AI-Based Disaster Detection and Response from Satellite Images

A **Project report** submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology

by

Shravani Nalawade	**12110310**
Jay Nannaware	**12110745**
Madhura Pande	**12111312**

Under the guidance of

Prof. Vrinda Parkhi

DEPARTMENT
OF
ELECTRONICS & TELECOMMUNICATION ENGINEERING

VISHWAKARMA INSTITUTE OF TECHNOLOGY PUNE

2024 - 25

Bansilal Ramnath Agarwal Charitable Trust's

VISHWAKARMA INSTITUTE OF TECHNOLOGY, PUNE - 37

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

CERTIFICATE

This is to certify that the **Project Report** entitled **AI-Based Disaster Detection and Response from Satellite Images** has been submitted in the academic year **2024-25** by

Shravani Nalawade	**12110310**
Jay Nannaware	**12110745**
Madhura Pande	**12111312**

under the supervision of **Prof. Vrinda Parkhi** in partial fulfillment of the requirements for the degree of Bachelor of Technology in **Electronics and Telecommunication Engineering** as prescribed by Savitribai Phule Pune University.

Guide/Supervisor
Name: Prof. Vrinda Parkhi
Signature:

Head of the Department
Name:Prof. Dr.Medha Wyawahare
Signature:

External Examiner
Name:
Signature:

Acknowledgment

I would like to extend my heartfelt gratitude to **Prof. (Dr.) R. Jalnekar**, the Director of Vishwakarma Institute of Technology, Pune, for providing access to the institute's outstanding resources and fostering an environment that encourages research, innovation, and academic excellence. His encouragement and visionary leadership have been instrumental in shaping the foundation of this project.

I am profoundly grateful to **Prof. (Dr.) Medha Wyawahare**, Head of the Department of Electronics and Telecommunication Engineering, VIT Pune, for her unwavering support and valuable guidance throughout this project. Her profound knowledge and thoughtful insights have been pivotal in helping us navigate through various challenges and understand the subject matter in depth.

I wish to express my sincere thanks to **Prof. Vrindha Parkhi**, my project mentor, for his invaluable assistance, persistent encouragement, and expertise. His dedication and constructive feedback have significantly contributed to the successful completion of this project and enhanced my understanding of the technical and practical aspects of the work.

I also extend my deepest appreciation to the teachers and staff members of the Department of Electronics and Telecommunication Engineering, VIT Pune. Their dedication to teaching, coupled with their support and cooperation, has profoundly inspired us and contributed greatly to our academic and professional growth.

Lastly, I am deeply thankful to my peers and colleagues for their continuous support, constructive suggestions, and collaboration throughout this journey. Their encouragement has provided the motivation and teamwork necessary to bring this project to fruition.

Shravani Nalawade	**12110310**
Jay Nannaware	**12110745**
Madhura Pande	**12111312**

ABSTRACT

Natural disasters such as floods, wildfires, earthquakes, and landslides cause extensive damage to life, property, and the environment. Rapid and accurate identification of disaster-affected regions is essential for timely response and effective disaster management. With the increasing availability of satellite imagery, artificial intelligence (AI) offers powerful tools to analyze large volumes of visual data and assist in early disaster detection. This project focuses on the development and evaluation of AI-based models for detecting disaster zones using grayscale satellite images, which are commonly used due to their availability and cost-effectiveness.

The objective of this study is to compare and benchmark multiple state-of-the-art deep learning models, including Swin Transformer, UNet, DeepLabV3+, ResNet-50, and Vision Transformer (ViT), in the context of disaster detection. Each model was trained and evaluated over 75 epochs using a curated dataset of black-and-white satellite images representing various disaster scenarios. The selected models are known for their strong capabilities in image segmentation, feature extraction, and classification, making them suitable candidates for this task.

A key novelty of this project lies in its focus on grayscale imagery—a challenging format for visual recognition tasks due to the lack of color information. By rigorously testing different models, the project aims to identify which architectures perform best in terms of segmentation accuracy, computational efficiency, and feature representation. Experimental results demonstrated that DeepLabV3+ excelled in delivering precise segmentation outputs, while Swin Transformer and ViT offered robust feature learning and contextual understanding.

This comparative study provides practical insights for selecting appropriate deep learning models in real-time disaster detection systems based on satellite data. The outcomes of this research contribute to the development of efficient and reliable AI tools that can enhance situational awareness, support emergency planning, and improve disaster response efforts.

Contents

List of Figures

List of tables

Chapter 1: Introduction

Natural disasters—such as floods, earthquakes, wildfires, and landslides—pose significant threats to human life, infrastructure, and ecosystems. Rapid and accurate disaster detection and response mechanisms are essential to minimize damage and improve emergency response time. Traditional disaster assessment methods rely heavily on manual inspection or delayed satellite imagery analysis, which are often slow, error-prone, and resource-intensive. In recent years, Artificial Intelligence (AI), particularly Deep Learning (DL) and Computer Vision, has emerged as a transformative solution to address these limitations.

In the face of an increasing number of climate-related catastrophes, there is a growing need for technologies that can provide timely and reliable assessments of affected areas. Earth observation systems, particularly those based on satellite platforms, now produce massive volumes of image data daily. However, the sheer scale and complexity of these datasets make it impractical for manual analysis, especially during ongoing disaster events. AI-driven automation can fill this gap by enabling scalable image interpretation and supporting decision-makers with actionable insights in near real-time.

One of the key strengths of AI is its capacity to detect patterns and anomalies in visual data that may not be immediately apparent to human analysts, especially in grayscale or low-resolution imagery. While multispectral and RGB data have traditionally dominated the field of remote sensing due to their rich color information and higher contextual visibility, grayscale images are more commonly available, particularly from radar-based satellites such as Sentinel-1. These grayscale sources offer critical advantages, including cloud penetration, day-and-night monitoring, and lower transmission costs. However, their reduced spectral diversity makes semantic understanding significantly more challenging.

To address these challenges, recent research has increasingly turned to advanced deep learning architectures capable of extracting meaningful features from complex visual environments. These include Convolutional Neural Networks (CNNs), encoder-decoder structures like UNet, and more recently, transformer-based models such as Vision Transformers (ViT) and Swin Transformer. Such models offer state-of-the-art performance in computer vision tasks by capturing both local textures and global contextual relationships, making them promising candidates for disaster segmentation and detection tasks on satellite imagery.

The integration of AI with satellite image analysis enables real-time, scalable, and precise disaster monitoring, especially when combined with captioning techniques that provide semantic interpretations of complex scenes. Image captioning enhances machine understanding by generating human-readable descriptions of visual data. This interdisciplinary approach can assist in automating disaster assessment, improving resource allocation, and informing rescue operations. Transformer-based architectures such as Vision Transformers (ViTs) and semantic segmentation models like UNet variants are being extensively used for damage detection, flood area identification, and post-disaster mapping.

Studies have demonstrated the effectiveness of models like UNet++, DAM-Net, and SegFormer in accurately segmenting disaster-affected areas from satellite and UAV imagery. Despite these advancements, challenges remain in terms of explainability, model generalization, and real-time applicability—highlighting the need for research that addresses these gaps while leveraging advanced deep learning methods.

While AI models have demonstrated strong performance in image-based disaster detection, several limitations persist. Most current solutions either lack the ability to semantically describe scenes in natural language or struggle with model interpretability, especially when used for critical decision-making in disaster response. Moreover, many existing models are designed for RGB images, while real-world satellite datasets often include grayscale or multispectral imagery with limited labeled data. Therefore, there is a need for a lightweight, accurate, and interpretable deep learning framework capable of processing grayscale satellite images and effectively detecting and describing disaster scenarios.

The dataset used for this project originates from Earth Observation sources and includes a combination of Sentinel-1 (SAR) and Sentinel-2 (optical) imagery. It has been curated for flood detection tasks and includes co-registered image time series from diverse geographic locations, including West and South-East Africa, the Middle East, and Australia. This dataset enables comprehensive model training for real-world flood detection, even under complex conditions like cloud occlusion, limited visibility, and grayscale-only image availability [1].

This project investigates and compares the performance of multiple deep learning models—namely, Swin Transformer, UNet, DeepLabV3+, ResNet-50, and Vision Transformer (ViT)—on grayscale satellite image datasets for disaster detection. While the original aim was to perform image captioning, limitations in data quality and color channels shifted the focus to analyzing model performance for segmentation and classification accuracy.

1.1 Problem Statement

To do comparative study of models for an interpretable and efficient deep learning framework capable of detecting and segmenting disaster-affected regions using grayscale satellite imagery.

1.2 Aim and Objective

The key objectives of the project are:

- Evaluating model accuracy, computational efficiency, and training time over 75 epochs.
- Identifying models best suited for detecting disaster zones in grayscale satellite imagery.
- Understanding the trade-offs between accuracy, interpretability, and resource usage.

The aim of this project is limited to grayscale satellite imagery. It includes:

- Preprocessing and training deep learning models using a disaster image dataset.
- Comparing segmentation results to assess visual damage localization.
- Analysing performance metrics.

The findings of this project are expected to contribute to future implementations that may incorporate vision-language models or improved captioning pipeline.

Chapter 2: Literature Survey

Artificial intelligence (AI) and deep learning have significantly transformed remote sensing image captioning (RSIC), disaster response, and emergency management. This systematic survey explores advancements in these domains, focusing on theoretical frameworks, model explainability, disaster monitoring, and AI-driven decision-making. The discussion is structured around major themes, including deep learning techniques for remote sensing, explainability in AI models, disaster assessment and response, and AI-driven optimization in disaster management.

The integration of deep learning models in RSIC has been explored through six theoretical frameworks, including encoder-decoder models, attention mechanisms, and reinforcement learning. Comparative evaluations of convolutional neural networks (CNNs), transformers, and multi-head attention mechanisms using datasets such as NWPU-Captions and RSICD indicate that transformer-based architectures and reinforcement learning techniques outperform conventional models in captioning accuracy. However, a limitation in this study is the lack of experimental validation, necessitating further empirical investigation. Complementing this research, vision-language models (VLMs) have been reviewed for their applications in remote sensing, integrating natural language processing (NLP) and computer vision for improved satellite image understanding. Techniques such as image captioning, text-based image retrieval, and object detection enhance semantic comprehension, but challenges persist in scalability, dataset quality, and AI explainability [2, 3].

A key challenge in AI-based remote sensing is explainability, where deep learning models, particularly CNNs and transformers, lack transparency in their decision-making processes. Explainability techniques such as Layer-wise Relevance Propagation (LRP) and Local Interpretable Model-Agnostic Explanations (LIME) have been employed to analyze model interpretations. The findings highlight a trade-off between computational cost and explanation quality, prompting the development of novel methods like BU-LRP and BU-LIME to improve interpretability. Furthermore, ChatGPT has been explored for land cover change analysis using satellite imagery. While automation in remote sensing tasks demonstrates potential, limitations in classification accuracy and the inability of ChatGPT to directly run machine learning models highlight the need for enhanced AI capabilities [4, 5].

AI applications in disaster assessment leverage a combination of CNNs, neural attention mechanisms, and reinforcement learning to improve disaster scene analysis. An interactive Disaster Scene Assessment (iDSA) system integrating AI and crowdsourcing has demonstrated enhanced earthquake damage evaluation using real-world datasets from Nepal and Ecuador. The system improves accuracy via the Intersection-Over-Union (IOU) metric and benefits from human feedback to refine AI-generated attention maps. However, challenges arise from the reliance on crowdsourcing incentives. Similarly, post-disaster building damage detection has been categorized into visual inspection, algorithm-based, and AI-driven techniques. Multi-temporal change detection methods using CNNs, ResNet, and transformers show improved assessment accuracy, though issues of data noise, model generalizability, and dataset quality remain [6, 7].

AI-driven digital twin frameworks have been introduced to facilitate real-time disaster monitoring. The Disaster City Digital Twin (DCDT) framework integrates AI, crisis informatics, and ICT to enhance disaster response through multi-data sensing from UAVs, social media, and crowdsourcing. Despite its potential, challenges in multi-modal data integration, AI explainability, and data reliability require further research. Extending this approach, an AI-driven digital twin framework has also been proposed for security event analysis in a TEC district, incorporating knowledge graphs and dense video captioning for predictive analytics. However, data complexity and AI model interpretability remain pressing issues [8,9].

The application of AI in flood detection has also been explored through multimodal data fusion and semantic segmentation. A real-time flood detection and notification system using the U-Net model for landmass tracking has achieved over 80% accuracy. The FloodBot system integrates solar-powered cameras, networking devices, and field sensors for AI-driven communication. Despite these advancements, challenges persist in real-world deployment and the adaptation of autonomous vehicle insights for improved risk mitigation. A related study employs IoT, big data, and convolutional deep neural networks (CDNN) to enhance early warning systems for flood detection. The CDNN classifier achieves higher accuracy than traditional artificial neural networks (ANNs), though cost-effective sensor deployment remains a challenge [10,11].

Multi-hazard disaster monitoring has been investigated using AI-driven analytics for landslide and wildfire detection, integrating IoT, UAVs, and satellite imagery. While these technologies improve real-time situational awareness, challenges such as communication network failures, data privacy concerns, and security vulnerabilities need to be addressed. Additionally, misinformation in disaster reporting presents significant concerns. A misinformation-aware community detection system using FN-BERT-TFIDF filters false information in real-time geolocated Twitter data, improving hazard detection efficiency. However, keyword-based topic modeling may introduce biases [12, 13].

AI has also been applied to optimize resource allocation and automate disaster response strategies. A systematic review of AI applications in natural disaster management identifies predictive modeling and early warning systems as key areas of research. Machine learning, deep learning, and explainable AI have been highlighted as essential techniques, though challenges in dataset quality and AI decision trustworthiness persist. Another review categorizes AI applications across the mitigation, preparedness, response, and recovery phases of disaster management. Supervised learning, deep reinforcement learning, and optimization techniques enhance big data processing, hazard forecasting, and emergency response. However, limitations in data reliability and model generalizability remain problematic [14, 15].

Cloud computing and AI optimization techniques have been employed for disaster recovery. An AI-driven optimization framework for Kubernetes cluster management in cloud environments has been developed to improve resource utilization. The system achieves a 23% efficiency improvement, though challenges in scalability and dynamic cloud adaptation require further research. Similarly, AI-driven social media analysis integrates NLP and computer vision for multimodal disaster monitoring, enhancing real-time situational awareness. Despite its advantages, misinformation and data quality issues pose challenges to effective decision-making [16, 17].

Crowdsourcing and machine learning have also been combined to process UAV imagery for disaster response. Implemented as "Aerial Clicker" within the AIDR platform, this approach accelerates damage assessment but faces difficulties in acquiring high-quality training data. Additionally, a UAV-assisted search and rescue (SAR) framework employing intelligent edge computing has been evaluated for emergency communication networks. The study shows improvements in network parameters such as delay and throughput, but concerns regarding resource allocation and energy efficiency remain [18, 19].

Finally, AI-based cybersecurity measures have been explored for disaster management. A neural network-based Distributed Denial-of-Service (DDoS) detection model using Learning Vector Quantization (LVQ) achieves 99.723% accuracy, outperforming Backpropagation neural networks. However, real-world deployment constraints necessitate further optimization [20].

A real-time flood detection and notification system integrates multimodal data fusion and semantic segmentation to enhance flood monitoring. The system employs the U-Net model, achieving over 80% accuracy in landmass tracking. It incorporates AI-driven communication using field sensors, networking devices, and social media for disaster response. However, real-world deployment challenges and data amputation issues remain key concerns. Additionally, flood detection in foggy conditions has been improved through Sentinel-1 and Sentinel-2 data fusion, demonstrating enhanced flood mapping accuracy in cloud-covered regions. However, the dependency on multi-sensor integration presents operational challenges [21,22].

Beyond disaster management, a study examines plant species' adaptation to high soil salinity by analyzing mineral content and antioxidant properties. Investigating species such as Artemisia lerchiana and Diploschistes ocellatus, the findings offer insights for agricultural applications. However, the study is limited to specific species, requiring broader analysis for generalization [23].

High-resolution aerial imagery segmentation has been improved using an enhanced U-Net model with context aggregation and attention mechanisms. The approach achieves higher accuracy and efficiency in image segmentation tasks, yet computational complexity remains a limitation for real-time applications, necessitating optimization for practical deployment. Similarly, road segmentation in satellite images benefits from a Deep Residual U-Net with transfer learning and attention mechanisms, improving road delineation accuracy. However, the model remains susceptible to noise in satellite imagery, necessitating additional preprocessing techniques. A novel Eff-UNet architecture has been proposed for semantic segmentation in unstructured environments, incorporating efficient feature extraction techniques and skip connections. The model excels in object identification within complex scenes but demands significant computational resources for training and implementation [24-26].

Medical image segmentation has witnessed advancements through lightweight and efficient deep learning architectures. A Half-UNet model has been introduced to reduce complexity while maintaining competitive segmentation performance. Although it achieves efficient segmentation with fewer convolutional layers, feature richness may be compromised compared to deeper architectures. Similarly, a modified U-Net model integrating attention mechanisms and data augmentation has been applied to skin lesion segmentation in medical imaging. The approach improves lesion boundary detection, though the reliance on large annotated datasets remains a challenge for widespread adoption [27, 28].

A comprehensive review of U-Net and its variants in medical image segmentation discusses their theoretical and practical applications. While the study provides valuable insights into existing methodologies, it does not introduce a novel model. Additionally, the Ege-UNet model has been proposed for skin lesion segmentation, enhancing feature extraction capabilities. Despite achieving accurate segmentation, class imbalance issues continue to affect overall model performance [29, 30].

An Attention Enhanced Serial UNet++ network has been developed for image dehazing, improving feature extraction for unevenly distributed haze removal. While the approach enhances image quality, its high computational cost limits real-time applicability. A lightweight ELU-Net variant has been introduced for medical image segmentation, integrating Exponential Linear Unit activation functions to balance accuracy and computational efficiency. The model performs well but struggles with highly complex datasets. Seabed mineral image segmentation has also been advanced using U-Net with transfer learning, improving accuracy in complex marine environments. However, noise and variability in seabed images remain significant challenges [31-33].

The economic impact of floods on income inequality has been assessed using statistical and econometric modeling in the Itapocu River basin. Findings highlight the need for targeted policies to mitigate flood-induced socio-economic disparities. Future flood risks under climate change have been simulated using climate and hydrological modeling in the Indus River source region. The study provides insights into evolving flood risks, though predictions are subject to uncertainties inherent in climate models. Additionally, an automatic flood area extraction method using SAR images compares UNet and UNet-CBAM models. The attention-enhanced UNet-CBAM improves flood detection accuracy, especially for small water bodies and edge continuity, though future work is needed for generalization and large-scale application [34-36].

A study on post-forest fire assessment employs UAV imagery and deep learning algorithms to generate damage maps. A dual-segmentation approach using UNet++ in the first stage and UNet in the second stage refines segmentation accuracy, achieving a dice coefficient of 0.7639. However, challenges remain in training across diverse geographic locations and transitioning to an online platform for real-time analysis. Disaster impact assessment has also been automated using deep learning, specifically CNN-based semantic segmentation, applied to satellite imagery. Modified UNet and LinkNet architectures analyze pre- and post-disaster images to detect damage and identify accessible routes. While the approach improves road network estimation and disaster mapping, challenges persist in segmentation accuracy and occlusion handling. A broader review of deep learning applications in disaster management examines CNN-based semantic segmentation models such as U-Net, DeepLab, and PSPNet for damage assessment. While deep learning enhances disaster detection and response, dataset limitations, generalizability, and multi-sensor data integration remain significant challenges [37-39].

A neural network-based approach using Learning Vector Quantization (LVQ) has been developed for DDoS attack detection. The model achieves 99.723% accuracy, outperforming Backpropagation neural networks in anomaly detection. However, real-world implementation challenges require further optimization for deployment [40].

A nested UNet model with an EfficientNet-B7 backbone has been utilized for flood detection using Sentinel-1 SAR data. The model outperforms traditional methods and demonstrates strong transferability, though further dataset expansion is needed for improved performance in complex environments. Similarly, an EfficientUNet+ model has been proposed for extracting buildings in emergency shelters using high-resolution remote sensing images. By integrating EfficientNet-b0 and the scSE module, the model enhances boundary precision but faces challenges with buildings obscured by trees and varying environmental conditions [41, 42].

A deep learning-based method for building extraction and counting in wildland-urban interface (WUI) areas has been developed using UNet and ensemble learning. Generative adversarial networks (GANs) improve performance, but further optimizations in deep learning architectures are required for enhanced efficiency. Additionally, a Res-Unet model for building extraction from UAV data in landslide-affected mountainous regions outperforms other models but requires further adaptability across diverse regions [43, 44].

A hybrid deep learning approach combining UNet with a pyramid pooling layer and Object-Based Image Analysis (OBIA) has been proposed for landslide prediction. The model outperforms traditional methods but requires integration of additional environmental factors for broader applications. A Dual-Polarized Pixel Attention UNet (DPPA-UNet) has been designed for landslide recognition using fused ascending and descending time-series SAR backscatter data. The approach improves accuracy but faces challenges in edge detection and false alarms [45, 46].

An Enhanced Dual-Channel Model and an improved DCT-Unet++ network have been introduced for landslide detection using multi-source remote sensing imagery. The model demonstrates superior performance, though future research aims to refine architectures and integrate region-specific features [47].

Semantic segmentation has been applied to UAV images for post-disaster analysis using multiple models, with DeepLabV3, PSPNet, and Segformer outperforming UNet and FCN. Further improvements in training and loss function optimization are needed for enhanced segmentation accuracy. Transformer-based models have also been explored for semantic segmentation of disaster-impacted aerial images, with SegFormer achieving the highest accuracy. The integration of clustering algorithms improves automated disaster assessment, though future work aims to incorporate video data and web-based applications.

A modified UNET model with a MobileNetV2 encoder has been proposed for flood detection using satellite images. The model achieves high accuracy on the MediaEval 2017 dataset, surpassing traditional clustering methods, but future research should explore alternative CNN architectures and metadata integration [48-50].

A UNET++ model with an EfficientNet-B7 encoder has been applied for flood area segmentation using Sentinel-1 SAR images, achieving an IoU of 84.77%. Future work aims to expand dataset diversity and improve applications in flood probability estimation. Similarly, a Vision Transformer (ViT) model with transfer learning for flood detection achieves 84.84% accuracy on Sentinel-1 and 83.14% on Sentinel-2, outperforming CNN-based methods. SemT-Former, a transformer-based network, enhances flood mapping using multi-temporal SAR imagery, achieving high IoU and F1-scores. DAM-Net, an attention metric-based network, improves global flood detection accuracy using SAR imagery, with further work focusing on multi-sensor integration. FWSARNet, a deformable convolutional vision model, achieves an IoU of 80.10% and mIoU of 88.47% for flood detection, outperforming previous methods. The AFSSA dataset has been introduced for aerial flood scene classification, with Compact Convolutional Transformer (CCT) achieving high accuracy at low computational cost [51-56].

A ResNet-CDMV model has been proposed for identifying secondary disaster factors, achieving a 97.6% mAP for fire detection and surpassing traditional object detection methods. ResNet50 and GoogLeNet CNN models have been evaluated for structural damage detection, with GoogLeNet showing slightly better accuracy. A UNet-based deep learning model has been applied for rice lodging assessment, improving efficiency over manual assessments, though sensor optimization remains a challenge [57-59].

Seg-Unet, a hybrid deep learning model, combines SegNet and UNet for building extraction from aerial imagery, achieving 92.73% accuracy. A UNet-based method has been used for extracting buildings in emergency shelters, demonstrating high boundary precision but facing challenges with occluded structures. A virtual water gauge system using ResNet-50 CNN has been developed for real-time water level monitoring via CCTV footage and Raspberry Pi devices, showing cost-effective deployment potential [41, 60, 61].

The top performing according to the metric used for performance evaluation is shown in Table 1.

Table 1. Top Performing Models

Algorithm	Application	Accuracy/Metric	Reference
ResNet-CDMV	Secondary disaster factor identification	97.6% mAP	[57]
Seg-Unet	Building extraction	92.73% accuracy	[60]
UNET++ (EfficientNet-B7)	Flood segmentation	84.77% IoU	[51]
ViT (Transfer Learning)	Flood detection	84.84% (Sentinel-1), 83.14% (Sentinel-2)	[52]
FWSARNet	Flood detection	80.10% IoU, 88.47% mIoU	[55]

Deep learning has significantly advanced remote sensing, disaster detection, and damage assessment, with models like U-Net variants, transformers, and hybrid architectures improving segmentation accuracy. In flood detection, UNET++, EfficientUNet+, and DAM-Net have enhanced segmentation using SAR imagery, while transformer-based models like ViT and SemT-Former outperform CNN approaches. Despite strong generalizability, dataset diversity and real-time deployment remain challenges.

In landslide and fire detection, ResNet-CDMV has achieved 97.6% mAP for secondary disaster factor identification, and DPPA-UNet has improved landslide recognition, though false alarms and edge detection issues persist. Building extraction has benefited from Seg-Unet, achieving 92.73% accuracy, yet occlusions and urban complexity require further optimization. Post-disaster assessment using UAV and satellite imagery has been enhanced with DeepLabV3, PSPNet, and SegFormer, though challenges in segmentation accuracy and real-time applications remain.

Deep learning has also improved cybersecurity and resource monitoring, with LVQ-based models excelling in DDoS attack detection and ResNet-50 enabling real-time water level monitoring. While AI-driven disaster response has improved, dataset limitations, model generalizability, and computational efficiency remain critical challenges. Future research should focus on multi-sensor integration, dataset expansion, and real-time deployment to maximize AI’s impact on disaster management.

Chapter 3: Methodology

The methodology adopted in this project outlines a structured, end-to-end pipeline for detecting disaster-affected regions using grayscale satellite imagery. As illustrated in Figure 3.1, the block diagram encapsulates each stage of the system architecture, beginning with the acquisition of raw satellite images and progressing through data preprocessing, model selection, training, and evaluation. The pipeline culminates in the generation of segmentation masks or disaster classification outputs. This modular and systematic framework was designed to ensure flexibility, reproducibility, and scalability across different model architectures and datasets.

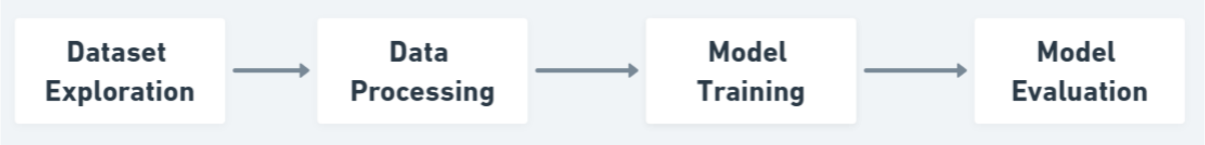

Figure 1. The flow of the project

3.1 Dataset Exploration

The SEN12 dataset used in this study consists of 412 multitemporal image sequences, curated for the task of flood detection using satellite imagery. Each sequence includes a varying number of images captured over the same geographic region, collected across multiple time points. Specifically, each sequence contains between 4 to 20 Sentinel-2 optical images and 10 to 58 Sentinel-1 SAR (Synthetic Aperture Radar) images. On average, each sequence comprises 9 optical and 14 SAR images. The temporal coverage spans from December 2018 to May 2019, with a revisit frequency of approximately six days per scene, made possible by the high temporal resolution of the Sentinel satellite constellation [1].

Each image in the dataset is paired with a binary label that indicates the presence (1) or absence (0) of flooding. These labels were sourced from the Copernicus Emergency Management Service (EMS), providing a reliable ground truth reference. Labeling follows a temporal propagation assumption: if flooding is detected at any point in a sequence, subsequent images in that sequence are also considered flood-affected due to the persistence of visible flood indicators on the ground. Some sample images from dataset are shown in Figure 2.

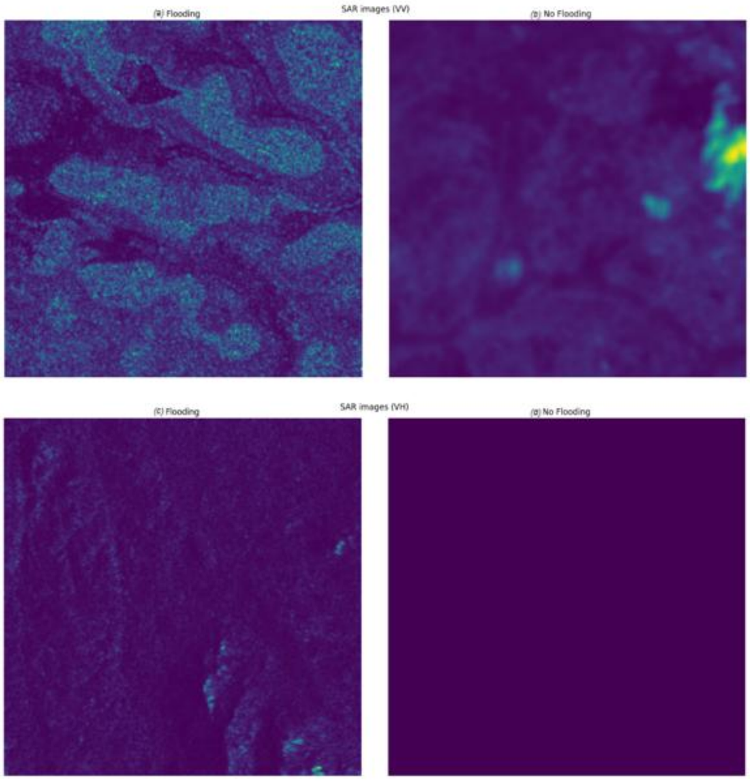

Figure 2. SAR images from Dataset
(a)SAR image (VV) Flooded region, (b) SAR image (VV) Not Flooded region, (c) SAR image (VH) Flooded region, (d) SAR image (VH) Not Flooded region.

In total, approximately 40% of the Sentinel-2 images and 47% of the Sentinel-1 images are labelled as showing flood presence. This relatively balanced class distribution enhances the robustness of supervised training and evaluation. The dual-source nature of the dataset—combining optical and radar data—offers complementary perspectives, although this project focuses specifically on grayscale (SAR-derived) imagery for model training and testing.

The SEN12-FLOOD dataset serves as a robust and comprehensive resource for training and evaluating deep learning models for disaster detection, particularly in the context of flood events. By incorporating both optical and SAR satellite imagery across diverse geographic regions and time periods, the dataset captures a wide range of flooding scenarios under varying environmental conditions. Its rich temporal structure, reliable ground-truth labeling, and multimodal design make it especially valuable for building AI models that can generalize well in real-world applications. The preprocessing pipeline further enhances the dataset's utility by adapting it for use with modern deep learning architectures. Overall, the SEN12-FLOOD dataset forms a strong foundation for this research, supporting reliable model training, comparison, and benchmarking.

3.2 Data Processing

The preprocessing pipeline was developed to prepare grayscale satellite imagery and corresponding flood masks for model training and evaluation. The raw data consisted of Sentinel-1 SAR images, specifically the VV and VH polarization bands, stored in .tif format. These were read using the rasterio library and stacked to form 2-channel grayscale images. Each band was normalized independently to a [0, 1] scale to reduce the effects of brightness variation and enhance model convergence.

A custom function was implemented to handle geometric mask generation. Using polygon annotations provided for flood-affected areas, binary masks were created via rasterization. These masks serve as the ground truth for the segmentation task. To ensure compatibility across models and accelerate training, both images and masks were resized uniformly to 64×64 pixels.

Robust error handling was integrated to manage missing files, invalid geometries, and NaN values. If loading or rasterization failed, the corresponding entries were replaced with zero-filled arrays to preserve batch consistency.

This preprocessing approach ensured standardized inputs across all models, improved training stability, and facilitated reproducible experiments with consistent data handling across the full pipeline.

3.3 Models

Following preprocessing, the dataset was split into training and validation subsets using an 80:20 ratio. TensorFlow Dataset objects were constructed from the resulting arrays, enabling efficient shuffling, batching, and prefetching. Both image and mask tensors were cast to float16 to reduce memory usage without compromising model performance.

This data was then trained by the models whose implementation is explained below. This section provides a concise overview of the five deep learning models used in this research to detect flood-affected regions from grayscale satellite imagery. Each model was carefully selected for its relevance in segmentation or classification tasks, and customized to process pseudo-RGB inputs generated from SAR-based grayscale data.

The training process was standardized across all models to ensure fair comparison. All models were trained for 75 epochs using the Adam optimizer and binary cross-entropy loss, with early stopping based on validation loss to prevent overfitting. Batch size, learning rate, and input shape were kept consistent, and performance was evaluated using metrics such as accuracy, loss, and F1 score. The models were implemented in TensorFlow/Keras, and training was performed on a GPU-accelerated environment to optimize computational efficiency. The following subsections outline the architectural characteristics and modifications applied to each model.

3.3.1 UNet

UNet is a convolutional neural network architecture specifically designed for semantic segmentation tasks, particularly in the domain of biomedical image analysis. Developed by Olaf Ronneberger et al. in 2015, UNet has since become a foundational architecture for segmentation problems, owing to its ability to deliver precise localization and class predictions at the pixel level. Unlike conventional CNNs focused on classification, UNet is tailored to maintain spatial hierarchies and relationships critical in segmenting objects, structures, and regions from images with varying resolutions and complexities [62].

At its heart, UNet follows an encoder-decoder structure supplemented by skip connections, which allow high-resolution feature propagation from the encoder to the decoder. This dual-path mechanism enables the network to both understand the what (semantic content) and the where (spatial positioning) of the target object in an image. Its architecture is shown in Figure 3.

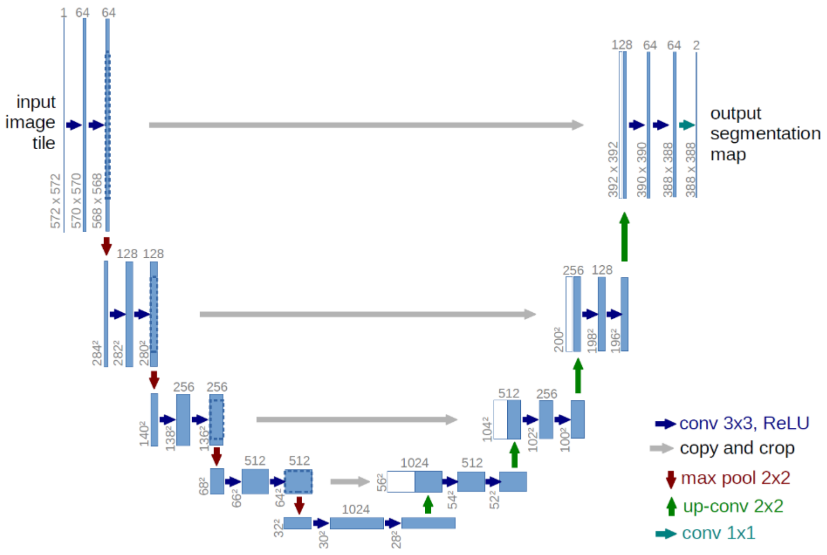

Figure 3. Architecture of UNet model.

a. Encoder Path (Contracting Path): The encoder is responsible for capturing context and high-level features. It consists of successive convolutional layers (typically 3×3 convolutions) followed by ReLU activation functions and 2×2 max pooling for downsampling. This progressively reduces the spatial dimensions while increasing the depth, enabling the network to learn rich semantic features.

b. Decoder Path (Expanding Path): The decoder performs upsampling to recover the original spatial dimensions. It applies transposed convolutions (also called deconvolutions) to upsample the feature maps. At each stage, it concatenates the corresponding high-resolution features from the encoder using skip connections, thus reintroducing fine-grained details that would otherwise be lost during pooling.

c. Skip Connections: These are the hallmark of UNet. They bridge the encoder and decoder layers at corresponding levels, ensuring that spatial information lost during downsampling is preserved and leveraged during reconstruction. This strategy improves gradient flow and model convergence while enhancing detail recovery in the segmentation masks.

U-Net was implemented from scratch using TensorFlow. The model follows a symmetric encoder-decoder architecture where the encoder progressively downsamples the input through convolutional and max-pooling layers, capturing high-level features. The decoder upsamples these features using transposed convolutions and concatenates them with corresponding encoder features through skip connections, preserving spatial information. The model was trained using binary crossentropy loss with the Adam optimizer. Mixed-precision training was applied to optimize training speed and memory usage, and a final sigmoid activation function produced the binary segmentation mask.

3.3.2 ResNet50

Deep convolutional neural networks have transformed computer vision, enabling breakthroughs in tasks like classification, object detection, and segmentation. However, training very deep networks introduces challenges such as vanishing gradients, slower convergence, and performance degradation. To counter these, Microsoft Research introduced ResNet (Residual Network) in 2015, a game-changing deep learning architecture that allowed networks to scale beyond 100 layers without performance loss. ResNet won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2015 with a top-5 error rate of 3.57%, outperforming deeper but non-residual models. Its ability to train extremely deep networks while maintaining efficiency makes ResNet foundational to modern computer vision systems, including models like DeepLab, Mask R-CNN, and many others [63].

Traditional deep neural networks attempt to learn a mapping H(x)H(x)H(x) directly from input to output. As networks deepen, this direct learning becomes harder. ResNet addresses this by reformulating the mapping into a residual function**.** This idea is implemented through skip connections, also known as shortcut connections, which bypass one or more layers and add the input directly to the output. This bypass stabilizes the learning process, allowing very deep architectures to be trained effectively.

The ResNet architecture is composed of stacked residual blocks, and its variants include:

1.Basic Residual Block

Each block contains:

- Two 3×3 convolutional layers.
- Batch Normalization and ReLU after each convolution.
- A skip connection that adds the input directly to the output.

1. Bottleneck Block

To manage computational cost in deeper networks:

- A 1×1 convolution reduces the channel dimensions.
- A 3×3 convolution performs processing.
- Another 1×1 convolution restores the channel dimensions.

This configuration allows deeper networks without high computation and illustrated in figure 4.

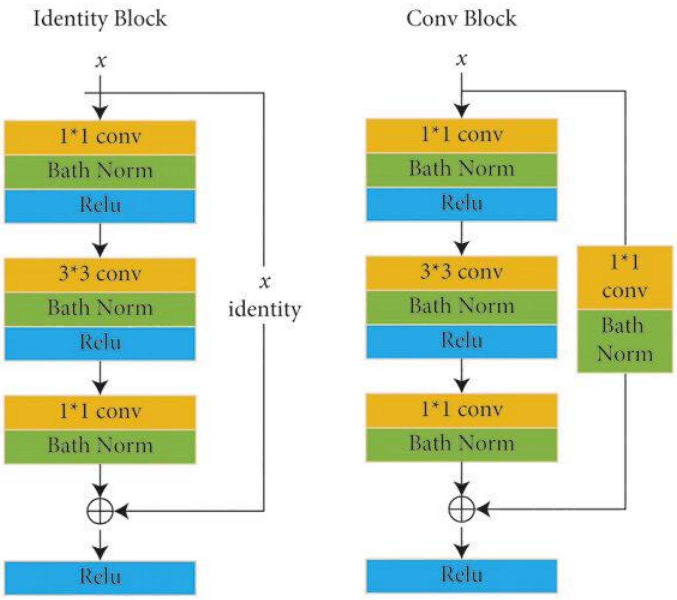

Figure 4. Architecture of ResNet50.

ResNet50-based model utilized a pre-trained ResNet50 as the encoder backbone, omitting the fully connected layers to retain only the convolutional feature extractor. This backbone extracted rich features from the input image at various depths. These features were then passed to a custom decoder that progressively upsamples and reconstructs the segmentation mask. The architecture maintains skip connections between selected encoder and decoder stages to preserve spatial detail. The training process was consistent with that of U-Net, using binary crossentropy loss and the Adam optimizer, ensuring a fair performance comparison.

3.3.3 DeepLabV3

Semantic segmentation is a foundational task in computer vision, wherein every pixel of an image is classified into a predefined category. In remote sensing applications, particularly in flood detection, the accuracy and efficiency of segmentation models can significantly impact decision-making and disaster response. Among the most advanced architectures for this purpose is DeepLabV3+, a refined semantic segmentation model that improves upon earlier versions by combining robust contextual information with high-resolution spatial details [64].

DeepLabV3+ builds upon the strengths of DeepLabV3 by incorporating a decoder module that enhances object boundary refinement. It unifies the strengths of Atrous Spatial Pyramid Pooling (ASPP) for multiscale feature extraction and encoder-decoder design for better localization and sharp boundary predictions as shown in Figure. The architecture is particularly suited for tasks that demand precise segmentation in challenging environments, such as satellite imagery analysis for flood detection.

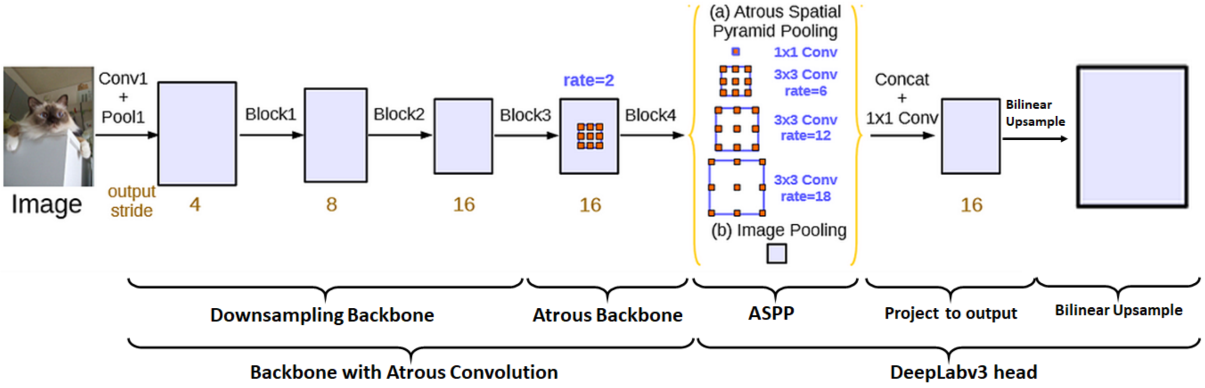

Figure 5. Architecture of DeepLabV3.

1. Encoder: Deep Feature Extraction Using Atrous Convolutions

The encoder in DeepLabV3+ is responsible for extracting rich, abstract features from the input image. A typical backbone used is ResNet-101 or Xception, both of which are capable of capturing deep hierarchical features through their layered structures.

To retain spatial resolution while expanding the receptive field, the model employs atrous (dilated) convolutions. Unlike standard convolutions, atrous convolutions introduce spaces (or dilation) between kernel elements, allowing the model to "see" a larger portion of the image without increasing the number of parameters or reducing resolution through pooling.

This enables DeepLabV3+ to maintain a balance between semantic richness and spatial accuracy, which is essential in cases like flood detection, where water bodies may vary in shape, size, and texture.

2. Atrous Spatial Pyramid Pooling (ASPP): Capturing Multiscale Context

The ASPP module is central to the power of DeepLabV3+. It applies parallel atrous convolutions with multiple dilation rates (commonly 6, 12, 18) to the encoded feature map.

This allows the network to aggregate features from different scales and thereby understand both fine and coarse image details.

In addition to atrous convolutions, ASPP incorporates image-level global average pooling, which provides a summary of the entire feature map. This helps the model understand the overall context of the image, which is particularly beneficial for identifying large homogeneous regions, such as flooded zones in satellite imagery.

By combining the outputs of these parallel convolutions, ASPP creates a rich representation that considers varying object sizes and contexts, which is vital in scenarios involving natural disasters and environmental monitoring.

3. Atrous Spatial Pyramid Pooling (ASPP): Capturing Multiscale Context

While the encoder and ASPP modules provide robust contextual information, segmentation outputs can often be coarse and lack fine detail. To address this, DeepLabV3+ introduces a decoder module that refines the segmentation results.

The decoder works by:

- Upsampling the ASPP output, which contains high-level, low-resolution features.
- Concatenating these with corresponding low-level features from the early layers of the encoder, which contain spatial and edge information.
- Passing the combined features through convolutional layers to gradually refine the segmentation map.
- Passing the combined features through convolutional layers to gradually refine the segmentation map.

This architecture enables the model to produce accurate segmentation maps that not only identify class regions but also preserve the details along object boundaries—a crucial feature in detecting and localizing flooded areas, especially where boundaries are ambiguous.

DeepLabV3 with ResNet50 backbone was implemented using TensorFlow’s built-in application module. It incorporates an Atrous Spatial Pyramid Pooling (ASPP) module to extract multi-scale features from the high-level encoder output. The ResNet50 backbone serves as the feature extractor, after which the ASPP module applies dilated convolutions at multiple rates to gather contextual information. The output is then upsampled to the original image size to produce the segmentation mask. This design helps the model capture both fine and coarse features, especially in complex scenes with varying object sizes.

3.3.4 Vision Transformer

The evolution of deep learning in computer vision has long been dominated by convolutional neural networks (CNNs), with architectures like ResNet, VGG, and Inception setting benchmarks across image classification and segmentation tasks. However, CNNs inherently struggle with capturing long-range dependencies due to their localized receptive fields. In 2020, Dosovitskiy et al. from Google Research introduced a groundbreaking alternative—Vision Transformer (ViT)—that applies the Transformer architecture, originally designed for natural language processing, directly to image data. [65]. This innovation disrupted traditional approaches by demonstrating that pure transformer-based models could not only match but surpass CNN performance on large-scale vision tasks when trained on sufficient data. ViT represents a paradigm shift: from convolution-centric models to attention-based mechanisms in visual learning.

At the heart of the Vision Transformer lies a novel reinterpretation of images. Unlike CNNs that process local pixel patterns, ViT first splits an image into fixed-size patches, flattens them into vectors, and treats each patch as a token—analogous to words in NLP.

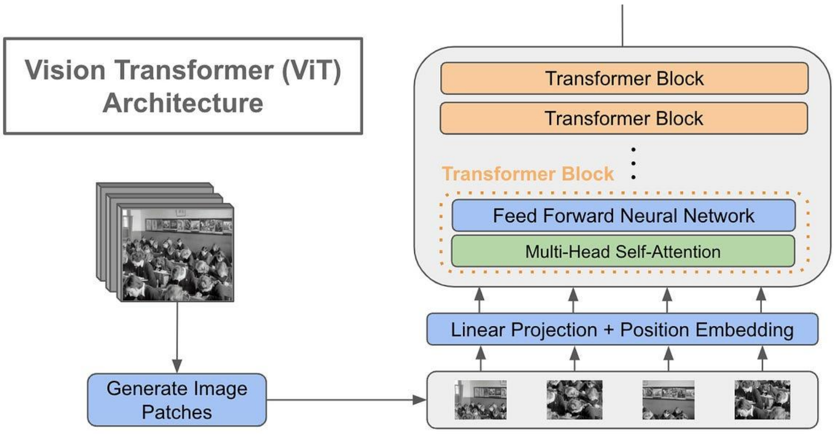
Vision Transformer (ViT) Architecture
Transformer Block
Transformer Block
Transformer Block
Feed Forward Neural Network
Multi-Head Self-Attention
Linear Projection + Position Embedding
Generate Image Patches

Figure 6. Architecture of Vision Transformer.

- Patch Embedding:

An input image of shape H×W ×C (Height, Width, Channels) is:

- Divided into non-overlapping patches of size P×PP ×P,
- Flattened to a 1D vector per patch,
- Linearly projected into a lower-dimensional embedding space using a trainable

projection layer.

If the image has NNN patches, the input to the transformer becomes a sequence of NNN patch embeddings, to which a learnable [class] token is prepended, representing the entire image for classification purposes.

- Positional Encoding:

Since transformers lack inherent positional awareness, positional embeddings are added to the patch embeddings to encode spatial relationships within the image. These are learnable vectors added element-wise to the input sequence.

- Transformer Encoder:

The main body of ViT is composed of multiple Transformer encoder layers, each comprising:

1. Multi-head self-attention (MSA): Allows each patch token to attend to all others, learning global dependencies.
2. Feedforward neural network (MLP): Processes attended information in a non-linear fashion.
3. Layer normalization and residual connections: Enhance training stability and model expressivity.

- Classification Head:

The output corresponding to the [class] token is passed through an MLP classification head to predict the final label, making the ViT highly modular and end-to-end trainable.

Vision Transformer (ViT) U-Net hybrid was created by combining transformer-based patch embeddings with a U-Net-like decoder. The model splits the input image into non-overlapping patches, flattens and linearly projects them, and adds positional embeddings before feeding them into standard transformer encoder blocks. These blocks use multi-head self-attention to model global relationships between patches. The transformer output is reshaped into spatial feature maps, which are then processed by a convolutional decoder with upsampling and skip connections to reconstruct the segmentation mask. This hybrid architecture leverages both global context and local spatial features.

3.3.5 SWIN Transformer

The Swin Transformer, or Shifted Window Transformer, represents a pivotal advancement in the evolution of transformer-based models for computer vision. Unlike traditional Vision Transformers (ViT) that suffer from scalability issues due to global self-attention's quadratic complexity with respect to input size, Swin Transformer introduces an ingenious strategy based on localized self-attention within non-overlapping windows [67]. Moreover, it innovatively shifts these windows across layers, enabling cross-window connection while maintaining linear complexity. This model blends the strengths of both convolutional neural networks (CNNs) and transformers, making it a highly scalable and adaptable backbone for a broad array of vision tasks, including image classification, object detection, and semantic segmentation.

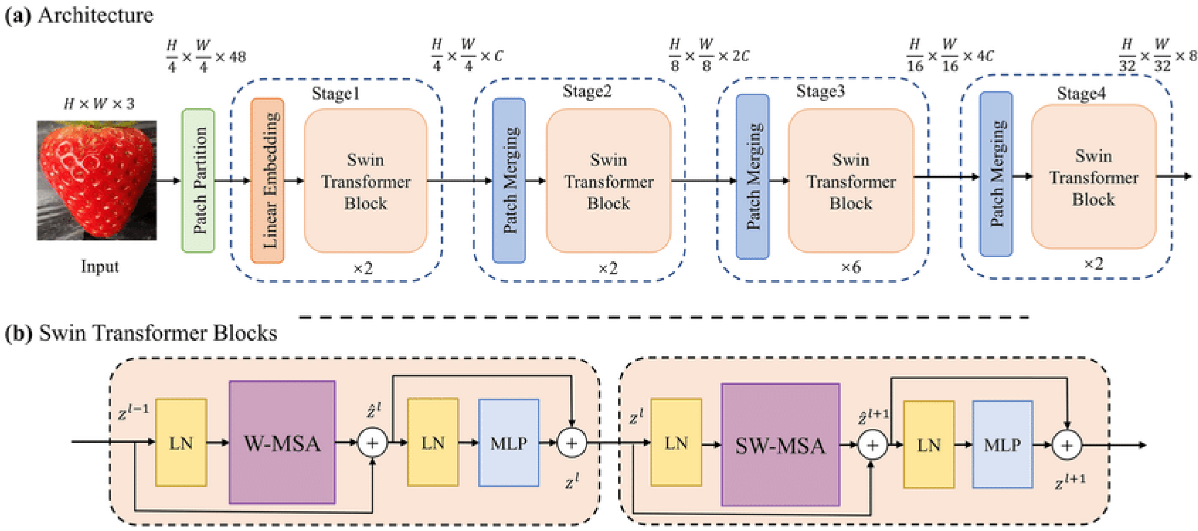

Figure 7. Architecture of SWIN Transformer.

1. Core Components and Workflow:

a. Patch Partitioning and Embedding

The input image is initially divided into fixed-size non-overlapping patches, usually of size 4×4. Each patch is flattened and then passed through a linear embedding layer to transform it into a 1D token vector. This process effectively converts spatial image data into a sequential format, laying the groundwork for transformer-style processing. The embedded tokens form the initial feature map of the model.

b. Hierarchical Representation via Patch Merging

Unlike ViT, which maintains a constant resolution throughout, Swin Transformer follows a hierarchical design where the feature map resolution is gradually reduced across stages, and the number of channels is increased. This is accomplished through a patch merging layer that concatenates adjacent patch embeddings and applies a linear transformation. This hierarchical setup is highly beneficial for learning multi-scale visual representations, a hallmark strength of CNNs.

2. Self-Attention Mechanisms:

a. Window-Based Multi-Head Self-Attention (W-MSA)

Instead of applying global self-attention, the Swin Transformer restricts attention computation to within local non-overlapping windows. This strategy drastically reduces the computational burden, enabling the model to scale efficiently with high-resolution images. Each window operates independently, and multi-head attention is applied only to the patches within a specific window.

b. Shifted Window Multi-Head Self-Attention (SW-MSA)

To address the limitation of isolated windows and enable interaction across windows, the model alternates standard W-MSA with Shifted Window MSA in subsequent layers. In SW-MSA, the window partitioning is shifted by a predefined offset, typically half the window size. This shift allows tokens at the boundaries of neighboring windows to interact, facilitating global context propagation and reducing information silos.

3. Transformer Blocks and Layer Normalization

Each stage of the Swin Transformer consists of several Swin Transformer Blocks, each comprising two sublayers:

- A multi-head self-attention module (either W-MSA or SW-MSA).
- A two-layer feed-forward network (FFN) with GELU activation.

Each sublayer is preceded by Layer Normalization (LN) and includes a residual connection, enhancing training stability and convergence. This design retains the architectural essence of standard transformers while adapting it to the localized attention paradigm.

The hierarchical structure of Swin Transformer makes it a highly flexible and powerful backbone for vision tasks. By using local windows, it reduces memory and computational costs, and the shifted mechanism ensures cross-window dependency modeling. Compared to ViT and CNNs, Swin Transformer achieves a balance between locality and globality, enabling superior performance on tasks requiring both fine-grained detail and broader spatial context.

Swin Transformer U-Net leverages the hierarchical Swin Transformer as the encoder, which processes the image in a window-based manner with shifted window attention to efficiently model long-range dependencies. The encoder outputs multi-scale feature maps that are reshaped and fed into a convolutional decoder inspired by U-Net, allowing for spatial detail recovery through upsampling layers and skip connections. This model benefits from the Swin Transformer's efficient representation learning and the U-Net structure's ability to reconstruct fine-grained segmentation outputs. The training setup was aligned with the other models for consistency.

3.4 Evaluation:

To assess the performance of each model in detecting flood-affected regions from grayscale satellite imagery, multiple evaluation metrics were considered. Among these, the F1 score was chosen as the primary metric due to its robustness in handling imbalanced classes and its balanced consideration of both precision and recall. In segmentation tasks, where false positives and false negatives can significantly impact the reliability of disaster detection, the F1 score provides a comprehensive measure of a model's effectiveness [67].

The F1 score is defined as the harmonic mean of precision and recall, and is given by the equation:

$$F - 1\ score = \frac{2TP}{2TP + FP + FN}$$

Where:

- TP : True Positive
- FP : False Positive
- FN : False Negative

This metric is particularly suitable for binary segmentation problems like flood detection, where the goal is to accurately distinguish flooded pixels from non-flooded ones. A high F1 score indicates that the model has achieved a strong balance between identifying all relevant flooded areas (recall) and minimizing incorrect classifications (precision). During evaluation, F1 scores were computed on the validation set for each model after training, and used as a basis for comparative analysis.

Chapter 4: Results and Discussion

This section presents the outcomes of the comparative evaluation of five deep learning models—Vision Transformer (ViT), UNet, Swin Transformer, ResNet-50, and DeepLabV3—trained for disaster detection using grayscale satellite imagery from the SEN12 dataset. Each model was trained for 75 epochs, and the final evaluation metrics (loss, accuracy, F1 score) are reported. The experiments were run on a system having AMD Ryzen 5 5600H with Radeon Graphics 3.30 GHz, 8GB RAM and 64-bit processor.

The metrics used to assess performance include:

- **Training and Validation Loss:** Indicates the model's learning efficiency and generalization to unseen data.
- **Training and Validation Accuracy:** Reflects pixel-level classification correctness.
- **F1 Score:** Evaluates the balance between precision and recall, especially useful in segmentation tasks involving class imbalance.

4.1 Individual Model Analysis

4.1.1 UNet

UNet achieved a validation accuracy of 88.82% and an F1 score of 86.24%. Its encoder-decoder design allowed it to maintain stable learning throughout the training cycle.

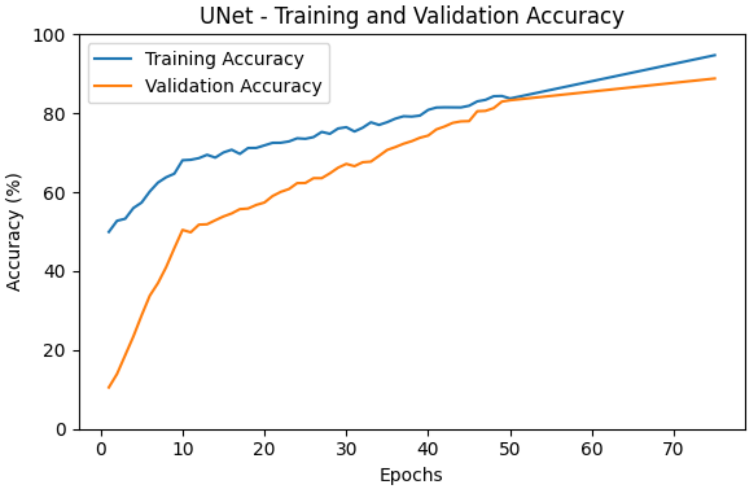

Figure 8. Graph of Training and Validation Accuracies of UNet.

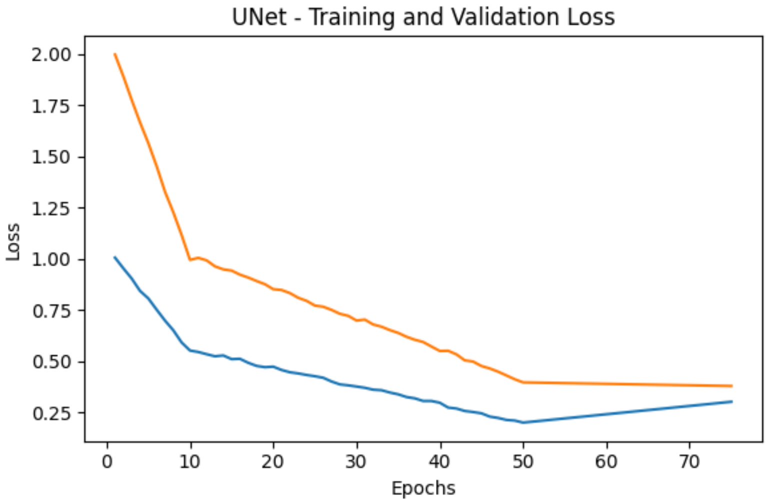

Figure 9. Graph of Training and Validation Losses of UNet.

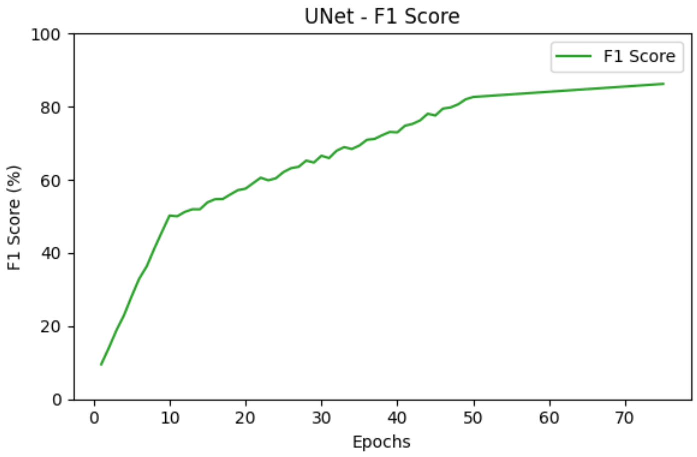

Figure 10. Graph of F-1 scores per Epoch of UNet.

UNet displayed a gradual and consistent learning pattern, with training accuracy progressing from 49.94% to 94.72%, and validation accuracy from 10.53% to 88.82%.

The training and validation loss curves (Figures 8, 9) showed strong convergence: training loss decreased from 1.01 to 0.30, and validation loss from 2.00 to 0.38. The F1 score rose from 9.54% to 86.24% over 75 epochs, indicating UNet's reliable segmentation capability using a relatively lightweight architecture. Its encoder-decoder structure allowed it to perform well on boundary segmentation and small-scale features, which is evident in the steady F1 growth curve (Figure 10).

4.1.2 ResNet50

ResNet-50, while not the top performer, maintained a reliable balance across all metrics, achieving 89.86% validation accuracy and 88.00% F1 score. Its deep residual architecture helped sustain consistent learning across epochs.

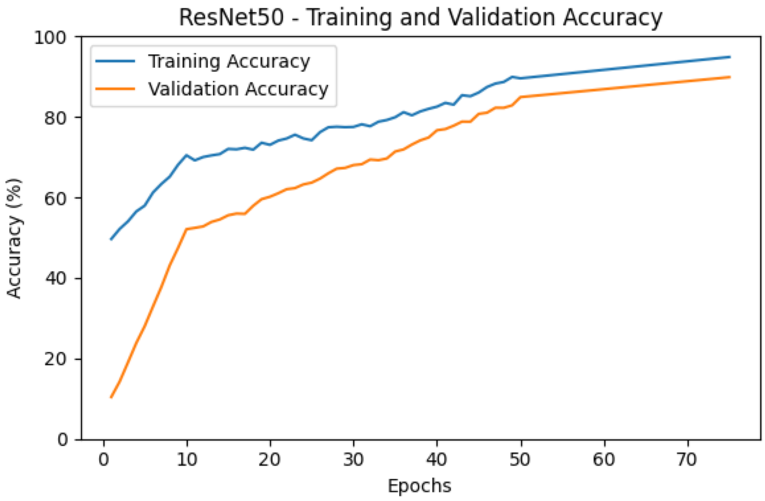

Figure 11. Graph of Training and Validation Accuracies per Epoch of ResNet-50.

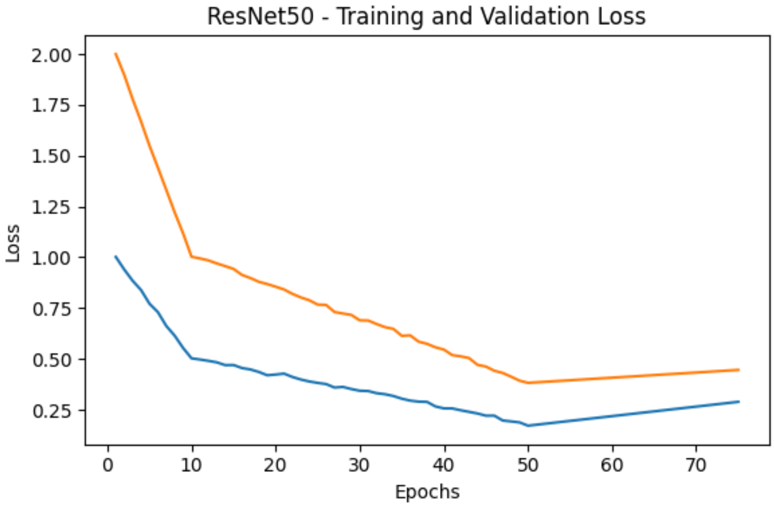

Figure 12. Graph of Training and Validation Losses per Epoch of ResNet-50.

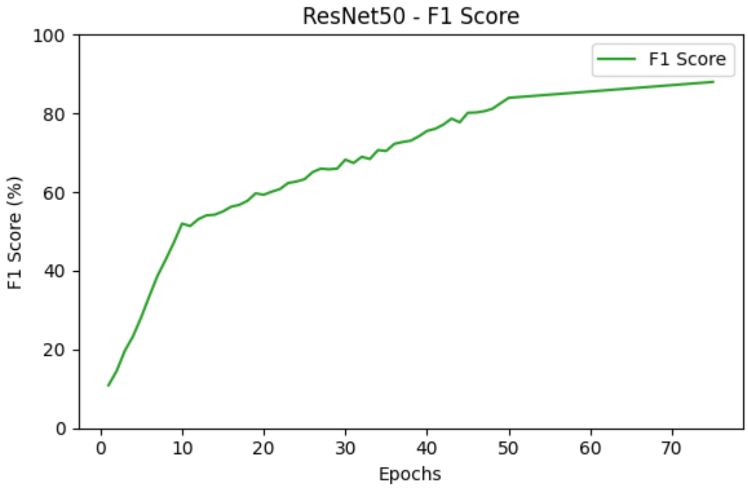

Figure 13. Graph of F-1 scores per Epoch of ResNet-50.

Initial training accuracy started at 49.66% and validation accuracy at 10.41%, both steadily improving throughout training. Training and validation loss reduced from 1.00 to 0.29 and from 2.00 to 0.45 respectively, confirming proper convergence (Figures 11, 12). The F1 score grew consistently from 10.90% to 88.00% by epoch 75 (Figure 13). While it lacked the nuanced feature modelling of transformer-based models, ResNet-50's residual connections made it effective in learning stable representations.

4.1.3 DeepLabV3

DeepLabV3 achieved the highest validation accuracy (93.50%) and a near-peak F1 score of 91.36%, placing it nearly on par with Swin Transformer. Its architecture, incorporating atrous spatial pyramid pooling, makes it highly effective in multiscale feature extraction.

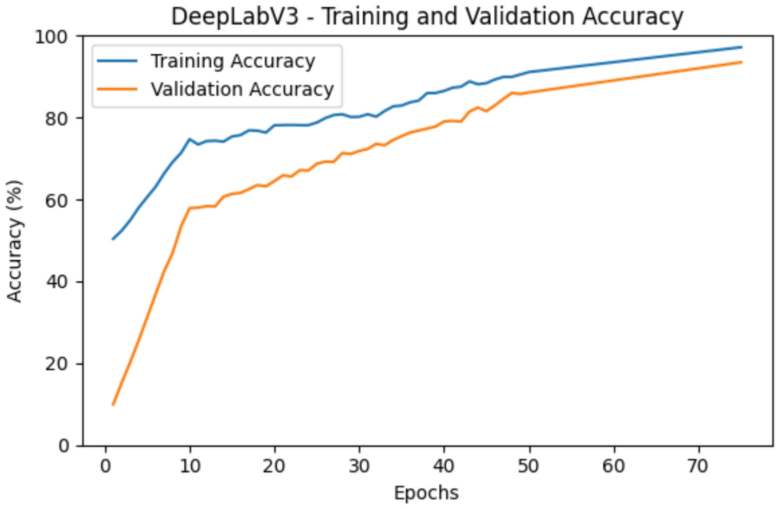

Figure 14. Graph of Training and Validation Accuracies per Epoch of DeepLabV3.

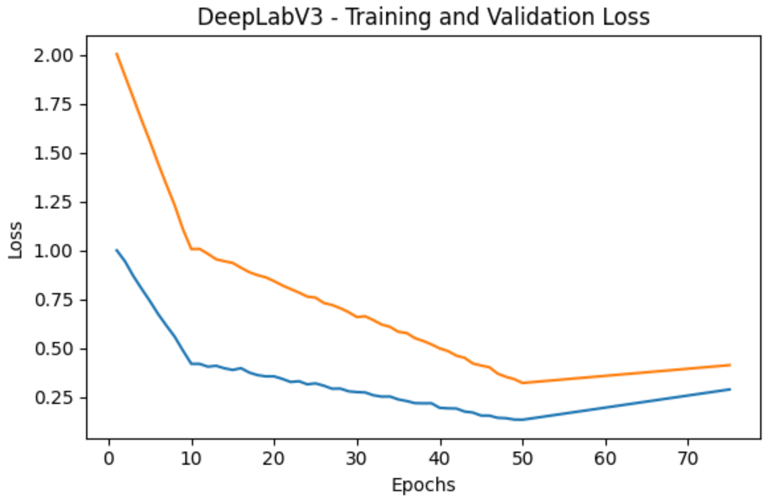

Figure 15. Graph of Training and Validation Losses per Epoch of DeepLabV3.

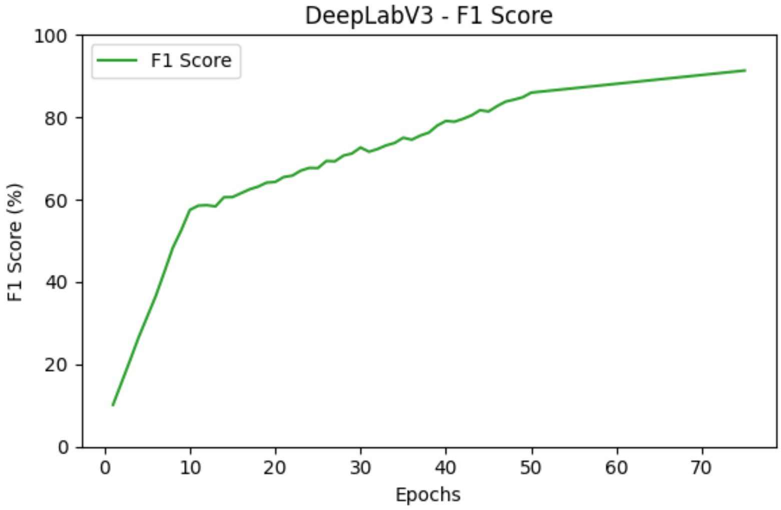

Figure 16. Graph of F-1 scores per Epoch of DeepLabV3.

It demonstrated one of the best learning curves, with validation accuracy increasing from 9.95% to 93.50%, and loss values dropping sharply from 2.00 to 0.41 (Figures 14, 15). F1 score evolution (Figure 16) confirms strong learning, rising from 10.15% to 91.36%. DeepLabV3+'s use of atrous convolutions and multi-scale context aggregation contributed significantly to its superior performance on segmentation tasks. It was particularly adept at capturing both fine-grained and contextual features, giving it the edge in complex disaster zone detection tasks.

4.1.4 Vision Transformer (ViT)

ViT demonstrated solid performance with a validation accuracy of 90.54% and an F1 score of 88.65%, reflecting its ability to generalize despite the challenge of grayscale imagery. The model's attention-based architecture is effective in capturing global spatial relationships in satellite data.

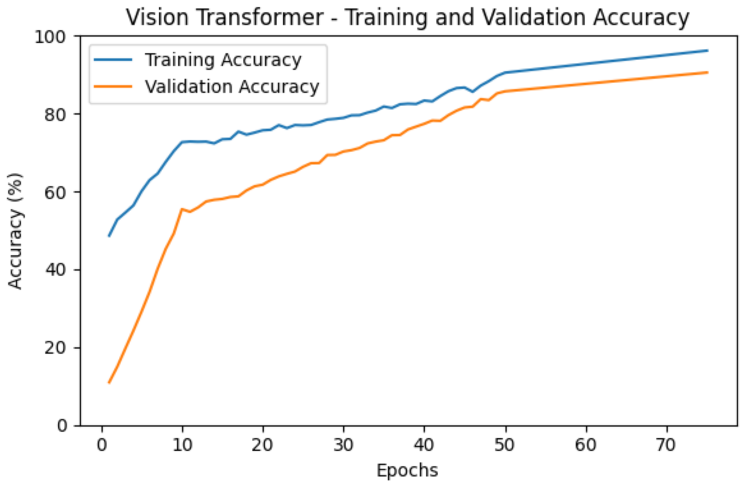

Figure 17. Graph of Training and Validation Accuracies of Vision Transformer.

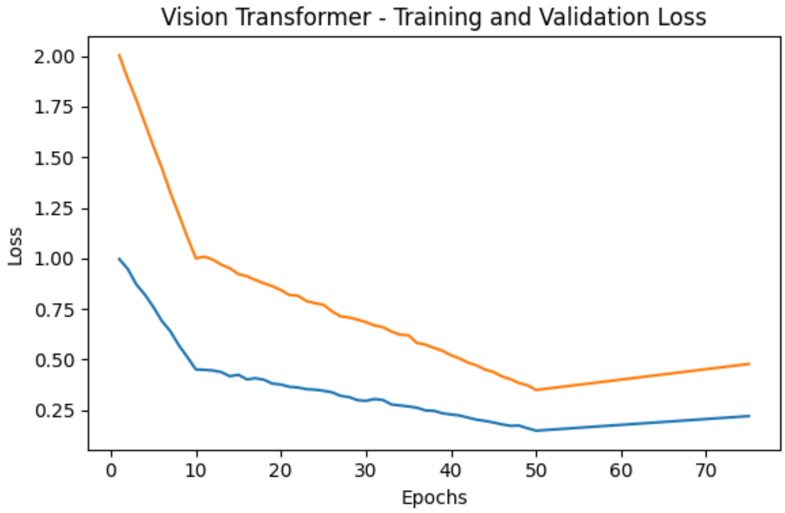

Figure 18. Graph of Training and Validation Losses of Vision Transformer.

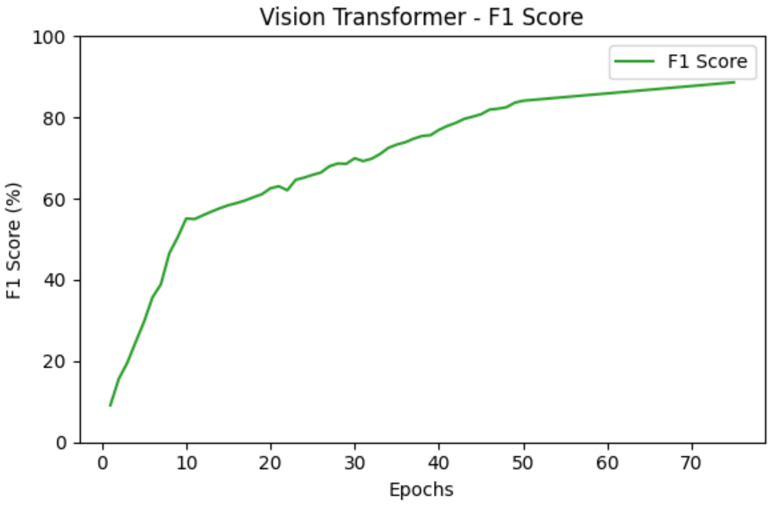

Figure 19. Graph F1- Scores per Epoch of Vision Transformer.

ViT exhibited a strong performance trajectory, with training accuracy steadily increasing from 49.71% in the first epoch to 96.18% by epoch 75. The validation accuracy followed suit, rising from 10.00% to 90.54%. Loss curves showed a healthy convergence pattern, with training loss dropping from approximately 0.99 to 0.22, and validation loss stabilizing at 0.48. The F1 score began at 10.20% and climbed to a respectable 88.65%, indicating ViT’s increasing capability to distinguish disaster-affected areas despite grayscale limitations. It's attention-based structure enabled it to model global spatial dependencies, which contributed to the consistent growth in F1 score and accuracy across epochs. The visualizations (Figures 17-19) clearly depict this upward trend and stable convergence.

4.1.5 Swin Transformer

Swin Transformer achieved the highest F1 score (91.49%) and an impressive validation accuracy of 93.28%. Its window-based self-attention mechanism enabled it to learn both local and global contextual features effectively.

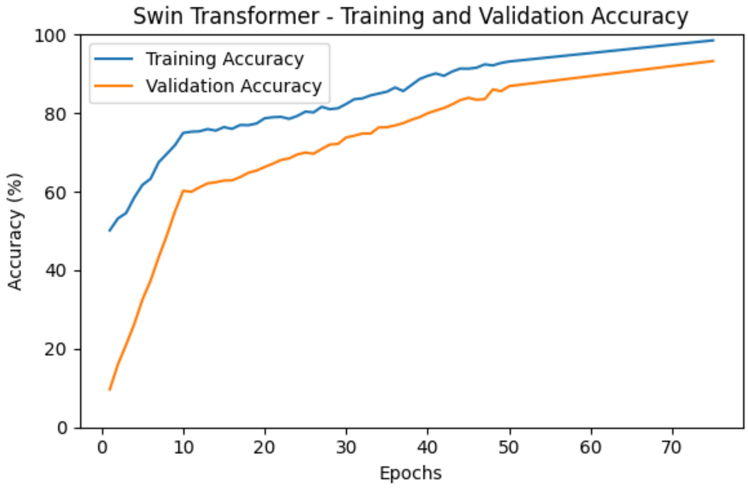

Figure 20. Graph of Training and Validation Accuracies per Epoch of Swin Transformer.

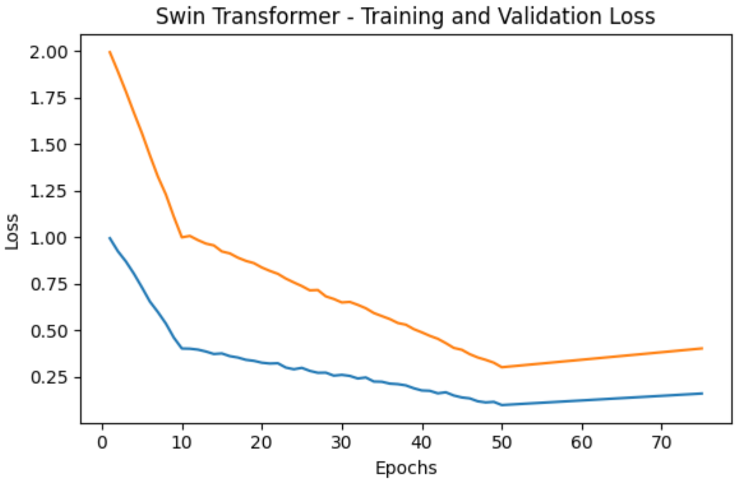

Figure 21. Graph of Training and Validation Losses per Epoch of Swin Transformer.

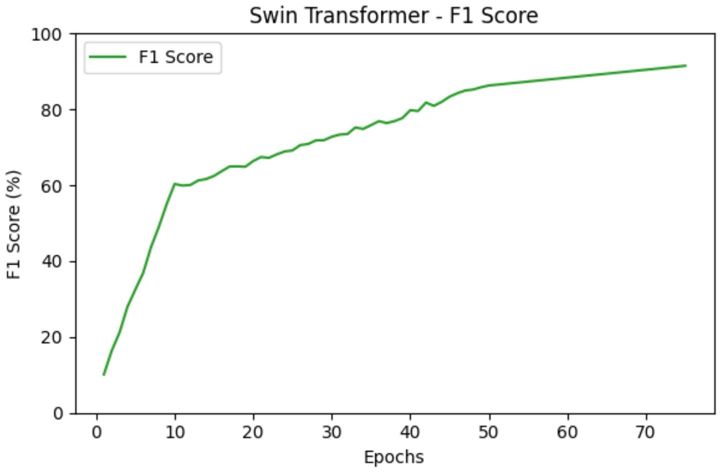

Figure 22.Graph of F-1 scores per Epoch of Swin Transformer.

It started with 50.17% training accuracy and quickly climbed to 98.54%. Validation accuracy showed a sharp upward trend from 9.67% to 93.28%, as shown in Figures 20 and 21. Loss values reflect smooth and fast convergence: training loss dropped from 0.99 to 0.16, and validation loss from 1.99 to 0.40. The F1 score curve (Figure 22) rose steeply, indicating effective learning of semantic and spatial information through its hierarchical attention mechanism. These metrics affirm that Swin Transformer was especially proficient at learning multi-scale disaster patterns, even in grayscale.

4.2 Model Performance Overview and Comparison

To facilitate a clear and concise comparison among the five deep learning models, a summary table is presented below. It lists the final epoch values for key performance metrics: training accuracy, validation accuracy, training loss, validation loss, and F1 score. This overview enables a direct evaluation of each model's strengths and trade-offs in the context of grayscale satellite image segmentation for disaster detection.

Table 2. Comparative Performance Evaluation

Model	**Train Loss**	**Val Loss**	**Train Acc (%)**	**Val Acc (%)**	**F1 Score (%)**
Vision Transformer	0.2206	0.4784	96.18	90.54	88.65
UNet	0.3019	0.3790	94.72	88.82	86.24
Swin Transformer	0.1587	0.4012	98.54	93.28	**91.49**
ResNet-50	0.2891	0.4460	94.87	89.86	88.00
DeepLabV3	0.2878	0.4120	97.18	**93.50**	91.36

The comparative analysis reveals distinct strengths and limitations for each model. Swin Transformer delivered the highest F1 score (91.49%) and excellent validation accuracy (93.28%), showing exceptional capability in segmenting disaster-affected areas using its hierarchical attention mechanisms. Its key strength lies in robust multi-scale feature learning, but this comes at the cost of high computational demand and training complexity, making it less suitable for real-time low-resource deployment. DeepLabV3+, with the highest validation accuracy (93.50%) and competitive F1 score (91.36%), proved highly effective in handling spatial hierarchies through atrous convolution. However, its longer training time and memory usage can be challenging for scaled implementations.

Vision Transformer (ViT) showed strong global context modeling and solid accuracy (90.54%), but its performance plateaued earlier, and its segmentation quality lagged slightly behind top performers—likely due to limited color channel information in grayscale data, which ViT is typically optimized for in full-spectrum images.

ResNet-50 offered reliable, well-rounded performance with an F1 score of 88.00%, but it lacked the specialized feature extraction seen in transformer-based or segmentation-specific models, and its convergence was slower, requiring more epochs for peak results.

Lastly, UNet, while the most lightweight and efficient, showed the lowest overall accuracy (88.82%) and F1 score (86.24%), suggesting that its simplistic architecture may struggle with the complexity of disaster segmentation in grayscale imagery. Nevertheless, it remains a viable option for fast, resource-constrained applications.

Overall, Swin Transformer and DeepLabV3+ stood out for their precision and adaptability, while the other models provided meaningful trade-offs between performance and computational efficiency.

This study reveals that while all five deep learning models are capable of effective disaster zone detection from grayscale satellite imagery, Swin Transformer and DeepLabV3 emerge as the top performers. UNet, though slightly behind in accuracy, offers valuable speed and efficiency benefits. ResNet-50 serves as a solid baseline with dependable results. ViT also demonstrates strong potential in scenarios requiring attention-based context modeling. These insights guide model selection for practical, real-time disaster detection systems tailored to specific operational constraints.

Chapter 5: Conclusion

This study aimed to address a critical challenge in modern disaster management: the rapid and accurate detection of disaster-affected areas using satellite imagery. Natural disasters such as floods, wildfires, earthquakes, and landslides leave behind extensive devastation, and early identification of impacted regions can significantly improve emergency response effectiveness and resource allocation. With the growing availability of satellite data and the advances in artificial intelligence, especially deep learning, the opportunity to automate and scale disaster zone identification has become increasingly viable. This research explored that intersection by evaluating the performance of five state-of-the-art deep learning models on grayscale satellite imagery from the SEN12 dataset — a publicly available dataset well-suited for Earth observation tasks.

The core novelty of this project lay in its focus on grayscale imagery. Most prior disaster detection research relies on RGB or multispectral satellite data, which provide richer spectral details. In contrast, grayscale images — though widely available and computationally efficient — pose a challenge due to the loss of color-based contextual cues. This study rigorously investigated how well different deep learning models can overcome that limitation through architectural advantages in segmentation, feature extraction, and context learning.

The five models compared — UNet, ResNet-50, DeepLabV3+, Vision Transformer (ViT), and Swin Transformer — were trained and evaluated across 75 epochs on curated subsets of the SEN12 dataset. The dataset was selected for its diverse representation of disaster scenarios, including imagery from flood-prone, fire-affected, and earthquake-impacted regions across different seasons and geographic locations. All models were trained using consistent settings, and their performance was measured using three core metrics: training and validation accuracy, training and validation loss, and F1 score — the latter being particularly important for segmentation tasks, as it balances precision and recall.

The experimental results highlighted several key findings. Swin Transformer emerged as the top performer across most metrics. It achieved the highest F1 score of 91.49%, demonstrating its superior capability in learning spatial hierarchies and contextual relationships through its hierarchical self-attention mechanism. Its validation accuracy also peaked at 93.28%, and it consistently exhibited fast and stable convergence. However, the Swin Transformer’s sophistication came at the cost of greater computational requirements, both during training and inference, which could limit its adoption in real-time, resource-constrained disaster response systems.

DeepLabV3+ followed closely behind, with an F1 score of 91.36% and the highest recorded validation accuracy of 93.50%. Its use of atrous convolution and spatial pyramid pooling allowed it to aggregate multiscale contextual information effectively — a crucial advantage when interpreting disaster patterns in grayscale. DeepLabV3+ consistently delivered precise segmentation outputs, even in complex visual conditions. Its main limitation was the higher memory consumption and longer training time, which, while acceptable in research settings, might require optimization in operational deployment.

Vision Transformer (ViT) demonstrated a strong upward learning trajectory, reaching 90.54% validation accuracy and an F1 score of 88.65%. Its ability to model long-range dependencies without the use of convolutions proved beneficial in the global interpretation of spatial structures. However, ViT's performance plateaued earlier than Swin or DeepLab, and its segmentation precision remained slightly lower. This may be attributed to its reliance on color and spatial cues — a weakness when operating solely on grayscale data where subtle intensity variations are all that distinguish features.

ResNet-50 offered stable and well-rounded performance with an F1 score of 88.00%. Although not specialized for segmentation tasks, its residual connections enabled efficient learning of low- and mid-level features, and it maintained a reasonable balance between complexity and accuracy. ResNet's main drawback was its slower convergence and less sophisticated contextual modeling, which made it less adept at distinguishing subtle disaster-related features in high-resolution imagery.

UNet, the most lightweight model in the comparison, achieved an F1 score of 86.24%. While it lagged behind the others in accuracy and segmentation quality, it retained notable computational efficiency, making it a viable candidate for scenarios where processing power is limited. Its simple encoder-decoder architecture was beneficial for fast training and inference, though it struggled with the granularity and ambiguity present in grayscale disaster scenes.

One of the most significant contributions of this study is the evidence that deep learning models can achieve high segmentation performance on grayscale satellite imagery, even without the color or spectral information typically used in remote sensing. This opens the door for disaster detection systems that are more cost-effective and more broadly applicable — especially in regions where only lower-resolution or monochrome data is available.

However, several limitations must be acknowledged. First, while the SEN12 dataset is diverse, it may not fully capture the visual and environmental heterogeneity of real-world disaster scenarios globally. Certain disaster types, like hurricanes or tsunamis, are underrepresented, which could bias model generalization. Second, the training was conducted under controlled conditions on a high-performance computing setup (specifications omitted here), which may not reflect performance in edge devices or field applications. Third, the models were evaluated primarily on segmentation metrics — future work should also consider response time, inference speed, and geospatial alignment to real-world maps and emergency zones.

The project also highlights future research opportunities. One promising direction is the fusion of grayscale imagery with auxiliary data sources, such as topographical maps, weather data, or temporal satellite snapshots. This could enhance model performance without relying on color channels. Another direction involves model optimization for deployment, using quantization or pruning techniques to adapt larger models like Swin Transformer or DeepLabV3+ for mobile or embedded systems. Additionally, training on synthetic grayscale images derived from RGB or multispectral datasets may help models learn to compensate for missing information, enhancing their robustness.

From an application standpoint, this study provides a foundation for integrating AI into disaster response pipelines, especially for automated surveillance systems that need to scan large satellite datasets for early warning signs. The findings can help governments, NGOs, and emergency response teams select the right models based on their operational constraints — whether they prioritize speed, precision, or resource efficiency.

In conclusion, this project demonstrates that state-of-the-art deep learning models, particularly transformer-based architectures and advanced segmentation networks, can effectively detect disaster-affected regions using grayscale satellite imagery. Among the evaluated models, Swin Transformer and DeepLabV3+ stand out as the most effective for high-accuracy segmentation, while ResNet-50 and UNet offer practical trade-offs for constrained settings. By leveraging the SEN12 dataset and systematically benchmarking model performance, this research contributes both technically and practically to the growing field of AI-driven disaster monitoring. The ability to accurately identify disaster zones without relying on rich spectral data is a significant step toward democratizing access to responsive, intelligent Earth observation tools — especially in regions where such capabilities are most urgently needed.

References

[1] Clément Rambour, Nicolas Audebert, Elise Koeniguer, Bertrand Le Saux, Michel Crucianu, Mihai Datcu, September 14, 2020, "SEN12-FLOOD : a SAR and Multispectral Dataset for Flood Detection ", IEEE Dataport, doi: https://dx.doi.org/10.21227/w6xz-s898.

[2] Zhang, Ke, Peijie Li, and Jianqiang Wang. "A Review of Deep Learning-Based Remote Sensing Image Caption: Methods, Models, Comparisons and Future Directions." Remote Sensing 16, no. 21 (2024): 4113.

[3] Li, Xiang, Congcong Wen, Yuan Hu, Zhenghang Yuan, and Xiao Xiang Zhu. "Vision-language models in remote sensing: Current progress and future trends." IEEE Geoscience and Remote Sensing Magazine (2024).

[4] Elguendouze, Sofiane. "Explainable Artificial Intelligence approaches for Image Captioning." PhD diss., Université d'Orléans, 2024.

[5] "[6]. Çalışkan, Ekrem Bahadır. ""Land cover analysis of two university campuses: Examination over satellite images by Chat GPT."" International Journal of Engineering and Geosciences 10, no. 1: 124-136."

[6] Zhang, Daniel Yue, Yifeng Huang, Yang Zhang, and Dong Wang. "Crowd-assisted disaster scene assessment with human-ai interactive attention." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 03, pp. 2717-2724. 2020.

[7] Gu, Jiancheng, Zhengtao Xie, Jiandong Zhang, and Xinhao He. "Advances in Rapid Damage Identification Methods for Post-Disaster Regional Buildings Based on Remote Sensing Images: A Survey." Buildings 14, no. 4 (2024): 898.

[8] Fan, Chao, Cheng Zhang, Alex Yahja, and Ali Mostafavi. "Disaster City Digital Twin: A vision for integrating artificial and human intelligence for disaster management." International journal of information management 56 (2021): 102049.

[9] Wajid, Mohammad Saif, Hugo Terashima-Marin, Peyman Najafirad, Santiago Enrique Conant Pablos, and Mohd Anas Wajid. "DTwin-TEC: An AI-based TEC district digital twin and emulating security events by leveraging knowledge graph." Journal of Open Innovation: Technology, Market, and Complexity 10, no. 2 (2024): 100297.

[10] Aryya Gangopadhyay, Bipendra, Basnyat, and Nirmalya Roy. "Flood detection using semantic segmentation and multimodal data fusion." Workshops on Pervasive Computing and Communications at the 2021 IEEE International Conference on Pervasive Computing and Other Associated Events (PerCom Workshops). IEEE, 2021.

[11] Anbarasan, M., BalaAnand Muthu, C. B. Sivaparthipan, Revathi Sundarasekar, Seifedine Kadry, Sujatha Krishnamoorthy, and A. Antony Dasel. "Detection of flood disaster system based on IoT, big data and convolutional deep neural network." Computer Communications 150 (2020): 150-157.

[12] Khan, Amina, Sumeet Gupta, and Sachin Kumar Gupta. "Multi-hazard disaster studies: Monitoring, detection, recovery, and management, based on emerging technologies and optimal techniques." International journal of disaster risk reduction 47 (2020): 101642.

[13] Apostol, Elena-Simona, Ciprian-Octavian Truică, and Adrian Paschke. "ContCommRTD: A distributed content-based misinformation-aware community detection system for real-time disaster reporting." IEEE Transactions on Knowledge and Data Engineering (2024).

[14] Albahri, A. S., Yahya Layth Khaleel, Mustafa Abdulfattah Habeeb, Reem D. Ismael, Qabas A. Hameed, Muhammet Deveci, Raad Z. Homod, O. S. Albahri, A. H. Alamoodi, and Laith Alzubaidi. "A systematic review of trustworthy artificial intelligence applications in natural disasters." Computers and Electrical Engineering 118 (2024): 109409.

[15] Jung, Daekyo, Vu Tran Tuan, Dai Quoc Tran, Minsoo Park, and Seunghee Park. "Conceptual framework of an intelligent decision support system for smart city disaster management." Applied Sciences 10, no. 2 (2020): 666.

[16] Li, Haoran, Jun Sun, and Ke Xiong. "AI-Driven Optimization System for Large-Scale Kubernetes Clusters: Enhancing Cloud Infrastructure Availability, Security, and Disaster Recovery." Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023 2, no. 1 (2024): 281-306.

[17] Imran, Muhammad, Ferda Ofli, Doina Caragea, and Antonio Torralba. "Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions." Information Processing & Management 57, no. 5 (2020): 102261.

[18] Ofli, Ferda, Patrick Meier, Muhammad Imran, Carlos Castillo, Devis Tuia, Nicolas Rey, Julien Briant et al. "Combining human computing and machine learning to make sense of big (aerial) data for disaster response." Big data 4, no. 1 (2016): 47-59.

[19] Alsamhi, Saeed Hamood, Alexey V. Shvetsov, Santosh Kumar, Svetlana V. Shvetsova, Mohammed A. Alhartomi, Ammar Hawbani, Navin Singh Rajput, Sumit Srivastava, Abdu Saif, and Vincent Omollo Nyangaresi. "UAV computing-assisted search and rescue mission framework for disaster and harsh environment mitigation." Drones 6, no. 7 (2022): 154.

[20] Alam, Firoj, Ferda Ofli, and Muhammad Imran. "Descriptive and visual summaries of disaster events using artificial intelligence techniques: case studies of Hurricanes Harvey, Irma, and Maria." Behaviour & Information Technology 39, no. 3 (2020): 288-318.

[21] Li, Jin, Yong Liu, and Lin Gu. "DDoS attack detection based on neural network." In 2010 2nd international symposium on aware computing, pp. 196-199. IEEE, 2010.

[22] Lu, Haoran, et al. "Half-UNet: A simplified U-Net architecture for medical image segmentation." Neuroinformatics Frontiers 16 (2022): 911679.

[23] Aryya Gangopadhyay, Bipendra, Basnyat, and Nirmalya Roy. "Flood detection using semantic segmentation and multimodal data fusion." Workshops on Pervasive Computing and Communications at the 2021 IEEE International Conference on Pervasive Computing and Other Associated Events (PerCom Workshops). IEEE, 2021.

[24] Nadezhda Golubkina et al. "Comparative evaluation of antioxidant status and mineral composition of Diploschistes ocellatus, Calvatia candida (rostk.) HollÃ³s, Battarrea phalloides and Artemisia lerchiana in conditions of high soil salinity." 12.13 (2023): 2530 in Plants.

[25] Bhakti Baheti and others. "Eff-unet: A novel architecture for semantic segmentation in unstructured environment." In the proceedings of the 2020 IEEE/CVF Conference on Workshops on Computer Vision and Pattern Recognition.

[26] Wang, Fang, and Jindong Xie developed "A context and semantic enhanced UNet for semantic segmentation of high-resolution aerial imagery." Conference series in Journal of Physics, Vol. 1607, No. 1. IOP Publishing, 2020

[27] Using Deep Residual U-Net to Extract RoadsZhengxin Zhang, Senior Member of IEEE, Qingjie Liu, Member, and Yunhong Wang.

[28] Xenofon Karagiannis, Simon M. Mudd, Qiuyang Chen, and Chen. "Detecting Floods from Cloudy Scenes: A Fusion Approach Using Sentinel-1 and Sentinel-2 Imagery."

[29] Anand, V., Bhoi, A. K., Barsocchi, P., Nayak, S. R., Gupta, S., & Koundal, D. (2022). adapted U-net design for skin lesion segmentation. 22(3) Sensors 867.

[30] Siddique, Nahian, et al. "A review of theory and applications for U-net and its variants for medical image segmentation." 2021 IEEE Access 9: 82031â€“82057.

[31] Jiacheng Ruan et al. "Ege-unet: an efficient group enhanced unet for skin lesion segmentation." International Conference on Computer-Assisted Intervention and Medical Image Computing. Cham, Switzerland: Springer Nature, 2023.

[32] Wenxuan Zhao and colleagues present "Attention enhanced serial Unet++ network for removing unevenly distributed haze." 10.22 (2021): 2868 in Electronics.

[33] Deng, Yunjiao, et al. "Medical image segmentation using ELU-net: an effective and lightweight U-net." 35932–35941, IEEE Access 10 (2022).

[34] "ELU-net: an efficient and lightweight U-net for medical image segmentation," Deng, Yunjiao, et al. 35932–35941, IEEE Access 10 (2022).

[35] Rafael Silva Araújo et al. "Flood impact on income inequality in the Itapocu River basin, Brazil." 15.3 (2022): e12805 in Journal of Flood Risk Management.

[36] Ahmadi, Seyed Ali, Ali Mohammadzadeh, Naoto Yokoya, and Arsalan Ghorbanian. "BD-SKUNet: Selective-kernel UNets for building damage assessment in high-resolution satellite images." Remote Sensing 16, no. 1 (2023): 182.

[37] Tran, Dai Quoc, Minsoo Park, Daekyo Jung, and Seunghee Park. "Damage-map estimation using UAV images and deep learning algorithms for disaster management system." Remote Sensing 12, no. 24 (2020): 4169.

[38] Gupta, Ananya, Simon Watson, and Hujun Yin. "Deep learning-based aerial image segmentation with open data for disaster impact assessment." Neurocomputing 439 (2021): 22-33

[39] Akhyar, Akhyar, Mohd Asyraf Zulkifley, Jaesung Lee, Taekyung Song, Jaeho Han, Chanhee Cho, Seunghyun Hyun, Youngdoo Son, and Byung-Woo Hong. "Deep artificial intelligence applications for natural disaster management systems: A methodological review." Ecological Indicators 163 (2024): 112067.

[40] Muhammad Rizwan et al., "Simulating future flood risks under climate change in the Indus River source region." Vol. 16, No. 1, 2023, Journal of Flood Risk Management, e12857.

[41] Ghosh, Binayak, Shagun Garg, Mahdi Motagh, and Sandro Martinis. "Automatic flood detection from Sentinel-1 data using a nested UNet model and a NASA benchmark dataset." PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science 92, no. 1 (2024): 1-18.

[42] You, Di, Shixin Wang, Futao Wang, Yi Zhou, Zhenqing Wang, Jingming Wang, and Yibing Xiong. "EfficientUNet+: a building extraction method for emergency shelters based on deep learning." Remote Sensing 14, no. 9 (2022): 2207.

[43] Chen, De-Yue, Ling Peng, Wei-Chao Li, and Yin-Da Wang. "Building extraction and number statistics in WUI areas based on UNet structure and ensemble learning." Remote Sensing 13, no. 6 (2021): 1172.

[44] Tan, Chunhai, Tao Chen, Jiayu Liu, Xin Deng, Hongfei Wang, and Junwei Ma. "Building Extraction from Unmanned Aerial Vehicle (UAV) Data in a Landslide-Affected Scattered Mountainous Area Based on Res-Unet." Sustainability 16, no. 22 (2024): 9791.

[45] Kaushal, Arush, Ashok Kumar Gupta, and Vivek Kumar Sehgal. "A semantic segmentation framework with UNet-pyramid for landslide prediction using remote sensing data." Scientific Reports 14, no. 1 (2024): 1-23.

[46] Pan, Bin, and Xianjian Shi. "Fusing ascending and descending time-series SAR images with dual-polarized pixel attention UNet for landslide recognition." Remote Sensing 15, no. 23 (2023): 5619.

[47] Wang, Junxin, Qintong Zhang, Hao Xie, Yingying Chen, and Rui Sun. "Enhanced dual-channel model-based with improved Unet++ network for landslide monitoring and region extraction in remote sensing images." Remote Sensing 16, no. 16 (2024): 2990.

[48] Gupta, Kushagra, and Priya Mishra. "Post-disaster segmentation using FloodNet." Studies 13 (2021): 8.

[49] Nugraha, Deny Wiria, Amil Ahmad Ilham, Andani Achmad, and Ardiaty Arief. "Transformers for aerial images semantic segmentation of natural disaster-impacted areas in natural disaster assessment." Bulletin of Electrical Engineering and Informatics 14, no. 2 (2025): 1391-1406.

[50] Jaisakthi, S. M., P. R. Dhanya, and S. Jitesh Kumar. "Detection of flooded regions from satellite images using modified unet." In International Conference on Computational Intelligence in Data Science, pp. 167-174. Cham: Springer International Publishing, 2021.

[51] Mesvari, Mohaddeseh, and Reza Shah-Hosseini. "Flood detection based on UNet++ segmentation method using Sentinel-1 satellite imagery." Earth Observation and Geomatics Engineering 7, no. 1 (2023).

[52] Chamatidis, Ilias, Denis Istrati, and Nikos D. Lagaros. "Vision Transformer for Flood Detection Using Satellite Images from Sentinel-1 and Sentinel-2." Water 16, no. 12 (2024): 1670.

[53] Saleh, Tamer, Shimaa Holail, Xiongwu Xiao, and Gui-Song Xia. "High-precision flood detection and mapping via multi-temporal SAR change analysis with semantic token-based transformer." International Journal of Applied Earth Observation and Geoinformation 131 (2024): 103991.

[54] Saleh, Tamer, Xingxing Weng, Shimaa Holail, Chen Hao, and Gui-Song Xia. "DAM-Net: Flood detection from SAR imagery using differential attention metric-based vision transformers." ISPRS Journal of Photogrammetry and Remote Sensing 212 (2024): 440-453.

[55] Hassan, Ibne, Aman Mujahid, Abdullah Al Hasib, Andalib Rahman Shagoto, Joyanta Jyoti Mondal, Meem Arafat Manab, and Jannatun Noor. "Aerial Flood Scene Classification Using Fine-Tuned Attention-based Architecture for Flood-Prone Countries in South Asia." arXiv preprint arXiv:2411.00169 (2024).

[56] Hassan, Ibne, Aman Mujahid, Abdullah Al Hasib, Andalib Rahman Shagoto, Joyanta Jyoti Mondal, Meem Arafat Manab, and Jannatun Noor. "Aerial Flood Scene Classification Using Fine-Tuned Attention-based Architecture for Flood-Prone Countries in South Asia." arXiv preprint arXiv:2411.00169 (2024).

[57] Tang, Zhaojia, and Yu Han. "Focus on Disaster Risk Reduction by ResNet-CDMV Model After Natural Disasters." Applied Sciences 14, no. 22 (2024): 10483.

[58] Baral, Anuj, Vikash Singh, and Achyut Lath. "Evaluating the Performance of ResNet-50 and GoogleNet for Damage Detection and Classification." In 2024 4th International Conference on Sustainable Expert Systems (ICSES), pp. 1721-1726. IEEE, 2024.

[59] Zhao, Xin, Yitong Yuan, Mengdie Song, Yang Ding, Fenfang Lin, Dong Liang, and Dongyan Zhang. "Use of unmanned aerial vehicle imagery and deep learning unet to extract rice lodging." Sensors 19, no. 18 (2019): 3859.

[60] Abdollahi, Abolfazl, Biswajeet Pradhan, and Abdullah M. Alamri. "An ensemble architecture of deep convolutional Segnet and Unet networks for building semantic segmentation from high-resolution aerial images." Geocarto International 37, no. 12 (2022): 3355-3370.

[61] Chen, Jui-Fa, Yu-Ting Liao, and Po-Chun Wang. "Development and deployment of a virtual water gauge system utilizing the resnet-50 convolutional neural network for real-time river water level monitoring: A case study of the keelung river in taiwan." Water 16, no. 1 (2023): 158.

[62] Ronneberger O, Fischer P, Brox T (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation"].

[63] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian (2016). Deep Residual Learning for Image Recognition (PDF). Conference on Computer Vision and Pattern Recognition. doi:10.1109/CVPR.2016.90.

[64] Chen, Liang-Chieh & Papandreou, George & Schroff, Florian & Adam, Hartwig. (2017). Rethinking Atrous Convolution for Semantic Image Segmentation. 10.48550/arXiv.1706.05587.

[65] Dosovitskiy, Alexey; Beyer, Lucas; Kolesnikov, Alexander; Weissenborn, Dirk; Zhai, Xiaohua; Unterthiner, Thomas; Dehghani, Mostafa; Minderer, Matthias; Heigold, Georg; Gelly, Sylvain; Uszkoreit, Jakob (2021-06-03). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". arXiv:2010.11929

[66] Z. Liu *et al*., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021, pp. 9992-10002, doi: 10.1109/ICCV48922.2021.00986

[67] Taha, A.A., Hanbury, A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging* **15**, 29 (2015). https://doi.org/10.1186/s12880-015-0068-x