Oversampling Social Media-Sourced Image Datasets for Better Deep Learning Classification of Natural Disaster Damage Levels

Nicholas Lau Kheng Seng[1], Goh Wei Wei[2], Tan Ee Xion[3]
School of Computer Science, Taylor's University, Subang, Malaysia[1, 2]
Dept. of Digital Health and Health Informatics, School of Business and Technology, IMU University, Kuala Lumpur, Malaysia[3]

***Abstract*—People in areas affected by natural disasters and use social media websites such as Facebook, Twitter (also known as "X") and Instagram tend to post images of damage to their surroundings. These social media sites have become vital sources of immediate and highly available data for providing situational awareness and organisation for natural disaster response. A few previous attempts at classifying the level of natural disaster damage in these images using image processing techniques had noted the challenge in producing robust classification models due to the effect of overfitting caused by a lack of observations and data imbalance in annotated datasets. This article shows an attempt to improve a training strategy within the data level for deep learning models such as VGG16, ResNetV2 and EffecientNetV2, used to estimate the level of disaster damage in images by training them with data generated using image data augmentation with data balancing, oversampling up to eight times and combining the oversampled image data collections. The F-1 score achieved for classifying damage on earthquake images and images from the Hurricane Matthew data collection by training EfficientNetV2 on a generated dataset made with a combination of oversampled data surpassed previous benchmark results. These results show that using data balancing and oversampling on the dataset prior to training deep learning models on these datasets result in increased robustness.**

Keywords*—Deep learning; image processing; oversampling; image data augmentation

I. Introduction

Natural disasters that have occurred from 1998 to 2017 have caused $2.9 trillion US in monetary damage and have cost the lives of 1.3 million people [4]. Worldwide insured losses from natural and man-made disasters in 2017 alone are estimated to cost $144 billion US according to a report by the Swiss Re [32]. Damages caused by Hurricane Ian in 2022 have been projected to cost exceeding $45 billion USD [10].

During the disaster response process, an assessment of damage done is made typically by door-to-door survey. This approach may cause a "cold start" issue to obtain and analyse the data. The time required to acquire and the complexity of annotating the data for damage assessment may also take several days to weeks when using the traditional "boots-on-the-ground" method.

Victims of natural disasters have been using social media posts, including image posts, to communicate and update their status during a disaster event [13]. The information extracted from these social posts have been useful in providing situational awareness in disaster response [27].

Data from social media websites are multi-dimensional; they are generally represented in four dimensions which are space, time, content and network [2]. Image data from within the content dimension in addition to the spatial and temporal dimensions may contribute to gaining situational awareness regarding ongoing natural disasters. Endsley [31] defined situational awareness as "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future".

Frameworks for collecting and annotating information regarding ongoing disaster events such as Artificial Intelligence for Disaster Response (AIDR) have been developed to collect data that can be used to combine human intelligence with Natural Language Processing (NLP) and Machine Learning (ML) models [26]. A system for collecting and annotating images with natural disaster damage from social media had been integrated into AIDR [35].

Deep Neural Networks (DNNs) such as Convolutional Neural Networks (CNNs) have been used to process digital images in various tasks such as image classification and object detection. VGG16 is a CNN that has been implemented to classify images from social media in the ImageNet challenge [9]. An adaptation of the VGG16 image classification model had been used to classify the level of disaster damage in images from social media based on intensity [15]. Other models such as ResNet50, InceptionNet, EfficientNet and MobileNet have been explored as replacements for VGG16 for classifying the severity of natural disaster damage [34].

This article explores the use of data balancing and oversampling in conjunction with image augmentation on a labelled image dataset containing images of natural disaster damage in various levels. This article aims to investigate the effects of using these data manipulation techniques with the goal of improving a training strategy for training image DNN disaster damage level image classifiers.

The rest of this paper is organized as follows. Section II details the works relating to DNNs, image processing methods used in disaster management, image data augmentation methods for oversampling, and issues in using image classification in disaster management. Section III presents the methodology

Spatial Big Data and IoT for coastal erosion and floods mitigation and prediction [grant number TUFR/2017/004/03].

aimed to improve the performance of current models. Section IV shows the results gained from this study, its limitations and discusses the achievements of this study. Lastly, Section V concludes this article and shows recommendations for further study.

II. Related Work

This section relays related work regarding literature on deep neural networks, their use within natural disaster response and current issues.

A. Deep Neural Networks

DNNs are a type of Artificial Neural Network (ANN) with many hidden layers, where in each layer the aggregation of input or activation signals of the prior layer is transformed [12]. DNNs have been used to implement image recognition and detection of intricate structures [21].

CNNs are a type of DNN which includes layers which use convolutions to transform the input. CNNs are often used in image processing tasks such as optical character recognition, image classification and object detection.

The addition of backpropagation to CNNs allows for the neural network to learn convolutional kernel coefficients from the dataset [20]. Prior to this, weights in CNN had to be designed manually. This led to the development of the LeNet model [19].

VGG16 is a relatively modern CNN which won second place in the ImageNet Large Scale Visual Recognition Challenge in 2012. This model uses groups of CNNs followed by fully connected layers to extract and classify combinations of features to classify images. Compared to other competitors, VGG16 was designed with smaller receptive fields (starting with 3x3) convoluted by a stride of 1 pixel and had a greater "depth" by increasing the number of convolutional weight layers [9].

ResNet is an image classification deep residual network consisting of a network of "residual units" where each of these units consists of skip networks [29]. An improved version known as ResNetV2 adds full pre-activation to the skipping vertices prior to addition [28]. Both ResNet and ResNetV2 have 50, 101 and 152 layer versions.

EfficientNet is an image classification deep neural network made as an attempt to create a scalable convolutional neural network [6]. EffecentNet's architecture is based on MnasNet which uses successive MBConv layers. The modification made by the authors of EfficientNet were that they had proposed a compound scaling method that scales the width, depth and resolution of these layers. An updated version of EfficientNet known as EfficientNetV2 which replaced MBConv layers with Fused-MBConv layers which resulted in faster and smaller models [5].

B. Data Collection and Annotation Methods in Natural Disaster Response

Disaster response is the second stage of natural disaster management carried out immediately after a natural disaster event. Traditional methods for collecting data for natural disaster response typically uses remote sensing or optical imagery from satellites but may be susceptible to noise from the effects of weather while being costly and time consuming to setup [25]. Social media provides an easy and immediate source of data to collect from. This allows for a quicker start to disaster response by collecting data from social media sites. This data includes text, images, videos, geospatial and temporal data [2]. Images of natural disaster damage are often uploaded to social media sites during and immediately after natural disaster events [13].

Image analysis for natural disaster response starts in the collection of image data. Image data is either scraped from social media sites or captured via aerial photography. This data is then annotated, based the parameters of the intended image processing method. For example, the AIDR [26] platform collects image data from Twitter (also known as "X"), uses volunteers to annotate the image data to form a dataset, then splits the data into a training set, validation set and a testing set, with the goal of developing a machine learning model that classifies the level of disaster damage shown in image data collected in the future. This platform has been used to gather and annotate data from social media on the regarding natural disasters events [14], [15], [16], [33], [34], [35].

Studies using data collected using AIDR have explored using various machine learning tasks such as NLP, computer vision tasks, and multi-modal learning [33]. For example, data collected during various natural disaster events were used to train machine learning models to detect natural disaster damage [14], classify the level of natural disaster damage [15], [35], classify the type of damage caused by natural disasters [16], [34].

C. Detection of Natural Disaster Damage in Images

There are multiple methods to detect natural disaster damage in images, often through the use of deep learning models trained as a binary classifier to detect the presence natural disaster damage in images.

A natural disaster damage image detector was used in [14] to filter posts based on the presence of natural disaster related content in conjunction with perceptual hashing with the intention of reducing the workload of human annotators. This classifier was implemented using a pre-trained VGG-16 CNN that has been fine-tuned to classify images that are relevant to disaster damage and achieved an almost perfect F-1 score of 0.98.

Another approach to detecting natural disaster damage in images is to use a (Single Shot-MultiBox Detector) SSD to detect natural disaster damage in images. A two-part SSD based on VGG-16 was used to detect natural disaster damage in aerial images [17]. This SSD was trained with a dataset that had been oversampled with augmented images.

Detection of urban flooding in crowd-sourced image data has been used to locate occurrences of urban flooding using the Clarifai object recognition model as a way to monitor urban flooding and to validate urban flooding models [3]. The Clarifai object recognition model was developed as a contender in the 2013 ImageNet LSVR challenge. This object recognition model was used via an online API and was used to provide a list of tags relevant to each image fed to the model as well as the probability of each tag.

Another effort relating to floods explored the use of InceptionNetv3 and DenseNet to classify images of flood based on severity as part of an image sorting system [24]. Both models used had an F-1 score of 0.82.

D. Classification of Natural Disaster Damage in Images

Damage depicted in image can classified into their respective classes following one or more taxonomies. Classification of posts often use modified versions deep learning models used to implement detection of damage depicted in social media posts. Social media posts that depict damage have been categorised by damage intensity, relevance, type of damage, or within a multi-dimensional taxonomy which may include a combination of the prior categorisation schemes.

Hence, image processing models have been used to classify disaster damage in images collected from social media. CNNs such as VGG16, InceptionNetV3, InceptionNetV4 and InceptionResNet have been used to classify the severity of disaster damage shown in the image into either the severe damage, mild damage or no damage classes. These models have been found to have comparable performance in various use-cases such as in classification of natural disaster damage levels [15], and classification of the types of natural disaster damage.

Earlier explorations at classifying images of natural disaster damage included training a VBoW model against an annotated disaster damage dataset but it was found that using pre-trained CNNs such as VGG16 pretrained on the ImageNet dataset performed better. Fine tuning this pre-trained VGG16 model with the same training data further increased the F-1 score [15].

TABLE I. PERFORMANCE OF BOVW AND VGG16 AGAINST DISASTER DAMAGE DATA COLLECTIONS

Event	Model	Accuracy	Precision	Recall
Nepal Earthquake	BoVW	0.78	0.77	0.78
	VGG-16-fc7	0.76	0.76	0.78
	VGG-16-fine-tuned	0.84	0.82	0.84
Ecuador Earthquake	BoVW	0.82	0.81	0.82
	VGG-16-fc7	0.82	0.82	0.84
	VGG-16-fine-tuned	0.87	0.86	0.87
Hurricane Matthew	BoVW	0.64	0.64	0.64
	VGG-16-fc7	0.63	0.63	0.64
	VGG-16-fine-tuned	0.74	0.73	0.74
Typhoon Ruby	BoVW	0.73	0.74	0.73
	VGG-16-fc7	0.79	0.80	0.80
	VGG-16-fine-tuned	0.81	0.81	0.80
Google images	BoVW	0.57	0.53	0.56
	VGG-16-fc7	0.60	0.63	0.64
	VGG-16-fine-tuned	0.67	0.67	0.67

According to study [1], training the last layer of a pre-trained model allows the use of a smaller dataset to transfer the capabilities of the model to train for a different task. This method also reduces the time spent on training the model as a smaller dataset is used. This technique is also known as transfer learning [1].

Table I compares the accuracy, precision, recall rate, and F-1 score (balance between precision and accuracy) of Bag-of-Visual-Words, VGG16-fc7 and the fine-tuned VGG16 models when trained with single event datasets with image data collected from social media related Nepal earthquake, Ecuador earthquake, Hurricane Matthew, Typhoon Ruby and images of damaged buildings from Google Images. The table shows that in some cases (e.g., Nepal earthquake, Ecuador earthquake), VGG16 without fine-tuning was comparable to visual bag-of-words for three of the image collections but surpassed the BoVW model's performance for the Typhoon Ruby and Google Images collections. The fine-tuned model achieved a higher F-1 score on all image collections.

A similar dataset was included in a benchmark image dataset compiled by [34] to benchmark various image classification models on various tasks. The F-1 scores obtained for classifying images of disaster damage levels using EfficientNetB1 was 0.758 compared to an F-1 score of 0.753 for VGG16.

Another notable implementation by [16] of a classifier for identifying damage in a social media post analysed both text and image data collected from Instagram posts via a fusion of image and text classifiers using a multi-modal approach. This multi-modal classifier was described using the Inception architecture to classify images in social media by several InceptionNet models to classify the type of disaster damage in images from social media. These classes were named: Infrastructure, Nature, Fires, Floods, Human and Non-Damage.

The inception network used was described as a layered stack of "Inception modules" with each layer consisting of multiple convolution filters with a variety of sizes. The Inception Network models have been described to have near state-of-the-art accuracies with the ImageNet dataset while having a relatively smaller model compared to other CNN models.

Mouzannar et al. [16] compared the performance of four DNN models which were InceptionNetv4, InceptionNetv3, VGG16 and InceptionResNet. InceptionNetv4 scored a higher validation accuracy while InceptionNetv3 scored a higher test accuracy. The higher validation accuracy of the InceptionNetv3 model led to its selection as a component for a multi-modal classifier.

The disaster damage severity task was revisited by [34] as part of an effort to build a consolidated benchmark dataset. A subset of this benchmark dataset includes the dataset by [15]. The F-1 score for classifying damage severity using EfficientNetB1 is 0.758, slightly higher compared to the F-1 score obtained using VGG16 which is 0.753.

Classification of natural disaster damage in images can also be used in image segmentation. Class activated mapping can be carried out via CNNs to classify areas of mild damage, severe damage or no damage done to an area depicted within these images [18]. This information can be used to generate a heatmap

visualisation of damage shown within a given image by finding gradients between segments of the image with damage, and segments without damage. This heatmap can then be used to calculate a Damage Assessment Value for each image.

E. Issues with Classifying Natural Disaster Damage Levels in Images

The research in [15] used a fine-tuned VGG16 model to classify image data by the level of disaster damage with a dataset that had a small amount of image data. The amount of image data in the dataset was limited due to the issues caused by the complexity of data annotation tasks, unintended collection of irrelevant data, the subjectivity of the data annotation tasks and time limitations when collecting and annotating the data. The dataset used is imbalanced; there are far fewer images labelled as mild within the dataset than images with other labels.

This issue was also highlighted by [34] when training other deep learning models with the same classification task. Alam had noted that the number of images labelled as "mild" was lower and that models trained for the damage severity task tend confuse images with this label as images with other labels. Efforts to overcome limitations in the robustness of deep learning models for various disaster informatics tasks was noted by [33].

The study in [14] revisited this and trained a fine-tuned VGG16 model to filter posts that are not related to natural disasters. This model is then combined with a perceptual hash function to filter posts that were irrelevant or duplicates. This filtering task reduced the amount of image data in the dataset by 62%. Training a model with a dataset that has a low amount of data can cause overfitting [11]. The dataset that was used by [15] was noted to be imbalanced such that the recall rate for minority classes was much lower than majority classes leading to many false negatives.

Training with imbalanced data can also increase the overfitting issue [23]. Overfitting is a phenomenon that causes deep learning models to strictly conform with a training dataset as a result of training with a training dataset that has a low amount of data. This causes the model to have a lower validation score, causing the number of false positives and false negatives to increase. The resulting trained VGG16 models by [15] achieved precision-recall rates for the minority class that were significantly lower compared to other classes and identified that the lack of labelled training data was the cause of this issue. [7] had shown that the effect of overfitting in CNNs used in image classification decreases as the number of observations increases. The VGG16 deep learning model uses dropout layers as a way to reduce this overfitting [9]. Dropout layers regulate overfitting by removing connections between layers [8].

F. Image Data Augmentation

Image Data Augmentation refers to the use of one or more image manipulation techniques, often used in conjunction with image data oversampling with the goal of reducing the effect of overfitting when training deep learning models. Overfitting is a phenomenon where a deep learning model is trained such that it has a high variance to fit the training data [11]. Overfitting can cause a stall in validation accuracy when training deep learning models to generalise the dataset. Larger datasets have been regarded as resulting in deep learning models with higher overall qualitative performance [7].

Transformations on images in the dataset include geometric, colour space manipulations and noise injection. Various geometric transformations can be applied to images in data augmentation. These geometric transformations include flipping, cropping, rotating and shifting. Colour space manipulations include applying a coloured tint or filters commonly found within photo editing applications. For example, training a classifier with the ImageNet or CIFAR-10 datasets would yield better results when vertical axis flipping is used while slight rotations can help in training with text recognition datasets such as MNIST [11].

The research in [30] used image data augmentation in combination with weight decay an various tasks and found it had significantly improved the performance of InceptionNetv3.

G. Oversampling

Oversampling is a data level technique which inflates the number of samples in a given dataset. Oversampling is often used to increase the signal to noise value by adding samples augmented with unrelated noise to the dataset before using any processing method. In the case of image processing, the dataset is inflated with augmented images [11]. Random oversampling can significantly improve classification of images and was found to have the best performance against other data-level methods such as undersampling, two phase training and thresholding [23].

H. Combining Data-Level Techniques

Data level techniques such as image data augmentation combined with oversampling can be used to improve the validity and robustness of social media image classification models especially within the limitations in natural disaster management. Image data augmentation utilises a collection of image manipulation techniques used as a technique for oversampling in image datasets which could play an important role in reducing the effect of overfitting brought about by training DNNs with smaller datasets. The next section discusses the methodology used in combining image data augmentation with image data oversampling.

III. Methodology

This section details the methodology in pre-processing a labelled dataset by oversampling it with augmented images, followed by training VGG16 with the pre-processed dataset, and testing the trained model and analysis of the results.

A. Experiments Carried Out

Three sequences of deep learning experiments are carried out. The initial sequence of experiments consists of a grid search of oversampling levels for each data collection using VGG16. This first sequence of experiments is also used to search for "early stopping" parameters.

The second sequence of experiments compares the performance of deep learning models selected after training each of theses models against the Nepal data collection. The deep learning models used in this sequence are VGG16, ResNet50V2 and EfficientNetV2B0.

The last of these experiments uses the optimal datasets acquired form the first experiment to train best performing model in the second experiment. The performance of the resulting trained model is then compared with the published performance of state-of-the-art models.

B. Equipment and Software used

The training of the image classification models was carried out on a computer with a graphics card capable of training deep learning models. This computer was built around an NVidia RTX2060 SUPER graphics card which has 8 GDDR6 RAM and 272 tensor cores.

As for the software, Ubuntu Server 20.04 LTS was installed without any desktop environment such that the computer can be operated headlessly (without a desktop GUI) to reduce GPU RAM usage. Python 3.7, TensorFlow version 2.2.0 and the included Keras library was used together with Jupyter Notebooks to implement image data augmentation, oversampling, model training, and model testing.

C. Dataset Details

The dataset used in this study is a data collection containing images collected from social media regarding several disaster events including the 2015 Nepal earthquake, the 2016 Ecuador earthquake, Hurricane Matthew in 2016 and Typhoon Ruby in 2014. These images were collected and annotated using the AIDR platform [26] which was modified to work with images [35].

This data collection contained three comma-separated value (CSV) files labelling 1,584 images with no natural disaster damage, 451 images showing mild natural disaster damage, and 1,785 images showing severe natural disaster damage together with these images. Each CSV file also divided the images into training (60%), cross-validation (20%) and testing (20%) sets. This data split is a common arrangement for training with cross-validation.

The dataset that contains this image collection can be downloaded from https://crisisnlp.qcri.org/ under Resource # 9. This dataset is also included in a benchmark dataset published by [34] from the same website under Resource #15.

D. Pre-Processing

During pre-processing, the images were first sorted based on their label to respective directories and split into either the training, validation or testing data split using the CSV files included such that the data collection can used with the ImageDataGenerator object from Keras.

After sorting the images, image data augmentation and various levels of oversampling were applied to generate augmented and oversampled datasets. The bulleted lists below show the augmentations applied to the training datasets and to validation datasets, while images in the testing data was rescaled to 1/255 only. These augmentations were selected to generate a variety of augmented images to prevent overfitting. The effect of these augmentations can be seen in Fig. 1.

List of Image Data Augmentations Used To Generate Training Data:

- Rescale values to 1/255.
- Random rotation range of -15° to 15°.
- Random width shift range of 10%.
- Random height shift range of 10%.
- Horizontal flipping.

List of Image Data Augmentations Used To Generate Training Data:

- Rescale values to 1/255.
- Horizontal flipping.

Table II shows the number of images after oversampling the images with augmented images. The control dataset does not contain any images with augmentations and does not contain any images generated for balancing or oversampling. The dataset with augmentations only contains augmented images but is not balanced or oversampled, preserving the same number of images as the control dataset. The balanced dataset contains images that have been augmented and balanced by oversampling images from the minority classes such that each class has the same number of images. The remaining datasets contain augmented images and were generated with two times, four times, and six times the number of images compared to the balanced dataset.

Fig. 1. Five augmented images were generated from an image from the Nepal Earthquake data collection.

TABLE II. NUMBER OF IMAGES IN EACH GENERATED DATASET AFTER PRE-PROCESSING THE NEPAL DATA COLLECTION

Name	Images labelled None	Images labelled Mild	Images labelled Severe	Total Images
No Augmentations (control)	4752	1354	5357	14463
With Augmentations	4752	1354	5357	14463
Augmented and Balanced	5357	5357	5357	16071
2X Oversampled	10714	10714	10714	32142
4X Oversampled	21428	21428	21428	64282
6X Oversampled	32142	32142	32142	96424

E. Deep Learning Model Implementation

This study will involve three deep learning models namely VGG16, ResNet50v2, and EfficientNetv2B0. The configuration of the VGG16 model used includes 224 pixel by 224 pixel inputs with three channels. This model was constructed using a pre-trained version of VGG16 supplied by the Keras software library which did not include dropout layers. The model was pre-trained with the ImageNet ILSVRC 2015 challenge dataset. Dropout layers were added back to the model as specified in the original implementation of VGG16 [9] with a dropout rate of 0.5 inserted before the FC1 and FC2 layers. These dropout layers were used for preventing overfitting by randomly dropping units during training [8].

The model was further modified by replacing the output layer which was originally used to classify 1000 classes in the ImageNet ILSVRC 2015 challenge, with a "dense" layer of three outputs with each output corresponding to each class found in the dataset. The L2 Kernel regularization rate for the output layer is set to 0.0005. The last layer was set to be trainable while the other layers were set to be not trainable.

The ResNet50V2 and EfficientNetV2B0 models were pre-trained with the ImageNet Dataset [22]. The ResNet50v2 model had its "head" replaced with a GlobalAveragePooling2D layer, followed by a Dropout layer with a dropout rate initially set to 0.5, and finally a Dense layer with three outputs. The head of the EfficientNetV2B0 model was replaced with a GlobalAveragePooling2D layer, followed by a BatchNormalization layer, then a Dropout layer with a dropout rate of 0.5, and finally a Dense layer with three outputs (see Fig. 2).

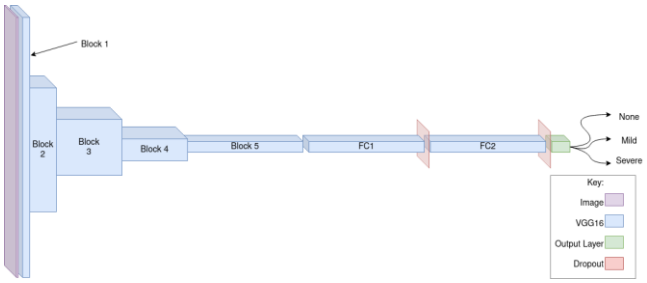

Fig. 2. VGG16 with three outputs and dropout layers.

F. Training, Validation and Testing

For the first round of training, VGG16 was initially selected as a control model. For each dataset generated in the prepossessing steps, an instance of this modified VGG16 model was trained on the dataset for up to 100 epochs. After each epoch, if the validation loss is lower than in all prior epochs, the weights of the model are saved.

During testing, the weights of each trained VGG16 instance was loaded, then tested by classifying images from the test set. The number of "true" and "predicted" occurrences is collected to calculate the precision, recall rate, F-1 score and to plot a confusion matrix. A "combined" F-1 score is also calculated to measure the overall performance of the trained model. Fig. 3 shows an activity diagram summarizing the process of training, validation and testing.

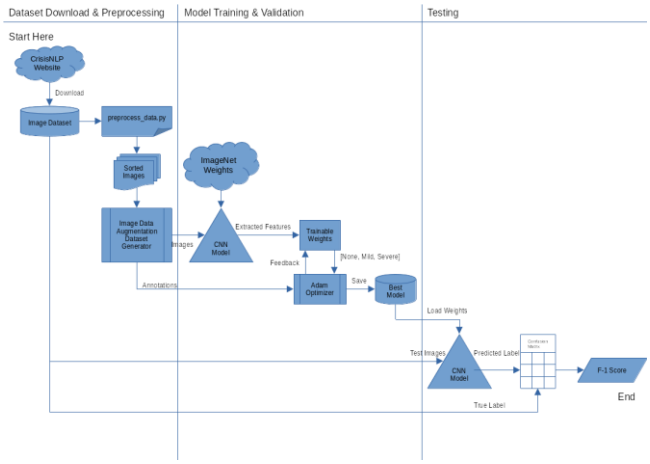

Fig. 3. Activity Diagram for training a CNN.

The second round of training, validation and testing involves ResNet50V2 and EfficientNetV2B0, the results of which are used to compare against each other (including VGG16) to determine which model achieves a higher F-1 score. This effort uses a similar process to the first round.

The third and final round of training involves the model selected from the second round of training against a dataset built by combining generated datasets that have obtained the highest combined F-1 score within each of the image data collections. This endeavour also uses a similar process to the first two rounds but has the addition of including a grid search of the dropout rate. The range of this dropout rate grid search starts from a dropout rate of 0.3 through 0.7 with a resolution of 0.1.

IV. Results, Limitations and Discussion

A. Results on VGG16 Single Event Tasks

Table III shows the F-1 scores for each class and the overall "combined" F-1 score obtained from testing VGG16 trained on single event generated datasets. The results in Table III demonstrate the effect of oversampling with image augmentations in training VGG16. With the exception of the Ruby image collection, models trained with generated datasets that had more oversampling tends to have a higher F-1 score.

Generated datasets that led to a higher combined F-1 score were selected to form a a combined dataset. An exception was made for the Nepal 6X generated dataset as combining it with the other selected datasets made it too large such that it caused an out-of-memory error. The final combined dataset included the balanced Nepal, Ecuador 8X, Matthew 8X, Ruby 2X datasets.

B. VGG16, ResNet50V2 and EffecientNetV2B0 against Nepal6X Dataset

Table IV shows the F-1 score obtained from testing VGG16, ResNet50V2 and EfficientNetV2B0 against the Nepal6X generated dataset. These three models were trained without a grid search of the dropout rate.

TABLE III. F-1 SCORE OF VGG16 AGAINST SINGLE-EVENT TASKS

Event	Generated Dataset	F-1 Score			
		None	*Mild*	*Severe*	*Combined*
Nepal	No Augmentations	0.76	0.02	0.80	0.53
	With Augmentations	0.75	0.05	0.79	0.53
	Balanced[a]	0.74	0.25	0.79	0.59
	2X Augmentations	0.72	0.18	0.76	0.55
	4X Augmentations	0.69	0.28	0.71	0.56
	6X Augmentations[b]	0.73	0.29	0.74	0.59
Ecuador	No Augmentations	0.80	0.00	0.85	0.55
	With Augmentations	0.78	0.00	0.84	0.54
	Balanced	0.72	0.19	0.81	0.57
	2X Augmentations	0.75	0.13	0.83	0.57
	4X Augmentations	0.73	0.18	0.79	0.57
	6X Augmentations	0.71	0.14	0.59	0.48
	8X Augmentations[a]	0.76	0.20	0.80	0.59
Matthew	No Augmentations	0.77	0.43	0.63	0.61
	With Augmentations	0.76	0.21	0.60	0.52
	Balanced	0.69	0.48	0.58	0.58
	2X Augmentations	0.73	0.45	0.59	0.59
	4X Augmentations	0.73	0.52	0.50	0.58
	6X Augmentations	0.70	0.46	0.61	0.59
	8X Augmentations[a]	0.67	0.54	0.62	0.61
Ruby	No Augmentations	0.74	0.73	0.11	0.53
	With Augmentations	0.77	0.72	0.36	0.62
	Balanced	0.72	0.64	0.47	0.61
	2X Augmentations[a]	0.75	0.66	0.46	0.62
	4X Augmentations	0.78	0.55	0.41	0.58
	6X Augmentations	0.74	0.67	0.38	0.60
	8X Augmentations	0.78	0.62	0.42	0.61

[a.] Generated datasets selected.

[b.] Too large to be combined with other datasets (causes out-of-memory error).

EfficientNetV2B obtained the highest F-1 score leading ResNet50V2 by 0.04 and VGG16 by 0.14. Both EfficientNetV2B and ResNet50V2 had significantly higher F-1 scores for all classes compared to VGG16. These results led to the selection of EfficientNetV2B0 for the next step.

TABLE IV. F-1 SCORE COMPARING VGG16 RESNET50V2 AND EFFICIENTNETV2B0

Model	F-1 Score			
	None	*Mild*	*Severe*	*Combined*
VGG16	0.73	0.29	0.74	0.59
ResNet50V2	0.82	0.40	0.84	0.69
EfficientNetV2B0	0.85	0.45	0.89	0.73

C. EfficientNetV2B0 against the Combined Dataset

The final model has managed to outperform state-of-the-art models in classifying the severity of damage in the Nepal, Ecuador and Matthew data collections which makes up the bulk of the damage severity dataset. Compared to [15], the F-1 score has increased from 0.76 to 0.782 for the Nepal data collection, 0.82 to 0.837 for the Ecuador data collection, and from 0.63 to 0.68 for the Matthew data collection. The combined overall F-1 score in the EfficientNetV2B0 model is close to the performance obtained with VGG16-fc7 by [34].

TABLE V. F-1 SCORE COMPARING VGG16 RESNET50V2 AND EFFICIENTNETV2B0

Data Collection	VGG16-fc7[cd]	EfficientNetV1B1[d]	EfficientNetV2B0
Nepal	0.76[c]	-	**0.782**
Ecuador	0.82[c]	-	**0.837**
Matthew	0.63[c]	-	**0.682**
Ruby	0.80[c]	-	0.709
Google Images	0.63[c]	-	0.576
Combined	0.753[d]	**0.758[d]**	0.752

[c.] from [15]

[d.] from [34]

D. Limitations

This study contains various technical limitations and time constraints which caused the scope of this study to be reduced.

This study was carried out using a single consumer grade Nvidia GPU with 8 Gigabytes of video RAM, namely an RTX 2060 Super. The computer used in this study had a solid state drive (SSD, not to be confused with Single Shot-MultiBox Detector) with a capacity of 256GB limiting the amount of generated image datasets that can be generated. This limits the size of both the model and the dataset that it can be trained on. These technical constraints restricted the scope of this study to models with 224 by 224 pixel inputs. There was not enough disk space to expand the study to include models with 240 by 240 input with this setup. These technical limitations also contributed to encountering an out-of-memory error when combining the six times oversampled Nepal dataset with other chosen generated datasets.

Another limitation is the amount of time needed to carry out model training. Training a deep learning model on one generated image dataset would take between three to six hours with the current setup assuming that the early stopping callback did not trigger. Each data collection would have seven generated datasets. Expanding the current scope to include the "CrisisMMD" dataset (containing 7 data collections) and "Damage Multimodal Dataset" dataset (treated as one data collection) would require an additional 56 single-event training and testing sessions, increasing the amount of time needed to include these. It is preferred that a data assessment task in response to the event of natural disasters takes place within 72 hours [25]. In the interest of time, the dataset used was limited to the "ASONAM2017" dataset.

This study did not explore fine-tuning as there is a lack of benchmarks for fine-tuned models trained on similar tasks to compare against.

E. Discussion

This paper shows a significant advancement in training deep learning models for classifying the level of natural disaster damage in images. The EfficientNetV2B0 model trained on the novel oversampling strategy was able to out-perform existing published benchmarks on classifying the level of disaster damage in the Nepal, Ecuador and Matthew data collections as seen in Table V.

These improvements were made possible using a combination of image data preprocessing techniques including a novel oversampling search strategy. This combination of techniques involve the use of image augmentation, data balancing and oversampling to address ongoing challenges in faced in data collection for disaster informatics leading to data imbalance and limited sample size for tasks involving image classification of disaster damage severity. The methods in order to obtain these results are:

1) For each image data collection, generate image datasets oversampled with augmented images ranging from balanced sampling to oversampling up to eight times the original sample size.

2) Use VGG16 to carry out a grid search of oversampling levels to identify which generated image dataset provides the highest F-1 score for each image data collection.

3) Combine the datasets identified in step 2 to create an optimized comprehensive dataset for training EfficientNetV2B0.

These steps have allowed for the creation of a dataset for training EfficientNetV2B0 such that it produces a more robust model with superior classification performance across various natural disaster scenarios.

These methods have demonstrated an importance in strategising the use of data preparation methods in machine learning when faced against situations caused by the nature of natural disaster events creating limitations that affect data collection. By using these image augmentation, balancing and oversampling methods, these issues that historically cause class imbalance and low sample size of images in this domain have been mitigated.

The findings in this study have shown several implications regarding disaster damage assessment through the classification of images. This study has demonstrated that the proposed training strategy improved the robustness and F-1 score of EfficienNetV2B0 in classifying the level of disaster damage. This in turn could increase the reliability of deep learning models used in the aftermath of a disaster event, potentially improving future efforts undertaken during disaster response and disaster resource allocation.

The proposed methodology in this paper could potentially be applied or be developed further in deep learning tasks facing similar issues on data imbalance and data scarcity. This combination of image data augmentation, data balancing and strategic data oversampling grid search could be implemented improve deep learning image classification tasks to counter the effects of imbalanced or scarce data.

V. Conclusion

This study has shown a novel strategy in countering the issue of data imbalance and data scarcity in classifying the level of disaster damage in images using deep learning models. By applying a mixture of techniques such as image data augmentation and image data oversampling, a trained EfficientNetV2B0 model that surpasses the performance of current models of similar input size in classifying the severity of natural disaster damage in some image collections has been obtained.

The methods in this study involved generating datasets of varying oversampling levels on different image data collections, ranging from balanced oversampling up to eight times the sample size. A search of an optimal amount of oversampling using VGG16 was carried out. The generated datasets with the optimal amount of oversampling were then combined to train the EfficentNetV2B0 model.

This success has netted a trained EfficentNetV2B0 model with improved F-1 scores of 0.782 on the Nepal data collection, 0.837 on the Ecuador data collection and a notable 0.683 on the Matthew data collection while maintaining a robust overall F-1 score of 0.752. These results show a major improvement on the classification of natural disaster damage levels in images, particularly on the Matthew data collection with some improvements on the Nepal and Ecuador data collections.

The findings in this study suggests that applying a combination of image data augmentation and oversampling techniques prior to model training helps in improving the robustness of deep learning classification models for classifying natural disaster damage. These methods have the potential to solve the challenges of data imbalance and data scarcity in image classification tasks involving natural disasters and offers a solution to improve the reliability and efficacy of natural disaster damage level classification in disaster response efforts.

References

[1] K. Weiss, T. M. Khoshgoftaar, and D. Wang, ‘A survey of transfer learning’, Journal of Big Data, vol. 3, no. 1, p. 9, May 2016, doi: 10.1186/s40537-016-0043-6.

[2] Z. Wang and X. Ye, ‘Social media analytics for natural disaster management’, International Journal of Geographical Information Science, vol. 32, no. 1, pp. 49–72, Jan. 2018, doi: 10.1080/13658816.2017.1367003.

[3] R.-Q. Wang, H. Mao, Y. Wang, C. Rae, and W. Shaw, ‘Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data’, Computers & Geosciences, vol. 111, pp. 139–147, Feb. 2018, doi: 10.1016/j.cageo.2017.11.008.

[4] P. Wallemacq and R. House, ‘Economic losses, poverty & disasters: 1998-2017 | UNDRR’. Accessed: Aug. 07, 2023. [Online]. Available: https://www.undrr.org/publication/economic-losses-poverty-disasters-1998-2017

[5] M. Tan and Q. V. Le, ‘EfficientNetV2: Smaller Models and Faster Training’, in International Conference on Machine Learning (ICML), 2021. [Online]. Available: https://arxiv.org/pdf/2104.00298.pdf

[6] M. Tan and Q. V. Le, ‘EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks’, 2019. [Online]. Available: https://arxiv.org/abs/1905.11946

[7] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, 'Revisiting Unreasonable Effectiveness of Data in Deep Learning Era', in 2017 IEEE International Conference on Computer Vision (ICCV), Oct. 2017, pp. 843–852. doi: 10.1109/ICCV.2017.97.

[8] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, 'Dropout: A Simple Way to Prevent Neural Networks from Overfitting', J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[9] K. Simonyan and A. Zisserman, 'Very Deep Convolutional Networks for Large-Scale Image Recognition', in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[10] B. K. Sillivan, 'Hurricane Ian 2022: Storm Is Set to Be One of Costliest in US History - Bloomberg'. Accessed: Aug. 09, 2023. [Online]. Available: https://www.bloomberg.com/news/articles/2022-09-27/hurricane-ian-is-set-to-be-one-of-costliest-storms-in-us-history#xj4y7vzkg

[11] C. Shorten and T. M. Khoshgoftaar, 'A survey on Image Data Augmentation for Deep Learning', Journal of Big Data, vol. 6, no. 1, p. 60, Jul. 2019, doi: 10.1186/s40537-019-0197-0.

[12] J. Schmidhuber, 'Deep learning in neural networks: An overview', Neural Networks, vol. 61, pp. 85–117, Jan. 2015, doi: 10.1016/j.neunet.2014.09.003.

[13] R. Peters and J. Porto De Albuquerque, 'Investigating images as indicators for relevant social media messeges in disaster management', in Proceedings of the ISCRAM 2015 Conference, Kristiansand, Norway, 2015.

[14] D. T. Nguyen, F. Alam, F. Ofli, and M. Imran, 'Automatic Image Filtering on Social Networks Using Deep Learning and Perceptual Hashing During Crises', in Proceedings of the 14th International Conference on Information Systems for Crisis Response And Management, T. Cornes, F.B., C. Hanachi, M. Lauras, and A. Montarnal, Eds., Albi, France: Iscram, 2017, pp. 499–511. [Online]. Available: http://idl.iscram.org/files/dattnguyen/2017/2038_DatT.Nguyen_etal2017.pdf

[15] D. T. Nguyen, F. Alam, M. Imran, and P. Mitra, 'Damage Assessment from Social Media Imagery Data During Disasters', in 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Aug. 2017, pp. 569–576.

[16] H. Mouzannar, Y. Rizk, and M. Awad, 'Damage Identification in Social Media Posts using Multimodal Deep Learning', in ISCRAM 2018 – 15th International Conference on Information Systems for Crisis Response and Management, 2018, pp. 529–543.

[17] Y. Li, W. Hu, H. Dong, and X. Zhang, 'Building Damage Detection from Post-Event Aerial Imagery Using Single Shot Multibox Detector', Applied Sciences, vol. 9, p. 1128, Mar. 2019, doi: 10.3390/app9061128.

[18] X. Li, H. Zhang, D. Caragea, and M. Imran, 'Localizing and quantifying damage in social media images', in Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Barcelona, Spain: IEEE Press, 2020, pp. 194–201.

[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, 'Gradient-based learning applied to document recognition', Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.

[20] Y. LeCun et al., 'Backpropagation Applied to Handwritten Zip Code Recognition', Neural Computation, vol. 1, no. 4, pp. 541–551, Dec. 1989, doi: 10.1162/neco.1989.1.4.541.

[21] Y. LeCun, Y. Bengio, and G. Hinton, 'Deep learning', Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, 'ImageNet Classification with Deep Convolutional Neural Networks', Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[23] J. M. Johnson and T. M. Khoshgoftaar, 'Survey on deep learning with class imbalance', Journal of Big Data, vol. 6, no. 1, p. 27, Mar. 2019, doi: 10.1186/s40537-019-0192-5.

[24] M. A. Islam, S. I. Rashid, N. U. I. Hossain, R. Fleming, and A. Sokolov, 'An integrated convolutional neural network and sorting algorithm for image classification for efficient flood disaster management', Decision Analytics Journal, vol. 7, p. 100225, Jun. 2023, doi: 10.1016/j.dajour.2023.100225.

[25] M. Imran, U. Qazi, F. Ofli, S. Peterson, and F. Alam, 'AI for Disaster Rapid Damage Assessment from Microblogs', AAAI, vol. 36, no. 11, pp. 12517–12523, Jun. 2022, doi: 10.1609/aaai.v36i11.21521.

[26] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, 'AIDR: Artificial Intelligence for Disaster Response', in Proceedings of the 23rd International Conference on World Wide Web, in WWW '14 Companion. New York, NY, USA: Association for Computing Machinery, 2014, pp. 159–162. doi: 10.1145/2567948.2577034.

[27] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, 'Processing Social Media Messages in Mass Emergency: A Survey', ACM Comput. Surv., vol. 47, no. 4, Jun. 2015, doi: 10.1145/2771588.

[28] K. He, X. Zhang, S. Ren, and J. Sun, 'Identity Mappings in Deep Residual Networks', in Computer Vision – ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 630–645.

[29] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[30] F. Alam, T. Alam, F. Ofli, and M. Imran, 'Robust Training of Social Media Image Classification Models', IEEE Transactions on Computational Social Systems, vol. 11, no. 1, pp. 546–565, Feb. 2024, doi: 10.1109/TCSS.2022.3230839.

[31] M. Endsley, 'Endsley, M.R.: Toward a Theory of Situation Awareness in Dynamic Systems. Human Factors Journal 37(1), 32-64', Human Factors: The Journal of the Human Factors and Ergonomics Society, vol. 37, pp. 32–64, Mar. 1995, doi: 10.1518/001872095779049543.

[32] L. Bevere, M. Schwartz, R. Sharan, and P. Zimmerli, 'sigma 1/2018: Natural catastrophes and man-made disasters in 2017: year of record-breaking losses | Swiss Re', 2018. Accessed: Aug. 07, 2023. [Online]. Available: https://www.swissre.com/institute/research/sigma-research/sigma-2018-01.html

[33] F. Alam, H. Sajjad, M. Imran, and F. Ofli, 'CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing', ICWSM, vol. 15, no. 1, pp. 923–932, May 2021, doi: 10.1609/icwsm.v15i1.18115.

[34] F. Alam, F. Ofli, M. Imran, T. Alam, and U. Qazi, 'Deep Learning Benchmarks and Datasets for Social Media Image Classification for Disaster Response', in 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2020, pp. 151–158. doi: 10.1109/ASONAM49781.2020.9381294.

[35] F. Alam, M. Imran, and F. Ofli, 'Image4Act: Online Social Media Image Processing for Disaster Response', in Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, in ASONAM '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 601–604. doi: 10.1145/3110025.3110164.