Article

Efficient Classification of Imbalanced Natural Disasters Data Using Generative Adversarial Networks for Data Augmentation

Rokaya Eltehewy [1,*], Ahmed Abouelfarag [2] and Sherine Nagy Saleh [3]

[1] Computer Engineering Department, College of Engineering & Technology, Arab Academy for Science & Technology (AAST), Cairo 2033, Egypt
[2] College of Artificial Intelligence, Arab Academy for Science & Technology (AAST), El Alamein 51718, Egypt; abouelfarag@aast.edu
[3] Computer Engineering Department, College of Engineering & Technology, Arab Academy for Science & Technology (AAST), Alexandria 1029, Egypt; sherine_nagi@aast.edu
* Correspondence: rokaya@aast.edu

Abstract: Rapid damage identification and classification in disastrous situations and natural disasters are crucial for efficiently directing aid and resources. With the development of deep learning techniques and the availability of imagery content on social media platforms, extensive research has focused on damage assessment. Through the use of geospatial data related to such incidents, the visual characteristics of these images can quickly determine the safety situation in the region. However, training accurate disaster classification models has proven to be challenging due to the lack of labeled imagery data in this domain. This paper proposes a disaster classification framework, which combines a set of synthesized diverse disaster images generated using generative adversarial networks (GANs) and domain-specific fine-tuning of a deep convolutional neural network (CNN)-based model. The proposed model utilizes bootstrap aggregating (bagging) to further stabilize the target predictions. Since past work in this domain mainly suffers from limited data resources, a sample dataset that highlights the issue of imbalanced classification of multiple natural disasters was constructed and augmented. Qualitative and quantitative experiments show the validity of the data augmentation method employed in producing a balanced dataset. Further experiments with various evaluation metrics verified the proposed framework's accuracy and generalization ability across different classes for the task of disaster classification in comparison to other state-of-the-art techniques. Furthermore, the framework outperforms the other models by an average validation accuracy of 11%. These results provide a deep learning solution for real-time disaster monitoring systems to mitigate the loss of lives and properties.

Keywords: data augmentation; deep neural network architectures; disaster classification; ensemble classifiers; generative adversarial networks

check for updates

Citation: Eltehewy, R.; Abouelfarag, A.; Saleh, S.N. Efficient Classification of Imbalanced Natural Disasters Data Using Generative Adversarial Networks for Data Augmentation. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 245. https://doi.org/10.3390/ijgi12060245

Academic Editors: Wolfgang Kainz, Christos Chalkias, Marinos Kavouras, Margarita Kokla and Mara Nikolaidou

Received: 20 April 2023
Revised: 9 June 2023
Accepted: 15 June 2023
Published: 17 June 2023

1. Introduction

Disasters and natural hazards require immediate intervention due to their severe negative impacts on human lives and properties, as well as permanent damage to the environment. The robust response and recovery efforts of responders and humanitarian organizations are vital to ensuring a minimum degree of losses and changes to the ecosystem [1]. Social media platforms have gained importance in the task of disaster classification as the continuous monitoring of content on different online streams can lead to the rapid identification of dangerous situations. The use of such resources helps in providing immediate disaster relief and rescue in critical cases [2].

Computer vision and deep learning real-time systems have rapidly become essential for active disaster response through tasks including incident identification, irrelevant image filtration, image classification into specific humanitarian categories, and damage severity assessment [3]. However, due to the complex and varying structures of disaster images, training an accurate and robust disaster classification network requires a large dataset, in addition to balanced sample sizes for each incident type.

1.1. Problem Background

Social media platforms, such as Twitter and Facebook, are considered vital sources of textual and imagery content. Despite extensive research that mainly focuses on textual content to extract useful information, other works have utilized multimodal architectures and enhanced the performance over existing baselines [4]. However, few studies have focused on the use of imagery content, although past research has proven that imagery content from social media during a disaster can be independently used to develop effective disaster classification frameworks [5].

The use of disaster-related social media imagery can help identify the spatial relationship between disastrous events and accidents from different angles. If such information is combined with the geographical characteristics of the incident, first aid personnel can be immediately directed to the incident location. This study proposes an approach to gather spatial data from social media platforms and apply image classification to detect disastrous events in real time. If the results are combined with geographical information, this approach can aid first-aid responders in cases of disasters.

In this study, we tackle the task of disaster classification using deep learning techniques with social media-based images. Class imbalance is a common issue in real-world classification problems, where a significant imbalance in the number of training samples between classes causes the learning algorithms to overgeneralize the samples of the majority classes and produce inaccurate network parameters. Certain categories of most disaster classification datasets show skewness in the overall data distribution, which adds to the difficulty of model training. In particular, since deep neural networks (DNNs) need a substantial amount of training data for multi-class problems, most learners tend to exhibit bias towards the majority classes or completely disregard the minority classes. In many problems, the minority group is the class of interest; however, in this current task, all classes of disasters are of equal importance.

Countermeasures against imbalance traditionally include data-level sampling techniques, such as the synthetic minority oversampling technique (SMOTE), which increases the number of class instances by interpolating neighboring points. Alternatively, algorithm-level techniques perform instance weighting or alter the loss function to penalize minority class errors [6]. However, data-dependent techniques may lead to the overlapping of classes and pose an impracticality for high-dimensional data. Similarly, algorithmic methods face the challenge of selecting an effective cost or penalty that continuously changes based on the domain of the task at hand.

To address the problems associated with traditional countermeasures, synthetic data can be created to mirror the statistical properties of the original dataset. The algorithm used for data generation is a crucial factor that affects the quality of the synthesized data. Data augmentation has become a crucial part of the deep learning process to generate sufficient samples of training data. Prior approaches mainly depend on data warping augmentation methods, including traditional geometric augmentation and color transformations [7], as well as more novel approaches similar to CutMix augmentation, which replaces regions from training images with patches from others originating from different classes to create additional data [8]. Although effective at increasing the sample size, these methods only apply simplistic manipulations that fail to produce any new meaningful features.

We conducted experiments to explore the validity of using synthetically generated samples to amplify a training dataset for the task of disaster classification. Generative adversarial networks (GANs) [9] have been shown to be effective for data augmentation in computer vision tasks. GANs are trained to fit the real distribution of data using a min–max game theory; once the correct distribution is reached, realistic synthetic samples are produced. A basic approach consists of the concurrent training of two models: a generative model to capture the data distribution from input noise and a discriminative model responsible for predicting the probability of a sample belonging to the training data or the generated output.

- This work proposes a disaster classification framework that maximizes classification performance by employing a generative network for synthetic disaster data generation in combination with an ensemble learner. This method eliminates the need to collect additional real data samples and aids in making informed decisions, which subsequently improves the results. The key contributions of this paper are summarized as follows: This work produces a comprehensive and balanced disaster classification dataset by training a conditional generative adversarial model. The model is designed to synthesize realistic disaster images determined using both qualitative and quantitative evaluations. This framework aims to provide extensive baselines for the crisis analysis research community.
- This work presents a deep learning framework for disaster classification using an ensemble learning approach. Using a new approach for generative augmentation, the framework aids in rapid damage assessment using real-time social media data.
- The proposed model has proven to improve the accuracy of disaster classification by up to 11% in comparison to other state-of-the-art research conducted for the same task. These superior results were obtained by training state-of-the-art convolutional architectures using a domain-specific dataset and experimenting with different augmentation methods, such as geometric augmentation, CutMix augmentation, GAN augmentation, and a combination thereof.

1.2. Related Work

Previous research has extensively studied disaster classification. However, the classification of imbalanced datasets that include less frequently occurring disasters, such as droughts and hurricanes, has rarely been studied directly.

1.2.1. Disaster Classification

Due to the lack of labeled images in the domain of disaster classification, recent studies have focused on the collection of disaster data through social media platforms to develop datasets for the research community. These datasets may include disaster images, textual information, or a combination of both.

The earliest publications include the CrisisLex dataset presented by Olteanu et al. [10], which consists of tweets related to six different disastrous events. As classification results show improvement when the datasets include images of the events, tweets combined with 5000 images are used to classify fires to aid emergency personnel in combating bushfires, with a classification accuracy of 86% [11]. Yang et al. [12] described a dataset of geotagged Twitter posts of Hurricane Sandy in 2012, yet the dataset lacks human annotations. Alam et al. [13] collected CrisisMMD, a large multimodal dataset from Twitter during different natural disasters with three different types of annotations. Other datasets combined and fused satellite imagery with social media content to extract meaningful insights for each disaster type [14,15]. Furthermore, the authors in [2] re-labeled the existing CrisisMMD dataset to detect the type of disaster, informativeness, and damage severity, which was tested for classification against state-of-the-art deep learning models to provide a baseline for future research in similar tasks.

Deep learning pre-trained models have shown positive results when combined with transfer learning. Valdez and Godmalin [16] created a lightweight convolutional network with two heads to classify disaster images and measure disaster intensity. Hong et al. [17] conducted a study to identify buildings damaged due to earthquakes using post-disaster aerial images and a proposed CNN that combines global and contextual features to determine the damage levels. Liang et al. [18] fine-tuned pre-trained convolutional and language models using multimodal inputs of text-image pairs and obtained effective results against multimodal classification benchmarks. Khattar and Quadri [19] utilized unsupervised domain adaptation to classify unlabeled data of a new disaster using the existing labeled images of relevant disasters.

In this study, we built a balanced disaster dataset, which is a combination of the previously discussed benchmark datasets, and further examples were added using a generative adversarial model for data augmentation to increase the number of samples for under-represented classes.

1.2.2. Generative Augmentation

Since the introduction of the generative adversarial network architecture by Ian Goodfellow and his colleagues in 2014 to produce synthetic visual samples, several variants of the original architecture have been proposed by the research community. A conditional model (cGAN) was proposed to handle multiple classes within the same network. Shahbazi et al. [20] studied the utilization of pre-trained cGANs to transfer knowledge across classes by injecting a network with a class label for each sample and applying conditional batch normalization. Additionally, the authors in [21] experimented with non-adversarial loss functions to enhance the quality of generated images. While CycleGAN is used for unpaired image-to-image translation, StyleGAN was designed to apply style transfer to the generated images by separating the high-level attributes of an image, such as pose and identity, and performing scale-specific interpolation operations on the output. Such networks have had a significant impact on areas of computer vision, including image-to-image translation, image generation, and similar domains [22].

GANs have recently been employed in the field of data augmentation since they can learn the original data distribution and produce similar realistic samples. A conditional GAN was used for multimodal audio-visual emotion recognition to address the class imbalance issue using generators and discriminators for both modalities [23]. A similar augmentation technique can be applied to medical tasks. Motamed et al. [24] augmented chest X-rays for the detection of pneumonia and COVID. They demonstrated that the GAN-based approach surpassed the results obtained by traditional augmentation methods in anomaly detection. Rui et al. [25] designed a Disaster GAN to generate data samples to identify building damage from remote sensing imagery. Additionally, a GAN-based model was used to replace the labor-intensive data collection process for human pose estimation [26] by entering the original limited dataset into the GAN model along with Gaussian white noise, which produced simulated human pose samples. The remainder of this paper is structured as follows: Section 2 presents our proposed methodology to enhance the classification results obtained on disaster classification datasets in detail. Section 3 elaborates on the conducted experiments to produce an effective deep-learning framework for disaster classification. Sections 4 and 5 discuss the results to verify the validity of the proposed framework. Finally, Section 6 concludes the paper by presenting a summary of the contributions and addressing future work.

2. Methodology

This paper proposes a technique to classify imbalanced disaster data into three main phases, as illustrated in Figure 1. First, we based our data augmentation on the state-of-the-art conditional GAN architecture to generate fake samples that are as realistic as possible. Once the samples are generated and evaluated using Fréchet Inception Distance and Inception Score [27], they are used to supplement the training dataset with the augmented information to avoid overfitting. This method eliminates the need to collect additional real data samples and aids in making informed decisions.

Second, we fine-tuned an ensemble of pre-trained VGG16 classifiers to perform disaster classification using the original dataset and the supplemented dataset. We also tested the models' performance on different types of augmentation. It is to be noted that all testing experimentations were conducted on the same subset of real data samples in order to avoid overfitting and ensure a fair comparison between different models, while the synthetic samples were only used during the training phase.

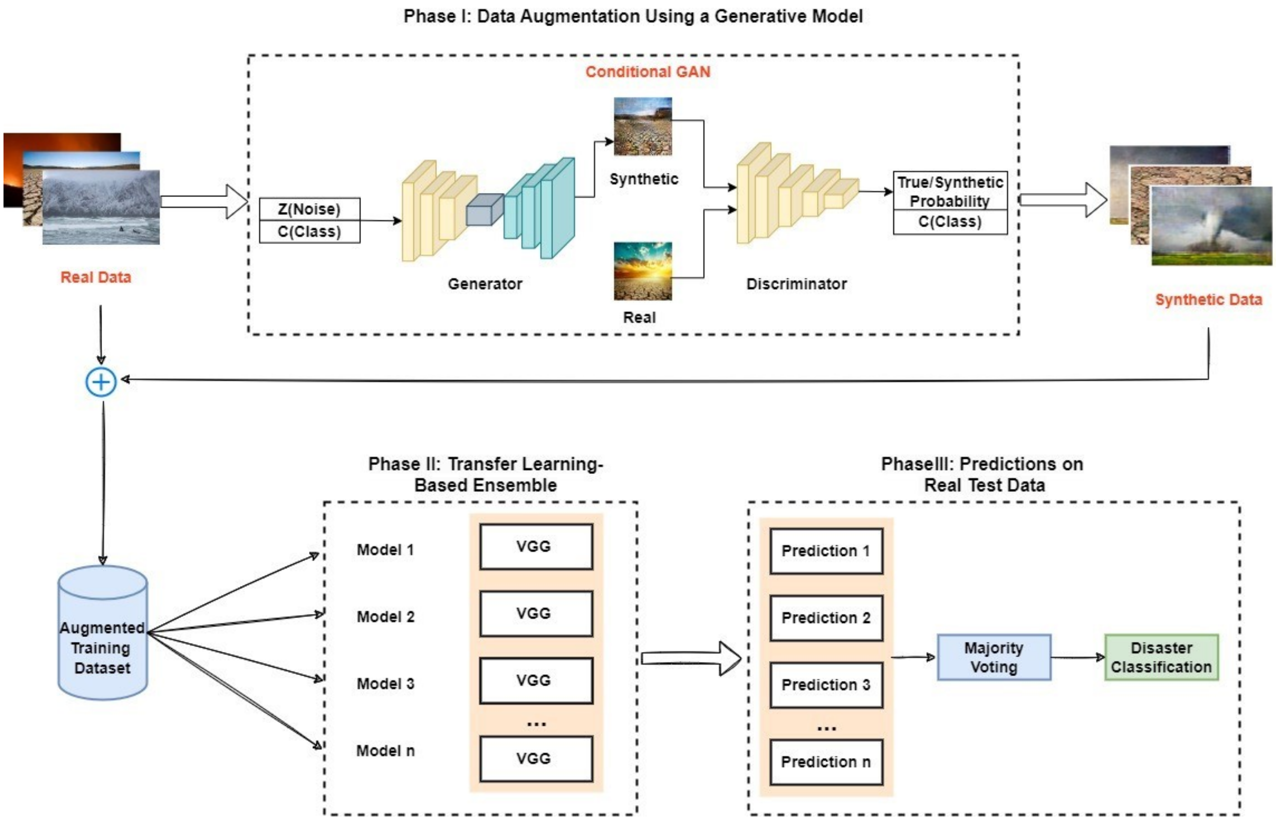


Figure 1. The proposed disaster classification framework.

Hyperparameters are manually set variables to help guide the learning process for each of the implemented architectures. All hyperparameters were obtained after applying a grid search, which defines a search space as a grid of hyperparameter values and evaluates every position in the grid to determine the optimal values.

As ensemble classifiers have the ability to produce performances superior and generally more stable than single models using imagery or textual datasets [28]. We applied an ensemble learning approach by combining predictions from multiple models trained on different subsets of the data by using bootstrap aggregation (bagging), which applies stratified data resampling with replacement.

Applying stratified random sampling with replacement creates a balanced and reliable representation of each class. After the training phase, in which transfer learning is implemented to gain the advantage of the previously learned weights, each model is ready to output predictions. In the final phase, a majority hard vote is used, where every individual classifier votes for a class with the highest output probability, and the class with the majority votes decides the final prediction.

The following subsections describe the generative model used for augmentation and the deep convolutional classifiers to be trained.

2.1. Data Augmentation

Data generation techniques using over- or under-sampling have proven to be effective for mixed or tabular data. However, GANs have additional merits regarding image generation. Generating high-quality synthetic images eliminates the dependency on a large training dataset. We have described the proposed framework of a disaster conditional GAN with the architecture shown in Figure 2. The model is conditioned on the class labels of each image to allow the targeted output of the images to be of a given class.

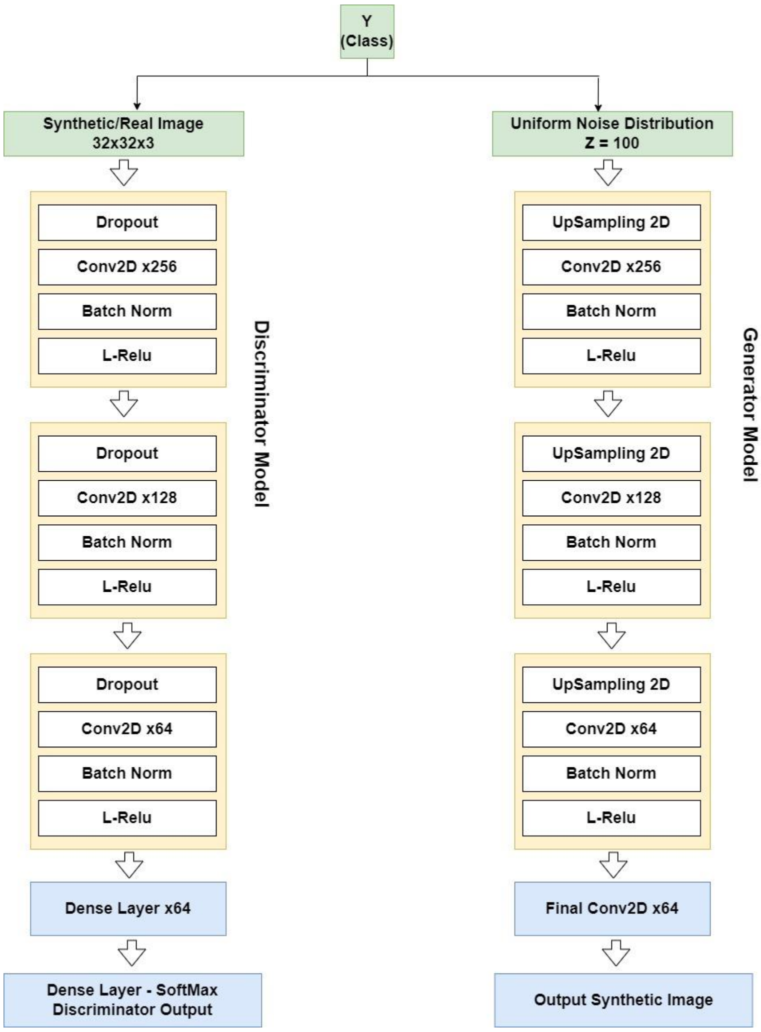


Figure 2. The architecture of disaster cGAN, including generator and discriminator models.

The generator model begins with a uniform noise distribution input that is passed to the first of the three convolution stacks, progressively converting it to a high-dimensional feature vector representing the new image. Batch normalization is used to unify the inputs to prevent all samples from collapsing into a single point. Finally, the model outputs $64 \times 64 \times 3$ images to be passed to the discriminator.

The discriminator is constructed with similar convolutional stacks, followed by final dense layers for classification. Average pooling down samples of the feature vectors are passed to the final fully connected dense layers to progressively map them to a lower-dimensional space for classification by a SoftMax layer. Finally, the model outputs a probability distribution over the class labels, and the most likely prediction is obtained.

2.2. Convolutional Classifiers

Deep learning is a leading approach for information extraction, and convolutional neural networks (CNNs) are commonly used architectures to process multidimensional vectors and achieve highly accurate results [29]. There are various CNN architectures, where each of the networks differs in terms of the internal layers and techniques used. We opted to use the Inception-V4 and VGG16 architectures since such networks can solve image-based problems while simultaneously reducing the required parameters in comparison to traditional networks. We have briefly described each architecture in the following subsections.

2.2.1. VGG16 Architecture

We employed a transfer learning-based approach, which uses the existing weights of a trained model to improve the performance of the targeted domain [30]. We fine-tuned a VGG16 pre-trained architecture [31] to adapt it to the task of natural disaster classification using our domain-specific collected dataset. The initial weights of VGG16 pre-trained on the ImageNet dataset, a dataset consisting of 1.2 million RGB images, were used to initialize the model. The architecture used was a 16-layer network composed of convolutional and fully connected layers, as shown in Figure 3, using only 3 × 3 convolutional layers stacked on top of each other for simplicity. The convolutional blocks were followed by a max pooling layer with stride 2, and the final layers consisted of three hidden layers of 4096 nodes followed by a SoftMax output layer of seven units to represent the class labels.

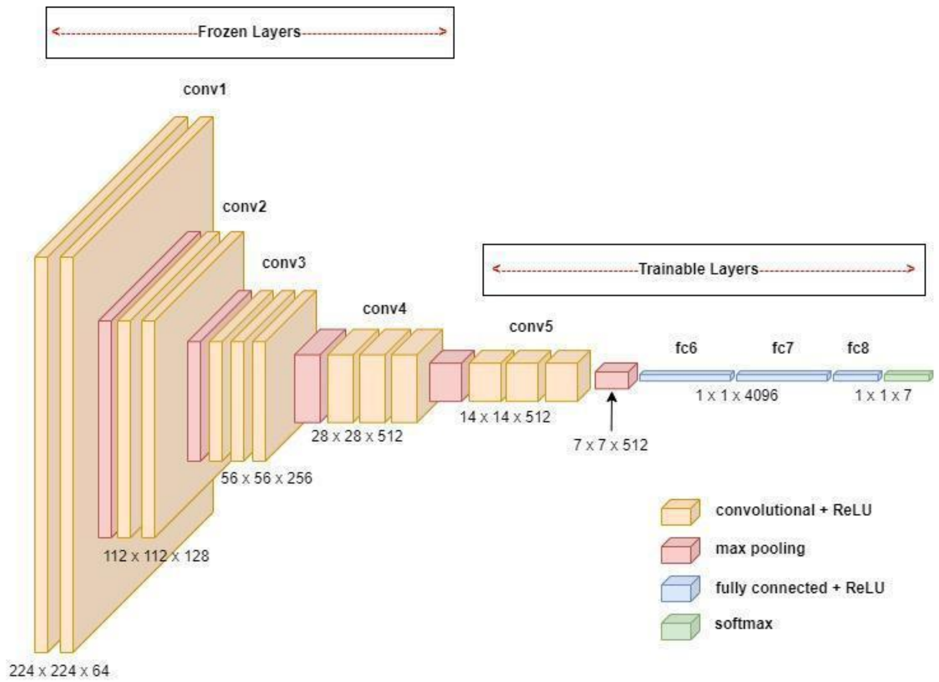


Figure 3. VGG16 Architecture.

The model was trained after adding additional classification layers to transfer the parameters from the broad domain to the problem-specific domain. Adam optimizer and SoftMax activation function are used. The learning rate was set to 10^{-4}, the momentum to 0.6, and the maximum number of epochs was set to 100 with an early stopping criterion. The batch size for each epoch was set to 128, as it provided an optimal balance between the training time and model accuracy.

2.2.2. Inception Architecture

The inception architecture, first introduced in 2015 by Szegedy et al. [32], was employed to be trained and tested for disaster classification. Although the model is relatively small in size in comparison to other state-of-the-art architectures, it provides satisfactory accuracies on the ImageNet dataset. Inception V4 comes as an extension of Inception V3 and Google LeNet modules [33]. The main aim of the Inception V4 model is to decrease the number of parameters to be trained, thereby decreasing the computational complexity. The architecture is based on the notion of building a wider rather than deeper network by using convolutional filters of different sizes operating on the same level. As the inception architecture is highly tunable, the model consists of three different types of inception blocks, each with different numbers of filters in the various layers, as shown in Figure 4.

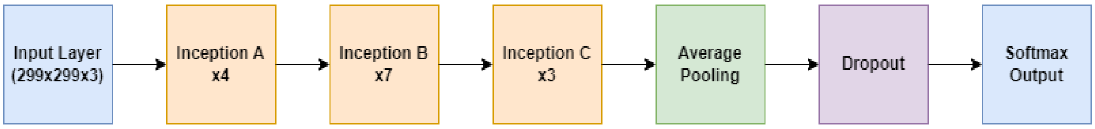


Figure 4. Simplified Inception V4 Architecture (Each colored block represents a different type of layer).

The Inception V4 model is a network of stacked inception modules, followed by a SoftMax layer. Each module has multiple scales containing different kernel sizes. Each filter size extracts different scales of feature maps and propagates the output to the next block. A 1 × 1 convolutional filter is used for dimensionality reduction, while 3 × 3 and 5 × 5 convolutions are applied to achieve optimal feature extraction. Each inception module is followed by a reduction layer to scale the dimensionality of the filter bank to match the depth of the input to the next layer. In terms of the hyperparameters, the network is trained with stochastic gradient descent, momentum, and a decay rate of 0.9. The initial learning rate is set to 0.045 for decaying every 2 epochs.

3. Experimental Settings

In this section, we discuss the original dataset followed by the augmented dataset. We performed a quantitative evaluation of the generated images to verify the quality of the synthesized features. The natural disaster dataset included seven different classes of disasters, each of which had distinct traits. However, two classes showed visible imbalance (hurricane images and drought images). Several state-of-the-art architectures, such as VGG19 and Inception V4, were tested [34]. To produce an effective classification model, the dataset was first balanced by training a generative adversarial network to produce additional samples for the minority classes.

We additionally compared the performance of GAN augmentation to both traditional and CutMix augmentations to verify its superiority, as illustrated by the different augmentation methods applied in Figure 5.

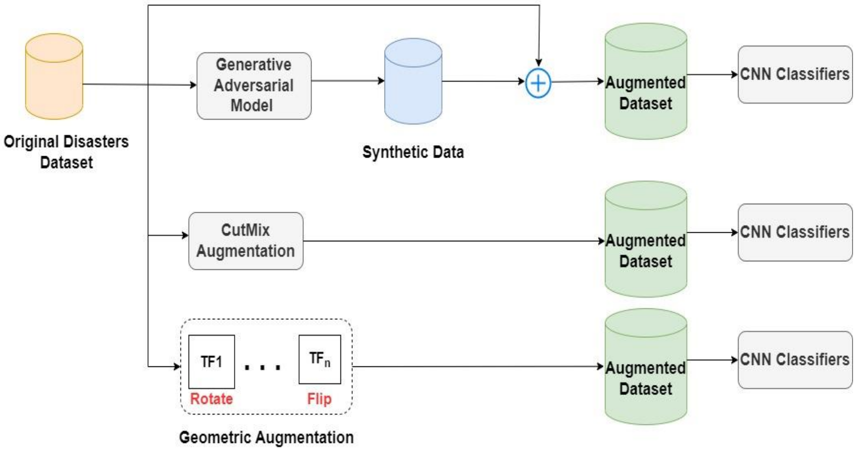


Figure 5. The different augmentation methods used for experimentation (Geometric augmentations highlighted in red include rotation and flipping).

3.1. Dataset Analysis

A subset of CrisisMMD [13], a real-world disaster-related dataset collected from Twitter during different natural disasters that took place in 2017, was used as the backbone of our dataset. Although the CrisisMMD dataset was collected several years ago, the samples remain relevant to the current times due to the consistent nature of natural disasters. In fact, the extreme weather events described by such datasets currently have an increasing toll on human and economic status due to the climate changes brought on by more frequent disasters. Additional samples were selected from other recent and publicly available datasets, including the damage identification multimodal dataset [35] and the damage severity assessment dataset [36].

The stated datasets consist of a combination of manually annotated tweets and images collected during seven major natural disasters, including earthquakes, hurricanes, wildfires, floods, etc., across different regions of the world. Images concerning certain natural disasters were selected and regrouped into new classes. Figure 6 shows samples from different classes within the dataset.

Figure 6. Sample images from the original imbalanced disaster dataset.

Since social media-based data can include irrelevant information, all samples were manually revised and any mistakenly labeled, noisy, or corrupted images were removed from the dataset before the data processing stage.

Additionally, text annotations were removed to focus only on the effects of synthetic images on image classification. As shown in Figure 7, the final dataset exhibits class imbalance in two different classes, which we later eliminated by augmenting the datasets with class-specific synthetic samples.

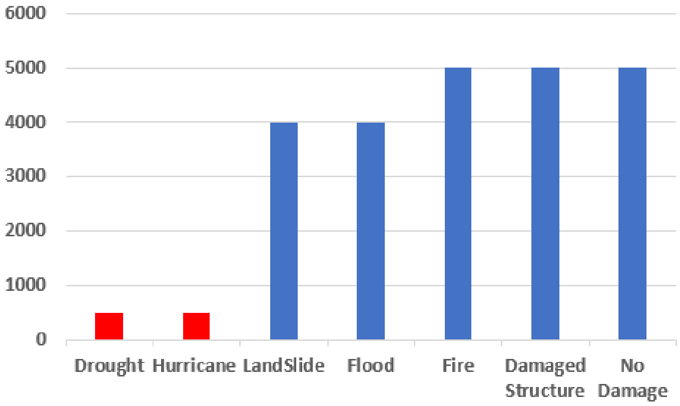


Figure 7. Number of samples per class showing the imbalanced nature of the dataset (Classes highlighted in red show severe imbalance).

The imbalance for multi-class classification is represented by the parameter (ρ), which represents the ratio between the number of examples in majority classes and the number of examples in minority classes. The ratio (ρ) can be expressed as follows, where C_i is a set of examples for each class i [37].

$$\rho = \frac{max_i\{|C_i|\}}{min_i\{|C_i|\}}, \tag{1}$$

For the given dataset, an imbalance can be seen for the drought and hurricane classes with an imbalance ratio of 10. Imbalanced classes were selected as inputs to the conditional GAN architecture to produce extra synthetic samples to balance the dataset while maintaining descriptive feature vectors for each class.

3.2. Implementation

The proposed framework was trained using different augmentation methods to determine the efficiency of GAN-based augmentation. After successfully balancing the dataset and verifying the quality of the generated images using inception score calculation, each generated dataset was used to train both the VGG16 and Inception V4 deep learning classifiers. Additionally, ensemble learning by bagging was applied to the best-performing models to further improve the classification results.

Since the introduction of the generative adversarial network architecture in 2014, several variants, including the cGAN model, have been introduced. A cGAN aims to condition the network with additional information, which allows for the generation of class-specific disaster samples after the network is injected with the class labels for each image.

In terms of data generation, the cGAN model was trained with 3000 epochs to generate 5000 new instances for the under-represented classes. Regarding hyper parameters, the Adam optimizer, along with binary cross entropy as a loss function, was used. Leaky Relu was used for activations except for the final layer, which used a Tanh activation. Generator dropout with a probability of 0.2, a learning rate of 0.0002, and a batch size of 4 were applied.

After adding the additional samples to the dataset, the class imbalance ratio was drastically reduced to a ratio of 2.5. It can be seen that the generated data are fairly similar to the true samples, as shown in Figure 8. After utilizing the synthetically generated data for classification, the results confirm that the generated samples produce similar feature vectors when compared to the true samples and can be used for accurate classification.

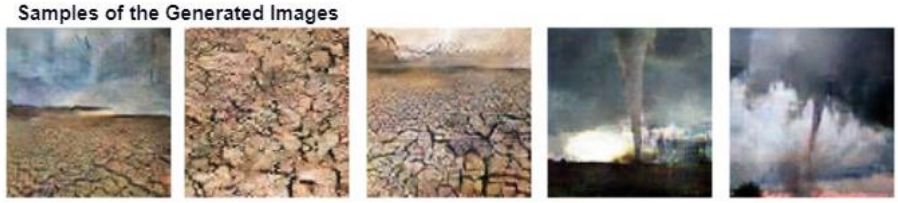


Figure 8. Images generated by cGAN for each imbalanced class (drought and hurricane images).

Other data augmentation methods of geometric and CutMix augmentation were also applied to the original dataset, and a sample of each method's output is shown in Figure 9.

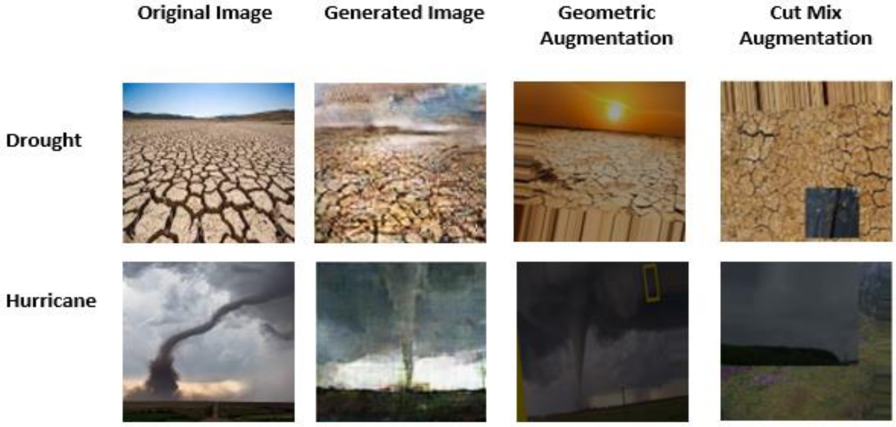


Figure 9. Samples of different augmentation methods (GAN, geometric, and CutMix augmentation).

To further evaluate the images generated by the conditional generative network, we applied the inception score method to measure the variation in images and the Fréchet inception distance (FID) evaluation method. The calculation of the FID is based on the features from the last average pooling layer of the inception model.

The equation to calculate the inception score for a set of generated images is as shown below, where $x \sim p_g$ indicates that x is an image sampled from p_g and $D_{KL}\ (p||q)$ is the KL divergence between the distributions p and q [38].

$$\mathrm{IS(G)} = \exp(E_{x \sim p_g} D_{KL}\ (p(y|x)\ \Big|\Big|\ p(y))), \tag{2}$$

The Fréchet inception distance is formulated as follows [39], where the distance between two Gaussians, real-world data $pw(.)$, and the generated data $p(.)$ were measured using the mean and covariance matrix (m, C) for each distribution.

$$D^2((m,C),\ (m_w,C_w)) = \left|\left|m - m_w\right|\right|\frac{2}{2} + Tr(C + C_w - 2(CC_w)^{\frac{1}{2}}), \tag{3}$$

We carried out three tests on batches of the generated images and averaged the results to obtain the inception score and FID values for the disaster dataset, as shown in Table 1.

Table 1. Synthetic images evaluation metrics.

	Inception Score	**Fréchet Inception Distance**
Evaluation score	41.3	6.72

It is to be noted that the inception and FID scores were only used as a reference to compare subsequent models working on the same task since the disaster classification images were relatively different from the ImageNet samples on which the baseline scores were determined.

Classifiers were trained and tested for every augmentation method in addition to the ensemble classifiers. A summary of the different experimentations is shown in Figure 10.

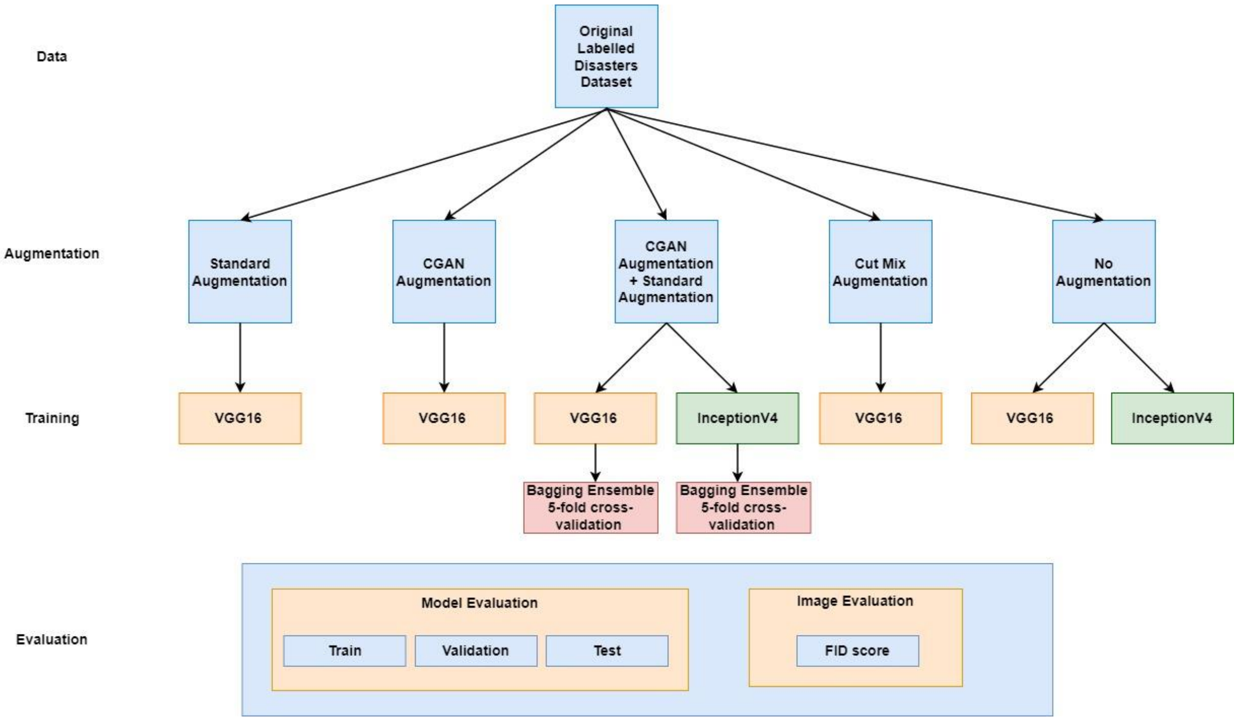


Figure 10. Experimentations on different classifiers combining data augmentation methods (Each colored block represents a different classifier or augmentation method).

4. Results

All the initial experiments used an imbalanced dataset with 24,000 images distributed over seven classes. The final experiments included the GAN-generated images, resulting in a significantly balanced dataset of 29,000 images. All proposed models were implemented using Python, Tensor Flow 2.3.1, and Keras. All training trials were completed on a Jupiter notebook using a local CPU, and the average epoch execution time of 5000 s could be significantly reduced if a GPU-based approach was utilized to speed up computations.

Geometric augmentation that was applied involved image flipping, a 30% zooming rate, and 30% horizontal and vertical shifting. Additionally, the dataset was divided according to common best practices into 70% training, 20% validation, and 10% testing samples with the batch size set to eight samples. Since the validation data provided information that directed the tuning of the model's hyperparameters and configurations, it was necessary to allocate sufficient validation samples. The test set required the least number of samples, as it measured the final model performance in terms of accuracy. The dataset initially suffered from an imbalance; thus, the majority of samples were directed towards the training set.

The two architectures employed in our experiments were VGG16 and Inception V4, which were evaluated individually using bagging ensembles to improve the resulting predictions by using several differently biased models and obtaining a final majority vote, as described in the previous sections.

In this paper, two hypotheses are evaluated: (1) whether the use of generative models for data augmentation is superior to that of traditional methods; (2) whether the use of an ensemble classifier improves the accuracy of the model for disaster classification.

4.1. Performance Evaluation

To measure the performance of each disaster classification architecture, we employed various performance metrics, such as accuracy, precision, recall, and F1 score, which were calculated as shown:

$$Accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN}\right)$$

$$Precision = \left(\frac{TP}{TP + FP}\right)$$

$$Recall = \left(\frac{TP}{TP + FN}\right)$$

$$F1 - Score = \left(\frac{2 * Precision * Recall}{Precision + Recall}\right), \tag{4}$$

where *TP* represents true positive, *TN* true negative, *FP* false positive, and *FN* false negative.

4.2. Compared Methods

This section discusses the results of each experiment. After testing several classifiers, ensemble learning, and different augmentation methods, the following results were obtained.

Table 2 reports the accuracy (mean ± standard deviation) and macro-averaged results of the validation set for each CNN model, which were used as baseline experiments to show the classification results before adding any method of data augmentation. It is observed that on the baseline dataset, the VGG16 ensemble classifier shows superior results after fine-tuning, with a validation accuracy of 78.2%.

Table 2. Performance of VGG and inception classifiers using the original imbalanced dataset.

	Accuracy	**AUC Score**	**Precision (Macro Avg.)**	**Recall (Macro Avg.)**	**F1 Score (Macro Avg.)**
VGG16	0.761 ± 1.35	0.869	0.594	0.574	0.584
Inception V4	0.716 ± 1.20	0.852	0.571	0.553	0.562
VGG16, Ensemble	0.782 ± 2.25	0.864	0.607	0.589	0.610
Inception, Ensemble	0.694 ± 1.80	0.822	0.541	0.524	0.532

In the following experiments, we mainly focused on the VGG16 ensemble classifier with a fine-tuning approach to train classifiers with different combinations of augmented data, as it was the best-performing learning scheme in all initial experiments. We also observed the performance of the Inception V4-based architecture.

Table 3 shows the macro-averaged results in terms of precision, recall, F1 score, and accuracy of all the experiments performed after augmentation using the three different augmentation techniques. Next, the improvement in performance for each class can be observed, specifically for the initially imbalanced classes. We can see that the original accuracy favors overrepresented classes, which leads to misleading results. To be more specific, the minority and imbalanced classes (drought and hurricane images) benefited the most from GAN augmentation, as we can see that precision and recall were, on average, six times higher after augmentation for drought images and doubled for the hurricane images, as shown in Table 4.

Table 3. Performance of VGG and inception classifiers after applying data augmentation.

	Accuracy	AUC Score	Precision (Macro Avg.)	Recall (Macro Avg.)	F1 Score (Macro Avg.)
VGG, Geometric Aug.	0.792 ± 1.55	0.915	0.633	0.646	0.641
VGG, CutMix Aug.	0.763 ± 1.50	0.837	0.627	0.571	0.598
VGG, GAN Aug.	0.862 ± 1.70	0.928	0.755	0.786	0.771
VGG, Ensemble, GAN Aug.	0.871 ± 2.25	0.942	0.787	0.804	0.794
VGG, Ensemble, GAN and Geometric Aug.	0.885 ± 2.20	0.952	0.827	0.834	0.831
Inception, GAN Aug.	0.785 ± 1.65	0.918	0.592	0.619	0.610

Table 4. Performance metrics per class for the best-performing model VGG ensemble (drought and hurricane classes highlighted in red initially displayed severe imbalance prior to augmentation).

	Augmentation Method			
	None		GAN Augmentation	
Class	Precision	Recall	Precision	Recall
Land Slide	0.775	0.775	0.857	0.826
Flood	0.852	0.751	0.837	0.857
Fire	0.752	0.722	0.825	0.839
Structures	0.782	0.669	0.802	0.825
Non-Damage	0.652	0.813	0.836	0.816
Drought	0.113	0.078	0.681	0.762
Hurricane	0.322	0.314	0.669	0.700

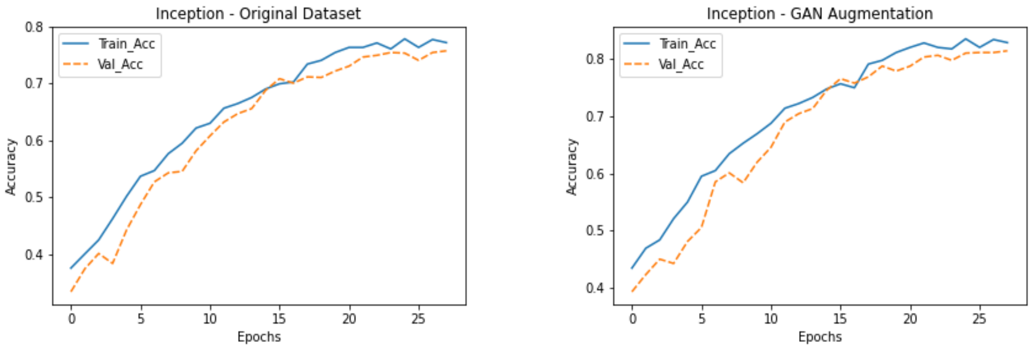
5. Discussion

This study aimed to produce a deep learning framework for disaster classification that outperforms current state-of-the-art models. The results show that our approach can be used to effectively classify damage incidents. We first compared the performance of different baseline models of inception and VGG19 in terms of accuracy and loss measures for each training and validation set.

Figure 11 shows that while the inception model is unable to reach the baseline accuracy achieved by the VGG16 model, an average improvement of 9% is detected after applying GAN augmentation. However, the model exhibits signs of overfitting when the number of epochs increases.

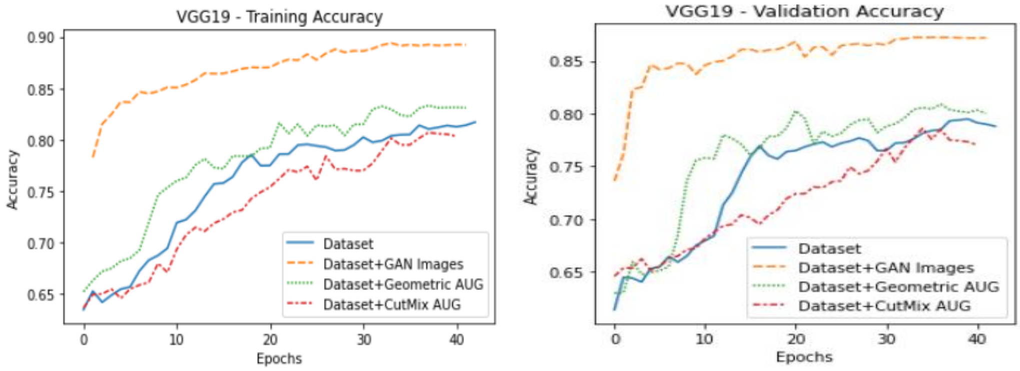
Further experiments focused on the use of the VGG architecture since it provided a noticeably higher performance on the original dataset in comparison to the inception model. After fine-tuning the VGG16 model using a transfer learning approach, the results were obtained for the different types of augmentation.

It can be seen that the baseline model trained on the original unbalanced dataset produced an accuracy of 76.1%. CutMix and geometric augmentation show improvement in accuracy, but they still visibly underperform in comparison to generative adversarial network-based augmentation. After applying GAN augmentation to balance the dataset, the accuracy increases to 86.2%, as shown in Figure 12.



(**a**) (**b**)

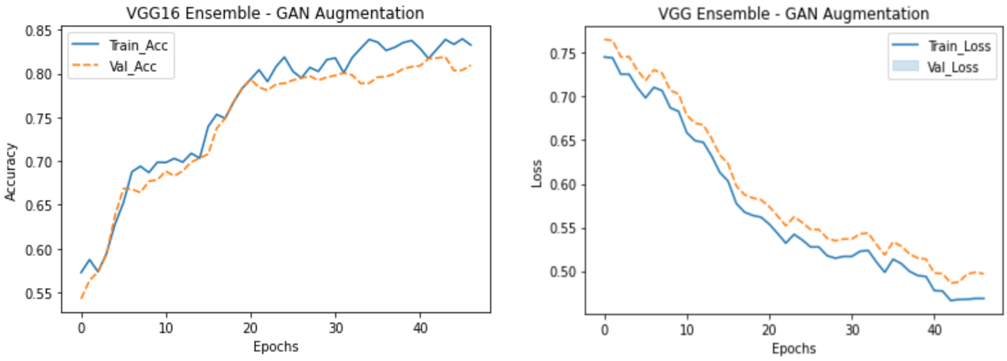
Figure 11. Inception V4 Training and Validation Accuracy. (**a**) Model trained on the original dataset. (**b**) Model trained on the augmented Dataset.



(**a**) (**b**)

Figure 12. VGG16 validation accuracy using original dataset, GAN augmentation, geometric augmentation, and CutMix augmentation. (**a**) training accuracy, (**b**) validation accuracy.

To further improve the classification results, a deep ensemble learning approach was applied to enhance the model's generalization ability, which significantly improved the performance over traditional models. Validation accuracy curves showed that using the augmented dataset to train an ensemble VGG classifier with a bagging technique outperformed the previous results using traditional architectures. From the results shown in Figure 13, we can observe that the framework reaches a validation accuracy of 88.5%, representing an 11% improvement compared to the original baseline model, which is the highest level of performance achieved on similar disaster datasets.



(**a**) (**b**)

Figure 13. VGG16 ensemble classifier training and validation results. (**a**) Model Accuracy. (**b**) Model loss.

Lastly, Figure 14 displays the confusion matrix of the best end-to-end model (VGG-based ensemble classifier with geometric and generative data augmentation). The matrix further highlights the improvement in the classification rate for the previously imbalanced classes.

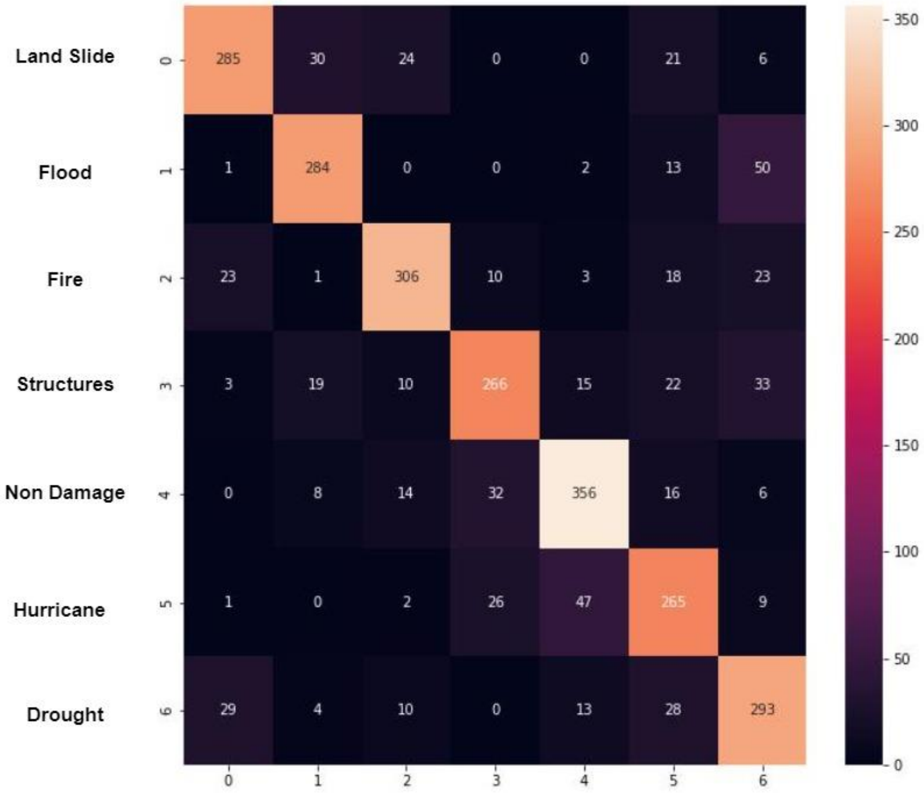


Figure 14. Confusion matrix of the best-performing model.

This finding verifies the effectiveness of the ensemble approach in adding bias to each individual model, thereby increasing the overall framework's generalization ability. We have additionally shown the validity of applying generative data augmentation to balance under-represented classes and improve the performance metrics obtained by different classifiers. In comparison with geometric transformation and CutMix augmentation, the proposed generative method has proven to be superior.

Compared to the traditional data augmentation methods, the model has shown its effectiveness; however, this approach requires a sufficient amount of data in initially unbalanced classes in order to adequately train the GAN model. If the number of samples per class is too small, this approach may not be applicable. The overall classification accuracy for the disaster classification problem showed an average improvement of 11% when the proposed method was applied.

As a result, the proposed framework can be implemented as a real-time disaster monitoring system that accurately and robustly detects the occurrence of any incident and automatically alerts first aid responders. Furthermore, the validity of the generative augmentation approach has been verified, which implies that this framework can be fine-tuned to suit any image classification task in which one or more classes similarly suffer from under-representation.

6. Conclusions

In this paper, we proposed a comprehensive framework for the classification of natural disasters by utilizing insights collected from social media. The initial dataset suffered from a severe class imbalance, which we tackled by training a generative adversarial network to generate high-quality synthetic samples to fortify the original dataset. The quality of the generated image was verified by measuring the inception score and the Fréchet inception distance for each synthetic sample. We conducted extensive experiments and trained an ensemble classifier of VGG16 models by applying a bagging approach. We verified the effectiveness of the proposed ensemble approach in combination with data augmentation by comparing the framework's performance metrics to those of the traditional convolutional neural models.

The final framework achieved an accuracy of 88.5%, which significantly exceeded the performance of all other approaches for tackling the same task by an average of 11%. This framework can be implemented to collect real-time data across all social media platforms and perform spatial analyses and classifications. As a result, dangerous situations and incidents can be rapidly identified and contained.

The proposed framework could be deployed in a web application to collect social media data in real time and accordingly update the weights of the trained model, thus enhancing its performance. This application could be licensed to institutions and authorities to send alerts in the case of sudden disasters.

Future work can be directed towards investigating the further development of the framework by including multimodal features. We believe that utilizing textual and geographical data describing each disastrous event in combination with disaster images could further improve the classification results.

Author Contributions: Rokaya Eltehewy, Ahmed Abouelfarag, Sherine Nagy Saleh: Conceptualization, methodology, and software; Rokaya Eltehewy: data curation, and original draft preparation; Ahmed Abouelfarag and Sherine Nagy Saleh: supervision, reviewing, and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work received no special funding.

Data Availability Statement: The complete Dataset is available in the following publicly accessible repository "github.com/Rokaya78/Imbalanced-Disaster-Classification (accessed on 5 June 2023)".

Conflicts of Interest: All authors of this paper certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

1. Dong, Z.S.; Meng, L.; Christenson, L.; Fulton, L. Social media information sharing for natural disaster response. *Nat. Hazards* **2021**, *107*, 2077–2104. [CrossRef]
2. Alam, F.; Ofli, F.; Imran, M.; Alam, T.; Qazi, U. Deep Learning Benchmarks and Datasets for Social Media Image Classification for Disaster Response. In Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), The Hague, The Netherlands, 7–10 December 2020; pp. 151–158. [CrossRef]
3. Ma, Z.; Mei, G. Deep learning for geological hazards analysis: Data, models, applications, and opportunities. *Earth-Sci. Rev.* **2021**, *223*, 103858. [CrossRef]
4. Hossain, E.; Hoque, M.M.; Hoque, E.; Islam, S. A Deep Attentive Multimodal Learning Approach for Disaster Identification from Social Media Posts. *IEEE Access* **2022**, *10*, 46538–46551. [CrossRef]
5. Aamir, M.; Ali, T.; Irfan, M.; Shaf, A.; Azam, M.Z.; Glowacz, A.; Brumercik, F.; Glowacz, W.; Alqhtani, S.; Rahman, S. Natural disasters intensity analysis and classification based on multispectral images using multi-layered deep convolutional neural network. *Sensors* **2021**, *21*, 2648. [CrossRef] [PubMed]
6. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 27. [CrossRef]
7. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
8. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6023–6032. [CrossRef]
9. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
10. Olteanu, A.; Castillo, C.; Diaz, F.; Vieweg, S. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, Ann Arbor, MI, USA, 1–4 June 2014. [CrossRef]

11. Lagerstrom, R.; Arzhaeva, Y.; Szul, P.; Obst, O.; Power, R.; Robinson, B.; Bednarz, T. Image classification to support emergency situation awareness. *Front. Robot. AI* **2016**, *3*, 54. [CrossRef]
12. Yang, W.; Zhang, X.; Luo, P. Transferability of convolutional neural network models for identifying damaged buildings due to earthquake. *Remote Sens.* **2021**, *13*, 504. [CrossRef]
13. Alam, F.; Ofli, F.; Imran, M. Crisismmd: Multimodal twitter datasets from natural disasters. In Proceedings of the Twelfth International AAAI Conference on Web and Social Media, Palo Alto, CA, USA, 25–28 June 2018. [CrossRef]
14. Bischke, B.; Helber, P.; Schulze, C.; Srinivasan, V.; Dengel, A.; Borth, D. The Multimedia Satellite Task at MediaEval 2017. In Proceedings of the MediaEval, Dublin, Ireland, 13–15 September 2017.
15. Benjamin, B.; Patrick, H.; Zhengyu, Z.; Damian, B. The multimedia satellite task at mediaeval 2018: Emergency response for flooding events. In Proceedings of the MediaEval, Sophia Antipolis, France, 29–31 October 2018.
16. Valdez, D.B.; Godmalin, R.A.G. A Deep Learning Approach of Recognizing Natural Disasters on Images Using Convolutional Neural Network and Transfer Learning. In Proceedings of the International Conference on Artificial Intelligence and its Applications, Suzhou, China, 15–17 October 2021; pp. 1–7. [CrossRef]
17. Hong, Z.; Zhong, H.; Pan, H.; Liu, J.; Zhou, R.; Zhang, Y.; Han, Y.; Wang, J.; Yang, S.; Zhong, C. Classification of Building Damage Using a Novel Convolutional Neural Network Based on Post-Disaster Aerial Images. *Sensors* **2022**, *22*, 5920. [CrossRef]
18. Liang, T.; Lin, G.; Wan, M.; Li, T.; Ma, G.; Lv, F. Expanding Large Pre-Trained Unimodal Models with Multimodal Information Injection for Image-Text Multimodal Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 15492–15501. [CrossRef]
19. Khattar, A.; Quadri, S.M.K. Generalization of convolutional network to domain adaptation network for classification of disaster images on Twitter. *Multimed. Tools Appl.* **2022**, *81*, 30437–30464. [CrossRef]
20. Shahbazi, M.; Huang, Z.; Paudel, D.P.; Chhatkuli, A.; Van Gool, L. Efficient conditional gan transfer with knowledge propagation across classes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12167–12176. [CrossRef]
21. Abu-Srhan, A.; Abushariah, M.A.; Al-Kadi, O.S. The effect of loss function on conditional generative adversarial networks. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 6977–6988. [CrossRef]
22. Wang, Z.; She, Q.; Ward, T.E. Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–38. [CrossRef]
23. Ma, F.; Li, Y.; Ni, S.; Huang, S.-L.; Zhang, L. Data Augmentation for Audio-Visual Emotion Recognition with an Efficient Multimodal Conditional GAN. *Appl. Sci.* **2022**, *12*, 527. [CrossRef]
24. Motamed, S.; Rogalla, P.; Khalvati, F. Data augmentation using Generative Adversarial Networks (GANs) for GAN-based detection of Pneumonia and COVID-19 in chest X-ray images. *Inform. Med. Unlocked* **2021**, *27*, 100779. [CrossRef] [PubMed]
25. Rui, X.; Cao, Y.; Yuan, X.; Kang, Y.; Song, W. DisasterGAN: Generative Adversarial Networks for Remote Sensing Disaster Image Generation. *Remote Sens.* **2021**, *13*, 4284. [CrossRef]
26. Yang, C.; Wang, Z.; Mao, S. RFPose-GAN: Data Augmentation for RFID Based 3D Human Pose Tracking. In Proceedings of the 2022 IEEE 12th International Conference on RFID Technology and Applications (RFID-TA), Cagliari, Italy, 12–14 September 2022; pp. 138–141. [CrossRef]
27. Borji, A. Pros and cons of GAN evaluation measures: New developments. *Comput. Vis. Image Underst.* **2022**, *215*, 103329. [CrossRef]
28. Zahid, Y.; Tahir, M.A.; Durrani, M.N. Ensemble learning using bagging and inception-V3 for anomaly detection in surveillance videos. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 588–592. [CrossRef]
29. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6999–7019. [CrossRef]
30. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A comprehensive survey on transfer learning. *Proc. IEEE* **2020**, *109*, 43–76. [CrossRef]
31. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
32. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]
33. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017. [CrossRef]
34. Khan, A.; Sohail, A.; Zahoora, U.; Qureshi, A.S. A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.* **2020**, *53*, 5455–5516. [CrossRef]
35. Mouzannar, H.; Rizk, Y.; Awad, M. Damage Identification in Social Media Posts using Multimodal Deep Learning. In Proceedings of the ISCRAM, Rochester, NY, USA, 20–23 May 2018.
36. Nguyen, D.T.; Ofli, F.; Imran, M.; Mitra, P. Damage assessment from social media imagery data during disasters. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Sydney, Australia, 31 July–3 August 2017; pp. 569–576.

37. Buda, M.; Maki, A.; Mazurowski, M.A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **2018**, *106*, 249–259. [CrossRef] [PubMed]
38. Buda, M.; Maki, A.; Mazurowski, M.A. A note on the inception score. *arXiv* **2018**, arXiv:1801.01973.
39. Yu, Y.; Zhang, W.; Deng, Y. *Frechet Inception Distance (FID) for Evaluating GANs*; China University of Mining Technology Beijing Graduate School: Beijing, China, 2021.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.