Check for updates

DOI: 10.1111/mice.12890

INDUSTRIAL APPLICATION

COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING WILEY

Post-disaster damage classification based on deep multi-view image fusion

Asim Bashir Khajwal | **Chih-Shen Cheng** | **Arash Noshadravan**

Zachry Department of Civil and Environmental Engineering, Texas A&M University, College Station, Texas, USA

Correspondence
Arash Noshadravan, Assistant Professor, Zachry Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX, USA.
Email: noshadravan@tamu.edu

Abstract

This study aims to facilitate a more reliable automated postdisaster assessment of damaged buildings based on the use of multiple view imagery. Toward this, a Multi-View Convolutional Neural Network (MV-CNN) architecture is proposed, which combines the information from different views of a damaged building, resulting in 3-D aggregation of the 2-D damage features from each view. This spatial 3-D context damage information will result in more accurate and reliable damage quantification in the affected buildings. For validation, the presented model is trained and tested on a real-world visual data set of expert-labeled buildings following Hurricane Harvey. The developed model demonstrates an accuracy of 65% in predicting the exact damage states of buildings, and around 81% considering ±1 class deviation from ground-truth, based on a five-level damage scale. Value of information (VOI) analysis reveals that the hybrid models, which consider at least one aerial and ground view, perform better.

1 | INTRODUCTION

Large-scale disaster events are often associated with vital damage to the built infrastructure, in addition to its effects on human life and the economy. The intensity and frequency of such large-scale events have been increasing worldwide over the last few decades (Institute for Economics & Peace, 2020). In the United States itself, the number of billion-dollar disaster events has increased significantly over the recent years (NOAA, 2021). This places enormous pressure on local and federal agencies and officials involved in different stages of disaster management, that is, mitigation, preparedness, response, and recovery. One of the key steps in post-disaster response and management is the assessment of damages to the built infrastructure. This exercise forms the basis for effective post-disaster planning, management, and fund allocations. A key objective of preliminary post-disaster building assessment is to communicate buildings' condition and suitability information to all stakeholders involved in post-disaster response and recovery. An early preliminary damage assessment (PDA) informs the prioritization of resource allocations, planning, and better execution of response and recovery operations. The information collected from PDA also aids engineers and researchers in understanding the trends and impacts of natural disasters on the built environment. The knowledge and information obtained from such assessments also provide very useful information to insurance and reinsurance agencies to make informed decisions on future insurance needs of the community.

In its current practice, PDA is mainly expert-centered and heavily relies on human resources. It is often carried out by trained professionals and expert reconnaissance teams by way of visual inspection and post-disaster surveys (Kang & Cha, 2018). These assessments are primarily based on door-to-door or windshield surveys, remote sensing data, and aerial imagery. The collected data are analyzed

© 2022 *Computer-Aided Civil and Infrastructure Engineering*.

by experts based on which the damage portfolio of the affected region is characterized. However, dependence on experts and trained professionals can be a bottleneck as it makes this process resource-intensive and inefficient in terms of time as well spatial coverage. Also, the most affected areas are the most inaccessible to the expert teams and inspectors for their assessment, thus further limiting or delaying a comprehensive evaluation of damaged buildings. As such, this process of performing PDA may take several weeks to complete. This may result in exacerbated loss due to delay in disaster management and recovery process.

Motivated by overcoming time, accessibility, and safety concerns, there is a need for more efficient practices for data collection and post-disaster damage assessment (Wang et al., 2021). With significant advances in data acquisition and remote-sensing technologies, the focus is now on alternative approaches that could reduce the reliance on human experts (Yuan et al., 2020). There have been several efforts in the past toward automating the detection and assessment of damages in the built infrastructure. At the individual structure or building level, damage detection and structural health monitoring have been extensively studied over the last few decades (Shwe & Adeli, 1993). Several sensor-based models (Oh et al., 2017) and vision-based models (Pan & Yang, 2020) have been developed to assess the buildings for their damage condition. Amezquita-Sanchez and Adeli (2015) presented a novel fractality-based model for detection, location, and quantification of damage state in smart high-rise buildings. Rafiei and Adeli (2017c) developed a novel machine learning model for damage detection in high-rise buildings using an unsupervised machine learning technique. Gao and Mosalam (2018) used deep learning (DL) to identify damages in different elements of the damaged building. Cha et al. (2017) proposed a vision-based DL model to detect damage in the form of cracks in concrete buildings. Pan and Yang (2021) developed a real-time DL model for detection and tracking of bolt loosening and rotation assessment. Liang (2019) performed the post-disaster damage assessment of bridges using DL. However, when it comes to regional scale, mostly vision-based approaches are relied on due to practical consideration. These efforts broadly fall into two categories: crowdsourcing or participatory damage assessment and automated damage assessment using artificial intelligence (AI). Many recent studies have highlighted the use of citizen science and participatory approach to enhance data collection for infrastructure monitoring (Li et al., 2021) and disaster impact assessment (Gharaibeh et al., 2021; Khajwal & Noshadravan, 2021). On the other hand, the promise of AI has not only been encouraging in solving scientific problems in other domains (Rafiei & Adeli, 2017a), it is also emerging as a promising solution for automation and enhancing the efficiency of structural health monitoring and post-disaster damage assessment (Xu et al., 2020). Specifically, in the domain of AI-assisted disaster damage assessment, several DL models have been proposed in the literature to perform post-disaster damage assessment at infrastructure level (Gao et al., 2021; Miura et al., 2020). Xu et al. (2021) proposed a framework for real-time seismic damage assessment at regional scale, based on long short-term memory (LSTM) neural network architecture. Xu et al. (2019) developed a damage recognition model based on convolutional neural network (CNN) that uses satellite images. Gupta et al. (2019) used satellite imagery and a CNN to perform classification of damage levels based on analysis of pre- and post-disaster photos. Sublime and Kalinicheva (2019) proposed a DL approach to evaluate the extent and severity of damage based on pre- and post-disaster images taken after the Tohoku tsunami of 2011. Nex et al. (2019) proposed a classification model based on DL to classify the patches of input aerial imagery into damaged building and undamaged buildings, or identify patches with no building. Chen et al. (2021) explored deep vision models for damage evaluation in the aftermath of a tornado event. Pi et al. (2020) adopt deep vision models to detect damaged and undamaged roofs in post-hurricane footage. Cheng et al. (2021b) proposed a stacked CNN architecture for automated PDA in buildings. The model was trained and tested on unmanned aerial vehicle (UAV) footage after the 2019 Hurricane Dorian (Cheng et al., 2021a) to classify the state of damage to six different damage states based on Federal Emergency Management Agency (FEMA) standard guidelines.

Although these developments are significant toward realizing an efficient and scalable automated PDA tool, there are several challenges that remain to be addressed in view of the effective and reliable use of these models in practice. Specifically, the current DL models adopted for PDA have the following two limitations: (1) Most of the models are restricted to qualitative damage identification rather than the refined quantification and classification of damage extent and (2) most of these models are based on individual images (usually satellite view), which may be indicative at times but not necessarily descriptive of a building's exact damage state. It is expected that the classification outcomes corresponding to multiple views of the same building be different because of factors such as directional damage, partial visibility, and image quality (Cheng et al., 2021b). Therefore, more comprehensive visual data in the form of multiple ground and aerial views of the damaged buildings can potentially lead to improved quantification of the damage level. The use of multiple views is also more consistent with human assessment

14678667, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/mice.12890 by Osaka University, Wiley Online Library on [24/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of the extent of the damage to a building. The development of automated DL models to make predictions based on the combined information from different views has been the subject of several recent studies. An early multiview DL model was developed for 3D-shape recognition of distinct objects (Su et al., 2015). Recently, researchers have also used it for finer classification problems like identification of plant species (Seeland & Mäder, 2021), prediction of crop yield from agronomy-related maps (Barbosa et al., 2020), enhanced clinical diagnosis like breast cancer screening (Geras et al., 2017). In the context of visual damage assessment, Presa-Reyes and Chen (2020) architected a dual-stream multiview convolutional neural network (MV-CNN) (two-stream) to classify the damage level using aerial-view images of a residential building taken before and after disasters. While such studies present novel applications and illustrate the use of multiview DL models, they are generally restricted to the classification of visually distinct, nonordinal objects or instances of the same object (Noshadravan et al., 2021). These do not classify the ordinally ranked manifestations of the same object, as is required in case of post-disaster damage classification, which involves ranking the buildings based on their damage level as per a predefined ordinal scale. As such, there is a need for an automated approach that not only facilitates refined and descriptive damage quantification in post-disaster scenarios but is also reliable, robust, and accurate in its prediction.

To address these challenges, this paper proposes constructing a spatially aware MV-CNN architecture that combines the information from different views of a post-disaster damaged building (aerial view, multiple street views), resulting in 3-D aggregation of the 2-D damage features from each view. This study presents significant advancement over the existing applications of DL for PDA, which are ubiquitously based on a single perspective imagery and hence lack the ability to assemble different channels of information for reliable estimation of damage condition. Also, different from the existing multiview DL approaches, this study accounts for the ordinality of damage classes by using a suitable loss function instead of traditionally used cross-entropy loss. Moreover, the complexity of the post-disaster scenes seriously deteriorates the AI performance by causing many false-positive and false-negative predictions. This study also attempts to overcome this problem by implementing a dual-step process to remove unnecessary information from the images before serving them as inputs of the AI model. The applicability and advantages of the proposed approach are presented and illustrated using a real-word post-hurricane damage data, consisting of the expert-evaluated building damage information along with the multiple visual imagery of the assessed buildings. This multiview model is evaluated and compared with the damage-level predictions based on the traditional single-view model using the aerial imagery data for training. The study also compares the model-predicted damage assessment with the crowd-based human assessment and makes recommendations about the future course of action toward developing a powerful human–AI partnered solutions for post-disaster damage assessment.

2 | RESEARCH SCOPE AND CORE CONTRIBUTIONS

This paper presents a step forward in AI-assisted disaster damage assessment by proposing a novel multiclass multiview damage classification model based on CNN. The study is based on an in-house annotated imagery database of multiple views of hurricane-damaged buildings, collected after Hurricane Harvey, in accordance with the standard damage scales used by the U.S. FEMA. These data are instrumental in training and benchmarking AI models for automated disaster damage assessment and can benefit other researchers in the natural hazards engineering community. In its core, the primary contribution of this paper is the development of the multiview, multiclass DL model for estimation of damage levels in disaster-affected buildings. The idea is to mimic the actual expert-conducted damage assessment wherein they have access to all the views of the damage building and hence base their judgment on the more comprehensive visual building information. In addition to this, understanding the data need and VOI is critical in enhancing model accuracy and interpretability. Thus, this paper also investigates the extent by which the prediction confidence is influenced by the combination of views used in building the model. The VOI analysis is performed by considering different scenarios with varying input data granularity (number and type of views). This can help to understand the performance of the model under different scenarios and have an optimal design of multiview models in terms of data need and computational expense. The model predictions are also compared with the human assessments to evaluate the scope for further improvements in enhancing the credibility of such automated models.

In light of the above objectives and contributions, this study leverages the visual damage data collected after a disaster event and adopts a standard damage scale used in practice by engineers and damage inspectors to simulate a realistic scenario. These damage scales are generally proposed by FEMA whose response system is regarded as the national standard for identification, evaluation, and communication of the structural condition of the disaster-damaged buildings (Peña-Mora et al., 2008). FEMA in

14678667, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/mice.12890 by Osaka University, Wiley Online Library on [24/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

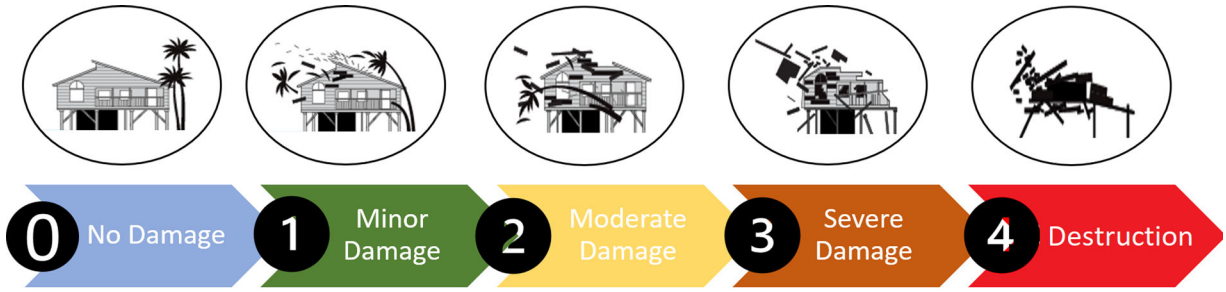

FIGURE 1 HAZUS-MH hurricane damage scale for residential buildings

Based on FEMA (2012)

collaboration with the National Institute of Building Sciences (NIBS) has developed different loss or damage evaluation models for quantifying the impact of different hazards on different types of buildings. These models, called HAZUS models provide practitioners and damage analysts guidelines to quantify the damage in built infrastructure into well-defined damage scales. This study intends to leverage one such descriptive damage classification scale for a specific building class and automate the process of damage classification using the proposed DL model.

The methodology and the modeling approach proposed in the paper is general and applicable to other kinds of natural disasters affecting the built environment. Although this can even be extended to different classes of built infrastructure in a community, its applicability and merit are demonstrated in this study by means of the post-hurricane visual damage data on residential buildings. For the model training and calibration, the expert assigned ground truth based on HUZUS-MH hurricane model classification for residential buildings is adopted. This scale presents five-category damage classification, based on level or damage and failure conditions. For example, if the roof or wall structure of a building is completely destroyed, the building is classified to be in Damage State-4 (DS-4). On the other hand, if the roof and wall structure are intact but the roof cover reflects greater than 50% damage, the building may be classified as DS-3. Similar metrics are defined for other damage states as well, in terms of damage to major components of a residential building viz-à-viz roof cover, windows and doors, roof deck, roof structure, and wall structure. Although there are several more refined and improved damage classification scales presented in the literature, the purpose of this study is limited to demonstrate the feasibility and advantages of the proposed methodology and hence the classification is based on FEMA guidelines for quantifying damage scale for residential buildings. Figure 1 shows a schematic representation of damage scale adopted in the present study, defined in terms of five levels of damage, ranging from 0 to 4 in the order of increasing severity.

3 | PROPOSED METHODOLOGY

Contrary to several visual recognition tasks involving detection of everyday objects in natural settings, identification and classification of post-disaster damage in buildings are highly complex tasks. This is mainly because (1) different damage states are not uniquely defined and hence inherently difficult to differentiate and (2) each building is more or less visually different than the other and the damage description in different buildings may not be unique. This can potentially result in a large number of incorrect predictions, negatively impacting the model performance. To minimize the erroneous detections, a dual-stage strategy is adopted in this work. The adoption of sequential configuration of models has been used as an effective strategy for recognition tasks in complex scenes (Narazaki et al., 2020). Specifically, prior to use of the proposed damage classification model (detailed in the next section), building localization is performed to differentiate "buildings" from "no building" objects. The idea is to get rid of all the background information in the image that is unnecessary for damage classification. Background subtraction will result in the segmented image of only the building, which can then be fed into the proposed classification model for better predictions. Further details about these models are discussed in subsequent sections.

3.1 | Background subtraction

It is worth noting that, for both ground and aerial images, complex and messy post-disaster scenes can significantly deteriorate the model performance. As such, prior to the classification of the damaged buildings into different damage levels, segmentation or localization of buildings in each image, irrespective of the level of damage incurred by the building, is performed. Each pixel belonging to the building in the images is annotated as a "building object" by masking the corresponding pixels manually. The idea is to recognize the pixels belonging to buildings in an image and filter out the visual information unnecessary for

14678667, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/mice.12890 by Osaka University, Wiley Online Library on [24/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

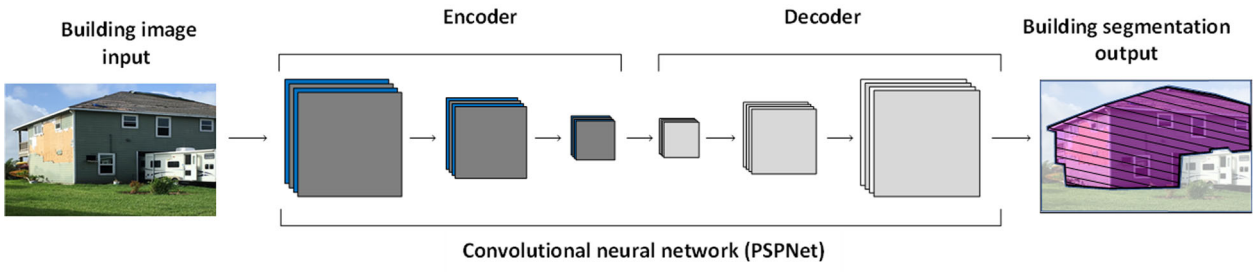

FIGURE 2 An overview of PSPNet architecture adopted for background subtraction

determining the damage state of the building like trees, sky, and roads. The output of this localization model is used to generate segmented regions corresponding to buildings only and the same is fed into the subsequent damage classification model in the second stage of the stacked PDA model. This task is performed using a semantic segmentation model architecture called PSPNet as shown in Figure 2 (Zhao et al., 2017). It is worth noting that semantic segmentation was used given the characteristics of the available data. In case each image consists of more than one building, instance segmentation might be a better option instead. PSPNet, a semantic segmentation model, is pretrained on the ImageNet data set (Russakovsky et al., 2015) and retrained on the training data compiled for this work. Due to the relatively small data size, real-time data augmentation and dropout techniques are implemented to prevent overfitting. In particular, for the data augmentation, we conduct various operations including horizontal and vertical flips, affine transforms, perspective transforms, color manipulations, image blurring and sharpening, and random crops. Moreover, the categorical focal loss is used as the loss function. The training and validation losses are monitored for 50 epochs. The termination criterion is activated when the validation loss increases, which indicates the sign of overfitting. The result shows that, with a 0.5 threshold of intersection over union (IoU) score, the model's average precision (AP) achieves 91.98%.

3.2 | Multiview classification model

The designed architecture of the multiview DL model consists of two network components: a multibranch feature extraction network and a damage assessment network, a fully-connected network as shown in Figure 3. The multibranch network contains multiple parallel input branches using convolutional layers. Recently, tons of researches have shown that the convolutional layers are powerful to extract image features effectively and finally learn to classify the images (Chollet, 2021). Thus, the objective of these branches is to independently extract the high-level (e.g., colors, outlines) as well as low-level (e.g., patterns, textures) features from multiple input images. Next, in the second component, a concatenation layer is designed that fuses a list of independently extracted information from numerous CNN branches into one layer. After this concatenation process, the concatenation layer is connected with a fully connected, multilayer perceptron with two hidden layers, named damage assessment network. This network focuses on predicting the damage state given the fused features information extracted from multiple images using the previous multibranch CNN. Finally, the output of the damage assessment network are softmax values of the five damage-level classes (0–4 levels), shown in Figure 1, where the damage states are based on the specifications of the residential damage scale given in the HAZUS-MH (Hazard United States - Multi Hazard) hurricane model (FEMA, 2012).

With the design ideas of the multiview classification model, let us introduce the model in the mathematical perspective. In the typical classification task with single image input, the damage classification model aims to infer which of the damage state the image input $x_j \in X$ belongs to, where the subscript j denotes index of the building under consideration, and $X = \{x_j\}_{j=1}^{N}$ denotes the set of all images in the data set. The damage classification model is thus an approximated function $f : x_j \rightarrow \{0, .., k-1\}$ that maps the input building image x_j to its respective damage level $y_j \in \{0, .., k-1\}$. This concept is extended to the damage assessment model based on the proposed composite network with multiple branches, named multiview damage CNN (MVD-CNN). The adopted MVD-CNN model is built on feature extraction from 2-D images using standard pretrained CNN architectures. The extracted feature maps from different building views are combined into a common 3-D feature space using a multiview pooling layer. The combined feature space is then passed into a series of fully connected layers for the final damage-level prediction (Figure 3). Mathematically, the underlying fact is that all the building images are organized in terms of a collection of images, $X_j = \{x_j^{(v)}\}_{v=1}^{n_v}$, where each collection corresponds to each individual building and consists

14678667, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/mice.12890 by Osaka University, Wiley Online Library on [24/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

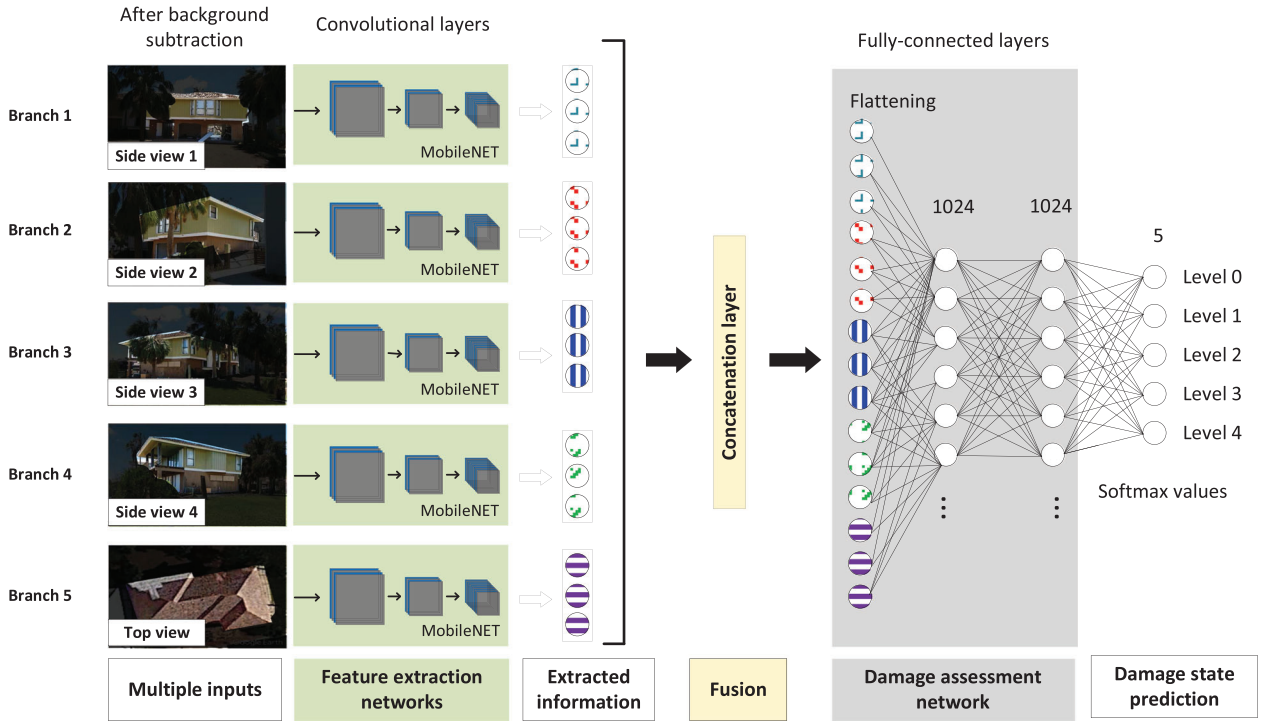

FIGURE 3 Proposed MVD-CNN architecture for multi-view damage classification

FIGURE 4 A sample of different views considered for a representative building

of n_V distinct views of the same building. Hence, the classification task at hand may now be reformulated as:

$$f : X_j \rightarrow \{0, .., k-1\} \mid X_j = \left\{ x_j^{(1)}, x_j^{(2)}, .., x_j^{(n_V)} \right\} \quad (1)$$

where $\{0, .., k-1\}$ denotes the k possible damage states a building can be classified into, $x_j^{(v)}$ denotes one image corresponding to a specific view of the building j in question. Each image $x_j^{(v)}$ is processed by a different branch of the similar CNN for feature extraction. The feature maps from different independent CNN branches are then stacked and the fused composite feature map containing information from all the views is then processed together to predict the damage class.

To train the MVD-CNN with the available data, transfer learning technique is adopted. MobileNet (Howard et al., 2017) is chosen as the base model for the feature extraction for each of the considered views separately. Pretrained on more than 1,000,000 images with diverse feature representations from the ImageNet data set (Deng et al., 2009), MobileNet has a light but powerful architecture and can be tuned efficiently for the available visual data.

4 | CASE STUDY

4.1 | Data description

The presented study uses the post-hurricane damage data collected after Hurricane Harvey, a category 4 hurricane made its landfall on August 25, 2017, causing a lot of destruction in Texas, Louisiana, and neighboring regions. The post-hurricane data have been collected by expert post-disaster reconnaissance teams from StEER

14678667, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/mice.12890 by Osaka University, Wiley Online Library on [24/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(Structural Extreme Event Reconnaissance) (StEER, 2022) network. The data consist of multiple geotagged photographs of the buildings inspected by the StEER experts and inspection teams, along with their detailed observations and overall rating based on the on-site assessment and evaluation of the damaged buildings. The data can be accessed from the Natural Hazards Engineering Research Infrastructure (NHERI) cyberinfrastructure Reconnaissance Portal (Roueche et al., 2018). The visual data collated for the current research include the geo-tagged ground images taken from different sides, as well as the aerial and satellite images of the affected buildings.

To make the structure of the data points uniform, five different images of each building considered for the study are selected from the database. The selected images consist of four different post-disaster ground images and a satellite image of the building. This selection is performed manually, given that the data used were not primarily collected for the intended purpose and hence need to be rearranged before they can be dependably used in the present study. Figure 4 shows the selected ground and satellite images for a representative building used in the study. Images containing unnecessary information or taken from inside of the building are filtered out and not considered in the study. While this may seem to be a time-consuming and laborious task, with more organized and planned data acquisition strategies, much of this effort can be significantly reduced. Following this, annotation of images is done using Labelbox (Labelbox, 2022), an online annotation and training data management platform. In total, images corresponding to 500 buildings (2500 images) were annotated by drawing masks around the building boundaries and were assigned labels representing their damage state. The damage labels (0 to 5) were based on the expert assessment reported in the original database. In addition to this, a testing data set of 80 buildings not included in the training data is set aside to evaluate the performance of the model. Similar to the training data, the test data consist of five different views of each building, including four ground views and one satellite view. This adds up to a total of 400 images in the test data.

4.2 | Model implementation

The implemented MVD-CNN model consists of the composite convolution base (MobileNet) with pretrained weights, and the newly added fusion layer and the fully connected network for final damage state prediction. Corresponding to each of the five views considered in the study, feature extraction is performed using transfer learning on a MobileNet CNN model. Following this, the convolutional feature maps obtained from different CNNs corresponding to each view are stacked and subsequently processed together. It should be noted that the fusion strategy adopted in this study is late fusion. The details about this fusion technique, referred to as late fusion can be found in Seeland and Mäder (2021). The adopted fusion technique maintains the correspondence between the feature maps of different views. The entire model, consisting of both the convolution base and fully connected network, is trained end-to-end as a single model. The newly added fully connected layers are initialized with randomly assigned weights, which are updated by retraining the model on the training data. The following methodology is adopted for training the model:

(1) First, the weights in the pretrained convolutional layer are frozen and the model is trained using an epoch size of 50 and learning rate = 0.0001. This enables the model to adapt the pretrained features to the new data.
(2) The second step involves unfreezing the weights in the convolution base and the entire model is retrained at a smaller learning rate of 0.00001. This is called fine-tuning and allows the incremental improvements and fine modifications of the pretrained weights to learn more effective features.

In addition to this, data augmentation technique is used to enhance the training data set. Different augmentation operations applied to the training data set include horizontal flip, affine transforms, perspective transforms, brightness/contrast/colors manipulations, image-blurring, sharpening, and Gaussian noise. This exposes the model to a more diverse training data, resulting in a more generalizable output. Since the present application deals with ordinal and inherently related classes, the most commonly used cross-entropy loss function is not an obvious choice. For example, damage states 0 and 4 are significantly distinct as compared to damage states 0 and 1. As such wrongly classifying a DS-1 building into DS-0 is not the same as classifying it as DS-4. In such cases, ranking or interclass relationships is not captured by cross-entropy loss function (Cheng et al., 2021b) and the training effectiveness using cross-entropy as loss function decreases significantly (Hou et al., 2016). Instead, a more suitable loss function for such problems, based on the earth mover's distance (EMD), is adopted. The EMD represents the cost of transforming one probability distribution to another and can be interpreted as a measure of distance or dissimilarity between the two distributions (Rubner et al., 2000). The computation of EMD is based on the solution to the famous Monge–Kantorovich transportation problem for which efficient algorithms are available (Rachev, 1985). However, for ordered-class classification problems, like the prediction of damage levels in the presented

14678667, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/mice.12890 by Osaka University, Wiley Online Library on [24/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

application, Hou et al. (2016) showed the EMD can be calculated in a closed form, which is directly applicable for ordinal ordered-class classification problems. They also showed that the square of earth moving distance (EMD^2) shows faster convergence as well as efficient optimization as compared with using EMD as the loss function. As such the simplified form of EMD^2, which is based on the assumption of interclass ordinality, is used as the loss function in the present study. Mathematically EMD^2 can be represented as the following closed form for ordered-class classification problems:

$$EMD^2(\hat{y}, y, j) = \sum_{k=0}^{K-1} [P(\hat{y}_j \leq k) - P(y_j \leq k)]^2 \quad (2)$$

where $\hat{y}_j$ denotes the softmax prediction vector obtained from the model, y_j represents the ground truth vector for each building j. $P(\hat{y}_j \leq k)$ and $P(y_j \leq k)$ denote the values of cumulative distribution function (CDF) corresponding to kth damage state for each building j. K refers to the total number of classes and ranges from 0 to 4 in this study. EMD^2 has been shown to be a suitable loss function for classification of highly related ordinal classes and exhibits smooth convergence (Hou et al., 2016) and hence is used in this work instead of cross-entropy loss function.

5 | RESULTS AND DISCUSSION

5.1 | Model testing and performance evaluation

The performance of the trained model was tested on a suite of 80 unseen buildings in the city of Rockport in Aransas County of Texas, which were damaged in the aftermath of Hurricane Harvey. The predictions from the model were compared with the actual door-to-door damage assessment performed by experts from StEER, which is regarded as the ground truth in this study. The test data consist of 13%, 12%, 22%, 27%, and 6% buildings in DS-4, DS-3, DS-2, DS-2, and DS-0, respectively. The distribution of actual damage levels in the testing data set is shown in Figure 5a. Majority of the considered buildings fall in DS-1 and DS-2, while very few buildings are completely undamaged by the hurricane (DS-0). This implies that the hurricane had significant impact on the area under consideration, which is expected, given that the Hurricane Harvey made its landfall just four miles east of Rockport, Texas, on San Jose island. The results of the damage prediction model are plotted in Figure 5b. The model predictions can be observed to be in a good agreement with the ground truth plotted in Figure 5a. To better visualize the similarity between the ground truth and the predicted damage states, their difference for each building is shown in Figure 5c. It can be seen that a significant majority of the buildings' predicted damage states match the expert-defined ground truth either exactly or differ by 1. For the quantitative comparison, the prediction accuracy was estimated as the percentage of the total number of buildings predicted correctly by the trained model as compared with the expert-assigned ground truth. The trained MVD-CNN model based on five different views including ground and aerial imagery resulted in the testing accuracy of 65%, when considering five-level classification of damage states. The level-wise prediction results are summarized by means of a confusion matrix, wherein each row represents the instances in actual damage level and each column represents the instances in the predicted damage level. The confusion matrix showing the comparison of the damage levels predicted by the MVD-CNN model as compared with the ground truth is shown in Figure 6a. It is observed that the majority of the buildings were classified accurately, despite limited number of training data as well as simplistic network architecture adopted in the study. The low accuracy in the identifying of undamaged state (DS-0) can be attributed to the fact that the both training and testing data contains only a few instances of DS-0 class. The low representation of undamaged buildings in a building stock affected by a high-intensity natural disaster event is expected and the same is reflected in the available data. Given that the task at hand is subjective to some extent and that there is no distinct demarcation between the adjoining damage classes, the achieved accuracy is deemed reasonably acceptable. In addition to this, receiver operating characteristic (ROC) curve is also plotted and shown in Figure 6b. ROC curve is a probability curve that plots true positive rate (TPR) versus false positive rate (FPR) at different classification thresholds, and represents the capability of the model to distinguish between different damage classes. Form the confusion matrix as well as the ROC curve, it can be seen that the trained model classifies DS-4 most accurately. Area under the curve of the ROC curve (AUC-ROC) for DS-4 is observed to be around 88%. Ideally AUC-ROC should be 1 (100%) and deviations from this ideal value indicate the degree inaccuracy in the model. The DS-0 is classified with the least accuracy, with AUC-ROC value of 72%. In addition to these metrics, other evaluation metrics including rrecision, recall, and F-1 score are also studied and reported for the model. These metrics are expressed in terms of the number of true-positive (TP), false-positive (FP), and false-negative (FN) predictions. Precision indicates the number of predictions matching the ground truth (TP) as a ratio of the total number of predicted positives (FP+TP), while recall is defined as the ratio of correct predictions (TP) over the number of actual positive instances (TP+FN)

14678667, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/mice.12890 by Osaka University, Wiley Online Library on [24/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

FIGURE 5 Map showing the study area and distribution of test buildings' (a) actual damage state, (b) predicted damage states, and (c) deviation of predicted damage states from the ground truth

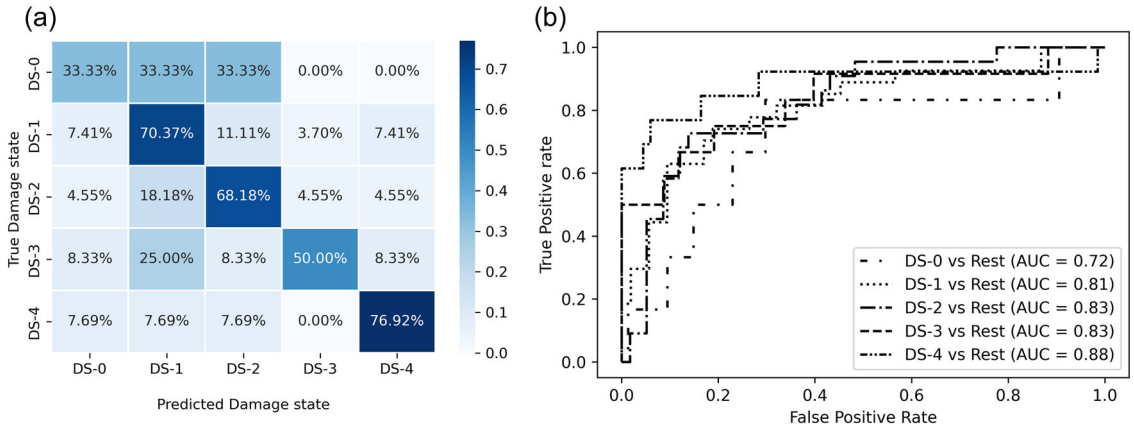

FIGURE 6 Evaluation of MVD-CNN (five-view) model (a) confusion matrix and (b) ROC curve

(Powers, 2020). F-1 score combines precision and recall into a single metric and is defined as the harmonic mean of precision and recall. The weighted average F-1 score of 65% is observed, while the weighted average values for precision and recall are observed to be 66% and 65%, respectively. The weighted AUC-ROC value is calculated to be 82.34%. The weighed average values of these metrics indicate the overall performance of the model, taking into account class imbalance in the testing data. Figure 7 shows the example of the model prediction for two different cases. Figure 7a corresponds to the TP case where the correct damage state (DS-4) of the building is predicted by the model, while in Figure 7b the damage state of a building actually in DS-0 is wrongly predicted as DS-4. Such extreme errors in prediction can be attributed to the noisy image pixels as observed in view 2 and view 3 of the building in Figure 7b.

It is apposite to mention here that prediction performance of the presented damage classification model is a function of the granularity of the damage classification as well. A finer classification scale is expected to be a more demanding task and intuitively result in lower accuracy as compared to the coarser damage scale. Figure 8 shows the modified confusion matrix observed in the case where a coarser three-class damage classification is adopted. Here, DS-0 and DS-1 are combined to represent "No or minor damage" class, DS-2 and DS-3 together represent "Moderate damage" class, and DS-4 represents "Complete damage" class. This coarser damage classification scale yields a significantly higher prediction accuracy of around 73%. Also, considering DS ±1 class deviation from the ground truth yields a significantly higher prediction accuracy of 81.25%. This is encouraging since predicting a DS-3 or DS-1 for a building, which is actually in DS-2, is a better

14678667, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/mice.12890 by Osaka University, Wiley Online Library on [24/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

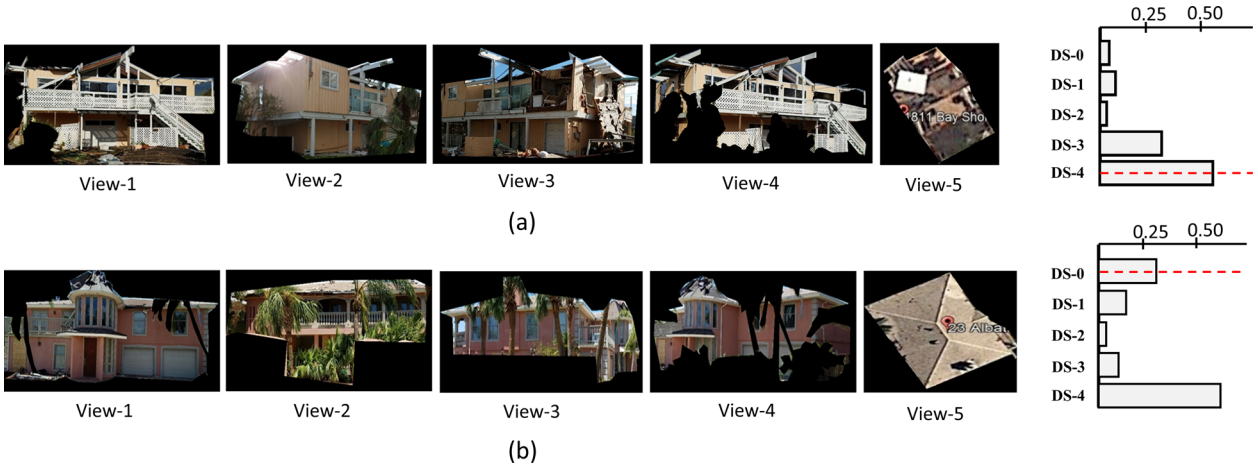

FIGURE 7 Example of samples (a) correctly classified by the model and (b) incorrectly classified by the model

Note: Dashed red line shows the actual damage state of the building

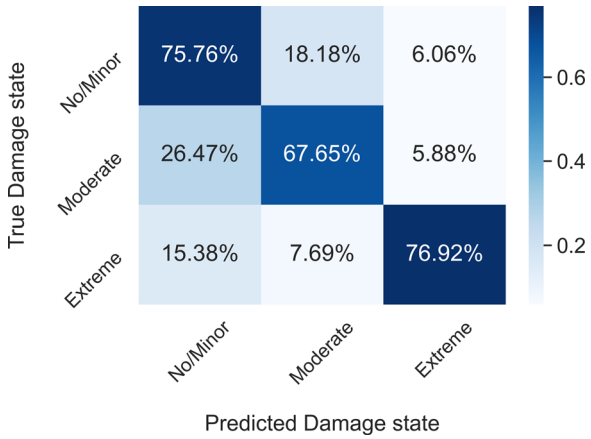

FIGURE 8 Confusion matrix for three-class MVD-CNN model

prediction than predicting the damage as DS-4. Given this ordinal nature of the damage classification scale, the presented model performance is significantly high.

5.2 | VOI

It is expected that including more channels of information in the model, here in terms of the fusion of multiple views, enhances the fidelity and prediction accuracy. However, this adds to the complexity of the model, which comes at the cost of higher computational expense and larger data requirements. For example, the computational demand of the presented five-view CNN model, is about five times that of the single-view prediction model. Another challenge is the availability of such multiview information and the limitation in resources to collect additional information. This raises the question of to what extent adding more views improves the automated damage assessment and that what is the minimum level of information to achieve a desired level of accuracy. The gain in accuracy with the additional visual data should be significantly enough to offset the cost of obtaining and fusing more data into the model. To that end, the VOI is analyzed by exploring different scenarios where the MVD-CNN model is trained and tested based on different views.

5.2.1 | Comparison with single-view CNN

First, the commonly used single-view prediction model was trained and tested using the same data, and the prediction performance was compared with the presented five-view CNN model as the benchmark. In the case of single-view CNN, two different models, one based on the aerial imagery and the other based on ground imagery, were trained and tested to evaluate the damage levels of the affected buildings in the testing data. These models were based on the same model architecture as each channel of the proposed MVD-CNN model. For consistency, the same set of buildings were used for training and testing the model in all cases. The selection of images for the single-view ground model was performed manually by selecting the view most representative of the damage in the building. In most of the cases, this corresponded to the front facade of the building, which was hence chosen to test the single-view ground model. The prediction accuracy of the single-view CNN model based on aerial and ground imagery were observed to be 55% and 53.25%, respectively. This is significantly less than the prediction accuracy observed with the proposed MVD-CNN model (65%), which confirms that damage prediction based on multiple views of the damaged buildings results in a more reliable AI model and higher prediction accuracy. The

14678667, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/mice.12890 by Osaka University, Wiley Online Library on [24/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

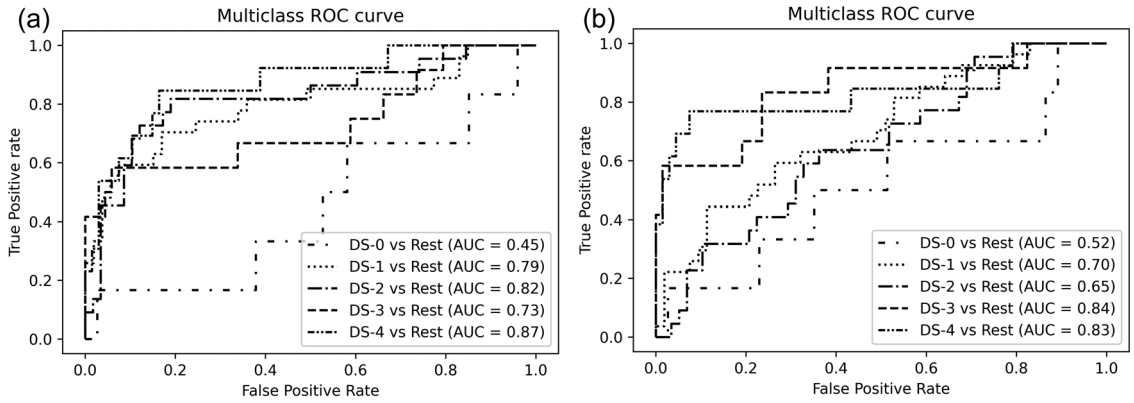

FIGURE 9 ROC Curve for single-view CNN model based on (a) aerial images and (b) ground images

relative comparison of the evaluation metrics of the two single-view models reveals that aerial imagery is better suited for training the damage prediction models and results in higher prediction accuracy as compared with a singe-view ground imagery. Here, it is pertinent to mention that the test area considered in this study consists predominantly wind-damaged buildings. With the recorded wind speed of over 200 km/h, the observed damage in the region is almost exclusively related to high winds (Fovenyessy & Patterson, 2019). Since unlike other hurricane-induced hazards, wind-related damage is distinctly reflected in the roof system of buildings, the aerial imagery provides a better description of the damage level of the affected building as compared with the limited ground views, which may show differential signatures of incurred damage. This explains why the aerial view–based model leads to higher accuracy. However, the difference in accuracy is expected to be more if the test area would have been much inland with wind as the only source of damage. Also, in case the damage is predominantly flood borne, or surge induced, the opposite trend may be expected as well. Figure 9 shows the ROC curves for the aerial as well as ground-view model based on the predictions on the testing data. As is evident from the AUC corresponding to each damage state, the aerial imagery specifically leads to better prediction of higher damage states as compared with lower damage states. This is expected given that the severe damage is distinctly manifested in the building roof system, specifically when the damage is predominantly wind borne. A completely destroyed building is easier to be identified from the visual condition of the roof system as compared with less damaged buildings with less distinct visual features.

5.2.2 | Comparison with other scenarios

In order to evaluate the accuracy with respect to the complexity of the desired damage prediction model, different models with varying fidelity were trained and tested on the available data. The level of input information in terms of number and combination of views used to train and calibrate the model was varied and the implication of that on model performance accuracy was analyzed. The question is whether there is an optimal scenario where the AI-assisted damage assessment is sufficiently accurate with minimum number of views and thus a less complex model. To answer this question, different combinations of two-, three-, and four-view models were developed, trained, and tested. The dual-view model with only ground imagery information resulted in the prediction accuracy of 60%. The three- and four-view models based only on ground-level imagery (no top aerial view) resulted in the testing accuracy of 58.75% each. Table 1 shows the comparison of performance metrics for different ground-based models. The weighted average ROC score for multiple-view models based only on ground imagery, was observed to be close to 80% for all cases. This indicates that there is no significant improvement offered by one of these models over the other. However, compared to the single-view model based on ground-based image data, the multiview approach presents a reasonable improvement. Table 2 shows the performance metrics for all the considered cases. Figure 10 presents the visual comparison of these performance evaluation metrics for different models. It can be clearly observed that the use of at least one aerial view and ground

TABLE 1 Performance metrics for different ground-based models

Model	Accuracy	Avg. F-1	Avg. ROC
One-view	53.75	50	72.7
Two-view	60	60.29	80.8
Three-view	58.75	58.61	80.27
Four-view	58.75	58.7	80.3

Note: All values are expressed as percentage (%).

14678667, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/mice.12890 by Osaka University, Wiley Online Library on [24/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Performance evaluation of different scenarios

Metric	One-view (ground)	One-view (aerial)	Multiview (ground)[a]	Hybrid (two-view)	Hybrid (five-view)
Accuracy	53.75	55	60	63.75	65
Avg. precision	49	55	61	67	66
Avg. recall	54	55	60	64	65
Avg. F-1 score	50	54	60.29	63	65
Avg. AUC-ROC	72.75	77.9	80.8	83.17	82.35

Note: All values are expressed as percentage (%).
[a]Corresponding to the case with maximum values.

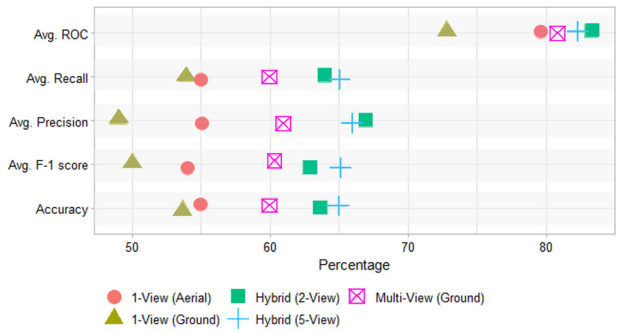

FIGURE 10 Comparison of evaluation metrics for different models

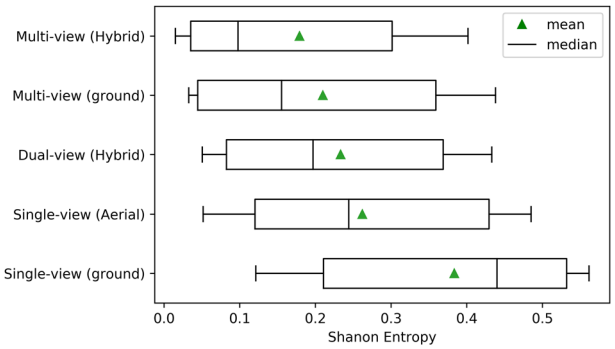

FIGURE 11 Barplot showing the distribution of entropy for different models

view for the damage prediction yields more accurate results. On the same testing data set, the hybrid model calibrated with dual-view and five-view information resulted in the testing accuracy of 63.75% and 65%, respectively. The average weighted ROC score obtained for these two models was around 83%, which is significantly greater than 72.75% and 77.9% obtained for both single-view ground and aerial models, respectively. This might be due to the fact that the damage features visible from these two views are considerably distinct and hence together lead to accurate predictions about the damage state. On the other hand, the inclusion of more than one ground image in the model does not result in a better accuracy. This counterintuitive observation can be attributed to the fact that many of the ground-view images used in the current study were actually not representative, being either masked by nonbuilding objects or had a lot of noise. This might have resulted in some reduction in the overall performance of the model when considering only the ground imagery. However, this reduction is deemed negligible and is attributed to the discrepancy in the data quality rather than model effectiveness.

The information content and the degree of uncertainty in the output of different models can be systematically quantified and compared using the entropy measure (Shannon, 1948) defined as follows: The entropy can be mathematically represented as:

$$H(\hat{y}_j) = -\sum_{i=0}^{K-1} P(\hat{y}_j^k) \log P(\hat{y}_j^k) \quad (3)$$

where index k refers to the kth damage class, K denotes the total number of damage states, $\hat{y}_j$ denotes the softmax prediction vector for building j obtained from the model, $P(\hat{y}_j^k)$ denotes the probability of building j to be in damage state k. Figure 11 shows the comparison of the entropy observed upon testing different models on the testing data. It is observed that the degree of uncertainty, which is proportional to the value of entropy, is higher for single-view models as compared with the multiview models. This is intuitive, given the fact that a more complex and better informed model is expected to output more certain predictions and compared with simplistic models calibrated using limited data channels. The multiview (hybrid) model, which is based on five different views of the building, results in the least entropy. The value of average entropy for the multiview model based on ground imagery is reported as the average of the values obtained for different combinations of such views. The single-view model based on ground images only results in the maximum entropy, hence the most uncertain prediction. This is expected given the information used in the model is very limited and potentially less descriptive.

14678667, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/mice.12890 by Osaka University, Wiley Online Library on [24/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

5.3 | Comparison with crowdsourced participatory damage assessment

The presented approach to leverage multiple views to enhance the performance of AI models comes from the intuitive understanding of how humans process imagery data to infer disaster damage assessment. It is thus reasonable to explore the performance of AI-assisted damage assessment as compared to nonexpert human assessment. Such comparison and insight are particularly important as a step toward a more effective and interpretable human–AI partnership for automated and rapid disaster assessment. The crowdsourcing and participatory sensing can be leveraged to reduce dependence on the domain experts for large-scale post-disaster PDA. The question is to what extent the prediction of AI model and nonexpert are close and consistent.

For the purpose of comparative study, the same testing data that were employed in the MVD-CNN model were utilized for conducting participatory damage assessment using online crowdsourcing. Anonymous nonexpert citizen participants were recruited using Amazon Mechanical Turk (M-Turk) crowdsourcing platform. A survey was designed where first the participants were provided with some basic instructions about the assignment of damage states to the buildings followed by some observational questions about the damage in the buildings. They were also tasked with assigning the damage class to each building based on the imagery inputs and the instructions provided. The response from around 100 anonymous participants with varying educational and professional background was collected and used to make the desired inferences. The crowdsourced assessment was based on multiple views of the damaged buildings, similar to the approach used for training the MVD-CNN model. With access to all the available views of the damaged buildings, the participants are provided with the same information content as the AI model, hence expected to result in better and more accurate damage evaluation.

The crowdsourced damage assessment adopted in this study is based on the methodology presented in Khajwal and Noshadravan (2021). The responses received from the participants are inferred using two approaches. First the majority vote (MV) heuristic is used for the inference of the underlying damage state of the buildings, given the noisy response obtained from the participants. This simplistic approach assigns the damage level to a building based on the maximum number of votes. For example, if majority of the participants attribute DS-3 to a building while others attributed DS-2, and some others DS-4, the DS-3 would be chosen as the final damage state of the building in question. Although simple and intuitive, this approach, however, has its own set of limitations. Primarily, it does not account for a different reliability among the participants and hence treats every participant response with equal weight, which may not always be true. In an anonymous crowdsourcing, like the one conducted in this study, and also in a general practical setting of citizen science project, it is very likely that the participants come from various backgrounds. Thus, their level of reliability is not homogeneous. This leads to uncertainty and subjectivity of the data quality. To account for this aspect, a probabilistic model was presented in Khajwal and Noshadravan (2021) to account and reduce the underlying uncertainty stemming from the unreliability of crowd participants. This model is based on maximum a posteriori (MAP) estimation of the underlying damage states from the noisy crowd responses. The model allows more reliable inference of the underlying true building damage levels while accounting for the subjectivity of the individual task as well as the participant reliability, which is not known a priori. In processing and inferring crowdsourcing data in this work, both these approaches are used to evaluate the damage predictions on the testing data and compared with the AI-based assessment of the same buildings. More details about the crowdsourced participatory damage assessment model can be found in Khajwal and Noshadravan (2021). Considering the scope of the paper, here only the outcome of comparison is presented, followed by the discussion and major takeaway from the results.

TABLE 3 Comparison of crowd-based assessment with the MVD-CNN model

Model	Accuracy
Crowdsourcing (MV)	40.0%
Crowdsourcing (MAP)	43.75%
MVD-CNN	65.0%

Abbreviations: MV, majority vote; MAP, maximum a posteriori. model.

The comparison between the accuracy of crowdsourced damage assessment and the MVD-CNN model is shown in Table 3. It can be seen that the MVD-CNN performs much better than the crowd. It is worth noting that the MVD-CNN model is trained using the data that are labeled by experts. On the hand, the human labelers in the crowdsourced damage assessment came from different backgrounds with no or very little domain knowledge and were also given minimal instructions before responding to the survey questionnaire. Although the crowdsourcing (MAP) model presented a low exact prediction accuracy of about 44%, considering ±1 class deviation from the ground truth yields a significantly higher accuracy of 70%. This is greater than even 60% observed in case of the single-view model. As mentioned earlier, the proposed multiview model yielded an accuracy of 81.25%, considering ±1 class deviation. This trend could be attributed to the fact that

14678667, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/mice.12890 by Osaka University, Wiley Online Library on [24/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

anonymous human participants are in a good position to roughly estimate the damage state of a building, rather than exact damage class identification. Another interesting observation from this comparative study reveals that only 27.5% instances were correctly predicted by both models, while 80% instances were predicted correctly by at least one model. The damage state of 36.25% buildings are predicted correctly by MVD-CNN only while 27.5% are predicted correctly by both the models. It can be seen that 52.5% instances correspond to the case where one of the model gives correct prediction while the other model gives incorrect prediction. This observation indicates the potential complementary scope for human and machine-driven models, wherein one can inform the other and the overall result will be enhanced accuracy compared to either of the approaches. Another important observation, validating the same point was the relative performance of the two models for DS-0. While it was observed that the MVD-CNN model did not produce higher accuracy when predicting DS-0, however the crowdsourced model predicted DS-0 with almost 100% accuracy.

5.4 | Potential directions for future research

Although the current study proposes a more robust and reliable methodology for automated damage assessment, like most other single-stream DL models, its output and parameters are deterministic. The inability to quantify uncertainties associated with the practical adoption of such a model, might be a concern when making crucial decisions (Begoli et al., 2019; Sajedi & Liang, 2021a). To that end, a possible direction of future research is to develop and explore probabilistic MV-CNN that enables more rigorous quantification of prediction uncertainty. Many previous studies have illustrated the importance of probabilistic design of neural networks (Ahmadlou & Adeli, 2010; Kendall & Gal, 2017). The concept of having an uncertainty-aware DL model for damage assessment in buildings has also been tested and advocated in some recent research efforts (Cheng et al., 2022; Sajedi & Liang, 2021b). It will be interesting to see how probabilistic settings in the MV-CNN model can lead to more informed and reliable damage assessment outcomes. In addition to this, different computational strategies like ensemble technique (Alam et al., 2020) and dynamic optimization (Rafiei & Adeli, 2017b) can be tested to enhance the performance of the proposed model.

Another potential direction for future research is to develop a more powerful solution for post-disaster damage assessment based on human–AI partnership. Given the high complexity of damage scenes after disasters, humans are usually expected to perform more reliably at identifying the damages of disaster sites. This may be partially because most humans have background knowledge and a fundamental insight into the messy scenes in disaster-affected sites. The current study also showed that while AI offers great promise in applications related to disaster damage assessment, human assessment based on crowdsourcing can prove to be complementary and help in enhancing the predictions from AI models. This calls for an integrated human–AI partnered system as a potential solution for more reliable AI-assisted disaster damage assessment. In such a system, AI processes the vast amount of field or aerial images, and humans (e.g., crowdsourcing) can help identify and improve the quality of the identified unreliable damage assessments. Overall, this system can lead to a better, efficient, and practical automated assessment of building damage states.

6 | CONCLUSION

This research is aimed to enhance AI-assisted post-disaster damage assessment by developing and benchmarking a multiview multimodal deep CNN model. An extensive training data set was compiled from an open-source data repository, consisting of multiple geotagged images of the hurricane-affected buildings along with their on-site evaluation and observations recorded by a team of experts. The annotated images depicting the damage in the considered buildings were used to train and test the proposed DL model. This model is an improvement over the existing practices in the use of AI for rapid and automated disaster damage assessment because it predicts the damage states based on the fusion of multiple images corresponding to different views of the damaged building. The framework involves a dual-stage model consisting of a localization model to identify and localize the building objects and a subsequent multiview classification model that uses the output of the localization model to predict the damages state based on FEMA standard guidelines. As a benchmark, a refined multiview CNN model based on five different views of the building was developed, including five side views captured in ground images and a top view captured by satellite imagery. Overall testing accuracy of 65% is observed when considering refined classification into five different damage states. A coarser three-level classification (minor, moderate, and extreme damage) resulted in a testing accuracy of 73%. Value for information in terms of the number of views as imagery input was studied by testing different models with different combinations of the building views. It was observed that the hybrid models involving at least one satellite view and single ground view perform significantly better than other configurations. The

14678667, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/mice.12890 by Osaka University, Wiley Online Library on [24/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

dual-view model (hybrid) showed a testing accuracy of 63.75%. The performance of the presented model was also compared with the results of a crowdsourcing experiment performed on the same test data. The AI model proposed in this paper is a step forward in improving existing DL models adopted for automated damage impact assessment. More quantitative and interpretatble predictions about the severity of the damages can in turn enable better informed modeling of regional damage and subsequent decision making. Moreover, the modeling strategy adopted in this paper is general in its applications and can be adopted to other kinds of natural disasters as well.

ACKNOWLEDGMENTS

The authors would like to thank Texas A&M University's High Performance Research Computing (HPRC) for providing necessary computing infrastructure for model training. Any opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily represent the views of the HPRC.

REFERENCES

Ahmadlou, M., & Adeli, H. (2010). Enhanced probabilistic neural network with local decision circles: A robust classifier. *Integrated Computer-Aided Engineering*, *17*(3), 197–210.

Alam, K. M., Siddique, N., & Adeli, H. (2020). A dynamic ensemble learning algorithm for neural networks. *Neural Computing and Applications*, *32*(12), 8675–8690.

Amezquita-Sanchez, J. P., & Adeli, H. (2015). Synchrosqueezed wavelet transform-fractality model for locating, detecting, and quantifying damage in smart highrise building structures. *Smart Materials and Structures*, *24*(6), 065034.

Barbosa, A., Marinho, T., Martin, N., & Hovakimyan, N. (2020). Multi-stream cnn for spatial resource allocation: A crop management application. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 58–59). https://openaccess.thecvf.com/content_CVPRW_2020/html/w5/Barbosa_Multi-Stream_CNN_for_Spatial_Resource_Allocation_A_Crop_Management_Application_CVPRW_2020_paper.html; https://www.thecvf.com/?page_id=31

Begoli, E., Bhattacharya, T., & Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, *1*(1), 20–23.

Cha, Y. J., Choi, W., & Büyüköztürk, O. (2017). Deep learning-based crack damage detection using convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering*, *32*(5), 361–378.

Chen, Z., Wagner, M., Das, J., Doe, R. K., & Cerveny, R. S. (2021). Data-driven approaches for tornado damage estimation with unpiloted aerial systems. *Remote Sensing*, *13*(9), 1669.

Cheng, C. S., Behzadan, A., & Noshadravan, A. (2021a). DoriaNET: A visual dataset from Hurricane Dorian for post-disaster building damage assessment. DesignSafe-CI. https://doi.org/10.17603/ds2-gqvg-qx37

Cheng, C. S., Behzadan, A. H., & Noshadravan, A. (2021b). Deep learning for post-hurricane aerial damage assessment of buildings. *Computer-Aided Civil and Infrastructure Engineering*, *36*(6), 695–710.

Cheng, C. S., Behzadan, A. H., & Noshadravan, A. (2022). Uncertainty-aware convolutional neural network for explainable artificial intelligence-assisted disaster damage assessment. *Structural Control and Health Monitoring*, e3019. https://doi.org/10.1002/stc.3019

Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE. https://ieeexplore.ieee.org/abstract/document/5206848?casa_token=5Re7v9fFy8AAAAAA:4NJm8Qx-5jzcYh1esEP1scGndW5DneKsCO8Fd6h1KYHF6Wgn4juaJuQozsgx2ZVsRLU4jmpZxg

FEMA (2012). *HAZUS-MH 2.1 Hurricane model technical manual*.

Fovenyessy, S., & Patterson, S. F. (2019). Wind or water? Hurricane Harvey's most destructive force. https://www.earthmagazine.org/article/wind-or-water-hurricane-harveys-most-destructive-force/

Gao, Y., & Mosalam, K. M. (2018). Deep transfer learning for image-based structural damage recognition. *Computer-Aided Civil and Infrastructure Engineering*, *33*(9), 748–768.

Gao, Y., Zhai, P., & Mosalam, K. M. (2021). Balanced semisupervised generative adversarial network for damage assessment from low-data imbalanced-class regime. *Computer-Aided Civil and Infrastructure Engineering*, *36*(9), 1094–1113.

Geras, K. J., Wolfson, S., Shen, Y., Wu, N., Kim, S., Kim, E., Heacock, L., Parikh, U., Moy, L., & Cho, K. (2017). High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv preprint arXiv:1703.07047*.

Gharaibeh, N., Oti, I., Meyer, M., Hendricks, M., & Van Zandt, S. (2021). Potential of citizen science for enhancing infrastructure monitoring data and decision-support models for local communities. *Risk Analysis*, *41*(7), 1104–1110.

Gupta, R., Hosfelt, R., Sajeev, S., Patel, N., Goodman, B., Doshi, J., Heim, E., Choset, H., & Gaston, M. (2019). xBD: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*.

Hou, L., Yu, C. P., & Samaras, D. (2016). Squared earth mover's distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916*.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Institute for Economics & Peace (2020). Ecological threat register 2020: Understanding ecological threats, resilience and peace, Sydney, September 2020. Available from: http://visionhumanity.org/reports (accessed 23 July 2022)

Kang, D., & Cha, Y. J. (2018). Autonomous UAVs for structural health monitoring using deep learning and an ultrasonic beacon system with geo-tagging. *Computer-Aided Civil and Infrastructure Engineering*, *33*(10), 885–902.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems, 30*.

Khajwal, A. B., & Noshadravan, A. (2021). An uncertainty-aware framework for reliable disaster damage assessment via

14678667, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/mice.12890 by Osaka University, Wiley Online Library on [24/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

crowdsourcing. *International Journal of Disaster Risk Reduction*, *55*, 102110.

Labelbox (2022). *Labelbox: The leading training data solution*.

Li, L., Bensi, M., Cui, Q., Baecher, G. B., & Huang, Y. (2021). Social media crowdsourcing for rapid damage assessment following a sudden-onset natural hazard event. *International Journal of Information Management*, *60*, 102378.

Liang, X. (2019). Image-based post-disaster inspection of reinforced concrete bridge systems using deep learning with Bayesian optimization. *Computer-Aided Civil and Infrastructure Engineering*, *34*(5), 415–430.

Miura, H., Aridome, T., & Matsuoka, M. (2020). Deep learning-based identification of collapsed, non-collapsed and blue tarp-covered buildings from post-disaster aerial images. *Remote Sensing*, *12*(12), 1924.

Narazaki, Y., Hoskere, V., Hoang, T. A., Fujino, Y., Sakurai, A., & Spencer Jr, B. F. (2020). Vision-based automated bridge component recognition with high-level scene consistency. *Computer-Aided Civil and Infrastructure Engineering*, *35*(5), 465–482.

Nex, F., Duarte, D., Tonolo, F. G., & Kerle, N. (2019). Structural building damage detection with deep learning: Assessment of a state-of-the-art CNN in operational conditions. *Remote Sensing*, *11*(23), 2765.

NOAA (2021). *National Centers for Environmental Information (NCEI) U.S. billion-dollar weather and climate disasters*.

Noshadravan, A., Khajwal, A. B., & Cheng, C. S. (2021). A novel automated post-disaster damage assessment based on multi-view imagery. In *Mechanistic machine learning and digital twins for computational science, engineering & technology: An IACM conference*. https://doi.org/10.26226/morressier.612f6737bc981037241008a5

Oh, B. K., Kim, K. J., Kim, Y., Park, H. S., & Adeli, H. (2017). Evolutionary learning based sustainable strain sensing model for structural health monitoring of high-rise buildings. *Applied Soft Computing*, *58*, 576–585.

Pan, X., & Yang, T. (2020). Postdisaster image-based damage detection and repair cost estimation of reinforced concrete buildings using dual convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering*, *35*(5), 495–510.

Pan, X., & Yang, T. (2021). Image-based monitoring of bolt loosening through deep-learning-based integrated detection and tracking. *Computer-Aided Civil and Infrastructure Engineering*, *37*(10), 1207–1222.

Peña-Mora, F., Aziz, Z., Chen, A. Y., Plans, A., & Foltz, S. (2008). Building assessment during disaster response and recovery. *Proceedings of the Institution of Civil Engineers-Urban Design and Planning*, *161*(4), 183–195.

Pi, Y., Nath, N. D., & Behzadan, A. H. (2020). Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. *Advanced Engineering Informatics*, *43*, 101009.

Powers, D. M. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Presa-Reyes, M., & Chen, S. C. (2020). Assessing building damage by learning the deep feature correspondence of before and after aerial images. In *2020 IEEE conference on multimedia information processing and retrieval (MIPR)* (pp. 43–48). IEEE. https://ieeexplore.ieee.org/abstract/document/9175304

Rachev, S. T. (1985). The Monge–Kantorovich mass transference problem and its stochastic applications. *Theory of Probability & Its Applications*, *29*(4), 647–676.

Rafiei, M. H., & Adeli, H. (2017a). NEEWS: A novel earthquake early warning model using neural dynamic classification and neural dynamic optimization. *Soil Dynamics and Earthquake Engineering*, *100*, 417–427.

Rafiei, M. H., & Adeli, H. (2017b). A new neural dynamic classification algorithm. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(12), 3074–3083.

Rafiei, M. H., & Adeli, H. (2017c). A novel machine learning-based algorithm to detect damage in high-rise building structures. *The Structural Design of Tall and Special Buildings*, *26*(18), e1400.

Roueche, D.B., Lombardo, F. T., Krupar- III, J. R., & Smith, D. J. (2018). *Collection of perishable data on wind- and surge-induced residential building damage during Hurricane Harvey (TX), in Collection of Perishable Data on Wind- and Surge-Induced Residential Building Damage During Hurricane Harvey (TX). DesignSafe-CI*. https://doi.org/10.17603/DS2DX22

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, *40*(2), 99–121.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Li, F. F. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Sajedi, S., & Liang, X. (2021a). Dual Bayesian inference for risk-informed vibration-based damage diagnosis. *Computer-Aided Civil and Infrastructure Engineering*, *36*(9), 1168–1184.

Sajedi, S. O., & Liang, X. (2021b). Uncertainty-assisted deep vision structural health monitoring. *Computer-Aided Civil and Infrastructure Engineering*, *36*(2), 126–142.

Seeland, M., & Mäder, P. (2021). Multi-view classification with convolutional neural networks. *Plos One*, *16*(1), e0245230.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Shwe, T., & Adeli, H. (1993). AI and CAD for earthquake damage evaluation. *Engineering Structures*, *15*(5), 315–319.

StEER. (2022). Structural extreme event reconnaissance. https://www.steer.network (Accessed July 2022)

Sublime, J., & Kalinicheva, E. (2019). Automatic post-disaster damage mapping using deep-learning techniques for change detection: Case study of the Tohoku Tsunami. *Remote Sensing*, *11*(9), 1123.

Wang, S., Zargar, S. A., & Yuan, F. G. (2021). Augmented reality for enhanced visual inspection through knowledge-based deep learning. *Structural Health Monitoring*, *20*(1), 426–442.

Xu, J., Gui, C., & Han, Q. (2020). Recognition of rust grade and rust ratio of steel structures based on ensembled convolutional neural network. *Computer-Aided Civil and Infrastructure Engineering*, *35*(10), 1160–1174.

Xu, J. Z., Lu, W., Li, Z., Khaitan, P., & Zaytseva, V. (2019). Building damage detection in satellite imagery using convolutional neural networks. *arXiv preprint arXiv:1910.06444*.

Xu, Y., Lu, X., Cetiner, B., & Taciroglu, E. (2021). Real-time regional seismic damage assessment framework based on long short-term memory neural network. *Computer-Aided Civil and Infrastructure Engineering*, *36*(4), 504–521.

14678667, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/mice.12890 by Osaka University, Wiley Online Library on [24/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Yuan, F. G., Zargar, S. A., Chen, Q., & Wang, S. (2020). Machine learning for structural health monitoring: Challenges and opportunities. In H. Huang (Ed.), *Sensors and smart structures technologies for civil, mechanical, and aerospace systems 2020* (Vol. 11379, pp. 1137903). International Society for Optics and Photonics. https://spie.org/Publications/Proceedings/Volume/11379?SSO=1

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881–2890). IEEE.

How to cite this article: Khajwal, A. B., Cheng, C.-S., & Noshadravan, A. (2023). Post-disaster damage classification based on deep multi-view image fusion. *Computer-Aided Civil and Infrastructure Engineering*, *38,* 528–544. https://doi.org/10.1111/mice.12890

14678667, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/mice.12890 by Osaka University, Wiley Online Library on [24/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License