Masters Program in Geospatial Technologies

# FEW-SHOTS LEARNING FOR POST-EARTHQUAKE BUILDING DAMAGE ASSESSMENT USING METRIC-BASED AND TRANSFER LEARNING METHODS

**Enayatullah Meskinyaar**

NOVA Information Management School

Master of Science in Geospatial Technologies

NOVA IMS Information Management School

Universitat Jaume I

Westfälische Wilhelms-Universität Münster

Supported by: Education and Culture ERASMUS MUNDUS

**NOVA Information Management School**

**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa

**TITLE**

# FEW-SHOTS LEARNING FOR POST-EARTHQUAKE BUILDING DAMAGE ASSESSMENT USING METRIC-BASED AND TRANSFER LEARNING METHODS

by

**Enayatullah Meskinyaar**

Master Dissertation presented as partial requirement for obtaining the master's degree in Geospatial Technologies

*Supervised by:*

**Leonardo Vanneschi**, PhD, NOVA Information Management School, Universidade Nova de Lisboa

*Co-supervised by:*

**Filiberto Pla Bañón**, PhD, GEOTEC, Universitat Jaume I Castellón

*Co-supervised by:*

**Christian Knoth**, PhD, ifgi, University of Münster

February 24, 2025

## STATEMENT OF INTEGRITY

I declare that the work described in this document is my own and not from someone else. All the assistance I have received from other people is duly acknowledged and all the sources (published or not published) are referenced.

This work has not been previously evaluated or submitted to NOVA Information Management School or elsewhere. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, February 20, 2025

Enayatullah Meskinyaar

### USE OF GENERATIVE ARTIFICIAL INTELLIGENCE

| Tasks | NO | YES | Generative Artificial Intelligence tools |
|---|---|---|---|
| Better understand issues related to the research | | × | SciSpace |
| Summarizing text from bibliography / resources | | × | SciSpace |
| Summarizing the method(s) used | | × | SciSpace |
| Translating text | × | | |
| Grammar check | | × | ChatGPT |
| Paraphrase or rewriting text from other people / resources | × | | |
| Coding in R, Python, etc. | | × | ChatGPT |
| Get help on a software | | | ChatGPT |
| Creating and editing images, maps, videos, etc. | × | | |
| Data analysis | × | | |
| Specify below other tasks not mentioned above: | | | |
| Improving writing quality | | × | ChatGPT, Grammarly |
| | | | |

# Acknowledgement

First and foremost, I extend my sincere gratitude to my supervisor, Professor Leonardo Vanneschi, for his invaluable guidance in a scientific field I was newly introduced to. His unwavering support and availability to clarify any doubts throughout the development of this thesis have been truly appreciated. And I would like to express my gratitude to Professor Filiberto Pla Bañón and Christian Knoth for accepting to be my co-supervisor and providing nice feedback.

I am also deeply grateful to Professor Marco Painho for his constructive feedback, mentorship, and continuous support not only during the thesis semester but all semesters of my Master's in Geospatial Technologies.

I would like to express my appreciation to my colleagues, Joseph Paintsil, Mohammad Qasim, Flavio Vata, Sebastian Andrade , Christopher Hubach , Hamna Khurshid , Sirak Transfmariam, Ammar Yusaf, Alonso Gonzalez , Lemesa Tolera Hirpha , Miguel Lucas , Carolina Cardoso , Bernardo Trovao and Guilherme Viegas, whose engaging discussions, ideas, and enthusiasm greatly enriched my academic experience.

As I reach the end of this journey, I feel profoundly thankful to everyone, especially my family, who supported me in any capacity, helping me embark on and navigate through these challenging times. I am especially grateful to those who assisted me in applying for this program, as well as those who welcomed me as a scholarship recipient.

# Few Shot Learning for Post-Earthquake Building Damage Assessment Using Metric-based and Transfer Learning Approach


## Abstract

Earthquakes are among the most devastating natural disasters, often causing widespread destruction and loss of life. A rapid and accurate assessment of building damage is crucial for effective disaster response and recovery. However, traditional methods are time-consuming and require significant resources, while deep learning approaches struggle with challenges such as limited and imbalanced datasets. This study focuses on the Mexico City Earthquake (2017), and it utilizes satellite imagery to classify building damage into varying levels of severity. Moreover, this study explores the use of Few-Shot Learning (FSL) to overcome the challenges of limited and imbalanced data. In the realm of FSL, we use metric-based learning and transfer learning approaches to improve classification performance in scenarios where data is scarce and imbalanced. To evaluate the effectiveness of our approach, we evaluate and compare Prototypical Networks, EfficientNetB7, and ResNet50, analyzing key metrics such as precision, recall, F-score, and overall accuracy. Our findings reveal that Prototypical Networks outperform other models, particularly in identifying severely damaged structures. Additionally, data augmentation and oversampling are proven to be effective techniques for handling data imbalance. As the dataset is very limited in this study, the challenge of data scarcity, however, persists. To provide deeper insights, we integrate our damage predictions with geospatial mapping, revealing a decent correlation between predicted damage severity and actual impacted areas. Ultimately, this research highlights the potential of Few-Shot Learning in enabling rapid and scalable damage assessment in data-limited scenarios and contributes to improving emergency response efforts and optimizing resource allocation.



**Key words:** Building Damage Assessment; Deep Learning; Few Shots Learning; Data Balancing; Remote Sensing


### Sustainable Development Goals (SGD):

# ACRONYMS

**CNN**- Convolutional Neural Networks

**DBUA**- Dual-Branch U-Net Architecture

**DNN**-Deep Neural Network

**FSL**- Few-Shot Learning

**GAP**-Global Average Pooling

**GSD**- Ground Sampling Distance

**ML**-Machine learning

**OCDPL**- Ordinal Class Distance Penalty Loss

**ProtoNets-** Prototypical Networks

**ReLU** - Rectified Linear Unit

**ResNet -** Residual Network

**SAR**- Synthetic Aperture Radar

**SKM**- Selective Kernel Module

**UAVs**- Unmanned Aerial Vehicles

**VHR**-Very High-Resolution Images

**Contents**

Abstract iii

Acronyms iv

List of Figures vii

List of Tables viii

1. Introduction ..... 1
   1.1 Background and Motivation ..... 1
   1.2 Objective and Research Questions ..... 2
2. Literature Review ..... 3
   2.1 Remote Sensing for Emergency Mapping ..... 3
   2.2 Deep Learning for building damage assessment ..... 4
   2.3 Few Shots Learning (FSL) ..... 5
   2.4 Transfer Learning ..... 5
   2.5 Convolutional Neural Networks (CNNs) ..... 6
   2.6 Prototypical Networks (ProtoNets) ..... 7
   2.7 Comparative Analysis ..... 8
   2.8 Data Balancing ..... 10
   2.9 Performance Metrics ..... 10
3. Data and Methodology ..... 12
   3.1 Dataset ..... 12
   3.2 Exploratory Data Analysis ..... 13
   3.3 Methodology ..... 14
      3.3.1 Data preparation ..... 14
      3.3.2 Oversampling ..... 15
      3.3.3 Metric-based Learning: Prototypical Network (ProtoNet) ..... 16
      3.3.4 Transfer Learning: ResNet 50 ..... 17
      3.3.5 Transfer Learning: EffecientNetB7 ..... 19
4. Results and Discussion ..... 20
   4.1 Learning process ..... 20
      Model 1 ..... 20

Model 2……21
Model 3……22
4.2 Performance Metrics ..... 23
Model ……..……23
Model 2……24
Model 3……25
Comparison……....26
4.3 Inference ..... 29
Rural Area……29
Urban Area……..30
Dense Urban Area…….....31
3.4.1 Limitation..... 33
5. Conclusion ..... 33
5.1 Implications and Future Directions..... 35
Data and Sources……34
6. Bibliography ..... 37

# List of Figures

Figure 2-1: An example of CNN architecture (Koukouraki et al., 2021) ..... 7

Figure 2-2: Show the ProtoNet Functionality (Snell et al., 2017) ..... 8

Figure 3-1: Post-Earthquake image and corresponding labels. In image right, Green shows No damage, yellow shows Minor Damage, Orange shows Major Damage and red shows Destroyed buildings. ..... 13

Figure 3-2: Flow chart for the thesis ..... 14

Figure 3-3: Shows the cropped building instances ..... 15

Figure 3-4: Show the ProtoNet's architecture ..... 17

Figure 3-5: shows the learning process ResNet50 ..... 18

Figure 3-6: Shows a shallow CNN architecture for ResNet 50 ..... 18

Figure 4-1: Shows the accuracy and loss over epochs for Model 1 ..... 21

Figure 4-2: Shows the accuracy and loss over epochs for Model 2 ..... 22

Figure 4-3: Shows the accuracy and loss over epochs for Model 3 ..... 23

Figure 4-4: A comparative view performance metrics for the 3 models ..... 28

Figure 4-5: Prediction based on Model 1 on test image from Mexico Earthquake (2017). The satellite image is overlaid with a) predicted label, b) ground truth labels, and c) the differences between predictions and labels. ..... 30

Figure 4-6: Prediction based on Model 1 on test image from Mexico Earthquake (2017). The satellite image is overlaid with a) predicted label, b) ground truth labels, and c) the differences between predictions and labels ..... 31

Figure 4-7: Prediction based on Model 1 on test image from Mexico Earthquake (2017). The satellite image is overlaid with a) predicted label, b) ground truth labels, and c) the differences between predictions and labels ..... 32

# List of Tables

Table 2-1: Different influential parameters on model performance .....9

Table 2-2: A general representation confusion matrix in binary classification.....10

Table 3-1: Shows the technical Specifications of xBD dataset .....12

Table 3-2: Number instances per class .....13

Table 3-3: Shows training set, validation set, test set and the number of times each minority class is resampled.....16

Table 3-4: Summary of hyperparameter settings for Model1, Model 2 and Model 3. .....19

Table 4-1: A Confusion Matrix of damage classes for Model 2 .....24

Table 4-2: Performance Metrics based on Precision, Recall and F-score for Model 1 .....24

Table 4-3: Performance Metrics based on Precision, Recall and F-score for Model 2 .....25

Table 4-4: A Confusion Matrix of damage classes for Model 2 .....25

Table 4-5: Performance Metrics based on Precision, Recall and F-score for Model 3 .....26

Table 4-6: A Confusion Matrix of damage classes for Model 3 .....26

Table 4-7: Averaged metrics' values and overall accuracy for models.....29

# 1. Introduction

## 1.1 Background and Motivation

Earthquakes are among the most devastating natural disasters, causing extensive destruction to infrastructure and loss of human lives. According to UNDRR and CRED (2019), earthquakes were the deadliest disasters of the past two decades, with their unpredictability posing significant challenges to emergency response efforts (Koukouraki et al., 2021; Wang et al., 2022). Consequently, rapid and accurate assessment of post-earthquake building damage is critical for effective disaster management and emergency response, as it enables the prioritization of relief efforts in the most affected areas (Lin et al., 2022).

Despite the urgent need for accurate assessments, traditional methods, such as ground surveys, remain labor-intensive, time-consuming, and often dangerous in post-disaster scenarios (Ma et al., 2019). To address these challenges, remote sensing methods, including satellite imagery and UAVs, have emerged as scalable, safe, and efficient alternatives for assessing building damage (Bouchard et al., 2022). These methods enable large-scale analysis of affected areas and prove invaluable in post-disaster mapping. However, extracting meaningful insights from such data introduces additional challenges, as the complexity and volume of imagery necessitate advanced computational techniques.

In this context, machine learning, particularly deep learning, has demonstrated its potential as a powerful tool for damage assessment. Convolutional Neural Networks (CNNs) have outperformed traditional machine learning approaches by automating feature extraction and classification tasks (Koukouraki et al., 2021; Ma et al., 2020). However, these models often require large, labeled datasets, which are typically unavailable in the immediate aftermath of a disaster (Bouchard et al., 2022). Moreover, the variability in construction materials, architectural styles, and damage patterns across regions further complicates the generalization of these models.

To overcome these limitations, Few-Shot Learning (FSL) method is used in this study. FSL offers a promising solution by enabling models to learn from a limited number of examples labeled. By using prior knowledge and advanced feature extraction techniques, FSL can generalize to new tasks with minimal data, which makes it particularly suitable for rare events like earthquakes (Koukouraki et al., 2021; Wang et al., 2021). Furthermore, metric-based methods, such as Prototypical Networks, have shown promise in learning effective distance metrics for classification in data-limited scenarios (Koukouraki et al., 2021). Complementing this, transfer learning complement this by providing additional capabilities as it uses pre-trained models to extract features, thereby it reduces the dependency on extensive labeled datasets (Bouchard et al., 2022; Lin et al., 2022).

Despite these advancements, most studies have focused on binary damage classification (e.g., destroyed vs. non-destroyed buildings) and relied on integrating pre- and post-event data, which may not always be accessible or feasible (Koukouraki et al., 2021; Wang et al., 2022). On the other

hand, multi-class damage classification, while more informative for nuanced disaster response, introduces significant challenges, such as class imbalance and increased computational complexity (Ma et al., 2020). Therefore, there is a need for an innovative approach such as FSL combining metric-based and transfer learning strategies to address those challenges to a considerable extent.

## 1.2 Objective and Research Questions

This study is concerned with developing a framework that uses FSL including metric-based methods, and transfer learning to perform multi-class post-earthquake building damage assessment. As we are dealing with multi-class problems in this study, the available dataset is highly imbalanced where one class has multiple times more samples than other classes. Additionally, as deep learning requires huge amounts of data, this study also faces the challenge of limited data. Therefore, this study aims to use techniques and models to mitigate the problem of scarce and imbalanced data.

There are four research questions that this study aims to address:

1. *How to address the challenges of limited and imbalanced data, where data is limited, and one class has multiple times more samples than other classes?*
2. *As this study is concerned with multi-class problems, how well do the proposed models perform class-wide?*
3. *Comparing the results based on performance metrics, which models are the winner?*
4. *Running the prediction based on the best model, to what extent is the prediction map indicative of the actual severity of the damage suffered by a region?*

To find answers to these research questions, the following objectives will be accomplished:

- **Review Existing Literature:** Investigate prior research and best practices related to handling imbalanced datasets, multi-class classification, and disaster damage assessment using deep learning.
- **Explore and Implement Suitable Methods:** Assess various data preprocessing techniques, model architecture, and learning strategies to address data limitations and imbalance challenges.
- **Evaluate Model Performance:** Measure the effectiveness of the implemented approaches using quantitative and qualitative evaluation metrics, ensuring a thorough comparison of model performance.
- **Assess Generalization Across Disasters:** run the prediction on different earthquake-stricken regions in Mexico City.

# 2. Literature Review

## 2.1 Remote Sensing for Emergency Mapping

Remote sensing has emerged as a transformative tool for emergency response and natural disaster mapping, which enables a rapid and detailed analysis of affected regions. By using satellite imagery, advanced computational techniques like deep learning, and innovative methods such FSL, remote sensing has significantly improved disaster assessment, response planning, and mitigation efforts.

At the core of remote sensing for disaster management lies satellite imagery, which provides large spatial and temporal coverage. This ability to capture large-scale data is critical during disasters like earthquakes, floods, or hurricanes, where quick and accurate damage assessment can save lives. High-resolution optical imagery, for instance, offers rich semantic details, that enables precise evaluation of urban damage. However, optical imagery is not without limitations; atmospheric disturbances such as clouds or low light conditions can reduce its effectiveness. In such cases, radar-based technologies like Synthetic Aperture Radar (SAR) step in, offering the ability to penetrate these obstacles and deliver vital ground-level information, ensuring continuity in disaster assessments (Ma et al., 2019).

To extract actionable insights from this wealth of data, deep learning has revolutionized the analysis of satellite imagery. CNN, known for its ability to extract fine features from visual data, has become essential in identifying collapsed buildings and flooded areas. Recent advancements, including lightweight architectures and optimized loss functions, have improved the efficiency of these models, which make them suitable for rapid deployment in real-time scenarios (Ma et al., 2019). Additionally, the emergence of transformer-based models has introduced new possibilities. These models excel in capturing both global and local spatial relationships, thus outperform traditional CNNs in disaster damage assessment tasks (Yu et al., 2023a).

Despite these advancements, a significant challenge in disaster mapping remains such as the scarcity of labeled data necessary for training deep learning models. This is where few-shot learning proves invaluable. By enabling models to generalize from only a few examples, FSL addresses the data deficiency problem head-on. Techniques like Prototypical Networks have demonstrated remarkable success in classifying damage levels with minimal training data, reducing the dependency on large datasets and allowing for quicker deployment of machine learning models in critical situations (Koukouraki et al., 2021). This capability is especially vital when time is of the essence, such as during the immediate aftermath of a disaster.

## 2.2 Deep Learning for building damage assessment

Deep learning has revolutionized building damage assessment by automating the extraction of meaningful features from remote sensing imagery. Pre-trained models and few-shot learning techniques have emerged as powerful tools for addressing challenges like data scarcity and class imbalance, which are common in post-disaster scenarios. Transfer learning, that uses pre-trained models, has shown good performance in adapting to new disaster contexts with limited labeled data. Bouchard et al. (2022) highlighted the importance of transfer learning in emergency scenarios, demonstrating how pre-trained CNNs fine-tuned with minimal disaster-specific samples can significantly reduce response time. This approach enables models to generalize across varying disaster conditions. Similarly, Toba et al.(2023) evaluated AlexNet, VGG-16, and ResNet-34 on the xBD dataset, concluding that ResNet-34 provided the best balance of accuracy and computational efficiency for building damage detection.

Pre-trained CNNs also play a critical role in multi-class damage assessment. Ahmadi et al. (2023) proposed a Selective Kernel U-Net for building damage assessment using pre- and post-disaster imagery. Their model demonstrated superior performance in accurately classifying damage levels, emphasizing the utility of adaptive receptive fields. Additionally, Lin et al. (2022) developed a data transfer algorithm that filters historical samples to improve model calibration for new disasters, and they achieved an 8% accuracy improvement when training on limited samples.

FSL has been explored to address the scarcity of labeled data. Koukouraki et al. (2021) tested Prototypical Networks for post-earthquake damage classification and found them effective in handling data imbalance. Their results confirmed the suitability of FSL for creating damage assessment maps with minimal training samples. Wang et al. (2021) emphasized the versatility of few-shot learning in leveraging prior knowledge to reduce reliance on large datasets, which is particularly beneficial in disaster scenarios with rare occurrences. Advanced architectures, such as attention mechanisms integrated with U-Nets, further enhance performance. Wu et al. (2021) applied attention U-Nets on the xBD dataset, achieving an F1 score of 0.792 for multi-class damage classification. Similarly, (Tsai & Lin, 2024)introduced a novel loss function addressing class imbalance in multi-class segmentation tasks, improving minority class predictions critical for disaster response.

Overall, the integration of pre-trained models and FSL approaches, coupled with innovative architectures, has significantly advanced building damage assessment. These methods not only improve accuracy and efficiency but also make automated systems more adaptable to diverse and challenging disaster scenarios.

## 2.3 Few Shots Learning (FSL)

FSL has gained attraction in addressing the challenge of limited labeled data in building damage detection. It enables the generalization of models to new tasks with minimal examples, an essential capability for disaster response scenarios where data scarcity is a prevalent issue. Koukouraki et al. (2021) demonstrated the application of FSL in post-earthquake urban damage detection. Using Prototypical Networks, they addressed multi-class damage classification, which is more informative than binary classification but introduces challenges like class imbalance. Their study highlighted the effectiveness of Prototypical Networks in learning meaningful class representations by averaging embedded support examples, making them particularly suitable for imbalanced datasets(Snell et al., 2017). Snell et al. (2017) also introduced Prototypical Networks as a metric-based approach to FSL, focusing on learning a metric space where classes are represented by their prototypes. These prototypes, defined as the meaning of the class embeddings, allow for efficient and accurate classification using distance metrics like Euclidean distance. This method simplifies the FSL pipeline while achieving state-of-the-art results in classification tasks. In disaster scenarios, Wang et al. (2021) emphasized the importance of using prior knowledge through transfer learning and meta-learning to reduce reliance on large datasets. FSL, particularly in its metric-learning form, bridges the gap between low data availability and high-performance requirements by focusing on embedding models that generalize across tasks. Hilliard et al. (2018) also proposed a flexible architecture that combines metric-agnostic conditional embeddings with FSL. This method adapts class representations based on query samples, enhancing the model's ability to differentiate subtle variations in damage levels. This approach complements Prototypical Networks by refining the representation of each class dynamically. FSL, particularly metric-based methods like Prototypical Networks, provides a robust framework for building damage detection in data-scarce environments. Its adaptability to multi-class scenarios and compatibility with transfer learning strategies positions as a critical tool in post-disaster assessment. Therefore, Prototypical Networks are used for this study.

## 2.4 Transfer Learning

Transfer learning has proven instrumental in enhancing building damage assessment by using pre-trained models to address data scarcity in post-disaster scenarios. Pre-trained architectures like InceptionV3, Residual Network (ResNet), and EfficientNet have shown considerable adaptability in extracting hierarchical features, and significantly improved classification and segmentation tasks.

Lin et al. (2022) emphasized the effectiveness of transfer learning in adapting models trained on historical disaster data to new earthquake scenarios. Their study introduced a data transfer algorithm, filtering useful historical samples to fine-tune models for new tasks. By using ResNet as the backbone, they achieved an 8% improvement in accuracy when only 10% of new disaster data was available, which underscores ResNet’s capability in generalizing across varying disaster

conditions. Similarly, Ma et al. (2020) demonstrated the superiority of InceptionV3 for classifying post-earthquake building damage. By integrating block-based vector data, the model achieved a 90.07% accuracy rate, that outperformed traditional machine learning approaches. InceptionV3's modular architecture effectively reduced overfitting while preserving essential features for damage classification.

(Ahmadi et al., 2023) explored the geographical transferability of EfficientNet within a dual-branch U-Net framework for building damage segmentation. The model successfully localized building footprints and classified damage levels across multiple disaster datasets, demonstrating EfficientNet's robust feature extraction capabilities and adaptability. Tsai & Lin, (2024) tackled class imbalance in multi-class damage segmentation using ResNet-based backbones with a novel loss function. Their study underscored the potential of transfer learning to mitigate performance biases and enhance predictions in minority classes, which are crucial in disaster response scenarios.

Transfer learning with pre-trained models such as ResNet50 and EfficientNetB7 has significantly advanced the field of building damage assessment. These models excel in adapting to diverse datasets, addressing data scarcity, and enhancing the accuracy and scalability of post-disaster mapping. Therefore, ResNet50 and EfficientNet that originally trained on millions of images with 10000 classes, known as *imagenet* used as pre-trained model in this study.

## 2.5 Convolutional Neural Networks (CNNs)

(CNNs are a specialized class of deep learning architecture designed for processing structured grid data, primarily images. CNNs utilizes spatial hierarchies of features, capturing patterns such as edges, textures, and complex structures at different levels of abstraction. Their primary applications include image classification, object detection, and segmentation, particularly in remote sensing and disaster assessment (Braik & Koliou, 2024; Duarte et al., 2018).

CNNs are structured into multiple layers, each playing a unique role in feature extraction and decision-making, Figure 2-1:

- **Convolutional Layers**: The core operation in CNNs is the convolution operation, where filters (kernels) slide over the input image, computing dot products between the filter weights and the local image region(Ji et al., 2020). These filters extract hierarchical features: early layers detect low-level features (e.g., edges, corners), while deeper layers detect high-level structures (e.g., building damage patterns in satellite images(Koukouraki et al., 2021).

- **Activation Functions**: Non-linearity is introduced via activation functions like ReLU (Rectified Linear Unit), which ensures that CNNs can model complex, non-linear relationships(Duarte et al., 2018).

- **Pooling Layers**: These layers reduce spatial dimensions while preserving important features, improving computational efficiency and robustness to small transformations. Max pooling is commonly used, selecting the highest value in a feature region, ensuring that key features remain dominant(Settou et al., 2022).

- **Fully Connected Layers (FC Layers)**: These layers take the flattened output from previous layers and apply dense connections to perform classification or regression. For instance, in building damage assessment, CNNs classify regions as no damage, minor damage, major damage and destroyed(Valentijn et al., 2020).
- **Dropout and Batch Normalization**: Dropout is used to prevent overfitting, randomly deactivating neurons during training. Batch normalization stabilizes learning, improving convergence speed(Braik & Koliou, 2024).

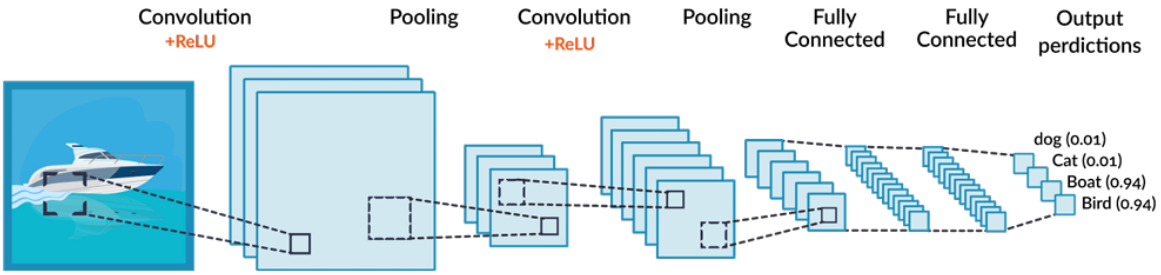


Figure 2-1: An example of CNN architecture (Koukouraki et al., 2021)

In general, CNNs process images in a hierarchical fashion: Convolutional layers extract spatial patterns, Pooling layers down sample feature maps, Dropout layers avoid overfitting, and fully connected layers predict object categories.

## 2.6 Prototypical Networks (ProtoNets)

ProtoNets are a metric-learning-based approach to FSL. Unlike traditional deep learning models that require large, labeled datasets, ProtoNets can generalize from a few labeled examples, making them ideal for tasks with limited data. This is particularly useful in applications such as post-earthquake building damage classification. ProtoNets learn a metric space, where classification is performed by measuring distances to prototype representations of each class. ProtoNet is considered to be a leading approach among state-of-the-art FSL algorithms(Koukouraki et al., 2021). It demonstrated the ability to identify entirely new classes not included in the training data. These networks integrate concepts from both meta-learning and metric learning. Meta-learning is a branch of machine learning, and it utilizes knowledge gained from related tasks to speed up the learning process for new tasks(Koukouraki et al., 2021).

In graphical representation in Figure 2-2, ProtoNet classifies data by computing class prototypes $C$ as the mean embeddings of support samples for each class. A query point $X$ is embedded into the same space, and its class is determined based on the shortest distance (e.g., Euclidean) between $X$ and the prototypes $C$, partitioning the space into regions

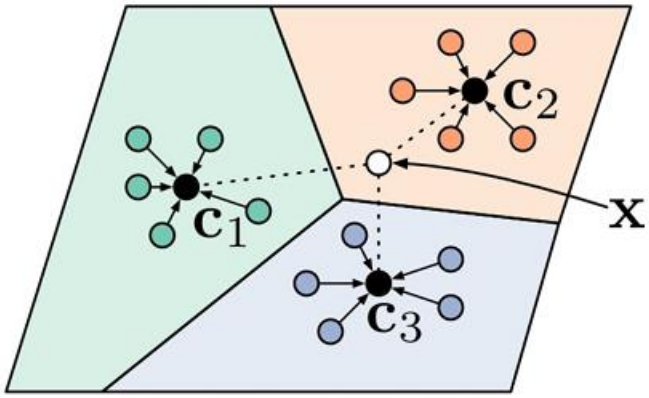


Figure 2-2: Show the ProtoNet Functionality (Snell et al., 2017)

## 2.7 Comparative Analysis

Machine learning (ML), particularly deep learning methods such as CNNs, has revolutionized building damage assessment in post-disaster scenarios. By leveraging advanced algorithms, diverse data types, and innovative techniques, ML has significantly improved the accuracy, efficiency, and scalability of damage detection and classification tasks. Despite these advancements, challenges related to data scarcity, class imbalance, and domain generalization persist, necessitating the development of novel strategies to address these issues effectively.

Data types play a critical role in shaping the performance of ML models. High-resolution optical imagery from satellites like WorldView and QuickBird is a common choice, providing detailed visual information essential for detecting structural changes(Ma et al., 2019; Wu et al., 2021). Complementary data sources, such as LiDAR, provide 3D structural information that enhances the detection of collapsed buildings by analyzing height differences before and after disasters(Amini Amirkolaee & Arefi, 2019). Bitemporal datasets, comprising pre- and post-event imagery, enable detection change, which is particularly effective in delineating the extent of damage(Yu et al., 2023). However, single-temporal data remains an option in scenarios where pre-disaster imagery is unavailable. The most common dataset used in these studies is known as xBD dataset which has become a widely used benchmark for building damage assessment, providing labeled pre- and post-disaster satellite imagery(Lin et al., 2022; Valentijn et al., 2020). Large-scale datasets like xBD enable nuanced multi-class classification; however, disaster-specific datasets are often limited in size, leading to challenges such as skewed predictions and class imbalances. To address

these issues, techniques like data augmentation and pre-trained model adaptation have been employed, enhancing the generalizability of models to diverse disaster scenarios(Ma et al., 2020; Toba et al., 2023).

To mitigate the impact of limited training data, FSL and transfer learning have emerged as promising methodologies. Pre-trained models like ResNet and EffecientNet play a critical role in enhancing feature extraction for damage classification tasks. For instance, Ma et al. (2020) adapted InceptionV3 for group-based damage detection, integrating block vector data to improve segmentation accuracy. Similarly, Toba et al. (2023) demonstrated the effectiveness of ResNet-34 in streamlining transfer learning through low-complexity preprocessing, achieving competitive F1 scores in classification tasks. These examples highlight how pre-trained models can serve as a foundation for robust damage assessment frameworks. FSL offers an additional layer of adaptability in data-scarce environments. Prototypical Networks, a metric-based FSL approach, have proven highly effective for urban damage classification. By representing classes through mean embeddings of training examples, these networks excel in handling imbalanced datasets. Table 2-1 highlights some of the most influential parameters mentioned in the papers that affect model performance.

Table 2-1: Different influential parameters on model performance

| Study | Data Source | Dataset Size | ML Models | classes |
|---|---|---|---|---|
| (Ahmadi et al., 2023) | Satellite (Bi-temporal) | 850,736 | Pre-trained EfficientNetB7 and ResNet34, DBUA, SKM | 4 |
| (Tsai & Lin, 2024) | Satellite(Bi-temporal) | 1219 | OCDPL | 4 |
| (Valentijn et al., 2020) | Satellite (Bi-temporal) | 175,289 | Pre-trained InceptionV3 | 4 |
| (Lin et al., 2022) | UAV, Areal (Bi-temporal) | 17281 | VGG-OR | 4 |
| (Koukouraki et al., 2021) | Satellite(post-event) | >150,000 | ProtoNet, Oversampling, Undersampling, cost-sensitive | 4 |
| (Toba et al., 2023) | Satellite(Bi-temporal) | >850,000 | Pre-trained AlexNet, ResNet34, VGG-16 | 2 |
| (Settou et al., 2022) | UAV(post-earthquake) | 19133 | Pre-trained AlexNet, GoogleNet, VGG-VD | 2 |
| (Ma et al., 2020) | Aerial (post-earthquake) | 16803 | Pre-trained InceptionV3 | 3 |

Multi-class segmentation, though valuable for detailed damage assessment, introduces further challenges such as class imbalance. Addressing this, (Tsai & Lin, 2024)proposed an ordinal loss function that penalizes misclassifications based on severity, improving predictions for minority classes. Additionally, hybrid feature representations, which combine outputs from fully connected layers of pre-trained CNNs, have demonstrated enhanced classification accuracy across diverse conditions.

As a result, this study proposes to use pre-trained ResNet50, EffecientNetB7 and Prototypical Network for building damage classification and data oversampling and augmentation for addressing the challenge of data imbalance.

## 2.8 Data Balancing

One of the drawbacks of the xBD dataset is that it is not balanced, meaning that each damage class has a significant inequal number of instances. The presence of this imbalanced dataset causes the model to generalize in favor of majority class, this problem is known as overfitting. However, there are some techniques to reduce the effect of overfitting by resampling and assigning different weights to the dataset. Famous techniques are oversampling, undersampling and cost-sensitive learning. Koukouraki et al (2021) applied all three methods, and the result of her study, however, shows that oversampling does a better job in data balancing compared to other techniques.

## 2.9 Performance Metrics

Performance metrics are essential for evaluating the effectiveness of classification models in building damage detection. Common metrics include precision, recall, and F1 score, which are derived from the confusion matrix. Precision measures the proportion of correctly identified positive instances relative to all predicted positives, while recall reflects the ability to correctly detect all actual positives. The F score balances precision and recall, making it particularly effective for imbalanced datasets where minority classes may otherwise be overlooked(Koukouraki et al., 2021; Toba et al., 2023). In remote sensing applications, Overall Accuracy is frequently used but can be misleading for imbalanced datasets, as it tends to favor majority classes(Koukouraki et al., 2021). As a result, precision, recall, and F score are preferred for a comprehensive evaluation of model performance. To better understand how these metrics are defined, consider the following two-class or binary classification scenario:

Table 2-2: A general representation confusion matrix in binary classification

| | | **Predicted** |
|---|---|---|
| **Ground Truth** | True Positive (TP) | False Negative (FN) |
| | False Positive (FP) | True Negative (TN) |

The following are the formulas for calculating performance metrics based on Table 2-2 (Koukouraki et al., 2021; Toba et al., 2023):

- **Precision:** Precision represents Positive Predictive Value (PPV) that measures the proportion of correctly predicted positive instances out of all predicted positives

$$PPV = \frac{TP}{TP - FP} \tag{2.1}$$

- **Recall:** Recall represents Sensitivity or True Positive Rate (TPR) that evaluates the model's ability to correctly identify all actual positive cases

$$TPR = \frac{TP}{TP + FN} \tag{2.2}$$

- **F Score (Harmonic Mean of Precision and Recall):** The F score provides a single metric that balances precision and recall, particularly useful for imbalanced datasets

$$F = 2.\frac{PPV.TPR}{PPV + TPR} \tag{2.3}$$

The F score is the harmonic means of precision and recall, providing a balanced evaluation when there is a trade-off between the two. It ranges from 0 (worst) to 1 (best.

These metrics are derived from the confusion matrix considering a binary classification scenario. However, the confusion matrix would be different in multi-class classification, where True Positives, False Positives, False Negatives, and True Negatives would be adjusted to the specific problems .

To optimize the model's weight during the training, the categorical cross entropy loss is implemented. Categorical Cross-Entropy Loss is a commonly used loss function for multi-class classification problems, where the model outputs a probability distribution over output classes (Koukouraki et al., 2021).

This loss function measures the difference between the true class labels (encoded as one-hot vectors) and the predicted class probabilities. It penalizes incorrect predictions by calculating the negative log likelihood of the predicted probability for the correct class.

For a dataset with mmm samples and $C$ classes, the categorical cross-entropy loss is calculated as equation 2.4.

$$Loss = -\sum_{i=1}^{m} y_i \,.\log \hat{y}_i \tag{2.4}$$

where $y_i$ is the $i$-th class in the model output, $\hat{y}_i$is the corresponding label and m is the total number of classes.

# 3. Data and Methodology

## 3.1 Dataset

The dataset used in this research is based on the Maxar Open Data Program and can be found at https://xview2.org/. This dataset, known as XBD, comprised of Very High Resolution (VHR) satellite imagery that were captured by the WorldView and GeoEye satellites(Koukouraki et al., 2021). Koukouraki et al (2021) further describes the technical specifications in the dataset in Table 3-1. The xBD dataset is a large-scale, publicly available dataset designed for building damage assessment in the context of natural disasters. It has been highly used for training and evaluating machine learning models. This dataset comprises high-resolution pre- and post-disaster imagery with their corresponding labels in JSON, which act as ground truth. The label consists of four building damage class: *no damage*, *minor damage*, *major damage,* and *destroyed*. (Valentijn et al., 2020; Wang et al., 2023; Zhang et al., 2023). The dataset includes data from a variety of disaster scenarios, including earthquakes, floods, hurricanes, wildfires, and volcanic eruptions. Its diverse nature allows models to generalize better across different types of disasters. Covering over 850, 736 annotated buildings, xBD provides a comprehensive benchmark for evaluating deep learning models in the domain of remote sensing and disaster assessment(Bouchard et al., 2022; Valentijn et al., 2020).

As this research focuses on detection of earthquake-induce damaged buildings, the earthquake dataset is isolated for the predictive models. This study aims to conduct prediction only on post - earthquake dataset from Mexico City 2017. The total images collected are 155, and each images has as standard dimensions of 1024 × 1024 pixels (Koukouraki et al., 2021).

Table 3-1: Shows the technical Specifications of xBD dataset

| Parameter | Values |
|---|---|
| Sensor Resolution | 0.66 m |
| GSD | 2.65 m |
| Off-nadir angle | 28.4 degrees |
| Sun azimuth angle | 143.6 degrees |
| Image Width | 1024 pixels |
| Image Height | 1024 pixels |

## 3.2 Exploratory Data Analysis

The Mexico Earthquake (2017) contains 386 VHR images both for pre-event and post-event. Since this study focuses on post-building damage detection, 155 images belonging to post-event are selected. As mentioned in previous sections, each image has its corresponding labels in JSON format, which consists of four damage class: *No Damage*, *Minor Damage*, *Major damage* and *Destroyed.* One example of image overlaid with its label is shown in Figure 3-1.

*Post-Earthquake (Not Annotated*)

*Post-Earthquake (Annotated* )

Figure 3-1: Post-Earthquake image and corresponding labels. In the image right, Green shows *No damage,* yellow shows *Minor Damage,* Orange shows *Major Damage* and red shows *Destroyed buildings.*

The number of instances for each class is shown in Table 3-2

Table 3-2: Number instances per class

| Damage Class | Number of Samples |
|---|---|
| No damage | 6650 |
| Minor damage | 1649 |
| Major damage | 1352 |
| Destroyed | 349 |

## 3.3 Methodology

The primary goal of this research is to investigate the transfer learning and metric-based method to detect damaged buildings. To achieve this goal, a workflow is developed which initially divides the labeled data into *training &validation* and *testing images*. Subsequently, the training and validation dataset undergone a preprocessing, and the training and validation datasets are balanced using resampling technique. Then, the performance of each model is evaluated, and the best model is chosen for the inference. The prediction is made on the parts of testing images based on the best model. Each step is discussed in more detail in later sections. Figure 3-2 shows the graphical representation of the workflow.

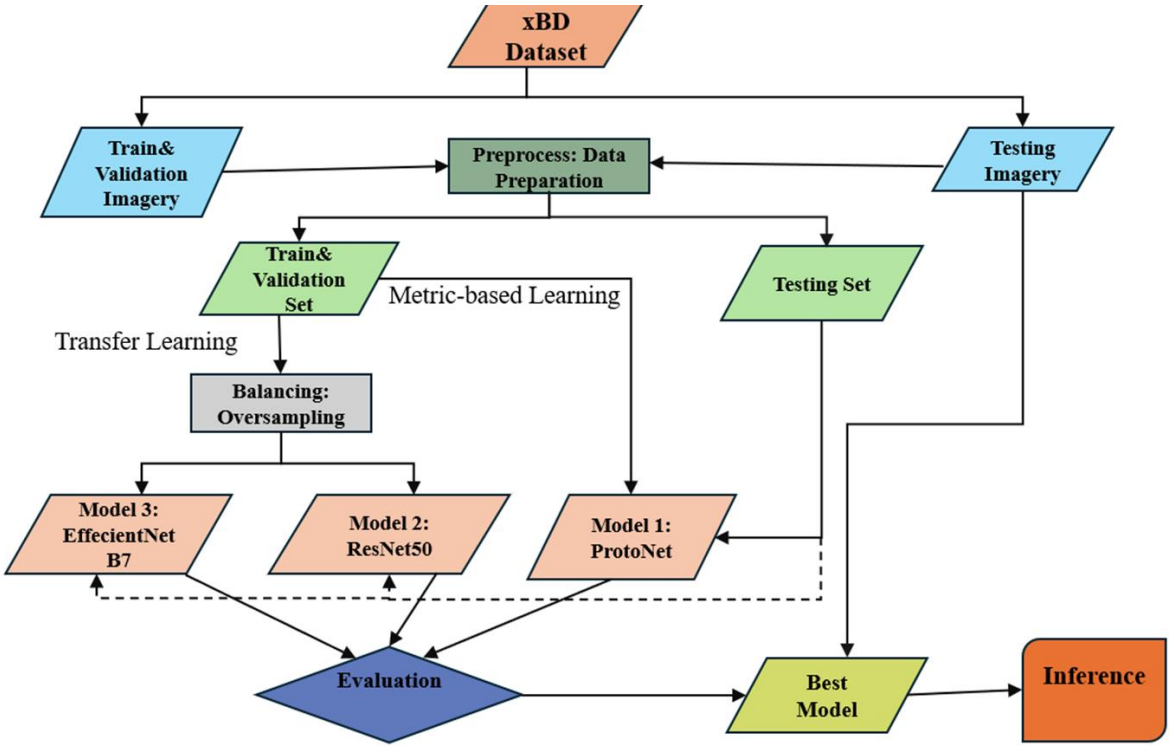


Figure 3-2: Flow chart for the thesis

### 3.3.1 Data preparation

The data preprocessing includes data preparation where the images are cropped by their corresponding labels or JSON file, Figure 3-3. The resulting images are saved in PNG, and their matching labels are saved in CSV file. Now, each cropped images instance acts as one sample for feeding the model. The resulting datasets are divided into three parts: training, validation and test, as shown in Table 3-3. The validation dataset consists of 20 % of the whole dataset. For each damage class, equal number of test samples are assigned since this balance ensures that the performance of the model remain unbiased on each class(Koukouraki et al., 2021).

Figure 3-3: Shows the cropped building instances

### 3.3.2 Oversampling

Data oversampling is a technique used in machine learning to address class imbalances in datasets. When a dataset contains significantly fewer examples of certain classes compared to others, it can lead to biased models that perform poorly in the minority class. Oversampling involves artificially increasing the representation of the minority class by generating synthetic data points or duplicating existing ones. This ensures a more balanced class distribution during the training phase. Oversampling can be accomplished in several ways. Simple replication duplicates existing examples of the minority class by applying transformation. The transformation used in this study includes horizontal flipping, vertical flipping, horizontal and vertical flipping and clockwise rotation and counterclockwise rotation. This method is repeated for instances of minority class until the dataset becomes balanced or each class possesses quarter of the overall dataset. Table 3-3 shows the number of times each minority classes are replicated. This approach can improve the generalization of the model, reducing the likelihood of overfitting. It is important to note the oversampling and data augmentation are applied only to train and validation set.

Table 3-3: Shows training set, validation set, test set and the number of times each minority class is resampled

| Damage Class | Dataset | Training (80%) | Validation (20%) | Test | Resample |
|---|---|---|---|---|---|
| No Damage | 6650 | 5320 | 1330 | 100 | 0 |
| Minor Damage | 1649 | 1319.2 | 329.8 | 100 | 3 |
| Major Damage | 1352 | 1081.6 | 270.4 | 100 | 4 |
| Destroyed | 349 | 279.2 | 69.8 | 100 | 16 |

### 3.3.3 Metric-based Learning: Prototypical Network (ProtoNet)

One of the candidate models used in this study is ProtoNet, which initially introduced by Snell et al. ( 2017). This model is a good choice for the task of few-shot learning in post-earthquake building damage assessment due to their simplicity and efficiency in learning from limited labeled examples. In a study conducted by Koukouraki et al (2021), ProtoNet outperformed other models in detecting new damaged buildings. The model works by embedding input data into a metric space and calculating a "prototype" for each class, which is the mean vector of all embeddings for that class. New data points are classified by finding the class prototype closest to their embedding, typically using a Euclidean distance metric. This approach enables Prototypical Networks to generalize effectively to new classes, as it relies on a simple inductive bias rather than complex architectural designs(Snell et al., 2017).

For this study, Prototypical Networks are implemented with specific parameters tailored to the task. The training setup includes 4-way classification (number of classes), 20 training queries per class, and a 20-shot configuration, where 20 labeled examples per class are provided. In Figure 3-4, you can see the architecture of the implemented model. This network includes two pathways or legs, one pathway calculates the embeddings for the support set, representing the class prototypes, while the other computes the embeddings for the query set. Each leg consists of 3 convolutional (Conv2D) backbone with ReLU activation and 3 Max Pooling layers for feature extraction. There is also one dropout layer to avoid overfitting.

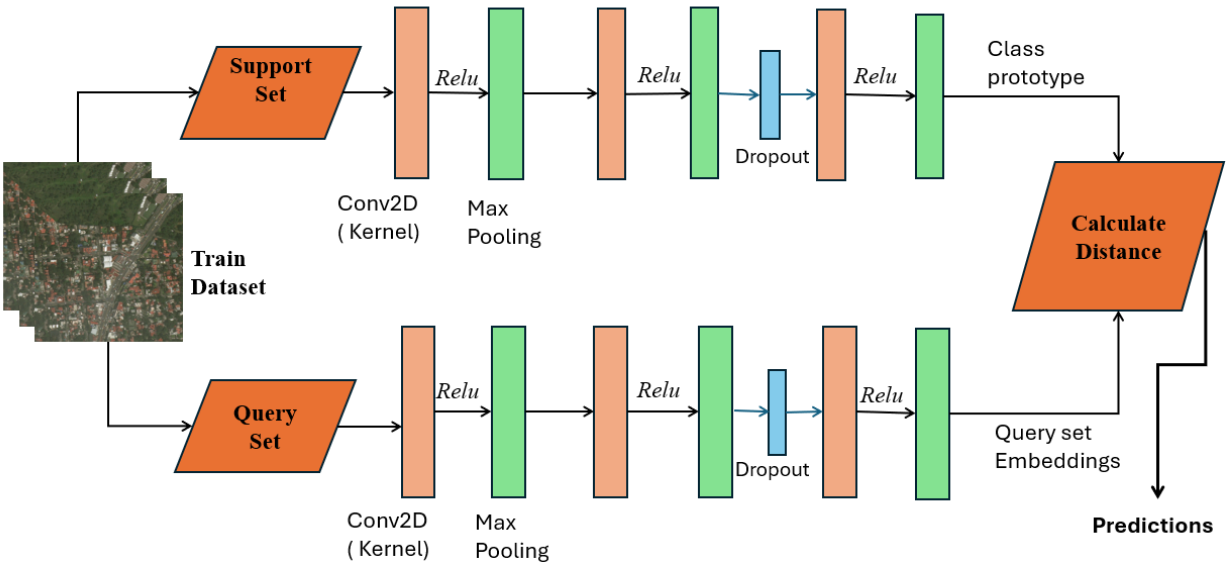


Figure 3-4: Show the ProtoNet's architecture

As mentioned earlier, the original images come with dimensions of 1024 × 1024 pixels and 3 channels (RGB). Training in this dimension would be computationally expensive and not considered a suitable choice for this model. therefore, all images are downscaled as Koukouraki et al (2021) proposes 128 × 128 × 3 dimension for this model. More hyperparameters settings are shown in Table 3-4

### 3.3.4 Transfer Learning: ResNet 50

ResNet, a deep convolutional neural network, has emerged as an effective model for building damage detection due to its residual learning framework (Baral et al., 2024; He et al., 2016). Another model candidate for this study refers to ResNet50, a Deep Neural Network (DNN) architecture trained on large image called *imagenet.* DNN typically consists of two main components, a shallow CNN with a few layers and deep architecture, such as ResNet50, with several layers( Koukouraki et al . ,2021). In this study, transfer learning method is used which uses ResNet50's learned weights to identify general patterns such as edges, textures, and shapes. These features are extracted through its layers, which are frozen to retain the pre-trained knowledge. New task-specific layers, shallow CNN, are added on top of ResNet50 to adopt in different problems. Figure 3-5 shows the learning process ResNet50.

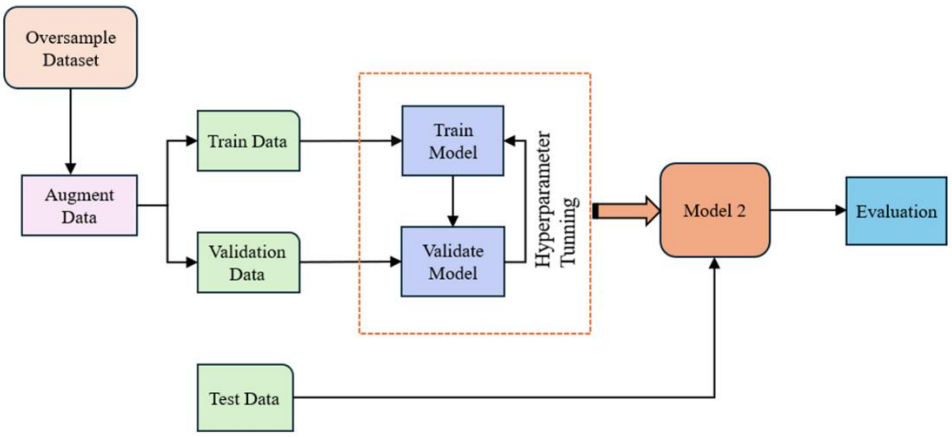


Figure 3-5: shows the learning process ResNet50

This architecture, Figure 3-6, processes 32 batches of input images using a hybrid design combining a shallow CNN and ResNet50 weights. It begins with a Conv2D layer followed by Max Pooling. Two additional Conv2D layers are used, each followed by Max Pooling. A Flatten layer converts feature maps into a 1D vector. Features from ResNet50 are fused with the shallow CNN outputs. The concatenated features pass through three fully connected layers with ReLU activation, followed by a SoftMax output for prediction.

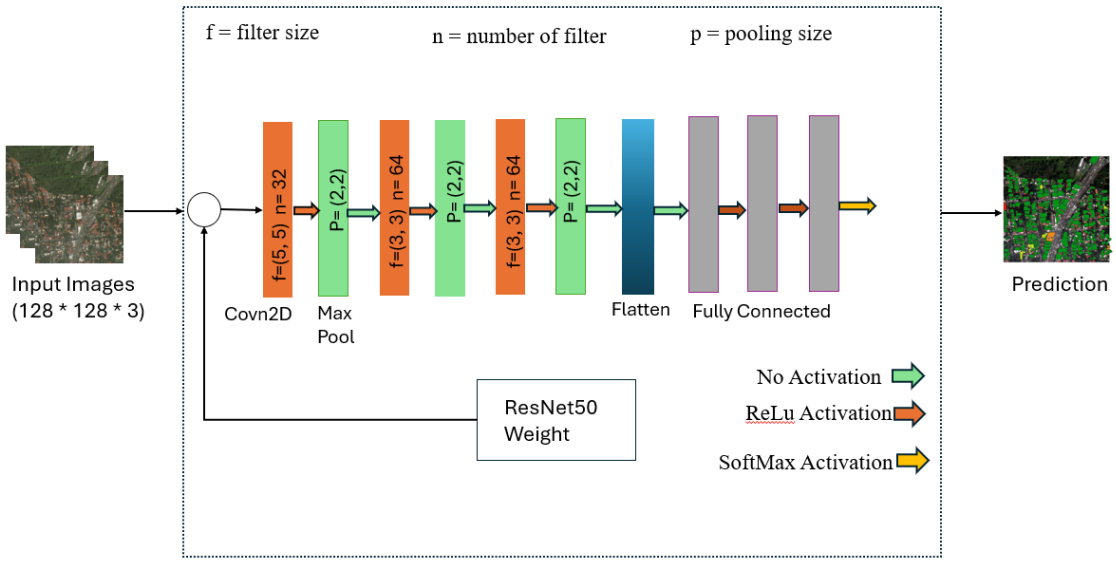


Figure 3-6: Shows a shallow CNN architecture for ResNet 50

### 3.3.5 Transfer Learning: EffecientNetB7

EfficientNetB7 is the largest and most powerful variant of the EfficientNet family, a deep learning model designed for efficient and accurate image classification. Developed by Google, EfficientNetB7 uses a compound scaling method, which optimally balances network depth, width, and resolution to enhance performance while maintaining computational efficiency (Ahmadi et al., 2023). Compared to other CNN architecture like ResNet, EfficientNetB7 achieves higher accuracy with fewer parameters and lower computational cost. It is widely used for computer vision tasks, including object detection, segmentation, and medical imaging. Due to its efficiency, it is ideal for real-world applications requiring high precision with minimal resources. In this study, this model is used as a pre-trained model where its original layers are frozen, and only its weight plus a certain number of custom layers are used for feature extraction from the input images. The training process for this model and custom layers added on top of its weight is similar to the ResNet50, Figure 3-5, 3-6. However, the input images for EfficientNetB7 have 254 × 256× 3 dimensions.

Table 3-4 provides hyperparameter settings for three deep learning models: ProtoNet, ResNet50, and EfficientNetB7. It outlines the input image size, showing that both ProtoNet and ResNet50 use 128 × 128 × 3, while EfficientNetB7 processes a larger input size of 254 × 256 × 3. The initial learning rates are 0.0001 for ProtoNet and 0.001 for the other models. All models utilize the Adam optimizer with parameters $\beta_1$=0.9, $\beta_2$=0.99 and employ Categorical Cross-Entropy as the loss function. These settings define the training configuration and impact model performance in post-earthquake building damage assessment.

Table 3-4: Summary of hyperparameter settings for Model1, Model 2 and Model 3.

| **Models** | **Input Image Size** | **Initial Learning Rate** | **Optimizer** | **Loss Function** |
|---|---|---|---|---|
| Model 1 (ProtoNet) | 128 × 128 × 3 | 0.0001 | Adam ($\beta$1 =0.9, $\beta$2 =0.99) | Categorical Cross Entropy |
| Model2 (ResNet50) | 128 × 128 × 3 | 0.001 | Adam ($\beta$1 =0.9, $\beta$2 =0.99) | Categorical Cross Entropy |
| Model3 (EffecientNetB7) | 256 × 256× 3 | 0.001 | Adam ($\beta$1 =0.9, $\beta$2 =0.99) | Categorical Cross Entropy |

# 4. Results and Discussion

This section illustrates the learning and evaluation process of the models, in which the number of epochs and how models perform in the training and validation set will be discussed. Moreover, it discusses how the trained models perform in unseen data and predict damage in different areas.

## 4.1 Learning process

The models are set to be trained for 60 epochs until the model converges to the minimum value or validation loss is less than 0.1 considering the presence of limited and imbalanced data. However, a scheduler is assigned to stop the training if the model does not learn anymore, specifically if there is no improvement in the validation loss in 15 epochs in arrow. The validation loss is an indicator of how a model learns from the training data or converges to a minimum across epochs, and it also shows whether the model is undergoing overfitting or not. This validation loss is the difference between predicted and actual values calculated using cross validation technique. In this study, categorical cross entropy is used for validation.

## Model 1

The accuracy and loss curves, Figure 4-1, provide insight into the learning behavior of Model 1. The left plot, showing accuracy, reveals that training accuracy (blue dashed line) increases consistently over 50 epochs, reaching approximately 75%. However, validation accuracy (green solid line) plateaus early around 64%, indicating that the model may have stopped improving on unseen data after an initial learning phase. The right plot, displaying loss, further supports this observation. The training loss (red dashed line) steadily decreases, which is expected as the model learns patterns in the data. However, validation loss (orange solid line) shows minimal improvement after about 15–20 epochs, stabilizing around 0.9. This gap between training and validation loss suggests the presence of overfitting, where the model learns training-specific patterns that do not generalize well. The best model, which has the minimum validation loss, is marked at epoch 50. This model is saved and used for testing and prediction.

Overall, Model 1 exhibits signs of overfitting, and its validation accuracy does not improve beyond a certain point, indicating limited generalization to unseen data. This mainly stems from the fact that the model is fed with around *25000* samples, after resampling, which is considered “too few” for FSL.

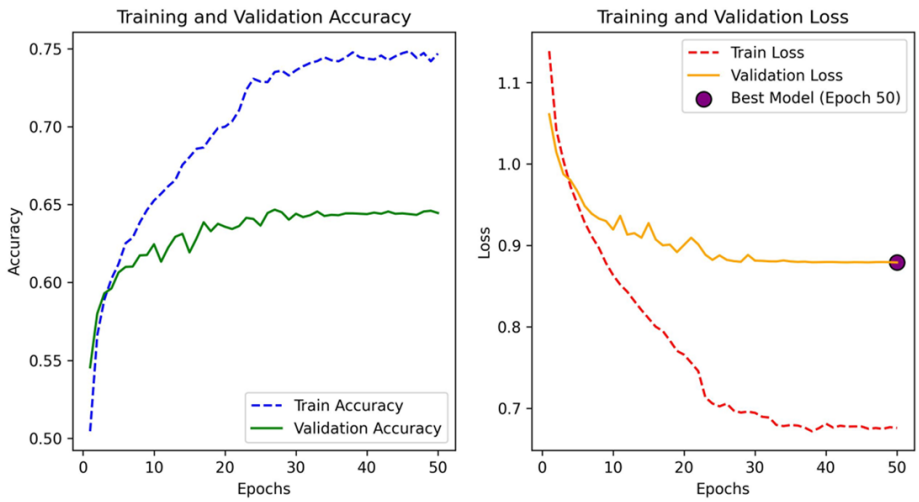


Figure 4-1: Shows the accuracy and loss over epochs for Model 1

## Model 2

The accuracy and loss curves, Figure 4-2, of Model 2 suggest a more stable and well-generalized learning process. The training accuracy (blue dashed line) and validation accuracy (green solid line) are closely aligned, with validation accuracy slightly exceeding training accuracy at certain points. This indicates that the model is not overfitting significantly and generalizes well to unseen data. The model reaches around 55-56% accuracy by epoch 50. The training and validation loss curves further support this. The training loss (red dashed line) and validation loss (orange solid line) both decrease rapidly in the initial epochs and gradually stabilize, maintaining a small gap between them. This suggests that the model has learned effectively without excessive overfitting. The best model is marked at epoch 47, indicating that early stopping may not be necessary since there are no clear signs of performance deterioration.

Compared to Model 1, Model 2 demonstrates better generalization, as evidenced by the closer alignment between its training and validation accuracy and loss curves. However, the model's performance plateaus after a certain number of epochs, as validation loss does not reach its

optimal minimum. This suggests that while Model 2 avoids overfitting more effectively than Model 1, further optimization may be needed to enhance its overall performance.

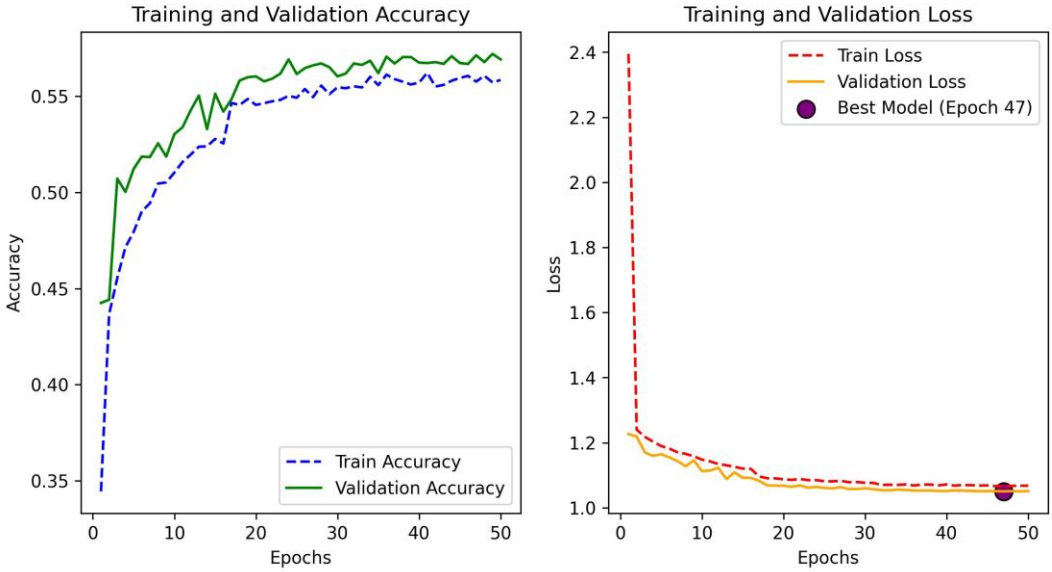


Figure 4-2: Shows the accuracy and loss over epochs for Model 2

## Model 3

The accuracy and loss curves of Model 3, Figure 4-3, indicate that it exhibits signs of overfitting, similar to Model 1. The training accuracy (blue dashed line) steadily increases, reaching approximately 75%, whereas the validation accuracy (green solid line) stagnates around 64% after an initial increase. This suggests that while the model learns patterns effectively on the training data, it does not generalize well to unseen data. The loss curves provide further evidence of overfitting. The training loss (red dashed line) continues to decrease consistently, while the validation loss (orange solid line) stabilizes at around 0.9 after early epochs. The widening gap between the training and validation loss suggests that the model is memorizing the training data rather than learning generalizable patterns.

Model 3 behaves similarly to Model 1, while Model 2 generalizes better due to its closer alignment of training and validation accuracy. Unlike Model 2, which demonstrates the most balanced learning stability, Models 1 and 3 achieve higher training accuracy and validation loss. A common limitation affecting all models is the constraint of limited data, which impacts their overall performance and generalization capability.

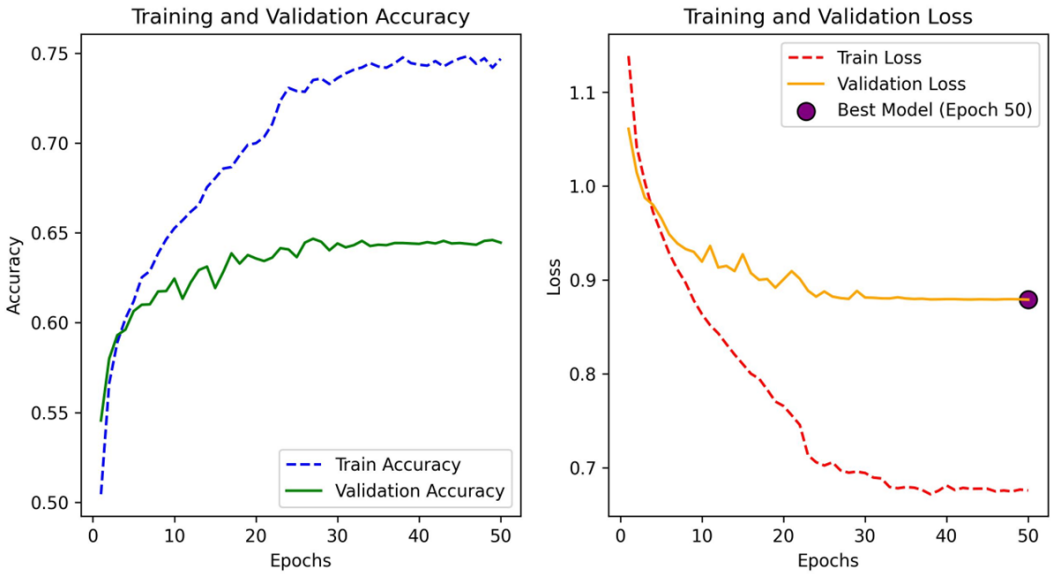


Figure 4-3: Shows the accuracy and loss over epochs for Model 3

## 4.2 Performance Metrics

Performance metrics are quantitative measures used to evaluate how well a machine learning model performs in achieving its intended task. These metrics provide insights into the model's ability to generalize from training data to unseen (validation/test) data, which helps identify whether the model is effective, overfitting, or underfitting. it measures the proportion of correctly classified instances out of the total instances. It provides an understanding of how well the model performs overall. Additionally, it represents the error between the predicted and true values. Lower loss indicates better model performance. In this study, the performance of a model is measured and presented by *Precision, Recall, F-Score and Confusion Matrix, see section 2.9*. And each model is tested against equal 100 samples for each class, see section 3.3.2.

## Model 1

Model 1's performance in classifying damage levels can be analyzed using the confusion matrix and precision-recall metrics.

From the confusion matrix, Table 4-1, "Destroyed" has the highest True Positive (78) and the least False Positive and False Negative, thus the best performance across all metrics. Accordingly, It has the highest precision, 0.72 and recall, 0.78, leading to an F-score of 0.75, Table 4-2. This indicates the model is effective in detecting severely damaged structures. For "Major-Damage," the model

achieves a precision of 0.66 and recall of 0.58, suggesting that while it makes relatively accurate positive predictions, it still misses some cases.

The "Minor-damage" class has a precision of 0.56 and recall of 0.59, meaning the model struggles to differentiate it from other categories, potentially confusing it with “No damage” and “Major-Damage.” "No damage" has the weakest performance, with a precision and recall of 0.52 and 0.51, respectively, indicating that the model often misclassifies undamaged structures.

Overall, while the model performs well in detecting severe damage, it struggles with minor and no-damage cases, likely due to overlapping features. Improvements could include refining class boundaries or incorporating additional features to reduce misclassification

Table 4-1: A Confusion Matrix of damage classes for Model 1

**Predicted**

| **Actual** | **No damage** | **Minor-damage** | **Major-Damage** | **Destroyed** |
|---|---|---|---|---|
| **No damage** | **51** | 30 | 7 | 12 |
| **Minor-damage** | 26 | **59** | 7 | 8 |
| **Major-Damage** | 20 | 12 | **58** | 10 |
| **Destroyed** | 2 | 4 | 16 | **78** |
| **Overall Accuracy: 0.63** | | | | |

Table 4-2: Performance Metrics based on Precision, Recall and F-score for Model 1

| Class | Precision | Recall | F-score |
|---|---|---|---|
| **No damage** | 0.52 | 0.51 | 0.51 |
| **Minor-damage** | 0.56 | 0.59 | 0.57 |
| **Major-Damage** | 0.66 | 0.58 | 0.62 |
| **Destroyed** | 0.72 | 0.78 | 0.75 |

## Model 2

The performance tables of Model 2 show a decline compared to Model 1, particularly in distinguishing less severe damage levels. The confusion matrix and performance metrics, Table 4-3 and Table 4-4, indicate that "Destroyed" remains the best predicted class, with a precision of 0.72 and recall of 0.65, though slightly lower than in Model 1. This suggests the model is still relatively effective at identifying severe damage but has more misclassifications.

For "Major-Damage," the precision is 0.57 and recall is 0.47, meaning the model has a weaker ability to correctly detect and retrieve instances of this category. It confuses many "Major-Damage" instances with "Minor-Damage" or "Destroyed." "Minor-Damage" has a precision of 0.40 and recall of 0.56, highlighting significant misclassification with other categories, particularly "No Damage" and "Major-Damage." The worst-performing class is "No Damage," with a precision of 0.43 and recall of 0.37, suggesting the model struggles to differentiate undamaged structures from minor damage.

In summary, Model 2 underperforms compared to Model 1, particularly in predicting "No Damage" and "Minor-Damage." The high misclassification rates indicate a need for feature improvement or rebalancing the training data to enhance differentiation between damage classes.

Table *4-3*: Performance Metrics based on Precision, Recall and F-score for Model 2

| Class | Precision | Recall | F-score |
|---|---|---|---|
| **No damage** | 0.43 | 0.37 | 0.40 |
| **Minor-damage** | 0.40 | 0.56 | 0.46 |
| **Major-Damage** | 0.57 | 0.47 | 0.51 |
| **Destroyed** | 0.72 | 0.65 | 0.68 |

Table 4-4: A Confusion Matrix of damage classes for Model 2

| | | **Predicted** | | |
|---|---|---|---|---|
| | **No damage** | **Minor-damage** | **Major-Damage** | **Destroyed** |
| **Actual** **No damage** | 37 | 47 | 5 | 11 |
| **Minor-damage** | 26 | 56 | 12 | 6 |
| **Major-Damage** | 19 | 26 | 47 | 8 |
| **Destroyed** | 4 | 12 | 19 | 65 |
| **Overall Accuracy: 0.51** | | | | |

## Model 3

Model 3 exhibits improvements in some areas but still struggles with class differentiation. The confusion matrix and performance metrics show that the model performs best in identifying "Destroyed" instances, with the highest precision (0.78) and a recall of 0.60, yielding an F-score of 0.68, Table 4-5 and Table 4-6. However, its recall is slightly lower than in Model 1.

"Minor-Damage" has a high recall of 0.73, meaning the model is effective at identifying most instances of this class, but it suffers from a low precision of 0.41. This suggests significant misclassification, as seen in the confusion matrix, where many "No Damage" and "Major-Damage" instances are labeled as "Minor-Damage." "No Damage" has a precision of 0.61, an improvement over Models 1 and Model 2, but with a recall of 0.42, showing a higher tendency to miss actual "No Damage" instances. "Major-Damage" remains a challenge, with a precision of 0.57 and recall of 0.43, indicating misclassification with "Minor-Damage" and "Destroyed."

Model 3 performs slightly better in "No Damage" and "Minor-Damage" detection compared to Model 2 but still struggles with class overlap. The next section illustrates more which models have higher performance.

Table 4-5: Performance Metrics based on Precision, Recall and F-score for Model 3

| Class | Precision | Recall | F-score |
|---|---|---|---|
| **No damage** | 0.61 | 0.42 | 0.50 |
| **Minor-damage** | 0.41 | 0.73 | 0.53 |
| **Major-Damage** | 0.57 | 0.43 | 0.49 |
| **Destroyed** | 0.78 | 0.60 | 0.68 |

Table 4-6: A Confusion Matrix of damage classes for Model 3

**Predicted**

| **Actual** | **No damage** | **Minor-damage** | **Major-Damage** | **Destroyed** |
|---|---|---|---|---|
| **No damage** | 42 | 48 | 3 | 7 |
| **Minor-damage** | 12 | 73 | 12 | 3 |
| **Major-Damage** | 12 | 38 | 43 | 7 |
| **Destroyed** | 3 | 19 | 18 | 60 |
| **Overall Accuracy: 0.55** | | | | |

## Comparison

Model 1 demonstrates the most balanced performance across all damage categories, with strong precision and recall for "Destroyed" (0.72 and 0.78, respectively) and "Major-Damage" (0.66 precision) Figure 4-4. It maintains a relatively high F-score of 0.62, showing it effectively captures the majority of "Major-Damage" cases with reasonable accuracy. However, its performance in "No Damage" (0.52 precision, 0.51 recall) and "Minor-Damage" (0.56 precision, 0.59 recall) indicate that it struggles to differentiate lower levels of damage, leading to frequent misclassifications.

Model 2 is the weakest among the three, with lower precision and recall scores across all categories. It struggles particularly in identifying "No Damage" (0.43 precision, 0.37 recall) and "Minor-Damage" (0.40 precision), resulting in higher misclassification rates. Though it maintains a decent performance in detecting "Destroyed" (0.72 precision), it does not significantly improve upon Model 1. Model 3 introduces improvements in "No Damage" precision (0.61) and significantly increases recall for "Minor-Damage" (0.73), meaning it captures more instances of minor damage. However, it misclassifies more cases into this category, leading to lower precision (0.41). While it maintains competitive scores in "Destroyed," it struggles with "Major-Damage" (0.57 precision, 0.43 recall), showing difficulty in correctly classifying heavily damaged structures.

Based on comparative analysis, Model 1 is the best overall model, followed by Model 3, due to its balanced performance across all damage classes. Table 4-7 further supports this as it shows most of the higher averaged values of performance metrics and higher values of overall accuracy is populated for Model 1

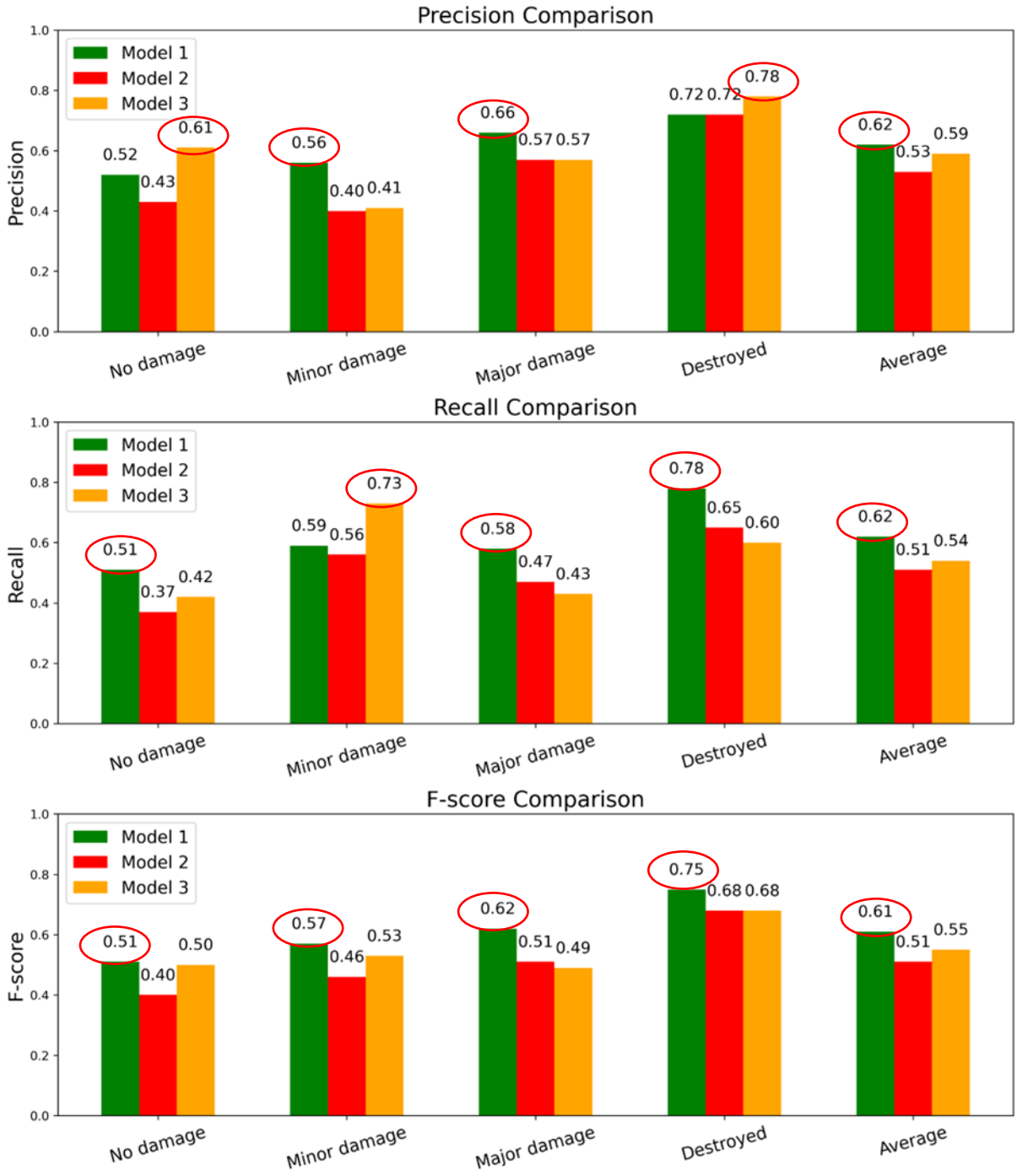


*Figure 4-4: A comparative view performance metrics for the 3 models*

The average values of each metric for each model are and overall accuracy is depicted in Table 4-7

Table 4-7: Averaged metrics' values and overall accuracy for models

| Metrics | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| Precision | 0.62 | 0.53 | 0.59 |
| Recall | 0.62 | 0.51 | 0.54 |
| F-Score | 0.61 | 0.51 | 0.55 |
| Overall Accuracy | 0.63 | 0.51 | 0.55 |

Given its higher values, frequency and balance across Precision, Recall, and F-score, Model 1 emerges as the best overall performer followed by Model 3.

## 4.3 Inference

In this section, the prediction is carried out using the best model, Model 1, on the test imagery. To understand how the model performs on predicting the test image, we test on different scenarios, rural areas, urban areas and dense urban areas. In each scenario, the satellite imagery is overlaid with three key components: a) predicted damage classifications, b) true damage labels, and c) the differences between predictions and ground truth.

**Rural Areas**

The Figure 4-5 shows how the model performs in detecting future or unseen damaged buildings in a rural area. As shown in map **c**, which portrays misclassification among the classes, the model was able to correctly classify most of **No Damage** labels, 51.4%. However, the significant numbers of labels belong to no damaged misclassified as other classes. As shown on map **c,** majority of the confusion occurred between close or similar classes. For example, there are more misclassifications between **No Damage** and **Minor Damage** labeled as 1 in map **c.** In contrast, as the severity-based classes become more distant, the confusion between the classes decreases. This prediction cannot represent how the model performs in detecting damaged buildings since the test dataset is heavily imbalanced in favor of the **No Damage** class. For instance, the only sample originally belongs to Minority Class correctly classified in predicted map, which is highlighted with red circle.

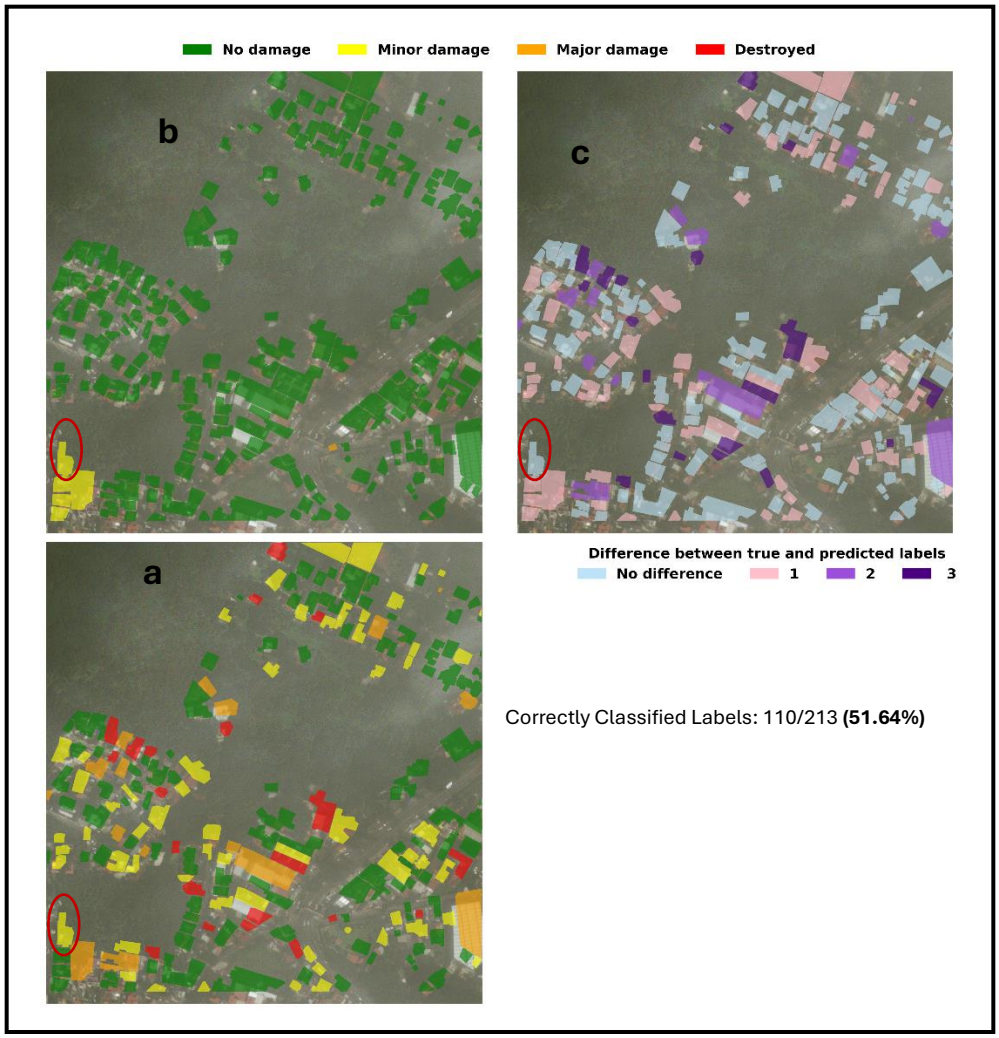


Figure 4-5: Prediction based on Model 1 on test image from Mexico Earthquake (2017). The satellite image is overlaid with a) predicted label, b) ground truth labels, and c) the differences between predictions and labels.

## Urban Area

Figure 4-6 shows the prediction made in an **urban** area based on Model 1. As shown in map c, the model incorrectly predicted most of the labels. As opposed to Figure 4-6, most of the confusion occurred between most dissimilar classes. For example, most labels that belong to **No Damage** are falsely classified as **Destroyed**. However, there are still some misclassifications between similar classes beyond the No Damage class. Despite the presence of limited number of samples in **Minor Damage**, **Major Damage** and **Destroyed,** the model was able to correctly predict at least one sample from each these classes. In total, the model correctly predicts **34.58%** of the labels. This demonstration suggests that the model's performance varies across the damage classes as it is further supported by the confusion matrixes in the previous sections.


Figure 4-6: Prediction based on Model 1 on test image from Mexico Earthquake (2017). The satellite image is overlaid with a) predicted label, b) ground truth labels, and c) the differences between predictions and labels

## Dense Urban Area

In Figure 4-7, the prediction is carried out on a dense **urban area.** Almost all the buildings in this area belong to **No Damage**, but they majority of them are classified as the most similar class, **Minor Damage.** However, the model was able to correctly predict **52.35%** of the total sample. As mentioned earlier, this is not an inclusive representation of how Model 1 performs in dense urban areas since the distribution of the true samples are not even for each class.


Figure 4-7: Prediction based on Model 1 on test image from Mexico Earthquake (2017). The satellite image is overlaid with a) predicted label, b) ground truth labels, and c) the differences between predictions and labels

### 3.4.1 Limitation

Few-shot learning (FSL) offers significant advantages for post-earthquake building damage assessment, particularly in scenarios where labeled data is scarce. However, it comes with several limitations. It is noteworthy to elaborate on some limitations of this study and FSL that would affect the performance of models in detecting damaged buildings. First of all, FSL relies on a minimal number of labeled examples to generalize predictions. However, in post-earthquake scenarios, the available data was insufficient to capture the variability in damage patterns across different damage class and geographic locations. The lack of diverse samples led to the reduction of model generalization. Additionally, this study involved multiple damage categories (e.g., no damage, minor damage, major damage, and destroyed). FSL models struggle with multi-class classification, especially distinguishing between similar damage categories. Another limitation is the presence of an imbalanced data set for prediction which prevents us from generalizing the model performance across the damage classes. The last but not least limitation would be randomly selecting samples for training and validation which might fail to adequately represent the full spectrum of building damage patterns. Additionally, random sampling may fail to capture rare yet critical types of damage, resulting in unreliable assessments when applied in real-world disaster scenarios.

# 5. Conclusion

This study successfully explored the application of Few-Shot Learning (FSL), combining transfer learning, and metric-based approaches for multi-class post-earthquake building damage assessment using VHR pan-sharpened satellite imagery. By addressing challenges such as data scarcity, class imbalance, and multi-class classification complexity, the research contributes to automated damage assessment methodologies. The literature review underscored the growing importance of integrating machine learning with remote sensing for disaster response. It highlighted the potential of advanced methods such as Prototypical Networks and pre-trained deep learning architectures like ResNet50 and EffecientNetB7 to generalize across diverse disaster scenarios. These insights formed the basis for selecting methodologies aligned with the study's objectives. As result, Prototypical Networks, ResNet50, and EffecientNetB7 were implemented and evaluated using comprehensive performance metrics, including Precision, Recall, and F-scores.

The methodological framework employed the XBD dataset, isolating earthquake-related instances for model training, validation, and testing. To counter the imbalanced distribution of damage classes, oversampling and data augmentation techniques were employed, ensuring a balanced dataset critical for fair model evaluation. Oversampling and data augmentation emerged as an effective technique for mitigating class imbalance, particularly for minority classes.

The learning behaviors of the models demonstrate that Model 3 exhibits similar behavior to Model 1, whereas Model 2 demonstrates superior generalization due to the closer alignment between its training and validation accuracy. Unlike Model 2, which maintains the most stable learning performance, Models 1 and 3 show higher training accuracy but also greater validation loss. A shared limitation among all models is the restricted availability of data, which affects their overall effectiveness and ability to generalize.

Based on the comparative analysis of precision, recall, F-score, and overall accuracy, Model 1 (Prototypical Network) demonstrates the best performance among the three models. It achieves the highest scores across all metrics, which outperforms Model 2 and Model 3. Model 1 also maintains a balanced trade-off between precision and recall, particularly excelling in identifying destroyed buildings. While Model 2 and Model 3 show moderate performance, they struggle with lower recall and precision in certain classes.

The inference results demonstrate that Model 1's performance varies across different environments due to class imbalance and misclassification between similar categories. In rural areas, around half of labels were correctly classified, but frequent confusion occurred with Minor Damage. In urban areas, only 34.58% of labels were accurately predicted, with greater misclassification across distant classes. In dense urban areas, the highest accuracy was achieved, but most No Damage buildings were misclassified as Minor Damage.

After fulfilling the objectives of the research, this study provides the following answers to the proposed research questions:

1. How to address the challenges of limited and imbalanced data, where data is limited, and one class has multiple times more samples than other classes?

In this study, we addressed the challenge of data scarcity and class imbalance through Few-Shot Learning (FSL), data augmentation, and oversampling techniques. Oversampling increased instances of minority classes, while data augmentation (flipping, rotation, scaling, etc.) introduced variations in the training and validation set. Moreover, Model 1 (Prototypical Network) effectively mitigated class imbalance by learning distance metrics between class prototypes, which reduces dependency on large datasets. Although transfer learning and metric-based learning techniques are proven to be effective in the presence of limited data, data scarcity remained a limitation since the amount of data used in this study is "too few".

2. As this study is concerned with multi-class problems, how well do the proposed models perform class-wide?

All three models showed different levels of performance across damage classes. Model 1, the Prototypical Network, emerged as the best overall performer, particularly excelling in detecting "Destroyed" buildings with a precision of 0.72 and a recall of 0.78. Model 2, ResNet50, demonstrated better generalization but struggled to distinguish between "No Damage" and "Minor Damage". Model 3, EfficientNetB7, improved recall for "Minor Damage" (0.73) but it had

low precision (0.41), that frequently misclassifying other categories. Across all models, there was consistent confusion between similar damage classes, such as No Damage and Minor Damage, highlighting the challenges of multi-class classification in post-earthquake damage assessment.

3. Comparing the results based on performance metrics, which models are the winner?

Model 1, the Prototypical Network, proved to be the best-performing model, which achieved the highest overall accuracy of **63%.** It maintained a well-balanced trade-off between precision and recall and demonstrated the strongest ability to detect destroyed buildings. Model 3, EfficientNetB7, ranked second with **55%** accuracy, which shows improvements over Model 2 but still struggling with class misclassification. Model 2, ResNet50, had the lowest performance, with only **51%** accuracy and significant confusion between "No Damage" and "Destroyed" buildings. The comparative analysis of precision, recall, and F-score further demonstrates Model 1's effectiveness in post-earthquake building damage assessment Figure 4-4.

4. Running the prediction based on the best model, to what extent is the prediction map indicative of the actual severity of the damage suffered by a region?

Model 1's predictions were impacted by class imbalance, making it less reliable in areas where certain classes were underrepresented. In rural areas, it correctly classified **51.4%** of the labels but frequently confused No Damage with Minor Damage, Figure 4-5 . In urban areas, accuracy dropped to **34.58%,** with significant misclassification between No Damage and Destroyed buildings, Figure 4-6. In dense urban areas, it achieved **52.35%** accuracy, though No Damage was often misclassified as Minor Damage, Figure 4-7. While the prediction map provides a general indication of damage severity, it is not entirely reliable in highly imbalanced areas. This highlights the need for more balanced datasets and improved class differentiation to improve model accuracy.

## 5.1 Implications and Future Directions

The findings emphasize the practical utility of Few-Shot Learning and transfer learning in rapid disaster response, which could offer scalable and adaptable solutions for post-earthquake scenarios. As Prototypical performed better than other models, it requires further investigation. Future research could also explore integrating hybrid architecture such as highbred data balancing. For example, future research would mix cost sensitive with oversampling method for data balancing. Additionally, extending this framework to other disaster types, such as hurricanes or floods, would validate its broader applicability. This study demonstrates a good step toward efficient, automated damage assessment, showcasing the potential of machine learning to support humanitarian efforts in disaster-stricken regions.

## Data and Sources

All the scripts can be found in the following repository:

https://github.com/Enayat1912/FSL_for_building_damage_detection

Another helpful scripts:

https://github.com/DIUx-xView/xView2_baseline

The original dataset can be found here:

https://xview2.org

# 6. Bibliography


Ahmadi, S. A., Mohammadzadeh, A., Yokoya, N., & Ghorbanian, A. (2023). BD-SKUNet: Selective-Kernel UNets for Building Damage Assessment in High-Resolution Satellite Images. *Remote Sensing*, *16*(1), 182. https://doi.org/10.3390/rs16010182

Amini Amirkolaee, H., & Arefi, H. (2019). CNN-based estimation of pre- and post-earthquake height models from single optical images for identification of collapsed buildings. *Remote Sensing Letters*, *10*(7), 679–688. https://doi.org/10.1080/2150704X.2019.1601277

Baral, A., Singh, V., & Lath, A. (2024). Evaluating the Performance of ResNet-50 and GoogleNet for Damage Detection and Classification. *2024 4th International Conference on Sustainable Expert Systems (ICSES)*, 1721–1726. https://doi.org/10.1109/ICSES63445.2024.10763274

Bouchard, I., Rancourt, M.-È., Aloise, D., & Kalaitzis, F. (2022). On Transfer Learning for Building Damage Assessment from Satellite Imagery in Emergency Contexts. *Remote Sensing*, *14*(11), 2532. https://doi.org/10.3390/rs14112532

Braik, A. M., & Koliou, M. (2024). Automated building damage assessment and large-scale mapping by integrating satellite imagery, GIS, and deep learning. *Computer-Aided Civil and Infrastructure Engineering*, *39*(15), 2389–2404. https://doi.org/10.1111/mice.13197

Duarte, D., Nex, F., Kerle, N., & Vosselman, G. (2018). Multi-Resolution Feature Fusion for Image Classification of Building Damages with Convolutional Neural Networks. *Remote Sensing*, *10*(10), 1636. https://doi.org/10.3390/rs10101636

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Hilliard, N., Phillips, L., Howland, S., Yankov, A., Corley, C. D., & Hodas, N. O. (2018). *Few-Shot Learning with Metric-Agnostic Conditional Embeddings* (No. arXiv:1802.04376). arXiv. https://doi.org/10.48550/arXiv.1802.04376

Ji, M., Liu, L., Zhang, R., & F. Buchroithner, M. (2020). Discrimination of Earthquake-Induced Building Destruction from Space Using a Pretrained CNN Model. *Applied Sciences*, *10*(2), 602. https://doi.org/10.3390/app10020602

Koukouraki, E., Vanneschi, L., & Painho, M. (2021). Few-Shot Learning for Post-Earthquake Urban Damage Detection. *Remote Sensing*, *14*(1), 40. https://doi.org/10.3390/rs14010040

Lin, Q., Ci, T., Wang, L., Mondal, S. K., Yin, H., & Wang, Y. (2022). Transfer Learning for Improving Seismic Building Damage Assessment. *Remote Sensing*, *14*(1), 201. https://doi.org/10.3390/rs14010201

Ma, H., Liu, Y., Ren, Y., Wang, D., Yu, L., & Yu, J. (2020). Improved CNN Classification Method for Groups of Buildings Damaged by Earthquake, Based on High Resolution Remote Sensing Images. *Remote Sensing*, *12*(2), 260. https://doi.org/10.3390/rs12020260

Ma, H., Liu, Y., Ren, Y., & Yu, J. (2019). Detection of Collapsed Buildings in Post-Earthquake Remote Sensing Images Based on the Improved YOLOv3. *Remote Sensing*, *12*(1), 44. https://doi.org/10.3390/rs12010044

Settou, T., Kholladi, M.-K., & Ben Ali, A. (2022). Improving damage classification via hybrid deep learning feature representations derived from post-earthquake aerial images. *International Journal of Image and Data Fusion*, *13*(1), 1–20. https://doi.org/10.1080/19479832.2020.1864787

Snell, J., Swersky, K., & Zemel, R. S. (2017). *Prototypical Networks for Few-shot Learning* (No. arXiv:1703.05175). arXiv. https://doi.org/10.48550/arXiv.1703.05175

Toba, H., Bunyamin, H., Widyaya, J. E., Wibisono, C., & Haryadi, L. S. (2023). Masking preprocessing in transfer learning for damage building detection. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *12*(2), 552. https://doi.org/10.11591/ijai.v12.i2.pp552-559

Tsai, F. J., & Lin, S.-Y. (2024). A Class Distance Penalty Deep Learning Method for Post-disaster Building Damage Assessment. *KSCE Journal of Civil Engineering*, *28*(5), 2005–2019. https://doi.org/10.1007/s12205-024-1587-1

Valentijn, T., Margutti, J., Van Den Homberg, M., & Laaksonen, J. (2020). Multi-Hazard and Spatial Transferability of a CNN for Automated Building Damage Assessment. *Remote Sensing*, *12*(17), 2839. https://doi.org/10.3390/rs12172839

Wang, Y., Jing, X., Xu, Y., Cui, L., Zhang, Q., & Li, H. (2023). Geometry-guided semantic segmentation for post-earthquake buildings using optical remote sensing images. *Earthquake Engineering & Structural Dynamics*, *52*(11), 3392–3413. https://doi.org/10.1002/eqe.3966

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2021). Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Computing Surveys*, *53*(3), 1–34. https://doi.org/10.1145/3386252

Wu, C., Zhang, F., Xia, J., Xu, Y., Li, G., Xie, J., Du, Z., & Liu, R. (2021). Building Damage Detection Using U-Net with Attention Mechanism from Pre- and Post-Disaster Remote Sensing Datasets. *Remote Sensing*, *13*(5), 905. https://doi.org/10.3390/rs13050905

Yu, C., Hu, B., Cheng, X., Yin, G., & Wang, Z. (2023). Remote sensing building damage assessment with a multihead neighbourhood attention transformer. *International Journal of Remote Sensing*, *44*(16), 5069–5100. https://doi.org/10.1080/01431161.2023.2242590

Zhang, Y., Yang, G., Gao, A., Lv, W., Xie, R., Huang, M., & Liu, S. (2023). An efficient change detection method for disaster-affected buildings based on a lightweight residual block in high-resolution remote sensing images. *International Journal of Remote Sensing*, *44*(9), 2959–2981. https://doi.org/10.1080/01431161.2023.2214274

# Masters Program in **Geospatial Technologies**

Supported by:
Education and Culture
ERASMUS MUNDUS