University of Cambridge

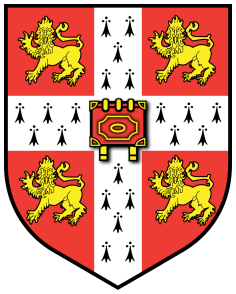
AI4ER MRes Project Report

A Deep Learning Approach to Post-Hurricane Automatic Damage Classification

July 2022

Author:
Grace Beaney Colverd[1]

Supervisors:
Dr. Emily So[2]
Dr Christian Geiss[3]



Abstract

When natural disasters hit urban areas, mass damage to infrastructure and loss of life can occur. Key to mitigating loss of life is a rapid assessment of the damage caused, in order to prioritise allocation of emergency services and humanitarian aid. Disaster events often damage communication channels, and local conditions can remain treacherous for some time, making the communication and assessment of damage both a low priority and a risky endeavour for people on the ground. Hence, ground truth surveys of damage frequently cannot be carried out in the timescales needed to prioritise relief. Automated damage assessment of infrastructure via remote sensing data offers a pathway for rapid damage classification without the need for an in-person ground survey. Here we apply automatic damage classification methods to infrastructure in the aftermath of Hurricanes Ida and Delta in Louisiana. We compare different deep learning architectures, and train a ResNet50 model on the Ida imagery to perform damage classification, achieving a building level F1 score of 84.1%. Model generalisation is tested on the unseen Delta imagery and achieves a building level F1 score of 55.6%. Damage maps at varying geographic levels are generated and their potential use in aid prioritisation is discussed.

[1] Department of Earth Sciences, University of Cambridge. Email: gb669@cam.ac.uk

[2] Department of Architecture, University of Cambridge. Email: ekms2@cam.ac.uk

[3] Department of Civil Crisis Information and Geo-Risks, German Aerospace Centre. Email: christian.geiss@dlr.de

Contents

1 Introduction

1.1 Background and Related Work

In the aftermath of natural disasters, infrastructure is assessed to provide a preliminary damage assessment based on visible damage to the structures. Depending on the extent of the damage, this can be time and resource intensive if completed on the ground, particularly if conditions remain treacherous. The use of remote sensing for rapid, large scale assessment of damage is an attractive proposition, and has generated a large amount of research interest over previous years [1], [2], [3], [4]. It is the prevalent view that extreme weather events such as hurricanes will increase due to anthropological impacts on the Earth. Given the potential for their increased frequency and severity, timely and automatic assessment of damage caused by hurricanes will only become more important. This report will add to the tools available for automatic damage assessment.

There is currently no universal mechanism for identifying damage across different types of disasters, due to their differing signatures [5]. In this report we focus on hurricane damage: specifically wind damage to infrastructure. Brown (2019) [6] found correlations between ground level damage to infrastructure with damage observed from earth observation methods, and REF [] found that severity of wind damage to buildings is correlated with the degree of change in the roof structure. Hence we can expect to infer with reasonable confidence, a buildings damage grade from earth observation imagery.

Traditional approaches to damage classification manually code features for extraction from remotely-sensed imagery, such as roof edges, roof texture and colour intensity. Thomas (2014) [7] manually coded features that were extracted from pre- and post-storm imagery, which were then used to predict damage type by comparing structures pre- and post-disaster. Other approaches for classification included the use of support vector machines (SVM). Romaniello (2017) [8] uses synthetic aperture radar (SAR) data and applied SVM algorithms to classify damage after the Haiti earthquake in January 2010. When classifying the most extreme damage, SVM achieves a classification accuracy of 84%.

Recent advances in image classification use convolutional neural networks (CNNs) to automatically extract features from images, removing the need to manually label features. The first use of modern CNNs was in analysing the hand-drawn digit dataset MNIST [9]. Deep architectures grew in popularity after the performance of AlexNet in the ImageNet classification task [10]. Since 2014, deep learning architecture has focused on deeper and wider networks and the improvements in classification performance has translated well into other domains [11].

The post-disaster damage classification community have made much use of deep architectures for extracting high-quality visual features from remotely-sensed imagery. The community have to contend with the 'messiness' of post disaster imagery, and most authors use some form of building or object detection before grading damage. Chen (2021) estimates damage after the 2018 Eureka-Kansas tornado, using imagery from Unarmed Aerial Vehicles (UAV). DenseNet architecture is used for image classification, after object detection. The authors note the novelty and utility in generating a full geo-tagged damage map, which would be ideal for use by aid agencies in supporting relief efforts. The modelling framework achieve a top classification accuracy of 84.8% [1].

Cheng (2021) applies a novel stacked model (SPDA) to UAV video imagery in the aftermath of Hurricane Dorian, achieving a 61% damage grade classification accuracy (90% with a +-1 class deviation). The model consists of two CNNs: a building localisation model (to distinguish buildings vs. other objects) and a fully-trained MobileNet CNN (Howard et al., 2017 [12]) for multi-classification of damage [2].

Francesco (2019) sources a large variety of multi-resolutional training data from different disasters (focusing on earthquakes and explosions) and evaluates the potential of deep learning in damage classification to deliver results in operational conditions (e.g. pre-trained networks with limited training and short (1-2 hour) analysis times. The authors use a custom CNN architecture tailored to multi-resolution data, with dilated convolutions and dense connections. For satellite data sources, accuracy when testing varied between 54.1% to 93.9% depending on the type of data, with UAV and airborne data generally outperforming satellite due to greater consistency in the quality of the data [5].

Authors in this area often cite the lack of labelled training data as a limitation in the application of deep learning approaches [4]. This shortage has several contributing factors. Firstly, the lack of raw imagery of appropriate resolution and timescale. Traditional satellites such as Sentinel do not have the appropriate

resolution to be of use in damage classification at the building level, and follow set orbital paths that are unlikely to pass over damaged areas in the days immediately following a disaster. Secondly, the lack of labels for appropriate raw imagery. Private satellites and UAV can offer high resolution data in the appropriate timescale, but a shortage of labels means that creating training data is a laborious process, often outsourced to non-experts. Whilst there are approaches for the automatic creation of training data for exposure risk purposes [13], we are grateful to Risk Management Solutions (RMS) for providing a large quantity of high-quality training data, in the form of building outlines and labels of damage classification per building for the infrastructure in the aftermath of Hurricanes Ida and Delta. The timely, high resolution data to which these labels apply is sourced from The National Oceanic and Atmospheric Administration (NOAA). The NOAA Remote Sensing division uses earth observation satellites to capture images in the immediate aftermath of hurricanes and tornadoes in the United States. These images are openly accessible online [14].

1.2 Project Approach

In this report, a damage classification algorithm is trained on the buildings in the New Orleans Basin, classifying damage in the aftermath of Hurricane Ida. Training data is based on the damage classification labels provided by RMS. After optimising the model architecture and running parameter sweeps, the model is tested in the same region. The model is then tested in a new region: the Creole region of Louisiana, in the aftermath of Hurricane Delta.

Hurricane Ida formed in the Atlantic basin on August 26th, 2021, making landfall in western Cuba as a Category 1 hurricane, before intensifying into a Category 4 hurricane on August 29th and making landfall on the Louisiana coast. Ida was the second most damaging hurricane to hit Louisiana in recorded history, after Hurricane Katrina in 2005, and caused an estimated $75 billion of damage in the US [15]. Hurricane Delta formed on October 4th, 2020 from a tropical wave, and made landfall in Creole as Category 2 hurricane. Total losses from Delta amounted to $3.09 billion [16]. Timelines for aid

Timelines for immediate aid responses in the wake of Hurricanes vary from days to weeks, depending on provider and capabilities. Local government support, when well prepared, can take 1-2 days. Federal government between 2 days - 2 weeks. NGO's and volunteers typically cannot enter until ground situations have stabilised. For Hurricane Katrina, more devastating than Ida, this was around the 1 week mark. Immediate aid takes the form of search & rescue, medical support, food and water supplies. To be of most use, we estimate that an automatic damage classification tool should take no more than 1 day to run, given the short timelines involved, and the 1-2 day delay in sourcing remote sensing imagery. This informs our later choices of evaluation criteria.

1.3 Report Outline

The structure of this report is as follows: an introduction to the datasets and model methods is given in Section 2. Results are given in Section 3 and discussions of the limitations and extensions of this work are given in Section 4.

2 Methodology

In this section we introduce the datasets used and the modelling methodology.

2.1 Datasets

2.1.1 Input data:

- **NOAA Aerial Imagery** High-resolution RGB images captured by the NOAA Remote sensing team via satellite. Ground sample distance (GSD) for each pixel is 15 cm.
  - Ida imagery captured between August 30th - September 2nd, 2021, data bounds in Figure 2
  - Delta imagery captured between October 10th-11th, 2020, data bounds in Figure 2

  Data is in RGB GeoTiffs of between $(3 \times 9\text{k} \times 9\text{k})$ to $(3 \times 14\text{k} \times 14\text{k})$ pixel size.

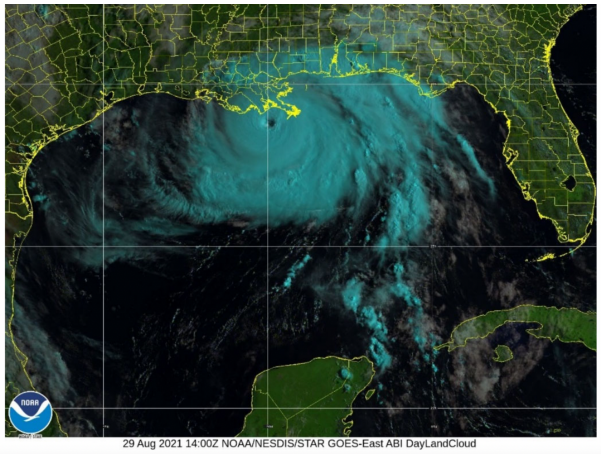


Figure 1: Cloud convection image of Ida, a few hours before landfall at Port Fourchon, Louisiana, on August 29th, 2021. Image courtesy of NOAA/NESDIS/STAR. [15]

NOAA generally starts to publish data a few days after an event, with full data for a large event taking up to a week. Due to the speed of data collection and publishing, NOAA note that it is possible that some clouds remain on the imagery. Visually, for the Ida imagery, cloud incidences seemed very low, especially after the merge of multiple days of imagery, and the decision was made to proceed without running any further cloud removal. This may need to be revisited if the methodology is applied elsewhere.

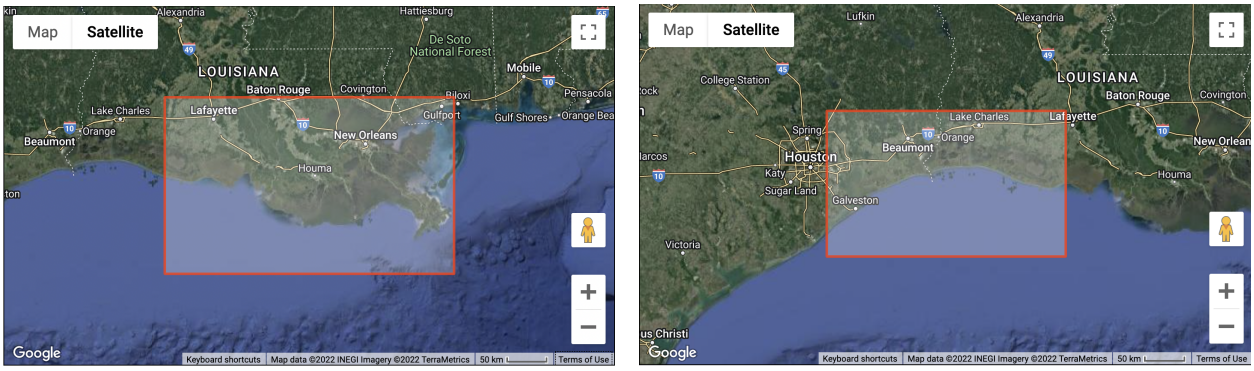


Figure 2: Left: Ida NOAA Data Boundaries [17] Right: Delta NOAA Data Boundaries [18]

2.1.2 Target data:

- **RMS Damage Classification** The ground truth for our classification is provided by RMS, in the form of a label of 'damaged' or 'not damaged' on a per building basis for buildings in the regions affected by Ida and Delta. These labels are generated from RMS proprietary methodology, which are validated using ground truth data. RMS methodology focuses on roof smoothness and surrounding debris in order to classify a building as damaged or not.

  Also provided by RMS are the building footprints for the region in the form of geo-tagged polygons, see Figure 3.

There is inherent uncertainty in the damage labels due to their origin as the output of the RMS model. The RMS model was not quantitatively evaluated, but manually checked that the most damaged areas from ground truth surveys were captured in the classification, and areas of sparse damage had some individual buildings checked to ensure they were correctly classified. Hence the accuracy scores of the models likely include some misclassifications where the original damage label was incorrect.

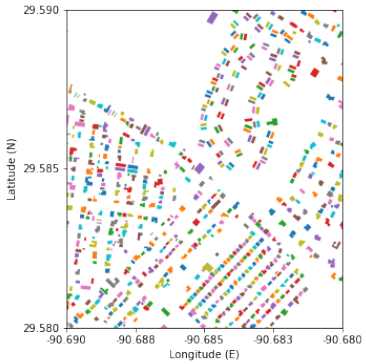


Figure 3: Example of Building Footprints

There is also inherent uncertainty in the building footprints we used to identify buildings These were provided from RMS who sourced them from Open Street Maps. These were manually verified to a certain extent but upon visual inspection still contain a small amount of incorrect footprints. We could have focused on cleaning the footprints using an object detection algorithm of the kind mentioned above, but chose to focus efforts on the damage classification given the vast majority of footprints were correct.

2.2 Building Types

Building type has a large impact on the level of damage a building will sustain, and will also affect the performance of any classification methodology. Figure 4 shows a selection of building types we have identified in the Southern Lousiana region. The 'patches' are of uniform area, generated from NOAA Ida imagery around the centroid of a building footprint. The reasoning behind uniform patch size is discussed below, but it is important to note that scale is preserved with this choice: the larger buildings are identifiably larger. The shotgun-style house with a long, thin body and pitched roof is classic French creole style **??**, as seen in columns 2 and 3. Given the use of roof smoothness as a proxy for damage in the RMS methodology, it is likely that the visual appearance of the striped rust on the corrugated-iron roofs affected the damage labels provided. Visual inspection indicates these are undamaged, which contradicts the damage labels provided. This will lead to inherent incorrectness in the model that accuracy metrics will not be able to catch.

Surrounding debris is also used as a damage proxy in RMS methodology. For the largest buildings in columns 5 and 6, the surroundings are obscured due to the median patch size. This will likely lead to a reduction in classification accuracy for those buildings which have surrounding debris but no obvious roof damage. Note also that the buildings are all relatively spaced out, which is typical both of the US and particularly of coastal regions. Hence our methodology will likely have limited success when applied to dense urban areas, given that visual similarity is the biggest driver of success of domain transferability CITE.

Finally note the variation in altitude angle across the patches. This arises due to the nature of NOAA imagery (private satellite, custom flight paths, multi-day image capturing and lower orbit). This will likely also add some bias to the classification results.

Four datasets were generated, bearing in mind that a balanced data set was found to improve transferability of network predictions and robustness of outputs [5].

1. **Hurricane Ida**

   (a) A small sandbox dataset with which to quickly test model architecture, consisting of 5k samples. Created from a 4km^2 area containing a high proportion of damaged buildings in order to create a balanced data set - Figure 8.

   (b) A full dataset of 100k samples containing all the damaged buildings from Ida. The 50k damaged buildings were used as a limiting factor, and 50k undamaged buildings were chosen randomly from the remaining 130k buildings.

1 2 3 4 5 6

Figure 4: Building types by column. 1: Small outhouses and informal buildings. 2: Corrugated-iron roof shotgun-style houses. 3: Wooden roof shotgun-style houses. 4: Medium residential houses. 5: Larger residential houses. 6: Large commercial buildings.

Subset	1a	1b	2a	2b
Train	3,640	63,963	N/A	300
Validation	911	15,991	N/A	N/A
Test	1,951	20,029	3,108	2808

Table 1: Dataset sizes

2. **Hurricane Delta**

   A balanced data set of 3k patches was generated, including all of the 1.5k damaged buildings and a random 1.5k subset of the other 100k undamaged buildings. This dataset is completely unseen in model training and is used to test domain transferability both with and without finetuning of the model. The finetuning training set is capped at 300, to simulate the amount of labelled patches that could be manually created in a few hours in the aftermath of a hurricane in a real-world scenario.

   (a) All patches used to test model

   (b) 10% of samples used to finetune model, and remaining patches used to test the updated model

A 0.64:0.16:0.2 split was applied to create the Train:Validation:Test datasets for datasets 1A and 1B, filtered out those buildings whose footprints did not wholly overlap with the aerial imagery available. The dataset sizes are given in Table **??**.

2.3 Models

Here, we describe problem formulation, model architectures investigated, training process and evaluation criteria.

2.4 Problem Statement

The problem is framed as patch-based image classification into the target label of damaged or undamaged. For each building footprint the corresponding NOAA imagery patch was generated thus:

1. Building footprint and aerial imagery aligned using their latitude and longitude. (Note that upon visual inspection a range of alignments were found - Figure 17.)
2. Square patch generated of a uniform area around the centroid of the building polygon, with the area of the patch chosen to fit around the median house size
3. Patch tagged with damaged or undamaged label

The decision for uniform patch sizes was made after domain expert discussion (C. Geiss, personal communication, June, 2022) indicated that this method generally produced the best results. This methodology ensures scale is preserved, i.e. large buildings appear larger. Patch size was capped due to the preference for mainly single buildings in view: in built up areas larger patch sizes would include multiple buildings. The disadvantage of this method is that for very large buildings, the surrounding area cannot be seen; see Figure 17. Surrounding debris is an important factor in damage classification so some classification ability may be lost for these very large buildings. This risk is mitigated by its small effect: upon investigation only the top 16% of buildings lack surrounding image data - examples of a building in the top 16th percentile is given in Figure 17. An extension of this work would be to create two separate networks for median and very large buildings and then to unify these results.

Patches were resized to between 180 $\times$ 180 - 299 $\times$ 299 pixels, dependant on model architecture, and pre-processing included normalisation based on full dataset mean and standard deviation.

2.5 Model Setup

Several model architectures were tested within this project.

Model 1A	5-Layer CNN w/ Batch Normalisation
Model 1B	5-Layer CNN w/o Batch Normalisation
Model 2A	ResNet 18
Model 2B	ResNet 50
Model 3	Inception V3

All models are implemented in PyTorch [19]

Model 1 was chosen as a simple introductory model, to explore the power of a simple CNN. Full model architecture is given in Appendix B.2. Model 1 was tested with and without batch normalisation to improve understanding of its impacts.

Models 2A & 2B use the residual architecture introduced in [20]. ResNet was chosen due to its strong performance in image classification tasks and generally fast training speed. Given the large training data and limited computational resources, optimising training speed was important. 2A was chosen to assess performance with a shallower network. 2B was chosen to compare shallow with very deep architectures.

Model 3 uses the updated Inception architecture introduced in [11]. Chosen due to its strong performance in ImageNet, and in particular its ability to learn both large and small features through the use of kernels of varying sizes. Given the high-resolution imagery, a model that discerns the fine detail of wind-damaged roof tiles and the large scale structural damage was predicted to perform well.

2.6 Training setup

Models 1A and 1B are trained with randomised initial conditions, whilst Models 2A, 2B and 3 are implemented as pre-trained feature-extractors. The 'transfer learning' approach has been found to produce better results than training from scratch, and reduces computational costs. (CITE). The classifier layers of Models 2 and 3 are updated to predict 2 classes rather than 1000.

Given the balanced datasets, binary cross entropy was chosen as the loss function.

$$\ell = 1\{y_i = 1\} \cdot \sum_{i=0}^{1} y_i \log(\hat{y}_i) \tag{1}$$

where y_i is the prediction probability and $\hat{y}_i$ is the target probability, corresponding to class i.

Model architectures were tested on dataset 1a with a full parameter sweep implemented with PyTorch Lightning and Wandb [21], using the 'train' and 'validation' subsets. Early stopping was based on validation loss. The parameters and their variation settings are given in Table 2.

Optimiser	LR	Batch Size	Weights
Adam	[0:1]	[34:500]	[1,1]
Adagrad	Uniform	LogUniform	[0.2,1]
RMSProp	Distrib	Distrib	[0.1,1]
SGD		step = 8	[0.01,1]

Table 2: Parameters for Sandbox sweep

This parameter sweep was used to inform the parameter sweep for dataset 1b. The two best performing model architectures, and a reduced parameter selection were tested on dataset 1b, again on the 'train' and 'validation' subsets.

2.7 Evaluation

Criteria for selecting the best model included validation accuracy and runtime. The problem formulation as patch-based image classification aims to develop a method and model which will generalise well to other regions without the need for time intensive retraining. However choosing a sensible trade-off in terms of accuracy and runtime retains the option to quickly fine tune on a new region.

The best model is then tested on the Ida test set for the New Orleans region. The accuracy, recall, precision and F1-score (metrics) are calculated at a building level, based on its corresponding patch. Patch predictions are then converted into damage maps of the area at varying geographic scales. Metrics are also calculated at this higher geographic level. This output choice is informed by desired use case for this work: in the immediate aftermath of a hurricane to prioritise aid. Finally the model is applied to datasets 2a & 2b, buildings in the aftermath of Hurricane Delta, and damage maps for this region are also generated.

In the literature, the Federal Emergency Management Agency (FEMA) rating system (0-4) is often used for classifying the scale of building damage [2]. Given the training data received from RMS, we use a binary classification of damaged or undamaged, and create custom 'Damage Levels' (DL). Shapefiles for 'census tracts' and 'blocks' are used to create damage maps, sourced from the USA Census Bureau [22]. (Blocks house up to approximately 600 people. Census tracts house approximately 3k-8k people). For each geographic region, all the patches within the region are pulled and a DL based on percentage of damaged buildings (x) is created, which is then plotted as a damage map:

Damage Level	Definition
DL 4	$x > 80\%$
DL 3	$60\% < x \leq 80\%$
DL 2	$40\% < x \leq 60\%$
DL 1	$20\% < x \leq 40\%$
DL 0	$0\% < x \leq 20\%$

Table 3: Damage Level definitions

Aid prioritisation based on this methodology relies on the assumption that severity of damage is correlated to % of damage: we assume that if a building is severely damaged, then buildings in its vicinity will also be damaged. All damage maps shown are based on % rather than absolute levels to enable the methodology to

be applied at different regional scales. It is also possible to generate damage maps based on the absolute number of damaged buildings - examples can be seen in the Appendix REF.

For the damage maps, only the 'test' sub-datasets are used, refer to Table **??**.

Accuracy, Precision, Recall and F1-Score are calculated at the geographic region level using the following definitions:

Positive	$= DL3 + DL4$
Negative	$= DL0 + DL1 + DL2$
TP, FP	True Positives, False Positives
TP, FN	True Negatives, False Negatives
Accuracy	$= \frac{TP+TN}{TP+TN+FP+FN}$
Precision	$= \frac{TP}{TP+FP}$
Recall	$= \frac{TP}{TP+FN}$
F1-Score	$= \frac{2 \times Precision \times Recall}{Precision+Recall}$

Table 4: Metrics

3 Results

3.1 Dataset 1a

For dataset 1a, full parameter sweeps were run for all models, with at least 30 parameter sweeps run per model.

The validation accuracy, model configuration and other key metrics at the building level for each top performing model on dataset 1a is given in Table 5.

Model	Optimiser	LR	Batch Size	Weights	Num. Epochs	Runtime	Val Acc	Val Precision	Val Recall
Model 1A	ADAM	0.000099	64	[0.1, 0.2]	4	2m 21s	73.2%	76.5%	74.5%
Model 1B	SGD	0.04202	104	[1,1]	19	6m 54s	74.1%	76.6%	73.5%
Model 2A	ADAM	0.003296	104	[1,1]	10	6m 6s	81.6%	81.9 %	81.8%
Model 2B	ADAM	0.003382	216	[0.1,1]	16	7m 37s	81.6%	81.6%	81.6%
Model 3	ADAGRAD	0.07671	88	[0.01,1]	13	9m 51s	77.0%	77.2%	77.1%

Table 5: Results of parameter sweeps for dataset 1a at building level

The damage map for dataset 1a was created using the configuration of Model 2B and is at the block level; see Figure 7.

We see excellent block level performance, with a recall of 93.1%. Only 7 blocks with severe damage were not identified. Full metric performance at block level is given in Table 6.

Metric	Building Level	Block Level
Accuracy	76.2%	76.5%
Precision	71.8%	61.8%
Recall	88.3%	93.1%
F1 Score	79.2%	74.3%

Table 6: Sandbox Data Results

3.2 Dataset 1b

On the basis of performance on dataset 1a, a parameter sweep was run for the Models 2A & 2B on dataset 1b. RMSProp was excluded as a potential optimiser. For Model 2B the configuration with the second highest

validation accuracy was chosen due to the massively reduced runtime - more details in Appendix REF.

The top models performance on validation accuracy and runtime were:

Model	Val Acc	Num Epochs	Runtime
Model 2A	85.0%	10	1h 57m 19s
Model 2B	84.8%	3	48m 52s

Table 7: Comparison of top perfomers for Models 2A & 2B

Given the intended uses of this work, in near-time disaster relief, a high accuracy, low training time combination is preferable over a very high accuracy and long training time.

For both these models, recall is higher than precision indicating a tendency to over-predict damage. This is favourable to under-predicting damage given the nature of intended use. Hence Model 2B was chosen as the final model architecture due to the reduced runtime and similarly high accuracy in classification.

Parameters for this run are given in Table **??**.

Weights: [0.1,1]
Batch Size: 280
Learning Rate: 0.06834
Optimiser: Adam

Henceforward this configuration shall be referred to as the 'final model'.

The final model was tested on the Test subset of dataset 1b and performance was analysed at the building, block and census tract level.

Metric	Building	Block	Census Tract
Accuracy	83.6%	80.3%	83.7%
Recall	90.7%	90.2%	92.9%
Precision	78.5%	71.9%	69.3%
F1 Score	84.2%	80.0%	79.4%

Table 8: Dataset 1b Test Results

The damage maps for the 1b Test subset were created at the census tract level to give a high-level damage map of the whole region. For any smaller region, a block-level map as generated for dataset 1a could be created. A similarly storng performance is seen for 1b, as for 1a. The census tract Recall is 92.9%: only 4 regions with severe damage were missed. The full damage maps can be seen in Figure 7.

3.3 Hurricane Delta Damage Classification

3.3.1 Dataset 2a

The final model was also tested on the Delta imagery. This dataset was completely unseen during model development and training. The damage map is at census tract level, and can be seen in Figure **??**. The metric results for building and census tract level are given in Table 9.

Metric	Building Level	Census Tract
Accuracy	62.3%	83.2%
Recall	46.6%	25.0%
Precision	68.9%	50.0%
F1 Score	55.6%	33.3%

Table 9: Delta Test Results - no finetuning

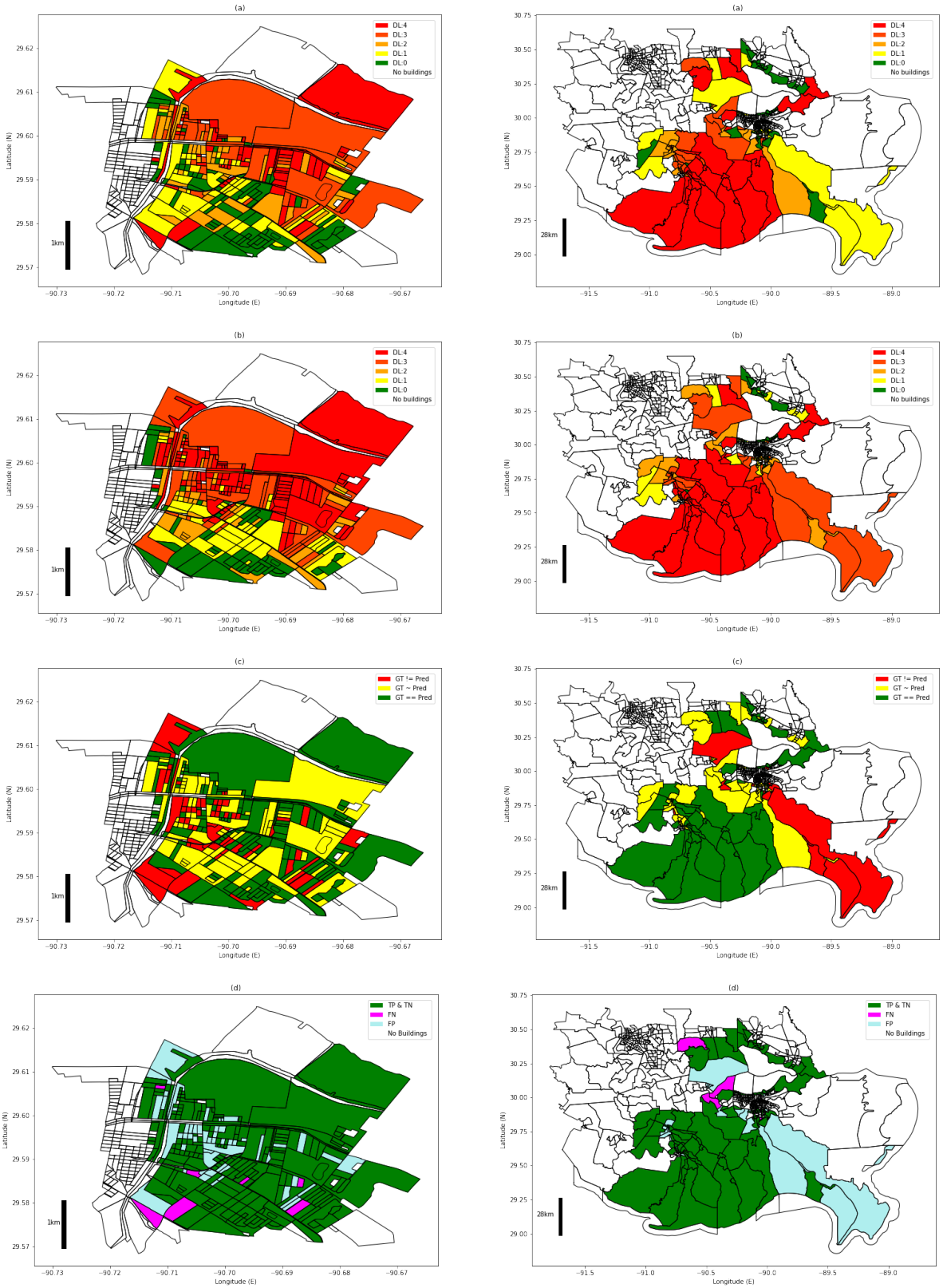


Figure 5: (Damage maps for datasets 1a (left) and datatset 1b (right), created on the Test subset. (a) and (b) show damage maps for Ground Truth and Model Prediction respectively (c) shows difference between Ground Truth and Prediction, with yellow representing a difference of 1 DL. (d) highlights the FP and FN, using definitions given in Section 2. Note that 'No Buildings' refers to within the test subset.

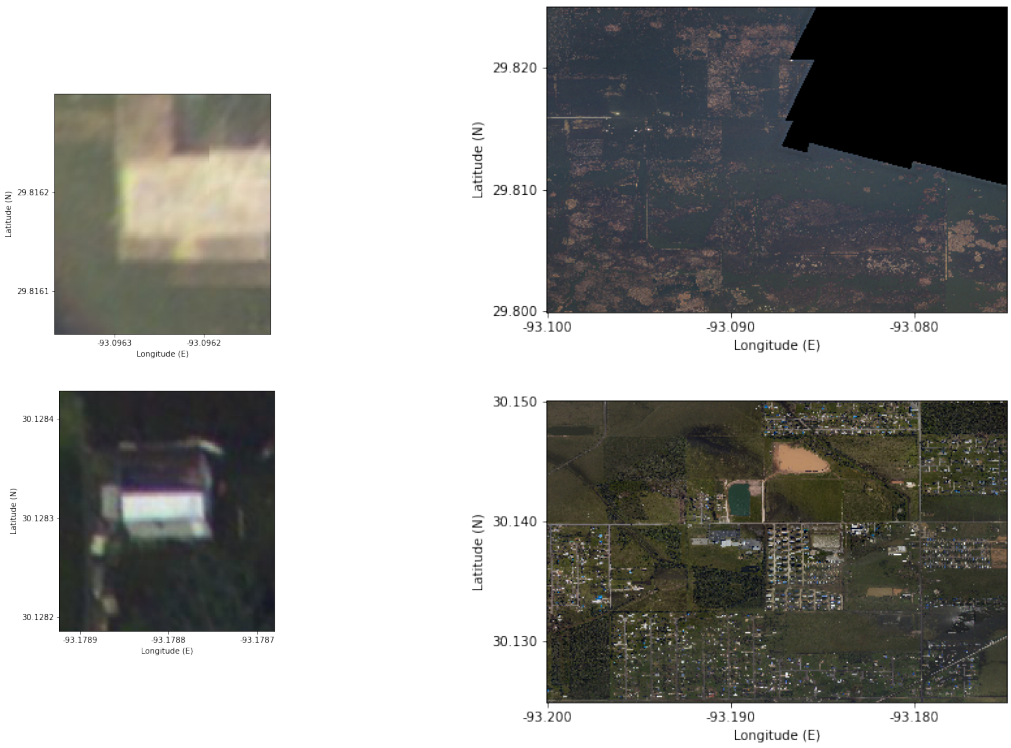


Figure 6: Example of bad NOAA patches

At a building and region level we see a good accuracy but a low recall. Unlike for dataset 1a & 1b, the model is significantly under-estimating damage. Looking at the NOAA source data for Delta, this seems to be caused by a strong visual difference in the patch images - see Figures **??** and **??**. Visual inspection reveals a much higher proportion of blurred patches, particularly for the damaged category. The smearing causes patches to appear 'smoother' which the model then mistakes for undamaged roofs and buildings. Given Delta imagery is from October, while Ida imagery from the end of August, it makes sense that visibility and weather conditions would have deteriorated causing the decline in data quality.

These results also concur with the results of Francesco (2019) who found that visual similarity between the training set and test set was the most important factor in classification performance [5].

3.3.2 Dataset 2b

Finetuning on dataset 2b was a very short process (runtime of 11m 31s, 19 epochs). The updated model achieved a 51% improvement in recall, and 15% improvement in accuracy. Recall at the building level after finetuning is 70.7%, which captures much more of the damage. The full building and region level results for the finetuned model on the remainder of the data are given in Table 10.

Metric	Building	Block	Census Tract
Accuracy	72.0%	75.0%	88%
Recall	70.7%	63.8%	57.9%
Precision	73.4%	35.9%	64.7%
F1 Score	72.0%	45.9%	61.1%

Table 10: Delta Test Results - with finetuning

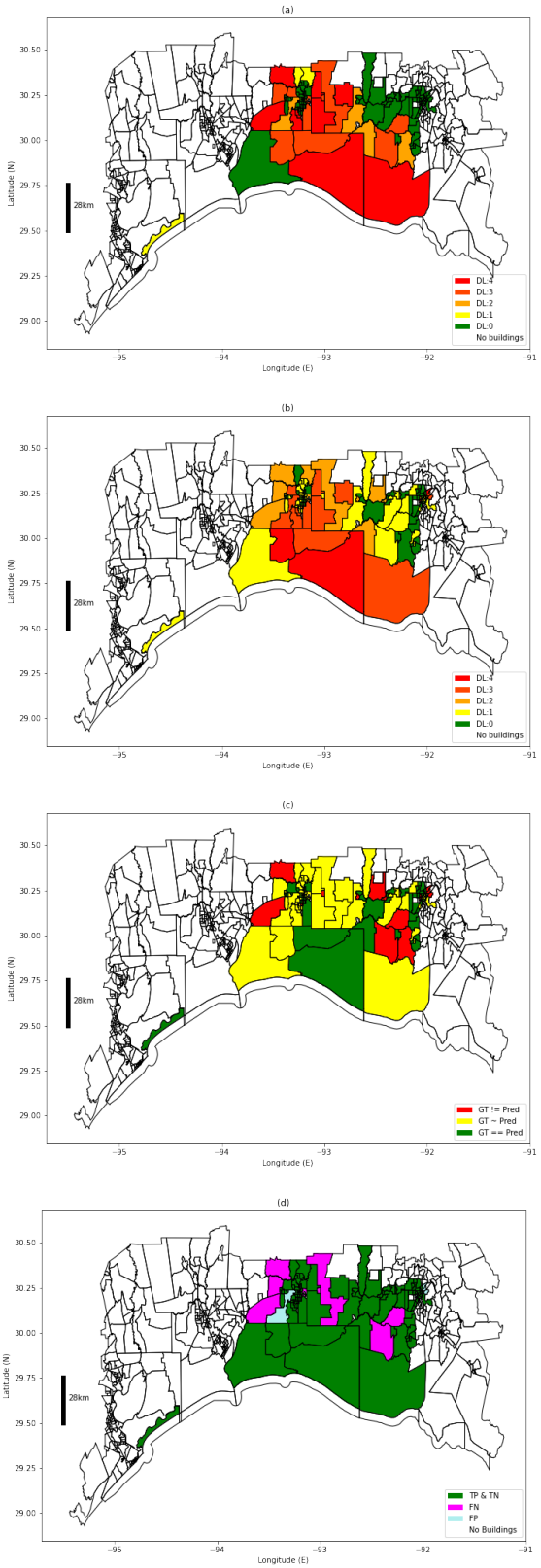


Figure 7: (Damage maps for dataset 2b, created on the Test subset. (a) and (b) show damage maps for Ground Truth and Model Prediction respectively (c) shows difference between Ground Truth and Prediction, with yellow representing a difference of 1 DL. (d) highlights the FP and FN, using definitions given in Section 2. Note that 'No Buildings' refers to within the test subset.

3.4 Comparisons with Previous Works

As mentioned in Sections 1.1, much of the literature uses multi-class damage classification instead of binary, limiting the amount of direct comparisons that can be made. The most direct comparison can be made between this work and Chen 2021 [1], which achieves a binary damage classification accuracy of 84.0% on their test dataset of the June 2018 Eureka-Kansas tornado, with DenseNet161 architecture. The relevant result here is the building level performance on the dataset 1b test subset, 83.6%. Despite giving the intersection-over-union (IOU) metric in earlier sections, [1] do not give a F1-score or IOU for this result, despite mentioning the issues of an unbalanced dataset. Based on visual comparisons of damage charts, we maintain that our work achieves a better recall on classifying damaged areas.

4 Limitations and Extensions

Here we discuss the technical and domain-transferability limitations, and subsequent extensions of this work.

4.1 Technical Limitations:

The process of creating the patches from building footprints was slow. For the Delta test patches (3k) the process took 8 hours. This is due to the merge of multiple GeoTiff files from multiple NOAA runs. Improvements were made by cropping the GeoTiff to the relevant patch size before merging, and by capping number of merged GeoTiff to 3. When generating dataset 1b, a manual multi-threading process was used, and generation took around 4 days. An extension would be to implement automatic multi-threading to speed up patch generation. Gridded data generation on urban areas (discussed further below) would also likely be a faster process.

4.2 Data Limitations

We have already discussed the inherent limitations in the damage labels as the outputs of the proprietary RMS model. A good further test of our work would be comparing performance after application to other high resolution datasets that have been labelled manually by experts.

Another limitation is in the NOAA imagery: the flight paths cover both urban and rural areas, and often focus on the coastlines. There is an implicit assumption that all of the damaged areas are covered by NOAA, but if this is not the case then some damaged areas risk being 'unseen'. When examining the Ida imagery, only the coastal sides of New Orleans were covered by the NOAA imagery, which prevented generation of a full damage map of the city. We decided to focus on the NOAA imagery given its high resolution, but an extension here would be to examine other data sources and seek to unify potentially lower resolution imagery with the NOAA imagery.

The missing-data concern applies equally if methodology is ported to another region: focusing aid efforts where damage can be mapped risks leaving unmapped communities behind. Methods of mitigating this risk depend on which organisation is using it: for small NGOs, partners with local knowledge can aid in finding the unmapped communities. For large NGOs or governments, focus should be on mapping these areas, and putting policy in place to fill in the gaps in the datasets. Using traditional satellites with fixed footprints can also play a part: whilst lacking the high resolution needed for damage classification, algorithms for building extraction could be used to verify total coverage by the UAV or private satellites.

4.3 Limitations in Domain Transferability

While results on Datasets 1a & 1b were very strong, the model struggled with classification on dataset 2a, which, despite having the same origin, was quite different visually. An extension would be to try and train a general model with source data from a variety of disaster events to improve performance on different regions. Furthermore, high resolution data is not available everywhere. Several potential avenues could be explored here to improve wider usage:

- Research and document all high resolution data sets available

- Test models performance on lower resolution data
- Partner with an NGO to fly UAVs in the immediate aftermath of hurricanes in areas with limited data availability. This would require significant financial outlay and other expertise. A potential avenue could be to partner with NOAA to provide support and domain expertise.

This model was built using building footprints to generate training and validation patches. Whilst these are available in the USA, building footprints are not easily available for many of the places which are vulnerable to hurricanes, such as the Caribbean. An excellent extension of this work would be to test the model performance when building footprints are not available. We see two options available for extension:

- Generic Grid Method: patches created by diving an urban area into a grid. We chose to do patch-wise image classification at a median scale with the aim that that the model would abstract well to places without building footprints available.
- Footprint Pipeline: use one of the existing building extraction methodologies in the literature [23], [24], [25] to first extract building outlines from imagery before the hurricane hits, and then apply the methodology in this report to the post-hurricane imagery. This method would potentially add considerable extra time to the damage classification process, and would require high-res imagery from before and after the hurricane.

Acknowledgements

The author is grateful to Giovanna Trianni and Robert Muir-Wood at Risk Management Solutions for their expertise and data they provided, and to both Emily and Christian for their excellent supervision.

5 Declarations

This report is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text and/or bibliography.

6 Code and Data Availability

All code is available in the GitHub repository LINK All data is available

References

[1] Zhiang Chen, Melissa Wagner, Jnaneshwar Das, Robert Doe, and Randall Cerveny. Data-Driven Approaches for Tornado Damage Estimation with Unpiloted Aerial Systems. *Remote Sensing*, 13, April 2021.

[2] Chih-Shen Cheng, Amir H. Behzadan, and Arash Noshadravan. Deep learning for post-hurricane aerial damage assessment of buildings. *Computer-Aided Civil and Infrastructure Engineering*, 36(6):695–710, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mice.12658.

[3] Milad Fallahian, Faramarz Khoshnoudian, and Viviana Meruane. Ensemble classification method for structural damage assessment under varying temperature. *Structural Health Monitoring*, 17(4):747–762, July 2018. Publisher: SAGE Publications.

[4] Yundong Li, Shi Ye, and Ivan Bartoli. Semisupervised classification of hurricane damage from postevent aerial imagery using deep learning. *Journal of Applied Remote Sensing*, 12:1, October 2018.

[5] Francesco Nex, Diogo Duarte, Fabio Giulio Tonolo, and Norman Kerle. Structural Building Damage Detection with Deep Learning: Assessment of a State-of-the-Art CNN in Operational Conditions. *Remote Sensing*, 11(2765), November 2019.

[6] Tanya M Brown, Daan Liang, and J Arn Womble. Development of a Statistical Relationship between Ground-Based and Remotely-Sensed Damage in Windstorms. Thesis, 11th Americas Conference on Wind Engineering, June 2019.

[7] Jim Thomas, Ahsan Kareem, and Kevin W. Bowyer. Automated Poststorm Damage Classification of Low-Rise Building Roofing Systems Using High-Resolution Aerial Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(7):3851–3861, July 2014. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[8] Vito Romaniello, Alessandro Piscini, Christian Bignami, Roberta Anniballe, and Salvatore Stramondo. Earthquake damage mapping by using remotely sensed data: the Haiti case study. *Jan–Mar 2017*, Vol. 11(1), March 2017.

[9] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. Conference Name: Proceedings of the IEEE.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[11] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. Technical Report arXiv:1512.00567, arXiv, December 2015. arXiv:1512.00567 [cs] type: article.

[12] MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, April 2017. Number: arXiv:1704.04861 arXiv:1704.04861 [cs].

[13] Patrick Aravena Pelizari, Christian Geiß, Paula Aguirre, Hernán Santa María, Yvonne Merino Peña, and Hannes Taubenböck. Automated building characterization for seismic risk assessment using street-level imagery and deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 180:370–386, October 2021.

[14] https://storms.ngs.noaa.gov/.

[15] John L. Beven II, and Robbie Berg, and Andrew Hagen. NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT HURRICANE IDA, April 2022.

[16] John P. Cangialosi and Robbie Berg. NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT HURRICANE DELTA, October 2020.

[17] National Geodetic Survey, 2022: 2021 NOAA NGS Emergency Response Imagery: Hurricane Ida. Technical report, National Geodetic Survey.

[18] National Geodetic Survey, 2022: 2020 NOAA NGS Emergency Response Imagery: Hurricane Delta,. Technical report, National Geodetic Survey, 2022.

[19] Paszke, Adam, Gross, Sam, and Massa, Francisco. PyTorch: An Imperative Style, High-Performance Deep Learning Library.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. Technical Report arXiv:1512.03385, arXiv, December 2015. arXiv:1512.03385 [cs] type: article.

[21] Biewald, Lukas. Experiment Tracking with Weights and Biases, 2020. Software available from wandb.com.

[22] US Census Bureau. Mapping Files. https://www.census.gov/cgi-bin/geo/shapefiles/index.php.

[23] Zhenfeng Shao, Penghao Tang, Zhongyuan Wang, Nayyer Saleem, Sarath Yam, and Chatpong Sommai. BRRNet: A Fully Convolutional Neural Network for Automatic Building Extraction From High-Resolution Remote Sensing Images. *Remote Sensing*, 12(6):1050, January 2020. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[24] Giorgio Pasquali, Gianni Cristian Iannelli, and Fabio Dell'Acqua. Building Footprint Extraction from Multispectral, Spaceborne Earth Observation Datasets Using a Structurally Optimized U-Net Convolutional Neural Network. *Remote Sensing*, 11(2803), 2019.

[25] Nitin L. Gavankar and Sanjay Kumar Ghosh. Automatic building footprint extraction from high-resolution satellite image using mathematical morphology. *European Journal of Remote Sensing*, 51(1):182–193, January 2018. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/22797254.2017.1416676.

[26] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8):5455–5516, December 2020.

[27] Vihar Kurama.

Appendices

A Data

The NOAA data for each hurricane event consists of multiple GeoTIFFs, taken over multiple flights over several days.

The detailed preparation of the training data:

1. List of building footprints filtered to include all damaged building (50k)
2. Out of the remaining 150k undamaged buildings, 50k selected after random shuffle
3. To ensure no edge effects from edge of aerial imagery, filter to only include buildings that sit within 99.99998% of the aerial imagery
4. For each building footprint, find relevant GeoTIFF that building polygon falls within
5. Crop GeoTIFF to patch aligned around the centroid of the building footprint
6. If there are multiple GeoTIFFs from different flights, merge these patches
7. Save patch into relevant folder based on Train / Test split and Damage / Non Damage label

B Models

B.1 Model Pre-Processing

The following transforms were applied to all datasets:

- Resize patches:
  - Model 1A & 1B: 180 x 180
  - Models 2A & 2B: 244 x 244
  - Model 3: 299 x 299
- Normalisation
  - All patches normalised using the following RGB mean and standard deviation: $(0.416, 0.416, 0.36)$, $(0.215, 0.21, 0.21)$

B.2 5-Layer-CNN

The model architecture for the 5-Layer-CNN is given in Fig 10. Model 1B was run first, and then Batch Normalisation (BN) was added after each of the five convolutional layers. The goal of testing BN was to see the impact on model performance and training time.

B.3 ResNet Model Architecture

ResNet architecture (short for Residual mapping) was introduced by Microsoft to solve the degradation of accuracy problem when training deep neural networks. With the use of residual mappings, seen in Figure 11, accuracy no longer plateaus when depth of network is increased, and optimisation is easier due to fewer parameters than networks of comparable depth.

The ResNet Models in this paper were implemented in Pytorch Lightning. ResNet in its 'raw' form outputs 1000 classes, based on the 1000 classes of ImageNet dataset. In order to use in binary classification task, the final classification layer was updated to 2.

Sandbox Region RGB Image

Figure 8: Image of the Sandbox data region, chosen for its high proportion of damaged buildings, stitching together 18 GeoTIFFs

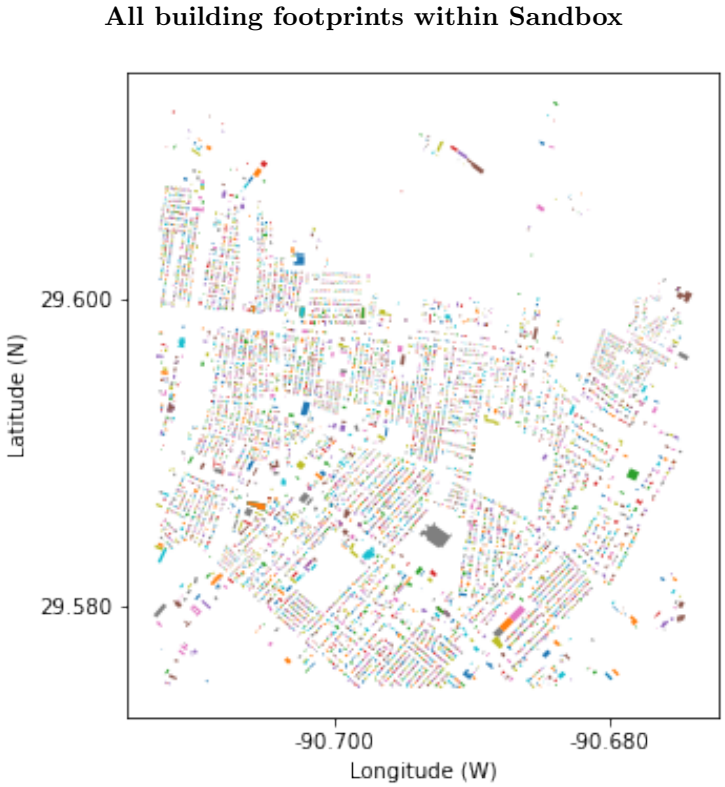


Figure 9: All of the building footprints within the Sandbox region

B.4 Inception V3 Model Architecture

Inception V3 architecture is built up progressively, containing all the following components:

1. Factorized Convolutions: reduces parameters and hence improves efficiency
2. Smaller convolutions: replacing larger kernels with several smaller ones
3. Asymmetric convolutions: Replacing 3 x 3 by 1 x 3 and 3 x 1, fewer parameters compared to 2 x 2
4. Auxiliary classifier: Small CNN inserted between layers, acting as a regularizer
5. Grid size reduction: usually done through max pooling, in an asymmetrical manner

In our Pytorch Lightning implementation, in order to use in binary classification, both of the main and auxiliary CNNs must have the final layer updated to output 2 classes rather than 1000.

Inception architecture is characterised by multiple convolutional kernels applied simultaneously, as seen in Figure 12.

B.5 Sandbox Data Results

The best performing model configuration was:

1. Batch Size: 184
2. Learning Rate: 0.006607
3. Optimiser: Adagrad
4. Weights: [0.2,1]

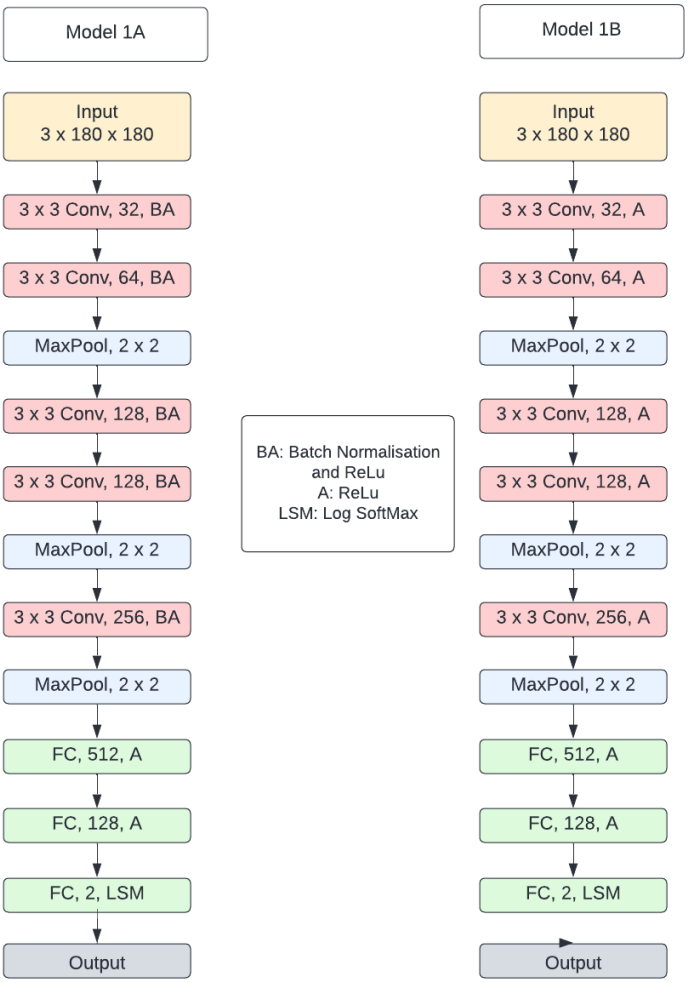


Figure 10: Model Architecture for 5-Layer-CNN

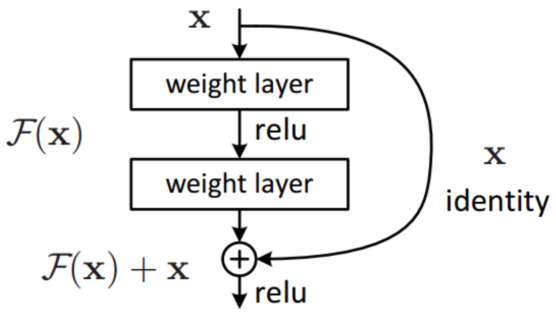


Figure 11: Example of the building block of ResNet: the Residual Block [20]

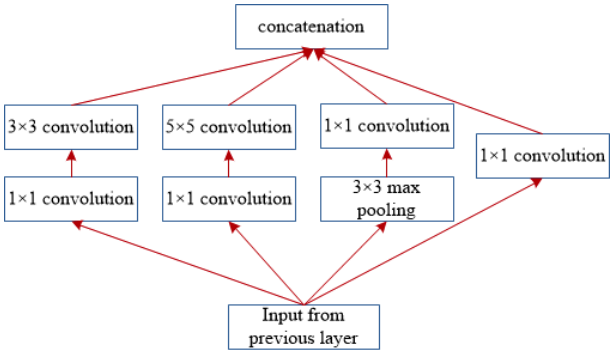


Figure 12: Inception Block [26]

This model ran for 49 epochs and took 10 hours to train, and achieved a validation accuracy of 84.9%. Due to the intended uses of this work, in near-time disaster relief, when compared to the second best model, this model was unfavourable due to the long training time.

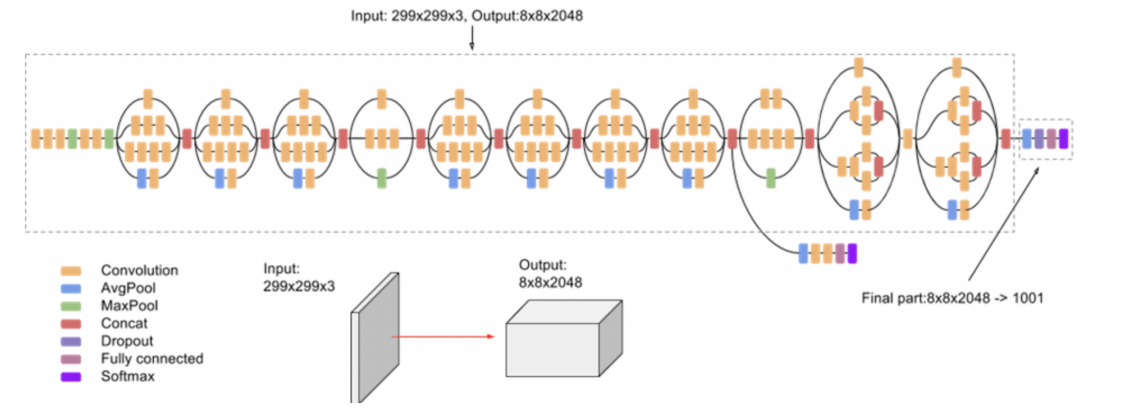


Figure 13: Full Inception V3 Architecture [27]

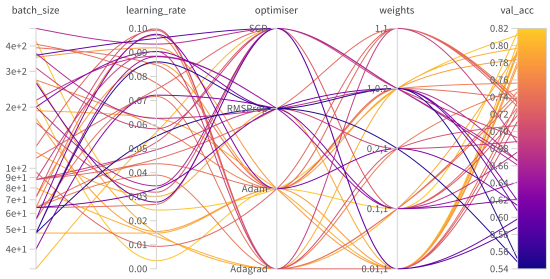


Figure 14: Parameter Sweep for Sandbox Data with Resnet50 model architecture

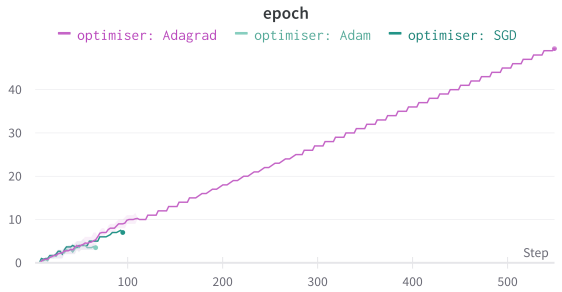


Figure 15: Epoch for Resnet50 per optimiser

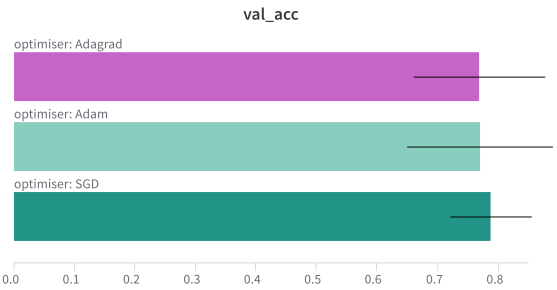


Figure 16: Validation Accuracy per Optimiser from Full Data Parameter Sweep

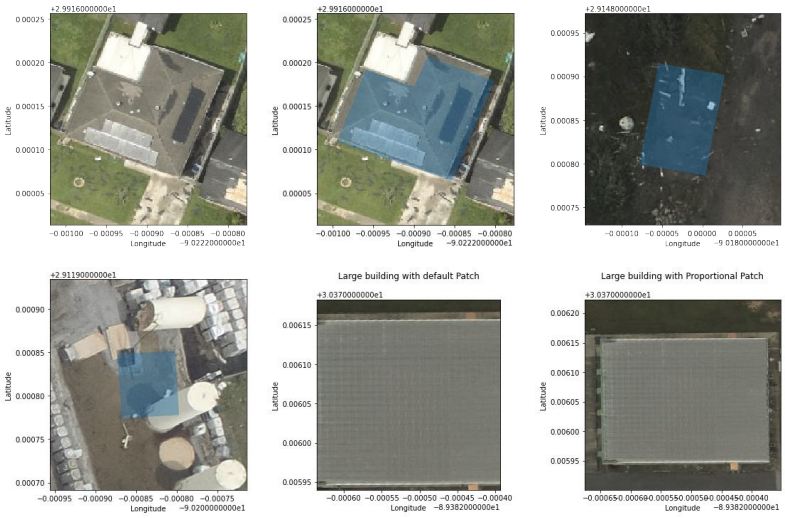


Figure 17: (a) Median Size house and default patch size. (b) Building footprint overlaid on Median house - good match (c) Example of missing building - incorrect footprint. (d) Building footprint over oil refinery - incorrect footprint. (e) Large house (top 16th percentile) with default patch size (f) Large house with proportional patch size