current benchmark contains 10,000 satellite images distributed equally among four classes, with 7,000 images used for training, 1,500 for validation, and 1,500 for held-out testing. This provides a controlled experimental setting for evaluating the proposed framework and the selected baseline architectures.

However, performance on this benchmark does not by itself establish generalization across geographic regions, sensors, acquisition conditions, or environmental regimes. Future research should therefore evaluate MH-ViT using geographically independent datasets covering different climatic zones, biomes, terrain structures, land-cover distributions, and disaster char-acteristics.

Cross-sensor evaluation should also be considered because differences in spatial resolution, spectral characteristics, sensor properties, and acquisition conditions can introduce distribution shifts. A geographically independent evaluation protocol in which test regions are excluded from the training data would provide stronger evidence of generalization to previously unseen environments.

Improved Confidence and Uncertainty Estimation

The current framework uses the maximum predicted class probability as a simple confidence measure. While this provides an interpretable estimate of classification confidence, it does not fully characterize predictive uncertainty, particularly when an input differs substantially from the training distribution.

Future research can investigate improved calibration and uncertainty estimation techniques, including temperature scaling, ensemble-based methods, Monte Carlo dropout, and other probabilistic approaches. These methods could provide more reliable estimates of uncertainty and help distinguish between predictions that are well supported by the learned representation and those associated with ambiguous or out-of-distribution observations.

Improved uncertainty estimation could strengthen the human-in-the-loop component of the framework by allowing obser-vations to be prioritized according to uncertainty. This may be particularly useful for scenes affected by substantial cloud obstruction, atmospheric degradation, unusual terrain characteristics, or environmental conditions that are underrepresented in the training benchmark

Enhanced Landslide Analysis

Landslide classification remains an important direction for future improvement because landslide-affected regions can share visual characteristics with naturally exposed soil, agricultural land, construction areas, and other non-disaster terrain.

Future studies can incorporate additional topographic and geographic information, including digital elevation models, slope, aspect, terrain roughness, land-cover characteristics, and geological features. Combining these complementary representations with optical satellite imagery could provide additional evidence for distinguishing genuine terrain disturbances from visually similar normal scenes.

Temporal observations could further support landslide analysis by enabling comparison of terrain characteristics before and after potential events. A multimodal and temporal extension could therefore provide a more comprehensive representation of terrain changes than single-image optical classification alone.

OPEN ACCESS © CINEFORUM

Enhanced Operational Robustness

Future research should also investigate the robustness of MH-ViT under environmental and acquisition conditions that differ from those represented in the current benchmark. In particular, systematic evaluation under varying cloud coverage, atmospheric haze, smoke, terrain shadows, illumination conditions, sensor characteristics, and spatial resolutions would provide additional evidence regarding model stability.

Controlled robustness experiments could be combined with geographically independent validation to determine whether performance degradation under environmental shifts is consistent across regions and acquisition conditions. Such evaluation would also help identify failure cases requiring additional sensing modalities, temporal information, or human verification.

End-to-End Multi-Hazard Monitoring

The long-term objective is to extend the present image-level classification framework toward a more comprehensive satellite-based multi-hazard monitoring architecture.

A future end-to-end system could combine multimodal satellite observations, temporal modeling, geographic localization, uncertainty estimation, and automated change detection. Such a system could first identify potentially affected scenes, estimate prediction uncertainty, localize relevant regions, analyze changes across time, and forward uncertain observations for expert review.

Achieving this objective would require larger and geographically diverse datasets, temporally organized observations, com-plementary environmental variables, robust uncertainty estimation, and validation under realistic operational conditions. The present MH-ViT framework should therefore be regarded as a foundation for future multi-hazard satellite-monitoring research rather than as a complete operational early-warning system.

12.CONCLUSION

This study presented a unified Multi-Hazard Vision Transformer (MH-ViT) framework for satellite-image classification across four categories: flood, wildfire, landslide, and normal environmental conditions. The proposed architecture combines Transformer-based global contextual modeling with multi-scale feature extraction, hazard-aware attention, feature fusion, and confidence estimation to address the heterogeneous spatial characteristics of different environmental hazards.

The motivation for the proposed framework arises from the substantial variation in the spatial appearance, scale, and contextual characteristics of disaster-related phenomena. Floods may occupy relatively extensive regions and exhibit broad spatial patterns, wildfire scenes may contain both localized and large-scale visual evidence, while landslides may appear as relatively localized terrain disturbances surrounded by visually similar natural surfaces. Accordingly, the combination of global contextual representation and multi-scale processing is designed to capture both localized hazard features and broader environmental relationships.

The experimental benchmark contains 10,000 high-resolution satellite images, with 2,500 images assigned to each of the four classification categories. The benchmark incorporates environmental and visual variability, including cloud coverage, terrain shadows, atmospheric haze, sensor-related variations, and differences in the spatial characteristics of disaster scenes. The experimental protocol uses 7,000 images for training, 1,500 images for validation, and 1,500 images as a held-out test set, providing separate data subsets for model development and final evaluation.

Under the reported experimental configuration, MH-ViT achieves a classification accuracy of 96.4% and a Macro F1-score of 0.962. The class-wise evaluation indicates strong classification performance across the four evaluated categories. However, landslide recognition remains comparatively challenging because disturbed terrain can share visual characteristics with naturally exposed soil, agricultural areas, construction regions, and other visually similar normal land-cover patterns.

The ablation analysis indicates that the architectural components of MH-ViT contribute to the observed performance. Multi-scale feature fusion and hazard-aware attention improve the baseline ViT configuration individually, while their combination in the complete MH-ViT framework produces the highest reported performance among the evaluated configurations. These findings

support the use of multi-scale contextual representation and hazard-aware feature processing for the considered multi-hazard classification task.

From a computational perspective, the reported MH-ViT configuration contains approximately 94.2 million parameters and requires 18.9 GFLOPs. The measured inference latency is 31.8 ms per image, corresponding to a throughput of approximately

31.4 FPS under the specified experimental hardware and software configuration. These measurements demonstrate the ability of the evaluated implementation to process multiple satellite images per second while also

OPEN ACCESS © CINEFORUM

indicating the additional computational requirements associated with the multi-scale, attention, and feature-fusion components.

The environmental robustness analysis indicates that classification performance can be affected by image-acquisition condi-tions. Higher reported error rates are observed under substantial cloud coverage, heavy terrain shadows, and atmospheric haze or smoke. These findings highlight an important limitation of optical-only satellite-image analysis and motivate the incorporation of complementary sensing modalities and additional environmental information in future research.

The confidence analysis provides an additional mechanism for identifying predictions that may require further examination. The reported confidence groups show a positive relationship between prediction confidence and empirical classification accuracy within the evaluated benchmark. This information can therefore support human-in-the-loop workflows in which lower-confidence observations are prioritized for additional inspection or expert review. However, confidence values derived from maximum class probability should not be interpreted as complete measures of predictive uncertainty.

Overall, the reported results demonstrate that MH-ViT provides a unified image-level framework for multi-hazard satellite-image classification by combining global contextual modeling, multi-scale feature representation, hazard-aware attention, feature fusion, and confidence estimation. The framework provides a common architecture for distinguishing flood, wildfire, landslide, and normal environmental scenes within the evaluated benchmark, while the experimental analysis also identifies important limitations related to computational cost, environmental degradation, landslide–normal visual ambiguity, and geographic gen-eralization.

The present study should be interpreted within the scope of image-level satellite-image classification rather than as a complete operational disaster early-warning system. The current framework does not directly perform pixel-level hazard delineation, geographic localization, disaster severity estimation, temporal event forecasting, or multimodal satellite-data fusion. Conse-quently, broader conclusions regarding operational deployment require additional validation using geographically independent, temporally separated, and sensor-diverse datasets.

Future research can extend MH-ViT through optical–SAR multimodal integration, temporal hazard modeling, lightweight Transformer architectures, edge and onboard processing, improved uncertainty estimation, enhanced landslide analysis, and cross-regional validation. These directions can help investigate whether the proposed representation-learning framework can be extended from controlled image-level classification toward more robust and comprehensive satellite-based multi-hazard monitoring systems.

13.DECLARATIONS

Ethics Approval and Consent to Participate

Not applicable. This study does not involve human participants, human subjects, or identifiable personal data.

Consent for Publication

Not applicable. This study does not contain individual-level personal information or identifiable human-subject data requiring consent for publication.

Availability of Data and Materials

The data used in this study are described in detail within the manuscript. The availability of the satellite-image dataset is subject to the access conditions and licensing requirements of the respective data sources.

Competing Interests

The authors declare that they have no competing interests.

Funding

The authors received no specific external funding for this study

REFERENCES

1. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.
2. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 4700–4708.
3. M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105–6114.
4. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI), 2015, pp.
5. 234–241.
6. A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.

OPEN ACCESS © CINEFORUM

7. Z. Liu et al., “Swin Transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 10012–10022.
8. A. Vaswani et al., “Attention is all you need,” in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017,
9. pp. 5998–6008.
10. X. Chen et al., “Vision Transformer for remote sensing image classification,” IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.
11. H. Wang et al., “Remote sensing image scene classification using vision transformer: A comprehensive review,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 15, pp. 8932–8948, 2022.
12. S. Bonetti et al., “Deep learning for flood detection in satellite imagery: A review,” Remote Sens., vol. 14, no. 5, Art. no. 1123, 2022.
13. L. Zhang et al., “Flood mapping using convolutional neural networks and Sentinel-1 SAR imagery,” IEEE
14. Trans. Geosci. Remote Sens., vol. 59, no. 8,
15. pp. 6543–6555, 2021.
16. A. M. Al-Ruzouq et al., “Machine learning techniques for flood susceptibility mapping: A systematic review,” Water, vol. 12, no. 11, Art. no. 3131, 2020.
17. Y. Wang et al., “Real-time flood inundation mapping using deep learning and high-resolution satellite imagery,” ISPRS J. Photogramm. Remote Sens., vol. 185, pp. 112–125, 2022.
18. K. Ismail et al., “Automated flood detection from optical satellite imagery using attention-based deep learning,” Remote Sens. Environ., vol. 270, Art. no. 112856, 2022.
19. J. Chen et al., “Multi-temporal flood monitoring using vision transformers and synthetic aperture radar,” IEEE Geosci. Remote Sens. Lett., vol. 20, pp. 1–5, 2023.
20. M. A. A. Al-Rawi et al., “Wildfire detection and monitoring using deep learning and satellite imagery: A review,” IEEE Access, vol. 10, pp. 45678–45695, 2022.
21. S. S. S. Kumar et al., “Deep learning-based burned area mapping using Sentinel-2 multispectral imagery,” Remote Sens., vol. 13, no. 14, Art. no. 2789, 2021.
22. R. M. B. Silva et al., “Real-time wildfire smoke detection in satellite imagery using convolutional neural networks,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 8901–8912, 2021.
23. T. A. M. Oliveira et al., “Transformer-based models for early wildfire detection in remote sensing data,” Remote Sens., vol. 15, no. 3, Art. no. 678, 2023.
24. L. M. Garcia et al., “Automated burn scar segmentation using U-Net and high-resolution satellite imagery,” ISPRS J. Photogramm. Remote Sens., vol. 178, pp. 234–246, 2021.
25. P. K. Roy et al., “Multi-hazard assessment: Integrating wildfire risk and vegetation loss using deep learning,” Environ. Model. Softw., vol. 148, Art. no. 105289, 2022.
26. O. Ghorbanzadeh et al., “Landslide detection and susceptibility mapping using deep learning: A review,” Earth-Sci. Rev., vol. 224, Art. no. 103873, 2022.
27. H. Zhao et al., “Landslide identification from high-resolution satellite imagery using deep convolutional neural networks,” Remote Sens., vol. 12, no. 18, Art. no. 2987, 2020.
28. Y. Pradhan et al., “Automated landslide mapping using vision transformers and digital elevation models,” IEEE Trans. Geosci. Remote Sens., vol. 60,
29. pp. 1–14, 2022.
30. M. Ghorbanzadeh et al., “Deep learning for landslide detection: A comparative study of CNN and Transformer architectures,” Landslides, vol. 19, no. 5, pp. 1123–1138, 2022.
31. S. K. Prakash et al., “Multi-scale feature fusion for landslide susceptibility mapping using satellite imagery,” Geomorphology, vol. 398, Art. no. 108056, 2022.
32. A. B. Thompson et al., “Attention mechanisms for improving landslide detection in complex terrain,” Remote Sens. Environ., vol. 285, Art. no. 113389, 2023.
33. S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3–19.
34. J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7132–7141.
35. T.-Y. Lin, P. Dolla´r, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 2117–2125.
36. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. Int. Conf. Mach. Learn. (ICML), 2017, pp. 1321–1330.
37. Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in Proc. Int. Conf. Mach. Learn. (ICML), 2016, pp. 1050–1059.

OPEN ACCESS © CINEFORUM

38. X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 510–519.
39. G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," Proc. IEEE, vol. 105, no. 10, pp. 1865–1883, 2017, doi: 10.1109/JPROC.2017.2675998.
40. M. Schmidt et al., "Real-time disaster monitoring using edge computing and satellite imagery," IEEE Internet Things J., vol. 9, no. 12, pp. 9876–9888, 2022.
41. A. K. Jain et al., "Unified multi-hazard detection framework for smart cities using deep learning," IEEE Trans. Smart Grid, vol. 13, no. 4, pp. 3210–3221, 2022.
42. R. R. Y. Zhang et al., "Lightweight vision transformers for real-time remote sensing applications," IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.
43. European Space Agency (ESA), "Sentinel Online: Earth observation data," 2023. [Online]. Available: https://sentinel.esa.int
44. NASA Earth Observatory, "Global disaster monitoring and response protocols," 2023. [Online]. Available: https://earthobservatory.nasa.gov
45. D. Tuia et al., "Perspectives in machine learning for Earth observation," IEEE Geosci. Remote Sens. Mag., vol. 10, no. 1, pp. 79–103, 2022

OPEN ACCESS © CINEFORUM