Building Damage Assessment Using Remote Sensing Data and Deep Learning Algorithms

A dissertation submitted by

Olalekan R. Sodeinde

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Civil and Environmental Engineering

Tufts University

February 2024

Adviser: Laurie Baise

Abstract

Climate change and wars bring catastrophic disaster events, which raise the need for quicker turnaround in building damage assessments. Disaster response personnels and non-governmental and governmental agencies require information on building damage to make timely, life-saving decisions such as resource allocations. To help automate this process, the xBD dataset was released for training deep learning algorithms. The dataset includes high-resolution satellite imagery annotated with building locations and damage classes ("destroyed", "major damage", "minor damage", and "no damage") before and after natural disasters. This thesis analyzed the quality of the xBD dataset for identifying building damage across a variety of natural hazards using deep learning algorithms.

In Chapter 1, models were created to evaluate the pros and cons of combining the training datasets across multiple natural hazards. Because the model overfits to the "no damage" class, merging classes was evaluated. The "no damage" class was retained and a second class ("damaged") was created by merging "minor damage," "major damage," and "destroyed" classes. Recommendations were then made on using the provided training dataset for optimizing classification accuracy for building damage assessment across hazards including: volcanoes, hurricanes, wildfires, floods, tsunamis, earthquakes, tornadoes, and fires.

Chapter 2 operationalizes the models from Chapter 1 for a new, unseen natural disaster event. It followed recommendations from Chapter 1 to select binary and multiclass models to perform damage assessments of the town of 'Eua, Tonga. The buildings in 'Eua, Tonga were affected by the January 2022 volcanic eruption and its resulting tsunami. However, the damages to the buildings were primarily caused by the tsunami. Post-disaster high-resolution (resolution of 0.8 m) satellite images from the Worldview-3 satellite (provided by Maxar Technologies through their the National Geospatial-Intelligence Agency (NGIA) access) were used with a pre-trained deep learning classification algorithm to classify the damages in 'Eua. Building footprint images are required for the damage classification models. The building footprints were downloaded from Open Street Maps (OSM). Validation of the results was based on United Nations Satellite Center (UNOSAT) visual interpretation of Pleiades images. We tested both a binary class and multiclass model trained using the multi-hazard xbd dataset. The result showed that the binary model trained on a multi-hazard xBD overfitted to the no-damage class. However, the multiclass model that was trained on a multi-hazard xBD dataset was transferable. It was able to detect 96% of no-damage buildings and 84% of damaged buildings.

Chapter 3 operationalizes the models from Chapter 1 for a non-natural disaster event. It examined using pre-trained classification models developed in Chapter 1 for classifying post-war images of buildings damaged by Russian invasion of Bucha, Ukraine in March 2022. We evaluated models trained using wind hazards and multi-hazards on a multi-class damage data. Building footprint images were also required for the damage classification models. For completeness, missing building footprints from OSM were added. The results from the models were validated against the UNOSAT post-disaster dataset, which only contained information about damaged buildings. The results indicate that the classification model trained with wind data detected minor damage (54%), but was unable to detect the major and destroyed classes. The multi-hazard model was able to detect some of the major (18%) and minor (27%) damages, but was unable to detect any of the destroyed buildings. However, both models overfit to the "no damage" class.

Introduction

One of the most important tasks post-disaster is the building damage assessment. For example, in the United States, the outcomes of the post-disaster building damage assessment or Preliminary Damage Assessment (PDA) help to determine the magnitude of damage and impact of a disaster. The Federal Emergency Management Agency (FEMA) uses the information from PDA to determine resource allocation to affected areas. These resources are vital for lifesaving and disaster recovery. The information from a PDA forms the basis of the advice FEMA gives to the President in response to a state governor's request for aid. This advice helps the President to decide on whether to declare a state of emergency, which allows FEMA to deploy public resources to the affected areas.

PDA information is usually collected by the state, tribe, or local government and they are categorized into the following degrees of damage: affected (if the damage to the home is mostly cosmetic); minor (a home with repairable non-structural damage); major (a home with structural damage or other significant damage that requires extensive repairs.); and destroyed (if a home is a total loss). Because PDAs are time-consuming, they create a key analytical bottleneck for post-disaster response and recovery efforts. For this reason, the Defense Innovation Unit, together with other Humanitarian Assistance and Disaster Recovery (HADR) organizations, released a high-resolution satellite dataset and a base algorithm (xBD) to train models to automate damage assessment and accelerate recovery from natural disasters. However, the xBD dataset is imbalanced; so, it overfits to its "no damage" class. Training a disaster model is time-consuming and costly. Therefore, it is important to know ahead of time if an existing training dataset is usable; some ideas to improve accuracy of the model; the type of data to collect for a future disaster event; if an existing model is operationalizable for a new unseen disaster event; and if a natural disaster model is transferable for use in a non-natural disaster event.

In Chapter 1, recommendations were made on class balancing and data augmentation, and base model modifications. Models were also created to analyze the quality of xBD dataset for identifying building damage across a variety of natural hazards using deep learning convolutional neural networks. Chapter 2 follows the recommendations in Chapter 1 to select and then analyze the transferability of some of these models for damage assessments of the January, 2022 volcanic eruption in the town of 'Eua, Tonga. Similarly, Chapter 3 follows the recommendations from Chapter 1 to analyze the transferability of some of the models in Chapter 1 for classifying post-war images of Bucha, Ukraine after its March, 2022 Russian invasion.

Table of Content

List of Tables

Chapter 1

Chapter 2

Chapter 3

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 1

One versus All: Identifiability with a multi-hazard and multiclass building damage imagery dataset and a deep learning neural network

Abstract

This paper analyzed the quality of the xBD image-training dataset for identifying building damage across a variety of natural hazards using deep learning convolutional neural networks. Specifically, we evaluated the pros and cons of combining training datasets across multiple natural hazards and provided recommendations on using the provided training dataset to optimize classification accuracy for building damage detection. The xBD dataset was rebalanced, using random oversampling and under-sampling methods. Random over-sampling randomly duplicates the minority class, while random under-sampling randomly cuts-off the majority class. With the balanced dataset, we used the xBD baseline architecture as a starting point in the classification and found that it overfit to the no damage class; therefore, we improved the base classification algorithm by modifying the top layers of ResNet50. We found that not all classes (destroyed, major damage, minor damage, and no damage) were uniformly identifiable across natural hazards; therefore, we retrained the weights from ImageNet, adding five new convolution, batch normalization, and max pooling layers on top of ResNet50. One dropout layer, with a rate of 0.5 was also added in-between the fully connected layers to reduce overfitting and improve performance. We also evaluate the identifiability of the four damage classes in the xBD dataset. Because classification performance was significantly higher for the "no damage" class as compared to "minor", "major", and "destroyed" classes, we evaluated merging classes. We kept the "no damage" class and created a second merged class ("damaged") representing "minor damage," "major damage," and "destroyed." We used the same architecture for the multiclass classification and the binary classification but without the ImageNet weights. Based on this work, we recommend that users be aware of performance differences across natural hazards and across damage classes. Earthquake building damage is extremely limited in the training data and, as a result, application of the trained algorithm on earthquake data cannot be evaluated given the xBD dataset. Building damage due to volcanoes and tsunami are also poorly represented in the training data, and do not have sufficient data for model validation (especially within all damage classes). Wind hazards are well-represented and therefore application of the algorithm trained using either the wind-only data or the multi-hazard dataset is reliable. The multi-class algorithm trained with wind hazard specific data slightly outperforms a multihazard trained multiclass model (F1 score 0.70 vs. 0.67). Both models have similar performance across all four classes (F1>0.5). For flood, fire, and tsunami hazards, we recommend using the binary damage classes as identifiability is low for at least two of the classes in each hazard. For flood building damage, binary classification performance resulted in a significantly higher F1 score when trained with the flood specific dataset versus the multihazard data (0.72 vs. 0.54). On the other hand, for fire building damage, classification performance is slightly higher when the model is trained on multi-hazard data, rather than trained using a fire specific dataset (F1 score 0.46 vs. 0.42).

Keywords: Damage Assessment, Disaster Response, Remote Sensing, Machine Learning, Deep Learning, Neural Networks, Convolution Neural Network, Imbalanced Dataset

Introduction

In the aftermath of a disaster, damage assessment maps containing information on the type and extent of damage are valuable for disaster response and recovery if they are available quickly after the event. Overhead images of the impacted region can be rapidly acquired and processed to create these maps. Current approaches to create these maps include in-person damage assessments, visual interpretations of overhead imagery, or automated classification of overhead imagery. In-person damage assessments are the most accurate; however, they are costly and time

consuming and are therefore usually limited in spatial coverage. Visual interpretation of overhead imagery can reduce the need to be on-site and can capture the image quickly after the event; however, interpretation and labeling still requires significant trained labor and is subject to human error and inconsistent labeling. Automated change detection, using optical and/or Synthetic Aperture Radar (SAR) images with traditional change detection analysis or machine learning, can detect damaged buildings and infrastructure and potentially, quantify the degree of damage. As an alternative to change detection (which requires pre- and post- imagery), an automated algorithm can be developed using a library of labeled imagery such as the xBD dataset evaluated in this study. When such an algorithm is pre-trained, it can be applied to new events with significantly less labor, thereby reducing response time and money, and resulting in consistent labeling. This study evaluates using the xBD training dataset and deep convolutional networks to produce labels for damaged buildings across different natural hazards.

Currently after natural hazards, there are a variety of damage maps produced by agencies such as NASA, UNOSAT, Copernicus or by individual research groups, each using different imagery and processing/classification methods. NASA's Damage Proxy Maps (DPMs), created by the Advanced Rapid Imaging and Analysis (ARIA) team at NASA's Jet Propulsion Laboratory and Caltech in Pasadena, CA, are created using a change detection technique and SAR images with the interferometry coherence method (Tay, et al., 2020). SAR is helpful for identifying damages, as its radar transmission in the microwave spectrum is independent of day-and-night visibility and weather conditions, such as persistent rain and cloud cover (Zhang, et al., 2012). At NASA, the post-imagery is registered, and histogram-matched to the pre-imagery. The difference between the pre- and post- coherence images is then used to create the damage proxy map. NASA's DPMs are used as guides for identifying damaged areas after a natural disaster but do not label damage (Tay, et al., 2020). NASA DPMs have a resolution of 30 m and do not differentiate between different types of damage (e.g., buildings, ground failure, flood). Some more recent SAR satellites, including ALOS-2, COSMO-SkyMed, TerraSAR-X, and TanDEM-X offer one-meter level spatial resolution SAR; however, the image availability is still limited in terms of spatial and temporal coverage (Anirudh, et al., 2022).

Other researchers have explored SAR as well as other data sources and advanced algorithms to develop damage maps. Sghaier et al. (2017) used texture descriptor, morphological operators, and Fuzzy logics to automate the extraction of flooded areas from SAR. Mason et al. (2021) combined SAR and Digital Elevation Models to identify flooded areas. Stephenson et al. (2021) used deep learning and timeseries of SAR images to delineate flooded areas. Xu et al. (2022) has used NASA DPMs with the USGS ground failure products and building footprints to create labeled damage maps using causal inference from satellite imagery.

Damage maps derived from optical imagery are often available at a higher resolution than those derived from SAR for the same event. For example, the United Nations Satellite Centre (UNOSAT) produces damage maps from high resolution optical images of up to 0.5 meters spatial resolution (Dell'Oro, 2017). UNOSAT produces damage maps using visual interpretation. Copernicus rapid damage mapping service, an optical method supported by the European Commission, can also produce damage maps from optical images of 10 m spatial resolution ("Copernicus: Sentinel-2 - Satellite Missions - eoPortal Directory", 2022). Copernicus produces damage maps, using satellite imagery and other geospatial data.

Other researchers have further explored advanced classification methods applied to optical imagery such as Yaun et al. (2021), Laigen et al. (2013) and Chen et al. (2016). Yuan et al. (2021), used two-step ensemble models to perform building segmentation and damage classification on a publicly available optical satellite imagery. Laigen et al. (2013), provided a comprehensive review of multi-temporal techniques that evaluate the difference between the pre- and post-disaster images to identify a building damage and mono-temporal techniques that use only the post-event images for damage detection. The weakness in a multi-temporal approach is that pre-disaster images are not always available immediately after a disaster event. Many cities, especially in developing countries, do not have pre-disaster images, but when they do, the spectral or the spatial resolution may be different from those of the post disaster images. High resolution aerial images have also been used to identify damaged buildings, and are generally only available as post-event images. Chen et al. (2016), proposed using high resolution aerial images, collected with Unmanned Aerial Vehicles (UAV), to create a depth difference map and a grayscale difference map to segment changed building areas. Classification with mono-temporal images requires extensive labeling for training and has not been developed for rapid classification.

Despite its higher resolution, supervised classification of optical images is slow and labor-intensive for extracting operationalizable information due to the need for labeling of training pixels (Xu, et al., 2019). It is for this reason that supervised machine learning algorithms, specifically deep-learning approaches, are now being explored and used to automate and speed up disaster damage detection (Rashidian, et al., 2021). Machine learning methods are also advantageous because they can require only post-disaster images for training datasets. Cooner et al. (2016) studied the effectiveness of several machine learning methods, including multilayer neural network, radial basis neural network and random forest in detecting earthquake damage caused by the 2010 Port-au-Prince, Haiti 7.0 moment

magnitude (Mw) event. Naito et al. (2018) also used Support Vector Machine (SVM), another machine learning algorithm, to detect damage to buildings, utilizing images of automobile running surveys after the 2016 Kumamoto Earthquake. In addition, Koshimura et al. (2020) discuss an application of machine learning to tsunami damage detection, using algorithms such as decision trees, random forest and K-Means clustering.

With unprecedented increase in the use of high-performance computers and high-resolution training dataset, neural networks are increasingly being leveraged to gain new insights from optical images. A regular neural network could be problematic for image-processing tasks. This is partially due to the large number of the weights needed. However, unlike the 1-D vector in a regular neural network, a Convolution Neural Network (CNN) carries information about the spatial relationships between image features and is therefore ideal for automated classification of images. To capture the complexity of an image feature, the neurons are organized into stacks of 3-D volumes, each connected to a later volume. In a classification problem, the later volumes are filtered and then flattened at the last layer to the number of classes in the problem. Like other deep-learning algorithms, CNN are computational models with multiple processing layers that learn representations of data, using multiple levels of abstraction (LeCun et. al., 2015). Li et al. (2019) described an approach that uses a convolutional auto-encoder that consists of VGG16 for post disaster damage detection. The exercise focused on detecting damage on images from Hurricane Sandy, using images from Hurricane Sandy as the training dataset. Siti et al. (2017) also used CNN with aerial imagery to detect landslides and floods. They created training data patches of pre-disaster and post-disaster by clipping and resizing aerial imagery obtained from Google Earth Aerial Imagery. Their paper focused on the application of CNN to a single disaster type and a single event.

To model the complex nature of disaster damage, deep learning requires a large quantity of labeled imagery for training (Gupta, 2019). The xBD dataset is a freely available multi-hazard dataset for building detection and damage classification. It combines data across 7 different disaster types and 19 different disaster events (Gupta, 2019). As a global dataset of over 40,000 km^2 of images, it contains over 850,000 building annotations, across 19 different disaster events. It was developed by DIU and HADR in 2019 for the xView2 challenge, a challenge to automate building damage assessment after a disaster. Its images were provided in two tiers. The first tier (Tier1) was provided by Maxar Technologies/DigitalGlobe as part of its open-data program, which provides free high spatial resolution of pre- and post-disaster images for humanitarian responses. The second tier (Tier3) was not part of the open-data program. It was released as part of a collaboration between Maxar Technologies and the National Geospatial-Intelligence Agency to motivate research with the xBD dataset.

The xBD dataset pools training data across multiple hazards and provides four class labels; however, the resulting dataset is unbalanced across natural hazard types and damage classes. The xBD dataset provides the opportunity to evaluate pooling as well as to explore the best practices for processing training datasets. In this paper, we analyzed the quality of a publicly available training dataset (xBD) for multi-hazard multi-class damage detection of buildings using deep-learning methods, discuss the pros and cons of combining training datasets across multiple natural disaster events, and provide recommendations on using the provided training datasets to optimize classification accuracy. We use class balancing techniques and data augmentation to improve model performance. To train the convolutional neural network for multiclass classification, we retained the weights from ImageNet, and added five new blocks of convolution, batch normalization, and max pooling layers on top of the ResNet50. One dropout layer, with a rate of 0.5 was also added in-between the fully connected layers to reduce overfitting and improve performance. For binary classification, we ignored the existing ImageNet weights and trained the whole network with xBD images by setting the trainable property of the ResNet50 architecture to "True". Model performance is assessed using the average F1 scores of the classes in the test dataset. We make recommendations on data training, data preparation and usage, including class balancing and data augmentation, and model modifications and model usage as related to the xBD training dataset and the identification of damaged buildings after natural hazard events.

With the continual frequency of natural disasters due to climate change, it is more important than ever to optimize classification accuracy of deep learning algorithms for building damage assessment. This work explores an underexamined aspect of the pros and cons of combining training datasets across multiple natural hazards and provided recommendations on using the provided training dataset to optimize classification accuracy for building damage assessment.

Data

xBD is a large-scale dataset of satellite imagery, which covers a diverse set of disasters. It also contains a base model for the xView2 challenge. The dataset contains bounding boxes for environmental factors such as fire, water, and smoke. Unlike other existing datasets, whose damage scales were limited to a single disaster type, xBD disaster scales cover multiple disaster types, allowing for easy comparison across multiple disaster types. Also, unlike existing datasets, xBD contains a damage scale that is more developed than a simple binary damaged/undamaged option. The non-binary nature of xBD damage scale was a requirement of disaster response experts. Since the difference between the damage scales are visually minute, to facilitate labeling these types of damage levels, high resolution images were lower than 0.8-meter spatial resolution. xBD was created to include damage caused by varying types of disasters, as well as a diverse set of buildings of different shapes and geographies. The "no damage" samples represent negative images, which are useful for learning the difference between the different levels of damage (Gupta, 2019). The dataset consists of three-band RGB formats imagery, annotations, and a base model.

The annotations were created using a web-based application developed by CrowdAI. The pre- and post-disaster image sets were first identified for each event before they were sent to the annotators, who drew polygons around visible building footprints on the pre-disaster images. Once the polygons were drawn, they were then labeled, using the post-disaster images. As part of preprocessing, to account for projection issues when overlaying the building polygons from the pre-imagery onto the post-imagery, the image pixels were shifted slightly. Some buildings were missing when overlaid on post-disaster images either because the building did not exist in the pre-disaster image, or the building did not meet the definition of a building in the pre-disaster image, or the building was occluded by cloud, haze, or vegetation.

The complete xBD dataset contains 22,068 and 850,736 polygons drawn around buildings across 45,361.79 km2 area. The data was split into train, test, and holdout datasets in the ratio of 80/10/10% respectively. The density of polygons is not evenly distributed across each occurrence of a natural disaster. For example, Mexico City earthquake and Palu tsunami have many polygons in comparison to their relatively low image area.

xBD Images

Each xBD image tile is a 1024×1024, three band RGB, with at least 0.8-meter spatial resolution. The radiometric resolution is 8 bits (0-255). The images were provided in a pixel coordinate system, but they can be georeferenced by using the information in the buildings' annotations or the labels' files (discussed later). There were a total of 22,068 images: 18,336 train images; 1,866 test images, and 1,866 holdout images. The dataset includes seven types of disasters labeled as: earthquake, fire, flood, hurricane, tornado, tsunami, and volcano (Figure 1). All the disasters occurred between 2011 and 2018.

Figure 1. Natural disasters represented in xBD dataset.

Filename Convention

Each image's filename includes its event name, a unique file number and a Post/Pre identifier, used as a suffix. For example, the first post-disaster image in the Midwest-Flooding event was named Midwest-Flooding _00000000_Post_Disaster.Png, and the second one was Midwest-Flooding _00000001_Post_Disaster.Png. The event name is 'Midwest-Flooding'; the file number is '00000000', and the suffix is 'Post'. This naming system allowed for the pairing of post-/pre-disaster images across all disaster events. It also allowed for the matching of images to their buildings' annotations in the label files.

Disaster Types

Nineteen events are included in the dataset across seven disaster types. The number of images in the training dataset ranged between 36 images for the Guatemala Volcano, and 3700 images for the Portugal Fire. The dataset included seven different disaster types: earthquake, fire, flood, hurricane, tornado, tsunami, volcano, and wind. We combined tornado and hurricane into a single category (wind) as the building damage for both types of natural disaster are controlled by wind and all the disasters related to fire were grouped into a single category (fire).

The Labels

Buildings' annotations are stored in GeoJSON files. These files are called labels, and they were named in a similar fashion as the image files, which makes it easier to associate them with their pre and post disaster images. It is possible to have multiple labels associated with an image. There are two objects at the root of each of the GeoJSON files: features, and metadata. A features' object contains coordinates of the vertices of a building. These coordinates were stored in geographic coordinates (longitude, latitude) as well as pixel coordinates (x, y). In addition to a buildings' coordinates, a features' object also contains a building's properties. These properties include feature types, subtypes, and unique identifiers (uids). The only feature type in an xBD label file is building. A subtype property contains information about a building's damage level and its uids. Unlike the feature object, a metadata object contains information about an image. Some of this information includes the name of the satellite sensor, the spatial resolution of the images, and the image's dimensions. Figure 2 describes the structure of the label file associated with each building.

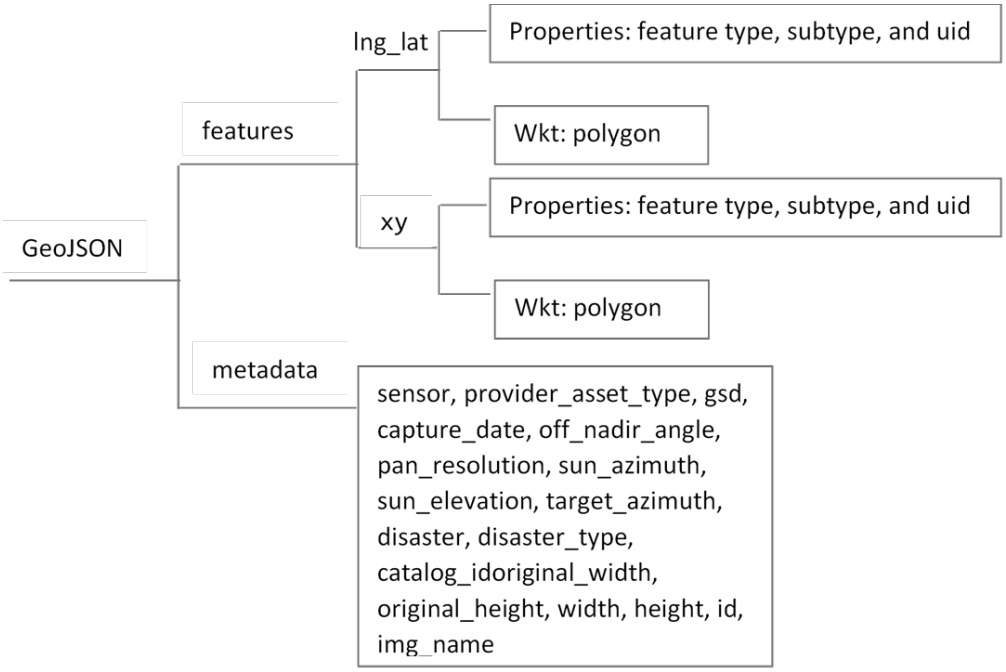


Figure 2. Structure of a label file

Damage Levels

The damage level is stored as a subtype in the label file with integer values ranging from 0 to 3: 0 is no-damage, 1 is minor damage, 2 is major damage, and 3 is destroyed. As part of the xView2 challenge, which was the source of the data, a uniform damage description was developed across disaster types as shown in Table 1. The uniform damage level enables a common understanding across disaster types and events. For example, destroyed means scorched in a fire, completely collapsed in an earthquake, partially or completely covered with water or mud in a flood. The image interpretation and damage labels were provided by different agencies responsible for each of the included events; therefore, the joint damage level was also created with insights from NASA, CAL FIRE, FEMA, and the California Air National Guard (Gupta, 2019).

Table 1. Uniform Damage Levels

Disaster Level	Structure Description
0 – No-Damage	Undisturbed, no signs of water, structural or shingles damage, or burn masks.
1 – Minor Damage	Building partially burnt, water surrounding structure, volcanic flow nearby, roof element missing, or visible cracks.
2 – Major Damage	Partial wall or roof collapse, encroaching volcanic flow, or surrounded by water/mud.
3 - Destroyed	Scorched, completely collapsed, partially/completely covered with water/mud, or otherwise no longer present.

According to FEMA, damage levels are useful for estimating life-safety consequences of building damage, monetary losses due to building damage, monetary losses that may result from business interruption, expected social impacts, and other economic impacts. FEMA uses HAZUS, a geographic information system-based tool, for analyzing

natural hazards. Overall, HAZUS uses 5 damage levels: “none”, “slight”, “moderate”, “extensive”, and “complete.” These levels denote the amount of damage for structural and nonstructural building features separately. For example, when there is structural damage to a ceiling due to an earthquake, HAZUS describes “slight” damage as when a few ceiling tiles have moved or fallen. “Moderate” non-structural damage includes more extensive tiles falling, the ceiling support framing (T-bars) has disconnected and/or buckled at a few locations, or lenses have fallen off some light fixtures and a few fixtures have fallen, and localized repairs are necessary. “Extensive” damage is described as when the ceiling system exhibits extensive buckling, disconnected T-bars, and falling ceiling tiles; ceiling partially collapses at a few locations and some light fixtures fall; repair typically involves removal of most or all ceiling tiles. “Complete” damage is described as when the ceiling system is buckled throughout and/or fallen and requires complete replacement; and many fallen light fixtures.

HAZUS describes structural damages caused by earthquakes for 16 basic specific building types. For example, damage levels for wooden and light frames were described separately. HAZUS indicated that “slight” damages to wooden and light frames includes small plaster or gypsum-board cracks at corners of door and window openings and wall-ceiling intersections, or small cracks in masonry chimneys and masonry veneer. “Moderate” damage occurs when large plaster or gypsum-board cracks at corners of door and window openings; small diagonal cracks across shear wall panels exhibited by small cracks in stucco and gypsum wall panels; large cracks in brick chimneys; toppling of tall masonry chimneys. “Extensive” damage occurs when there are large diagonal cracks across shear wall panels or large cracks at plywood joints; permanent lateral movement of floors and roof; toppling of most brick chimneys; cracks in foundations; splitting of wood sill plates and/or slippage of structure over foundations; partial collapse of “room-over-garage” or other “soft-story” configurations; small foundation cracks. “Complete” damage is described as when a structure may have large permanent lateral displacement, may collapse, or be in imminent danger of collapse due to cripple wall failure or the failure of the lateral load resisting system; some structures may slip and fall off the foundations; large foundation cracks.

HAZUS describes the damage levels for Hurricane damages using the U.S. Department of Housing and Urban Development (HUD) categorization. HAZUS based the damage levels description on the percentage of the roof that is damaged. For example, “no damage” level is when there is 0% damage to the roof; “slight” damage is when there is between 0% and 33% roof damage; “moderate damage” is when there is between 33% and 66% roof damage; “extensive damage” occurs when the damage to the rooftop is between 66% and 99%, and “Complete damage” occurs when there is 100% damage to the rooftop. These descriptions are comparable to other agencies' categorizations. For example, FEMA and Red Cross categorize “slight” damages as “affected”, “moderate” damages as “minor”, “extensive” damages as “major”, and “complete” damages as “destroyed.” xBD joint damage levels were created from insights from these categorizations.

Of all the xBD damage levels, a deep learning classification algorithm has the most difficulty capturing the nuance in minor and major damages. xBD describes minor and major damage as partial damage to a building or the presence of some environmental factors such as lava flow around a building. Although “major damage” is expected to describe a higher degree of damage or a larger number of these environmental factors, it is challenging for analysts to correctly determine the degree of partial damage, or the number of environmental factors needed to classify a building as a minor or a major damage. In addition to the challenges presented by the xBD damage level descriptions for differentiating minor and major damages, the modality of satellite imagery, such as resolution, azimuth, and smear also presented some challenges (Gupta, 2019).

The ‘no-damage’ label is consistent across natural hazards and events. Appendix 3 provides examples of ‘no-damage’ buildings across all six disaster types. They are usually undisturbed, with no signs of water, structural or shingles damage, or burn masks. The ‘minor damage’ is illustrated in Appendix 4 and can be described as buildings that are partially burnt, water surrounding structure, volcanic flow nearby, roof element missing, or visible cracks. As shown in the figure, different hazard types have different visual cues for the ‘minor damage’ label. The ‘major damage’ is illustrated in Appendix 5 and can be described as having partial wall or roof collapse, encroaching volcanic flow, or surrounded by water/mud. Similarly, the ‘major damage’ label has different visual cues across hazard types. Destroyed buildings are illustrated in Appendix 6 and can be described as buildings that are scorched, completely collapsed, partially/completely covered with water/mud, or otherwise no longer present. The minor and the major labels are the most difficult to identify visually (Table 2).

Table 2. Minor and Major damages are the most difficult to identify.

	Flood	Fire	Tsunami
Minor Damage			
Major Damage			

Methods

Data Processing

The xBD base classification algorithm requires training images of buildings and their labels. To extract the buildings from the images, the coordinates of buildings were extracted from the label files and were then converted into rectangular bounding boxes. These boxes were used to clip the training post-disaster images containing the buildings. 80% offsets (in relation to the building footprint) were added to the bounding boxes to capture the area surrounding the buildings. These building clips were then resized to 128 x 128 pixels so that all buildings are represented on a tile of the same size (but different resolution). Along with files of the buildings' images, a CSV file was also created to store buildings' uids and the damage level (Appendix 7). The uids are unique identifiers for each building.

Improving xBD Dataset Through Class Balancing and Data Augmentation

Class Balancing

Figure 3 shows the damage labels by natural hazard event and Figure 4 shows the damage labels by natural hazard type. The distribution of the original xBD training dataset is highly imbalanced across events, natural hazard type, and damage classes. The "no-damage" label (0) is significantly overrepresented in the dataset across all events and natural hazard types. Machine learning algorithms are sensitive to class imbalance and will overfit to the majority class if class balancing is not taken into account. In terms of natural hazard type, wind-related events have the highest number of labels and the best distribution across classes. Volcano has the lowest number of total labels (3,644 "no damage," 58 "minor damage," 35 "major damage," and 529 damage buildings), and earthquake has the worst distribution of classes with only 2 "destroyed," 110 "minor," and 18 "major," damage buildings out of a total of 32,271 buildings. Tsunami has 1 "minor," 671 "major," 5145 "destroyed," and 42,524 "no damage" buildings. Since the distributions of volcano and earthquake datasets are poor, they will not be considered in the analysis of multiclass and binary classification results. We do not recommend using the xBD dataset for classifying damaged buildings for volcanoes or earthquakes. Also, since Tsunami only has 1 minor damage label, it will only be considered in the analysis of results of binary classification.

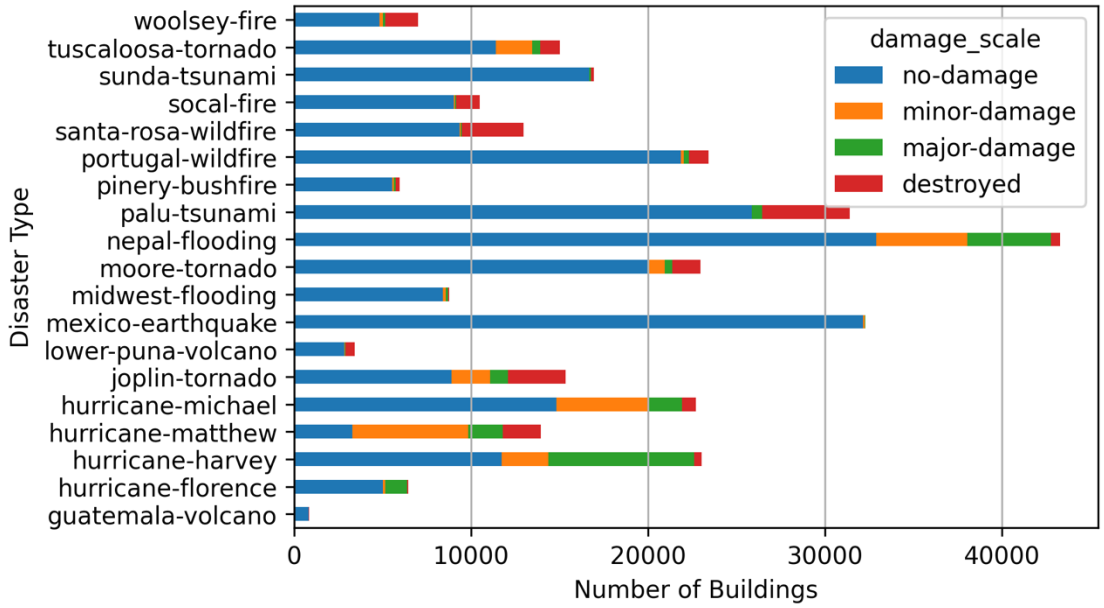


Figure 3. Damage level per disaster events

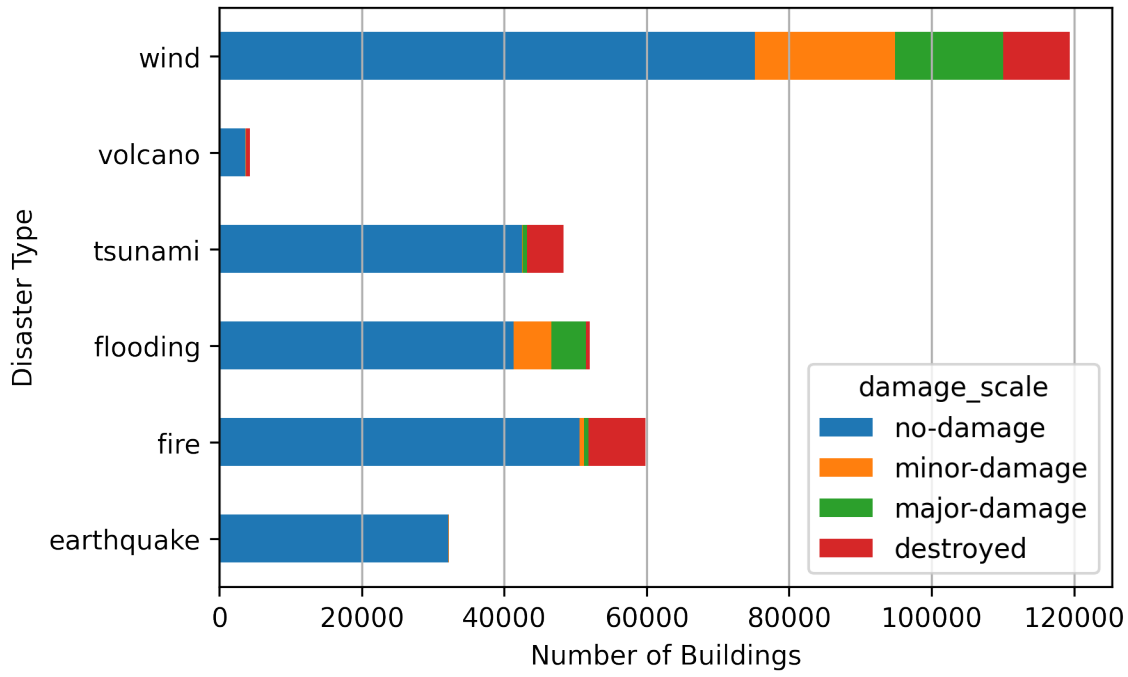
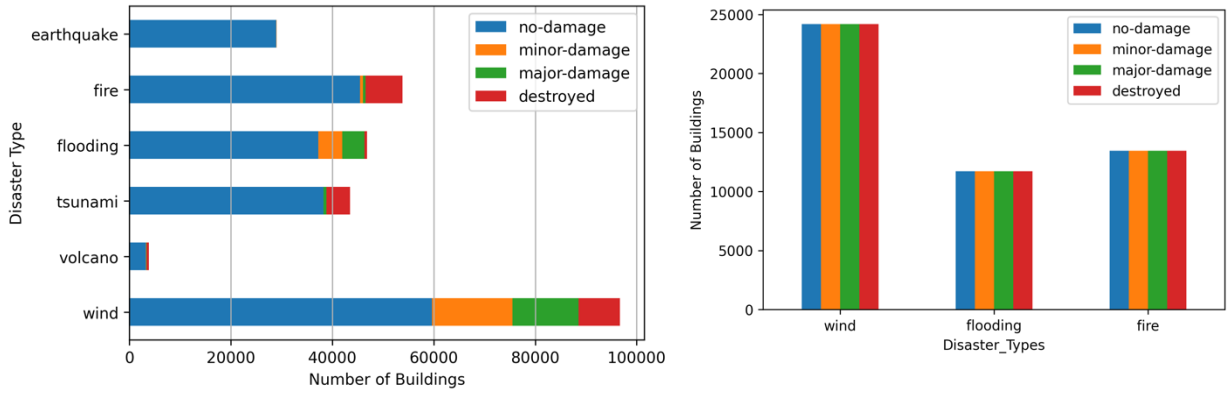


Figure 4. Damage level per disaster types

In an imbalanced dataset, the distribution of classes is uneven. Johnson et al. (2019) indicates that there are three groups of methods for handling class imbalance in machine learning: data-level techniques, algorithm-level methods, and hybrid approaches. The data-level techniques use different sampling approaches to balance input data. The algorithm approaches work on the algorithms themselves by adjusting weights, costs and/or the algorithm itself. The hybrid approaches combine both data and algorithm level approaches. In addition to sampling approaches, Mateusz et al. (2018) studied imbalanced problems in Convolution Neural Network by using two-phase training, and thresholding that compensates for prior class probabilities.

Data-level approach uses random over-sampling and under-sampling methods to reduce data imbalance. Random oversample randomly duplicates the minority class. While random under sampling randomly cuts-off the majority class. Oversampling will increase the number of training dataset, increasing the training time. Under sampling on the other hand, will remove data and reduce the amount of data available to training the deep neural network algorithm. To combat the impact of under-sampling, Zhang, and Mani (2003) suggested some intelligent under-sampling approaches (Zhang J, 2003). One of these approaches called the Near-Miss algorithm uses the K-nearest neighbor algorithm to select the majority sample that should be removed. In 2000, Kubat and Matwin suggested a one-sided selection method to remove noise in a majority class (Kubat M, 2000). Synthetic Minority Over-sampling Technique (SMOTE) is another over-sampling technique that strengthens class boundaries, reduces overfitting and increases class discrimination. Transfer learning has been used to deal with class imbalance issues. It uses information gained from training a deep learning algorithm to solve similar problems. In 2016, Al-Stouhi S et al. discussed how to use transfer learning to deal with class imbalance in rare datasets.

Of all the methods, data-level methods are most popular in addressing class imbalance issues, with Random Oversampling and Under-Sampling methods showing better overall results. The algorithm-level approaches show inconsistent and conflicting results. They are limited in evaluated techniques and lacking comprehensive, comparative studies (Leevy, Khoshgoftaar, & Bauder, 2018). For these reasons, the Random Oversampling and Under-Sampling methods approach was chosen for this research. To balance the xBD training dataset, the weights of the inputs into the model were adjusted to be inversely proportional to the class frequencies in the input data therefore increasing the importance of less prevalent classes to the training process. The class weights define the relative importance of each class to the training process. For example, in the flood dataset, class 0, the 'no-damage' class was under-sampled from about 37,240 to about 12,000. Meanwhile, classes 1 and 2 were oversampled by about three folds (Figures 5). The result is that for each hazard type, we used an equal number of labels for each class. Due to the poor distribution of the tsunami dataset, its performance was only analyzed when the algorithm was trained and classified into two classes, using the multi-hazard dataset. Earthquake and volcano datasets were not analyzed.



Number of labels before balancing the dataset

Number of labels after balancing the dataset

Figure 5. Number of buildings in each of the classes after weights were applied.

Data Augmentation

Augmentation is a technique to increase the diversity of a training set by randomly applying realistic transformation to images in a training set (TensorFlow, 2021). This operation does not result in additional images for the training data. However, it helps to improve the performance of the algorithm. Data augmentation can also help to reduce overfitting. In this study, augmentation was performed on batches of 32-image samples at a time. Because running optimization algorithms on a full dataset of large data could be expensive, batch training is used to train the model, using only a subsample of the data at a time. For each batch, predictions were made, and errors were computed to optimize the algorithm. The predictions were made, using an optimization algorithm (Adam), which looked for the internal model's parameters and then calculated their performance measure or errors (e.g., mean square error) by comparing them to their real expected outcomes, and using the error to update the internal model parameters.

The images in a batch were randomly flipped horizontally; shifted to 0.05 fraction of the height and of the width; zoomed into the 0.01 range; and rotated by 2 degrees. Finally, all the images in the batch were rescaled by 1/255 and then resized to 128 by 128.

Measuring the Performance of a Model

An overall accuracy of a model is usually not the best measure of its performance on an imbalanced dataset. F1 score gives a better view of the performance within the classes. In Equation (1), F1 score is defined as the harmonic mean of precision and recall. It reaches its best score at 1 and worst score at 0. Precision is the ratio of true positive to the sum of true positive and false positive. Recall is the ratio of true positive to the sum of true positive and false negative. F1 score is calculated for each of the classes in the classification. These scores can be averaged into a single F1 score. There are four averaging approaches. One is the macro averaging, which calculates F1 score, using unweighted mean. This approach ignores the imbalance in the dataset. A second approach is micro averaging. This is a global approach. It counts all true positives, false negatives and false positives. A third approach is a weighted average approach. It considers the imbalance in the dataset by finding the number of true instances for each label. The last approach is sample averaging, which calculates the F1 score for each instance of the labels and averages them (Scikit-learn, 2020). We use the macro averaging approach shown in Equation (4) because we are using a balanced dataset as discussed later.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$ ---------- Equation (1)

$$recall = \frac{TP}{TP + FN}$$ ---------- Equation (2)

$$precision = \frac{TP}{TP + FP}$$ ---------- Equation (3)

$$Macro\ F1\ = \frac{1}{n}\sum_{i=1}^{n} (F1_i)$$ ---------- Equation (4)

Where:

TP = True Positive
FP = False Positive
FN = False Negative
n = number of classes
$F1_i$ = F1 score of the individual classes

One of the drawbacks of F1 score is that it does not show the counts or distribution of a model's performances within classes. A confusion matrix provides a solution to this drawback by providing the model's performances within classes. It is a grid whose diagonals are the numbers of true positives, and the off-diagonals are the miss-classified classes. The sum of each row represents the total number of true labels in each class while the sum of each column represents the total number of predicted labels per class. We use the confusion matrix to assess model performance across classes.

Improving Performance of xBD Baseline Deep Learning Architecture for Multiclass and Binary Classification

xBD Architecture

xBD provided a baseline architecture, which was based on ResNet50 (Figure 6). This architecture was a starting point for performing the damage level classification. ResNet50 consists of five groups of convolution and identity blocks (Figure 6). Each convolution and identity block consists of three convolution layers. The convolution layers are made-up of a set of filters or kernels whose weights need to be learned. These weights were learned when the filters convolve with input data and then pass the results to an activation function (Qiuhong Ke, 2018). The identity function, on the other hand, allows a layer to be represented as a function of the original input, keeping the input information fresh through the network. The xBD dataset baseline algorithm was built on a transfer learning from ResNet50. The xBD algorithm retained the weights from training ImageNet, an image database used in the ILSVRC competition. For its transfer learning, the ResNet layers were left untrained, but new layers were concatenated on top to make it learn the xBD dataset. The new layers were carefully selected to avoid overfitting and to achieve an acceptable accuracy. This selection process is called fine-tuning and results in the xBD architecture shown in Figure 7.

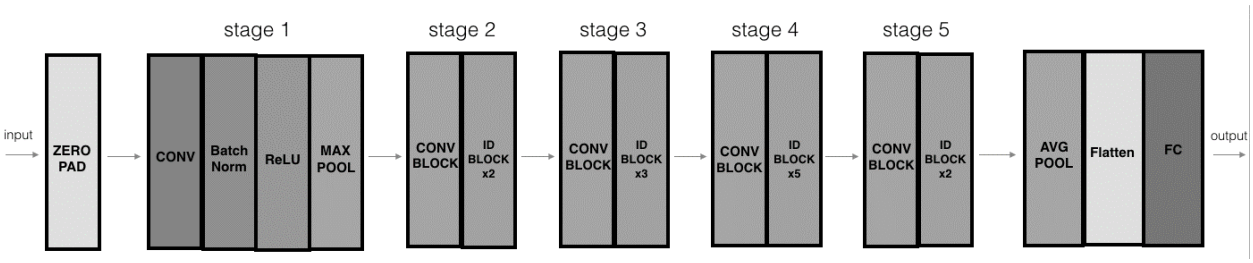


Figure 6. ResNet 50 architecture

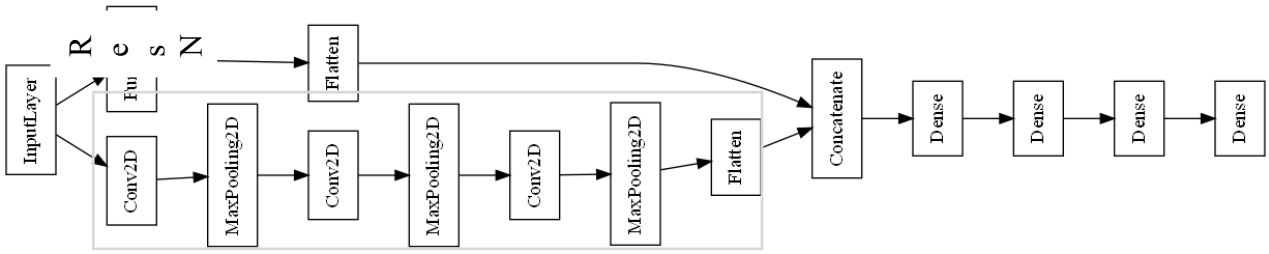


Figure 7. xBD baseline architecture

A model trained by pooling together all the images within the training dataset was used to check the performance of the xBD baseline architecture. The xBD baseline architecture uses an Adam optimizer, a stochastic gradient optimizer that only requires first-order gradients with little memory requirement. Unlike classical stochastic gradient descent optimizers, which use a gradient-based optimization to update and measure new weight for each neuron in a neural network, Adam has the advantage of being simple to implement, computationally efficient, uses little memory, is well suited for problems with large datasets, and its parameters are intuitive to interpret. It computes adaptive learning rates for each of the parameters, using estimates of first and second order moments. The algorithm updates exponential moving averages of the gradient and the squared gradient. Beta1 and Beta2 are the hyper-parameters that control the exponential decay rates of the moving averages. Unlike traditional stochastic gradient descent, which uses a single learning rate for the weight updates, Adam's learning rates are adaptive and are maintained for each parameter. According to the authors of Adam, good default values for Adam's hyperparameters are Alpha = 0.001, Beta1 =0.9, and Beta2 = 0.999. The alpha or learning rate is a tuning parameter that determines the step size at each iteration, while working towards minimum loss in the optimization process. Beta1 and Beta2 are the exponential decay rates for the first and second moment estimates respectively. The model for this work was compiled with categorical cross entropy function, F1 metrics, and the Adam optimizer.

The categorical cross entropy is a function that was used for calculating the differences between predicted values and the expected values for the multiclass classification. For this project, it is expected that as loss is minimized, the F1 score is maximized. The default values of Beta1 and Beta2 were used and held constant while the Alpha value (i.e., learning rate) was tuned. Tuning of the learning rate showed that a minimum for F1 scorers lies between 0.0001 and 0.00001, So, Adam optimizer with piecewise learning rates of 0.0001 and 0.00001, and boundary set to 50% of epochs was selected to optimize the model. A boundary is an integer that defines the value of the epoch at which the learning rate changes. In each iteration, cross-validation was used to randomly hold-back 10% of the multi-hazard training dataset for the purpose of assessing the performance of the model during training. To check the performance of the xBD baseline architecture, the model was trained on the multi-hazard dataset and then assessed on the multi-hazard testing datasets. A confusion matrix was used to explore the performance of the model. Figure 8 shows that even after balancing the dataset, the model still performed poorly and overfits the training datasets since all the buildings were classified as class (0), 'no-damage'.

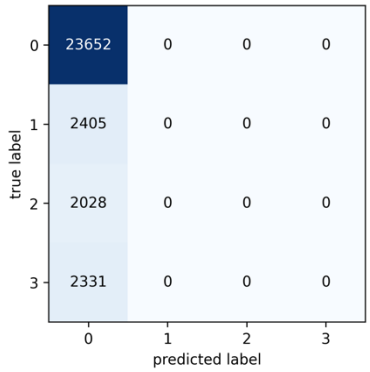


Figure 8. A confusion matrix, created from the entire test dataset; it shows that xBD baseline architecture overfitted on a balanced multi-hazard training dataset.

Modifying xBD Architecture for Multiclass Classification

Because the xBD baseline architecture overfits to the "no-damage" class, a new model was created by modifying only the top layers of a ResNet50. With the weights from ImageNet retained, using fine-tuning, five new blocks of convolution, batch normalization, and max pooling layers were added on top of the ResNet50. One dropout layer, with a rate of 0.5, was also added in-between the fully connected layers to reduce overfitting and improve performance. Dropout was used to randomly change the architecture of the network such that the architecture will not easily overfit to the training dataset. Figure 9 shows the modifications to xBD baseline architecture used in this study.

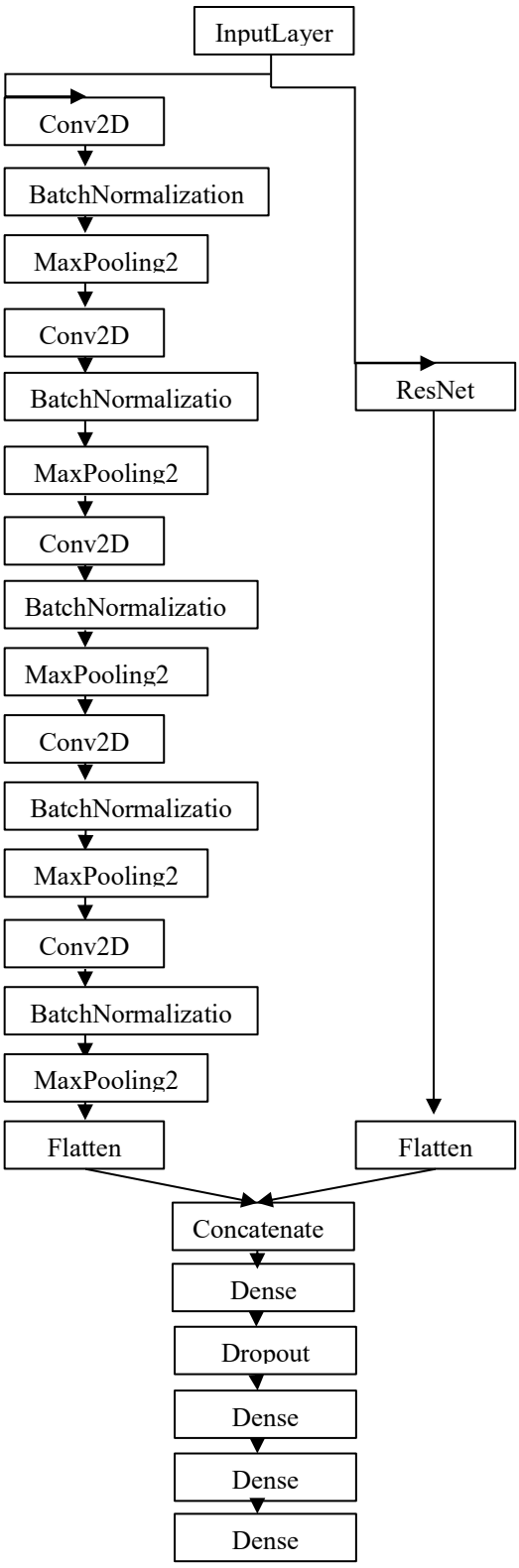


Figure 9. Five blocks of convolution, batch normalization, and max pooling layers were added on top of the ResNet50 in the modified xBD baseline architecture.

Still using the Adam optimizer, the modified model generalized better when trained with the multi-hazard dataset over 10 epochs and then tested on the multi-hazard testing dataset. Unlike the unmodified model, which detected only the no-damage class, the modified model was able to detect 'no-damage' (0), 'minor-damage' (1), and 'major-damage' (3) resulting in average F1 score of 0.40 and F1 scores of 0.75,0.31,0.00,0.55 for "no damage", "minor damage", "major damage", and "destroyed" buildings respectively. This improvement can be seen in the multi-hazard testing dataset's confusion matrix and shown on a map for a region that experienced flooding in Nepal in Figure 10.

(a)

(b)

Figure 10. (a) Confusion matrix for the multi-hazard test data and demonstration of labels for flooding in Nepal. (b) Modified model results with 10 epochs. Average F1=0.40.

Combining Longer Epochs with Early Stopping

With the same optimizer and learning rates, model performance improved when longer epochs were combined with early stopping. Early stopping was needed to save the best models during the training. It was also used to monitor model performance and stop training when model performance failed to improve after 50 epochs. Compared to 10 epochs, the model improved significantly when trained over 80 epochs as shown in Figure 11 (Average F1 score increased to 0.66) and F1 for each class (0.87,0.48,0.58,0.69) for "no damage", "minor damage", "major damage", and "destroyed" buildings respectively. For this reason, subsequent trainings were performed over 100 epochs with early stopping set at 50 epochs if the F1 score did not improve. More could be done to improve the performance of this model; however, we will use this model as a baseline for multiclass classification of all separate and multi-hazard disaster types.

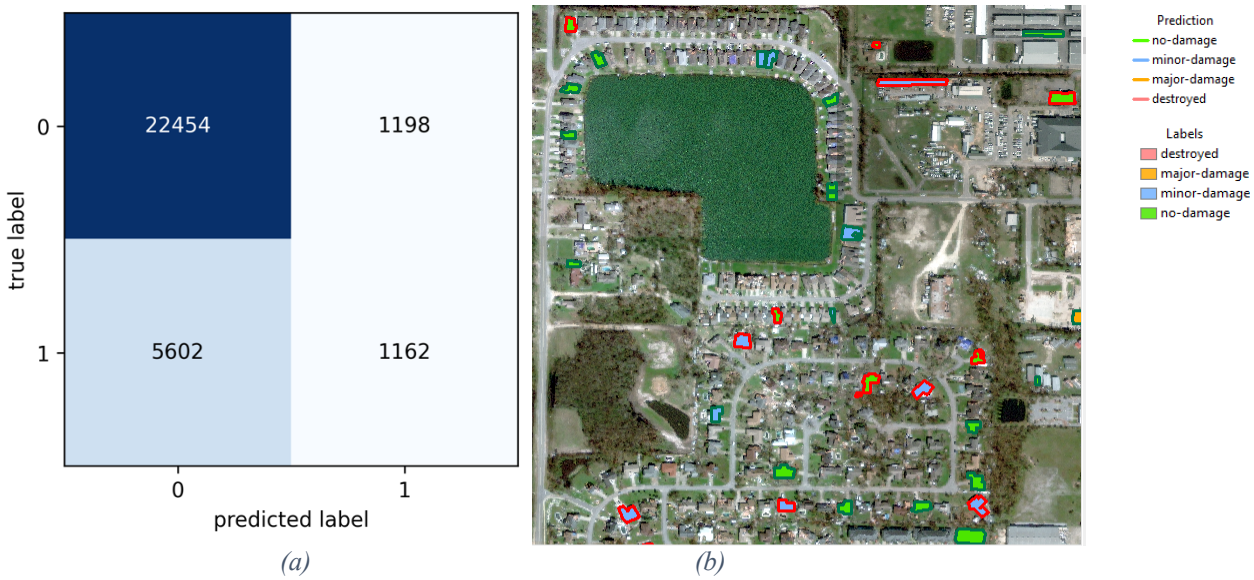
(a) *(b)*

Figure 11. (a) Confusion matrix shows that longer epochs and early stoppings work for (b) Modified model with 80 epochs. Average F1=0.66.

Modifying xBD Architecture for Binary Classification

To modify the xBD deep learning architecture for binary classification, we first rebalanced the training data using the merged class 0 (no damage) and 1 (minor, major and destroyed), and using the same oversampling and undersampling approaches used in multiclass classification. When used for binary classification, the model from modified xBD baseline architecture classified all the buildings in the multi-hazard testing dataset as "no-damage".

To modify the architecture for binary classification, we ignored the existing ImageNet weights and trained the whole network with xBD images by setting the trainable property of the ResNet50 architecture to "True". This approach improves the identifiability of the classes in the binary classification. With this change, both the "no-damage" and "damage" classes in the multi-hazard testing dataset are now identifiable for all the disasters, resulting in average F1=0.57 (Figure 12) ("no damage" = 0.87 and "damage" = 0.26). With this improved result, the modified xBD baseline architecture with trainable set to "True", was used for binary classifications of multi-hazard and separate dataset in this study. Like in multi-class classification, augmentation was performed on batches of 32-image samples at a time. The images in a batch were randomly flipped horizontally; shifted to 0.05 fraction of the height and of the width; zoomed into the 0.01 range; and rotated by 2 degrees. Finally, all the images in the batch were rescaled by 1/255 and then resized to 128 by 128. The training was performed over 100 epochs with an early stopping of 50 epochs.



(a) *(b)*

Figure 12. (a) Confusion matrix of binary classifications of multi-hazard dataset, using modified xBD baseline architecture, with trainable set to "True". F1=0.57. (b) modified binary class model results applied to Hurricane Michael

Results and Analysis

Identifiability Of Classes

Identifiability of each of the disaster's damage classes was evaluated by using the multi-hazard training dataset, training the model based on the multi-hazard dataset, using the model to perform multiclass classifications with the testing data, and then comparing the F1 scores from these classifications. The same thing was then done using the binary classification of the multi-hazard training dataset. Separately, new models were created from each of the hazard type training dataset, using the same algorithm that was used to create models based on the multi-hazard dataset. In the algorithm, RESNET50 weights were retained for multiclass classification, however, they were overwritten for binary classification. These models were also used to perform multiclass as well as binary classifications of their testing datasets. Unlike in the multiclass classification, the modified xBD architecture used for training the binary models ignored the existing ImageNet weights in its ResNet50 by setting the trainable property of the ResNet50 architecture to "True". The F1 scores from these classifications were also compared to those from the multi-hazard datasets. For all the models, training was performed over 100 epochs with an early stopping set of 50 if the F1 score did not improve.

Multiclass Classification of Multi-hazard Dataset

Using the multi-hazard training dataset, the modified xBD baseline architecture, and the other techniques as described, a multiclass classification accuracy was assessed on the testing dataset of each of the hazards. The model's performance was measured using F1 scores across the four classes: "no-damage", "minor-damage", "major damage" and "destroyed". The F1 scores are summarized across each hazard type in Table 3.

Table 3. F1 scores for multi-hazard datasets, classified into "no damage", "minor damage", "major damage", and "destroyed", using the modified xBD baseline architecture. The average F1 score for all data combined is 0.51.

Disaster Types	Flooding	Wind	Fire
Average_F1_Score	0.40	0.66	0.42
no_Damage	0.71	0.77	0.91
Minor_Damage	0.23	0.56	0.03
Major_Damage	0.45	0.65	0.05
Destroyed	0.22	0.66	0.70

The wind-related dataset had the highest overall F1 score of 0.66 (Table 3), with all its classes above F1=0.5. Fire dataset is identifiable for "no-damage" and "destroyed" classes, but not identifiable for "minor damage" and "major damage" classes. Flooding is identifiable for "no-damage" and maybe "major damage" but had F1 of 0.22 and 0.23 for "minor damage" and "destroyed" classes respectively. Generally, the F1 scores of "major", "minor", and "destroyed" classes are lower than the F1 score of "no damage" class. This indicates that the algorithm is unable to differentiate these classes.

Binary Classification of Multi-hazard Dataset

Using the multi-hazard training dataset with binary damage classes, the modified xBD baseline architecture for binary classification, and the other techniques as described, a binary class classification accuracy was analyzed, using the testing data of each of the hazard types, separately. The F1 scores for binary classification of the multi-hazard dataset is summarized across each hazard type in Table 4.

Table 4. F1 scores of binary classification of multi-hazard data. The average F1 score for all data combined is 0.72 and 0.71 (without Tsunami).

	Wind	Flood	Fire	Tsunami
Average_F1_Score	0.77	0.54	0.81	0.75
No damage	0.83	0.83	0.93	0.95
Damage	0.70	0.22	0.68	0.54

In Table 4, the fire dataset had the highest average F1 score (0.81) and the most identifiable "no damage" class (0.93). The wind dataset had the second highest F1 score (0.77) with similar F1 scores across damage and no damage (0.70 versus 0.83, respectively). The tsunami dataset had a similar F1 score (0.75) with a high F1 score for no damage (0.95) and a lower F1 score for damage class (0.54). The flood dataset had the lowest F1 score (0.54), divided between a well identified "no damage" class (0.83) and low identifiability for the "damage" class (0.22).

Classification of Dataset, Separately

The same architectures used for multiclass and binary classifications of the multi-hazard dataset were retrained using only images from a single hazard, resulting in individual models for wind, flood, tsunami, and fire. Due to the imbalance of the training dataset, we did not train a multi-class model for tsunami. The training was performed over 100 epochs with an early stopping of 50 epochs if the F1 score did not improve. The model parameters remained the same, except that the training and testing of the algorithms were performed on the individual hazard datasets separately.

Multiclass Classification by Hazard

Table 5 shows F1 scores, using the modified model trained only using hazard specific training images. When trained with hazard specific training images, the wind building damage was the most identifiable and the wind-trained model successfully identified damage across all four classes (F1>0.5). The wind-trained model had the highest average F1 score (0.70) and the highest F1 scores for all the four classes (0.83 for 'no damage', 0.58 for 'minor damage', 0.67 for 'major damage', and 0.71 for destroyed). Similar to the multihazard model presented above, the fire specific model was able to identify "no damage" and "destroyed" classes, but was not able to identify minor or major classes. The flood specific model performance was also similar to the multi-hazard model, with a high F1 score for "no damage" (0.76) but low F1 scores for the damage classes (0.21-0.32).

Table 5. F1 scores for multiclass classification of datasets using hazard specific training data. The average F1 score for all data combined is 0.52.

Disaster Types	Flooding	Wind	Fire
Average_F1_Score	0.39	0.70	0.46
no_Damage	0.76	0.83	0.91
Minor_Damage	0.21	0.58	0.08
Major_Damage	0.32	0.67	0.14
Destroyed	0.28	0.71	0.72

Binary Classification of Dataset, Separately

Table 6 shows the F1 scores when the modified binary models were trained using hazard specific training images. Binary classification models trained with fire, wind, and flood specific images all result in average F1 scores above 0.70. The "no damage" class is consistently more identifiable than the "damage" class; however, all F1 scores are above 0.5. The F1 scores of "damage" buildings were 0.65 for fire; 0.72 for wind; and 0.56 for flood.

Table 6. F1 scores of binary classification using hazard-specific trained models

	Wind	Flood	Fire	Tsunami
Average F1 Score	0.77	0.72	0.78	0.50
No damage	0.82	0.87	0.93	1.00
Damage	0.72	0.56	0.65	0.00

Multiclass or Binary Classification

The "no damage" class is the most identifiable class across all of the models. The F1 scores of the "no damage" classes were higher when using a binary class rather than when classified into four using the multi-hazard dataset (Tables 3 and 4). These gains could be attributed to ResNet50 layers that were made trainable for binary classifications, making them use weights from the images as a replacement for weights from ImageNet: wind (0.77 vs 0.83); flood (0.71 vs 0.83); and fire (0.91 vs 0.93). Since the number of "no damage" buildings were the same in separate datasets for binary and multiclass classifications, the gains in F1 scores (Tables 5 and 6) were not due to changes in the number of "no damage" buildings, but instead due to the ResNet50 layers of the xBD baseline architecture, which were made trainable for binary classification.

To evaluate the models overall and compare multiclass or binary class models, we use the average F1 score. The average F1 score for wind specific models was higher when separated for binary classifications than multiclass classifications (0.77 vs. 0.70). The average F1 score for flood specific models improved significantly for binary classification than for multiclass classification (0.72 vs. 0.39). Fire specific models resulted in similar significant improvements in F1 for binary over multiclass models (0.78 vs. 0.46). Perhaps, the ability of the algorithm to detect the "no damage" class in binary classifications was diluted by the merged classes.

In summary, average F1 scores were higher for binary classification (modified from the xBD labels) than for multiclass classification for all three hazards (Figure 13). Wind damage is the only damage that can be successfully identified across all four classes. When using the multiclass models, the destroyed class is identifiable for fire, volcano, and tsunami (F1=0.72, 0.83., 0.67, respectively) but the minor and major classes are not well identified.

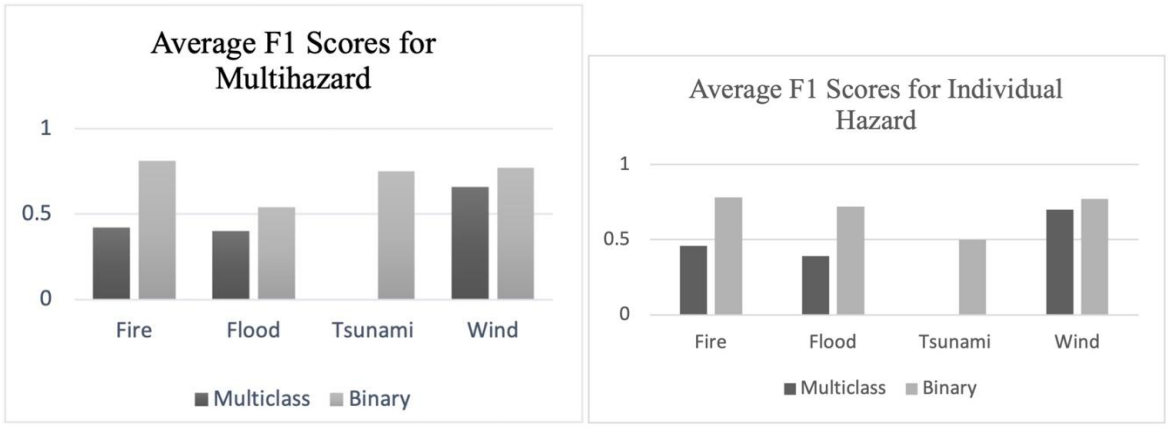


Figure 13. Average F1 scores are higher for binary than multiclass classifications.

Multi-hazard Or Individual Hazard

For wind building damage, the F1 scores improved for all classes when the model was trained with wind-specific data, where the "minor damage" and "major damage" labels are not well identified, the average F1 scores were higher when trained with hazard specific data than those resulting from multi-hazard training data. On the other hand, for flood building damage, "minor damage" and "major damage" levels benefitted from the pooling the training dataset (Table 7). The small training dataset such as the tsunami training datasets used herein benefited from the increased training data size in the pooled multi-hazard training data (Table 4).

Table 7. Highlighted cells are the gains in F1 scores when datasets were pooled instead of separated for multiclass classification.

F1 Scores	Flooding	Wind	Fire
Average_F1_Score	0.01	-0.04	-0.04
no_Damage	-0.05	-0.06	0.00
Minor_Damage	0.02	-0.02	-0.05
Major_Damage	0.13	-0.02	-0.09
Destroyed	-0.06	-0.05	-0.02

For the binary classification summarized in Tables 4 and 6, the F1 scores were not significantly impacted by pooling data across hazards versus using hazard specific training data for wind and fire related building damage. On the other hand, flood specific trained models performed significantly better for detecting flood related building damage than binary models trained with multi-hazard data. F1 scores increased for the "damage" class in binary classification of flood from 0.22 to 0.56. Fire related building damage detection benefited slightly from pooling the datasets (F1 score increased from 0.65 to 0.68) (Figure 8).

Table 8. Highlighted cells are the gains in F1 scores when datasets were pooled instead of separated for binary classification.

F1 Scores	Flooding	Wind	Fire
Average_F1_Score	-0.18	0.00	0.03
no_Damage	-0.04	0.01	0.00
Damage	-0.34	-0.02	0.03

In 2020, Valentijn et al. performed experiments that showed that the performance in terms of area under the curve does not correlate with the type of natural hazard, geographical region, and satellite parameters such as the off-nadir angle, which support our findings that models' performances vary across natural hazards. However, unlike the CNN model used by Valentijn et al., which concatenates features extracted from the pre- and post-disaster images, the CNN model used in this study did not concatenate features, but instead relied on only the abilities of the network to learn and identify damages from the post-disaster images. Valentijn et al. chose the concatenated CNN design because it

enables the model to use the specifics of the before and after imagery to learn the degree of damage. In February of 2023, Kaur et al. proposed a similar algorithm that included a hierarchical transformer architecture (DAHiTrA). This algorithm differs in that it concatenates only the changes between pre- and post-disaster images. Kaur et al. argued that directly concatenating the features complicates the task of localizing pixels for damage assessment. This study takes a different approach; we use only the post disaster images to train a fine-tuned xBD base model, developed from RESNet50. In addition, our approach did not perform any localization of pixels. We demonstrated the use of only the post-disaster images for training damage assessment models. This is useful in disaster events where pre-disaster images are not readily available.

Other studies have also shown that models trained on xBD dataset can easily overfit to the "no damage" class. In 2023, Malmgren et al. trained a dual task UNET algorithm on an xBD dataset for damage assessment after the Turkey Earthquake and almost only correctly identified "no damaged" buildings, achieving F1 scores of 0.95 during training as well as 0.93 and 0.83 in the damage assessments of Kahramanmaras and Antakya, respectively, which shows that their model overfitted to the no-damage buildings. It also shows that it was challenging for their model to detect the minor and major damages.

Deep learning algorithms struggle to detect "minor" and "major" damages effectively in the xBD dataset for several reasons. One of these reasons is the imbalanced nature of the xBD dataset. Even when all the xBD disasters are combined, the "minor" and "major" damage classes are underrepresented in the whole dataset. The wind dataset has most of the "minor" and "major damages", but the total number of buildings in both classes combined still pales in comparison to the "no damage" class. This imbalance causes the deep learning algorithm to overfit to the "no-damage" class since it is over-exposed to the "no damage" class and under-exposed to the relatively fewer "minor" and "major damage" classes. The low accuracy in the minor and major classes might also be due to the misclassification of some of the training dataset by the analysts. Errors might also come from the nuance in the difference in the description between the major and minor classes. Finally, when compared to the "destroyed" and the "no damage" classes, the minor and major damages are also more challenging to visualize on a satellite image, making it equally challenging for a deep learning algorithm to detect these classes.

Conclusions

Convolutional neural networks when paired with labeled image libraries such as the xBD dataset can be used for automating the creation of building damage events after natural hazards. The task of training a convolutional neural network to classify buildings' damage levels due to a disaster is a challenging task. This challenge is further compounded by an imbalanced dataset, which causes overfitting, or has insufficient training data to train the algorithm. In this project, we evaluate the xBD dataset and available CNNs for this purpose. We balanced the xBD dataset to reduce overfitting and achieve a better classifier performance. We used transfer learning from ResNet50 to reduce overfitting, by fine-tuning the algorithm with batch normalization and dropout layers. In addition, we improved the performance with long epochs, combined with early stopping. The resulting model was created by adjusting only the top layers of a ResNet50. With the weights from ImageNet retained, five new blocks of convolution, batch normalization, and max pooling layers were added on top of the ResNet50. This modified architecture was used to evaluate the identifiability of damage classes across the hazard types available in the xBD dataset: flood, wind, fire, volcano, tsunami, and earthquake. We did not evaluate building damage due to earthquake and volcano as there is insufficient training and testing data related to the damage classes for these events in xBD. Due to the limited damage labels in earthquake and volcano events, it is not possible to evaluate performance. Building damage labels (especially the "minor damage" class) were also limited for Tsunami events. Tsunami was therefore considered only for the multi-hazard binary classification.

The xBD dataset has significant class imbalance, especially in the minor damage and major damage classes; therefore, we explored reducing the dataset to a binary class (no damage and damaged buildings). The damaged class includes minor, major, and destroyed. To perform binary classification, we used the same architecture but with the ImageNet weights turned off.

After comparing models trained with hazard specific data versus multi-hazard data and binary versus multiclass labels, we can make recommendations for future use. Only the wind hazard had sufficient training data across all four classes to consistently train the modified architecture to accurately label all four original damage classes. Wind

building damage is best classified when using a multiclass model trained on wind specific data from xBD. Flood building damage is best classified using a binary classification model and hazard specific training images from xBD. For fire and tsunami building damage, our recommendation is to use the pooled dataset and a binary classification for building damage.

Funding

This research was funded by the National Geospatial Intelligence Agency, grant number # HM04762010006.

References

A. Garcia-Garcia, S. O.-E.-M.-R. (2017, April 22). A Review on Deep Learning Techniques.

Al-Stouhi, S. &. (2016). Transfer Learning for Class Imbalance Problems with Inadequate Data. *Knowledge and information systems*, 48(1), 201–228.

Anirudh, R., Jungkyo Jung, J., Vitor Silva, S., Giuseppe, M., &. (2022). Earthquake Building Damage Detection based on Synthetic Aperture Radar Imagery and Machine Learning. *Natural Hazards and Earth System Sciences*.

Aston Zhang, Z. C. (2021). *Dive into Deep Learning.*

Bello, O., & Aina, Y. (2014). Satellite Remote Sensing as a Tool in Disaster Management and Sustainable Development: Towards a Synergistic Approach. *Procedia Soc. Behav. Sci.*, 365–373.

Boyle, T. (2019, February 3). Methods For Dealing With Imbalanced Data. pp. https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18.

Chen, B., Chen, Z., Deng, L., Duan, Y., & Jie, Z. (2016). Building change detection with RGB-D map generated from UAV images. *Neurocomputing*, 208.

Cheryl W. J.Tay, S.-H. S. (2020). Rapid food and damage mapping using synthetic aperture radar in response to Typhoon Hagibis, Japan. *nature*, 1-9.

Cooner, A., Shao, Y., & Campbell, J. (2016). Detection of Urban Damage Using Remote Sensing and Machine Learning Algorithms: Revisiting the 2010 Haiti Earthquake. *Remote Sensing*, 8(10):868.

Copernicus: Sentinel-2 - Satellite Missions - eoPortal Directory". (2022, June 18). Retrieved from COPERNICUS: SENTINEL-2 - SATELLITE MISSIONS - EOPORTAL DIRECTORY: https://earth.esa.int/web/eoportal/satellite-missions/c-missions/copernicus-sentinel-2

Dell'Oro, L. (2017, February 21). *Overview of UNITAR-UNOSAT and GDACS Satellite Mapping Coordination System.* Retrieved from Appliedsciences.nasa.gov: https://appliedsciences.nasa.gov/sites/default/files/2020-11/gdacspart1bv2.pdf

Dong, L., & Shan, J. (2013). A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. *ISPRS J. Photogramm. Remote Sens*, 85–99.

Ehrlich, D. G. (2009). Identifying damage caused by the 2008 Wenchuan earthquake from VHR remote sensing data. *International Journal of Digital Earth,* 309–326.

El Naqa I., M. M. (2015). What Is Machine Learning? *Machine Learning in Radiation Oncology*.

Federal Emergency Management Agency (FEMA, 2018). Hazus hurricane model user guidance. Technical report,

Federal Emergency Management Agency, Apr. 2018. 2

Fraile, J. G. (2019, March 7). *Data Science and Satellite Imagery*. Retrieved from towards Data Science: https://towardsdatascience.com/data-science-and-satellite-imagery-985229e1cd2f

Gao. (1996). A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of The Environment*, 257-266.

Garbin, C. Z. (2020). Dropout vs. batch normalization: an empirical study of their impact to deep learning. *Multimedia Tools and Applications*, 79.

Gupta, R. H. (2019). xBD: A Dataset for Assessing Building Damage from Satellite Imagery. *Computer Vision and Pattern Recognition*.

Japkowicz, N. (2000). The class imbalance problem: Significance and strategies. *In proceedings of the 2000 international conference on artificial intelligence (ICAI)*, 111–117.

Johnson, J. K. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 27.

Jude W. Shavlik, T. G. (1990). *Reading in machine learning.* San Mateo, California: Morgan Kaufmann Publishers, Inc.

Kaur, Navjot, et al. "Dahitra: Damage Assessment Using A Novel Hierarchical Transformer Architecture." Texas A&M University, 2022.

Koshimura, S., Moya, L., Mas, E., & Bai, Y. (2020). Tsunami Damage Detection with Remote Sensing. *Geosciennces*, 177.

Kubat M, M. S. (2000). Addressing the curse of imbalanced training sets: one-sided selection. *Fourteenth international conference on machine learning.*

Laigen Dong, J. S. (2013). A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. *ISPRS Journal of Photogrammetry and Remote Sensing*, 85-99.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, pp. 436-444.

Leevy, J. L., Khoshgoftaar, T. M., & Bauder, R. A. (2018). A survey on addressing high-class imbalance in big data. *J Big Data, 5*, 42.

Li Yundong, W. H. (2019, March 13). Building Damage Detection from Post-Event Aerial. *Applied Sciences*. lidarradar.com. (2022, Febraury 11). *LiDAR and RADAR Information*. Retrieved from LiDAR and RADAR Information: lidarradar.com

Lillesand, K. &. (2004). *Remote Sensing and Image Interpretation.* Hoboken, NJ, USA: John Wiley and Sons, Inc.

Linlin Ge, A. H.-M. (2015). Near real-time satellite mapping of the 2015 Gorkha earthquake, Nepal. *Annals of GIS*, 175-190.

Maithra Raghu, J. K. (n.d.). Transfusion: Understanding Transfer Learning for Medical Imaging. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019).* Vancouver, Canada.

Malmgren, J., & Karlberg, T. (2023, 10 18). *Deep Learning for Building Damage Assessment of the 2023 Turkey Earthquakes.* Retrieved from KTH Royal Institute of Technology: https://kth.diva-portal.org/smash/get/diva2:1795168/FULLTEXT01.pdf

Mason, D. C. (2021). Floodwater detection in urban areas using Sentinel-1 and WorldDEM data. *J. Appl. Remote Sens.*, 5 (03).

Masroor Hussain, D. C. (2013, March 22). Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS Journal of Photogrammetry and Remote Sensing*, pp. 91-106.

Mateusz Buda, A. M. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Computer Vision and Pattern Recognition*, 13.

Mehryar Mohri, A. R. (2018). *Foundations of Machine Learning.* The MIT Press.

Naito, S., Tomozawa, H., Mori, Y., Nakamura, H., & Fujiwara, H. (2018). Damage Detection Method for Buildings with Machine-Learning Techniques Utilizing Images of Automobile Running Surveys Aftermath of the 2016 Kumamoto Earthquake. *Journal of Disaster Research*, 928-942.

NASA. (2021, September 15). *Haiti Earthquake, Landslides & Flooding 2021*. Retrieved from Haiti Earthquake, Landslides & Flooding 2021: https://appliedsciences.nasa.gov/what-we-do/disasters/disasters-activations/haiti-earthquake-landslides-flooding-2021

Qiuhong Ke, J. L. (2018). Chapter 5 - Computer Vision for Human–Machine Interaction. In *In Computer Vision and Pattern Recognition* (pp. 127-145). Academic Press.

Ramirez-Herrera, M., & Navarrete-Pacheco, J. A. (2012). Satellite Data for a Rapid Assessment of Tsunami Inundation Areas after the 2011 Tohoku Tsunami. . *Pure Appl. Geophys.*, 1067–1080.

Rashidian, V. a. (2021). Detecting Demolished Buildings after a Natural Hazard Using High Resolution RGB
Satellite Imagery and Modified U-Net Convolutional Neural Networks. *Remote Sensing*, 2176.

Scikit-learn. (2020). *Scikit-learn F1-Score*. Retrieved from Scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

Sghaier et al., 2. M. (2017). River extraction from high-resolution SAR images combining a structural feature set and mathematical morphology. *IEEE J. Sel. Topics. Appl. Earth Obs. Remote Sens.*, 1025-1038.

Shunichi Koshimura, L. M. (2020). Tsunami Damage Detection with Remote Sensing: A Review. *geosciences*, 2-28.

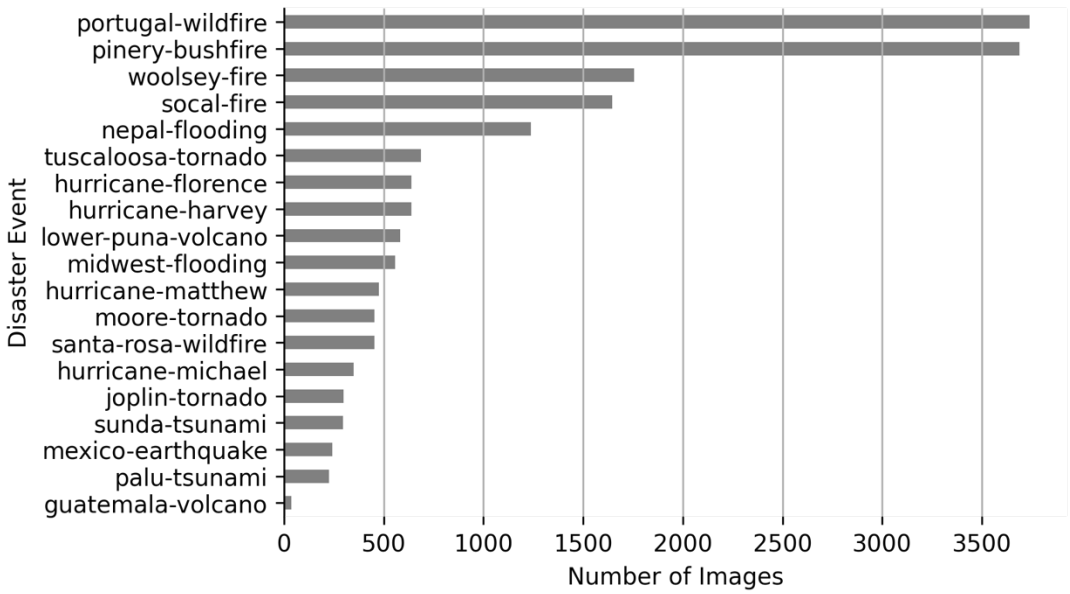
Siti Nor Khuzaimah Binti Amit, Y. A. (2017, September). Disaster Detection from Aerial Imagery with Convolution Neural Network.

Stephenson, L. O., Kohne, T., Zhan, E., Cahill, E. B., Yun, S.-H., Ross, Z., & Simons, M. (2021). Deep Learning-Based Damage Mapping With {InSAR} Coherence Time Series. *{IEEE} Transactions on Geoscience and Remote Sensing*, 1-17.

Suppasri, e. a. (2012). Remote Sensing: Application of remote sensing for tsunami disaster. *Remote Sensing of Planet Earth*, 143-168.

Tay, C., Yun, S.-H., Chin, S., Bhardwaj, A., Jung, J., & Hill, E. (2020). Rapid Flood and Damage Mapping Using Synthetic Aperture Radar in Response to Typhoon Hagibis, Japan. *Sci. Data*, 100–108.
TensorFlow. (2021, October 3). *Data augmentation: TensorFlow Core*. Retrieved from Data augmentation: TensorFlow Core: https://www.tensorflow.org/tutorials/images/data_augmentation
Tewkesbury, A. J. (2015, January 9). A critical synthesis of remotely sensed optical image change detection techniques. *Remote Sensing of Environment*.
Théau. (2008). Change Detection. In X. H. Shekhar S., *Encyclopedia of GIS.* Boston: Springer.
Valentijn, T. (2020). Multi-Hazard and Spatial Transferability of a CNN for Automated Building Damage Assessment. *remote sensing*.
Xu Lu, W. J. (2019). High-Resolution Remote Sensing Image Change Detection Combined With Pixel-Level and Object-Level. *IEEE*, 78909-78918.
Xu, S. D. (2022). Seismic multi-hazard and impact estimation via causal inference from satellite imagery.*Nat Commun* , 13.
Ya-Lun S. Tsai, A. D. (2019). Remote Sensing of Snow Cover Using Spaceborne SAR: A Review. *Remote Sensing.*, 1456.
Yann LeCun, Y. B. (2015). Deep learning. *Nature*, 436.
Yuan, X. &. (2021). Automated Building Segmentation And Damage Assessment From Satellite Images For Disaster Relief. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 741-748.
Yun. (2015). *United States of America Patent No. 9207318B2.*
Zhang J, M. I. (2003). KNN approach to unbalanced data distributions: a case study involving information extraction. *Proceedings of the ICML'2003 workshop on learning from imbalanced datasets.*
Zhang, J. X., Zhao, Z., Huang, G. M., & Lu, Z. (2012). CASMSAR: An integrated airborne SAR mapping system. *Photogrammetric Engineering & Remote Sensing,* 78(11):1110–1114.

Appendix



Appendix 1. Total number of pre- and post-disaster images per event in training dataset

	Type
volcano	volcano
hurricane	wind
tornado	wind
earthquake	earthquake
flooding	flooding
tsunami	tsunami
bushfire	fire
wildfire	fire
fire	fire

Appendix 2. Disaster Type

Flood

Fire

Tsunami

Hurricane

Volcano

Earthquake

Appendix 3. "no-damage" buildings are usually undisturbed, with no signs of water, structural or shingles damage, or burn masks

Flood

Fire

Tsunami

Hurricane

Volcano

Earthquake

Appendix 4. Buildings with "minor damage" are building with partially burnt, water surrounding structure, volcanic flow nearby, roof element missing, or visible cracks

Flood

Fire

Tsunami

Hurricane

Volcano

Earthquake

Appendix 5 "major damage" can be described as having partial wall or roof collapse, encroaching volcanic flow, or surrounded by water/mud.

Flood

Fire

Tsunami

Hurricane

Volcano

Earthquake

Appendix 6: "Destroyed" buildings can be described as buildings that are scorched, completely collapsed, partially/completely covered with water/mud, or otherwise no longer present.

A building	A row in training data CSV	
	uuid	Label
	e3ebb55b-6334-46ee-98e7-207096b3d163.png	2

Appendix 7: Output from data processing

Chapter 2

Operationalizing the models from Chapter 1 for a new, unseen natural disaster dataset: Case Study of Tonga Volcanic Eruption

Abstract

In recent times, machine learning and deep learning algorithms have proven to be more efficient than traditional change detection techniques in extracting changes or damage due to a natural or a man-made disaster from a satellite image. However, these algorithms need a large number of images and post-disaster change or damage labels for training to accurately label damage. These images and labels may be challenging to collect, costly, and not readily available. Transfer learning has been used in deep learning to reduce the need for large datasets of labeled images. In transfer learning, labeled images for a different set of objects can be used to pre-train an algorithm. Then a smaller set of labeled images with the desired objects labeled is used as a second step. Transfer learning is an efficient approach for post-disaster damage assessment, especially in developing countries.

In 2019, xBD dataset was one of the highest-quality publicly available datasets of high-resolution satellite imagery, released to advance the fields of change detection, humanitarian disaster response, and building damage assessment (Gupta, 2019). The dataset contains pre- and post-disaster satellite images across multiple disaster events. It also includes building polygons and their corresponding damage level. One drawback of the xBD dataset is that it exhibits class imbalance. So, prior work by Sodeinde et al. (2023) looked at improving the xBD dataset and its base algorithm and made recommendations on using either a multi- or an individual hazard model for a post-disaster damage identification and classification. The multi-hazard model was trained on earthquake, fire, flooding, tsunami, volcano, and wind hazards. The individual hazard models were trained on each of the hazards, separately. For each of the disaster types, the study also made recommendations on whether to use a binary or a multiclass model. To avoid overfitting to the "no damage" class, the binary models were trained on dataset created from merging classes by keeping the "no damage" class. A second merged class ("damaged") was created to represent "minor damage," "major damage," and "destroyed." This paper followed these recommendations to select a binary model trained on the multi-hazard xBD dataset to perform damage assessments of the town of 'Eua, Tonga. The buildings in 'Eua, Tonga were affected by the January 2022 volcanic eruption and its resulting tsunami. However, the damages to the buildings were primarily caused by the tsunami.

Post-disaster high-resolution satellite images from Maxar Technologies were used to classify the damages in 'Eua. Validation of the results was based on United Nations Satellite Center (UNOSAT) visual interpretation of a different satellite imagery, which are different from xBD dataset classification levels. Because of this, we merged the UNOSAT dataset from five into two classes. These classes represent level of damage. We discovered that the binary model trained on a multihazard xBD dataset (i.e., dataset of earthquake, fire, flooding, tsunami, volcano, and wind hazards) was not transferable; it overfitted to the no-damage class. However, the multiclass model that was trained on a multihazard xBD dataset detected 96% of no-damage buildings and 84% of damaged buildings from the post-disaster image of Ohonua, Eau, Tonga when grouped into two classes.

Introduction

Usually, machine learning and deep learning models are built with the assumptions that training and test data are from the same feature space, and that the data distribution are the same. However, when there are differences between the training and the test data feature space and data distribution, the performances of the models degrade (Weiss, Khoshgoftaar, & Wang, 2016). To combat this deficiency, transfer learning uses models trained on a related source domain on a target domain. Many transfer learning models have been applied to the field of image classification (Garcia-Garcia, et al., 2017). Some of these models were made popular due to their successes in computer vision

challenges. A popular challenge is the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which produces some of the most popular deep learning architectures for image classification, including AlexNet, VGG-16, GoogLeNet, and ResNet. The ILSVRC architectures were built using the ImageNet library. The ImageNet library contains 1,000 object classes, 1,281,167 training images, 50,000 validation images and 100,000 test images (ImageNet, 2023). Some of the ImageNet classes include buildings, animals, and food. Other popular image datasets are CIFAR and MNIST. The CIFAR-10 dataset consists of 60,000 images in 10 classes (CIFAR-100, 2023). The MNIST is a database of handwritten digits. It has a training set of 60,000 examples, and a test set of 10,000 examples (THE MNIST DATABASE, 2023).

ResNet, which won the 2015 ImageNet ILSVRC challenge, is one of the most popular ImageNet architectures. It solves the problem of training a deep neural network by introducing identity skip connections so that layers can copy their inputs to the next layer. This makes it possible for the next layer to learn new and different information from the input layer since it is provided with both the output of the previous layer and the unchanged input layer (Garcia-Garcia, et al., 2017). ResNet architectures are usually named according to their depths. Some of the popular ones include ResNet18, ResNet34, ResNet-50, ResNet-101, ResNet-110, ResNet-152, ResNet-164, and ResNet-1202. The numbers at the end of the names indicate the number of layers in the architecture. For example, ResNet-50 is a 50-layer ResNet architecture. It also happens to be one of the most vibrant ResNet architectures. ResNet50 forms the backbone for the baseline algorithm of xView2, a challenge which uses satellite imagery to identify buildings and classify them according to damage levels. ResNet34 is popular as well. In 2021, Abdi et al customized ResNet34 to identify damaged and non-damaged buildings, using post-event images of the 2010 Haiti earthquake and the Woolsey fire with overall accuracy of 91%.

In transfer learning, a pre-trained CNN model could be easily used as a base model or adapted to classify a new dataset, saving time, effort and cost needed for training a completely new model. For example, a pre-trained CNN model of the 2010 Haiti earthquake can be used as a base model to classify the 2023 Turkey-Syria earthquake. If the accuracy of the classification is poor (i.e., the model is not transferable), the pre-trained model can be fine-tuned. Fine-tuning is the process of removing the final layers of a pre-trained model and then replacing them with new ones (Shabbir, et al., 2021). The new layers of the pre-trained model are then trained on the new dataset.

The process of assessing damaged buildings after a natural or manmade disaster is a costly and time-consuming one. It is costly because it requires significant human efforts. It is time-consuming because the process is mostly manual, requiring people to identify damages from satellite or aerial images, which may last several months and cost a lot of money. For example, after the 2010 Haiti earthquake, it took thousands of volunteers and months of effort to label the damaged building. However, the advent of more computational processing powers has allowed for research into faster and more efficient approaches to detecting changes from satellite images. Deep learning is one of these approaches, which is now being researched to automate the process of classifying a post-disaster image into damage classes. In 2017, Nia K.R and Mori G. performed building damage assessment using deep learning and ground-level image data. They focused on building damage assessment, using only post-disaster images. In 2022, Bouchard et al used transfer learning techniques to fine-tune a model and estimated the number of labeled samples required for the model to be functional in operational situations. Also in 2022, Berezina P. et al used a couple of deep learning-based models for damage assessment of Hurricane Michael, using a large, very high-resolution satellite images dataset. Deng L. and Wang Y. (2022) also designed a two-staged building damage assessment network; the first stage was to segment the buildings while the second stage was to classify them into damage levels. Calantropio A. et al (2021) used images from Unmanned Aerial Vehicles (UAV) and deep learning to automatically classify building footprints, which helps in the rapid creation of cartography for emergency response purposes. Saman G. et al (2019) used OpenStreetMap and multi-temporal high resolution satellite images to update maps of damaged buildings during the damage and recovery stages after the 2013 Super Typhoon Haiyan in Tacloban, the Philippines. Alidoost F. et al (2017) proposed a system that combines deep learning and UAV for rapid location of degraded areas after a disaster. Cheng C. et al. (2021) used deep learning to classify post-hurricane UAV images of Hurricane Dorian for the purpose of building damage assessment. Su, et al., (2022) compared the results of deep learning, Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest models in hazard assessment of earthquake disaster chains and concluded that deep learning has the highest accuracy. Sublime S. and Ekaterina K. (2019) applied deep learning to satellite images taken before and after the Tohoku tsunami of 2011 and compared its result to those of other machine learning methods, concluding that deep learning beats existing approaches due to its good performance and relative speed of analysis. Deep learning approaches have been used to assess damages due to a tsunami disaster as well (Meilano, et al., 2020; Andraud, et al., 2022; Fauzi et al., 2019; and Mulia et al., 2020).

In 2020, Meilano et al. studied building damage due to tsunamis in Palu, Indonesia. They used deep learning to extract pre- and post- disaster buildings from satellite images and then analyzed them using tsunami height zone map to see correlations between the damaged buildings and the tsunami heights. Andraud et al. (2022) explored using deep

learning for rapid forecasting of coastal flooding in real time. They performed a preliminary comparison of the coastal impact captured by the nonlinear time-consuming tsunami approach to the one generated by the deep learning approach for the French Mediterranean coastline. According to Fauzi et al., (2020), the nonlinear tsunami inundation forecasting approach also has the disadvantage of having relatively high computational cost. So, they performed a study using deep learning, to estimate tsunami inundation in real time. The case study areas were Atashika and Owase Bays in Japan and the results indicated that the deep learning approaches are extremely fast when compared to the nonlinear tsunami inundation forecasting approach. Mulia et al., 2020 applied deep learning to tsunami inundation database of megathrust earthquake, finding that the deep learning approach has comparable accuracy to the conventional physics-based simulation but achieves approximately 90% reduction of real-time computational efforts.

In 2017, Fujita, et al., explored Convolution Neural Networks (CNNs) for classifying whether a building was washed away or not, using pre- and post-tsunami aerial images. This exploration was based on the 2011 Tohoku tsunami in Japan. In the work, the building footprint was used to select the target image patch area. Their experiment concluded that AlexNet and VGG achieved 94-96% accuracy. They also discovered that the use of the pre-tsunami image does not improve detection accuracy. Since building footprints are not always accessible, Bai, et al. (2017) performed an experiment with only the post-disaster TerraSAR-X data from 2011 Tohoku earthquake and tsunami in Japan. They classified the built-up and non-built-up regions as well as the damage-level. They achieved accuracy of 80.4% for the built-up region extraction, and accuracy of 74.8% for the damage level classification. In 2018, Bai, et al. also proposed a U-net CNN to rapidly map damage. They used pre- and post-disaster high-resolution WoldView-2 images of 2011 Tohoku Earthquake-Tsunami for the study. Their proposed method achieved accuracy of 70.9% in classifying the damage into “washed away”, “collapsed”, and “survived” at the pixel level.

Most of the damage assessment methods require urban base maps (e.g., building footprints data) and/or building heights to assess building damage after a disaster (Koshimura, Moya, Mas, & Bai, 2020). When these maps are not accessible, official cadastral information, open-source data (e.g., OpenStreetMap (OSM)), referenced sources (e.g., Google Street Map, Google Earth) and built-up extraction methods are used. OSM is especially useful for deep learning because it is free. In 2017, Chen et al. incorporated OSM with deep neural networks for active detection of buildings or roads for humanitarian aid purposes. Many public prize competitions have also released building footprints along with high resolution images for automated detection of building footprints after a natural disaster. One of these competitions is the building extraction challenges by SpaceNet and its partners: CosmiQ Works, Radiant Solutions, and NVIDIA. SpaceNet challenges provided building footprints for five areas of interest: Rio, Vegas, Paris, Shanghai, and Khartoum. Their goal was to advance automated feature extraction in support of humanitarian efforts and disaster response (Wenrich, 2023). This dataset contained 685,000 building footprints. xBD dataset, a freely available multi-hazard dataset for building detection and damage classification was released for the xView2 challenge. The dataset combines data across 7 different disaster types and 19 different disaster events (Gupta, 2019). As a global dataset of over 40,000 km2 of images, it contains over 850,000 building annotations, across 19 different disaster events. It was developed by DIU and HADR in 2019 for the xView2 challenge and its images were provided in two tiers: Tier1 and Tier3. The first tier (Tier1) was provided by Maxar Technologies/DigitalGlobe as part of its open-data program, which provides free high spatial resolution of pre- and post-disaster images for humanitarian responses. The second tier (Tier3) was not part of the open-data program. It was released as part of a collaboration between Maxar Technologies and the National Geospatial-Intelligence Agency to motivate research. The xBD building footprints were released for clipping images used for training models for building detection and damage classification. However, these building footprints are only useful in areas covered by the xBD dataset. OSM is handy for downloading free building footprints of new disaster events for transfer learning purposes. One drawback of OSM is that not all building footprints are available for a disaster event (Pisla, Li, Lautenbach, Herfort, & Zipf, 2021). If pre-disaster images exist, a transfer learning approach could be used to detect building footprints, using instance segmentation to identify pre-disaster building footprints from pre-disaster images. However, in the absence of pre-disaster images, OSM may be the only option. Using OSM building footprints with missing data can affect the accuracy of a building damage classification. For completeness, these missing footprints data could be sourced from other open-source data such as the open building project from Google Research, which is an effort that currently has 817 million buildings across Africa, South Asia, and South-East Asia (Sirko, et al., 2021). Missing building footprints could also be filled-in from local cadastral information. This paper looks at the effectiveness of building damage classification for a new event using pre-trained models and OSM.

In 2023, Sodeinde et al. analyzed the quality of the xBD image-training dataset for identifying building damage across a variety of natural hazards using deep learning convolutional neural networks. Specifically, they evaluated the pros and cons of combining training datasets across multiple natural hazards and provided recommendations on using the provided training dataset to optimize classification accuracy for building damage detection. They rebalanced xBD dataset, using random oversampling and under-sampling methods and then created pre-trained models from fine-tuned

ResNet50. Initially, they evaluated the identifiability of all the damage classes (i.e., multiclass classification) in the xBD dataset. Because classification performance was significantly higher for the "no damage" class as compared to "minor", "major", and "destroyed" classes, they also evaluated merging classes by keeping the "no damage" class and creating a second merged class ("damaged") to represent "minor damage," "major damage," and "destroyed." They used the same architecture for the multiclass classification for the binary classification but without the ImageNet weights. Based on their work, they recommended that users be aware of performance differences across natural hazards and across damage classes. They concluded that for the multiclass classification, building damage due to tsunami is poorly represented in the training data, and that there is insufficient data for model validation (especially within all damage classes).

In this work, the recommendations from Sodeinde et al were applied to the 15th of January 2022 volcanic eruption on Tonga Island in the southern Pacific Ocean. The volcanic eruption reached a very large and powerful climax, producing the biggest atmospheric explosion recorded on Earth in more than a century and generating tsunamis across the Pacific and other ocean basins around the world. One of the places affected by the resulting tsunami was 'Eau, an island about 41 km southeast of the volcanic eruption (Figure 1 and 2). 'Eua is about 34 km square in area and it has a population of about 5,000 people. So, for this study, we used models from Sodeinde et al (2023).

Figure 1. Map of 'Eua Island

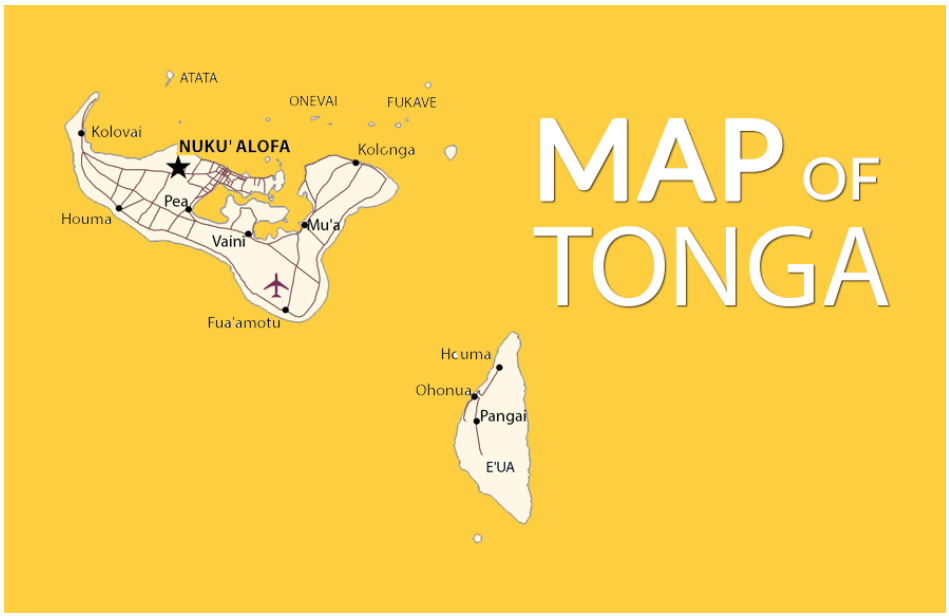


Figure 2. Map of Tonga, showing 'Eua Island

There are two major settlements in 'Eau:'Ohonua and Pangai. Of these two settlements, 'Ohonua was disproportionately affected by the volcanic ashes from Tonga island after its 2022 volcanic eruption because it is along the shore. In this work, we classified the buildings in the town of 'Ohonua into damage classes, using the recommendations and pre-trained models from Sodeinde et al. We applied the multi-hazard models to 'Eau to study the performances of these models on a different, non-xBD tsunami event.

The post disaster images of 'Eua were downloaded from the Maxar Technologies website. The images are NextView, orthorectified, 8 bands, 8 bits per pixel, radiometrically-corrected multiband images, which were pan-sharpened to improve their resolutions. The bounding boxes of the buildings were derived from OpenStreetMaps building footprints. These bounding boxes were used to clip-out the images which were passed to the pre-trained models. The pre-trained model was not fine-tuned or retrained. Each clipped image was reshaped into a 128 X 128 and then passed into both the multi-hazard and tsunami-specific binary models, separately. Validation was based on Unitar visual interpretation from satellite imagery, which only presents a single damaged building label. Because of this binary label, we will only evaluate the binary classification. The confusion matrices from these classifications were used to analyze the identifiability of the damaged buildings.

Data

The town of Eau is located between latitudes -21.28 N and -21.46 S and between longitudes -174.98 W and -174.91 E. Its soil is volcanic in nature, aggregated from the deposit of the volcanic eruption on Tonga Island. Underlying the soil are solid coral rock. The most recent volcanic eruption on Tonga Island occurred on the 15th of January 2022. To identify the damage to Eau, pre- and post-disaster satellite images were downloaded from Digital Globe under the NEXTVIEW agreement between Digital Globe and National Geospatial Intelligence Agency ("NGA"). The pre-disaster images were taken on the 26th of May 2021, while the post-disaster images were taken on 1st of February 2022. Both pre- and post-disaster images were taken, using a WorldView-3 commercial satellite sensor. WorldView-3 operates at an altitude of 617 km, and it can collect up to 680,000 square km per day. The images were orthorectified, 8-bands, GeoTIFFs as listed in Table 1. They were also radiometrically corrected, with pixel depth of 8-bit. The image multispectral 8-bands in VNIR (Visible Near Infrared) include C (Coastal Blue), B (Blue), G (Green), Y (Yellow), R (Red), RE (Red Edge), N (Near IR1), and N2 (Near IR2).

Table 1. Pre- and Post-disaster images

Filename	Date	Type	Spatial Resolution (m)
22FEB01220334-M3DS_R1C2-014988117010_01_P001.TIF	02/01/2022	Post-disaster	1.4
22FEB01220334-M3DS_R2C1-014988117010_01_P001.TIF	02/01/2022	Post-disaster	1.4
22FEB01220334-M3DS_R2C2-014988117010_01_P001.TIF	02/01/2022	Post-disaster	1.4
22FEB01220334-M3DS_R3C1-014988117010_01_P001.TIF	02/01/2022	Post-disaster	1.4
22FEB01220334-M3DS_R3C2-014988117010_01_P001.TIF	02/01/2022	Post-disaster	1.4
22FEB01220334-M3DS_R4C1-014988117010_01_P001.TIF	02/01/2022	Post-disaster	1.4
22FEB01220334-M3DS_R4C2-014988117010_01_P001.TIF	02/01/2022	Post-disaster	1.4
21MAY26214815-M3DS_R1C1-014987729010_01_P001.TIF	05/26/2021	Pre-disaster	2
21MAY26214815-M3DS_R1C2-014987729010_01_P001.TIF	05/26/2021	Pre-disaster	2
21MAY26214815-M3DS_R2C1-014987729010_01_P001.TIF	05/26/2021	Pre-disaster	2
21MAY26214815-M3DS_R2C2-014987729010_01_P001.TIF	05/26/2021	Pre-disaster	2
21MAY26214815-M3DS_R3C2-014987729010_01_P001.TIF	05/26/2021	Pre-disaster	2

The images are projected to Universal Transverse Mercator (UTM) Zone 1S, in meters as shown in Figures 2a and 2b.

Figure 2a. Post-disaster Images

Figure 2b. Pre-disaster Images

The Study Area

Ohonua is the capital of Eau and it is located along the western shore of Eau. It is the bigger of the two settlements. After the volcanic eruption in Tonga, the ashes were transported by the ocean and the air to ‘Ohonua on ‘Eau Island. This study focuses on ‘Ohonua, which is the administrative center of ‘Eau and the largest settlement in the country. ‘Ohonua is situated at a height of 25 m above sea level. According to the Tonga Population Census (2021), there are 632 males and 657 females on the island. The centroid of the island is at 21° 20′ 31.7″ S, 174° 57′ 21.1″ W.

Open Street Maps (OSM)

The building footprints, which capture the pre-disaster state of ‘Ohonua, Eau, were downloaded from OSM (Figure 3). OSM is a free and publicly available geographic database. Included in the database are features such as transport, boundaries, places, and manmade (e.g., buildings). These features are crowdsourced from multiple people across the world from sources such as surveys, aerial images and other freely available geodatabases.

The OSM building layer comes from multiple sources and people all over the world. For this reason, it contains errors due to multiple data sources. Some of these errors lead to misalignments and shifts in building

footprints. Other errors include error of omission, when buildings that should have been included in the dataset are missing in the dataset. Similarly, there were errors due to commission; in these cases, building footprints were incorrectly recorded, when they were actually missing. There were also errors from when the shapes and sizes of buildings were incorrectly represented. People are encouraged to submit the most recent information to correct the wrong data. They are also encouraged to make the edits to fix the data as well. In this study, misaligned building footprints were fixed by realigning them with the building footprint on the pre-disaster images. Errors due to commission and omission were not corrected.

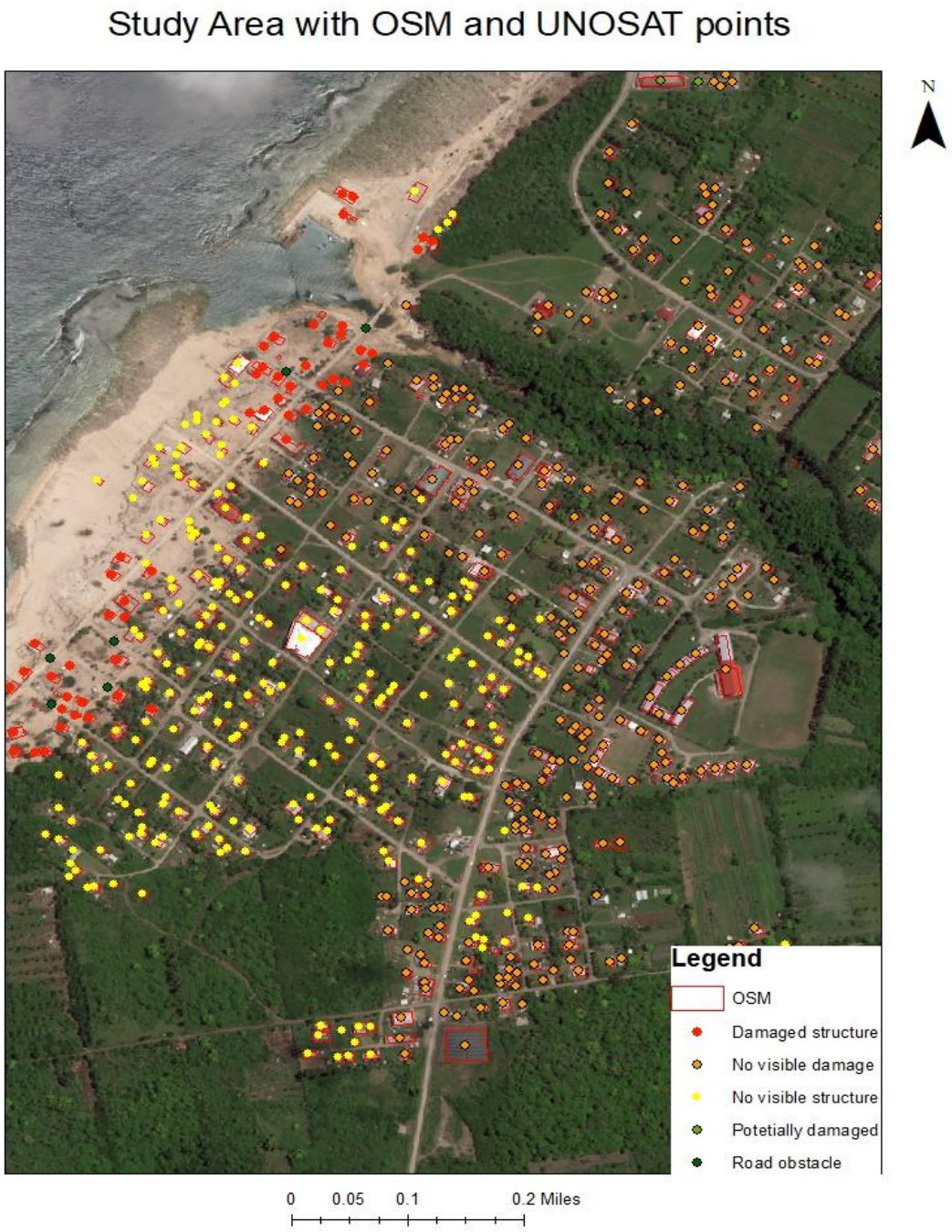


Figure 3. Study Area with building footprint and UNOSAT points

UNOSAT Validation Data

As of 20 January 2022, the United Nations Institute for Training and Research (UNITAR), through United Nation Satellite Center (UNOSAT), observed extensive damage to Eau Island caused by tsunami waves triggered by volcanic eruption. It was observed that the settlement along the shoreline of Ohonua town has been washed away by the tsunami waves. Using the post-event images, dated January 20 and 21, 2022, from the satellite - Pleiades, a preliminary assessment was performed by UNOSAT and maps were created and made available in both shapefile and geodatabase formats, whose metadata contains attributes such as sensorDate, Settlement, and Notes (Table 2.). Among the 2,028 structures that were analyzed in Eau, 1,231 were in the cloud-free area of the satellite images. Among the 1,231 structures, 54 were identified as damaged or destroyed, 1,174 had no visible damage, 797 were not visible on the satellite image due to cloud cover, 3 were potentially damaged, and there were 6 road obstacles as summarized in Table 3. According to UNOSAT, these were preliminary observations that have not yet been verified by ground feedback. This data will be used as the validation dataset for the project.

Table 2. Ohonua UNITAR data metadata

	Field	Description
1	SensorDate	Date of satellite image
2	Settlement	Name of town or city
3	Notes	Type of data
4	EventCode	Disaster event code
5	siteID	Site ID
6	SensorID	Sensor ID
7	Confidence	How reliable is the assessment
8	FieldValid	If verified by ground feedback
9	Main_Dmg	
10	Grouped_Da	
11	StaffID	ID of the staff that perform the assessment
12	ImageID_Nu	ID of the image

Table 3. Count of the Notes

Notes	Count (Eau)	Count (Ohonua)
Damaged structure	54	53
No visible damage	1,174	342
No visible structure	797	292

Potentially damaged	3	3
Road obstacle	6	6

Data Preparation

The binary model trained on pooled dataset required a 3-band clipped image, which encompasses the building whose damage level will be determined. These 3 bands are expected to be the Red, Green, and Blue bands (i.e., RGB). However, because the downloaded post-disaster images are 8 bands, they were first converted to the 3-band RGB images by creating a composite of only the RGB bands to be used in the model. After the conversion, the new 3-band images were saved in Universal Transverse Mercator (UTM) Zone 1S, the coordinate system of the input 8-band images. For clipping, the OSM building footprints, which were in World Geodetic System (WGS) 84, were projected to match the coordinate system of the images.

The building footprints were edited and corrected by shifting them horizontally and vertically to align with the post-disaster image. After the building footprints have been corrected, their bounding boxes were calculated, extended by 80% offset, and then used to clip the post-disaster images. Each clipped image contains a building whose damage level is to be determined. These images were stored as geotiff file format in a personal geodatabase and later exported into PNG file format, making sure that the pixel depth was 8-bit unsigned.

Method

The clipped images of the post-disaster buildings were loaded into pre-made models recommended for classification by Sodeinde et al (2023). At first, a binary model trained on a xBD multi-hazard dataset was used for the classification of the clipped images as recommended by Sodeinde et al (2023) (Figure 4). Later, a multiclass model trained on a xBD multi-hazard dataset was used. In both cases, the UNITAR dataset was used for the validation of the results.

Sodeinde et al (2023) analyzed the quality of xBD training dataset for identifying building damage across a variety of natural hazards using deep learning convolutional neural networks. Specifically, the study evaluated the pros and cons of combining training datasets across multiple natural hazards and provided recommendations on using the provided training datasets to optimize classification accuracy for building damage detection.

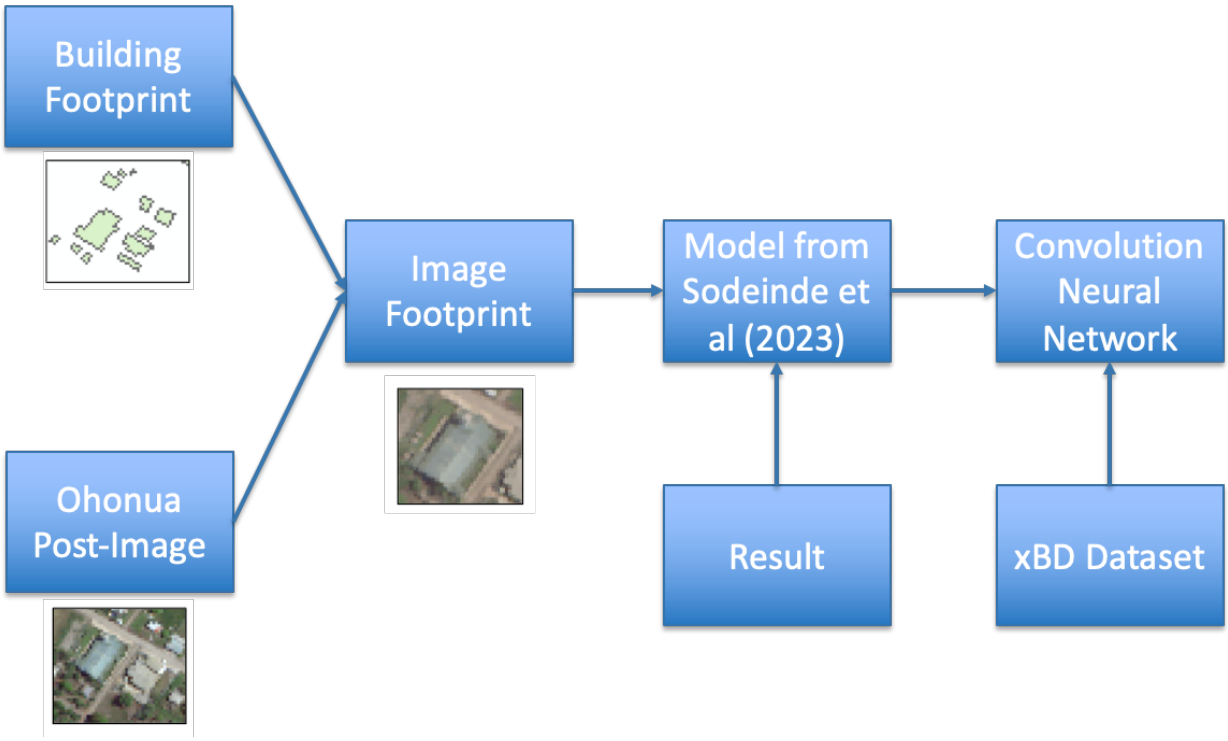


Figure 4. Approach for classifying and verifying Ohonua post-disaster dataset.

The Algorithms

The models from Sodeinde et al (2023) were developed from modified versions of xView2 base algorithm and xBD dataset. The base algorithm was a CNN algorithm released along with xView2 challenge as a starting point for developing solutions for xView2 challenge. The base algorithm was developed using ResNet50 architecture with ImageNet weights, fine-tuned with three sets of convolution and MaxPooling layers (Figure 5). Since this base algorithm overfitted to the "no-damage" class of the xBD dataset, modifications were made in Sodeinde et al (2023) to improve its performance.

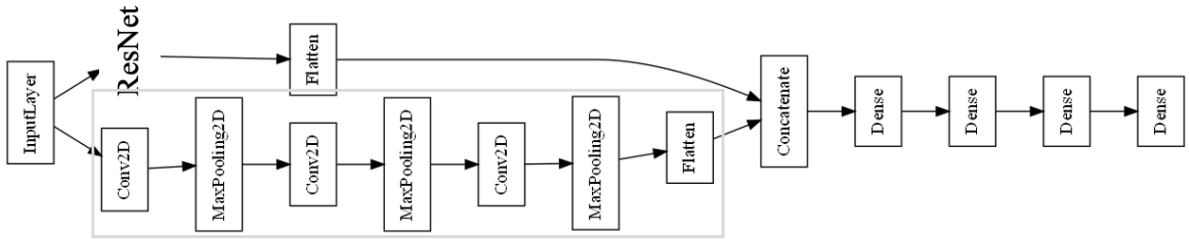


Figure 5. xView2 base algorithm

Modifications for Multi-class Classification

To create a four-class model ("no-damage", "minor-damage", "major-damage", and "destroyed"), the xView base algorithm was further tuned by adding two sets of convolution and MaxPooling layers interleaved with five BatchNormalization layers; one for each of the five blocks of convolution and MaxPooling layers. A Dropout layer with a rate of 0.5 was also added on top to reduce overfitting. This new modification performed better at identifying the "minor-damage", "major-damage", "destroyed", in addition to the "no-damage" class. Longer epochs and early stopping were also added to improve the performance of the models from this algorithm.

Modifications for Binary Classification

Since classification performance was significantly higher for "no-damage"classes as compared to "minor", "major", and "destroyed" classes, Sodeinde et al (2023) also evaluated merging classes and performing a binary classification: one class represented "no damage" and the second class represented "minor damage," "major damage," and "destroyed." The binary model was trained on a similar algorithm as the multiclass model, except that the ImageNet weights from ResNet50 were ignored by setting the trainable property of ResNet50 to true.

Predictions

Without further fine-tuning and additional training, the binary and multiclass models trained on multi-hazard xBD dataset from Sodeinde et al (2023) were used for classification in this study. A building's damage class was predicted by loading a model into the computer memory using the TensorFlow libraries; once in the memory, the model was compiled using an Adam optimizer with an ordinal loss function. Clipped images of the buildings, which were resized into 128X128, converted to an array and normalized over 255 served as inputs into the models. The models' outputs contain probabilities of a building belonging to a damage class. The class with the highest probability represents the predicted class for the building.

Results and Analysis

The classification results for all the clipped images of the buildings were saved in a CSV (Comma Separated Values) file. Each row in the CSV file contains row number, Feature Identification (FID), and the predicted class. The FID is a unique identification number for a building; it maps a building to its ObjectID in the OSM shapefile. The UNOSAT dataset was used to validate the results of the classification.

UNOSAT Dataset

The UNOSAT dataset came in five classes (damaged structure, potential damage, no visible damage, no visible structure, road obstacle) (Figures 6a and 6b). These classes are different from the xBD damage classes (no-damage, minor, major, and destroyed), making it difficult to directly validate the results of the models. For this reason, the UNOSAT dataset was collapsed into two classes (damage and no-damage); first, the "No visible structure" and the "Road obstacle" classes were removed; second, two new classes were created from the remaining classes (damage and no-damage). The buildings with "No visible structure" were obstructed by cloud cover. The new "damage" class consists of the "Damaged structure" and "Potentially damaged" classes from UNOSAT and it is assumed to

correspond to the minor, major and destroyed classes in the xBD dataset. The new "no-damage" class consists of only the "No visible damage" class, and it is assumed to correspond to the no-damage class in the xBD dataset.

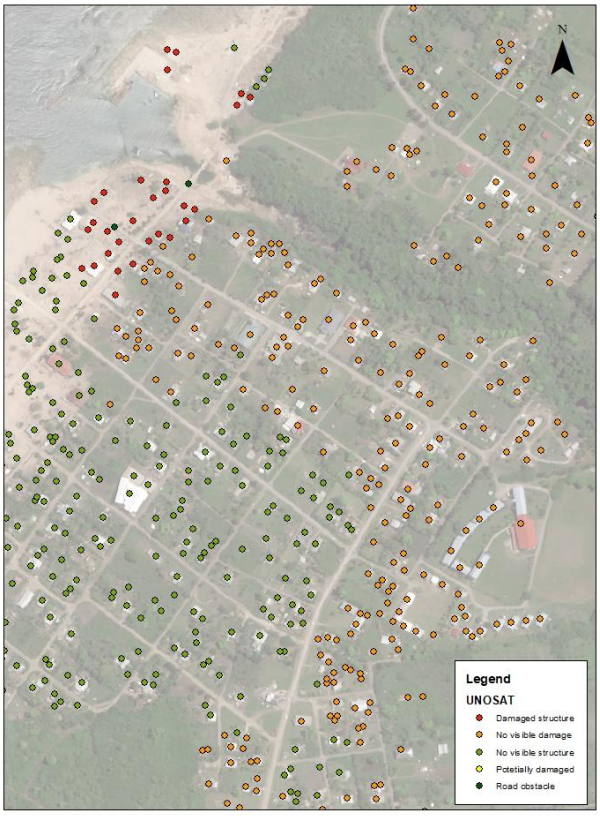


Figure 6a. The five classes of the UNOSAT dataset

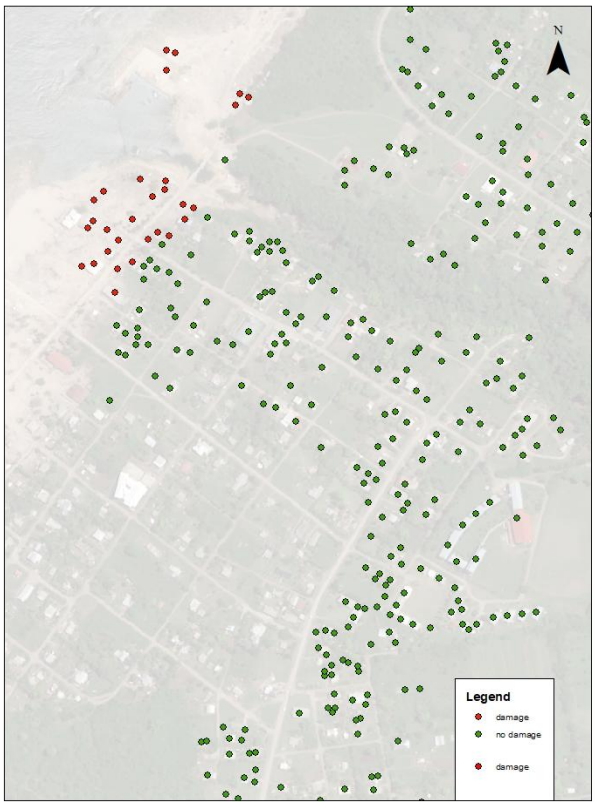


Figure 6b. The merged classes of the UNOSAT dataset

Binary Classification

The binary model trained on a multi-hazard xBD dataset outputs two values for each of the buildings. The class with the higher value was taken as the predicted class. The result of the binary classification is shown in Figure 6. The spatial overlay of UNOSAT on the classification result shows that 342 out 342 no-damage buildings were identified correctly (Figure 7). However, none of the 56 damaged buildings were identified by the model. This indicates that the binary model is not transferable and overfits to the no-damage class. This overfitting might be due to the fact that the dataset is from a non-xBD dataset; it might also be due to the presence of fewer debris in the Ohonua images compared to the xBD images. This difference in amount of debris may affect how the model identifies damaged buildings. To a model, the presence of debris might also indicate the presence of damaged buildings. The confusion matrix and overall accuracy are shown in Table 4.

Table 4: Confusion matrix and overall accuracy for binary classification

	damage	No-damage
damage	0	56
No-damage	0	342

Overall accuracy is 86%

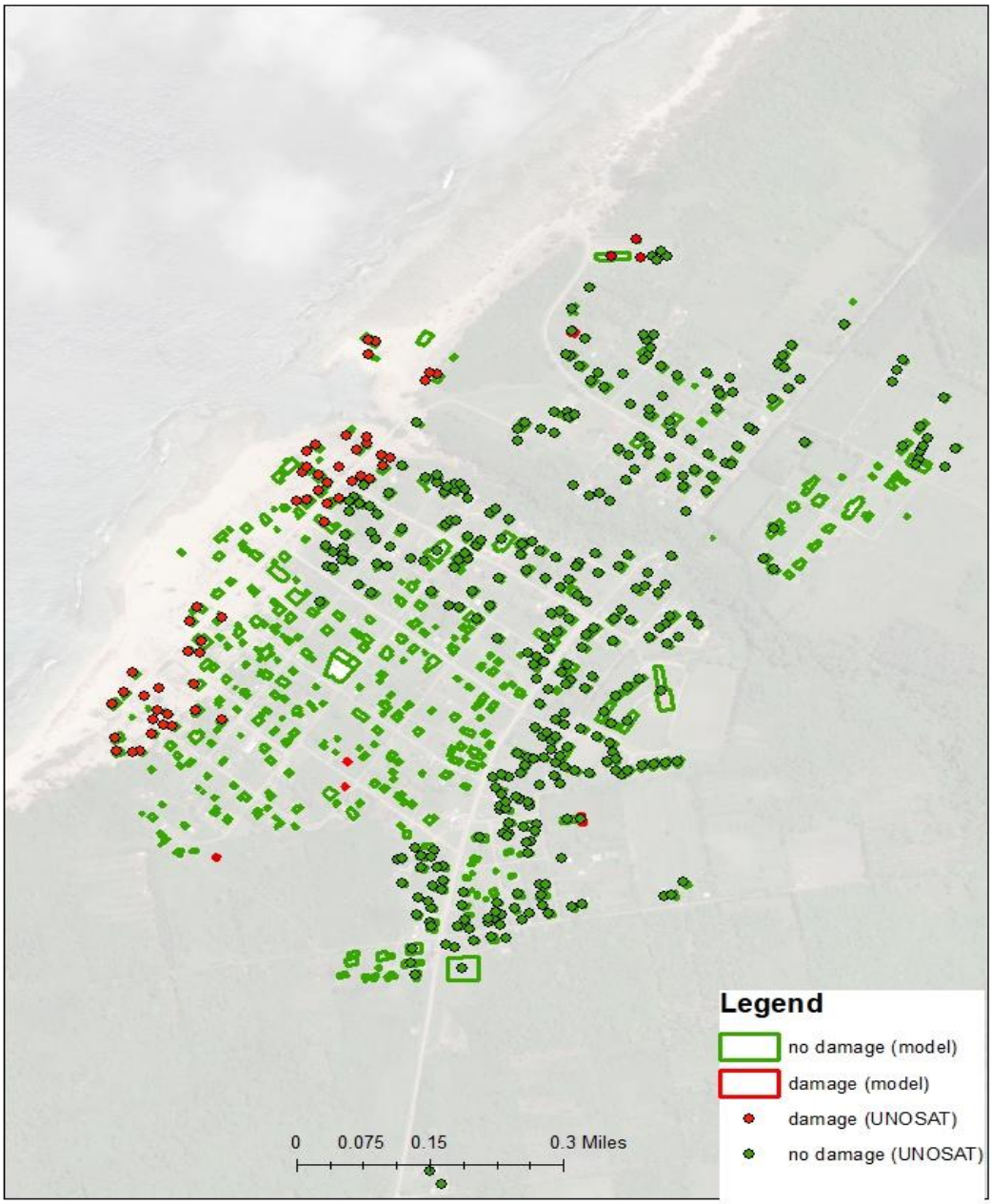


Figure 7. UNOSAT overlaid on binary classification result.

Multiclass Classification

The multiclass model trained on a multi-hazard xBD dataset outputs four values, one for each of the classes. The highest value represents the damage level of the building. Unlike the binary model, the multiclass model performed better. It predicted more damaged buildings than the binary model. This improvement in performance was probably because the model was able to pick up nuances in the damage levels from the images (Figure 8a). However, in order to effectively validate the result, the result was first converted into binary data by merging its classes into two (damage and no-damage) such that the new "damage" class is equivalent to the minor, major, and destroyed classes. We found out that only 9 out of the 56 damaged buildings were misclassified as no-damage and 15 out of the 342 no-damage buildings were misclassified as damaged (Figure 8b). The confusion matrix and overall accuracy are shown in Table 5. This is a substantial improvement over the binary classification model. The multiclass multi-hazard model detected most of the damaged buildings along the coastline, where most of the destruction occurred.

Table 5: Confusion matrix and overall accuracy for multiclass classification

	damage	No-damage
damage	47	9
No-damage	15	327

Overall accuracy is 94%

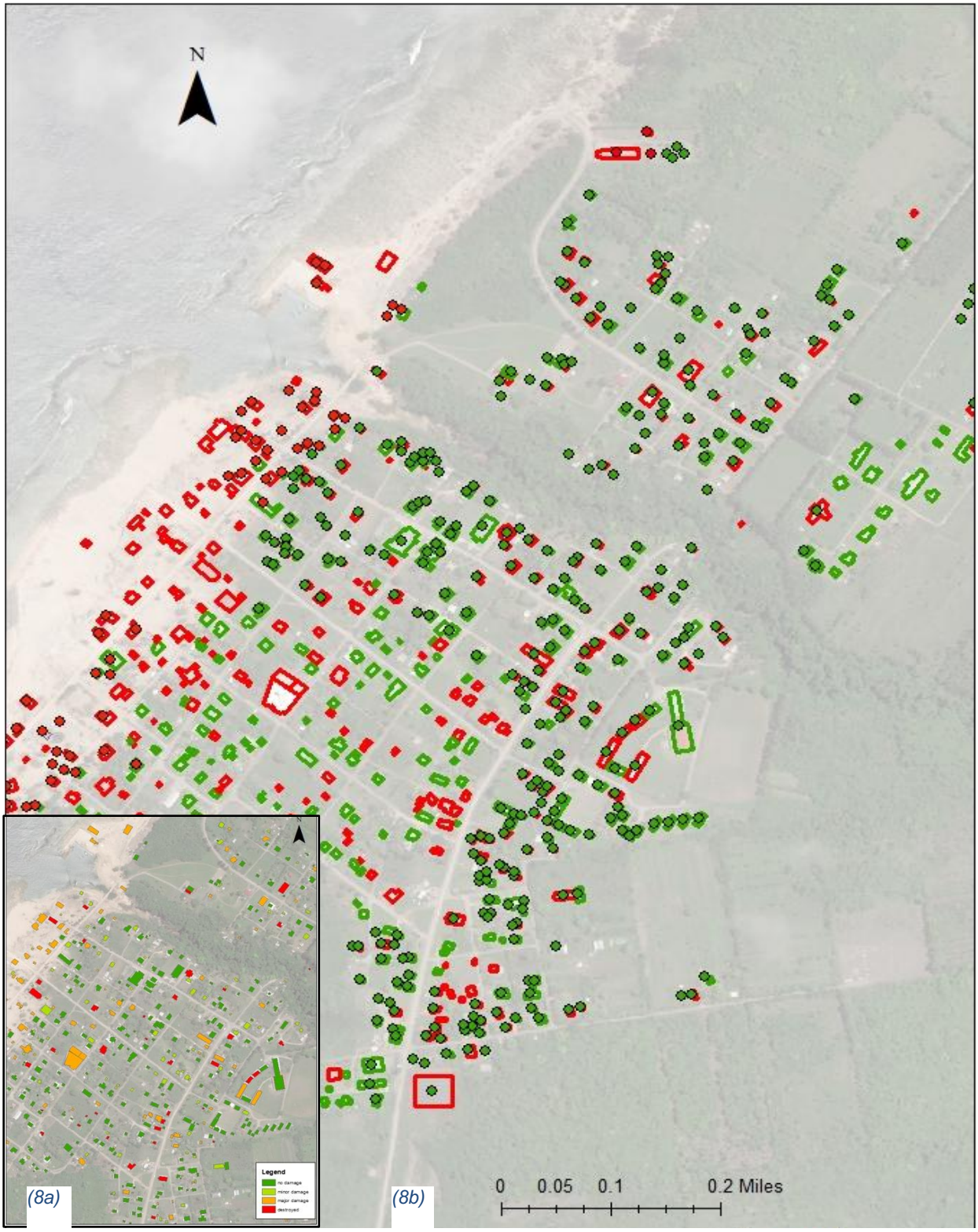


Figure 8a. Multiclass classification result; 8b. UNOSAT overlaid on merged multiclass classification result.

Conclusions

The binary model trained on xBD multi-hazard dataset is not transferable for classification of Ohonua tsunami images as it seems to overfit the no damage building class. However, the multiclass model trained on xBD multi-hazard dataset was transferable; it was able to detect the four building damage classes across the image. In order to validate the results, the multiclass model's result and the UNOSAT validation dataset were grouped into two classes, separately. The multi-class model detected 96% of no-damage buildings and 84% of the damaged buildings in Ohonua. For classification, this paper finds that the pretrained multiclass model developed by Sodiende et al. (2023) can be used on a new tsunami event. For validation, the damage classes were merged into two classes (damage or no damage).

References

Abdi, G., Esfandiari, M., & Jabari, S. (2021). Building Damage Detection in Post-Event High-Resolution Imagery Using Deep Transfer Learning. *International Geoscience and Remote Sensing Symposium* IGARSS (pp. 531-534). Brussels, Belgium: IEEE.

Alidoost, F., & Arefi, H. (2017). Application of Deep Learning for Emergency Response and Disaster Management. *Proceedings of the AGSE Eighth International Summer School and Conference. University of Tehran*.

Andraud, P., Gailler, A., Sprunck, T., & Vayatis, N. (2022). Deep learning models exploration for rapid forecasting of coastal tsunami impact in near field context - application to the French Mediterranean coastline. *24th EGU General Assembly.* Vienna, Austria: EGU.

Ardiansyah, F., & Norimi, M. (2020, March 2020). Machine Learning Algorithms for Real-time Tsunami Inundation Forecasting: A Case Study in Nankai Region. *Pure Appl. Geophys.*, pp. 1437–1450.

Bai, Y., Mas, E., & Koshimura, S. (2018, October 5). Towards Operational Satellite-Based Damage-Mapping Using U-Net Convolutional Network: A Case Study of 2011 Tohoku Earthquake-Tsunami. *Remote Sensing Image Processng*, p. 1626.

Berezina, P., & Liu, D. (2022). Hurricane damage assessment using coupled convolutional neural networks: a case study of hurricane Michael. *Geomatics, Natural Hazards and Risk*, 414-431.

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,* 2532.

Calantropio, A., Chiabrando, F., Codastefano, M., & Bourke, E. (2021). Deep learning for automatic building damage assessment: application in post-disaster scenarios using UAV data. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 113-120.

Cheng, C.-S., Behzadan, A. H., & Noshadravan, A. (2021). Deep learning for post-hurricane aerial damage assessment of buildings. *Computer-Aided Civil and Infrastructure Engineering*, 695-710.

CIFAR-100, C.-1. a. (2023, March 19). *CIFAR-10 and CIFAR-100 datasets*. Retrieved from CIFAR-10 and CIFAR-100 datasets: https://www.cs.toronto.edu/~kriz/cifar.html

Deng, L., & Wang, Y. (2022). Post-disaster building damage assessment based on improved U-Net. *Sci Rep.*

Fujita, A., Sakurada, K., Imaizumi, T., Ito, R., Hikosaka, S., & Nakamura, R. (2017). *Damage detection from aerial images via convolutional neural networks.* Nagoya, Japan: IEEE.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Garcia-Rodriguez, J., , , . . . , . (2017). A Review on Deep Learning Techniques Applied to Semantic Segmentation. *arXiv*, 1704.06857.

Ghaffarian, S., Kerle, N., Pasolli, E., & Arsanjani, J. J. (2019). Post-Disaster Building Database Updating Using Automated Deep Learning: An Integration of Pre-Disaster OpenStreetMap and Multi-Temporal Satellite Data. *remote sensing*.

Gupta, R. H. (2019). xBD: A Dataset for Assessing Building Damage from Satellite Imagery. *Computer Vision and Pattern Recognition*.

ImageNet. (2023, March 19). *ImageNet*. Retrieved from ImageNet: https://image-net.org/

Koshimura, S., Moya, L., Mas, E., & Bai, Y. (2020). Tsunami Damage Detection with Remote Sensing: A Review. *geosciences*, 177.

Meilano, I., Rahadian, A. I., Suwardhi, D., Suminar, W., Atmaja, F. W., Pratama, C., . . . Haksama, S. (2020). Analysis of Damage to Buildings affected by the Tsunami in the Palu Coastal Area Using Deep Learning. *2020 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS)* (pp. 95-97). Jakarta, Indonesia: IEEE.

Mori, G., & Karoon, R. N. (2917). Building Damage Assessment Using Deep Learning and Ground-Level Image Data. *Computer and Robot Vision (CRV)*, 95-102.

Mulia, I. E., Gusman, A. R., & Satake, K. (2020, September 13). Applying a Deep Learning Algorithm to Tsunami Inundation Database of Megathrust Earthquakes. *Solid Earth Journal of Geophysical Research*.

Pisla, J., Li, H., Lautenbach, S., Herfort, B., & Zipf, A. (2021, February 29). Detecting OpenStreetMap missing buildings by transferring pre-trained deep neural networks. *24th AGILE Conference on Geographic Information Science,*

Shabbir, A., Ali, N., Ahmed, J., Zafar, B., Rasheed, A., Sajid, M., . . . Dar, S. H. (2021). Satellite and Scene Image Classification Based on Transfer Learning and Fine Tuning of ResNet50. *Mathematical Problems in Engineering*, 18.

Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Bouchareb, Y. S., Dauphin, Y., . . . Quinn, J. (2021). Continental-Scale Building Detection from High Resolution Satellite Imagery. *Computer Vision and Pattern Recognition*.

Su, Y., Rong, G., Ma, Y., Chi, J., Liu, X., Zhang, J., & Li, T. (2022). Hazard Assessment of Earthquake Disaster Chains Based on Deep Learning—A Case Study of Mao County, Sichuan Province. *Front. Earth Sci., Sec. Geohazards and Georisks*.

Sublime, J., & Kalinicheva, E. (2019). Automatic Post-Disaster Damage Mapping Using Deep-Learning Techniques for Change Detection: Case Study of the Tohoku Tsunami. *Remote Sens.* .

THE MNIST DATABASE. (2023, March 19). *MNIST handwritten digit database*. Retrieved from THE MNIST DATABASE: http://yann.lecun.com/exdb/mnist/

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3, 9.

Wenrich, S. (2023, February 19). *SpaceNet 2: Building Detection v2*. Retrieved from SpaceNet: https://spacenet.ai/spacenet-buildings-dataset-v2/

Appendix

How to download OSM data

OSM data could be downloaded, using QGIS, a freely available open-source Geographic Information System (GIS) application for working with geospatial data. The OSMDownloader plugin of QGIS was used to download the data. The plugin required the bounding box coordinates of Ohonua and the location where the data should be saved. At first, the file was in OSM file format, an XML file format that contains several features, which were represented in lines, points, and polygons. The buildings are stored as polygons, and they were exported into shapefile formats for further processing and analysis in ArcGIS.

Chapter 3

Operationalizing the models from Chapter 1 for a non-natural disaster dataset: Case Study of Bucha, Ukraine

Abstract

Like natural disasters, wars can cause destruction to urban fabrics and lead to massive casualties and humanitarian needs. For natural disasters, post-disaster building damage assessments have been useful for analyzing post-disaster needs and appropriating assisting resources effectively and efficiently. Similarly, post-disaster damage assessments are useful for analyzing post-disaster needs for man-made disasters. The current prevalence of wars and their consequential humanitarian disasters have made the needs for post-disaster assessments to be essential. These assessments require a quick turnaround to meet the urgent needs of saving lives and appropriating resources. Traditionally, in-situ or field verifications and visual explorations of satellite and aerial images were used for disaster damage assessment. These methods are time intensive. Recently, deep learning approaches are now being explored as an effective means for extracting post-disaster building damage information. Deep learning requires huge training data to accurately classify post-disaster images to extract building damage information. There have been recent efforts to create and make available training and label datasets to the public to teach post-natural-disaster deep learning algorithms. One of such datasets is the xBD dataset, which was released in 2019. Another is the Low Altitude Disaster Imagery (LADI) Dataset, which was released in 2019 as well. LADI is a collection of labeled images by the Civil Aviation Patrol. This study examines the performance of models trained on xBD dataset when used to classify satellite images from a post-war disaster.

Transfer learning has been used for problems where training and label dataset are not available; It uses models trained on datasets from a related source domain on a target domain. For example, ResNet50, trained on ImageNet library, a collection of 10,000 object classes, has formed the bases for several transfer learning models. Even the xBD dataset was released with a base model trained on ResNet50. Transfer learning can be used to classify images for a man-made post-disaster building damage assessment. In this study, wind hazards and multi-hazards models created from xBD dataset (Sodeinde et al, 2023) were used, separately, for classifying post-war images of buildings damaged by Russian invasion of Bucha, Ukraine in March 2022.

Building footprint images are required for the damage classification models. The building footprints were downloaded from Open Street Maps (OSM). For completeness, missing building footprints from OSM were drawn-in. The results from the models were validated against the UNOSAT post disaster dataset. The UNOSAT dataset only contained information about damaged buildings. As a result, only damaged buildings were analyzed. Confusion matrices were created to analyze the performances of the models on the post-disaster satellite images. The results indicate that the wind model is not transferable for classification of Bucha post-war images; it was only able to detect half of the minor damages, but was unable to detect the major and destroyed classes. The wind model classified 58 out of 117 damaged buildings as “no damage”. The multi-hazard model was able to detect more of the minor and major damage classes, but was unable to detect any of the destroyed classes. The multi-hazard model classified 72 out of the 117 damaged buildings as “no damage”. All the OSM buildings in the study area that were not damaged according to UNOSAT were used as “no damage” buildings. Both models overfit to the “no damage” buildings, probably because the “no damage” buildings were overrepresented in their training data. Of the 2,329 “no damage” buildings, the multi-hazard model identified 1,689, while the wind model identified 1,886.

Introduction

Man-made disasters such as wars can have similar impacts as natural disasters on built-environments; they can cause building damages, casualties, and humanitarian needs. The efficiency of damage assessment efforts immediately following these disasters is crucial for humanitarian response and disaster recovery. The human cost of war could be very significant. Between 2003 and 2023 alone, the estimated deaths due to the United States war in Iraq was between 250,000 and 300,000. The United States war in Syria against ISIS also cost between 250,000 and 300,00 deaths (Crawford, 2023). About 11,000 people were killed in the 2023 West Dafur, Sudan conflict, where satellite images showed that multiple villages were burnt down (Formanek et al., 2023). According to the Uppsala Conflict Data Program, about 82,637 people were killed in the 2022 Russian invasion of Ukraine. During a war, both artillery and aerial bombardments can lead to destruction of buildings in a city. However, aerial bombardments of cities, especially, have become a major cause of casualties during wars, affecting civilians. In addition, they have caused destruction to buildings and urban fabrics such as roads, hospitals and schools. The advent of drone usage on the battlefield have also increased the amount of destruction caused to cities during warfares. The Russian invasion of Ukraine has led to the first large-scale war where both sides have extensively deployed both commercial and military drones. In addition to surveillance, these drones were also used for dropping bombs on cities' infrastructures and civilian buildings. For example, on the last night of 2022, Russia attacked Ukrainian Energy Facilities and apartment buildings in the Kyiv Oblast province with dozens of loitering munitions (Kunertova, 2023). The early days of the Russian-Ukraine war also witnessed attacks on health institutions, schools, humanitarian (evacuation) corridors, churches, civilian shelters, residential buildings, and energy infrastructures (Tin et al., 2023). Like natural disasters, post-disaster damage assessments are crucial for saving lives and aiding humanitarian assistance following these war attacks. High resolution remote sensing images have been very useful for identifying and classifying damaged buildings. Recently, deep learning approaches have been explored for accelerating the process of post-disaster damage assessment after a natural disaster (Gupta, 2019).

High resolution satellite images are efficient for assessing buildings damaged due to war. Many postwar building damage assessments are performed through visual exploration of satellite images. For example, based on satellite images collected on 31st of March, 2022, UNOSAT analysts performed postwar damage assessment of the city of Irpin, Ukraine by exploring high resolution satellite images for visible damages to buildings (UNOSAT, 2022). The analysts found out that 1,060 buildings sustained damage which were visible on the satellite images. During the Syrian war in 2014, UNOSAT performed assessment of damaged buildings in Syrian cities of Aleppo, Ar-Raqqa, Dar'a, Deir ez-Zur, Hama, Homs, and Idlib. In these cites, they identified a total of 10,126 destroyed, 12,170 severely damaged, and 9,462 moderately damaged buildings. These damage assessments were performed through visual observation of high-resolution satellite images. A building whose all or most of the structure has collapsed was classified as destroyed (i.e. 75% - 100% structurally destroyed). When a significant part of a building structure has collapsed, the building is classified as severely damaged (i.e. 30% - 75% structurally damaged). A moderately damaged building is a building with limited observable damage (i.e. 5% - 30% structurally destroyed) (UNITAR, 2015). To assess building damage due to the 50-day long conflict that took place in Gaza strip over the summer of 2014, UNOSAT used very high-resolution commercial satellite images to map damaged infrastructure such as residential and industrial buildings, health and educational facilities, roads, and agricultural areas. This work, which was performed by United Nations (UN) workers and one UNOSAT expert who was deployed to Gaza for in-situ support, identified 6,761 destroyed, 3,565 severely damaged and 4,938 moderately damaged structures (UNITAR, 2014). Visual exploration of satellite images could be slow and inaccurate. For this reason, automated approaches have been explored to save time and improve accuracy.

Some automated approaches use optical images, while others use Synthetic Aperture Radar (SAR) or combination of optical images and SAR to detect damages to buildings after a war. In 2018, Ghandour and Jezzini performed building damage estimation using shadow information and Gray Level Co-occurrence Matrix features from optical images over a Syrian war-affected zone near Damascus. Their approach set a threshold on the difference

between pre- and post-war building damage to detect a damaged building. They also used Gray Level Co-occurrence Matrix (GLCM), a popular statistical method of extracting textural features from images to determine post-war damaged buildings. Two features were used: variance was used to measure the image homogeneity, while correlation was used to measure linear dependency of gray levels on neighboring pixels. Buildings without shadows in the post-war images were classified as damaged buildings. Their results showed acceptable building damage estimation results. In 2021, Boloorani et al used Sentinel SAR and Interferometric SAR to estimate post-war damage to Mosul City in Iraq. The results were validated using the UNOSAT maps, and findings showed that 40% of the entire city was affected by the Islamic State of Iraq and Syria (ISIS) war. Their method found 4,773 were destroyed, 8,233 severely damaged, and 6,882 moderately damaged buildings in Mosul. In 2022, Aimaiti et al used Sentinel-1 Radar and Sentinel-2 Optical Images to assess post-war building damage in Kyiv, Ukraine. Their method uses SAR log ratio of intensity for the Sentinel-1 and texture analysis of Sentinel 2 data. They used masks for OpenStreetMap building footprints and World Settlement Footprint (WSF) to suppress unwanted changes that were not related to the urban areas. When compared to UNOSAT damage assessment of the same area, their study assessment showed that 58% of the damaged buildings were correctly classified. Machine learning techniques, specifically deep learning, are now being explored to improve the turnaround time and the accuracy of post-war building damage assessment.

Deep learning models are useful for automated building damage detection and monitoring of war-related building damage. In 2021, Mueller et al demonstrated the use of Convolution Neural Network (CNN), a deep learning algorithm, to detect buildings destroyed by artillery and bombing in satellite images. They were able to demonstrate that the persistent nature of destroyed buildings can improve the training of models for automated monitoring of damaged buildings during wars. They also proposed additional machine learning stages that use the images of surrounding areas and multiple successive images of the same area to improve the accuracy of the models substantially. Studies by Zhang et al in 2023 incorporated a temporal knowledge-guided detection scheme (TKDS) for satellite urban destruction monitoring. They further incorporated a pixel-based transformer model into TKDS to achieve the detection performance in Syrian civil war and Russian-Ukraine war with medium or high resolution satellite imagery. Their model outperformed the state-of-the-art methods by 44.4 in terms of F1 score in six Syrian cities and four Ukraine cities. The scheme detected 1,652 damaged buildings in Mariupol, Ukraine and it ignored monitoring moderate and severely damaged buildings because their destruction appearances may not be visible in satellite images.

Deep learning requires a huge number of label datasets to train a successful model. These label datasets may not be readily available. For this reason, transfer learning, a technique in deep learning, could be used to transfer knowledge from a model trained on a related source domain to a target domain. If the accuracy of the classification is poor (i.e., the model is not transferable), the pre-trained model can be fine-tuned. Fine-tuning is the process of removing the final layers of a pre-trained model and then replacing them with new ones (Shabbir, et al., 2021). The new layers of the pre-trained model are then trained on the new dataset.

Many recent models have used transfer learning techniques for building successful models. For example, ResNet50, a very successful deep learning architecture, is used in many computer vision applications as a base architecture for training new models. xView2, a challenge to detect and classify buildings damaged by natural disaster, released a baseline architecture containing ResNet50 along with a huge labeled dataset (xBD) for building detection and damage classification algorithms. xBD contains data across 7 different disaster types and 19 different disaster events (Gupta, 2019). It has four damage levels: no-damage, minor-damage, major-damage, and destroyed. The no-damage class represents buildings that are undisturbed, with no signs of water, structural or shingle damage, or burn marks. Minor damages are for buildings that are either partly burnt, with water surrounding their structure, volcanic flow nearby, roof elements missing, or with visible cracks. Major damage are buildings with partial wall or roof collapse, or buildings with encroaching volcanic flow or surrounded by mud or water. Destroyed buildings are completely collapsed, scorched, partially or completely covered with water or mud, or buildings that are no longer present after the event (Gupta, 2019). xBD dataset is highly imbalanced since no-damage class is overrepresented in the dataset. Imbalance in a dataset can lead to a poor performing model, where the model overfits to the overrepresented class. In 2023, Sodeinde et al looked at improving the xBD dataset and its base algorithm and made recommendations on using either a multi- or an individual hazard model for a post-disaster damage identification and

classification. For each of the disaster types, the study also made recommendations on whether to use a binary or a multiclass model. Since the models overfitted to the no-damage class, they experimented with binary classification, which trained models on the two new classes: the original "no-damage" class, and a second merged class ("damaged") to represent "minor damage," "major damage," and "destroyed." This study used models trained on samples from natural hazards domain (Sodeinde et al, 2023) to detect and classify man-made damages due to war in Bucha, Ukraine.

On 24th of February 2022, Russia invaded Ukraine. As part of the offensive on Kyiv, Ukraine's capital city, the battle of Bucha started on February 27th and ended on March 31st. In addition to the casualties, the war in Bucha left several destructions in its wake. Bucha is a city of about 37,000 people and is about 25 km northwest of Kyiv (Figures 1 and 2). According to the city data on destruction, 1,354 buildings were destroyed (Ukraine Gate, 2022).

Figure 1. Kyiv and the surrounding City

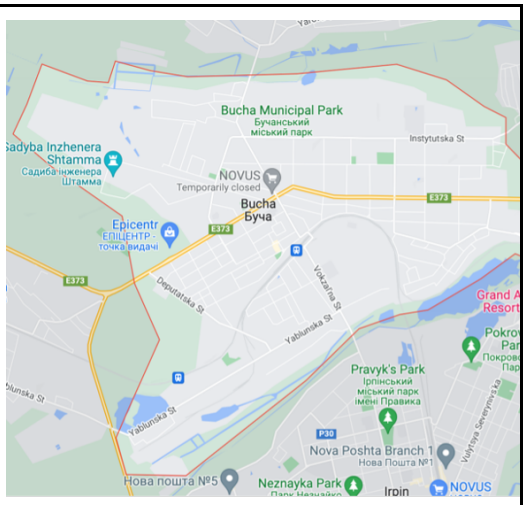


Figure 2. City of Bucha

In this study, wind hazards and multi-hazard models from Sodeinde et al (2023) were used, separately, to classify buildings damaged by the Russian bombardment of Bucha, Ukraine. The identifiability of damaged buildings in satellite images were also evaluated. Applying the models to Bucha was useful for studying the performances of the models created from xBD dataset, on samples from a manmade disasters event.

The post-disaster images of Bucha were downloaded from Maxar Technologies website as DigitalGlobe WorldView2, orthorectified 3 bands, 8 bits per pixel, radiometrically-corrected, natural images. The images were pan-sharpened to improve their resolutions and their bounding boxes were clipped, using OpenStreetMaps building footprints. Missing building footprints were added. Each clipped image was reshaped into a 128 x 128 and then passed into the models from Sodeinde et al. (2023) for classification. Validation data from the Unitar website in terms of damaged buildings were used. The confusion matrices from the classifications were used to analyze the identifiability of the damaged buildings.

Data

The town of Bucha is located between latitudes 50.61 N and 50.48 S and between longitudes 30.14 W and 30.28 E. Bucha has a population of about 37,321 people. It is located in Kyiv Oblast (i.e. administrative region), and is 25 km west of the city of Kyiv, Ukraine. It also shares borders with the cities of Irpin and Hostomel and the villages of Vorzel, Mykhailivka-Rubezhivka, and Blystavytsia. Bucha's main landmark is a 19th-century railway station located at the south edge of the city. Other important landmarks in Bucha include the Yuvileiny Stadium; Bucha Glassworks, which is a glass factory; a small train stop called Sklozavodska and the major highway that runs through the city, called M07.

To identify the damage to Bucha, pre- and post-war satellite images were downloaded from Digital Globe under the NEXTVIEW agreement between Digital Globe and National Geospatial Intelligence Agency (NGA). The pre-war images were taken on the 28th of February 2022, while the post-war images were taken on 14th of March 2022. Both pre- and post-war images were taken, using a WorldView-2 commercial satellite sensor. WorldView-2 operates at an altitude of 770 km. The pre-war images were orthorectified, 8-bands, GeoTIFFs as listed in Table 1. They were also radiometrically corrected, with pixel depth of 8-bit. The image multispectral 8-bands in VNIR (Visible Near Infrared) include C (Coastal Blue), B (Blue), G (Green), Y (Yellow), R (Red), RE (Red Edge), N (Near IR1), and N2 (Near IR2). The post-war images were orthorectified, 3-bands, GeoTIFFs as listed in Table 1. They were also radiometrically corrected, with pixel depth of 8-bit. The image 3-bands natural colors include B (Blue), G (Green), and R (Red).

Table 1. Pre- and Post-images

Filename	Date	Type	Spatial Resolution (m)
22MAR14085032-S2AS_R3C3-015097301010_01_P001.TIF	03/14/2022	Post-war	0.46
22MAR14085032-S2AS_R3C2-015097301010_01_P001.TIF	03/14/2022	Post-war	0.46
Multispectral_22FEB28090606-M3DS-015042718010_01_P001.TIF	02/28/2022	Pre-war	1.84

The images are projected to Universal Transverse Mercator (UTM) Zone 36N, in meters as shown in Figures 3a and 3b.

Figure 3a.

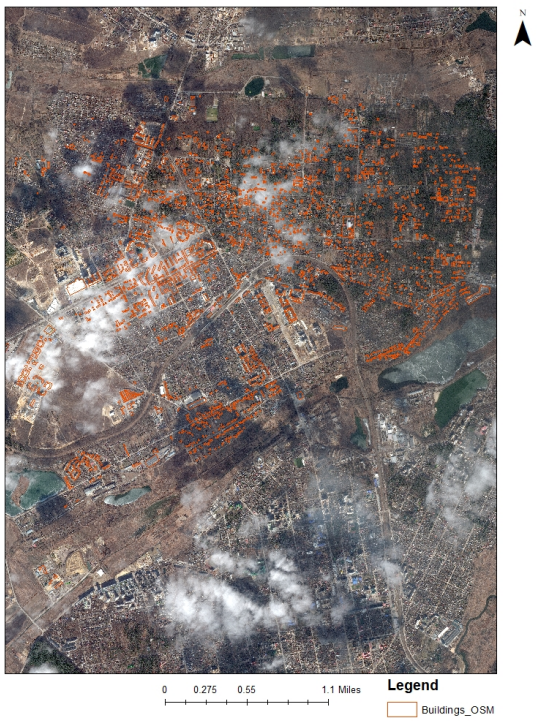


Figure 3b.

The Study Area

The study area is Bucha, one of the 12 communities under the Kyiv region. It is the administrative center of the Bucha territorial community and the district center of the Buchansk district of the Kyiv region. According to Yukhnovskiy et al. (2019), the landscape public areas of Bucha are represented by parks, squares, boulevards and urban forests of 99.6 hectares. Of its 2,658 hectares land area, 305 hectares are residential apartment buildings; 567 hectares are homestead buildings; 88 hectares are roads, streets and squares; 71 hectares are industrial buildings; 40 hectares is water surfaces; 476 hectares is agricultural area; and 42 is forest. The centroid of the city is at 50° 32 '33.7"N, 30° 12' 34.2"E.

Open Street Maps (OSM)

The building footprints, which capture the pre-war state of Bucha, Ukraine, were downloaded from OSM (Figure 4a and 4b). OSM is a free and publicly available geographic database. Included in the database are features such as transport, boundaries, places, and manmade structures (e.g., buildings). These features are crowdsourced from multiple people across the world from sources such as surveys, aerial images and other freely available geodatabases. OSM buildings were downloaded on March 20, 2022. The study area contained 2,446 buildings. The data was in shape file format and its coordinate system was Geographic Coordinate Systems (GCS) WGS 1984. The bounding box of the data is: Top: 50.568987, Left: 30.163407, Right: - 30.260329, and Bottom: 50.519776.

The OSM building layer comes from multiple sources and people all over the world. For this reason, it contains errors due to multiple data sources. Some of these errors lead to misalignments and shifts in building footprints. Other errors include error of omission, when buildings that should have been included in the dataset are missing.t. Similarly, there were errors due to commission; in these cases, building footprints were incorrectly recorded, when they were actually missing. There were also errors from when the shapes and sizes of buildings were incorrectly represented. People are encouraged to submit the most recent information to correct the wrong data. They are also encouraged to make the edits to fix the data.. In this study, misaligned building footprints were fixed by realigning them with the building footprint on the pre-disaster images. Errors due to commission were not corrected. However, errors due to omission were corrected by adding building footprints that were missing in the dataset.

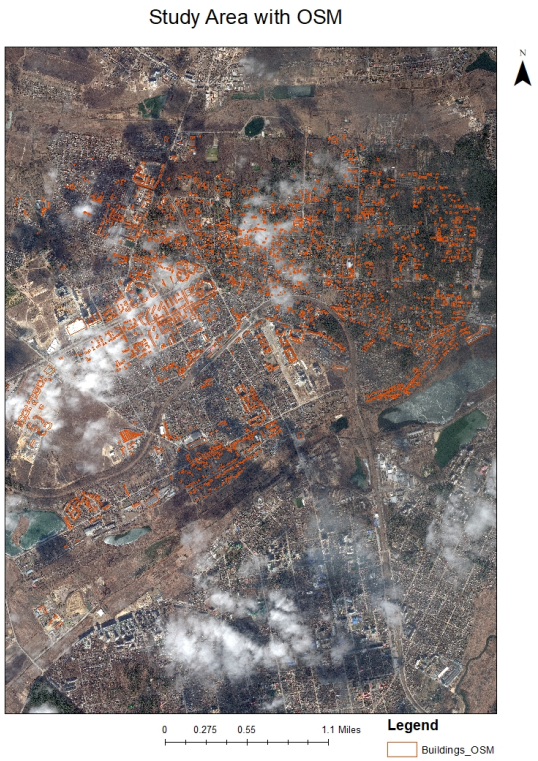


Figure 4a.

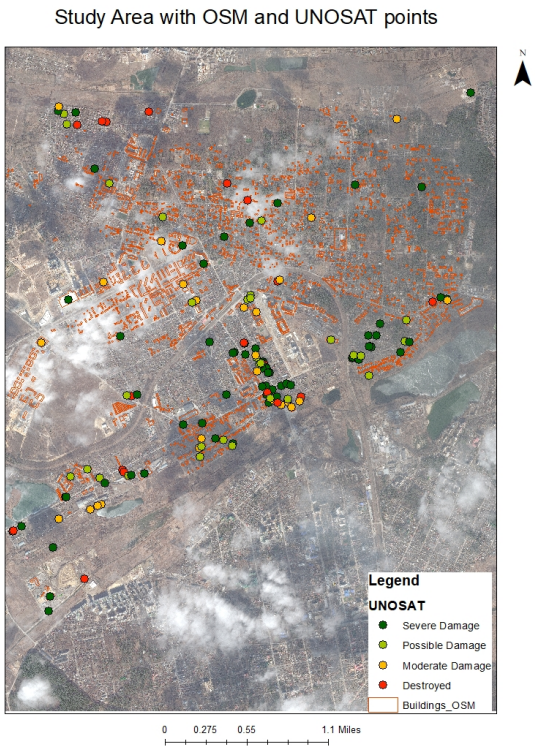


Figure 4b.

UNOSAT Validation Data

Analysis performed by UNOSAT as of 31st of March 2022, using marginally degraded images affected by clouds and other limiting factors, showed extensive damage to the city of Bucha. These WorldView-3 images, which were collected on 31st of March and 20th of February 2022, showed that 147 structures sustained damage. Out of these, 19 were destroyed, 72 severely damaged, 26 moderately damaged and 29 possibly damaged (Table 2). This result was from a preliminary assessment by UNOSAT and maps were created and made available in both shapefile and geodatabase formats, whose metadata contains attributes such as sensorDate, Settlement, and Notes. According to UNOSAT, these were preliminary observations that have not yet been verified by ground feedback. This data was used as the validation dataset for this study.

Table 2. Count of the Notes

Notes	Count
Destroyed	19
Severely damage	72
Moderate damage	26
Possible damaged	29

"No-damage" Buildings

There are 2,329 OSM buildings in the study area that were not classified as damaged in the UNOSAT validation data. All of these buildings were assumed to be in the no-damage class for this study.

Data Preparation

The models required a 3-band clipped image, which encompasses the building whose damage level will be determined. These 3 bands are expected to be the Red, Green, and Blue bands (i.e., RGB). Luckily, the post-war images were already 3-bands (Red, Green, and Blue) so they can be clipped directly. For clipping, the OSM building footprints, which were in World Geodetic System (WGS) 84, were projected to match the coordinate system of the images.

The building footprints were edited and corrected by shifting them horizontally and vertically to align with the post-war image. Missing buildings were added for UNOSAT points that do not have corresponding buildings. After the building footprints have been corrected, their bounding boxes were calculated, extended by 80% offset, and then used to clip the post-war images. Each clipped image contains a building whose damage level was to be determined. These images were stored as geotiff file formats in a personal geodatabase and later exported into PNG file format, making sure that the pixel depth was 8-bit unsigned.

Method

The clipped images of the post-war buildings were loaded into pre-made models recommended for classification by Sodeinde et al (2023). Multiclass models trained, separately, on xBD multi-hazard and wind hazards were used for the classification of the clipped images in accordance with recommendations from Sodeinde et al (2023) (Figure 5). In both cases, the UNITAR dataset was used for the validation of the results.

Sodeinde et al (2023) analyzed the quality of xBD training dataset for identifying building damage across a variety of natural hazards using deep learning convolutional neural networks. Specifically, the study evaluated the pros and cons of combining training datasets across multiple natural hazards and provided recommendations on using the provided training datasets to optimize classification accuracy for building damage detection.

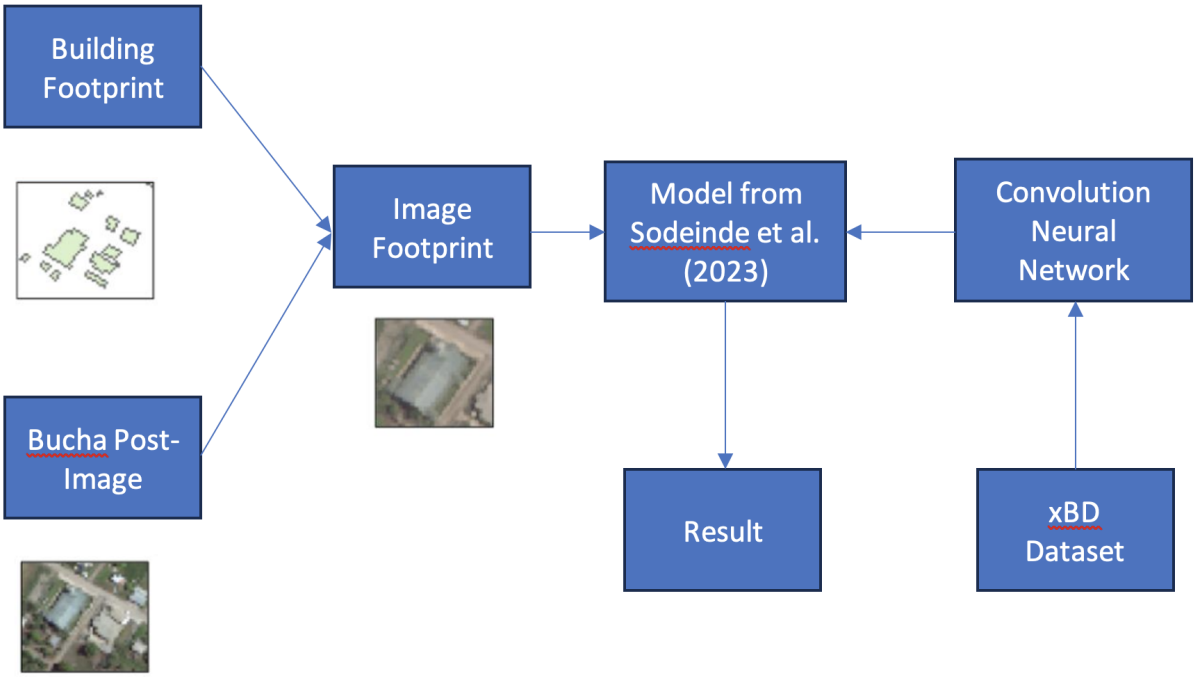


Figure 5. Approach for classifying and verifying Bucha post-war dataset.

The Algorithms

The models from Sodeinde et al (2023) were developed from modified versions of xView2 base algorithm and xBD dataset. The base algorithm was a CNN algorithm released along with xView2 challenge as a starting point for developing solutions for xView2 challenge. The base algorithm was developed using ResNet50 architecture with ImageNet weights, fine-tuned with three sets of convolution and MaxPooling layers (Figure 4). Since this base algorithm overfitted to the "no-damage" class of the xBD dataset, modifications were made in Sodeinde et al (2023) to improve its performance.

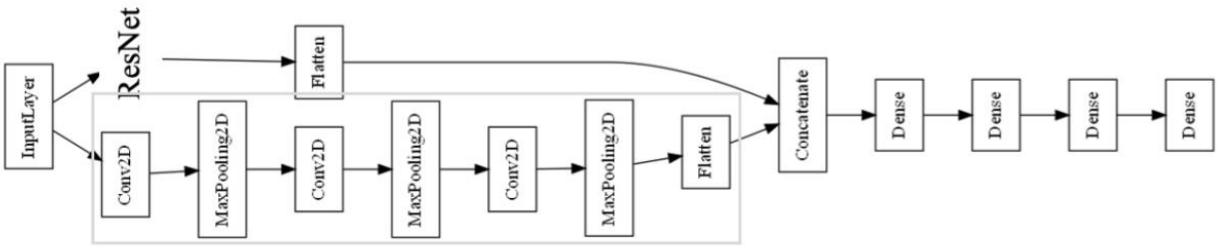


Figure 4. xView2 base algorithm

Modifications for Multiclass Classification

To create a four-class model ("no-damage", "minor-damage", "major-damage", and "destroyed"), the xView base algorithm was further tuned by adding two sets of convolution and MaxPooling layers interleaved with five BatchNormalization layers; one for each of the five blocks of convolution and MaxPooling layers. A Dropout layer with a rate of 0.5 was also added on top to reduce overfitting. This new modification performed better at identifying the "minor-damage", "major-damage", and "destroyed" class, in addition to the "no-damage" class. Longer epochs and early stopping were also added to improve the performance of the models from this algorithm.

Multi-hazard Model

The multi-hazard model was trained on earthquake, fire, flooding, tsunami, volcano, and wind hazards. It contained about 316,000 buildings. The wind hazard represents about 38% of the dataset. The "no damage" class has the highest number of buildings; it represents about 78% of the dataset. This created an overfitting challenge for the model. Volcano is the least prominent, representing only about 1% of the dataset, while wind is the most prominent, representing about 38% of the dataset (Table 3). In this study, the wind and multi-hazard multiclass models from Sodeinde et. al. (2023) were used, separately, to classify the post-war images of Bucha.

Table 3. Distribution of xBD dataset for multi-hazard model

	destroyed	major	minor	no-damage	Total
earthquake	2	18	110	32,141	32,271
fire	7,995	635	698	50,576	59,904
flooding	577	4,840	5,283	41,321	52,021
tsunami	5,145	671	1	42,524	48,341
volcano	529	35	58	3,644	42,66
wind	9,314	15,250	19,664	75,173	119,401
Total	23,562	21,449	25,814	245,379	316204

Wind Hazards Model

Unlike the multi-hazards model, which was trained on all the hazards in the xBD dataset, the wind hazard model was trained only on the tornado and hurricane hazards because the building damage for both types of natural disaster are controlled by wind. In this study, the wind multiclass model from Sodeinde et al. (2023) was also used to classify the post-war images of Bucha's because the destruction caused by the war appears similar to the destruction caused by wind hazards when viewed on a satellite image. There are about 120,000 buildings in the wind hazard dataset (Table 3). The wind hazard dataset includes 75,173 "no damage", 19,664 "minor damage", 15,250 "major damage", and 9,314 destroyed buildings.

Predictions

Without further fine-tuning and additional training, the multiclass models trained on multi-hazard and wind-only xBD datasets from Sodeinde et al (2023) were used for classification in this study. A building's damage class was predicted by loading a model into the computer memory using the TensorFlow libraries; once in the memory, the model was compiled using an Adam optimizer with an ordinal loss function. Clipped images of the buildings, which were resized into 128x128, converted to an array and normalized over 255 served as inputs into the models. The models' outputs contain probabilities of a building belonging to a damage class. The class with the highest probability represents the predicted class for the building.

Results and Analysis

The classification results for all the clipped images of the buildings were saved in a CSV (Comma Separated Values) file. Each row in the CSV file contains a row number, Feature Identification (FID), and the predicted class. The FID is a unique identification number for a building; it maps a building to its ObjectID in the OSM shapefile (Table 4). The UNOSAT dataset was used to validate the results of the classification.

Table 4. CSV file containing buildings FID and Classes

	FID	class
0	0	3
1	1	0
2	10	3
3	100	1
4	101	0
5	102	1
6	103	2
7	104	2
8	105	1
9	106	2
10	107	2

UNOSAT Dataset

The UNOSAT dataset came in four classes (destroyed, severely damaged, moderate damage, possible damage) (Figures 6a and 6b). These classes are different from the xBD damage classes (no-damage, minor, major, and destroyed), making it difficult to directly validate the results of the models. For this reason, the classes in the UNOSAT dataset were mapped to xBD classes: "moderate damage" were mapped to "minor damage" and "severely damaged" were mapped to "major damage." The destroyed class in UNOSAT was mapped to the destroyed class in

xBD. The “possible damage” class was not used because its damage levels were not certain. The “no damage” class was not used because they were not contained in the UNOSAT data.

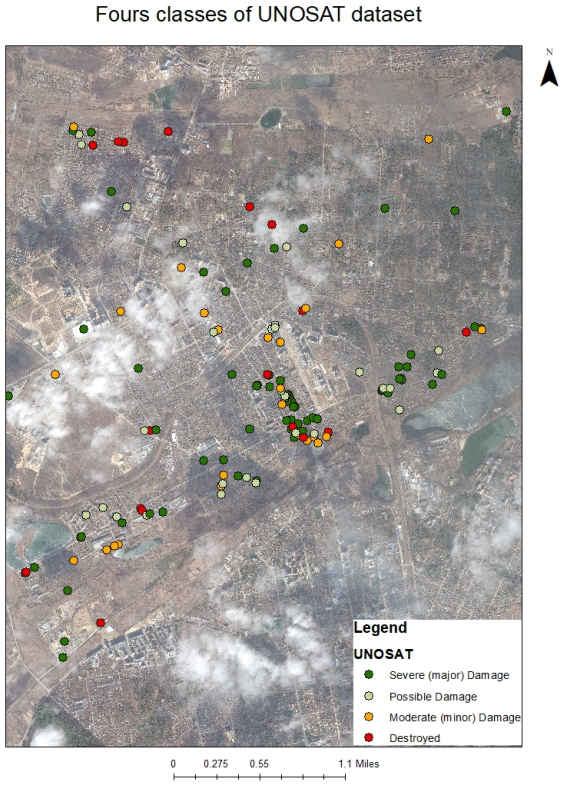


Figure 6a.

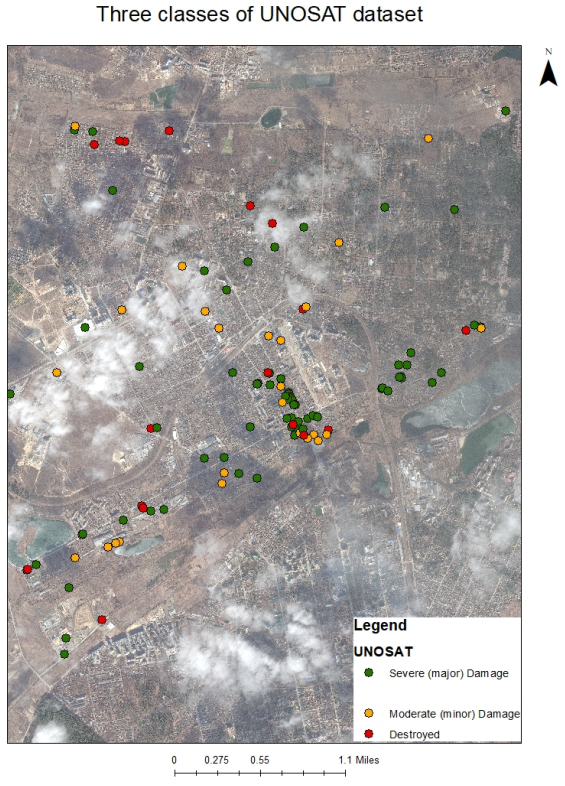


Figure 6b.

Wind Hazards Multiclass Classification

The multiclass model trained on a wind hazards xBD dataset outputs four values, one for each of the xBD classes. The highest value represents the damage level of the building. The classification accuracy was poor. The model classified 58 (12 minor, 37 major, and 9 destroyed) of the 117 UNOSAT data as “no damage,” when 0 “no damage” was expected. It also classified 59 (14 minor, 34 major, and 10 destroyed) buildings as “minor damage,” when there were only 26 “minor damage” buildings in the UNOSAT data. Only 1 (major damage) building was identified as destroyed, when there were actually 19 destroyed buildings in the UNOSAT data. None of the 72 “major damage” buildings were detected (Tables 5a and 5b). Figures 7a and 7b show the result of the classification overlaid on UNOSAT data. Of the 2,329 “no damage” buildings from the OSM layer, 1,886 were classified as “no damage”, 440 were classified as “minor damage”, none was classified as “major damage”, and 3 were classified as “destroyed”. Figure 7c shows the result from the classification of the “no damage” buildings.

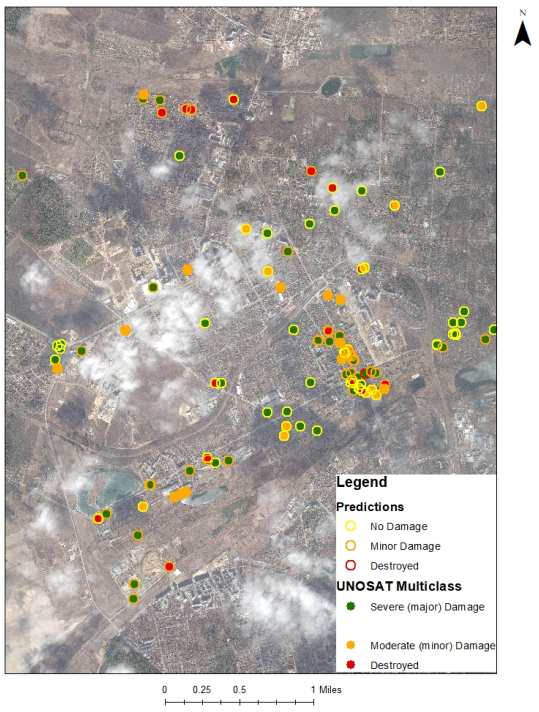


Figure 7a.

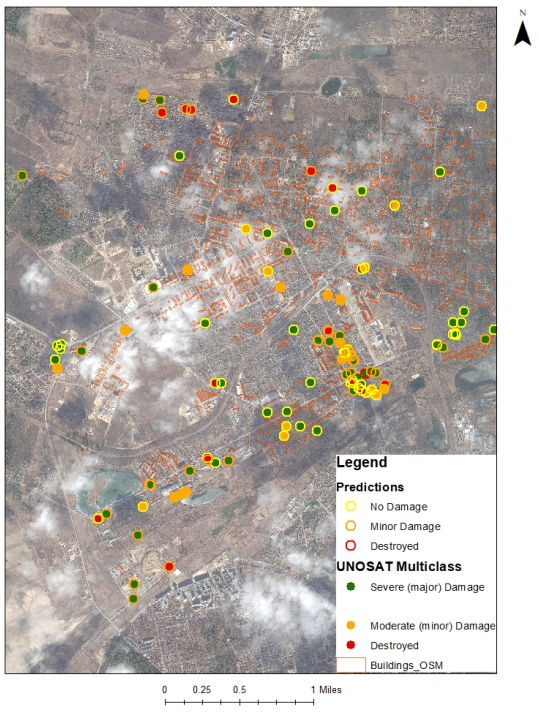


Figure 7b.

Table 5a. Expected vs. Actual counts of damages (wind hazards)

	Expected	Actual
no damage	2,329	1,944
minor damage	26	499
major damage	72	0
destroyed	19	4

Table 5b. Confusion Matrix of wind hazards

	no damage	minor damage	major damage	destroyed
no damage	1,886	12	37	9
minor damage	440	14	34	10
major damage	0	0	0	0
destroyed	3	0	1	0

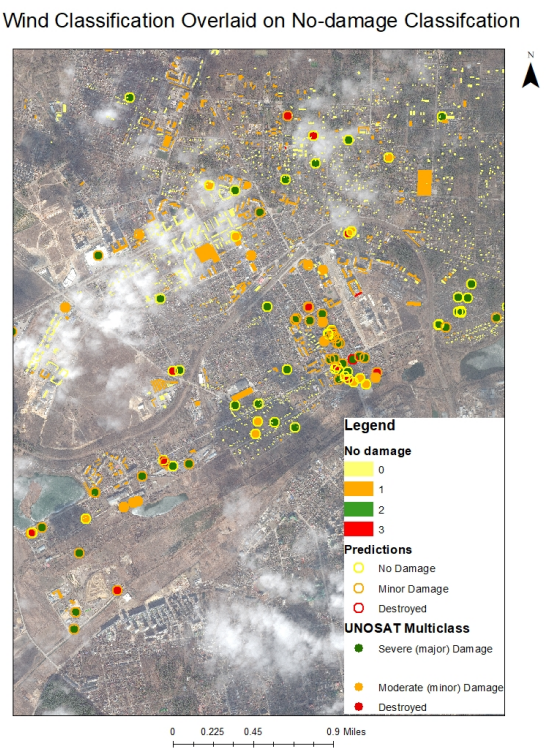


Figure 7c.

Multi-hazard Multiclass Classification

The multiclass model trained on a multi-hazard xBD dataset outputs four values, one for each of the classes. The highest value represents the damage level of the building. The multiclass model performed better. Although it was unable to identify any of the destroyed buildings, it predicted more minor and major damaged buildings. This improvement in performance was probably because the model was able to pick up nuances in the damage levels from the images (Figure 8a).

The model classified 72 buildings (17 minor, 43 major, and 13 destroyed) of the 117 UNOSAT data as “no damage”, when there were 0 “no damage” in the validation data. It also classified 29 (7 minor, 18 major, and 4 destroyed) buildings as “minor damage,”when there were only 26 “minor damage” buildings in the UNOSAT data. The model identified 17 (2 minor, 13 major, and 2 destroyed) of the 117 UNOSAT data as “major damage,” when there were actually 72 buildings with major damage according to UNOSAT. None of the 19 “destroyed” buildings were detected (Tables 6a and 6b). Figures 8a and 8b show the results overlaid on UNOSAT data.

Of the 2,329 “no damage” buildings from the OSM layer, 1,689 buildings were classified as “no damage”, 545 buildings were classified as “minor damage”, 75 buildings were classified as “major damage”, and 20 were classified as “destroyed”. Figure 8c shows the result of the classification of buildings with no damages.

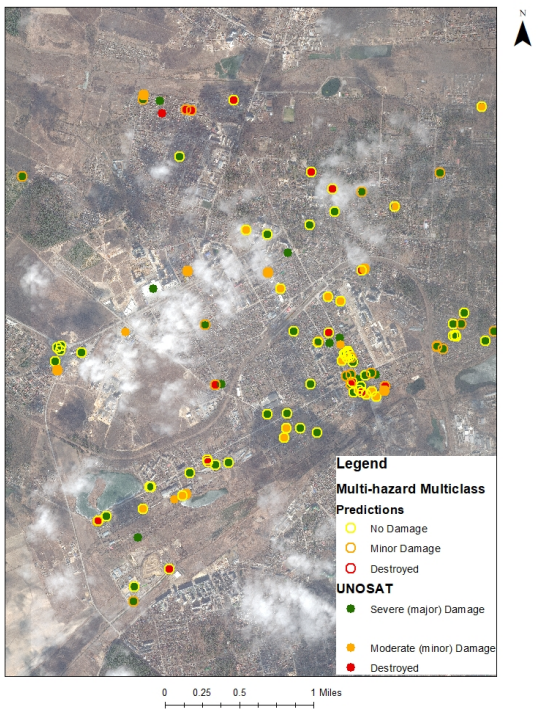


Figure 8a.

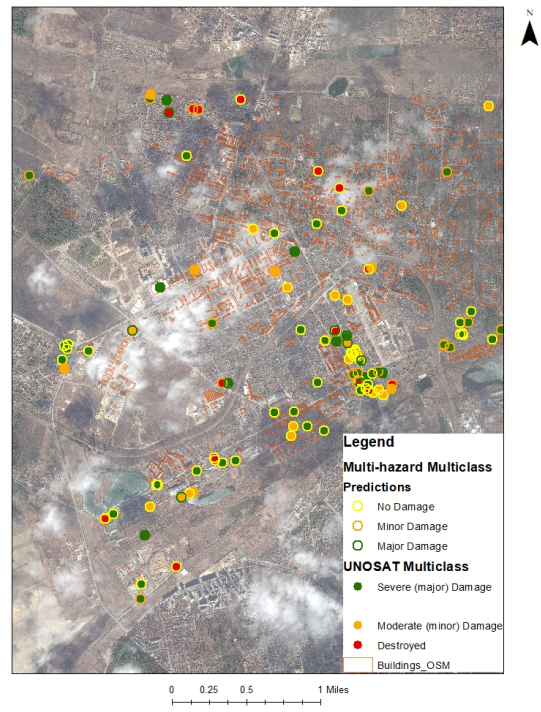


Figure 8b.

Table 6a. Expected vs. Actual counts of

	Expected	Actual
no damage	2,329	1,761
minor damage	26	574
major damage	72	92
destroyed	19	20

Table 6b. Confusion Matrix of multi-hazard damages (multi-hazard)

	no damage	minor damage	major damage	destroyed
no damage	1,689	17	43	13
minor damage	545	7	18	4
major damage	75	2	13	2
destroyed	20	0	0	0

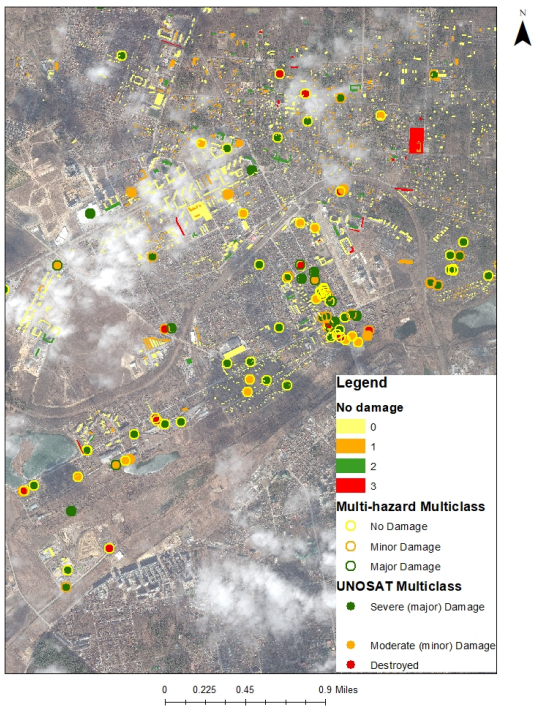


Figure 8c.

Discussion

Both models overfit to the "no damage" class. This was probably due to the overrepresentation of the "no damage" class in the xBD training dataset. The wind model was trained on 63% "no damage" data, and it predicted the "no damage" class with an accuracy of about 80%. The multi-hazard model was trained on about 78% "no damage" training data and it predicted the "no damage" class with an accuracy of about 71%. The wind model performed better in spite of its lower percentage of "no damage" training data probably because the "no damage" buildings are easily differentiable on a wind hazard satellite image.

The wind model also identified more minor damages than the multi-hazard model. It identified 14 out of the 26 UNOSAT "minor damage." The multi-hazard model identified only 7 "minor damage" buildings. For major damages, the multi-hazard model performed better. It identified 13 out of the 72 UNOSAT "major damage" buildings. The wind model was unable to identify any of the major damages. Therefore, the wind model might be better for identifying minor damages, while the multi-hazard model might be better for identifying major damages. However, both models were unable to identify destroyed buildings.

Conclusions

The results indicate that the wind model is not transferable for classification of Bucha post-war images; it was only able to detect half of the minor damages, but was unable to detect the major and destroyed classes. The multi-hazard model was able to detect more of the minor and major damage classes, but was unable to detect any of the destroyed classes. The inability of the models to detect "destroyed" buildings might be due to the fact that the characteristics of a building destroyed by a war may be different from those destroyed by a natural disaster. A multi-hazard model is more transferable to problems with spatial patterns, roof appearances, and/or lateral patterns of damaged buildings similar to those in the images used to train it. The spatial pattern of buildings damaged by wars may be different from those damaged by natural disasters. Buildings destroyed by a natural disaster usually have homogenous patterns, where damaged buildings are more spread across due to severe weather conditions such as wind. This is quite different from buildings damaged by wars, where the damages are heterogeneous, and the damaged buildings may be more spread sporadically. The appearance of a roof damaged by a war may also be different from the one damaged by a natural disaster. For a war, rather than the roof blown away, the damage to the roof may be contained; the roof may sustain holes due to an aerial bombardment. However, for a natural disaster, the roof may be blown away as a result of wind or flood. The pattern of lateral damages to buildings due to wars may be different from those due to natural disasters. This may be one of the reasons why the multi-hazard model was not transferable. For example, combats and artillery shelling of buildings may create direct impacts on walls, especially those facing the streets, if the war zone is an urban area. These direct impacts may cause the sides of the buildings facing the street to collapse, creating patterns of collapsed walls along the street.

This poor performance may be improved by fine-tuning the multi-hazard model and then training it with labeled images of buildings that were damaged by war. Ancillary data such as layers differentiating military installations from civilian structures could also be added to the algorithm to enhance its classification accuracy. For the best performance, the recommendation is that higher resolution images of buildings damaged by wars should be used to train a model for classifying buildings in a war zone. The lesson here is that the nature of a disaster that caused damage to a building determines the spatial and lateral patterns, as well as the type of damage to a building's roof. The multi-hazard model was more transferable than the wind hazards model probably because it contains more samples of the minor and major damages, leading the model to be able to detect the nuances in these classes. Another reason for the poor tranferability of the models might be the differences between the annotation or labeling of the damage levels by the UNOSAT analyst versus the xBD. In this case, for better transferability, a model that is trained on labeled or annotated post-war buildings should be used for classifying post-war images or the labeling or annotation of the post-war buildings should be made to match those of the natural disaster or the xBD dataset.

References

Aimaiti, Yusupujiang, et al. "War Related Building Damage Assessment in Kyiv, Ukraine, Using Sentinel-1 Radar and Sentinel-2 Optical Images." *remote sensing*, vol. 14, no. 24, 2022, p. 6239.

Boloorani], Ali, et al. ``Post-War Urban Damage Mapping Using InSAR: The Case of Mosul City in Iraq." *International Journal of Geo-Information*, vol. 10, no. 3, 2021, p. 140.

Crawford, Neta C. *Blood and Treasure: United States Budgetary Costs and Human Costs of 20 Years of War in Iraq and Syria, 2003-2023*. WATSON INSTITUTE INTERNATIONAL AND PUBLIC AFFAIRS. 2023. *COST OF WAR*, BROWN UNIVERSITY, https://watson.brown.edu/costsofwar/.

Formanek, Ingrid, et al. "10,000 reported killed in one West Darfur city, as ethnic violence ravages Sudanese region." *CNN*, 26 July 2023, https://www.cnn.com/2023/07/26/africa/sudan-west-darfur-thousands-killed-intl/index.html. Accessed 26 August 2023.

Ghandou, Ali J., and AbdulKarim A. Jezzini. “Post-War Building Damage Detection.” *Proceedings of The 2nd International Electronic Conference on Remote Sensing*, vol. 2, no. 7, 2018, p. 359.
Gupta, Ritwik. “A Dataset for Assessing Building Damage from Satellite Imagery.” *Computer Vision and Pattern Recognition.*, 2019.
Gupta, Ritwik H. “xBD: A Dataset for Assessing Building Damage from Satellite Imagery.” *Computer Vision and Pattern Recognition.*
Kunertova, Dominika. “The war in Ukraine shows the game-changing effect of drones depends on the game.” *Bulletin of the Atomic Scientists*, vol. 79, no. 2, 2023, pp. 95-102.
Mueller, Hannes, et al. “Monitoring war destruction from space using machine learning.” *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES*, vol. 118, no. 23, 2021, p. e2025400118.
Shabbir, Amsa, et al. “Satellite and Scene Image Classification Based on Transfer Learning and Fine Tuning of ResNet50.” *Mathematical Problems in Engineering*, vol. 9, 2021.
Tin, Derrick, et al. “Hybrid warfare and counter-terrorism medicine.” *Eur J Trauma Emerg Surg*, vol. 49, no. 1, 2023, pp. 589–593.
Ukraine Gate. “The damage caused by the Russian invasion of Bucha is estimated at 450 million dollars.” *Ukraine Gate*, 2022, https://www.ukrgate.com/eng/?p=37366. Accessed 04 October 2023.
UNITAR. “Four Years of Human Suffering – the Syria Conflict as Observed through Satellite Imagery.” *UNITAR*, March 2015, https://unitar.org/learning-solutions/publications/four-years-human-suffering-syria-conflict-observed-through-satellite-imagery. Accessed 23 September 2023.
UNITAR. “UNOSAT latest reports show state of the art in satellite analysis.” *UNITAR*, 2014, https://unitar.org/about/news-stories/news/unosat-latest-reports-show-state-art-satellite-analysis. Accessed 24 September 2023.
UNOSAT. “UNOSAT Damage Assessment Overview Map - Ukraine: Irpin, Kyiv Oblast, Imagery Analysis: 31 March 2022 Published: 18 April 2022 V2 - Ukraine.” *ReliefWeb*, 15 September 2022, https://reliefweb.int/map/ukraine/unosat-damage-assessment-overview-map-ukraine-irpin-kyiv-oblast-imagery-analysis-31-march-2022-published-18-april-2022-v2. Accessed 23 September 2023.
Uppsala Conflict Data Program Department of Peace and Conflict Research. “UCDP.” *Uppsala Conflict Data Program*, 2023, https://ucdp.uu.se/country/369. Accessed 26 August 2023.
Zhang, Liqiang, et al. “Satellite monitoring of war urban damage with a temporal knowledge-guided deep learning scheme.” *Research Square*, vol. 10, 2023.

Dissertation Summary

The goal of this work is to examine xBD dataset and evaluate if its models could be operationalized for future disasters. In evaluating the operationalizability of the models, we concluded that OSM is a good replacement for pre-disaster images and that UNOSAT datasets are useful for a model's validation, but may need to be restructured. The evaluation results show that xBD models can be extended beyond natural hazards and used for man-made hazards.

Acknowledgement

I would like to thank my esteemed supervisor, Dr. Laurie Baise, for her invaluable supervision, support, and tutelage during my PhD degree. Additionally, I would like to express gratitude to my committee members: Dr. Magaly Koch, Dr. Babak Moaveni, and Dr. Sumeeta Srinivasan for their treasured support which was influential in shaping my methods and critiquing my results. The Civil and Environmental Engineering Department, Tufts University has been a very welcoming home during my studies. I would also like to thank the National Geospatial Intelligence Agency (NGIA) for their support. Finally, I would like to thank all my family members, friends, and colleagues.