RESEARCH & INNOVATION DIVISION
IJIST

Analysis of Social Media Imagery for Crisis Management Applications

Gul e Rana Iftikhar[1], Masood Ahmad Arbab[1], Muhammad Iftikhar Khan[2], Atif Sardar Khan[3]
[1] Department of Computer Systems Engineering, University of Engineering Technology, Peshawar, Pakistan
[2] Department of Electrical Engineering, University of Engineering Technology, Peshawar, Pakistan
[3] US–Pakistan Center for Advanced Studies in Energy, University of Engineering & Technology, Peshawar, 25000, Pakistan
*Correspondence: maagul1986@gmail.com, arbabmasood@uetpeshawar.edu.pk, miftikhar@uetpeshawar.edu.pk, atifsardarkhan@uetpeshawar.edu.pk

Citation| Iftikhar, G.R., Arbab. M. A Khan. M. I., Khan. A. S, "Analysis of Social Media Imagery for Crisis Management Applications", IJIST, Vol. 07 Issue. 02 pp 1320-1334, June 2025
Received| June 05, 2025 **Revised**| June 27, 2025 **Accepted**| June 29, 2025 **Published**| June 30, 2025.

Social media data holds immense potential for real-time disaster response. This study explores leveraging deep learning to automatically detect disaster-related information across various social media platforms. By analyzing the performance of different models in identifying relevant content, we aim to reduce information gathering delays and support timely rescue efforts. Faster information gathering translates to quick deployment of rescue teams, potentially saving lives and minimizing property damage. We evaluate these models on a benchmark dataset and explore the potential of combining them for even greater accuracy. Among the models, VGG16 achieved an accuracy of 81% in identifying disaster-related content. Additionally, exploring different fusion techniques for combining these models further improved accuracy to 83% with Hybrid Fusion. This research offers valuable insights for future exploration of deep learning techniques in disaster management.

Keywords: CNN, Disasters, Fusion, Social media networks, SVM.

Introduction:

Disasters such as earthquakes, wildfires, hurricanes, floods, and other disasters have devastating consequences for human lives and property. When such tragedies strike, having timely access to vital information is critical for conducting rescue operations. This stresses the necessity of quick access, which makes it easier for the appropriate authorities to respond when such events occur. Collecting accurate information in the aftermath of such disasters is a difficult task. Humanitarian and news organizations often struggle to provide timely updates due to limited access to the affected areas. As social media continues, its role as a key platform for disseminating information has become increasingly evident. As a result, in such emergencies, researchers are particularly interested in deducing facts from visual and textual content posted on social media [1]. Much effort has been made to develop effective methods and techniques for analyzing social media content for disaster detection. This demonstrates the growing interest of researchers in this field. Performance analysis of several feature sets retrieved by different deep learning models and their combination for disaster-related social media photo classification is the subject of this research effort.

Any disaster requires timely action to avoid huge destruction and loss. Timely action and access to relevant data in case of any disaster are important for the authorities to carry out their activities and rescue operations [2]. Social media is increasingly showing its potential as a tool that can save lives. Qualitative, quantitative, and behavioral research shows that social media has proven to be helpful and of great service during disasters. Social media can be useful for the authorities as well as the public to assess the situation and degree of disaster. Since collecting such data is not an easy task, it is important to come up with efficient tools and techniques to collect visual data. This research focuses on the issue of identifying disasters using images from social media and inspecting them from two different angles. Firstly, it assesses the effectiveness of various deep learning models in handling this specific task. Secondly, it investigates the synergies between these diverse models to enhance the overall system performance.

The research aims to propose efficient methods and techniques for disaster management that can assist rescue authorities in obtaining timely information and responding appropriately to evolving situations. Our goal is to utilize various machine learning and computer vision techniques to create a system that collects updates about disasters during emergencies and delivers timely information.

Literature Review:

In recent years, the growing use of social media has made it a leading platform for rapidly and easily obtaining information. Various social media platforms have proven to be a vital source of information transmission in various emergency circumstances. Researchers are trying to develop tools that can figure out information from what people post on social media during emergencies. The use of social media during disasters has garnered attention. Several benchmark initiatives, such as the MediaEval Challenge, have emerged in this field. Numerous strategies and approaches have been proposed in the literature for analyzing social media content to support disaster detection and identification [1]. The information used for this purpose includes text, photos, and videos. To address this issue, a range of feature extraction and classification algorithms have been developed. The content posted on social media platforms is usually in the form of text or visuals such as videos and photographs, depending on the nature of the network. This study focuses on Twitter as a social medium for our research. The study, however, can be applied to other social media sites such as Instagram, Flickr, and Facebook, among others.

Images and videos from social media have been used for catastrophe research in the same way as textual data. Visual content is usually associated with two types of information: Visual features and Metadata about the visual content. Both pieces of knowledge have been

frequently used, either separately or in combination. One study showed the combination of textual data with visual characteristics of the images provided in the form of user tags for Flickr photos [3]. This research utilizes two types of visual features: Low-level color characteristics based on HSV and mid-level object features extracted through SPCPE [4]. The word frequency is used to derive textual properties (word count). Another approach that can be utilized is transfer learning. The system utilizes a pre-trained CNN model to extract visual features along with metadata such as user tags, location, and date. The accuracy is enhanced by 1.7 percent when the feature vectors of both characteristics are fused together. The goal of this study is to determine the type of disaster by analyzing images on social media. We utilized both types of features individually and in combination for evaluation purposes. To extract visual features from the provided images, our approach leverages deep architectures, advanced computational frameworks designed for high-level feature representation. An approach proposed in [5] used a pre-trained CNN model on both the ImageNet and Places datasets [6]. Scene-level and item-specific information were extracted from the photos using the CNN model.

The scores from each separate classifier are combined using the late fusion method, and these features are used to train SVM classifiers. Authors [7] use three pre-trained models for feature extraction and combine results with Induced Ordered Weighted Averaging Operators. Author [8]. suggested a domain-specific approach to late fusion for merging various visual features and metadata for disaster detection tasks. This research has combined different strategies, such as late fusion, tuning, and ensemble learning, to address the specific needs of the project. The job of detecting disaster occurrences from social media is viewed as an ensemble learning and tuning problem, with supervised learners using visual features. A Feedforward neural network is trained using textual features for retrieving disaster photos based on metadata [9]. To enhance efficiency, the metadata is processed beforehand to eliminate unneeded information like URLs and image names. Another study proposes the use of a CNN model to extract visual features [10], combined with a Relational Network (RN) to process metadata. The trials reveal that when CNN and RN are coupled for visual characteristics and metadata, the greatest results are produced. Author [11], presented an image analysis framework for analyzing social media photos during emergency circumstances by integrating human experts and machine learning algorithms. The framework's main goal is to complete two tasks: collecting and filtering images from social media that are related to emergencies and to extract useful information from social media content.

The AIDR system, which is freely accessible, is used for collecting images. These images are later annotated by human annotators to provide relevant labels and context. A pre-trained VGGNet-16 CNN model is employed to fine-tune the collected images for classification purposes [12]. Any classification assignment necessitates the selection of features. A study by author (2016) provides an in-depth comparison of two distinct types of features, namely global and deep features, as well as alternative classification approaches for a big dataset to address this difficulty. The dataset includes a significant number of photos relating to various calamities.

A recent study [13] highlights a pressing challenge, the overwhelming surge of information during disasters such as floods. The research utilizes machine learning techniques to analyze large volumes of social media data, with a primary focus on tweets. By automatically classifying flood-related tweets, the system enables emergency services to swiftly identify and prioritize critical information.

Aims and Objectives:

The following are the goals of this research project:

1. To apply deep learning models to analyze images posted on social media to detect and respond to disasters.

2. To implement a fusion of deep learning models to assess images shared on social media for disaster detection.
3. To investigate the application of several fusion approaches to recognize disaster-related photographs from user-generated social media images.

Research Methodology:

Dataset Description: The dataset used for experiments in our research problem has been taken from the Qatar Computing Research Institute (QCRI). The dataset contains a total of 18082 images of humanitarian disasters. For the first task, we divided the images into test and training sets randomly. In this task, we extracted deep features through different pre-trained CNN models, i.e., VGG16, ResNet50, and DenseNet121, and then we performed different fusion techniques. The second task for the research problem was to divide the humanitarian disaster images into different categories of disasters. We labeled each humanitarian disaster with numeric labels to help us identify the type of disaster at a specific place. Thus, the system not only analyzed the image dataset but also identified the disaster type from the images by using different techniques and algorithms.

Convolutional Neural Networks:

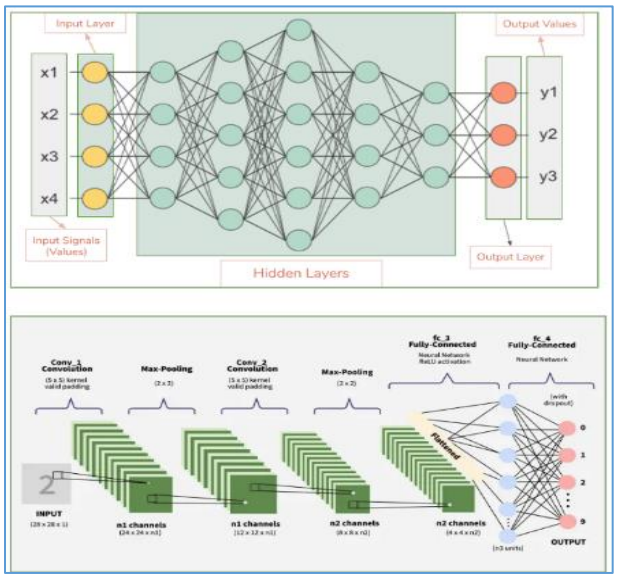

Figure 1. CNN Layers Visualization Source [14]

Feed-forward neural networks, such as Convolutional Neural Networks (CNNs), are widely used in the field of artificial intelligence, particularly renowned for their effectiveness in image recognition tasks. A CNN treats the input as a multidimensional array. A standard architecture contains an input layer, multiple hidden layers, and an output layer, with the hidden layers typically formed by consecutive convolutional layers. CNNs perform effectively when a large amount of labeled data is available. In Convolutional Neural Networks (CNNs), the receptive field refers to the specific region of the input image that a neuron is sensitive to. It determines how much spatial context a neuron "sees". Global features are captured by larger receptive fields, while smaller receptive fields help capture fine details. During training, weights of the convolutional filters are learned through backpropagation, enabling the model to identify and emphasize important patterns within each receptive field. Recent updates about CNN [15] include the discovery of extraordinary capabilities in sequential data analysis, like natural language processing. These are built on the assumption that the network's input is an image. This assumption allows the network to incorporate specific attributes to improve feed-forward computation efficiency by reducing network parameters.

CNN Architecture:

Figure 1 shows two types of networks. On the top is a conventional neural network,

while at the bottom is a Convolutional Neural Network (CNN). In a traditional neural network, the input is received as a single vector and then transformed through multiple hidden layers according to the model's architecture [16]. There is a specific number of neurons in each hidden layer that connect to all neurons in the last layer. One layer of neurons, in contrast, acts fully on their own, and there is no communication of information between any of them. The "output layer" is the final layer of the network and, in classification tasks, represents the scores for each class. As expected, this architecture does not scale efficiently when dealing with large images. In a standard neural network, a single neuron in the first layer would require approximately 3,072 weights to process an image with a resolution of 32x32x3. For example, a 200x200x3 image will result in a network with 120,000 weighted neurons.

The presence of many neurons results in a network with many parameters. As a result, a completely connected design is inefficient and frequently leads to overfitting. The arrangement of neurons in three dimensions becomes visible when viewing one of the network layers. In this example, the red layer is considered as the input image, the dimensions of the image show its width and height, while depth is set to 3 in this very example (RGB Image). In summary, a CNN is composed of multiple layers [17], each applying a specific function to transform a 3D input into a 3D output. These transformations may or may not involve learnable parameters.

Layers Of CNN:

Every CNN is made up of a series of layers [18], each of which is responsible for converting a three-dimensional input into a three-dimensional output volume using a differentiable function. In CNN, there are three sorts of layers.

- Input layer.
- Convolution layer.
- RELU layer.
- Pooling layer.
- Fully connected layer.

Input layer: Within a CNN architecture, the input layer encompasses all the CNN's data. This layer typically symbolizes the pixel array of an image in a neural network for image processing.

Convolution layer: A convolutional neural network relies on the convolution process, which is made possible by the convolution layer. The convolutional layer's parameters consist of small learnable filters (kernels) that cover the entire depth of the input volume. The most common filter sizes are 3x3, 5x5, and 7x7. The number of channels in the input is the third dimension of the filter. The color image has three (RGB) color channels, and the grayscale image depth is one. During forward propagation, each filter convolves with the input volume across width and height, calculating dot products at each position. Next, a nonlinear activation function such as sigmoid, tanh, or ReLU is applied, and the resulting outputs are known as feature maps. The feature map (sometimes called an activation map) displays the filter's responses at each spatial location. To produce the output volume, the activation maps are stacked along the depth dimension. The size and structure of this output volume are influenced by three key hyperparameters: depth, stride, and padding.

The output volume's depth corresponds to the number of filters used in the convolution. Each filter learns different features from the input, like edges, blobs, and colors. The stride indicates how many steps the filter moves over the input. When the stride value is set to 1, the filter moves one pixel at a time. When we use a stride of 2, the filters move 2 pixels at a time, reducing production volume in terms of space. Padding allows you to regulate the size of the output. When convolution is applied to an input, the output size is reduced, resulting in information loss. We pad the input volume with zeros at the border to avoid this. Valid convolution and the same convolution are two popular options. In the case of the same convolution, the output size remains the same as the input size, while in a valid convolution,

there is no padding. The output size is determined as

$$(n + 2p – f) / s + 1.$$

Where n denotes the number of filters, p the padding, f the filter size, and s the stride length.

The RELU layer: The RELU layers use an element-by-element activation function to keep the volume [32x32x12] constant. For instance, max (0, x) sets the input to zero.

Pooling layer: After convolution layers, CNNs frequently use the pooling layer technique to reduce the dimension, which is also known as subsampling or down-sampling. When designing ConvNet, it is typical to include a Pooling layer after successive Convolution layers. The Pooling layer's primary objective is to gradually decrease the spatial dimension of the representation, which reduces the number of parameters and computations required in the network and helps prevent overfitting. This layer uses the MAX operation to resize the input's spatial dimensions independently for each depth section. One common version involves using a pooling layer with 2x2 filters and a stride of 2. This down samples each layer in the input by 2 along both width and height, removing 75 percent of the activations. As a result, each MAX operation takes a maximum of four values within a 2x2 region of a given layer.

Fully connected layer: A fully connected layer is a type of neural network that consists of a series of interconnected layers. Each layer is fully linked, meaning that every neuron in one layer is connected to every neuron in the next. Fully connected (FC) layers are usually positioned at the end of CNN architecture and are used to optimize specific objectives, such as class scores, when such targets are available. By removing the last layer of a CNN model, it can be repurposed as a feature extractor.

Our proposed approach comprises three main steps: feature extraction, classification, and fusion. In the first stage, different pre-trained deep learning models are utilized to extract features from the image dataset. These models help capture important visual representations, which serve as the foundation for the subsequent classification and fusion processes. Once these features were extracted, the next step was to train a classifier to classify the user-generated photos using these features. Support Vector Machine (SVM) was selected as the classification method for this study due to its ability to effectively handle non-linear data through the use of various kernel functions. Several types of fusion approaches were also employed to boost classification accuracy. Different classification scores or feature descriptors were integrated into the fusion procedure, resulting in improved final classification accuracy. The following subsections provide a detailed explanation of the methods used for extracting features, classifying the data, and combining model outputs.

Feature Extraction:

In image classification tasks like disaster image identification, feature extraction plays a critical role. It refers to the process of transforming raw image data into a more compact and informative representation that captures the essential characteristics relevant to the classification task. [19]

In this study, we utilized pre-trained neural networks for extraction by taking advantage of their ability to capture complex visual patterns from large-scale datasets.

Pre-trained Models: We utilized pre-trained Convolutional Neural Networks (CNNs) as feature extractors. CNNs are a powerful class of deep learning models specifically designed for image recognition. These models are trained on massive datasets like ImageNet, which contain millions of labeled images across diverse categories.

Transfer Learning: Instead of training a CNN from scratch on your specific disaster image dataset (which can be resource-intensive), we employed the concept of transfer learning. This technique leverages the knowledge already learned by a pre-trained CNN on generic image features like edges, textures, and shapes. These pre-trained models act as feature extractors, and we can then utilize the learned features for our specific disaster classification task.

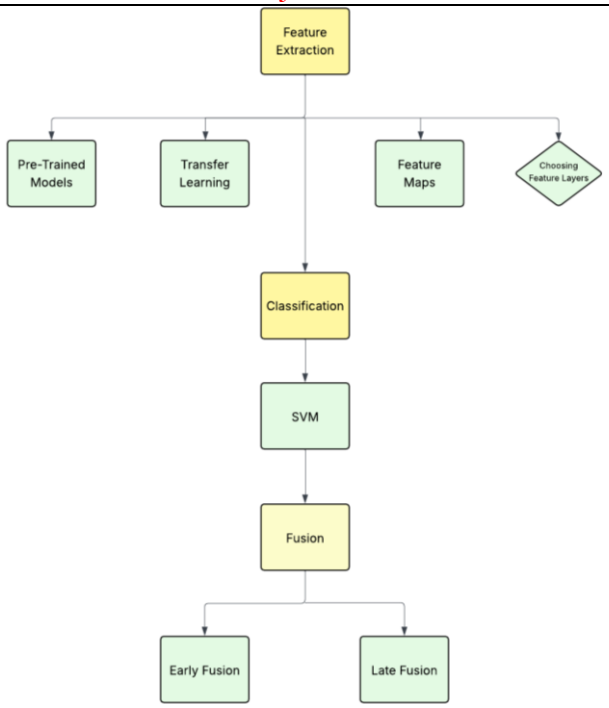

Figure 2. Flowchart diagram of the Methodology

Feature Maps: During the forward pass of a pre-trained CNN, the network processes the input image through multiple convolutional layers. Each convolutional layer generates a feature map, which essentially represents the image's activation at different levels of abstraction. The earlier layers capture low-level features like edges and textures, while later layers learn more complex and class-specific features.

Choosing Feature Layers: Based on the complexity of the disaster classification task, feature maps can be selected from specific layers within the pre-trained CNN to best capture the relevant information. For example, for broad disaster categories like floods or earthquakes, earlier layers capturing general image properties might suffice. For more granular classification (e.g., differentiating between types of floods), features from deeper layers that encode more specific object details might be necessary.

Classification:

After extracting the features, the next step was to classify the obtained data. Among the various classifiers considered, we selected the Support Vector Machine (SVM) due to its proven superior classification performance, as highlighted in recent studies. Several recent studies show that support vector machines (SVMs) generally outperform other data classification algorithms in terms of accuracy. Once the pre-trained models have been used to retrieve the features, the data is classified using these features. In this study, Support Vector Machines (SVMs) were chosen due to their capability to model both linear and nonlinear data effectively. The following section offers a more detailed overview of SVM and its application in our approach.

Support Vector Machine (SVM):

Support vector machines are based on supervised learning used for regression and classification analysis. For classification, SVM forms several hyperplanes in high-dimensional feature space between two classes to classify them. When data is linearly separable, SVM constructs a set of hyperplanes, but the hyperplane with maximum margin is chosen. Maximum margin gives better performance while classifying data. When data is not separable linearly, kernel tricks or kernel functions are used. Kernel functions transform data from a low-dimensional feature space into a high-dimensional one, allowing previously non-linearly

separable data to become linearly separable. There are many kernel functions, and the function used in this research was linear.

Support-vector machines (SVMs) are machine-learning models within the domain of supervised learning. These models utilize learning algorithms to evaluate data for tasks involving classification and regression analysis. While SVMs can be employed for both classification and regression, they are mainly applied to address classification challenges [20]. In Support Vector Machines (SVM), data points are represented in an n-dimensional space, where *n* corresponds to the number of features. Each feature value determines a specific coordinate in this space, allowing the algorithm to plot and separate the data effectively. The value of each feature is mapped to a specific coordinate in the SVM algorithm. This study aims to determine a hyperplane that effectively differentiates the two classes in the data through classification. The support vectors are the vectors that are employed to characterize the hyperplane. The primary objective of this research work is to execute different fusion methods of features/ classification score and then evaluate all of them based on the results.

Fusion:

This research aims to utilize various fusion techniques (combining features and classification scores) and assess their effectiveness based on the outcomes. It can occur either before classification at the feature level or after sorting at the classification score level. Fusion can play a good role in improving the accuracy of the model. In this study, we performed five different fusion techniques. The details of these different techniques are given in the following subsections:

Early Fusion: In the Early Fusion Method, [21] all the local features were extracted and were then combined (points, edges, or objects) to form one large feature vector set to use for classification [22]. Early Fusion was performed by extracting features from each category (in our case we have 8 categories/modalities) and combining them into a single vector set such that, If the feature vector extracted from model 1 is z1 and from model 2 is z2, then the concatenated feature vector is Z= [z1, z2]. Feature fusion and kernel space fusion are the two most used Early Fusion techniques. We used feature fusion for classification in our research project. [23] The key benefit of this technique is that it requires only one learning phase, but it can be difficult to combine all the features into a single common representation.

Late Fusion: Late Fusion combines prediction scores from multiple classifiers, each trained with distinct features or models, to enhance recognition accuracy [24]. In this research work, we targeted three types of score-level fusions given in the upcoming sections.

Late Fusion with Equal Weights: Late fusion is a method where the probabilities from different classifiers trained on different models are combined by concatenating them. For example, if z1 is the probability of a specific class from classifier 1 and z2 is the probability from classifier 2, the probability of the class after late fusion would be z1+z2. While late fusion is simpler than early fusion, it requires more computational effort.

Results And Discussion:

The experiments conducted and the evaluation of the resulting outcomes are discussed in the following section.

Our research work was primarily divided into two main parts. Firstly, itfocused on the classification of humanitarian images and the analysis of the results obtained from this process. Secondly, the images classified during the initial phase, which were further categorized into different types of humanitarian disasters, allowing for a more detailed understanding and assessment of the visual data.

Experimental Setup:

In this study, we analyzed the performance of feature descriptors obtained from three pre-trained CNN models individually. We also combined feature descriptors and classification scores of the classifier using different fusion techniques.

We carried out a few experiments to test the performance of different feature representations and fusion methods. First, we used three different pre-trained CNN models to extract features and see how good the feature descriptors are. Then, we combined these features using early fusion at the feature level to find out if merging them helps improve results. Lastly, we applied score-level fusion on the same feature descriptors to compare it with early fusion. These steps helped us understand which method works better for our task.

Experimental Results of Humanitarian Tasks:

The outcomes of the experiments covered in this research are presented as follows:We evaluated the performance of individual CNN models on humanitarian tasks and presented the outcomes of various fusion techniques used to enhance classification accuracy.

Performance Analysis of Individual CNN Models: The results of evaluating the performance of descriptors derived from the three CNN models are shown in Figure 3. As mentioned earlier, we used CNN models trained on ImageNet datasets to extract deep features. These models are trained to recognize objects within images.

Our initial experiments using three CNN models; VGG16, ResNet50, and DenseNet121 achieved accuracies of 70%, 69%, and 69%, respectively. The influence of these features on overall performance is demonstrated by the results shown in Figure 3.

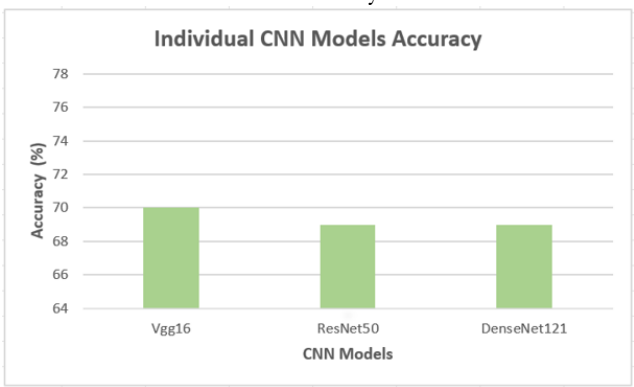

Figure 3. Results of Humanitarian Task

Performance Analysis of Different Fusion Techniques:

After evaluating the performance of individual CNN models, we experimented with different fusion methods on the dataset: early fusion, late fusion, and hybrid fusion. These fusion techniques proved to be beneficial, as they significantly improved the accuracy achieved compared to the previous experiment. Early fusion yielded an accuracy of 75%, late fusion reached 79%, and hybrid fusion achieved the highest accuracy of 80%. The outcomes of using several fusion techniques are presented in Figure 4.

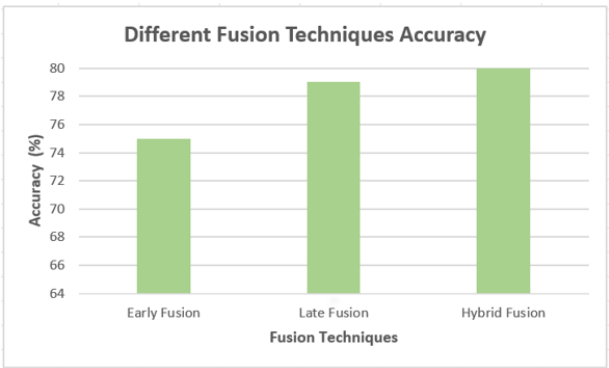

Figure 4. Accuracy for different fusion techniques

Performance Analysis of Individual CNN Models After Removing Imbalanced Classes: During our experiments, we observed potential class imbalances in the data. Class imbalance occurs when the distribution of classes (categories) within a dataset is uneven, with

some classes having significantly fewer examples than others. To address this issue, we removed data points to create a more balanced dataset and repeated the experiment. This approach proved helpful, as the results improved: VGG16 achieved an accuracy of 81%, ResNet50 reached 71%, and DenseNet121 obtained an accuracy of 75%. The minimal f1-score for class 0, class 2, and class 3 is presented in Table 1, emphasizing the data imbalance issue. Additionally, Figure 5 illustrates the accuracy achieved when various classes were excluded from the dataset.

Table 1. Confusion Matrix

	Precision	Recall	F1-score	Support
0.0	0.73	0.39	0.51	175
1.0	0.86	0.98	0.92	1058
2.0	1.00	0.15	0.25	62
3.0	0.00	0.00	0.00	5
4.0	0.76	0.95	0.85	2641
5.0	0.88	0.53	0.66	757
6.0	0.81	0.47	0.60	711
7.0	0.84	0.43	0.56	87
Accuracy			0.80	5496
Macro Avg	0.74	0.49	0.54	5496
Weighted Avg	0.81	0.80	0.78	5496

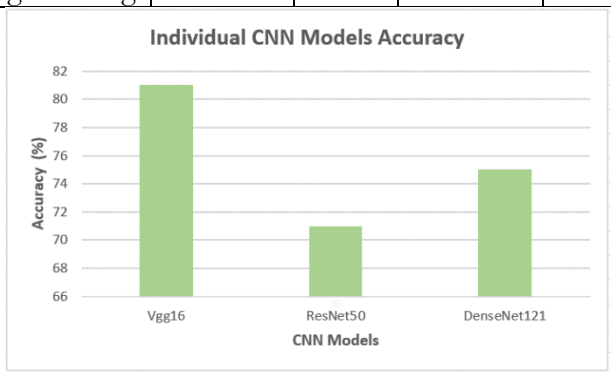

Figure 5. Accuracy obtained from removing different classes.

Performance Analysis of Individual CNN Models After Data Augmentation: To further improve accuracy, we employed data augmentation. This technique involves generating new data from existing data, essentially expanding the training dataset. The more data, the better the accuracy, and data augmentation helps us achieve this. Consequently, a larger and more diverse training dataset led to improved overall average model performance. The final accuracies achieved after data augmentation are presented in Figure 5: VGG16 reached 75%, ResNet50 reached 75%, and DenseNet121 reached 70% Figure 6 presents the accuracy obtained following the application of data augmentation techniques.

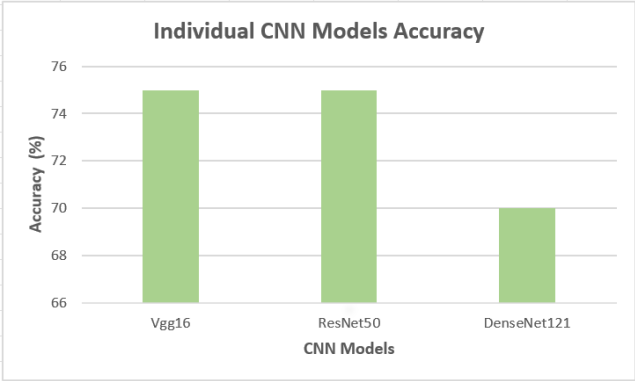

Figure 6. Accuracy obtained from Data Augmentation

Performance Analysis of Different Fusion Techniques After Data Augmentation: We evaluated several fusion techniques and found that, following data augmentation, the Hybrid approach demonstrated outstanding accuracy in this experiment. Figure 7 displays the accuracy graph for the various fusion techniques evaluated in this study.

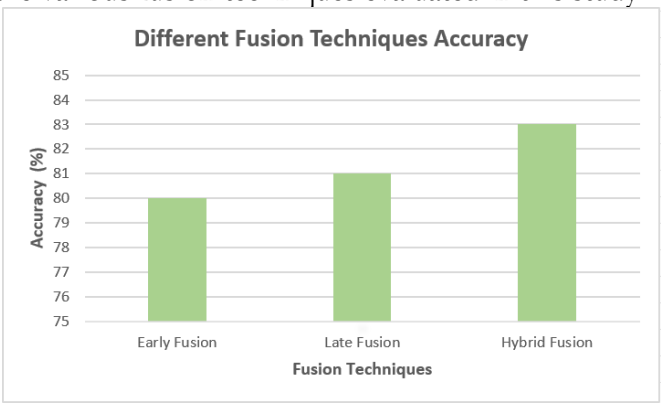

Figure 7. Accuracy obtained from different fusion techniques after Data Augmentation

This highlights the effectiveness of combining data augmentation and fusion techniques in enhancing the overall performance of the model.

Discussion:

The results of our experiments show a consistent improvement in disaster image classification using fusion techniques and data augmentation. Our initial experiments using three pre-trained CNN models, VGG16, ResNet50, and DenseNet121, each achieved approximately 70% accuracy. With the application of fusion methods, performance improved significantly, with hybrid fusion achieving the highest accuracy of 80%, outperforming early (75%) and late fusion (79%). Data augmentation further improved model performance, with VGG16 and ResNet50 reaching 75% accuracy post-augmentation. Even after augmentation, hybrid fusion continued to yield the best results.

When comparing our findings to those of author [1] and the study by author [3], which demonstrated improved retrieval performance using feature fusion, our results align well with the conclusion that combining multiple feature types leads to better performance. More specifically, our results closely resemble those reported in the paper [25] *"Analysis of Social Media Data Using Multimodal Deep Learning for Disaster Response"* (arXiv:2004.11838). That study used multimodal deep learning to combine textual and visual features from social media posts related to disasters. Their best performance in informativeness classification was an F1-score of **84.2%** using combined (text + image) input, while the best result for humanitarian-category classification was an F1-score of **78.3%**. Although our work focuses exclusively on image data, we similarly observed significant improvements when combining features from multiple CNN models. In contrast to the multimodal approach of using both text and images in the

referenced study, we demonstrate that even within the visual modality alone, meaningful gains can be achieved through model-level feature fusion. Moreover, their use of a single CNN (VGG16) for visual feature extraction is expanded in our approach by leveraging multiple CNN architectures, leading to more diverse and complementary features.

In summary, while existing studies emphasize the value of combining multiple data modalities (e.g., text and image), our findings reinforce the effectiveness of intra-modal fusion (within image data) using multiple CNN architectures. Both approaches highlight that fusion, whether across models or data types, is a key strategy for improving classification accuracy in disaster-related tasks.

Conclusion:

During disasters, multimedia content on social media sites delivers vital information. During a crisis or an emergency, information available on social media can help with a variety of humanitarian tasks. Reports of injured or deceased people, infrastructure damage, and missing or found people are among the types of information shared. Although several studies have demonstrated the utility of both text and visual information for disaster response, our research has primarily focused on visual information. Our experiments confirmed that the method of data augmentation showed improvements in the results. Our aim in developing those pipelines was to assess different models for better performance. Our experimentation with data augmentation improved the results. Our work on improving the accuracy of actual datasets, trained through different CNN models, revolved around the idea of introducing additional data to the model through data augmentation. We can conclude that additional data improved our results.

Future Work:

Our research opens doors to further exploration in a few key areas:

Datasets for Specific Severity Levels: While existing research primarily focuses on datasets with generic disaster classifications, there's a gap in exploring datasets labeled for mild and severe disasters. This area presents a valuable opportunity for future work. Developing and utilizing such data sets could enable the creation of models that predict the severity of a disaster based on social media images. This information would be crucial for prioritizing humanitarian assistance, allowing aid teams to rapidly deploy resources to areas experiencing the most critical situations.

Text Data Integration: Our current study primarily focused on visual features extracted from social media images. However, social media posts often include text data alongside the images, potentially containing valuable details about the disaster. Future research could investigate incorporating text data alongside visual features in the classification process. Analyzing both elements together could potentially improve the accuracy and richness of the information extracted from social media posts.

Global Image Features: In addition to deep learning features, exploring global image features like color distribution, texture analysis, and the presence of specific objects could also enhance the classification of disaster severity in social media images. Investigating these features alongside deep learning techniques presents another avenue for future exploration.

By delving into these areas, we can further refine models to provide more specific information regarding the location and severity of disasters. This, in turn, would empower humanitarian teams to target their assistance more effectively and deliver crucial aid to those in dire need more rapidly.

Acknowledgement:

We extend our heartfelt gratitude to all those who played a crucial role in contributing to this research endeavor, each in their unique capacity.

Author contributions:

All authors contributed to the study conception and design. Material preparation, data collection, training, and testing the different forecasting algorithms were performed by Gul-E-Rana Iftikhar. Masood Ahmad Arbab reviewed and edited the draft of the paper. All authors read and approved the final manuscript.

Data Availability:

The most widely used dataset is CrisisMMD: Multimodal Crisis Dataset. The CrisisMMD multimodal Twitter dataset consists of several thousand images collected during seven major natural disasters, including earthquakes, hurricanes, wildfires, and floods that happened in the year 2017 across different parts of the World. [26]

The datasets generated and/or analyzed during the current study are available on the CrisisNLP website with the name.

CrisisMMD dataset version v2.0:

Competing Interests:

The authors have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Funding:

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

References:

[1] Z. Wang and X. Ye, "Social media analytics for natural disaster management," *Int. J. Geogr. Inf. Sci.*, vol. 32, no. 1, pp. 49–72, Jan. 2018, doi: 10.1080/13658816.2017.1367003.

[2] M. Y. Kabir and S. Madria, "A deep learning approach for tweet classification and rescue scheduling for efective disaster management," *GIS Proc. ACM Int. Symp. Adv. Geogr. Inf. Syst.*, pp. 269–278, Nov. 2019, doi: 10.1145/3347146.3359097;WGROUP:STRING:ACM.

[3] S. G. Sevil, O. Kucuktunc, P. Duygulu, and F. Can, "Automatic tag expansion using visual similarity for photo sharing websites," *Multimed. Tools Appl.*, vol. 49, no. 1, pp. 81–99, Aug. 2010, doi: 10.1007/S11042-009-0394-5/METRICS.

[4] M. L. Shyu, S. C. Chen, M. Chen, C. Zhang, and K. Sarinnapakorn, "Capturing high-level image concepts via affinity relationships in image database retrieval," *Multimed. Tools Appl.*, vol. 32, no. 1, pp. 73–92, Jan. 2007, doi: 10.1007/S11042-006-0059-6/METRICS.

[5] N. N. C. Ahmad, Sheharyar Ahmad, Kashif Ahmad, "Convolutional Neural Networks for Disaster Images Retrieval," *MediaEval*, vol. 17, pp. 13–15, 2017, [Online]. Available: https://ceur-ws.org/Vol-1984/Mediaeval_2017_paper_11.pdf

[6] P. H. Kashif Ahmad, Pogorelov Konstantin, Michael Riegler, Nicola Conci, "CNN and GAN Based Satellite and Social Media Data Fusion for Disaster Detection," *MediaEval*, vol. 17, pp. 13–15, 2017, [Online]. Available: http://slim-sig.irisa.fr/me17/Mediaeval_2017_paper_15.pdf

[7] R. R. Yager and D. P. Filev, "Induced Ordered Weighted Averaging operators," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 29, no. 2, pp. 141–150, Apr. 1999, doi: 10.1109/3477.752789.

[8] D.-T. D.-N. Dao, Minh-Son Pham, Quang-Nhat-Minh, "A Domain-based Late-Fusion for Disaster Image Retrieval from Social Media," *MediaEval*, vol. 17, pp. 13–15, 2017, [Online]. Available: https://ceur-ws.org/Vol-1984/Mediaeval_2017_paper_24.pdf

[9] M. S. Dao, P. Quang Nhat Minh, A. Kasem, and M. S. Haja Nazmudeen, "A context-aware late-fusion approach for disaster image retrieval from social media," *ICMR 2018 - Proc. 2018 ACM Int. Conf. Multimed. Retr.*, pp. 266–273, Jun. 2018, doi: 10.1145/3206025.3206047;PAGE:STRING:ARTICLE/CHAPTER.

[10] Y. Wei *et al.*, "Cross-Modal Retrieval With CNN Visual Features: A New Baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017, doi: 10.1109/TCYB.2016.2519449.

[11] A. Eutamene, H. Belhadef, and M. K. Kholladi, "New Process Ontology-Based Character Recognition," *Commun. Comput. Inf. Sci.*, vol. 240 CCIS, pp. 137–144, 2011, doi: 10.1007/978-3-642-24731-6_13.

[12] M. A. I. Osama A. Shawky, Ahmed Hagag, El-Sayed A. El-Dahshan a, "Remote sensing image scene classification using CNN-MLP with data augmentation," *Optik (Stuttg).*, vol. 221, p. 165356, 2020, doi: https://doi.org/10.1016/j.ijleo.2020.165356.

[13] B. R. Eike Blomeier, Sebastian Schmidt, "Drowning in the Information Flood: Machine-Learning-Based Relevance Classification of Flood-Related Tweets for Disaster Management," *Information*, vol. 15, no. 3, p. 149, 2024, doi: https://doi.org/10.3390/info15030149.

[14] GeeksforGeeks. (n.d.), "CNN in machine learning", [Online]. Available: https://www.geeksforgeeks.org/cnn-in-machine-learning/

[15] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018, doi: 10.1016/J.PATCOG.2017.10.013.

[16] B. Dickson, "What are convolutional neural networks (CNN)?," *Bdtechtalks.Com.*, 2020, [Online]. Available: https://towardsml.wordpress.com/2018/10/16/deep-learning-series-p2-understanding-convolutional-neural-networks/

[17] P. Mahajan, "Fully Connected vs Convolutional Neural Networks," *Medium*, 2020, [Online]. Available: https://medium.com/swlh/fully-connected-vs-convolutional-neural-networks-813ca7bc6ee5

[18] N. B. D. Timea Bezdan, "Convolutional Neural Network Layers and Architectures," *Int. Sci. Conf. Inf. Technol. Data Relat. Res.*, 2019, doi: https://doi.org/10.15308/Sinteza-2019-445-451.

[19] I. K. Georgios Kontonatsios, Sally Spencer, Peter Matthew, "Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews," *Expert Syst. with Appl. X*, vol. 6, p. 6100030, 2020, doi: https://doi.org/10.1016/j.eswax.2020.100030.

[20] A. Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, 2020, doi: https://doi.org/10.1016/j.neucom.2019.10.118.

[21] H. Ergun, Y. C. Akyuz, M. Sert, and J. Liu, "Early and Late Level Fusion of Deep Convolutional Neural Networks for Visual Concept Recognition," *https://doi.org/10.1142/S1793351X16400158*, vol. 10, no. 3, pp. 379–397, Nov. 2016, doi: 10.1142/S1793351X16400158.

[22] H. Gunes and M. Piccardi, "Affect recognition from face and body: Early fusion vs. late fusion," *Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern.*, vol. 4, pp. 3437–3443, 2005, doi: 10.1109/ICSMC.2005.1571679.

[23] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," *Proc. 13th ACM Int. Conf. Multimedia, MM 2005*, pp. 399–402, 2005, doi: 10.1145/1101149.1101236;TOPIC:TOPIC:CONFERENCE-COLLECTIONS>MM;PAGE:STRING:ARTICLE/CHAPTER.

[24] S. Balaji, "Binary Image classifier CNN using TensorFlow," *Medium*, 2020, [Online].

Available: https://medium.com/techiepedia/binary-image-classifier-cnn-using-tensorflow-a3f5d6746697

[25] N. Thakur, E. Bhattacharjee, R. Jain, B. Acharya, and Y. C. Hu, "Deep learning-based parking occupancy detection framework using ResNet and VGG-16," *Multimed. Tools Appl.*, vol. 83, no. 1, pp. 1941–1964, Jan. 2024, doi: 10.1007/S11042-023-15654-W/METRICS.

[26] M. I. Firoj Alam, Ferda Ofli, "CrisisMMD: Multimodal Twitter Datasets from Natural Disasters," *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 12, no. 1, 2018, doi: https://doi.org/10.1609/icwsm.v12i1.14983.