

MSc. Thesis

The impact of spatial resolution and satellite data type characteristics on Automated Damage Assessment using a Convolutional Neural Network

M.A.F. Verschoor



The impact of spatial resolution and satellite data type characteristics on Automated Damage Assessment using a CNN

MSc thesis, Geoscience and Remote Sensing

by

Fleur Verschoor

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday February 17, 2022 at 15:00.

An electronic version of this thesis is available at <http://repository.tudelft.nl/>.

Image Source Coverpage: Nasa - UN Photo/UNICEF/Marco Dormino



Department of Geoscience and Remote Sensing
Faculty of Civil Engineering and Geosciences
Delft University of Technology



510 - initiative of the Netherlands Red Cross
Anna van Saksenlaan 50
2593 HT Den Haag

Student number: 4467159
Project duration: February 8, 2021 – February 17, 2022
Thesis committee: Prof. Dr. S. Lhermitte TU Delft, supervisor
Dr. Ir. P. López-Dekker TU Delft
Dr. R. Lindenbergh TU Delft
Dr. M. van den Homberg 510
Dr. J. Margutti 510

Abstract

Satellite data, such as optical and Synthetic Aperture Radar imagery, can provide information about the location and level of destruction caused by natural hazards. This information is essential to optimise the rescue mission logistics by humanitarian aid organisations and save people in need. Currently, many Automatic Damage Assessment (ADA) methods exist, designed explicitly for one data type with corresponding spatial resolution. However, the weather and satellite coverage conditions can hinder rapid and complete data acquisitions after large events. Therefore, it is essential to identify the limits and capabilities of novel methodologies testing various data availability scenarios and adjusting them to become robust and widely deployable.

In this research, the Convolutional Neural Network *Caladrius of 510* - an organisation of the Red Cross Netherlands is selected to perform experiments. Initially, the model was designed to input high-resolution imagery and based on the Siamese Architecture, including two Inception-V3 modules followed by three connected layers. The multiple experiments are based on single-, dual-, and cross-mode scenarios, representing data characteristics with varying resolutions, satellite sources and observation sensor types. The xBD dataset provides pre- and post-event high-resolution optical imagery of numerous disasters with corresponding validated damage labels of the included buildings. Subsequently, this dataset is replicated in three down-sampled versions and using Sentinel-2 1C and Sentinel-1 GRD data. With the use of the Macro F1-score and the Cohen's Kappa coefficient, the performances are compared and the predictions' reliability is determined in operational situations.

The results indicate that a lower resolution of the input data has a negative effect on the correct classified buildings. A linear relation does not express the loss in performance, as most damage properties are captured between 0.5- and 2.5-meter. Consequently, this implies that the Sentinel 10-meter resolution datasets provide few recognisable features. The Sentinel-2 1C experiment outperforms the Sentinel-1 GRD, which equals the output of a random classifier. However, no final conclusion is drawn between the true prediction rate of the model compared to the input data type; optical and SAR imagery due to the non-optimal experiment circumstances and limited included datasets. Furthermore, the results from the dual-mode mapping showcase the importance of identical data characteristics between train and test datasets. Conversely, with the use of the cross-mode experiments, it is found not essential to match the pre- and post-event resolution imagery. This latter is very promising for the Red Cross and creates flexibility to construct datasets quickly after the disaster has struck.

Preface

After nine months of hard work, this graduation project comes to an end. In this period, I had the opportunity to learn and experience the process of academic research within a humanitarian aid organisation. All phases were accompanied by specific approaches, insights, and struggles. It allowed me to put the Master's degree Geoscience and Remote Sensing in the context of real-world implementations.

Six years ago, I started studying the Bachelor Aerospace Engineering at the TU Delft, where my interest grew for satellites and the corresponding applications. I made the choice to switch to the Geoscience and Civil Engineering faculty to dive deeper into the technology to observe the Earth instead of space. I enjoyed all the knowledge I gained about measuring and forecasting nature processes and dealing with large data structures. For my electives, I selected Machine Learning and Deep Learning courses linked to a growing field of innovations and comprehensive employment at companies and in society. I appreciated all the chances I got to develop myself broadly at the TU Delft, within my study and with extracurricular activities. One of the highlights consisted of the exchange to Hong Kong.

Since I was young, I was determined to work at a charity or aid organisation, to apply my understanding with the result of helping people all over the globe. Everything came together by doing this project, using a convolutional neural network to assess damage after a natural hazard inputting satellite imagery. For this reason, I really want to thank the 510 team from the Netherlands Red Cross. The work, the organisation, accomplishes is an inspiration. A special thanks to Marc van den Homberg for the support of the project. Furthermore, I would like to express my gratitude to Jacopo Margutti, who helped me with his expertise and technical understanding of the *Caladrius* model and datasets. I hope that, in whatever small way, my thesis might contribute to the development and implementation of an accurate damage assessment model to speed up the search for survivors and to improve aid logistics after natural hazards.

Next, a special thanks to Niels Jansen, who did help me enormously by getting familiar with the Linux environment and the capacities of the virtual computer of the TU Delft, called VRLab, to get access to the GPU and run my model on.

Above all, I would like to thank my supervisor, Stef Lhermitte, for guiding me through the graduation process. He helped me innumerable times with challenges surrounding every facet of my thesis, from programming to writing, and was always available for a quick meeting. All the feedback, brainstorming sessions and encouragement supported me to reach the finish.

Overall, it was an exciting journey, and I am curious what the future brings.

*M.A.F. Verschoor
Rotterdam, February 2022*

Contents

1	Introduction	1
1.1	Automatic Damage Assessment	2
1.1.1	Previous Research	2
1.2	Problem Statement	3
1.2.1	Research Objective	3
1.3	Overview of Chapters	4
2	Literature Study	5
2.1	Satellite Imagery Types	5
2.1.1	Optical Imagery	6
2.1.2	Synthetic Aperture Radar	6
2.2	Convolutional Neural Network	7
2.2.1	Feature Learning	8
2.2.2	Classification	10
2.2.3	Training a CNN	10
2.2.4	Regularization Techniques	12
3	Data	13
3.1	High-resolution optical imagery	14
3.1.1	xBD	14
3.1.2	Study Area	15
3.2	Low-resolution optical imagery	17
3.2.1	Down-sampled xBD	17
3.2.2	Sentinel-2	17
3.3	Synthetic Aperture Radar	19
3.3.1	Sentinel-1	20
4	Method	23
4.1	Caladrius Model	23
4.1.1	Extract Buildings	23
4.1.2	Convolutional Neural Network	25
4.2	Data Pre-processing	27
4.2.1	Down-sampling	28
4.2.2	Resampling	28
4.3	Experiments	29
4.3.1	Single-Mode	29
4.3.2	Dual-Mode	30
4.3.3	Cross-Mode	30
4.4	Performance Metrics	31
5	Results	33
5.1	General Performance	33
5.1.1	Effect of Class Imbalance	34
5.1.2	Reasoning of Miss-classifications	35
5.2	Relation between Resolution and Performance	40
5.2.1	Single-Mode	40
5.2.2	Dual-Mode	45
5.2.3	Cross-Mode	46
5.3	Relation between Satellite Imagery Type and Performance	46

6 Discussion	49
6.1 Interpretations	49
6.2 Implications	51
6.3 Limitations	51
6.4 Further Research	52
7 Conclusion	53
A Appendix	55

List of Figures

1.1	Comprehensive disaster management [15]	1
2.1	Sensor types of satellites to observe the Earth or third bodies [62]	5
2.2	Imaging Radar Geometry [26]	6
2.3	Example of a structure of a Convolutional Neural Network [18]	7
2.4	Filter applied to a two-dimensional input, with the corresponding output of the feature map	8
2.5	Example of maximum and average pooling extractions	9
2.6	Relation between the prediction function over time and the real values of the input data	10
2.7	Mathematical model of an artificial neuron, including input data, weights, bias and activation function [46]	11
2.8	The validation and train set error with respect to the iterations in the training phase of the CNN model	12
3.1	Data source overview subdivided into two types: optical and SAR imagery	13
3.2	xBD data-frame: Pre-event imagery (left) and Post-event imagery (right), with the corresponding JSON files (Example: Hurricane Michael 78)	14
3.3	Visualisation of the Joint Damage Scale description of four-level granularity scheme	15
3.4	Study area: Central and North America (left), Indonesia (right)	16
3.5	The effect of down-sampling the xBD input imagery to 2.5, 5.0 and 10.0-meter resolution (Example: Hurricane Michael 43)	17
3.6	Timeline of Sentinel-2 1C and xBD data collection, pre- and post-event	19
3.7	The comparison of an extracted Sentinel-2 1C image with the xBD 0.5 and 10.0-meter resolution visualisation (Example: Hurricane Matthew 03)	19
3.8	Timeline of Sentinel-1 GRD and xBD data collection, pre- and post-event	21
3.9	The comparison of an extracted Sentinel-1 GRC image with Sentinel-2 1C and the xBD 0.5 resolution visualisations (Example: Tsunami Palu 18)	21
4.1	The pipeline of the data flow, building extraction and damage classification	23
4.2	Visualisation of the effect of down-sampling on polygon scale (Example: Tsunami Palu 138)	24
4.3	Architecture of the <i>Caladrius</i> model. The number in the squares on the left side of a block represents the input size of each block, the number on the right side indicates the output size. N refers to the number of damage classes [58].	25
4.4	a) Naive inception module, b) Inception module with dimension reduction [54]	26
4.5	The three blocks, A, B and C of the Inception V3 and the [54]	26
4.6	Training dataset size and composition, before and after re-sampling	29
5.1	The loss and F1-score plots, trained on the xBD dataset, 100 epochs	34
5.2	The confusion matrices of the imbalanced and balanced dataset, trained on the xBD dataset	35
5.3	The loss and F1-score plots, trained on the imbalanced and balanced xBD dataset, 50 epochs	35
5.4	Left: visualisation of image including cloud coverage (Example: Tsunami Sunda Strait 82-84 (labeled: 3, prediction: 3)) Right: visualisation of image including a destroyed village (Example: Tsunami Palu 20-37 (labeled: 0, prediction: 3))	36
5.5	Visualisation of miss-classified polygons due to the acquisition of imagery three years in advance of the disaster (Example: Hurricane Matthew)	36
5.6	Box-plot of the satellite acquisition characteristics per disaster, including the median, standard variation and outliers	37

5.7	The F1-score, Recall and Precision plots per disaster and class	38
5.8	The confusion matrices per disaster dataset, trained on the xBD dataset	39
5.9	Visualisation of <i>Major damage</i> labelled polygons and classified with the <i>No damage</i> class (Example: Left Tsunami Palu 59-47 Right Tsunami Palu 60-225)	39
5.10	F1-score plot with respect to the resolution setting of the xBD dataset	40
5.11	The confusion matrices of multi- and binary-classification, trained on the xBD dataset with varying resolutions	42
5.12	Visualisation of polygons with varying resolution settings to detect when miss-classification occurs	43
5.13	Distribution of True and False predictions versus the building footprint, per resolution dataset [m^2]	44
5.14	Visualisation of polygons representing the visual differences between 10-meter resolution imagery	45
5.15	The confusion matrices of the multi- and binary-classification experiments Sentinel-1 and Sentinel-1* (*=without a frozen first layer)	47
5.16	Visualisation of polygons representing the four damage types, comparing the xBD and Sentinel-1 dataset	48

List of Tables

3.1	Joint Damage Scale descriptions on a four-level granularity scheme [23]	15
3.2	Details of the disasters included in the study area	16
3.3	Sentinel-2 product types [17]	18
3.4	Details of the extracted Sentinel-2 data included in this research	18
3.5	Details of the extracted Sentinel-1 data included in this research	21
4.1	Architecture of Inception-V3 [54]	27
4.2	The six designed experiments including xBD, Sentinel-2 and Sentinel-1 imagery in Single-Mode data scenarios	29
4.3	The six designed experiments including xBD imagery in Dual-Mode data scenarios	30
4.4	The six designed experiments including xBD imagery in Cross-Mode data scenarios	30
5.1	Results of multi-classification per disaster (epoch=100), sorted by the F1 scores and recall value per damage type; No = No-damage, Min.= Minor-damage, Maj.=Major-damage and Des.=Destroyed	33
5.2	Results of multi-classification of the imbalanced and balanced dataset (epoch=50), sorted by the F1 scores and recall value per damage type	34
5.3	Results of multi-classification per disaster (epoch=50), trained on the xBD dataset, sorted by the F1 scores and recall value per damage type	38
5.4	Results of multi- and binary-classification with varying resolution settings of the xBD dataset, sorted by the F1 scores and recall value per damage type	40
5.5	Cohen's Kappa coefficients (κ) per xBD experiment, multi-label classification	41
5.6	Results of multi- and binary-classification of the optical imagery datasets consisting of 10-meter resolution, sorted by the F1 scores and recall value per damage type	45
5.7	Results of Dual-Mode experiments with use of the xBD dataset, multi-label classification	45
5.8	Results of Cross-Mode experiments with the use of the xBD dataset, multi-label classification	46
5.9	Results of the multi- and binary-classification experiments Sentinel-1 and Sentinel-1* (*=without a frozen first layer)	46
5.10	Cohen's Kappa coefficients (κ) of the multi-classification experiments Sentinel-1 and Sentinel-1* (*=without a frozen first layer)	47
1	The 19 included disasters in the xBD dataset [23]	55

Introduction

The frequency of natural related disasters is increasing due to the effect of climate change. The average number of events recorded per year has quintupled worldwide in just fifty years time. Rising air and water temperatures are leading to amplified storms, extreme droughts, longer wildfire seasons, heavier precipitation, higher sea levels and subsequent floods. The extent and level of destruction and devastation have impacted, in 2019 alone, more than 95 million people and damage costs running up to 140 billion US Dollars [43]. Especially, the under-developed countries are affected and assistance is becoming a constant necessity.

Numerous non- and governmental organisations are motivated to mitigate and control the impact of these disasters, prioritising decreasing the death rate.

A comprehensive disaster management system is designed and integrated within humanitarian aid operations to structure the logistics in the time period of a natural hazard, showcased in Figure 1.1. The system can be subdivided into two phases concerning pre-event and post-event countermeasures [29]. The timeline starts with harm mitigation and preparation for a future disaster, followed by prediction and early warning, all based on preventing damage. The four post-event countermeasures include damage assessment, disaster response, recovery, and reconstruction. At every phase of the management system, high-quality information is crucial to understand the natural phenomena and contribute to immediate response.

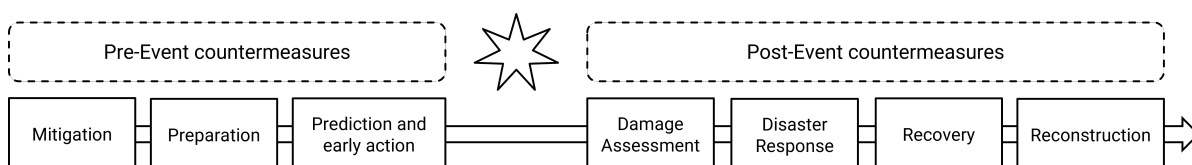


Figure 1.1: Comprehensive disaster management [15]

The first four days after the disaster has struck, the data indicating the level of damage is essential to locate the regions with the most vulnerable victims and to execute adequate rescue and relief actions [50]. Currently, the information is collected with the use of different methodologies and tools [21]. The most fundamental approach is conducting field surveys directed by observation teams to assess the type and scale of damage in detail [3]. These ground-based observations are time-consuming, costly and only possible to acquire in passable areas, causing delay and incomplete data. Furthermore, this method limits quick updates and provides inconsistent quality due to human errors [35].

To overcome these limitations, remotely sensed data is introduced to map the affected region. Drone-, air-, or space-based observations can be gathered by optical or radar sensors. Each of these platforms and sources is accompanied by its own ad- and disadvantages. The drone- and air-based systems can be deployed on the spot after the disaster, when necessities are available and capabilities are met, which are often limited in emergent regions. Additionally, imagery in advance of the disaster is not observed consistently due to different prioritisation and the unpredictable timing of events.

Satellite imagery has the ability to map a big region because of its high spatial coverage. Furthermore, the data is collected continuously, making it suitable to access pre- and post-event data. The revisit time and orbit position specify the availability and usability of the data with respect to the target location. A variety of satellite imagery is provided by private and governmental institutions, accessible in a range of resolutions; the shortest distance between two points on a specimen that the observer can still distinguish as separate entities. The type of sensor of the satellite mission defines how to interpret the visualisation and the implementation in the comprehensive disaster management system. Optical imagery is based on a passive sensor, observing the reflectance of visible wavelengths provided by Sunlight. The biggest downside is the quality dependency on weather and day-night cycles. The space-borne sensing method Synthetic Aperture Radar data is unaffected by clouds, smoke and darkness. The active sensor collects data by actively illuminating the ground with radio waves and measuring the reflected signal. The observed backscatter differences indicate the properties present at the surface. This technique is able to provide information with an off-nadir and side-view orientation.

To perform damage assessment, the first post-event countermeasure of the comprehensive disaster management system, satellite images of both types are reviewed by specialised organisations such as UNOSAT and manually annotated in affected regions [57]. This approach does resolve the safety limitations of field survey damage predictions. However, restrictions still exist regarding time, scalability, and human labelling subjectivity. Therefore, automatic damage assessment is introduced in innovative research to take full advantage of the data available and subsequently improve rescue missions.

1.1. Automatic Damage Assessment

Automatic Damage Assessment (ADA) involves models which classify building polygons with binary- or multi-class labels, by only inputting imagery. The models have the potential to produce more accurate, quicker and cheaper damage maps of the impacted regions.

In previous research, a range of techniques is investigated to obtain reliable results, based on different satellite imagery types and mono- and multi-temporal data scenarios. Results show varying success.

1.1.1. Previous Research

The first damage assessment technique based on optical input data was represented by texture- and segmentation-methods, which highlighted and classified buildings by their visual characteristics [59]. Rapidly, these methods were followed by supervised machine learning techniques. The support vector machine modelling application proposed a good damage prediction rate with the help of multiple parameters, established in advance [24]. Additionally, the Bayes Decision theory was tested on numerous disaster datasets. The method assigned labels based on the decision rule of conditional probabilities, created by assumptions of feature distributions. Both supervised learning techniques required a-priori knowledge to direct the classification process.

With the introduction of unsupervised learning, the innovations were boosted and the application of Convolutional Neural Networks (CNN) was explored. A CNN is a type of artificial neural network, specifically designed to process pixel data to recognise features. The model is fed with training data, consisting of images and corresponding ground truth labels. After the training phase, the self-learning CNN model is ready to predict classes on unseen data, named the test-set.

Numerous different architectures were designed, including varying order and number of convolution-, activation layers and type of augmentation [2] [11] [19] [31] [32] [44] [65]. It was found to be difficult to directly compare performances of the models due to non-similar class balances, disaster types and data sources. In most researches, class imbalance occurred between the binary classes *no-damage* and *damage*, resulting in unreliable high accuracy scores. Nonetheless, it was identified that all models did learn damage-related features, since their accuracy was superior compared to a naive classifier that always predicts the majority class.

In most researches, the input data contained solely one use case of a previous earthquake, tsunami or hurricane. This created non-robust models to real-life situations, as all disaster types showcase different visible damage characteristics. Subsequently, Tinka Valentijn et al. (2020) [58] investigated the relationship between the performance of a CNN and a range of hazard types. Although the results varied between datasets, no relation was determined to be conclusive. The quality differences between pre- and post-event imagery of the selected disasters made it complex to substantiate an explanation.

The ADA methods inputting SAR data consist of non-similar fundamental characteristics. SAR imagery does present unique challenges for computer vision algorithms and human comprehension due to the non-literal imagery type visualisation [67]. In previous research, SAR data was interpreted using their grey bands or by exploiting phase differences between two radar observations, called interferometric SAR (InSAR). This latter was commonly used to estimate the interferometric coherence and intensity correlation. These methods required three images of the location of interest acquired by the same satellite mission and with identical imaging geometry. The coherence and correlations were computed of the created pre-disaster and co-disaster image pairs [47]. Subsequently, the damage caused by the natural disaster was able to assess by detecting the change between the two image pairs. Both the interferometric coherence and the intensity correlation showed a decreasing value with an increasing damage level. This value was based on the varying phase and intensity of the complex observed SAR back-scatter [20].

One of the biggest hurdles of innovating ADA techniques is the low quality and limited amount of validation data observed by any sensor type. Annotating ground truth labels of buildings linked to previous disasters is a labour-intensive job and requires consistent protocols of damage scales. However, to create a reliable model for various operational situations, an extensive input dataset must be present, including multiple hazard types originated all over the globe to showcase a variety of constructions. In 2019, a new opportunity within the research field appeared by the start of the xView2 challenge held by the Defence Innovation Unit [66]. The largest and highest-quality publicly available dataset xBD was released in this challenge, containing high-resolution satellite optical imagery with specified building locations and damage scores. The xBD dataset consists of 19 disasters, including six different data types; volcano eruptions, wildfires, floods, tsunamis, earthquakes and hurricanes [23].

1.2. Problem Statement

Despite the academic research on satellite observed imagery driven ADA, the implementation after natural hazards shows a delayed effect. Multiple models exist, specifically designed for one data type with respective characteristics. Humanitarian aid organisations are lacking resources and procedures to be able to apply and master all methods.

The organisation *510 - an initiative of the Red Cross Netherlands* has designed a CNN named *Caladrius*, trained and tested on high-resolution optical imagery. Under perfect data availability conditions, the model would be applicable to detect damage of buildings. Unfortunately, this is often hindered by circumstances such as the satellite revisit time, cost of information or weather forecasts. Whilst at the same time, different sources, types or lower-resolution data are available to implement. However, this is not possible due to the missing knowledge on how to interpret the predictions. The reliability is not tested and determined, which is not permitted in life and death situations.

1.2.1. Research Objective

This research is focused on eliminating the deficit to augment the implementation of ADA. In cooperation with *510 - An Initiative of the Netherlands Red Cross*, the *Caladrius* model is prepared and improved to apply to various data availability scenarios, including optical and SAR data.

The research will provide insights into the reliability of classifications in all-weather situations and by inputting openly accessible data sources in addition to expensive and exclusive high-quality data.

With these results, the *Red Cross* can judge how and when to implement the damage assessment map within the comprehensive disaster management system to improve humanitarian aid's efficiency by facilitating data-driven aid prioritisation. As such, the research questions is phrased as follows:

"What is the influence of different input data characteristics to the true prediction rate of assessing damage on building level, using the Convolutional Neural Network 'Caladrius'?"

To compare the different data characteristics, various resolution imagery and satellite sources are included to train and test the *Caladrius* model. Sub-questions are determined, to specify the research objective:

1. What is the influence on the true prediction rate of the *Caladrius* model by training and testing on optical imagery with varying resolution settings, ranging from 0.5 to 10.0-meter?
2. What is the influence on the true prediction rate of the *Caladrius* model by mixing the input imagery with non-identical resolution settings?
3. What is the influence on the true prediction rate of the *Caladrius* model by training and testing on Synthetic Aperture Radar imagery?

Experiments are set up to answer these sub-questions, consisting of single-, dual- and cross-mode data availability scenarios, imitating operational cases. The data scenarios are based on the xBD dataset, which is down-sampled in lower resolutions and replicated using openly available datasets, originating from the Copernicus missions Sentinel-1 and Sentinel-2. The provided coordinates of the image bounds enable the recreation and the ground truth labels ensure the validation of the model. The performance and true prediction rate of the *Caladrius* model are measured using the F1-scores, the Cohen's Kappa coefficient and the Area Under the Curve (AUC).

1.3. Overview of Chapters

This thesis is split up into seven chapters. The Literature Study will provide information about the acquiring process of satellite imagery and the respective interpretation. In addition, the functioning of a CNN is explained, divided into feature learning concepts and the operation of classification. The Data chapter includes a detailed description of the datasets used to train and test the *Caladrius* model. Plus, the extraction method of the openly accessible datasets Sentinel-1 and Sentinel-2 is provided. Subsequently, the Methodology chapter concerns the required steps to identify buildings and classify the corresponding polygons with damage types, along with data pre-processing steps, the experimental setups and selected performance metrics. Next, the Results chapter showcases and discusses the performance differences of the *Caladrius* model by executing the experiments. Found relations and correlations are elaborated to reason the reliability of predictions in varying data availability scenarios. Last, the Discussion and Conclusion chapters aim to answer the research questions and provide advice for improvements and future work.

2

Literature Study

In this chapter, background information is given to gain knowledge about fundamental concepts regarding the research topic. Firstly, the two satellite imagery types are elaborated in more detail, in section 2.1. The visualisations observed by optical and Synthetic Aperture Radar sensors require different interpretations, which is essential to understand before inputting into the *Caladrius* model. Subsequently, in section 2.2, the functioning of a Convolutional Neural Network (CNN) is discussed, including its layers and components.

2.1. Satellite Imagery Types

Satellites observing the electromagnetic spectrum provide three imagery types; visible, infrared and water vapour. Waves of charged particles produced by vibration travel through the atmosphere and the vacuum of space. These waves are linked to different wavelengths and frequencies. Instruments are required to detect the electromagnetic energy and to utilize the range of the spectrum to explore and understand processes [61]. The applied instruments can be subdivided in two types; active and passive sensors, visible in Figure 2.1. Both types acquire data at different spectrum wavelengths, applicable to various use cases.

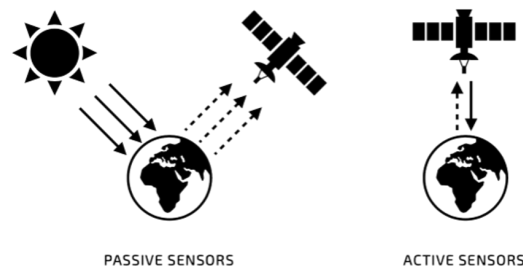


Figure 2.1: Sensor types of satellites to observe the Earth or third bodies [62]

The resolution of the imagery defines how the data from both sensors can be used and is dependent on the satellite orbit and sensor design. Four types of resolution can be defined, radiometric, spatial, spectral and temporal. The first describes the number of bits representing the energy recorded. The higher the resolution, the more information can be stored, showing a great discrimination between pixels. The spatial resolution is often referred to as the commonly used definition of resolution. It specifies the area on Earth's surface covered by one pixel. Details become visible with higher resolution imagery. The spectral resolution explains the ability of a sensor to discern finer wavelengths. The narrower the range of wavelengths belonging to a band, the higher the resolution. It is used to detect densities and property details of soils. The last resolution description represents a satellite's time period to revisit the exact location. The temporal resolution is based on the mission's orbit height and swath width. Unfortunately, a trade-off is required to optimize one of the resolution settings, based on the respective use-case.

In this research, multiple datasets are included to review the *Caladrius* model. The datasets vary in spatial resolution and imagery type. The two types used are elaborated in the next sections.

2.1.1. Optical Imagery

Optical imagery is a passive sensing method that measures naturally available energy. It has equalities with visuals of a standard camera, using wavelengths of visible light and thermal infrared. The electromagnetic emissions can be produced locally from vegetation or exist due to reflected Sun illuminations. This latter causes the undesired dependency on the day-night cycles. Additionally, clouds and shadows can contaminate the imagery and can make it non-usable in many situations.

Four different systems exist to channel the received light, containing panchromatic, multi-spectral, super-spectral and hyper-spectral imaging systems. Most openly available datasets such as Landsat and Sentinel consist of a multi-channel detector including a few spectral bands. Each band is sensitive to radiation within a narrow wavelength range, representing the brightness and colour information. An oscillating mirror continuously scans the surface of the Earth perpendicular to the velocity of the satellite. Every mirror sweep scans six lines simultaneously in each of the spectral bands [53].

The multi-spectral scanners can be further divided into two types; whiskbroom and pushbroom scanners. The first is also known as the across-track scanner, which uses a rotating mirror and a single detector to scan the scene along with a long and narrow band, one pixel at a time. The pushbroom scanner does not have a movable mirror, but uses several detectors placed perpendicular to the flight direction. The imagery of the second scanning technique is of higher quality due to the longer observation times, absorbing a stronger signal [56].

Optical imagery is collected by governmental and private organisations and available in various resolutions.

2.1.2. Synthetic Aperture Radar

Synthetic Aperture Radar is an active data collector, producing energy in form of radio waves and recording the reflected energy after interacting with the Earth. The received signal's strength, direction, and travel time provide information of properties linked to the surface and object observed. Diffused scattering will be measured when the surface is rough, visualised by varying pixel brightness. Specular reflection appears when a smooth surface reflects the beam [63].

The radar sensors utilise long wavelengths with a range of centimetres to meters, making it possible to penetrate through clouds. Different wavelengths are referred to as bands, with corresponding letters such as X, C, L and P. The length of the wave determines how the radar signal interacts with the surface and how far a signal can penetrate a medium, for example, soil, ice or canopies of forests [42].

One of the limitations of radar imagery is the achievable azimuth resolution, parallel to the flight direction of the satellite, visualised in Figure 2.2. The width of the beam's footprint on the surface is proportional to the antenna length and determines the resolution. Large antennae are obstructive and therefore restrict radar imagery's visible detail.

This undesired effect can be mitigated by introducing a moving antenna to synthesise the working of a long antenna. The many radar pulses of the same object provide information, which improves the resolution in the azimuth direction.

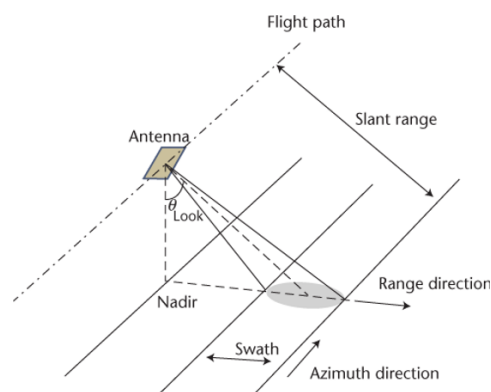


Figure 2.2: Imaging Radar Geometry [26]

The transmitted and received path of the signal can differ in polarisation, referring to the plane's orientation in which the transmitted electromagnetic wave oscillates. Typically, SAR transmits linearly polarised, with the horizontal direction indicated by the letter H and the vertical by V.

The polarisation setting can be installed at the receiver and transmitter, making it possible to create various combinations; VV, HH, VH, HV. The strength of these combinations carries information about the structure of the imaged surface, based on the types of scattering, such as rough surface, volume and double bounce.

SAR visualisations are a non-literal imagery type and should be interpreted accurately, using a different approach than optical imagery. Colours are not visible, but intensity ranges depending on the amount of energy the SAR sensor measures. This intensity is related to the dielectric constant of the scattering object, which stands for the sensitivity of the material to the reflectance of electromagnetic waves and the roughness of the surface [36].

Shadows, foreshortening and layover effects can cause distortion within the SAR visualisations. The shadows are formed by objects blocking the path of the radar beam. These areas will return no signal and appear black. Foreshortening causes misalignments when the radar beam reaches the base of a tall feature, such as a mountain or high building, tilted towards the radar before reaching the top. The slope will appear compressed in comparison with reality. The layover effect is the opposite of foreshortening, by contacting first the top of a tall feature. The top will be displaced towards the radar, creating a 'lay over' the base. These geometric limitations occur due to relief displacement, which is one-dimensional perpendicular to the flight path. The sensor's look angle can influence these phenomena. A larger angle will increase the shadows' length while minimising the layover effect [13].

Additionally, speckle can arise in SAR imagery, meaning a salt and pepper variation in the pixel brightness, which degrades the quality of the images. Speckle occurs due to the possibility of many scatter signals in a given pixel, which will lead to positive and negative interference.

SAR imagery is popular because of the 24-hour all-weather observations, deployable for various use-cases. Many private companies respond to this growing demand by providing high-resolution imagery on request, such as Capella space [26]. These high-detail gray visualisations could replace optical imagery in many situations.

2.2. Convolutional Neural Network

Artificial Intelligence (AI) is introduced expandingly within many situations, to bridge the gap between the capabilities of humans and computers. Machine- and deep learning, branches of AI, are based on learning and adapting models through experience to execute tasks such as image recognition and classification. The corresponding input data can be labelled or unlabeled.

A Convolutional Neural Network (CNN) is an algorithm based on deep learning techniques, analysing input data and assigning importance to features and characteristics. A CNN can be subdivided into two parts; feature learning and classification, visualised in Figure 2.3.

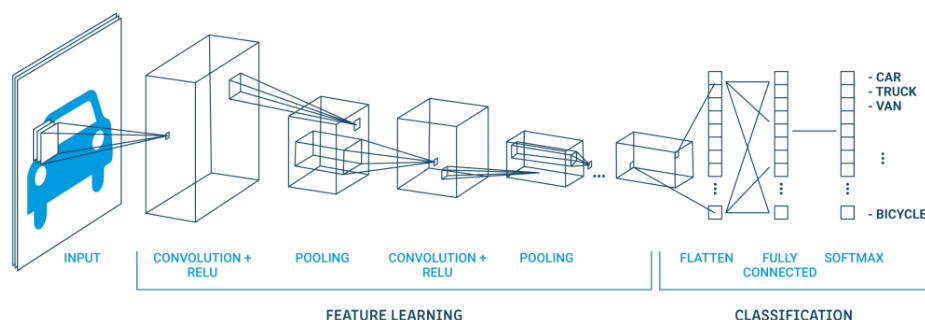


Figure 2.3: Example of a structure of a Convolutional Neural Network [18]

The feature learning exists of several connected layers, including three phases. Starting with the convolutions, which produces a set of linear activations and searches for patterns with the use of filters.

Followed by the detection phase, consisting of a non-linear activation function, such as the Rectified Linear Unit (ReLU). The last phase equals the pooling function to modify the layer's output. The second part of the CNN, the classification, includes the flattening transformation, a fully connected layer and a Softmax function to classify the object with a probabilistic value between 0 and 1. In the next subsections, the characteristics of the two parts are further explained. Subsequently, an elaboration about the training phase of a CNN is given and possible regularization techniques to prevent overfitting.

2.2.1. Feature Learning

Convolution Layer

A convolution layer is introduced to scan the input imagery and detect patterns with the use of filters, also called kernels. A kernel is a smaller sized two-dimensional array compared to the input data and filled with specified weight values, learned during the training phase by the gradient descent to minimise the loss function. By moving systematically over the input data, features can be identified at every location. In Equation 2.1, the mathematical relation is given of the convolution, when continuous or discrete, respectively [4]. The integral expresses the overlap of filter function g shifting over the input function f . The output of the expression is equal to the feature map. In Figure 2.4, the flow of the mathematical convolution is visualised.

$$(f \cdot g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t - \tau)d\tau$$

$$(f \cdot g)(t) = \sum_{\tau=-\infty}^{\infty} f(\tau)g(t - \tau)$$
(2.1)

Multiple filters can be included in a convolution layer, indicating the depth dimension of the feature map. A trade-off is introduced, as an extra filter also increases the computing power and training time. Every filter has its own learnable parameters, which remain equal while shifting over the input image. This latter is called parameter sharing and one of the most significant advantages of the convolution layers compared to traditional Neural Networks. Instead of learning a parameter for each location, only one parameter is learned for the feature map.

A second quality of the convolution layer equals the property of equivariant representations. This means that the feature map will reflect any affine transformations occurring in the input data, useful when a local function can be applied everywhere.

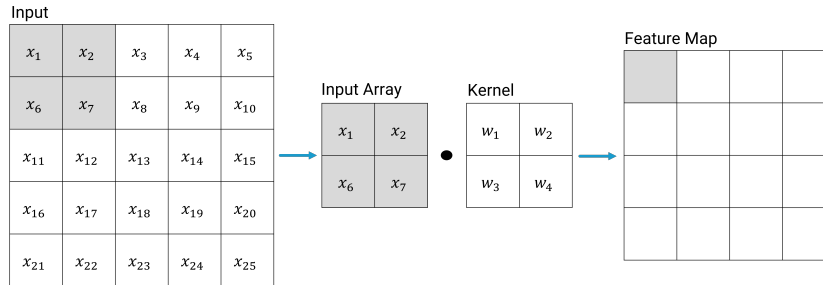


Figure 2.4: Filter applied to a two-dimensional input, with the corresponding output of the feature map

The output size of the feature map $((n_h - k_h + 1) \cdot (n_w - k_w + 1))$ is dependent on the size of the input shape $(n_h \times n_w)$ and filter size $(k_h \times k_w)$. By applying a smaller filter than input size, the image shrinks every time a convolution operation is performed.

In general, pixels in the middle of the input image are filtered more often than pixels on corners and edges. Consequently, the information on the borders is under-represented. To overcome this problem, padding is introduced. Padding is a process of adding layers of zeros to the borders of the input image, transforming the input size to $(n_h + 2p_h) \cdot (n_w + 2p_w)$ and output size to $(n_h - k_h + p_h + 1) \cdot (n_w - k_w + p_w + 1)$, in which p is equal to the number of layers of zeros.

In default situations, the filter starts at the upper-left corner of the input array and slides one pixel at a time over all locations. By initiating a stride, equal to the number of pixels shifts over the input matrix, intermediate pixels can be skipped, optimal for computational efficiency or down-sampling wishes. Again, this has an influence on the output shape of the feature map, $(n_h - k_h + p_h + s_h) / s_h \cdot (n_w - k_w + p_w + s_w) / s_w$, where s_h and s_w are the corresponding height and width step sizes of the stride [4].

Activation Function

The activation function defines how the weighted sum of the input, originated from the convolution layer, is transformed to the next layer. Often the activation is performed with a non-linear expression, which is beneficial for learning complex tasks.

Multiple activation functions in Machine Learning are designed and still innovating. Popular examples are the rectified linear unit function ReLU [52], Max-out and Channel-out [60], the Sigmoid, hyperbolic tangent and arc-tangent functions [1]. The selection of the best-suited activation function has a large impact on the capability and performance of the network.

The ReLU is the most integrated activation function in many types of neural networks. It has proven to optimise the performance and ease the training, reducing the computational complexity expressed in time and space. Additionally, the ReLU can be used by multi-layer perceptrons and within convolutional neural networks, despite the risk of the vanishing gradient problem.

The mathematical relation of the ReLU layer is stated in Equation 2.2, in which the output is always equal to a positive value [7].

$$ReLU(x) = \max(0, x) \quad (2.2)$$

When the condition of $x > 0$ is not met, the output will equal zero. The first-order derivative can be computed, allowing the execution of back-propagation. Back-propagation is the process of computing the gradient of the loss function with respect to the model's weights backwards through the network. The ReLU is resistant to the vanishing gradient problem, because the gradient of the loss function will never approach zero. Nevertheless, this activation layer is accompanied by flaws, consisting of the exploding gradient problem and the possible presence of the dying ReLU; the situation when most output values are equal to zero and back-propagation cannot be performed.

Pooling

Multiple feature maps are created and stored in the CNN. The low-level features close to the input and high-order learnable and abstract features are represented in deeper layers [8].

The limitation of these feature maps is recording the precise position of patterns within the input data. Just with minor movements in the input data, due to cropping, shifting or/and rotation, the output of the feature map is heavily changed. A common approach to minimize this sensitivity is to down-sample the feature maps. A lower resolution version of the input signal will be created, including fewer details but retaining important structural elements. As mentioned earlier, down-sampling can be achieved by increasing the stride within the convolutional layer.

However, a more robust and suitable approach is introducing a pooling layer. Different pooling functions exist to summarize surrounding outputs within the feature map. The maximum pooling and average pooling are the most commonly used and showcased in Figure 2.5. The maximum pooling function outputs the highest value within the window and the average pooling function computes the average of all tiles within the window.

Feature Map				Max Pooling: Output		Average Pooling: Output	
5	4	7	9	5	9	3	5
0	3	1	3	6	8	4	4
3	5	0	2				
6	2	8	4				

Figure 2.5: Example of maximum and average pooling extractions

Implementing a pooling layer after a convolution and ReLU activation layer is a typical pattern used for ordering layers and can be repeated multiple times. The capability added by pooling is called the model's invariance to local translation. Additionally, the pooling layer reduces the dataset size, processing time and minimises the risk of overfitting [8].

2.2.2. Classification

The second part of the CNN consists of the classification, subdivided into three steps.

First, a feature map is transformed to one column during flattening. The column is filled with the values of the feature map matrix, row by row. With many feature maps, all will be flattened and placed beneath each other, resulting in one long vector of inputs [8].

Next, the flattened vector is inputted into a fully connected layer, which learns non-linear combinations of the high-level features found by the convolution layers. The layer connects the first part of the CNN with the last activation function.

The last activation function is the third step called the Softmax; also known as the normalized exponential function. A vector of arbitrary real values is turned into a vector of probability factor values adding up to 1, presenting predictions for each label. The mathematical expression of the Softmax S can be found in Equation 2.3.

$$S(y)_i = \frac{e^{(y_i)}}{\sum_{j=1}^n e^{(y_j)}} \quad (2.3)$$

Where y_i equals the i -th element of the input vector and n the number of classes [6].

2.2.3. Training a CNN

After the architecture of the CNN is built, the data can be collected and pre-processed to train the model. The input data is divided into three subsets to optimise and test the performance. Often, a split ratio is chosen of 80/10/10 representing the training, validation and test dataset, respectively. The training dataset functions to fit the model and determines the weights and biases of the neural network. Subsequently, the validation dataset unbiasedly evaluates the established fit during the training phase, from which the results are reviewed to update the hyper-parameters. Next, the test dataset is inputted into the trained model to classify unseen data.

The split ratio depends on the samples in the dataset and the trained model type. When little hyper-parameters are included, a smaller validation set can be chosen.

In general, it is beneficial for the model's performance to include a big training set. Although, it is important to detect and prevent overfitting, which can occur when the model learns specific characteristics of the training set that do not occur in unseen data. For example, the features could identify noise or irrelevant information and learn to classify a specific label.

Figure 2.6 shows the prediction function with respect to the real values of the input data. In the underfitted scenario, the model has not trained long enough or on too few samples, detecting no meaningful relationships between the input and output variables. The overfitted scenario shows the opposite.

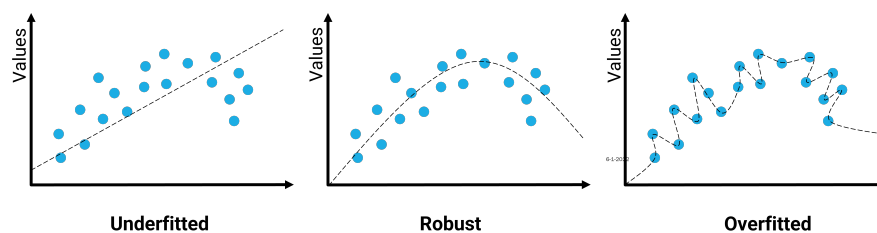


Figure 2.6: Relation between the prediction function over time and the real values of the input data

A good balance should be found between the model's training set size and complexity to optimise the performance. This could be obtained using the error computation of the training- and test-set. If the training set has a low error rate and the test data has a high error rate, it signals overfitting.

The error can be expressed using a loss function. Multiple functions exist, linked to their own characteristics and implementations. In this research, multi-classification is introduced and, therefore, the cross-entropy loss is selected [37]. In Equation 2.4, the mathematical expression is given to compute the difference between true values (y_{ic}) and the predicted values (\hat{y}_{ic}). The subscript i stands for the data point belonging to class c .

$$J = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{ic} \log(\hat{y}_{ic}) + (1 - y_{ic}) \log(1 - \hat{y}_{ic}) \quad (2.4)$$

The loss function is the model's objective function, which must be minimized during the iterations in the training phase.

The training phase can be subdivided into a forward and backward pass. In the forward pass, the input is running through all network layers. Followed by the backward pass, where the gradients of the loss function are propagated and weights are updated.

The prediction values per class are obtained after the forward pass by inputting batches of the complete training set. The predictions are rated with the loss function, which indicates if the weight and bias parameters of the neurons should be tweaked and iterated. The bias represents the shift of the activation function, and is equal to the constant in a linear function. Bias units are independent of the previous layer but connected to their own weights. The weights define the strength of a connection between neurons, it affects the influence of a change in the input upon the output. In Equation 2.5, the computation of a neuron is given of the weighted sum of the inputs, where x_n is equal to the input, w_n to the weights and b_n to the bias.

$$Y = \sum (w_n \cdot x_n) + b_n \quad (2.5)$$

Subsequently, the output value Y is fed into the activation function, which prepares a prediction value, visualised in Figure 2.7.

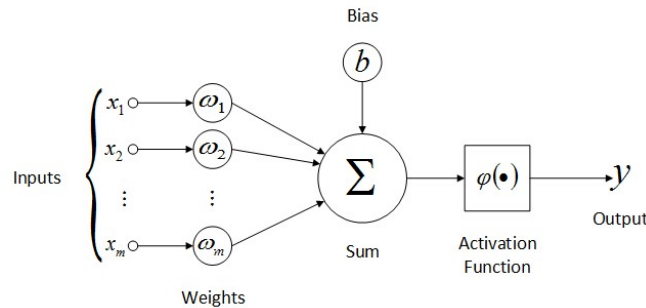


Figure 2.7: Mathematical model of an artificial neuron, including input data, weights, bias and activation function [46]

After the loss is computed, the backward pass starts to seek the optimal change of weights and biases to reduce the error in the next iteration. The gradient of the loss function is reviewed and specifies the rate of change and the sign shows the direction. The old weights w_n are updated with use of this gradient ($\partial J / \partial w_n$), as can be seen in Equation 2.6 [45]. The learning rate is equal to a , and is a hyperparameter that controls how much to change the model in response to the estimated error each time the model weights are updated [64].

$$*w_n = w_n - a \left(\frac{\partial J}{\partial w_n} \right) \quad (2.6)$$

While back-propagating, the chain rule is applied to compute the gradient one layer at a time. The iteration from the last layer to the start, prevents redundant calculation of intermediate terms. Unfortunately, no global optimum is guaranteed while updating the weights, a local optimum can confuse.

One epoch consists of one completed forward- and backward-pass for all batches in the training set. Again, an optimum must be found between the number of epochs and the final permitted loss. Increasing the number of epochs enlarges the chance of converging towards an optimal performance value, but also increases the computer time.

2.2.4. Regularization Techniques

Regularization techniques are invented to prevent overfitting by reducing the test error at the expense of increasing the train error. The four most applied techniques contain the L1 and L2 regularisation, dropout, data augmentation and early stopping.

The first method adds an extra term to the cost function, known as the regularisation term. This term reduces the values of the weight matrices, which simplifies the model.

The L1 term is selected when the absolute value of the weights should be penalized or even be reduced to zero. This application could be beneficial to compress the model. The L1 term is given in Equation 2.7, in which the λ represents the regularization parameter. This hyper-parameter should be optimised for the best result [33].

$$L1 = \lambda \cdot \sum ||w|| \quad (2.7)$$

The L2 term, also called the weight decay, is almost identical to the L1 regularization but will never equal zero.

$$L2 = \lambda \cdot \sum ||w||^2 \quad (2.8)$$

Secondly, dropout is a popular regularisation technique that randomly selects nodes at every iteration and removes them with their corresponding in- and outgoing connections. Every iteration includes a different set of nodes, resulting in various outputs. The hyper-parameter of the dropout function states the number of nodes that will drop. The method replicates the effect of training many neural networks with different architectures, in parallel.

The dropout regularisation technique can be applied at many stages of the neural network. For example, in the beginning the dropout can prevent the hierarchy of importance per incoming batch [52].

By increasing the size of the training dataset, overfitting could also be prevented. The method of data augmentation is introduced, which does not require new input data with expensive and time-consuming labelling. The method replicates existing data and tweaks the input using rotating, flipping, scaling and shifting.

Last, the early stopping approach reviews the validation score and indicates when to stop training the model. The moment is selected when the performance of the validation set does not improve or stagnates. The sweet stop, expressed in number of epochs is visualised in Figure 2.8 with the blue dotted line.

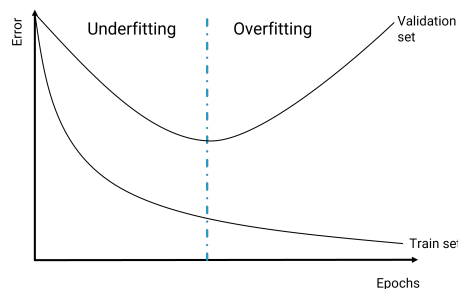


Figure 2.8: The validation and train set error with respect to the iterations in the training phase of the CNN model

3

Data

To investigate the research objective, multiple datasets are selected and tested with the *Caladrius* model, listed in Figure 3.1. The used data can be subdivided into two types: optical and Synthetic Aperture Radar (SAR) imagery. Both have specific characteristics, which lead to ad- and disadvantages. The optical imagery is represented by the xBD dataset released in the xView2 challenge of the Humanitarian Assistance Disaster Recovery organisation, plus the openly accessible Sentinel-2 data. The Sentinel-1 mission provides the SAR data. Unfortunately, no high-resolution SAR imagery was available for free and included within this research.

The influence of the input data with respect to the model's true prediction rate can be compared using equal disasters and regions to minimise the variation and reasoning of miss-classifications. The xBD dataset is determined as a benchmark and replicated in down-sampled versions, and with the use of Sentinel data. The longitude-latitude coordinates of the image bounds are extracted to create identical datasets, in addition, the provided ground truth labels of the damage types per polygon are used to validate the experiments.

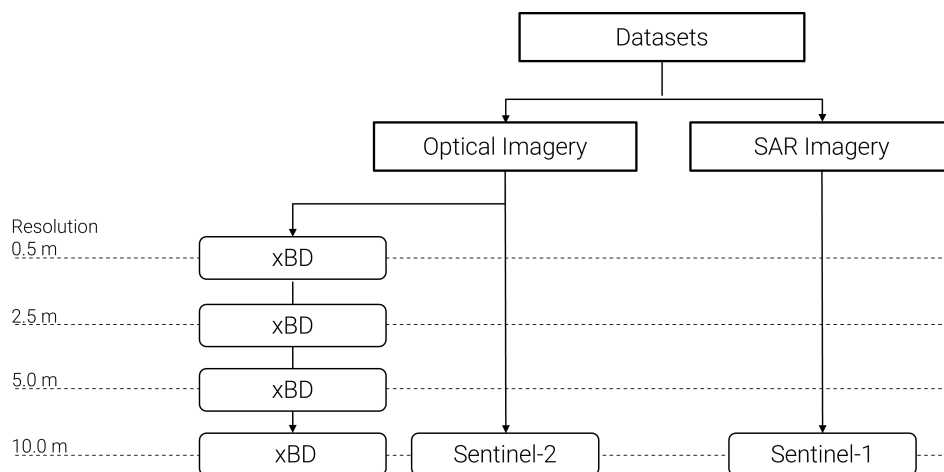


Figure 3.1: Data source overview subdivided into two types: optical and SAR imagery

This chapter covers the description of all data types and datasets. Firstly, the high-resolution optical imagery representation of the xBD dataset is explained, in section 3.1, including the chosen study area. Section 3.2 elaborates the down-sampled xBD version and the extraction of Sentinel-2 data. It is interesting to parallel the xBD 10-meter resolution and Sentinel-2 imagery to identify differences. Finally, section 3.3 contains information about the SAR data type and the implementation in this research.

3.1. High-resolution optical imagery

Imagery collected by optical sensors has similar characteristics to visuals of a standard camera, using wavelengths of visible light and thermal infrared. The spatial resolution defines the detail detectable in the figures, available in a wide range. High-resolution imagery (0.1-1.0 meters) is beneficial while applying change detection methods. The biggest disadvantage equals the dependency on the day-night cycle and weather conditions. More information about this sensing method is stated in subsection 2.1.1. In this research, the three colour bands; Red, Green and Blue are selected to picture the landscape and the built-up areas.

3.1.1. xBD

The xBD dataset was released in 2019 by the Humanitarian Assistance & Disaster Recovery for the xView2 challenge, with the specific purpose to detect damage at building level. The challenge encouraged the progress of accurate and efficient deep learning models, to distinguish dangerous situations with pre-and post-disaster satellite imagery [66].

The dataset consists of 19 disasters linked to 22.068 images, sourced from the Maxar open data program. This organisation provides imagery of major crisis events observed by multiple satellite missions: Worldview02, Worldview03 VNIR and Geoeye01 [38]. The data ranges from 0.5 to 1.0-meter resolution.

The disasters within the xBD dataset are located across the globe, to ensure a variation of shape and size of constructions based on landscape differences. Furthermore, multiple natural hazard types are included to showcase a range of visible damage features. The seven disaster types contain earthquakes, wildfires, tornadoes, hurricanes, floods, tsunamis and volcanic eruptions. In Appendix A, the details of the location and time period are listed of the 19 disasters. The included images per disaster event are unevenly represented concerning the covered surface area km^2 and the number of polygons.

In Figure 3.2, the data-frame of xBD is showcased, linking pre- and post-event images to corresponding JSON files. The before and after visualisations enable classification based on change detection.

The images are released in PNG and GeoTIFF versions. The JSON files include the x/y pixel- and longitude/latitude coordinates of the polygons, valid for the respective image types. The drawn polygons of the visible building footprints are estimated on the pre-imagery by data annotators and overlaid on their matching post-event imagery pair. This provides the ideal box before the damage occurred, since the footprint may be significantly shifted during a disaster. However, this method can result in missed buildings, firstly, because the building did not exist yet in the pre-imagery, secondly, due to coverage of clouds, haze, or vegetation. These missed buildings are neglected within the research.

In the post-event JSON files, the damage labels per polygon are given, assuming that all buildings before the events are non-damaged. Additionally, metadata is listed within the JSON files to represent background information about the image collection: sensor, grounds sampling distance, capture date, pan resolution, sun azimuth-, sun elevation-, target azimuth- and off-nadir angle.

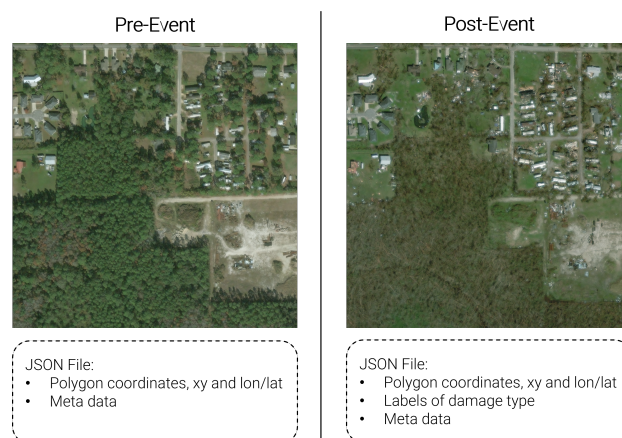


Figure 3.2: xBD data-frame: Pre-event imagery (left) and Post-event imagery (right), with the corresponding JSON files (Example: Hurricane Michael 78)

Each post-event facility is classified with a damage label based on the Joint Damage Scale, shown in Table 3.1 and Figure 3.3. The scale consists of four classes, ranging from *no-damage* (0) to *destroyed* (3), and is generalised in an iterative process to represent various disaster types, structure categories, and geographical locations.

Two review sessions of labelling ensure consistency, checked by experts on random samples. The California Air National Guard, NASA, and FEMA estimated that 2-3% of the annotations were mislabeled and were subsequently corrected manually.

The xBD dataset is heavily imbalanced towards the negative imagery, which results in a high level of *no-damage* buildings. This division is not preferred, as it can influence the true prediction rate of the minority classes. The classification model will skew towards the majority class to reach high accuracy.

Table 3.1: Joint Damage Scale descriptions on a four-level granularity scheme [23]

Disaster Level	Structure Description
0 (no-damage)	Undisturbed. No sign of water, structural or shingle damage, or burn marks
1 (minor-damage)	Building partially burnt, water surrounding structure, volcanic flow nearby, roof elements 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 no longer present

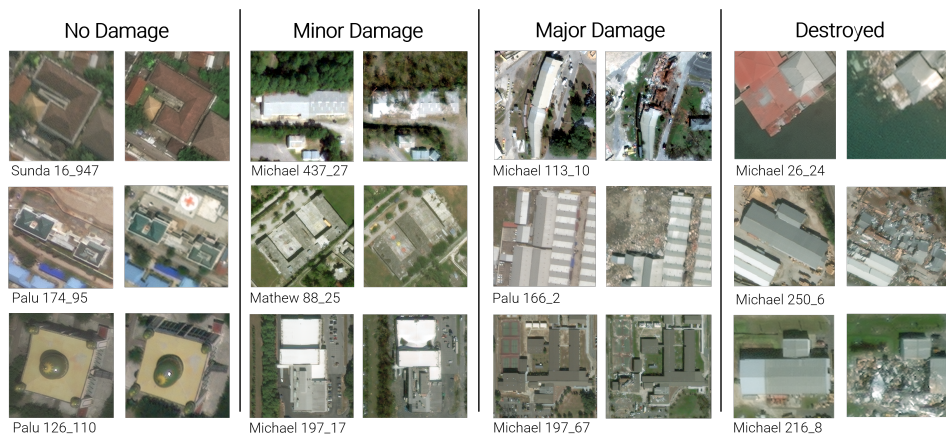


Figure 3.3: Visualisation of the Joint Damage Scale description of four-level granularity scheme

3.1.2. Study Area

In this research, it is decided to focus on a section of the total xBD dataset. A selection is made to reduce the model's input for two reasons. Firstly, the xBD dataset is replicated in down-sampled versions, in Sentinel-2 and Sentinel-1 imagery, creating multiple reprocessing steps and requiring much storage space. Secondly, the model is trained iteratively. The *Caladrius* model runs very computer power-intensive and time-consuming. To make the research executable within the time frame of the thesis, the input data is minimised.

To select specific disasters, two criteria are set: (1) similar visual damage characteristics, (2) events from 2015 until the present. The first restriction is created to ensure no correlation between performances and disaster types. By only selecting similar scenarios, the comparison can be made between different data input types. The second restriction is related to the launch period of the Copernicus mission. Before 2015, the Sentinel-2 sensor was not launched into space and, therefore, unable to collect images.

This selection procedure results in a dataset including four disasters; Hurricane Matthew, Hurricane Michael, Tsunami Palu and Tsunami Sunda Strait. All visual damage is wind-related, and therefore, the two different disaster types meet the first criteria of the selection. This specific damage type is chosen due to the obvious changes within the construction of the buildings. In contrast, floods show water and mud surrounding the houses and facilities. Furthermore, volcano eruptions and forest fires are less suitable because of the smoke and cloud formation after the event struck.

Details of the data linked to each event are listed in Table 3.2, including the number of image pairs, polygons and class distributions. Additionally, Figure 3.4 visualises the study area and the corresponding locations of the selected datasets. The real impact of the events did hit a larger region than shown, but the xBD images were only gathered above the large urban areas, covered by the indicated boxes.

Table 3.2: Details of the disasters included in the study area

Disaster Event Name	Images	Polygons	Class Distribution
Hurricane Matthew	405	9,506	18/54/12/16
Hurricane Michael	550	20,046	64/24/8/3
Tsunami Palu	196	24,119	85/0/2/13
Tsunami Sunda Strait	148	11,682	98/0/1/1
Total	1299	65,352	69/16/5/10

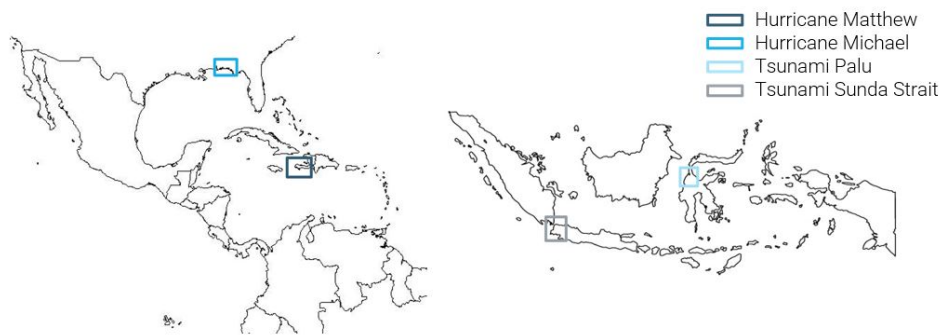


Figure 3.4: Study area: Central and North America (left), Indonesia (right)

Hurricane Matthew is well represented within the dataset, regarding a percentage of 30% of the total images. Nevertheless, the polygon portion is significantly lower. The geographical location of Haiti, with smaller and fewer cultivation, can be the reason for the non-polygon dense areas. The impact of the hurricane, which struck the 4th of October in 2016, was very violent and categorised with the value 4 on the Saffir-Simpson Hurricane Wind Scale ranging from 0 to 5. The combination of the effects of wind, coastal flooding and rain caused heavy flooding, landslides and destruction of the infrastructure, agriculture and natural ecosystem. In total, 2 million people were affected, 546 people did not survive the event, and many more lost their homes and searched for shelters [49].

Hurricane Michael was the seventh in the Atlantic hurricane season of 2018 and labelled with category 5. It hit the region of Central America and the US along Florida's Panhandle. In the xBD dataset, Panama City is captured with 550 images before and after the disaster. Due to the solid constructions of the buildings within this area, most remained intact [12].

The third disaster; Tsunami Palu, was located at Sulawesi in Indonesia, nearby the city Palu. It resulted from an earthquake with a 7.5 on the Richter magnitude scale, caused by increased stress on the Palu-Kora fault [28]. The six-meter high wave had a speed of 400 km/h and caused 4340 deaths and 14.000 injuries. Within the xBD dataset, the Tsunami Palu illustrates the most *destroyed* buildings.

Last, the tsunami near the area of Sunda Strait between Java and Sumatra in Indonesia was originated by a massive landslide into the caldera of the volcanic island Anak Krakatau. The tsunami struck during December 2018, at popular tourist's places. Locally, waves erased beach shores about 1-meter vertically and 20 to 30-meters wide. Within the xBD dataset, this specific disaster visualises a limited amount of *destroyed* and *major-damaged* buildings.

In Table 3.2, the class distribution is given per disaster, expressed in the percentage of samples belonging to the labels. It can be noticed that there is an understanding of a high imbalanced dataset, which can create difficulties within the research. In section 4.2, a method is elaborated to overcome this issue by resampling the dataset to create a balance.

3.2. Low-resolution optical imagery

In this research, the performance of the *Caladrius* model is tested on various resolution datasets to determine the relationship between both parameters. The lower resolution data will show fewer details, which can cause difficulties during the recognition of damage features. In some scenarios, the buildings will be smaller than the resolution in meters, resulting in a polygon representation with just one or a few pixels.

The low-resolution optical imagery is represented by the down-sampled xBD and Sentinel-2 data, extracted with the use of Google Earth Engine. Both are elaborated in this section.

3.2.1. Down-sampled xBD

The 0.5-meter resolution xBD dataset is down-sampled using a transformation algorithm, explained in subsection 4.2.1. The original dimensions of the xBD images are equal to 1024x1024 pixels, which is altered to 204x204, 102x102 and 51x51 pixels. The alteration creates images with the spatial resolution of 2.5, 5.0 and 10-meter, respectively. The effect of the down-sampling is visualised in Figure 3.5, more and more detail of the region and on building level is discarded.

The reason for the chosen maximum resolution setting of 10-meter is to compare the Sentinel-2 and Sentinel-1 data with the 'perfect' down-sampled xBD dataset, provided with related ground truth labels.

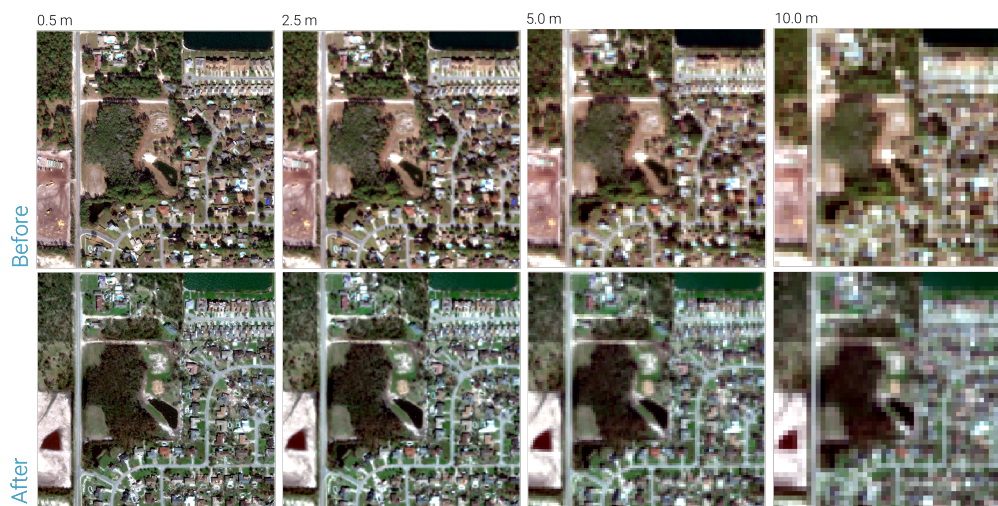


Figure 3.5: The effect of down-sampling the xBD input imagery to 2.5, 5.0 and 10.0-meter resolution (Example: Hurricane Michael 43)

3.2.2. Sentinel-2

The replication of the openly available Sentinel-2 dataset is created using the tool Google Earth Engine (GEE). Google Earth Engine is a platform providing satellite imagery and geospatial datasets with multiple analysis tools, enabling it to detect variabilities and trends on the surface all over the globe [48].

Sentinel-2 is part of the European Copernicus mission, including a constellation of two polar-orbiting satellites stationed in equal sun-synchronous trajectories, phased at 180 ° to each other. The mission aims to monitor the differences in land surface conditions using optical imagery. This can be realised with a wide swath width of 290 km and a high revisit time [17]. Sentinel-2 offers two products available for users, listed in Table 3.3. Due to the start date of the available products and the selected natural disasters of the xBD dataset, the Level-1C product is chosen for the research. This is unfortunate because the Bottom-Of-Atmosphere reflectance Level-2A could be clearer from noises.

Table 3.3: Sentinel-2 product types [17]

Name	High-Level Description	Production and Distribution	Data Volume	Data availability
Level-1C	Top-Of-Atmosphere reflectances in Cartographic geometry	Systematic generation and Online distribution	~600 MB (each 100 km x 100 km ²)	23-06-2015 - Present
Level-2A	Bottom-Of-Atmosphere reflectances in Cartographic geometry	Systematic and on-User side (using Senitnel-2 Toolbox)	~800 MB (each 100 km x 100 km ²)	28-03-2017 - Present

Every disaster of the dataset is handled separately to find the specific region, extraction date and filter settings. Firstly, the longitude and latitude coordinates of the xBD GeoTIFF's are extracted to specify the region observed within the dataset. This same region is then visualised in Google Earth Engine, making it possible to select the suitable Sentinel-2 1C figures before and after the disaster. The decision is made not to apply a reduced statistic, such as the median, mean, sum or variance, but to show the first image available. By reducing an image collection to one embodiment, each pixel is composed, which creates an illusory figure and makes it non-applicable to change detection in a short period of time.

The biggest hurdle of finding the right set of Sentinel-2 1C figures is the cloud coverage and associated shadows, often present after a hurricane. A cloud filter and mask are not applied, since the cloud mask would output zero values for the covered pixels. The *Caladrius* model will not recognise this.

Manually, the dates before and after the disaster are selected within an iterative process. By zooming-in on the urban area's included in the xBD dataset, the cloud coverage is rated. The goal is to choose the clearest Sentinel-2 1C image closest to the disaster date.

The clipped images are saved in the GeoTIFF file format, referred to the EPSG:4326 frame and include the B4, B3, B2 bands, equal to the Red, Green, Blue bands, respectively.

In Table 3.4, the dates of the extracted Sentinel-2 1C images are given, per disaster. Additionally, Figure 3.6 visualises the dates on a scale, together with the disaster date and collected xBD imagery. It can be noticed that the xBD data is gathered at multiple timestamps and pre-event imagery originate from 1-3 years before the event. The latter is not advantageous, since this can cause differences when comparing pre- and post-event polygons, without it being damaged. For example, due to construction or dismantling of structures. The Sentinel-2 1C pre-event imagery is extracted in less than 2 months before the event, the post-event imagery within 30 days.

Table 3.4: Details of the extracted Sentinel-2 data included in this research

Disaster	Disaster Date	Before Date	After Date	Images
Hurricane Matthew	28-09-2016 10-10-2016	19-09-2016	09-10-2016	1/1
Hurricane Michael	07-10-2018 16-10-2018	29-06-2018	17-10-2018 28-10-2018	1/2
Tsunami Palu	18-09-2018	18-08-2018	27-09-2018	1/1
Tsunami Sunda Strait	22-12-2018	16-11-2018	05-01-2019 10-01-2019	1/2

In the specific case of the Hurricane Michael disaster, a trade-off is made regarding the cloud cover percentage and the rapid data collection. On the 17th of October in 2018, the first image available shows small thick clouds. Due to the polygon dense area of Panama City, many buildings became non-recognisable. The disfigurement is visible on a small scale when extracting the GeoTIFF's bounds of the individual images. In response to this, thirty per cent of the clipped Sentinel-2 1C photos are replaced with the representation of the 27th of October.

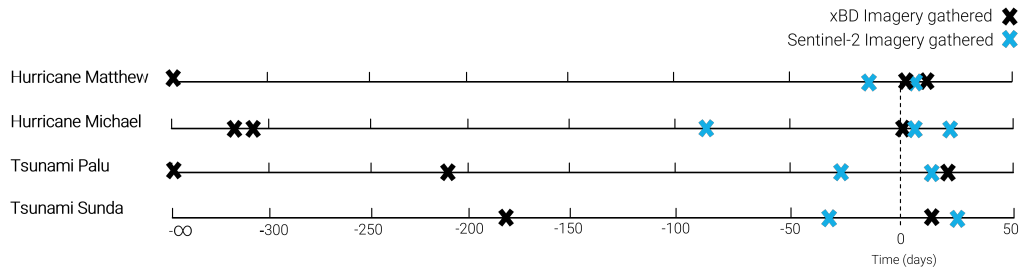


Figure 3.6: Timeline of Sentinel-2 1C and xBD data collection, pre- and post-event

The pictured area in xBD of the Sunda Strait Tsunami is scattered over the island of Java, including three cities: Bojong Kidul, Labuhan and Situpotong. The considerable distance between the cities makes it difficult to select a Sentinel-2 1C image without clouds above all urban areas. The first time-stamp, the 5th of January in 2019, is a good fit for the two cities located at the coast but cloudy within the mountains. Therefore, the second time stamp, the 10th of January, is chosen for the Situpotong region.

In Figure 3.7, an example image of the Hurricane Matthew disaster is pictured of the 0.5 and 10.0-meter resolution xBD dataset compared to the Sentinel-2 1C. This input data-point is an area covered with agriculture, afforestation and buildings. Similarities can be detected; the river in the left upper corner and the urban area on the right side. The most significant difference is the colour palette, by means of the high contrast within the Sentinel-2 1C image. The xBD representations contain white, brown and greenish colours, whereas the Sentinel-2 1C also outputs black and purple shades. The polygons will be zoomed in to assess the damage on the building level. Within the xBD 10-meter resolution and Sentinel-2 1C imagery, these will equal several pixels, without detail or a visible structure. This can cause problems, which will be examined in the neural network results, in chapter 5.



Figure 3.7: The comparison of an extracted Sentinel-2 1C image with the xBD 0.5 and 10.0-meter resolution visualisation (Example: Hurricane Matthew 03)

3.3. Synthetic Aperture Radar

Synthetic Aperture Radar visualisations are a non-literal imagery type and should be interpreted accurately, using a different approach than with optical imagery. Intensity ranges are visualised, depending on the amount of energy the SAR sensor measures. The energy can be observed 24-hours a day in all-weather conditions, deployable for various use-cases. More information on how SAR imagery is constructed can be found in subsection 2.1.2.

In this research, no high-resolution SAR data is integrated due to the non-available open sources which could be linked to the selected xBD disasters. Only low-resolution data is represented, originating from the satellite mission Copernicus Sentinel-1.

3.3.1. Sentinel-1

The Sentinel-1 mission operates day and night in a constellation of two polar-orbiting satellites, Sentinel-1A and Sentinel-1B. The radar functions with a C-band sensor linked to a wavelength and frequency of 7.5–3.8 cm and 4–8 GHz, respectively [16].

The goal of the first Copernicus mission is to cover the entire world bi-weekly and observe sea-ice zones, Europe's coastal zones, and shipping routes. The short revisit time and precise position information make the mission very successful to monitor land and investigate climate change.

Three products are openly available, consisting of Level-0, Level-1 and Level-2. The first represents the raw data, which should be decompressed and processed before use.

Level-1 includes a Single Look Complex (SLC) and Ground Range Detected (GRD) product. The SLC set is geo-referenced SAR data based on orbit and attitude data provided in zero-Doppler slant-range geometry. A single look in each dimension is created by the full transmit signal bandwidth with phase information. The GRD product only includes amplitude information and is based on the multi-look setup with a ground range projection using an Earth ellipsoid model. A square spatial resolution pixel is the result. The final product, Level-2, is helpful to detect Ocean Swell spectra, including Ocean Wind Fields and Surface Radial Velocities.

On the platform Google Earth Engine only Level-1 GRD data is available and, therefore, selected for this research. The image collection is pre-processed by removing thermal noise, radiometric calibration and terrain correction. Each scene within the collection has four instrument modes and four polarisation combinations; single-band and dual-band [14].

The specific images that are extracted to replicate the xBD set are filtered with the use of three settings. The first equals the instrument mode. The Sentinel-1 operates with four acquisition settings; Interferometric Wide swath (IW), Extra-Wide swath (EW), Strip Map (SM) and Wave (WV). The first three options are available in single and dual polarisations. The WV instrument mode can only be linked to the single polarisation, VV and HH. The different settings correspond to respective swath widths and spatial resolutions. The IW mode is the most commonly used and applicable, meeting many service requirements and covering all countries around the globe. The EW mode is primarily selected for the application of wide-area coastal monitoring, for example, ship traffic and sea-ice measuring. SM only provides visualisation of small islands and on request for emergency organisations. In this research, the IW swath is chosen, in which the beam is electronically steered in the azimuth direction for each burst, resulting in homogeneous image quality. The IW SLC product has one image per sub-swath and one per polarisation channel.

Secondly, the orbit property could be filtered. SAR satellites have two-orbit direction trajectories, from the North to the South Pole and the other way around, corresponding to a descending and ascending orbit, respectively. The same area is revisited by both orbit properties, but will output different visualisations because of the sight view of objects. When extracting the Sentinel-1 GRD imagery, there is tried to create an ascending and descending pair of the disasters. Unfortunately, this was not possible due to the accessibility of the data. The Sentinel-1 GRD does revisit most places on Earth within five days, but the data is not collected continuously.

At last, the reducer type could be selected. Again, the first image available, before and after the disaster, is extracted.

Without any hinder of clouds within the SAR visualisations, the selection process of image pairs is easy. The only essential requirement implies that the orbit property must be equal for both images, pre- and post-event. In Table 3.5, the dates of collected imagery are listed and in Figure 3.8 shown on a timeline. One problem did occur by extracting imagery for the Hurricane Matthew disaster, consisting of missing data after the event of half of the region. Two cities are pictured within the Hurricane Matthew dataset; Cayes and Jeremy. The city Cayes is not captured within four months after the disaster, with an ascending orbit direction. Furthermore, no descending data is openly available above Haiti in any of the years before the event. Therefore, there is chosen to include the data of the 17th of October in 2016, which only represents 62% of the original image pairs. The remaining pairs are ignored.

Table 3.5: Details of the extracted Sentinel-1 data included in this research

Disaster	Disaster Date	Before Date	After Date	Orbit direction	Images
Hurricane Matthew	28-09-2016 10-10-2016	02-07-2016	17-10-2016	Ascending	1/1
Hurricane Michael	07-10-2018 16-10-2018	27-09-2018	21-10-2018	Ascending	1/1
Tsunami Palu	18-09-2018	08-06-2018	06-10-2018	Descending	1/1
Tsunami Sunda Strait	22-12-2018	11-12-2018	29-12-2018	Descending	1/1

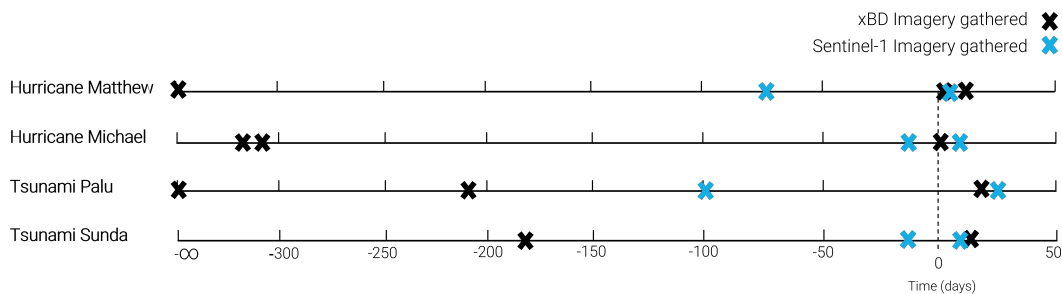


Figure 3.8: Timeline of Sentinel-1 GRD and xBD data collection, pre- and post-event

The images within this research are created using the RGB colour-composite of VH, VV and VH/VV polarisation channels. An example of Tsunami Palu is showcased in Figure 3.9. It is striking that the visualisation is entirely different compared to the xBD and Sentinel-2 1C images. However, back-scatter differences can be detected due to the damaged region in the lower part of the image. The pixels output fewer dark colours, representing a more smooth surface than before.

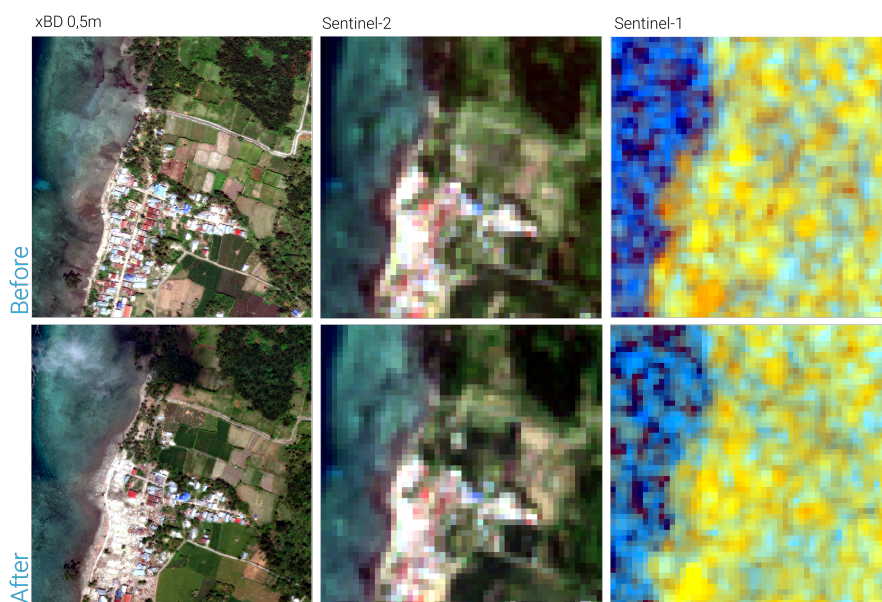


Figure 3.9: The comparison of an extracted Sentinel-1 GRC image with Sentinel-2 1C and the xBD 0.5 resolution visualisations (Example: Tsunami Palu 18)

4

Method

This chapter describes the methodology used to answer the research questions and fulfil the objective. In Figure 4.1, the pipeline of the Automatic Damage Assessment (ADA) is visualised, including the data input, output and models. The flow starts with the input pre- and post-event imagery pairs, equal to the input of the building extraction model. In subsection 4.1.1, all details are given of the functioning of this extraction step, creating polygons by using the provided building coordinates. Subsequently, all clipped polygon images are resampled to equal the dimension of 299x299 pixels to fit into the *Caladrius* model. In subsection 4.1.2, a description is given of the Convolutional Neural Network (CNN), which outputs a list of the ground-truth labels and predictions per polygon.

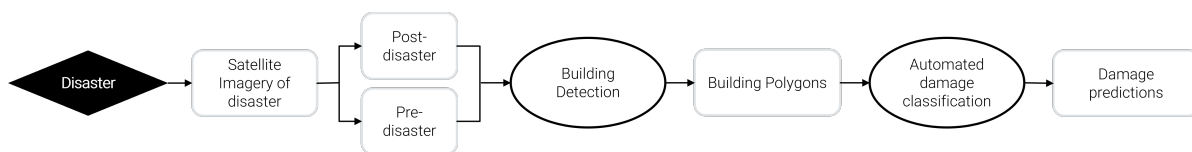


Figure 4.1: The pipeline of the data flow, building extraction and damage classification

To resolve the restrictions of the imbalanced xBD dataset, data pre-processing is applied and a balance is created, the process of which is described in section 4.2.

Next, the experiment set up is included in section 4.3, to exhibit the runs executed with the *Caladrius* model, consisting of single-, dual-, and cross-modal data scenarios. Finally, the metrics to evaluate the performance of the model are proposed.

4.1. Caladrius Model

The pipeline includes two models; to extract the buildings from the input images and to classify the polygons with a damage label. The CNN *Caladrius*, based on the Siamese architecture, executes the second step. Both are explained in this section.

4.1.1. Extract Buildings

Within the JSON files of the xBD dataset, longitude/latitude and x/y coordinates are provided of the building's corners to extract from the GeoTIFF and PNG images, respectively. Firstly, the shape formed by the coordinates is checked and resized to a rectangle, with the longest original side as the base. Next, an extra border is added around the polygon area. This extra space can give interesting information about the context and, therefore, the damage type, such as flooded water and debris. The extension factor, a percentage of the total polygon size, is varied within the research of Ritwik Gupta [22] to estimate the optimal factor. The extension factor is set at 5%, added at all four sides to minimise the risk of including neighbouring buildings within the polygon image.

Before saving the created polygons as images, the polygons are examined to meet two restrictions. Primary, the sum of all RGB values of the polygon requires to be positive. Secondly, a threshold of non zero pixels of 0.90 should be fulfilled.

Each individual saved polygon is renamed, and a text file is created, including a column with filenames and their corresponding ground-truth label.

In this research, the GeoTIFF version of the xBD dataset is selected, because of the replication possibility of Sentinel-2 and Sentinel-1. It is noticed that the longitude/latitude polygon footprints do not overlap the actual building area; however, no fixed offset is found and resolved. Therefore, the x/y coordinates are translated to new longitude/latitude values with the help of the equations below.

First, the longitude/latitude coordinates of the xBD imagery bounds are estimated and used to determine the size of the pixels ($w_{\lambda_{\text{pixel}}}$, $h_{\phi_{\text{pixel}}}$) expressed in degrees ($^{\circ}$). The minimum and maximum values of the coordinates are divided by the number of pixels (1024) in the height and width direction of the image shape, showed in Equation 4.1.

$$\begin{aligned} w_{\lambda_{\text{pixel}}} &= \frac{\lambda_{\text{max}} - \lambda_{\text{min}}}{1024} \\ h_{\phi_{\text{pixel}}} &= \frac{\phi_{\text{max}} - \phi_{\text{min}}}{1024} \end{aligned} \quad (4.1)$$

Next, the width and height values of the pixels ($w_{\lambda_{\text{pixel}}}$, $h_{\phi_{\text{pixel}}}$) are integrated in Equation 4.2, to compute the new longitude/latitude coordinates of the polygon bounds ($[\lambda_1, \lambda_2]$, $[\phi_1, \phi_2]$). The coordinates can be computed with a linear equation, using the minimum longitude and maximum latitude values (λ_{min} , ϕ_{max}) to represent the offset coefficient and the pixel size to characterize the multiplication coefficient.

$$\begin{aligned} [\lambda_1, \lambda_2] &= \lambda_{\text{min}} + w_{\lambda_{\text{pixel}}}[x_1, x_2] \\ [\phi_1, \phi_2] &= \phi_{\text{max}} - h_{\phi_{\text{pixel}}}[y_1, y_2] \end{aligned} \quad (4.2)$$

The minimum longitude (λ_{min}) and maximum latitude (ϕ_{max}) are selected, since the x/y coordinates $[x_1, x_2, y_1, y_2]$ are estimated from the origin located at the left upper corner.

All pixels touched by the polygons are extracted to create the building images. This results in bigger polygon images for lower resolution datasets, the effect is shown in Figure 4.2. In addition, the extension factor is, in specific cases, enlarged by applying the model on the lower resolution datasets. The extension factor is based on the difference between the size of the building and the resolution setting, plus the original 5% on all sides. This design choice is implemented to ensure polygons exist of more than 1 pixel when the building is smaller than the spatial resolution.

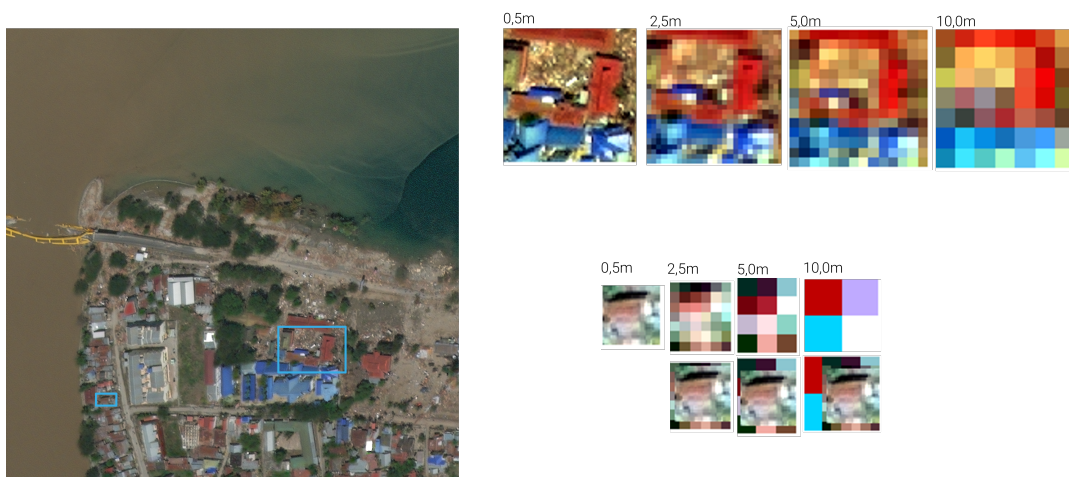


Figure 4.2: Visualisation of the effect of down-sampling on polygon scale (Example: Tsunami Palu 138)

4.1.2. Convolutional Neural Network

The CNN model named *Caladrius* was created in 2019 by *510 - an initiative of the Red Cross Netherlands*, with the mission to minimise the time taken by aid organisations to reach the victims of the disaster. The damage detection tool is openly available on Github¹, where iterations and improvements are constantly made to achieve reliable predictions and reach optimal performance to implement the tool in real-life situations.

The architecture of the CNN model, shown in Figure 4.3, is inspired by the Siamese architecture, including two identical networks with separate pre- and post-event imagery as input data. The Siamese architecture is designed to discriminate characteristics of inputs and uses the same weights and parameters within the twin subnetworks [39]. The *Caladrius* Network differs on this aspect since it updates the weights in parallel due to the distinct features of the pre- and post-event input data, which should be learned separately. In addition, the outcome is not solely established by the differences of the output vectors of the linear networks, but is assembled by implementing three fully connected layers.

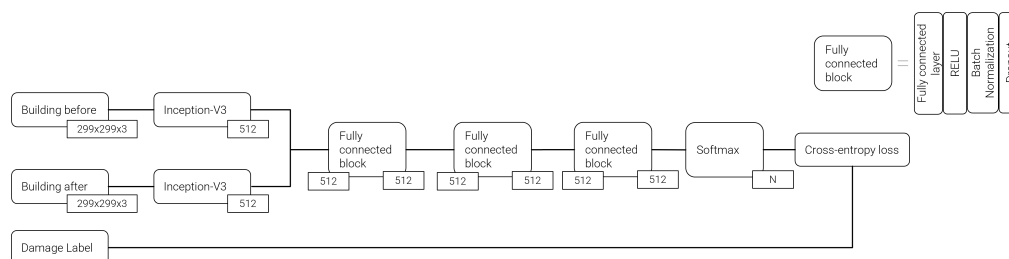


Figure 4.3: Architecture of the *Caladrius* model. The number in the squares on the left side of a block represents the input size of each block, the number on the right side indicates the output size. N refers to the number of damage classes [58].

The fully connected layers consist of a ReLU activation function, batch normalisation and dropout, all explained in detail in section 2.2. In the training phase of the neural network, the final fully connected layer is linked directly to the cross-entropy loss computation, which is used to examine the model's performance and is optimised by altering the weight values. The Softmax finalises the pipeline in the test phase and produces the prediction values between 0 and 1 for every damage label.

The settings of the model's hyper-parameters are tested with the Adam optimiser in previous research [58], resulting in a batch size of 32 and a learning rate of 0.0001. The augmentation is applied at every training batch with a scheme containing flipping, rotation and translation. The transformations are different for every epoch.

Inception-V3

The two separate CNN models follow the Inception-V3 architecture, successfully applied to a larger variety of computer vision tasks. The Inception architecture family is designed to solve two issues of neural networks. The first arises due to a large size variation of the target within the input imagery, making the search difficult for a fitted kernel size to overlap the information. A larger kernel detects information globally distributed, and a smaller kernel is preferred to learn small details. The second issue concerns the computationally expensive convolutions existing in deep neural networks. The Inception architecture is built with smart design choices to reduce the computer time and storage required.

The general Inception module, also called the naive version, resolves the first issue by including filters on the same level with multiple sizes, which creates a wider network, visualised in Figure 4.4. Next to the convolution layers, a max-pooling layer is added, the functioning of which is explained in subsection 2.2.1.

To reduce dimensions and at the same time computational intensity, extra 1x1 convolution blocks are placed before the larger sized convolution blocks. This action does limit the number of channels and therefore decreases the number of parameters. Simultaneously, the width and height of the input imagery are not affected.

¹<https://github.com/rodekruis/caladrius>

The first Inception version was designed during the ImageNet Large-Scale Visual Recognition Challenge in 2014, and further research evaluated the architecture, resulting in already five versions [51].

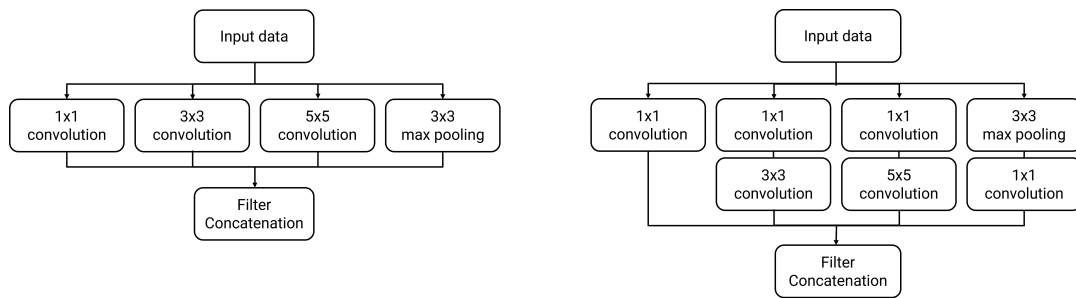


Figure 4.4: a) Naive inception module, b) Inception module with dimension reduction [54]

The Inception-V3 architecture is based on three building blocks (Figure 4.5), which are iterations of the naive Inception module. It improves accuracy and uses smart factorisation methods to be more efficient in computational complexity. Within block A the convolutional layer of the dimension 5x5 is transformed into two series blocks of a 3x3 size. This multi-layer network uses 28% less parameters with the same input size and output depth, due to the weight sharing between adjacent tiles.

Block B introduces an asymmetric combination of 1×7 convolutions followed by 7×1 convolutions. This factorization does perform well on medium grid-sizes, ranging from 12 to 20, where computational cost saving increases as the grid grows.

At last, block C factorises the 3x3 convolution layers by creating an asymmetric parallel combination of 1×3 and 3×1 convolutions with the same receptive field, reducing 33% of the parameters while maintaining the existing performance.

The spatial dimension of the input is reduced within the inception Dim, while increasing its number of channels [41].

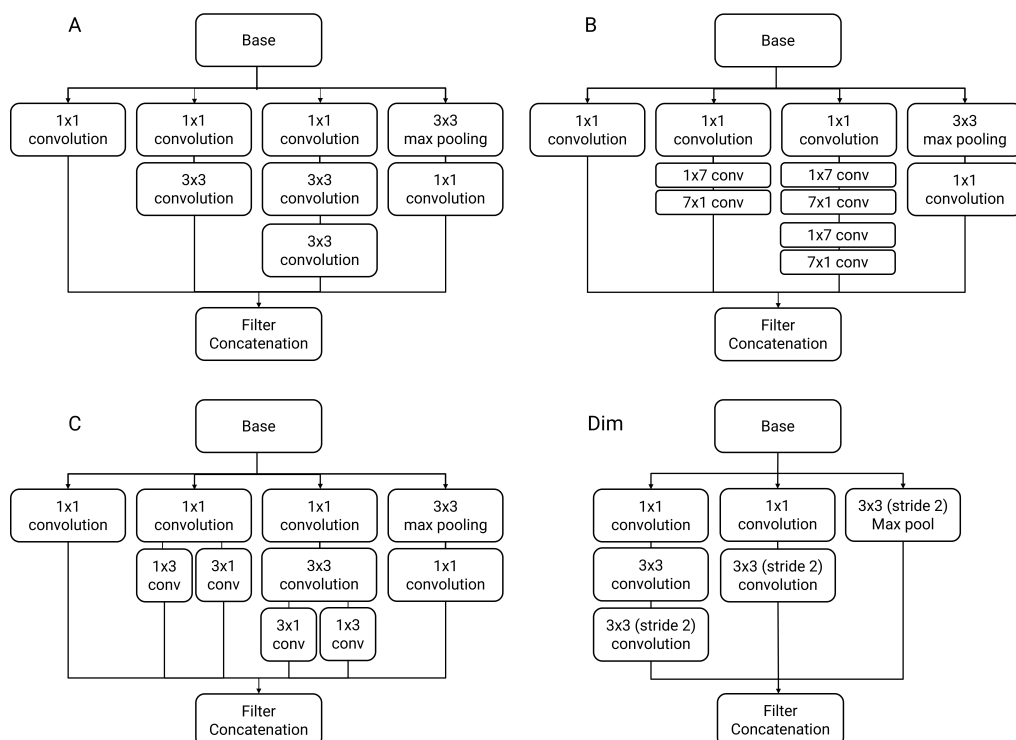


Figure 4.5: The three blocks, A, B and C of the Inception V3 and the [54]

In Table 4.1, the original architecture of Inception-V3 is listed, including 6 standard convolutional layers and 12 inception layers.

In the training phase, the model is extended with an auxiliary classifier that injects gradients between the last layers of the inception block B and the inception dim layer. This extra classifier prevents the problem of the vanishing gradient. In addition, a dropout layer is implemented between the output of the average pooling layer and the fully connected layer.

The Inception-V3 is trained on a million images from 1000 classes of the ImageNet dataset, learning general features of various animals and objects. This transfer learning process sets the initial weights of the model and results in a significant decrease in training time and the size of the dataset required [51].

Table 4.1: Architecture of Inception-V3 [54]

Type	Patch size/stride or remarks	Input size
conv	$3 \times 3/2$	$299 \times 299 \times 3$
2× conv	$3 \times 3/1$	$299 \times 299 \times 3$
max pool	$3 \times 3/2$	$147 \times 147 \times 64$
conv	$3 \times 3/1$	$147 \times 147 \times 32$
conv	$3 \times 3/2$	$147 \times 147 \times 32$
conv	$3 \times 3/1$	$147 \times 147 \times 32$
3× Inception A	Figure 4.5	$35 \times 35 \times 288$
3× Inception dim	Figure 4.5	
5× Inception B	Figure 4.5	$8 \times 8 \times 1280$
3× Inception dim	Figure 4.5	
2× Inception C	Figure 4.5	$8 \times 8 \times 1280$
Avg pool	8×8	$8 \times 8 \times 2048$
Fully connected + Softmax		$1 \times 1 \times 2048$

Although the Inception-V3 is adopted in the Siamese inspired *Caladrius* network, a minor adjustment was made. The output size of the final fully connected layer in the Inception module is reduced from 2048 to 512 features. The reason for the downsizing is the small dimension of the polygon input images of 299×299 pixels, creating a high risk of overfitting. In addition, three fully connected layers are added on top of the Inception-V3 architectures to link the twin sub-networks, combine the learned insights and create a deeper network, to finally predict the damage of the polygon images.

In this research, no significant modifications to the *Caladrius* model are required by running the multiple experiments, including different resolutions and types of input imagery. The three colour bands of the input polygon images are all normalised, solving the varying ranges of the RGB values in the xBD, Sentinel-2 and Sentinel-1 datasets.

Due to the non-typical visualisation of the Synthetic Aperture Radar data, the first layer of the Inception-V3 is unfreezed in this specific experiment, undoing the transfer learning. The features of the ImageNet dataset will not match the characteristics of the SAR imagery.

4.2. Data Pre-processing

The down-sampling method is discussed in this section, to elaborate on the creation of the 2.5, 5.0, 10.0-meter xBD datasets. Subsequently, the images of the xBD, Sentinel-2 and Sentinel-1 datasets require a few pre-processing steps before being put into the *Caladrius* model. Firstly, the transformation of the GeoTIFF to the PNG file version should be made to read the polygon images properly within the CNN. Next, the imbalanced datasets are resampled to create similar-sized damage classes, elaborated in subsection 4.2.2.

4.2.1. Down-sampling

The average resampling technique is selected, to shrink the input rasters. The method computes the mean of all original pixel values X_{RGB_i} included within the new pixel area. This method makes sure no information is lost, compared to other approaches. In Equation 4.3, the relation is given of the assigned value X_{RGB} to each band; Red, Green and Blue.

$$X_{RGB} = \frac{\sum_{i=1}^n X_{RGB_i}}{n} \quad (4.3)$$

4.2.2. Resampling

A high-class imbalance is detected within the labelled xBD dataset; substantially more images represent the *no-damage* buildings compared to the remaining damage types. An imbalance can have a negative influence on the performance of classification [34]. Models tend to classify input data with the most common label, due to the best chance of being correct. Nevertheless, minority labels are often and in this specific use case the critical targets to identify.

The imbalance can be addressed with several methods, subdivided into two categories. The first category is based on data level methods operating on the training dataset by changing its class distribution. The second category covers classifier-level methods that keep the training dataset unchanged and adjust the loss functions.

Over-sampling and under-sampling belong to the first category, replicating and removing randomly selected samples of minority and majority classes, respectively. The factor of the over- and under-sampling can be specified beforehand or chosen to re-balance the classes with an even distribution. In both cases, the training set is resized. The method of over-sampling has emerged as dominant in most analysed scenarios [10]. Nevertheless, the risk of overfitting can arise with a high replication factor. Characteristics of the same images within the minority class will be learned, resulting in an inaccurate performance on unseen data, defeating the purpose of the model. The significant disadvantage of applying the under-sampling method is discarding a portion of available data.

The second category includes thresholding and cost-sensitive learning. The first adjusts the decision threshold and changes the output class probabilities. The optimal threshold is found by trial and error, using the ROC curve to examine performance differences [9]. The last method, cost-sensitive learning, assigns different costs to miss-classifications. The learning rate is modified such that higher cost examples contribute more to the update of weights.

Tinka Valentijn et al. [58] researched the effect of the cost-sensitive learning on the xBD dataset, no significant effect was shown, based on the Macro F1-scores and recall values of the four damage classes. Meanwhile, resampling, a combination of over- and under-sampling, did show improvements and left the resampled train-data size unchanged.

The original ratio of classes within Hurricane Matthew, Michael, Tsunami Palu and Sunda Strait dataset equals 69/16/5/10. In this research, there is chosen to resample the dataset to resolve the imbalance, with a combination of over- and under-sampling. The order of the resampling steps and splitting the dataset in training, test, and validation portions is elaborated below. Furthermore, the resulting distribution of the classes within the training set is visualised in Figure 4.6.

1. Firstly, the majority class is under-sampled to balance the *no-damage* buildings originated from America and Asia. The set of American *no-damage* buildings consists of 26,649 polygons compared to the 60,680 Asian polygons. An equilibrium is created between both polygon numbers $x_{\text{America}_1} = x_{\text{Asia}_1}$, reducing the total dataset with 34,031 data-points. Two reasons thrive this alteration, primarily to limit the bias towards learning characteristics of constructions related to environmental and location-specific circumstances. Secondly, to prevent a prohibitive replication factor required to over-sample the minority class to create a balanced dataset.
2. The total dataset is randomly split into a train-, test-, and validation dataset to realise a ratio of 80/10/10, respectively. Every dataset in this research is created with the same random seed.
3. The third step of the resampling process includes the combination of over- and under-sampling to retain the dataset's size. A new ratio of the damage type classes is established to be 35/25/20/20, to create a balanced training dataset. No equally distributed ratio is specified to imitate a real-life scenario. Accordingly, the replication factors are equal to 0.6095/1.1306/2.6956/1.5335.

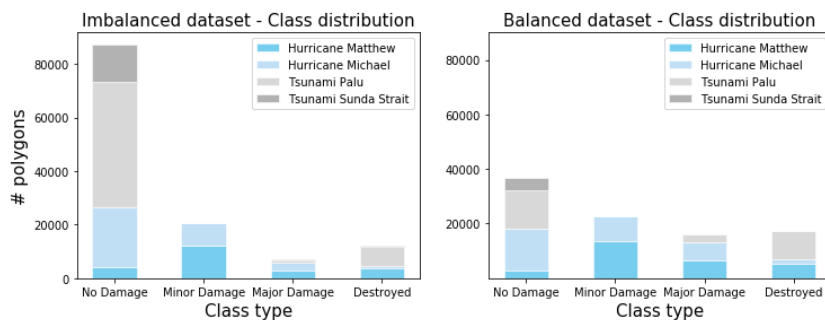


Figure 4.6: Training dataset size and composition, before and after re-sampling

4.3. Experiments

Eighteen experiments are set up and performed, subdivided into three modality scenarios; single-mode, dual-mode and cross-mode to address all possible data combinations for damage mapping. In change-detection research, is it beneficial to use pre- and post-event images taken under almost identical acquisition conditions to train and test the model, represented by single-mode experiments. However, it is unpredictable if this is feasible within a restricted time period. The dual-mode and cross-mode experiments investigate the deployability of the *Caladrius* model inputting a mix of non-similar data.

4.3.1. Single-Mode

The single-mode scenarios, in Table 5.4, are designed to answer the first and third research questions, regarding the relationship between the optical imagery resolution and imagery type with respect to the true prediction rate of the *Caladrius* model, respectively. The original and down-sampled xBD datasets are inputted to train and test the model, representing the first four experiments.

Subsequently, the 5th experiment using Sentinel-2 data is created to estimate differences between satellite sources, despite having equal resolution and optical sensors. Additionally, this experiment reviews the capability of replicating a dataset. Last, the functioning of the *Caladrius* model is examined inputting SAR imagery of the Sentinel-1 dataset. The results of the last three experiments are interesting to compare, to identify the differences between the sensor types and determine if the *Caladrius* model is able to function in all-weather circumstances.

Table 4.2: The six designed experiments including xBD, Sentinel-2 and Sentinel-1 imagery in Single-Mode data scenarios

		Optical				Sentinel-2	SAR
		xBD					Sentinel-1
Res [m]		0.5	2.5	5.0	10.0	10.0	10.0
1	Pre-event	x					
	Post-event	x					
2	Pre-event		x				
	Post-event		x				
3	Pre-event			x			
	Post-event			x			
4	Pre-event				x		
	Post-event				x		
5	Pre-event					x	
	Post-event					x	
6	Pre-event						x
	Post-event						x

All six experiments, listed in Table 5.4, are repeated by binary-classification between the labels *no-damage* and *damage*. The threshold is determined by a qualitative analysis of the visualisations of the polygons and drawn between the *minor* and *major-damage* classes. The binary distinction can indicate which regions are hit instead of specifics on the damage type, which might suffice operational use. This application can be interesting for low-resolution data due to the loss of detail in the images.

The approach of binary classification equals training and testing on multi-class labels and grouping the

predictions into binary labels. In previous research, this approach outperformed training and testing on binary labels, expressed in AUC values [58].

4.3.2. Dual-Mode

In the six dual-mode experiments, the input data of the training phase consists of non-similar characteristics compared to the test data. These experiments represent the scenario in which the training process can not be repeated due to time restrictions, and the available data of the affected region has a lower resolution than the train dataset. Training the model can take days, which is not permitted to serve rapid relief actions. Therefore, this process is performed in advance. However, when the accessible test data does not suit the trained model, *Caladrius* is unserviceable. It is essential to understand the reliability of the damage predictions in these situations and how sensitivity the model reacts to the mix of data. All six experiments, given in Table 4.3, only focuses on the different resolution xBD datasets.

Table 4.3: The six designed experiments including xBD imagery in Dual-Mode data scenarios

		Optical xBD					Sentinel-2	SAR Sentinel-1
Res [m]		0.5	2.5	5.0	10.0	10.0	10.0	
7	Train-set	x						
	Test-set		x					
8	Train-set	x						
	Test-set			x				
9	Train-set	x						
	Test-set				x			
10	Train-set		x					
	Test-set			x				
11	Train-set		x					
	Test-set				x			
12	Train-set			x				
	Test-set				x			

4.3.3. Cross-Mode

Again, six experiments are defined to consider the cross-mode data scenario, listed in Table 4.4. The cross-mode configurations consist of optical high-resolution pre-event imagery and low-resolution post-event imagery, to train and test on. After the disaster has struck, it is crucial to collect the post-event imagery as soon as possible to visualise the damage affected. The date of pre-event imagery is less restricted by the time period of the disaster, creating flexibility to select clear high-resolution data. For this reason, it may happen that pre- and post-event imagery are acquired by non-identical characteristics. To note, it is important to train the model on equal data properties as the test-set.

Table 4.4: The six designed experiments including xBD imagery in Cross-Mode data scenarios

		Optical xBD					Sentinel-2	SAR Sentinel-1
Res [m]		0.5	2.5	5.0	10.0	10.0	10.0	
13	Pre-event	x						
	Post-event		x					
14	Pre-event	x						
	Post-event			x				
15	Pre-event	x						
	Post-event				x			
16	Pre-event		x					
	Post-event			x				
17	Pre-event		x					
	Post-event				x			
18	Pre-event			x				
	Post-event				x			

4.4. Performance Metrics

To measure and compare the performances of the experiments executed with the *Caladrius* model, it is necessary to define when one outcome is superior to the other. Different Machine Learning models are evaluated with varying performance metrics, depending on their type and goal. The most common metrics can be subdivided into threshold and rank types [30]. Threshold metric types predict a sample with a positive or negative value with respect to the threshold level. The distance of the prediction score compared to the threshold can be neglected. The rank metric type depends on the ordering of the samples with corresponding prediction values.

Three metrics are highlighted of the first type, including the accuracy, Cohen's Kappa coefficient and F1-scores. The accuracy is equal to the ratio between the number of correct and total predictions, given in Equation 4.4. This is not a preferred measure when working with an imbalanced dataset, since a bad model will tend to classify all samples with the majority class, resulting in a high accuracy score. In this research, the accuracy score does not correctly represent the performance of the classification model.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (4.4)$$

The Cohen's Kappa coefficient, given in Equation 4.5, measures the inter-rater reliability, corrected by the agreement of chance. The index can be computed using the confusion matrix and is applicable to imbalanced datasets. The numerator of the ratio is based on the difference between the observed overall accuracy P_{Obs} and the overall accuracy, which will be reached by chance P_{Chance} . The denominator is equal to the maximum value of this difference. The Kappa value 1 represents a good performing model, and a zero indicates a random classification. In theory, a negative value can be scored, when the performance is even lower than obtained by a random guess [5]. The Cohen's Kappa metric has the downside of outputting a higher coefficient for balanced datasets than imbalanced datasets. In this research, the experiments all have the same class distribution, created by the resampling steps, elaborated in subsection 4.2.2. Therefore, the Cohen's Kappa coefficient can be used to compare the varying resolution datasets, using the original xBD dataset as the base. The Kappa coefficient can especially be interesting to examine the lower resolution data and determine if random classification occurs or characteristics of the images are still learned.

$$K = \frac{P_{Obs} - P_{Chance}}{1 - P_{Chance}} \quad (4.5)$$

The F1-score is interpreted as the weighted average of the precision P and recall R values. The precision represents the True Positive examples divided by total positive classified examples. The recall, also called sensitivity, is the fraction of True Positives divided by the sum of True Positives and False Negatives. Within the multi-class experiments, the Harmonic F1-score is computed per class. The true class is equal to the class the calculation is related to, and the negative class refers to the summation of the remaining classes. The individual Harmonic $F1_i$ -scores show if a class under-performs and can be used to calculate the overall Weighted F1-score of the model. The Weighted F1-score is sensitive to an imbalance within the dataset, since it is based on the sum of the number of images belonging to the specific classes n_i times the linked Harmonic $F1_i$ -scores, divided by the total number of images within the test dataset n .

To treat every class equally, the Macro F1-score is the best suited performance measure, giving the same importance to each class. All Harmonic $F1_i$ -scores are added and divided by the total included classes N , 4 in multi-class and 2 in the binary-class setting.

In previous research [58], the Harmonic F1-score of the total dataset, also called a Pythagorean mean, is introduced to estimate the true prediction rate of the *Caladrius* model. To compare results, this specific metric is together with the Macro F1-score included in this research. The Harmonic F1-score can be computed by dividing the number of classes by the sum of reciprocals of each Harmonic $F1_i$ score.

In Equation 4.6, all relations of the F1-score computations are stated.

$$\begin{aligned}
 P &= \frac{TP}{TP + FP} \\
 R &= \frac{TP}{TP + FN} \\
 \text{Harmonic } F1_i &= 2 \times \frac{P \times R}{P + R} = \frac{TP}{TP + 0.5(FP + FN)} \\
 \text{Weighted } F1 &= \frac{1}{n} \sum_{i=1,2,3,4}^n n_i \times \text{Harmonic } F1_i \\
 \text{Macro } F1 &= \frac{1}{N} \sum_{i=1,2,3,4}^N \text{Harmonic } F1_i \\
 \text{Harmonic } F1 &= N \frac{1}{\sum_{i=1,2,3,4}^N \text{Harmonic } F1_i^{-1}}
 \end{aligned} \tag{4.6}$$

The metrics belonging to the rank metric type include the Area Under the ROC Curve (AUC-ROC) and Area under Precision-Recall curve (AUC-PR). In this research, the AUC-ROC is better suited, since it is less sensitive to imbalance. The AUC of the curves is only applicable to measure performance for the binary classification setting.

The AUC measures how well predictions are ranked and examines the quality of the model's predictions irrespective of what classification threshold is chosen. The vertical axis of the ROC plot represents the True Positive rate and the horizontal axis ranges the False Positives.

The metrics results, computed on the test set, will not be used to rank the performances of the *Caladrius*, but will provide insights into how reliable the predictions are in operational situations after a natural disaster. In addition, the miss-classifications are reasoned by qualitative analysis and interpreting confusion matrices.

5

Results

In this chapter, the results of the performed experiments are showcased and discussed. Firstly, the general performance of the *Caladrius* model is reviewed and verified with previous research, in section 5.1. Next, the results of the single-, dual- and cross-mode experiments are investigated in section 5.2, to elucidate the relation between the resolution of optical imagery and the true prediction rate of the Convolutional Neural Network (CNN). Multiple reasons for miss-classifications are discussed to recommend improvements in further research. In section 5.3, the effect of inputting Synthetic Aperture Radar (SAR) data into the *Caladrius* model is examined.

5.1. General Performance

To verify the *Caladrius* model and compare performances with previous research [58], the model is separately trained and tested on the selected disasters; Hurricane Matthew, Hurricane Matthew, Tsunami Palu and Tsunami Sunda. In Table 5.1, the Macro and Harmonic F1-scores, together with the recall values per damage type, are listed. The test runs are paramount to identify if the true prediction rate is similar and create a baseline for further alterations to the *Caladrius* model. All runs are trained for 100 epochs, and the datasets are still intact without applying the resampling steps.

In Table 5.1, a minor inconsistency between metric scores can be detected, explained by three possible reasons. Firstly, the runs are operated using other computational resources; GPU and processing card. This can cause mathematical rounding differences. Secondly, the *Caladrius* model is previously trained and tested on xBD PNG image files. In this research, the GeoTIFF version is selected, as elaborated in chapter 3. Last, the random seed could be non-identical between comparable runs, assigning other images to the train-, test-, and validation-dataset. This effect is translated in the *minor-damage* recall value, linked to disaster Tsunami Palu and Sunda Strait. In previous research, both scores were equal to 1.00. In the verification runs of this research, no samples were included of this damage type; therefore, the recall value outputs 0.00.

Table 5.1: Results of multi-classification per disaster (epoch=100), sorted by the F1 scores and recall value per damage type; No = No-damage, Min.= Minor-damage, Maj.=Major-damage and Des.=Destroyed

	F1		Recall			
	Harm.	Macro	No	Min	Maj	Des
Hurricane Matthew [58]	0.54	0.58	0.38	0.86	0.32	0.68
Hurricane Matthew	0.49	0.55	0.26	0.87	0.29	0.74
Hurricane Michael [58]	0.50	0.54	0.91	0.40	0.41	0.35
Hurricane Michael	0.49	0.54	0.94	0.47	0.36	0.36
Tsunami Palu [58]	0.00	0.70	0.98	1.00	0.00	0.82
Tsunami Palu	0.00	0.43	0.99	0.00	0.00	0.67
Tsunami Sunda [58]	0.00	0.33	1.00	1.00	0.00	0.29
Tsunami Sunda	0.00	0.40	1.00	0.00	0.00	0.44

To reduce the training time of the model, the loss and score functions are reviewed in Figure 5.1 to estimate the required number of epochs. The running time of the verification experiments ranged between 12 hours and 3 days, depending on the number of polygons included in the disaster dataset. In this research, the data of the four datasets are combined, enlarging the training phase to 8 days.

The train- and validation loss curve indicates how well the *Caladrius* model is learning and if the model is able to generalise predictions on a hold-out dataset. In addition, the plot can show signs of under- and overfitting characteristics, both effects are preferably avoided [7].

The cross-entropy loss is computed every epoch and consulted to update the weights within the CNN. In Figure 5.1, heavy spikes are visible in the validation loss trend-line, the unrepresentative train dataset can cause these outliers. The imbalanced dataset provides insufficient information to learn the minority class features.

The loss value starts at 0.90 and converges to 0.45. In the following experiments, the *Caladrius* model is trained on 50 epochs due to time restrictions and the minimal improvement in the remaining epochs.

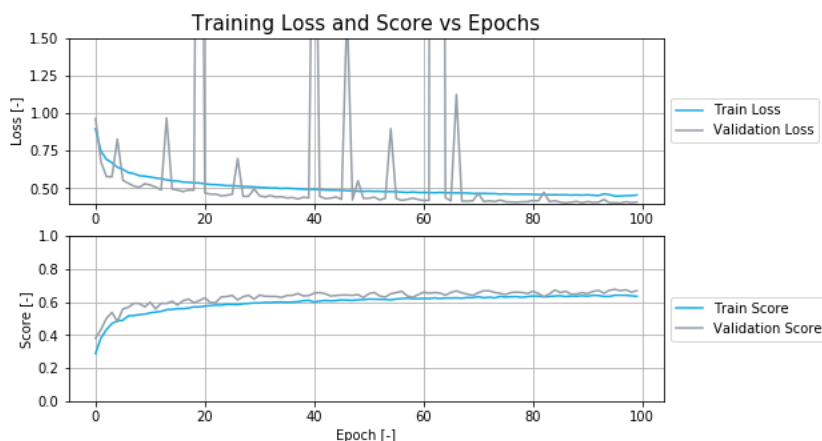


Figure 5.1: The loss and F1-score plots, trained on the xBD dataset, 100 epochs

5.1.1. Effect of Class Imbalance

The class imbalance causes many miss-classifications, this is harmful since the minority classes are the most important to be predicted correctly, equal to the *major-damage* and *destroyed* buildings. To limit the reduction in performance due to the class distribution, resampling is applied, as explained in subsection 4.2.2. In Table 5.2, the effect of combining over- and under-sampling can be detected. The Macro F1-score does improve significantly. The recall values show variable results. The classes, which contain damage characteristics, do reach better prediction rates.

Table 5.2: Results of multi-classification of the imbalanced and balanced dataset (epoch=50), sorted by the F1 scores and recall value per damage type

	F1		Recall			
	Harm.	Macro	No	Min	Maj	Des
Imbalanced	0.66	0.55	0.94	0.69	0.21	0.69
Balanced	0.63	0.68	0.86	0.68	0.42	0.76

To examine the results further, confusion matrices are created, including the number of predicted polygons and normalised scores per true label, in Figure 5.2. Preferably the diagonal is highlighted, indicating the number of True Positives plus the corresponding recall value.

Comparing both matrices, it can be noticed that fewer polygons are assigned to the *no-damage* type by the *Caladrius* model, trained on the balanced dataset. The miss-classified *no-damage* polygons are shifted towards the neighbouring classes of the true label. This is improved behaviour because the model is learning damage features instead of classifying most samples with the majority class. Consequently, the recall value and True Positives of the *no-damage* class did reduce.

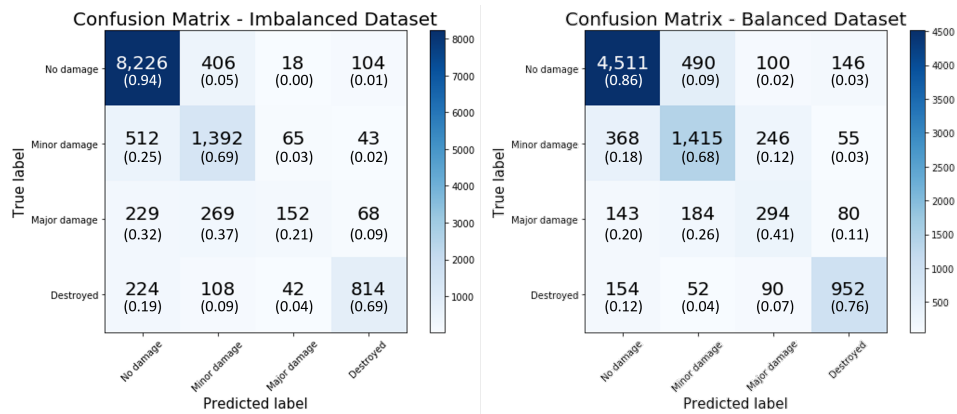


Figure 5.2: The confusion matrices of the imbalanced and balanced dataset, trained on the xBD dataset

In Figure 5.3, the training and validation process are showcased of the imbalanced and balanced dataset. The training loss of the balanced dataset starts and converges towards a higher value. This can be explained by the dataset's lower number of *no-damage* buildings. This specific class represented the majority within the imbalanced dataset, resulting in small loss values and high accuracy, despite the low overall performance expressed in Macro F1-score.

The validation loss plot presents a smoother trend line in the balanced dataset, without any outliers. Nevertheless, an offset occurs between the training and validation loss function. This effect arises because of the different class distributions within the training and validation dataset, no resampling is applied to the latter. Additionally, validation loss functions are often lower than training loss functions due to the regularisation application during the training phase. Also, the training loss is estimated during each epoch, while validation loss is measured after each epoch [7].

The overall score of the training and validation process do improve and converge towards a higher value due to the resampling steps, this behaviour is desired.

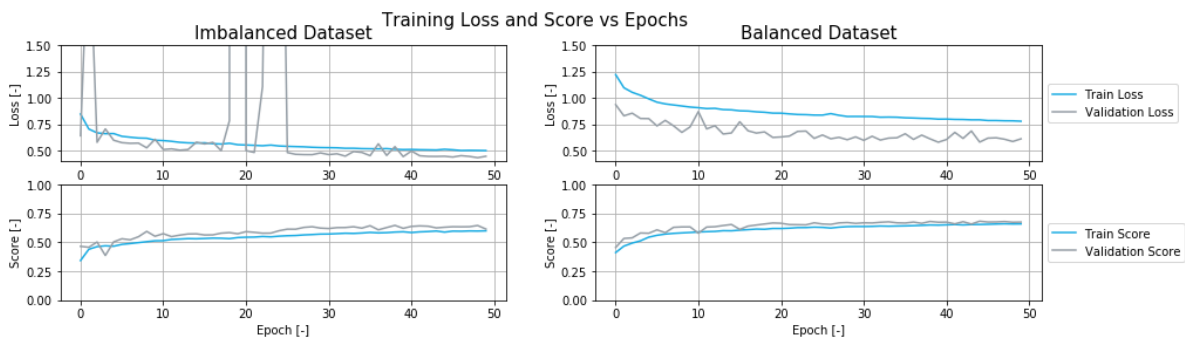


Figure 5.3: The loss and F1-score plots, trained on the imbalanced and balanced xBD dataset, 50 epochs

5.1.2. Reasoning of Miss-classifications

Evaluating why the *Caladrius* model is miss-classifying specific polygons is essential to detect if the flaw is coming from the CNN, the quality of the images or labels. Subsequently, it is crucial to resolve the issue and prevent false predictions.

Miss-classifications due to the Quality of the xBD Dataset

The imagery of the xBD is of high quality, however, shortcomings are detected. The xBD data is collected with optical sensors, which creates the possibility of cloud covered figures. In the composed dataset of this specific research, clouds are also present. Especially, the Tsunami Sunda Strait images show this effect, pre- and post-event. It is observed that almost all post-event polygons, covered with clouds, are on default labelled with the *no-damage* type. This can be explained by the human annotators, who interpreted the labels based on these images.

However, something interesting is found, as many of these covered polygons are classified by the model as *destroyed*. The similarity between the visualisations of a building that is wiped away or covered with clouds can confuse. In Figure 5.4, an example is given of a cloudy region and of a destroyed coastal town. In both polygon images, no construction is visible anymore after the Tsunami has hit. Therefore, the *Caladrius* model did classify both buildings as *destroyed*, in which the cloud covered polygon was labelled with *no-damage*.

For further research, it is essential that polygon images covered with clouds are removed, not exclusively to reduce miss-classification but also to prevent the *Caladrius* model from learning characteristics of cloudy pixels linked to the default class type *no-damage*.

The second shortcoming of the xBD dataset is the considerable time period in between the pre-imagery collection and the actual disaster impact. The data of Hurricane Matthew was acquired three years in advance, discarding constantly evolving constructions within urban areas. Consequently, multiple buildings visible in post-imagery are not recognised due to the non-existence in the pre-imagery. Furthermore, situations occur where buildings are labelled with the *no-damage* type, but predicted as *destroyed*. In the 3 years time, these buildings were rebuilt or reconstructed, causing significant changes, without being damaged. Examples of miss-classified polygons due to reconstruction alterations are visible in Figure 5.5.



Figure 5.4: Left: visualisation of image including cloud coverage (Example: Tsunami Sunda Strait 82-84 (labeled: 3, prediction: 3)) Right: visualisation of image including a destroyed village (Example: Tsunami Palu 20-37 (labeled: 0, prediction: 3))

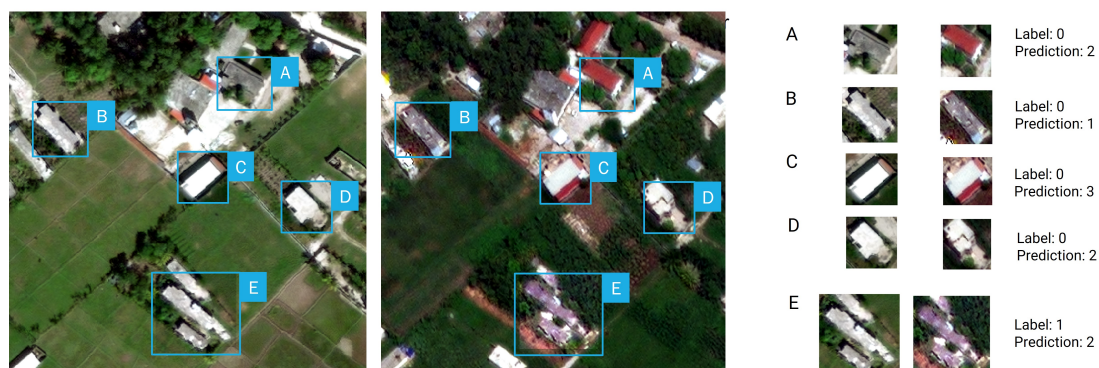


Figure 5.5: Visualisation of miss-classified polygons due to the acquisition of imagery three years in advance of the disaster (Example: Hurricane Matthew)

Thirdly, the satellite parameters vary between images within pre- and post-event pairs. The provided metadata in the JSON-files consists of the off-nadir-, target azimuth-, sun azimuth- and sun elevation-angles, defining the position of the satellite and Sun with respect to the Earth. Non-similar parameters can cause different illumination, side views and visible detail within the images. In Figure 5.6, the variety is showcased in a box plot per disaster. The more significant the difference between pre- and post-event the more effect it could have on the prediction accuracy. In previous research [58], it was found that the Sun azimuth and elevation do influence the AUC score. When the lighting is too bright, too dark or when the difference between the image pair is too large, the performance of the *Caladrius* classification degrades.

In addition, a big difference in off-nadir angle between pre- and post-event imagery causes non-accurate polygon outlines. The building bounds were estimated on the pre-event data, as explained in subsection 3.1.1. The safety border partly compensates for the misalignment; however, this does not resolve the shift in some samples. Consequently, only partial information of the building is given to the *Caladrius* model.

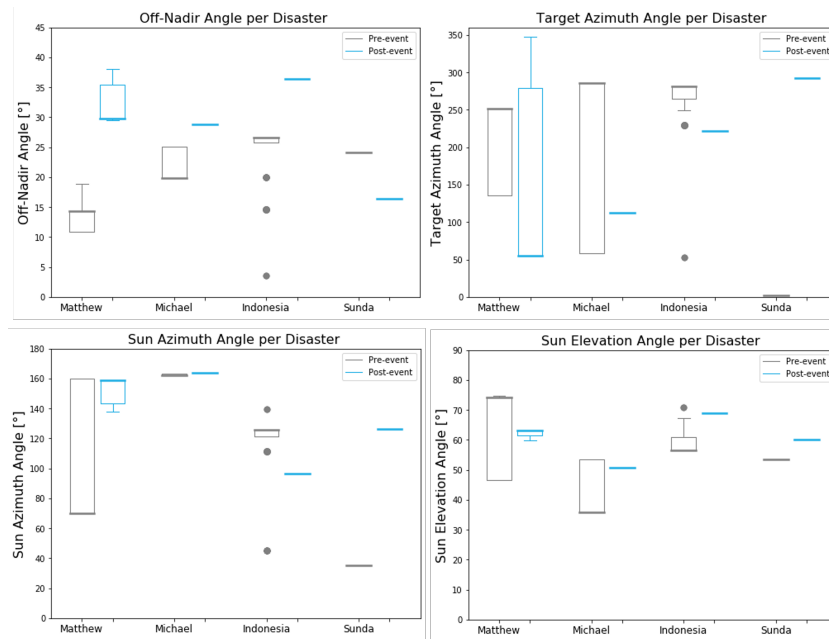


Figure 5.6: Box-plot of the satellite acquisition characteristics per disaster, including the median, standard variation and outliers

Lastly, the correctness of the manual labelled buildings is debatable. The boundaries between the neighbouring classes are specified using the Joint Damage Scale (Table 3.1), but are subjective. The distinction between the *no-damage* and *destroyed* labels is clear, however, the visual difference between *minor-* and *major-damage* buildings are sometimes hard to recognise in qualitative analysis, the same goes for the *Caladrius* model.

Furthermore, when zooming into the miss-classifications, some *major-damage* and *destroyed* labels were given to buildings in heavily impacted regions. However, no damage properties could be detected by inspecting the specific polygon image. The human annotators were probably biased by the visualisation of the total image, instead of examining only the characteristics of the construction. The *Caladrius* model did classify the respective buildings as *no-damage*, which were wrongly identified as miss-classification.

Miss-classifications due to the Differences between Disaster Datasets

It is interesting to determine if the *Caladrius* model under-performs on one specific disaster, and if this downgrades the overall performance of the xBD dataset. Earlier, the scores per disaster were given in Table 5.1. However, these results are based on training and testing solely on one specific disaster, whereas in this research, there is trained on the total dataset, including the four selected disasters. In Table 5.3, the prediction accuracy per disaster is listed and the recall values per damage type are showcased in Figure 5.7.

Table 5.3: Results of multi-classification per disaster (epoch=50), trained on the xBD dataset, sorted by the F1 scores and recall value per damage type

	F1		Recall			
	Harm.	Macro	No	Min	Maj	Des
Hurricane Matthew	0.44	0.54	0.16	0.84	0.46	0.75
Hurricane Michael	0.49	0.54	0.88	0.43	0.49	0.30
Tsunami Palu	0.49	0.51	0.94	0.00	0.17	0.85
Tsunami Sunda Strait	0.00	0.40	0.98	0.00	0.00	0.71
Total	0.63	0.68	0.86	0.68	0.42	0.76

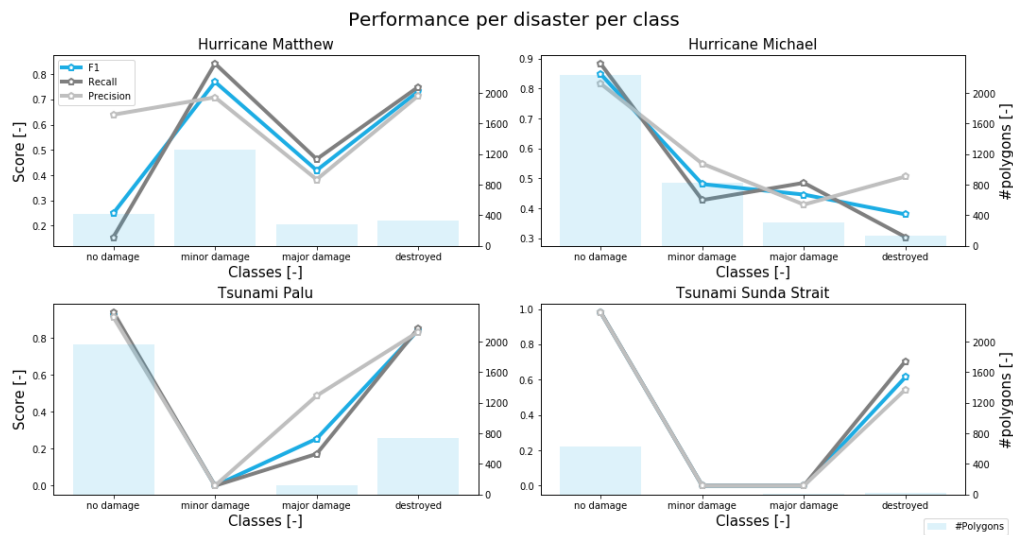


Figure 5.7: The F1-score, Recall and Precision plots per disaster and class

The quality of the predictions, linked to the Hurricane Matthew disaster, did score the highest Macro F1-value. Nevertheless, the performance is hardly perfect. The *no-damage* class underachieves because the model labels these polygons with the *minor-damage* type, the majority class within the Hurricane Matthew dataset. In Figure 4.6, it can be noticed that more than half of the *minor-damage* polygons within the total resampled training set belong to Hurricane Matthew. This can lead to a scenario where the *Caladrius* model is not only learning characteristics of this damage type but also linking buildings in this region to the *minor-damage* class, without being so. This is an undesired behaviour.

The second best scored damage type concerns the *destroyed* class, the prediction threshold of this label has the biggest distance to the *minor-damage* class and performs for this reason better than the *no-damage* and *major-damage* class.

The metric values based on the Hurricane Michael dataset are still affected by the imbalance, despite the resampling of the training-set. The F1-score follows the trend of the class type distribution, as can be noticed in Figure 5.7. The *no-damage* class outperforms the remaining classes.

In the confusion matrix in Figure 5.8, no clear diagonal can be recognised. By observing the *destroyed* predictions, substantially more False Negatives are counted than True Positives. It can be concluded that the *Caladrius* model does not perform well on this specific disaster.

In the Tsunami Palu test-set, no samples linked to the *minor-damage* label are present. For this reason, the corresponding Macro F1-score, recall and precision values equal zero.

The performance of the *major-damage* class is shockingly low, the model does not recognise characteristics linked to this damage type. When performing qualitative analysis, it is detected that miss-annotations cause most miss-classifications. All buildings surrounding the respective polygons are wiped away, however, no damage is visible in the polygon image itself.

In addition, other *major-damage* labelled samples are miss-classified without any determined explanation, clear visible details are showcased in the polygons hinting to damage related features. In Figure 5.9, examples of wrongly predicted *no-damage* polygons are given. In these specific samples, the *Caladrius* model lacks expertise.

Surprisingly, the *destroyed* class performance is above average, despite the low number of polygons included in the train and test-set.

The last disaster reviewed is the Tsunami Sunda Strait. Within the test-set, almost 94% of the polygons belong to the *no-damage* type. Again, the *destroyed* class scores high in terms of recall value.

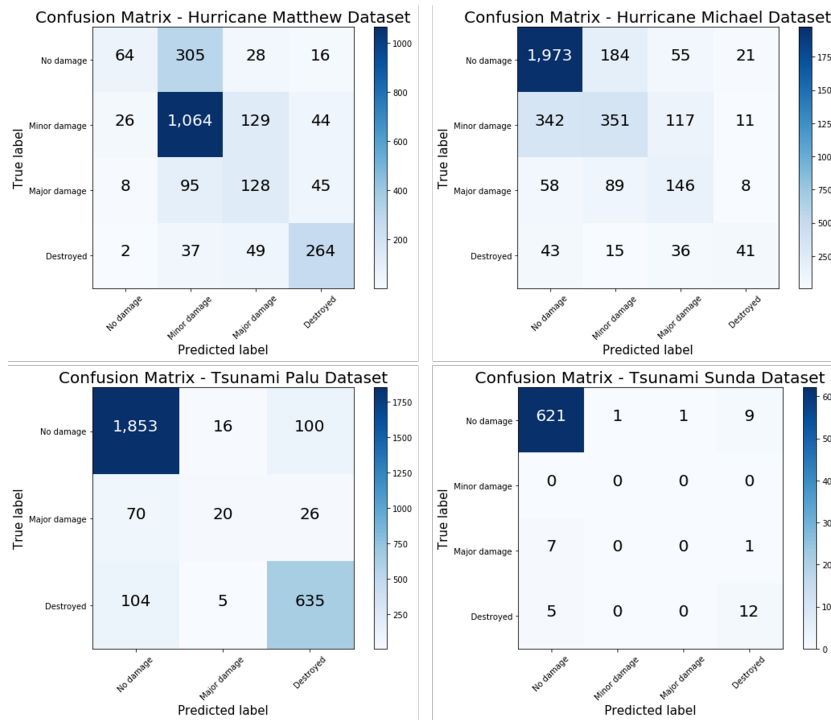


Figure 5.8: The confusion matrices per disaster dataset, trained on the xBD dataset



Figure 5.9: Visualisation of *Major damage* labelled polygons and classified with the *No damage* class (Example: Left Tsunami Palu 59-47 Right Tsunami Palu 60-225)

In general, it can be concluded that the *Caladrius* model encountered varied issues at each of the disaster datasets. Testing on the total dataset scores the highest Harmonic and Macro F1-scores. More consistent performance across disasters would have been preferred, as this would indicate reliable predictions in operational situations. In real life, a dataset of a disaster can be specified with one of the troubling characteristics discussed in this section, resulting in a scenario where the *Caladrius* model can have problems by predicting one of the damage types.

5.2. Relation between Resolution and Performance

Varying the resolution of the input imagery can have an influence on the performance of the *Caladrius* model. The drawn hypothesis equals a positive relationship between performance and resolution setting, implicating that a decrease in input data resolution will result in a reduction in correct predictions. It is interesting to find out if there is a specific resolution threshold, from which the model stops recognising damage characteristics and the classification becomes random. This latter is tested with the use of the Cohen's Kappa coefficient.

The tree data scenario's, single-, dual- and cross-mode are created to perform multiple experiments and to determine the relation. In this section, only the optical input imagery is discussed.

5.2.1. Single-Mode

The single-mode data scenario consists of similar resolution pre- and post-imagery, to train and test the *Caladrius* model on to investigate the searched relations. The original xBD dataset has a ground resolution of 0.5-meter, and is down-sampled to investigate the results of 2.5, 5.0, 10.0-meter resolution settings. In Table 5.4, the performance is expressed in F1-scores and recall values per damage type. In addition, Figure 5.10 showcases the reduction of F1-scores per resolution setting, in which it can be noted that the loss is not linear through the points. The biggest drop of score occurs between the 0.5 and 2.5-meter resolution setting. The difference in performance between the 2.5 and 5.0-meter resolution is two times smaller per meter. This indicates that the most recognisable features of damage are smaller than 2.5-meter. The slope between 5.0- and 10.0-meter is similar to the antecedent.

The 30-meter resolution dataset is added to the experiments, to investigate if the score would converge towards a certain minimum performance. However, the F1-scores do still reduce comparing the 10.0-meter resolution dataset. This finding is interesting because this means that the classification on the 10.0-meter dataset still functions in a degree.

In addition, all single-mode experiments are repeated with the binary distinction of *no-damage* and *damage* buildings. By observing the scores in Table 5.4, the multi-class and binary performance can be compared. The binary-class setting does outperform the multi-class. The Harmonized and Macro F1-score of the 30-meter resolution setting are even higher than the respective scores of the original xBD 0.5-meter multi-class experiment.

It is no surprise that the recall values of the *no-damage* labelled polygons exceed the *damage* cases, due to the small imbalance of the 60/40 ratio in the training dataset. In Figure 5.10, a similar behaviour and slope variation can be detected comparing both class settings.

Table 5.4: Results of multi- and binary-classification with varying resolution settings of the xBD dataset, sorted by the F1 scores and recall value per damage type

	Multi-classification						Binary-classification			
	F1		Recall				F1		Recall	
	Harm.	Macro	No	Min	Maj	Des	Harm.	Macro	No	Dam
0.5 m xBD	0.63	0.68	0.86	0.68	0.42	0.76	0.81	0.83	0.93	0.73
2.5 m xBD	0.54	0.61	0.81	0.62	0.33	0.69	0.75	0.78	0.90	0.65
5.0 m xBD	0.49	0.57	0.75	0.60	0.29	0.72	0.72	0.75	0.86	0.66
10.0 m xBD	0.41	0.52	0.66	0.62	0.23	0.61	0.67	0.71	0.86	0.57
30.0 m xBD	0.33	0.48	0.78	0.49	0.11	0.48	0.62	0.68	0.92	0.41

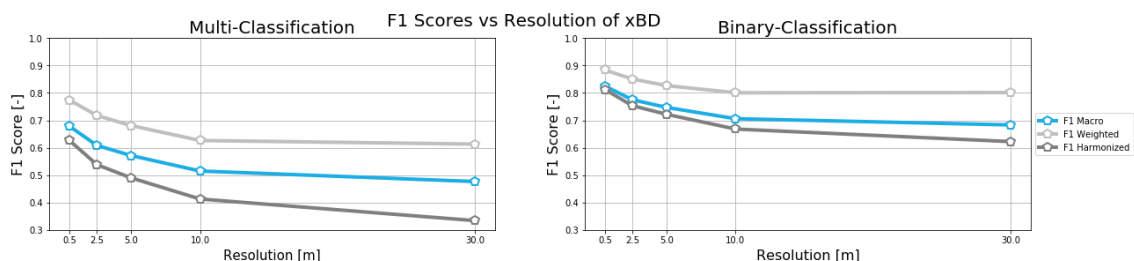


Figure 5.10: F1-score plot with respect to the resolution setting of the xBD dataset

All xBD datasets with varying resolutions are examined with the use of the Cohen's Kappa coefficient, elaborated in section 4.4. This coefficient indicates if the classification is to an extent based on the agreement of chance. When the value appears close to zero, a random classification occurs, a value equal to one represents a perfect performing model.

In Table 5.5 the metric scores are listed. It is interesting to estimate that even the 10.0 and 30.0-meter resolution xBD datasets surpass the performance of a random classifier. With the eye, no visible buildings can be found within most polygon images, linked to these low resolutions.

Table 5.5: Cohen's Kappa coefficients (κ) per xBD experiment, multi-label classification

Dataset	Res [m]	κ [-]
xBD	0.5	0.63
xBD	2.5	0.53
xBD	5.0	0.48
xBD	10.0	0.40
xBD	30.0	0.36

To examine the effect of reducing the resolution per class, the confusion matrices are given per dataset representing the multi- and binary-classification, in Figure 5.11.

First, the multi-class performances are investigated. The *no-damage* class originally scores the highest recall value, however, this value reduces the most by lowering the resolution. When the model misclassifies a *no-damaged* building, it is most often with the prediction *minor-damage*. In operational situations, this is preferred over the other two damage types due to the neighbouring position and the otherwise misplaced effort to help people, without utility.

The performance of the *minor-damage* type is the most constant in the four experiments. However, an increase in wrongly predicted *minor-damage* samples arises by lowering the resolution, specifically of the *no-damage* true labelled buildings.

The *major-damage* class is evidently under-performing compared to the rest of the classes. The model finds it hard to recognise the specific characteristics. Surprisingly, the assigned *major-damage* samples are often predicted to belong to the *no-damage* and *minor-damage* classes, despite the clear visible destruction of the constructions. It would be more obvious and preferred when the miss-classified *major-damage* samples were predicted as *destroyed*, otherwise dangerous situations could occur. Overall, the outcome of the *major-damage* class is not reliable and reaches 0.42 recall in the original run, which means the *Caladrius* model predicts more wrong than correct. In the matrix of the 10-meter resolution, this recall value is halved to 0.23.

Remarkably, the *destroyed* class scores more adequately than the *major-damage* class, despite similar visible features. Again, most miss-classified buildings are labelled with the *no-damage* type. The threshold of this specific class has the greatest distance to the *destroyed* class, nevertheless, the model can be influenced by the imbalance within the training set.

No identical trend for all classes is detected by lowering the resolution. For example, the 5.0-meter resolution dataset provides some improvements with respect to the *destroyed* class compared to the 2.5-meter resolution dataset.

When inspecting the confusion matrices created by the binary-classification, it can be seen that the *no-damage* class is performing less with every resolution drop. Unlike the *damage* class, which improves in the 5.0-meter resolution experiment. When the *damage* class is equal to the Positive class, the absolute numbers indicate more False Positives compared to False Negatives. However, when observing the normalised ratio, the False Negatives are more common. This ratio is not preferred, as this means the *Caladrius* model assigns the *no-damage* label to the *damage* buildings, which defeats the purpose of the damage assessment model. The *damage* buildings are crucial to locate.

In the experiment based on the original xBD dataset, the performance shows usable predictions after an disaster. The aid organisations would provide help for the labelled *damaged* houses, with an effective rate of 73%. The 27% of wasted effort can be taken for granted. However, when using 10-meter resolution data, almost 50% percent of the labelled *damaged* houses would be approached, without being so. Moreover, around 800 households would not be reached.

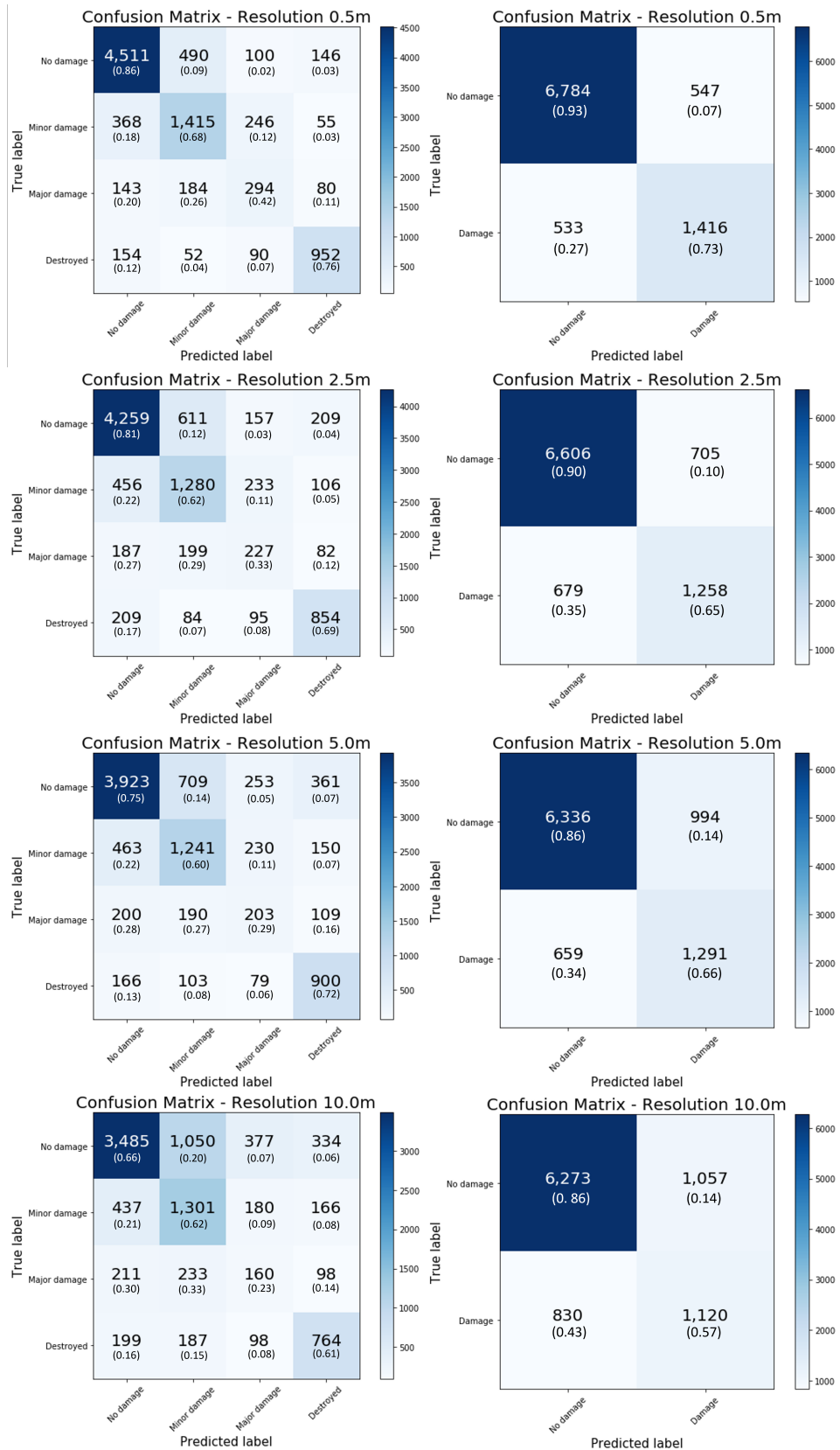


Figure 5.11: The confusion matrices of multi- and binary-classification, trained on the xBD dataset with varying resolutions

Subsequently, in Figure 5.12 visualisations are created of polygons to perform a qualitative analysis and to detect when miss-classifications occur. The first example represents a polygon classified correctly in all resolution settings. The second example is only recognised as *minor-damage* within the three higher-resolution datasets; 0.5, 2.5, 5.0-meter. This threshold of true predictions shifts over the four examples, to detect when detail is lost or characteristics can not be learned anymore. By reviewing the polygon visualisations, no trend or reasoning can be estimated.

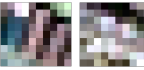
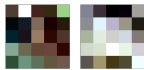










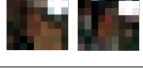


Example	Label	0.5m	2.5m	5.0m	10.0m
Palu 66_11	Destroyed	Pred: Destroyed ✓ 	Pred: Destroyed ✓ 	Pred: Destroyed ✓ 	Pred: Destroyed ✓ 
Michael 413_108	Major Damage	Pred: Major Damage ✓ 	Pred: Major Damage ✓ 	Pred: Major Damage ✓ 	Pred: No Damage ✗ 
Matthew 158_8	Minor Damage	Pred: Minor Damage ✓ 	Pred: Minor Damage ✓ 	Pred: Major Damage ✗ 	Pred: Major Damage ✗ 
Sunda 18_51	No Damage	Pred: No Damage ✓ 	Pred: Minor Damage ✗ 	Pred: Destroyed ✗ 	Pred: Destroyed ✗ 

Figure 5.12: Visualisation of polygons with varying resolution settings to detect when miss-classification occurs

A hypothesis was drawn linking miss-classifications to the building footprint, expressed in square meters. The smaller the building the less detail is visible and the more complex feature extraction becomes, consequently this can result in a lower classification rate. In addition, if the relation between footprint size and performance does exist, it is predicted to have a more significant effect on the lower resolution data.

In Figure 5.13, the distribution of True and False predictions versus the building footprint is plotted. The area size of the buildings is computed by extracting the length and width of the created polygons, based on the provided coordinates in the JSON-files. The longitude λ and latitude ϕ coordinates are translated to distances d , expressed in meters. Below, the integrated computation is stated to estimate the area A of the buildings. The notation R equals the radius of the Earth 6371 km.

$$\begin{aligned}
 a &= \sin^2\left(\frac{\Delta\phi}{2} \frac{180}{\pi}\right) + \cos(\phi_1 \frac{180}{\pi}) * \cos(\phi_2 \frac{180}{\pi}) * \sin^2\left(\frac{\Delta\lambda}{2} \frac{180}{\pi}\right) \\
 d &= R * 2 * \text{atan}(\sqrt{a}) * \text{atan}(\sqrt{1-a}) \\
 A &= d^2
 \end{aligned} \tag{5.1}$$

The calculated area A of the polygons, included in the test-set, ranges from 9 to 44244 m^2 , with the mean of 519.54 m^2 and standard deviation of 1177.52 m^2 . Due to the divergent footprints, outliers are erased by only including the 95th percentile of polygons.

By reviewing Figure 5.13, the first hypothesis can be investigated. An overlap is detected between the distributions of the True and False classified buildings. This demonstrates that a larger building size does not significantly enlarge the chance of correct classification. This non-existing relation is also confirmed by the results of the Hurricane Michael dataset, given in Table 5.3. At this specific location, the constructions are notably bigger than in the other three regions. Nevertheless, the Macro F1-score of the Hurricane Michael dataset does not excel.

In the zoomed-in plots, a difference between high- and low-resolution datasets can be detected. The maximum peak of the True predictions is higher for high-resolution data and decreases for lower resolution datasets. The opposite behaviour is identified in the False prediction probability density function plot. This does confirm the hypothesis of lower resolution data having a bigger chance of mispredicting small buildings. However, the effect is minor.

There must be emphasised that the weak relation between polygon size and performance exists due to the loss of detail in lower resolution data. Not because the *Caladrius* model can learn, for example, that small buildings are less resistant to an impact and therefore collapse earlier. All polygon images are resized to 299x299 pixels before inputting them into the CNN.

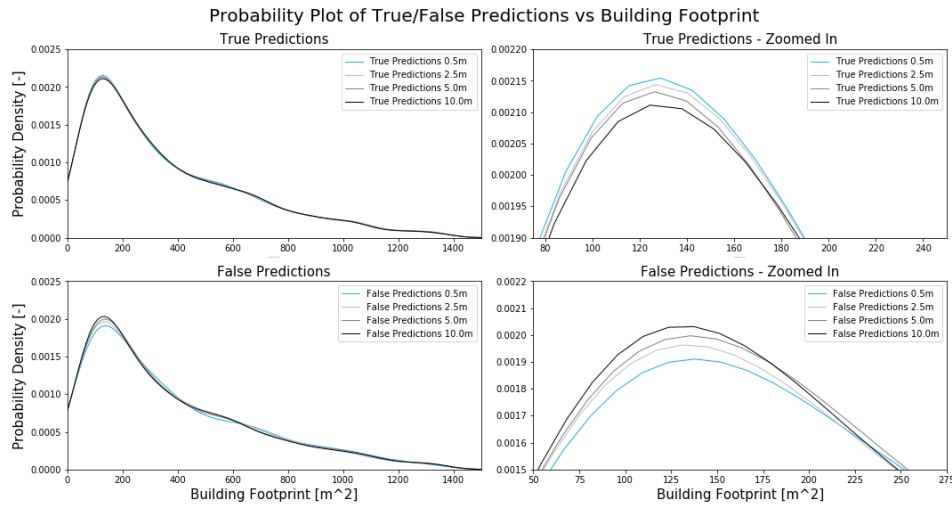


Figure 5.13: Distribution of True and False predictions versus the building footprint, per resolution dataset [m²]

Sentinel-2

Next, a single-mode experiment is executed using the Sentinel-2 1C dataset, consisting of 10-meter resolution optical imagery. The xBD dataset is collected with private satellites of high quality, whereas the Sentinel-2 mission is a governmental project providing openly available data. It is interesting to compare the performances by inputting the two data sources of equal resolutions.

By plotting the images on a reference frame using their longitude and latitude coordinates, a slight offset appeared between the visualisations of the xBD and the Sentinel-2 1C data. No constant shift was detected, and the large number of images in the dataset made it impossible to fix it manually. To make sure the possible difference between metric scores of the Sentinel-2 1C and 10.0-meter xBD experiments are not caused by the misalignment, an extra experiment is designed. The experiment equals the single-mode data scenario using the 10-meter xBD dataset. However, while creating the polygon images, the bounding boxes are shifted with one pixel representing 10-meters.

In Table 5.6, the F1-scores and recall values are listed of the respective optical data, including the 10-meter resolution imagery. It can be noticed that the Sentinel-2 1C experiment performs a bit lower than the 10.0-meter xBD dataset. Especially, the classes which represent damage properties are affected by the switch to another satellite source. This difference in metric scores is estimated to be not exclusively due to the misalignment of the polygons.

The Cohen's Kappa coefficient is checked to indicate if the Sentinel-2 experiment is based on chance, which equals 0.32. The coefficient of the shifted xBD dataset has a value of 0.39, which means the *Caladrius* model under-performs by inputting the Sentinel-2 dataset but reaches the same order of magnitude. This is very promising.

Examples of polygons of the 10-meter resolution datasets are showcased in Figure 5.14. It can be detected that the Sentinel-2 building visualisations have a different size compared to the xBD polygons. This can be explained by the large pixel size and the relative position of the polygons within the imagery. The clipping is based on the exact longitude and latitude coordinates, however, the relative position of those coordinates within a pixel window decides which pixels are included.

By reviewing the given polygons, the identified difference in colour palette could be the first reason for the reduced performance of the *Caladrius* model. In the xBD dataset, the RGB ranges are stored using 8-bits, including values between 0-255. The Sentinel-2 colours are provided with wavelength values of 0 to 10,000. The minimum and maximum of both ranges do not represent the same colour, meaning the transformation could have caused a loss in the true prediction rate.

The second reason for the reduction in the quality of the classifications can be linked to the replication of the images and copying the polygon coordinates plus ground truth damage labels, which are specifically created for the xBD dataset.

Table 5.6: Results of multi- and binary-classification of the optical imagery datasets consisting of 10-meter resolution, sorted by the F1 scores and recall value per damage type

	Multi-classification						Binary-classification			
	F1	Recall					F1	Recall		
		Harm.	Macro	No	Min	Maj		Des	Harm.	Macro
10.0 m xBD	0.41	0.52	0.66	0.62	0.23	0.61	0.69	0.71	0.86	0.57
10.0 m xBD + shift	0.39	0.50	0.71	0.54	0.17	0.64	0.68	0.72	0.87	0.57
Sentinel-2	0.27	0.44	0.72	0.55	0.07	0.44	0.56	0.63	0.88	0.39

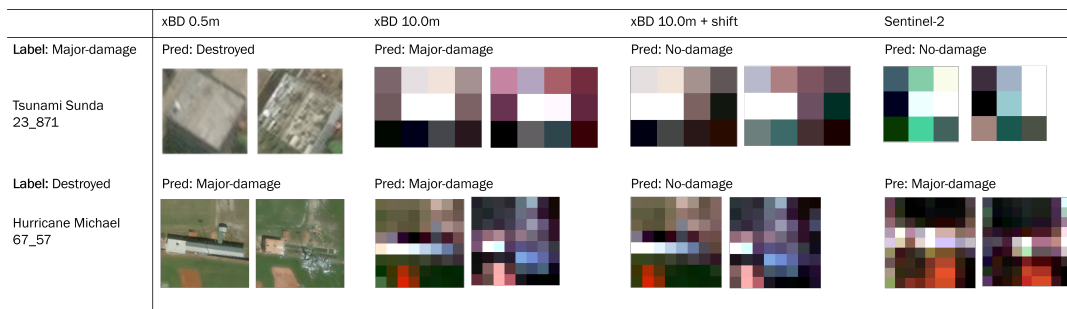


Figure 5.14: Visualisation of polygons representing the visual differences between 10-meter resolution imagery

5.2.2. Dual-Mode

The dual-mode experiments check if the characteristics of the training and test dataset are required to be identical. In situations of natural hazards where time is crucial, it is beneficial when the model's training process does not have to be repeated and the affected region can be assessed immediately. The experiments are based on the multi-classification of the *Caladrius* model, trained on high-resolution and tested on low-resolution data. All results of possible combinations are given in Table 5.7. The diagonal represents the Macro F1-scores and Cohen's Kappa coefficients of the single-mode experiments, inputting the four different resolution xBD datasets. It can be noticed that the performance reduces significantly while combining datasets of different resolutions. The scenario of training and testing the model on 10-meter resolution data is even more beneficial than training on 0.5-meter and testing on 2.5-meter resolution data. The higher the resolution difference between the train and test-set, the lower the performance. For example, the metric scores higher for the combination of 5.0- and 10.0-meter resolution data compared to 2.5- and 10.0-meter resolution data, representing the train and test-set, respectively.

The drop in Macro F1-scores per combination is related to the trend detected and showcased in Figure 5.10. Most recognisable damage features were captured between the 0.5- to 2.5-meter resolution data. In Table 5.7, it can be noticed that the maximum reduction of the prediction rate per meter occurred while combining 0.5 and 2.5-meter data. In the remaining cases, the average drop in Macro F1-score, listed horizontal, equals 0.8.

Table 5.7: Results of Dual-Mode experiments with use of the xBD dataset, multi-label classification

(a) Macro F1-Score

		Test			
		0.5 m	2.5 m	5.0 m	10.0 m
Train	0.5 m	0.68	0.49	0.38	0.34
	2.5 m	x	0.61	0.54	0.42
	5.0 m	x	x	0.57	0.49
	10.0 m	x	x	x	0.52

(b) Cohen's Kappa score (κ)

		Test			
		0.5 m	2.5 m	5.0 m	10.0 m
Train	0.5 m	0.63	0.37	0.22	0.16
	2.5 m	x	0.53	0.47	0.33
	5.0 m	x	x	0.48	0.38
	10.0 m	x	x	x	0.40

This reduction in metric values can be explained by the different visualisations used to learn damage characteristics from in the training phase and test and assign labels to. The *Caladrius* model could not identify patterns it mastered.

It can be concluded that it is essential to acquire equal, when possible, or similar resolution imagery to train and test the *Caladrius* model on.

5.2.3. Cross-Mode

The cross-mode experiments check if the characteristics of the pre- and post-event imagery are required to be identical. In situations after a natural hazard, it is beneficial to select the first dataset available to identify damage and act quickly to save lives. The data collection before the disaster is less time-restricted and can be optimised.

The experiments are based on combinations of varying resolutions within the image pairs, to train and test the *Caladrius* model. In Table 5.8, the Macro F1- and Cohen's Kappa scores are listed and show promising results. The output indicates that it is always better to pick a high-resolution pre-event image and a low-resolution post-event image, compared to selecting the respective low-resolution imagery pre- and post-event. For example, the 0.5-meter dataset can be combined with the 5.0-meter dataset, which will outperform the pair of similar 5.0-meter resolution pre- and post-event imagery. Furthermore, the difference between resolutions of pre- and post-event imagery does not require to be minimal. The experiment including 0.5- and 10.0-meter imagery surpassed the combination of 5.0- and 10.0-meter resolution data, pre and post-event, respectively.

All these findings show that the *Caladrius* model can learn innovative features when trained on a dataset with identical characteristics as it is tested on. Due to the deep layers in the CNN, the model can recognise damage properties, despite the non-similarities of pre- and post-event imagery. This strength is owned by the Siamese architecture, in which both weights are initiated and updated separately.

Overall, it is preferred to acquire high-resolution data in advance of the disaster, despite the resolution of the post-imagery, to train and test the *Caladrius* model.

Table 5.8: Results of Cross-Mode experiments with the use of the xBD dataset, multi-label classification

(a) Macro F1-Score

		Post-Event Imagery			
		0.5 m	2.5 m	5.0 m	10.0 m
Pre-	0.5 m	0.68	0.63	0.61	0.57
	2.5 m	x	0.61	0.58	0.57
	5.0 m	x	x	0.57	0.53
	10.0 m	x	x	x	0.52

(b) Cohen's Kappa score (κ)

		Post-Event Imagery			
		0.5 m	2.5 m	5.0 m	10.0 m
Pre-	0.5 m	0.63	0.56	0.53	0.49
	2.5 m	x	0.53	0.50	0.48
	5.0 m	x	x	0.48	0.42
	10.0 m	x	x	x	0.40

5.3. Relation between Satellite Imagery Type and Performance

The relation between the satellite imagery type and the true prediction rate of the *Caladrius* model is examined by a single-mode experiment inputting Sentinel-1 GRD data. This SAR data could replace the optical imagery when cloud coverage occurs after a natural disaster. The Sentinel-1 GRD visuals consist of 10-meter resolution imagery, comparable with the Sentinel-2 1C and 10-meter xBD datasets. As explained in subsection 4.1.2, Inception-V3 model is pre-trained on the ImageNet dataset. This signifies that the first layer of the network is frozen and memorises the features learned on the ImageNet. Due to the visual difference between optical and SAR imagery, the first layer is unfrozen to train and test the *Caladrius* model with the Sentinel-1 data. The results of both configurations are listed in Table 5.9 to compare the effect of the frozen first layer.

Table 5.9: Results of the multi- and binary-classification experiments Sentinel-1 and Sentinel-1* (*=without a frozen first layer)

	Multi-classification						Binary-classification			
	F1	Recall					F1	Recall		
		Harm.	Macro	No	Min	Maj		Des	Harm.	Macro
Sentinel-1	0.00	0.31	0.88	0.00	0.04	0.42	0.54	0.62	0.88	0.35
Sentinel-1 *	0.04	0.32	0.88	0.01	0.07	0.42	0.54	0.62	0.87	0.35

By reviewing the classification settings, no clear winner can be declared. Surprisingly, the untrained model does not outperform the pre-trained model. In both experiments, the model does perform inadequately by assessing the damage.

To determine if these metric scores are established by chance, the Cohen’s Kappa coefficient is computed, given in Table 5.10. Both values are around 0.17, which is close to zero, equal to a random classification. This indicates that the *Caladrius* model is not able to learn and recognise features by inputting Sentinel-1 GRD data.

Table 5.10: Cohen’s Kappa coefficients (κ) of the multi-classification experiments Sentinel-1 and Sentinel-1* (*=without a frozen first layer)

Dataset	Res [m]	κ [-]
Sentinel-1	10.0	0.17
Sentinel-1 *	10.0	0.17

In Figure 5.15, the confusion matrices visualise the distribution of polygons classified by the *Caladrius* model. The most significant difference between the un- and pre-trained model is linked to the predicted label *minor-damage*. Despite being trained with this distinction, the pre-trained model did not assign this damage type to any sample.

Overall, both models did perform poorly on the four label scenario. The *no-damage* type is predicted the most frequently, which can be explained by the data availability issues of the Hurricane Matthew dataset. The missing images altered the balanced ratio of the training dataset to 40/18/19/23, causing low recall values of the *minor-damage* and *major-damage* labels, equal to the minority classes.

Furthermore, the outcome of the binary classification is investigated. A recall value of ~ 0.35 is reached to allocate *damaged* polygons, which signifies more False than True predictions .

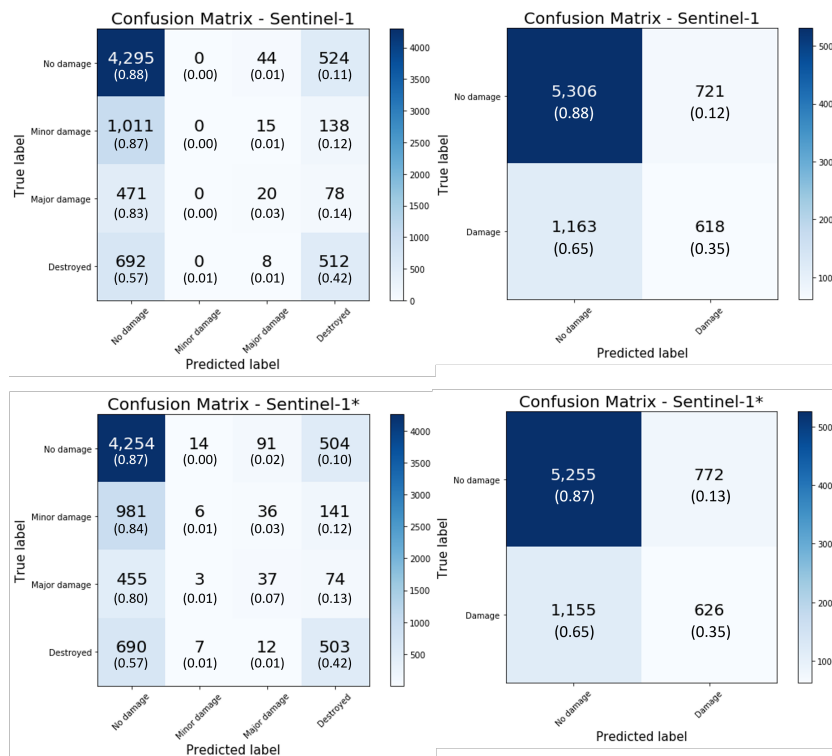


Figure 5.15: The confusion matrices of the multi- and binary-classification experiments Sentinel-1 and Sentinel-1* (*=without a frozen first layer)

In Figure 5.16, four visualisations of polygon images labelled with the four damage types are showcased. The examples are selected carefully to display the most apparent features of that specific label and to detect characteristics within the respective SAR figures. All examples are predicted correctly by inputting the original resolution xBD dataset.

The colours of the pixels indicate the backscatter intensity of the area. Without any changes on the ground, the colours would be assumed similar pre- and post-event. However, this theory is not validated by analysing Figure 5.16. The first example represents a building of the Tsunami Palu dataset, in which the original image shows no damage or change. Yet, the SAR visualisation does output colour variations and the *Caladrius* model does classify the sample as *major-damaged*. Also, in the second example, an extensive colour range is visible, but in this case, the model predicts the building with the *no-damage* label. The last two original xBD polygon images show damage properties due to the effect of the disasters. Nevertheless, both SAR images are predicted differently and incorrect. No clear pattern can be recognised in the visualisation of the Sentinel-1 imagery. Damaged buildings do not solely and always show considerable differences in colour pallet, and the trigger for the *Caladrius* model to classify a sample with a specific label is not found.

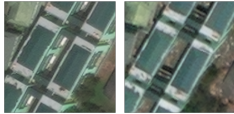

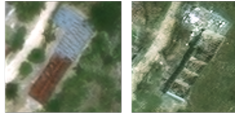

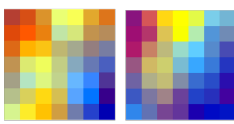
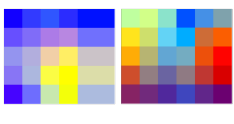
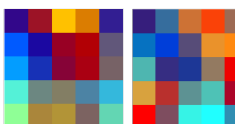
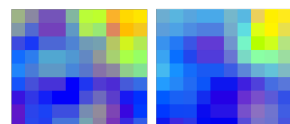
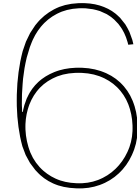
Example	Tsunami Palu 122_21	Hurricane Michael 427_32	Hurricane Matthew 236_28	Hurricane Michael 22_19
Label	No-damage	Minor-damage	Major-damage	Destroyed
xBD 0.5m	Pred: No-damage 	Pred: Minor-damage 	Pred: Major-damage 	Pred: Destroyed 
Sentinel-1	Pred: Major-damage 	Pred: Minor-damage 	Pred: Destroyed 	Pre: No-damage 

Figure 5.16: Visualisation of polygons representing the four damage types, comparing the xBD and Sentinel-1 dataset

Additionally, it is interesting to compare the metric scores of the *Caladrius* model by inputting Sentinel-1 GRD, Sentinel-2 1C and the xBD 10-meter resolution dataset. The xBD 10-meter dataset predicts significantly more polygon images correctly and reaches a Macro F1-score of 0.52. The Sentinel-2 experiment follows this performance ranking and the Sentinel-1 finished last with the value of 0.32. The SAR dataset included in this research provides no learnable features and equals a random classifier. The unconventional colour palette can have caused this difference.

It is important to note that this does not imply that all SAR datasets will be outperformed by optical imagery and the imagery type is non-suitable as input data of the *Caladrius* model to assess damage automatically.

Additionally, the research did not answer the question if SAR data could be used by an Automatic Damage Assessment tool, because this was already proven by many existing methods based on interferometric coherence and intensity correlations. In previous research, Sentinel-1 GRD did output promising results by mapping damage after a natural hazard [25]; however, this is not the case by using the *Caladrius* model.



Discussion

This chapter provides a discussion of the results and how they could be interpreted and implemented within operational situations after a natural hazard to optimise rescue missions. Additionally, the encountered limitations of this research are listed with recommendations for further research.

6.1. Interpretations

This research aimed to gain insights into the functioning of the *Caladrius* model in real-life data availability scenarios. It was initiated as the first step to create an all-weather and all-around applicable Automatic Damage Assessment (ADA) tool on building level.

Originally, the Convolutional Neural Network (CNN) was designed by *510 - an organisation of the Red Cross Netherlands* to classify four damage types, inputting satellite high-resolution optical imagery. However, in emergencies, this data was not always accessible or cloud-free, making the *Caladrius* model non-usable. In these situations, the *510* organisation relied back on field surveys to identify destruction, accompanied by many disadvantages. Additionally, not using available low-resolution optical and/or Synthetic Radar Aperture (SAR) data to map the affected region is a waste.

Eighteen experiments were designed and executed, including single-, dual- and cross-mode data scenarios. The input data were characterised by the resolution and observation sensor type to investigate the relations between the two parameters and the true prediction rate of the *Caladrius* model.

Single-Mode Experiments

Within the single-mode experiments, the model was trained and tested on multiple datasets, containing the original and down-sampled xBD, plus the replicated Sentinel-2 1C and Sentinel-1 GRD versions. Due to the identical study area and provided annotated damage labels, a comparison could be made. The hypothesis was drawn that changing the input characteristics from the design settings of the *Caladrius* model would lower the performance, which was confirmed to be conclusive.

The first experiment served as a benchmark and reached a Macro F1-score of 0.68, based on the original xBD 0.5-meter resolution dataset. Preferably, the accuracy measure would have been closer to 1, representing a perfect classification. The performed pre-processing steps of the data could have affected the non-optimal performance, such as the composition of the selected four disasters and the resampling method, which was proposed to resolve the imbalance issues.

Next, the predicted decreasing trend of the true prediction rate of the *Caladrius* model by down-sampling the xBD dataset was substantiated. Two key takeaways were identified to be interesting for the Red Cross. First, the loss per meter differed between the resolution transformations, indicating that most damage features were sized between 0.5-2.5 meters. Secondly, a random classification only occurred by inputting a dataset with a lower resolution than 10-meters. This latter came as a surprise because the clipped polygons consisted of just a few pixels. Still, the *Caladrius* model was able to learn and recognise related properties.

Despite the resampling of the imbalanced training dataset, the minority class *major-damage* was detected to perform inadequately and this deteriorated by lowering the resolution. Additionally, an increased number of samples was assigned to the majority label *no-damage*.

Identical trends were estimated by the binary-classification experiments, where the threshold was drawn between the *minor-* and *major-damaged* classes. However, the Macro F1-scores demonstrated more reliable and promising results. By reviewing the confusion matrices, the division of the existing miss-classifications of all resolution datasets showcased a higher absolute number of False Positives compared to False Negatives, indicating that more houses were predicted as *damaged* while being labelled as *no-damage*. In real-life situations, this will result in wasted effort of the Red Cross; however, too much given aid is preferred over missed dangerous situations. Unfortunately, the overlooked *damaged* buildings increased from 27% to 43% of the total *damaged* buildings, by inputting 0.5- and 10-meter resolution imagery, respectively. This defeated the purpose of the ADA.

One indicator was tested to explain the miss-classifications caused by down-sampling, represented by the size of building footprints. A minor effect was detected, containing a bigger chance of mispredicting small buildings for lower resolution data. Nevertheless, besides the loss of visible detail within the polygon images, no extra theoretical reason was determined to explain the positive relationship between the resolution and performance of the *Caladrius* model.

The Sentinel-2 Level-1C dataset was introduced to investigate the quality of an open-source compared to the equal resolution xBD dataset collected with private satellites. The free-accessibility and self-service of extracting the region of interest provide many opportunities to the Red Cross to implement this data source within the ADA process.

The experiment did reach a 15% lower Macro F1-score compared to the multi-classification of the 10-meter xBD dataset. Particularly the polygons containing *damaged* characteristics were misidentified, which were crucial in this use case to locate. By conducting an extra experiment, it was determined that the difference in accuracy measures was not exclusively caused by the misalignment of the polygon footprints due to the non-similar off-nadir angle of the satellites. Other reasons for the non-equal performances could be the colour translation required from wavelength values to RGB ranges or the variation in created polygon sizes. However, it was promising to notice that the *Caladrius* model learned features by inputting the B4, B3, B2 bands of this data source. Furthermore, it would be interesting to investigate the addition of the infrared band to expand the visuals' information and the outcome of an experiment based on the Level-2A product, including the Bottom-Of-Atmosphere reflectance.

Lastly, the Sentinel-1 GRD dataset provided the SAR imagery to train and test the *Caladrius* model. Unfortunately, the result appeared similar to the performance of a random classifier. Even when the operation of the frozen first layer of the CNN was reversed, no patterns were mastered to recognise damage features. In theory, the back-scatter values of the SAR observations must have been transformed to a colour visualisation, indicating differences with a range of intensity varieties. However, no trend was identified after a qualitative analysis to confirm this theory within the pre- and post-event polygon pairs. The poor prediction capability could be explained by the different visual interpretations required, the resampled resolution of the GRD product and the selected colour representation of the visuals by the polarisation bands; VH, HH and VH/VV.

Even though the performance of training and testing on the Sentinel-1 GRD dataset was inadequate, it has been proven that the *Caladrius* model was able to run by inputting SAR data.

Overall, fully relying on the predictions after a natural hazard is still discouraged based on the found results of the single-mode experiments. First, the initial performance level of the *Caladrius* model inputting high-resolution imagery must be improved before low-resolution imagery and Sentinel-2 1C would be deployable. Additionally, binary classification is found better suited than multi-classification.

Dual-Mode Experiments

The dual-mode experiments examined the capability of the *Caladrius* model to immediately test the dataset of the affected region, without retraining the CNN with similar data characteristics. The training process of the network takes a couple of days, which is not permitted in rescue missions where time is crucial. Unfortunately, it was estimated that this reduces the true prediction rate drastically. The bigger the difference between the resolution of both datasets, the lower the accuracy measure.

Practically, this implies that the organisation 510 should train the *Caladrius* model with multiple different resolution datasets and choose the suited version which matches the available satellite imagery after a disaster has struck.

Cross-Mode Experiments

The last cross-mode experiments mimicked the data scenario of high-resolution pre-event imagery and low-resolution post-event imagery. After a natural hazard, the first dataset available is highly appreciated as it includes the just wrought havoc, regardless of the corresponding characteristics. Pre-event imagery has no restricted time period to acquire, creating the opportunity to select high-quality and resolution data.

The results of the experiments outputted minimal difference, during training and testing on the resolution combinations. This increased the flexibility of matching imagery pairs in operational situations. It was found beneficial to include the highest-resolution data possible in advance of the disaster. In contrast, the resolution of the post-event imagery impacted the performance less strongly. This simplifies the trade-off between rapid data collection and resolution setting.

It can be concluded that the *Caladrius* model can learn innovative features when trained on a dataset with identical characteristics as it is tested on, whatever the build-up of the datasets may be.

6.2. Implications

All these experiments created insights into how to implement the *Caladrius* model in an emergency use case and how to rely on the predictions. Last month, the Hunga Tonga-Hunga Ha'apai volcano erupted causing a subsequent tsunami. The coral reefs and surrounding islands were affected by the ash and the force of the water. In this situation, the ADA tool of the Red Cross could have been used to map the destruction and to identify the islands locating the most people in need.

The highest resolution optical data, preferably higher than 5,0-meters, should have been collected pre- and post-event to create image pairs, prioritising the quality of the data in advance and the speed after the hazard. It is not obliged to match the data characteristics, learned from the cross-mode experiments. In addition, by performing the dual-mode experiments it was found essential to equal the resolution of the test- (imagery of Tonga) and the training dataset. Unfortunately, the models trained in this research do not apply to this specific use case due to the dataset's missing volcano eruption disaster types. However, the xBD does provide previous natural events of this category.

Since the training process is time-intensive, it is preferred this process is taken care of in advance. By down-sampling the xBD dataset in multiple resolution settings and combinations, pre- and post-event, various trained *Caladrius* versions should be created to select the suited model when needed.

In this research, the building polygons were provided with longitude and latitude coordinates. In emergency operations, the buildings should be extracted from the imagery using a model or with the help of information on the internet. For example, HOTOSM and Microsoft Maps Team invested in locating building footprints on the Open Street Map of various areas around the globe [27] [40].

The reliability of the binary-classification predictions of the Tonga use case would be related to weather conditions, the data resolution, the composition of the training dataset and the quality of the *Caladrius* model itself. It defines if the predictions could be used as leading information or as an additional reference when outlying the rescue mission.

6.3. Limitations

This research is associated with limitations, which hindered the search for conclusive answers to the initiated research questions and the quality of the predictions. The three most significant limitations are described below.

Quality of the xBD dataset

All experiments were based on the xBD dataset, in its original state, down-sampled and replicated. This dataset was selected because of the provided coordinates of the building footprints and validation labels of damage types. However, this 'high-quality' dataset was accompanied by multiple flaws, leading to unsatisfactory results.

Primarily, the high class-imbalance made it challenging to train the model and learn the minority class's damage properties accurately. The *no-damage* class dominated the prediction ratio, despite attempting to create an equilibrium by a resampling method.

Secondly, the included data could have been selected with higher standards. A portion of the images was covered with clouds, which overshadowed the targets. Additionally, the acquisition period of disasters took place years in advance, creating non-reliable image pairs pre- and post-event due to missing or renovated buildings, which is undesired in change detection processes. The variable satellite parameters also made a shortcoming in the quality of the image pairs. Illumination differences caused a reduction in performance, expressed in the Area Under the Curve (AUC) value. Henceforth, this could be corrected within the pre-processing steps of the data.

At last, the polygons were labelled subjectively by the annotators using the Joint Damage Scale. Qualitative analysis identified inconsistencies due to misinterpretations of visual characteristics and the vague description of the distinctions. The mislabelling led to inputting misinformation into the *Caladrius* model to learn features from during the training phase.

Quality of the SAR dataset

To test the potential of the *Caladrius* model with the active and passive sensor based satellite data, the Sentinel-1 and Sentinel-2 datasets were included. The Sentinel-1 Ground Range Detection (GRD) imagery generated by Synthetic Aperture Radar observations were only based upon amplitude measures without phase information. The Sentinel-1 Single Look Complex (SLC) product includes both details, which can be beneficial in damage detection. However, this latter dataset was not openly accessible on Google Earth Engine and could not be tested within this research.

Furthermore, the performances of both satellite sensor types were examined with 10-meter resolution figures, which created non-optimal circumstances to make a comparison. No detail and contours were visible in the low-resolution data, restricting the use of side-view characteristics, which equalled one of the main advantages of the SAR data. Unfortunately, no high-resolution SAR data was available to replicate the disasters in the xBD dataset and conclusively investigate the relation between data sensor type and true prediction rate of the *Caladrius* model.

Quality of the Caladrius Model

In this research, the *Caladrius* model, owned by 510 - an organisation of the Red Cross Netherlands, was used to assess damage automatically. By inputting the high-resolution xBD imagery no optimal performance was found. After investigating the visuals of the miss-classifications, some samples were identified with apparent features belonging to the annotated label. This concluded that the inadequate quality of the model was one of the reasons for the undesired true prediction rate.

The computer scientists of the Red Cross did research opportunities to improve the architecture. First, a trial was implemented to replace the cross-entropy loss function with a log F1-score computation, which showed robust behaviour to an imbalanced input dataset. Secondly, it was found that the Inception-V3 was not the best-suited fit to pre-train on the ImageNet dataset. The EfficientNet-4 scored higher metric measures [55] and was integrated within the *Caladrius* model as a pilot.

6.4. Further Research

Further research is recommended to resolve the aforementioned limitations to achieve the *Caladrius* model's optimal performance on high-resolution optical imagery. Additionally, the pre-processing steps and training phase, performed in this research, should be criticised and optimised.

Without this being accomplished, the predictions are not deployable within emergency operations. Plus, searched relationships and correlations will not be substantiated using the CNN.

Subsequently, in-depth research must be performed to investigate the capability of inputting SAR imagery into the *Caladrius* model. Sentinel-1 GRD can be explored using different polarisation bands to recreate the colour-composite, and the SLC product can be tested for improvements. But most importantly, a high-resolution SAR dataset should be accessed as it could have much potential to create reliable damage assessment maps in all-weather situations and replace the optical imagery when necessary.



Conclusion

In this thesis, the influence of different input data characteristics was tested on the true prediction rate of the Convolutional Neural Network *Caladrius* to assess the damage on building level. Three relationships were investigated to specify the different input data characteristics concerning various resolutions and satellite imagery types.

Relationship between input imagery resolution and multi-classification performance

The relation between the resolution of the optical input data and the true prediction rate of the *Caladrius* model was estimated to be positive. By down-sampling the spatial resolution of the xBD imagery to 2.5, 5.0 and 10.0-meter, the Macro and Harmonic F1-scores decreased due to the loss of visible details within the polygon images. The function between both parameters was not associated with a constant linear coefficient, as the slope between the resolution transformations differed per meter. This indicated that most recognisable features were sized smaller than 2.5-meter. Nevertheless, damage properties still existed within the 10-meter resolution data. The agreement of chance was tested with a 30-meter resolution dataset and Cohen's Kappa coefficient, showing that the experiment outperformed the capability of a random classifier.

A trend became visible of the model predicting more and more samples to the majority class of the label *no-damage*, by reducing the resolution. The model no longer recognised the existing damage properties of buildings, and had yield to the imbalance of the dataset.

Equal behaviour was demonstrated by the binary classification of the xBD datasets, in which multi-class labels were grouped to *no-damage* and *damage* distinctions. However, the *Caladrius* model did reach higher F1-scores by switching to this distribution setting. The detailed distinction of the four classes demanded too much detail.

Relationship between mixed input data characteristics and multi-classification performance

The classification quality did reduce drastically when non-equal resolution optical imagery was acquired to train and test the CNN, respectively. It was found more optimal to input equal low-resolution data than a high-resolution data combination in all situations.

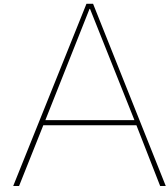
Mixing different resolutions of pre- and post-event optical imagery showed promising results to become more flexible in creating matching imagery pairs and speeding up the damage mapping. Preferably, the highest resolution data possible was collected in advance of the disaster, independent of the resolution value of the post-event imagery, to train and test the *Caladrius* model on.

Relationship between input imagery type and multi-classification performance

In this research, no substantiated comparison could be made between the capabilities of the *Caladrius* model by inputting optical and Synthetic Aperture Radar (SAR) imagery due to the inadequate resolution of the data sources.

The optical imagery was represented by the down-sampled 10.0-meter xBD and Sentinel-2 1C dataset. Despite the identical data characteristics, the two experiments showed different classification performances. Especially, the identification of buildings containing *damage* properties under-performed by training and testing with the Sentinel-2 1C dataset. Still, the order of magnitude of metric measures was similar and showcased that the *Caladrius* model was able to distinguish classes.

The *Caladrius* model based on the SAR data, originated from the Sentinel-1 GRD mission, outputted a true prediction rate comparable to the competence of a random classifier. This low performance indicated the *Caladrius* model was not capable of learning features from the provided SAR visuals. From the experiments in this research, the conclusion could be drawn that the implementation of the Sentinel-2 1C data was superior to the Sentinel-1 GRD data. However, it is not determined if the difference in the observation sensor type exclusively caused the variation in accuracy measures. Moreover, other SAR datasets with higher resolutions or originating from a different source could provide adequate performances using the *Caladrius* model. It is important to investigate this further to eliminate or prove the functionality of SAR data, as it would create many opportunities when optical observations are unusable.



Appendix

Table 1: The 19 included disasters in the xBD dataset [23]

Disaster Event Name	Location	Event Dates		
Mexico City Earthquake	North America	19 Sep		2017
Midwest US Floods	North America	03 Jan	31 May	2019
Pula Tsunami	Asia	18 Sep		2018
Sunda Strait Tsunami	Asia	22 Dec		2018
Hurricane Michael	North America	07 Oct	16 Oct	2018
Hurricane Florence	North America	10 Sep	19 Sep	2018
Hurricane Harvey	North America	17 Aug	02 Sep	2017
Hurricane Matthew	Central America	28 Sep	10 Oct	2016
Monsoon in Nepal, India, Bangladesh	Asia	01 Jul	30 Sep	2017
Joplin, MO Tornado	North America	22 May		2011
Moore, OK Tornado	North America	20 May		2013
Tuscaloosa, AL Tornado	North America	27 Apr		2011
Carr Wildfire	North America	23 Jul	30 Aug	2018
Woolsey Fire	North America	09 Nov	28 Nov	2018
Pinery Fire	Oceania	25 Nov	02 Dec	2018
Portugal Wildfires	Europe	17 Jun	24 Jun	2017
Santa Rosa Wildfires	Central America	08 Oct	31 Oct	2017
Lower Puna Volcanic Eruption	North America	23 May	14 Aug	2018
Guatemala Fuego Volcano Eruption	Central America	03 Jun		2018

Bibliography

- [1] F. Agostinelli et al. "Learning Activation Functions to Improve Deep Neural Networks". In: *International Conference on Learning Representations* (Dec. 2014), p. 9. arXiv: 1412.6830v3 [cs.NE].
- [2] Y. Bai, E. Mas, and S. Koshimura. "Towards Operational Satellite-Based Damage-Mapping Using U-Net Convolutional Network: A Case Study of 2011 Tohoku Earthquake-Tsunami". In: *Remote Sensing* 10.10 (2018). ISSN: 2072-4292. DOI: 10.3390/rs10101626.
- [3] A. Bárcena et al. *Handbook for Disaster Assessment*. Santiago, Chile: ECLAC United Nations, 2014.
- [4] C. M. Bishop. "Pattern Recognition and Machine Learning". In: Sprint Street, New York, US: Springer Science+Business Media, 2006, p. 758. ISBN: 10: 0-387-31073-8,
- [5] J. Bland. *Cohen's Kappa -Percentage agreement: a misleading approach*. Ed. by University of York Department of Health Sciences. 2008. DOI: 10.1109/ACII.2013.47.
- [6] S. van den Boogaart. *The use of Active Learning in Automated Damage Assessment*. MSc. Thesis. Maastricht University, 2021.
- [7] J. Brownlee. "Better Deep Learning". In: *Machine Learning Mastery*, 2018, p. 575.
- [8] J. Brownlee. "Deep Learning for Computer Vision". In: *Machine Learning Mastery*, 2019, p. 563.
- [9] J. Brownlee. "Imbalanced Classification with Python". In: *Machine Learning Mastery*, 2020, p. 463.
- [10] M. Buda, A. Maki, and M. Mazurowski. "A systematic study of the class imbalance problem in convolutional neural networks". In: *Neural Networks* 106 (Oct. 2018), pp. 249–259. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2018.07.011.
- [11] Q. D. Cao and Y. Choe. "Building damage annotation on post-hurricane satellite imagery based on convolutional neural networks". In: *Natural Hazards* 103.3 (July 2020), pp. 3357–3376. ISSN: 1573-0840. DOI: 10.1007/s11069-020-04133-2.
- [12] *Catastrophic Hurricane Michael Strikes Florida Panhandle*. URL: <https://www.weather.gov/tae/HurricaneMichael2018>. (accessed: 09.11.2021).
- [13] X. Chen, Q. Sun, and H. Jun. "Generation of Complete SAR Geometric Distortion Maps Based on DEM and Neighbor Gradient Algorithm". In: *Applied Sciences* 8 (Nov. 2018), p. 2206. DOI: 10.3390/app8112206.
- [14] *Copernicus-S1-GRD*. URL: <https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS-S1-GRD>. (accessed: 05.06.2021).
- [15] T. Cova. "GIS in Emergency Management". In: Dec. 1999, pp. 845–858.
- [16] *Data Products Sentinel-1*. URL: <https://sentinel.esa.int/web/sentinel/missions/sentinel-1>. (accessed: 09.09.2021).
- [17] *Data Products Sentinel-2*. URL: <https://sentinel.esa.int/web/sentinel/missions/sentinel-2>. (accessed: 09.09.2021).
- [18] S. Ejaz and R. Islam. "Masked Face Recognition Using Convolutional Neural Network". In: (Dec. 2019), pp. 1–6. DOI: 10.1109/STI47673.2019.9068044.
- [19] A. Fujita et al. "Damage detection from aerial images via convolutional neural networks". In: *2017 Fifteenth IAPR international conference on Machine Vision Applications (MVA)* (May 2017), pp. 5–8. DOI: 10.23919/MVA.2017.7986759.
- [20] P. Ge, H. Gokon, and K. Meguro. "A review on synthetic aperture radar-based building damage assessment in disasters". In: *Remote Sensing of Environment* 240 (2020), p. 111693. ISSN: 0034-4257. DOI: 10.1016/j.rse.2020.111693.

- [21] I. Gortzak. *An assessment of damage detection methods in post-disaster humanitarian response: the opportunities and challenges for innovation*. Internship Report. University of Twente, 2021.
- [22] R. Gupta et al. "Focusing on the Big Picture: Insights into a Systems Approach to Deep Learning for Satellite Imagery". In: (2018). arXiv: 1811.04893 [cs.CV].
- [23] R. Gupta et al. "xBD: A Dataset for Assessing Building Damage from Satellite Imagery". In: *Carnegie Mellon University* (2019). arXiv: 1911.09296.
- [24] E. Harirchian et al. "Application of Support Vector Machine Modeling for the Rapid Seismic Hazard Safety Evaluation of Existing Buildings". In: *Energies* 13.13 (2020). ISSN: 1996-1073. DOI: 10.3390/en13133340.
- [25] J. van Heyningen. *Rapid Disaster Response, building damage detection using the Google Earth Engine*. MSc. Thesis. Delft University of Technology, 2018.
- [26] *High resolution SAR: Capella*. URL: <https://www.capellaspace.com/>. (accessed: 25.07.2021).
- [27] *Humanitarian OpenStreetMap Team*. URL: <https://www.hotosm.org/>. (accessed: 11.02.2022).
- [28] V. Sevilgen J. Patton R. Stein. "Tsunami in Sulawesi, Indonesia, triggered by earthquake, landslide, or both". In: *Temblor* 10 (Oct. 2018).
- [29] G. Jedlovec. "Advances in Geoscience and Remote Sensing". In: Rijeka, Croatia: InTech, 2009, p. 740. ISBN: 978-953-307-005-6, DOI: 10.5772/46139.
- [30] L. Jeni, J. F. Cohn, and F. De La Torre. "Facing Imbalanced Data—Recommendations for the Use of Performance Metrics". In: (2013), pp. 245–251. DOI: 10.1109/ACII.2013.47.
- [31] M. Ji et al. "A Comparative Study of Texture and Convolutional Neural Network Features for Detecting Collapsed Buildings After Earthquakes Using Pre- and Post-Event Satellite Imagery". In: *Remote Sensing* 11.10 (2019). ISSN: 2072-4292. DOI: 10.3390/rs11101202.
- [32] M. Ji et al. "Discrimination of Earthquake-Induced Building Destruction from Space Using a Pre-trained CNN Model". In: *Applied Sciences* 10.2 (2020). ISSN: 2076-3417. DOI: 10.3390/app10020602.
- [33] T. van Laarhoven. "L2 Regularization versus Batch and Weight Normalization". In: *CoRR* (2017). arXiv: 1706.05350.
- [34] X. Liu, J. Wu, and Z. Zhou. "Exploratory Undersampling for Class-Imbalance Learning". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.2 (2009), pp. 539–550. DOI: 10.1109/TSMCB.2008.2007853.
- [35] S. Loos et al. "G-DIF: A geospatial data integration framework to rapidly estimate post-earthquake damage". In: *Earthquake Spectra* 36.4 (2020), pp. 1695–1718. DOI: 10.1177/8755293020926190.
- [36] S. Manickam et al. "Estimation of Snow Surface Dielectric Constant From Polarimetric SAR Data". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10.1 (2017), pp. 211–218. DOI: 10.1109/JSTARS.2016.2588531.
- [37] M. Martinez and R. Stiefelwagen. "Taming the Cross Entropy Loss". In: (2019). Ed. by Thomas Brox, Andrés Bruhn, and Mario Fritz, pp. 628–637.
- [38] *Maxar Open Data Program*. URL: <https://www.maxar.com/open-data>. (accessed: 06.11.2021).
- [39] I. Melekhov, J. Kannala, and E. Rahtu. "Siamese network features for image matching". In: (2016), pp. 378–383. DOI: 10.1109/ICPR.2016.7899663.
- [40] *Microsoft has released new and updated building footprints*. URL: <https://blogs.bing.com/maps/2022-01/New-and-updated-Building-Footprints?s=09>. (accessed: 30.01.2022).
- [41] P. M. Milano. "The Power of Inception: Tackling the Tiny ImageNet Challenge". In: (2015).
- [42] A. Moreira et al. "A tutorial on synthetic aperture radar". In: *IEEE Geoscience and Remote Sensing Magazine* 1.1 (2013), pp. 6–43. DOI: 10.1109/MGRS.2013.2248301.
- [43] E. Neumayer and F. Barthel. "Normalizing economic loss from natural disasters: A global analysis". In: *Global Environmental Change* 21.1 (2011), pp. 13–24. ISSN: 0959-3780. DOI: 10.1016/j.gloenvcha.2010.10.004.

- [44] F. Nex et al. "Structural Building Damage Detection with Deep Learning: Assessment of a State-of-the-Art CNN in Operational Conditions". In: *Remote Sensing* 11.23 (2019). ISSN: 2072-4292. DOI: 10.3390/rs11232765.
- [45] H. Nielsen. "Theory of the backpropagation neural network". In: (1989), 593–605 vol.1. DOI: 10.1109/IJCNN.1989.118638.
- [46] R. De Oliveira et al. "A System Based on Artificial Neural Networks for Automatic Classification of Hydro-generator Stator Windings Partial Discharges". In: *Journal of Microwaves, Optoelectronics and Electromagnetic Applications* 16 (Sept. 2017), pp. 628–645. DOI: 10.1590/2179-10742017v16i3854.
- [47] S. Plank. "Rapid Damage Assessment by Means of Multi-Temporal SAR — A Comprehensive Review and Outlook to Sentinel-1". In: *Remote Sensing* 6.6 (2014), pp. 4870–4906. ISSN: 2072-4292. DOI: 10.3390/rs6064870.
- [48] *Platform Google Earth Engine*. URL: <https://earthengine.google.com/>. (accessed: 25.05.2021).
- [49] *Rapidly Assessing the Impact of Hurricane Matthew in Haiti*. URL: <https://www.worldbank.org/en/results/2017/10/20/rapidly-assessing-the-impact-of-hurricane-matthew-in-haiti>. (accessed: 08.11.2021).
- [50] H. Ritchie and M. Roser. "Natural Disasters". In: *Our World in Data* (2014).
- [51] O. Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *CoRR* 43 (Feb. 2021). arXiv: 1409.0575.
- [52] N. Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15 (June 2014), pp. 1929–1958.
- [53] D. Summers. "Optical Remote Sensing: Principles". In: Australian National University, 2016.
- [54] C. Szegedy et al. "Rethinking the Inception Architecture for Computer Vision". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* abs/1512.00567 (2015). arXiv: 1512.00567. URL: <http://arxiv.org/abs/1512.00567>.
- [55] M. Tan and Q. V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *CoRR* abs/1905.11946 (2019). arXiv: 1905.11946.
- [56] P. Turon and T. Depost. "LEO thermal imagers: push broom or whisk broom". In: *Defense, Security, and Sensing*. 1992.
- [57] *UNOSAT Rapid Mapping Service*. URL: <https://www.unitar.org/maps/unosat-rapid-mapping-service>. (accessed: 21.08.2021).
- [58] T. Valentijn et al. "Multi-Hazard and Spatial Transferability of a CNN for Automated Building Damage Assessment". In: *Remote Sensing* 12.17 (2020). ISSN: 2072-4292. DOI: 10.3390/rs12172839.
- [59] A. Vetrivel et al. "Segmentation of UAV-based images incorporating 3D point cloud information". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XL-3/W2 (2015), pp. 261–268. DOI: 10.5194/isprsarchives-XL-3-W2-261-2015.
- [60] Q. Wang and J. JaJa. "From maxout to channel-out: Encoding information on sparse pathways". In: 19 (2013). arXiv: 1312.1909 [cs.NE].
- [61] *What is Remote Sensing?* URL: <https://earthdata.nasa.gov/learn/backgrounders/remote-sensing>. (accessed: 06.09.2021).
- [62] *What is Satellite Remote Sensing?* URL: <https://www.skyrora.com/post/remote-sensing>. (accessed: 08.01.2022).
- [63] *What is Synthetic Aperture Radar?* URL: <https://earthdata.nasa.gov/learn/backgrounders/what-is-sar>. (accessed: 18.11.2021).
- [64] B. J. Wythoff. "Backpropagation neural networks: A tutorial". In: *Chemometrics and Intelligent Laboratory Systems* 18.2 (1993), pp. 115–155. ISSN: 0169-7439. DOI: 10.1016/0169-7439(93)80052-J.

-
- [65] J. Z. Xu et al. "Building Damage Detection in Satellite Imagery Using Convolutional Neural Networks". In: (2019). arXiv: 1910.06444 [cs.CV].
- [66] *xView2: Assess Building Damage*. URL: <https://xview2.org/>. (accessed: 24.03.2021).
- [67] S. Yun et al. "Rapid damage mapping for the 2015 Mw 7.8 Gorkha Earthquake Using synthetic aperture radar data from COSMO-SkyMed and ALOS-2 satellites". In: *Seismological Research Letters* 86 (Nov. 2015), pp. 1549–1556. DOI: 10.1785/0220150152.