Received: 12 November 2021 | Revised: 30 March 2022 | Accepted: 20 May 2022

DOI: 10.1002/stc.3019

RESEARCH ARTICLE

WILEY

Uncertainty-aware convolutional neural network for explainable artificial intelligence-assisted disaster damage assessment

Chih-Shen Cheng[1] | **Amir H. Behzadan**[2] | **Arash Noshadravan**[1]

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

[2]Department of Construction Science, Texas A&M University, College Station, TX, USA

Correspondence
Arash Noshadravan, Zachry Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX 77843, USA.
Email: noshadravan@tamu.edu

Abstract

Accurate damage assessment is a critical step in post-disaster risk assessment, mitigation, and recovery. Current practices performed by experts and reconnaissance teams in the form of field evaluation require considerable time and resources. Recent advances in remote sensing imagery, artificial intelligence (AI), and computer vision have enhanced automated and rapid disaster damage assessment. Recent literature has shown promising progress in AI-assisted aerial damage assessment. However, accounting for the uncertainty in the outcome for improved quantification of confidence and enhanced model explainability for human decision-makers remains one of the key challenges. Overlooking uncertainty can lead to erroneous decisions, especially in highly-consequential tasks such as damage assessment. The aim of this study is to develop uncertainty-aware deep learning models for the assessment of post-disaster damage using aerial imaging. Within the framework of variational Bayesian inference, Monte Carlo dropout sampling technique is used to propagate epistemic uncertainty in model predictions. With this stochastic setting, the model produces damage prediction labels with softmax as random variables, which helps quantify confidence in the model outcome using appropriate measures of uncertainty. Two networks are implemented and trained separately on two different disaster damage datasets consisting of unmanned aerial vehicle building footage as well as satellite-captured post-disaster imagery. The first network attains 59.4% accuracy in building classification, and the second network gives an accuracy of 55.1%. Results from uncertainty analysis, model confidence quantification, and analyzing model attention zone can lead to more explainable and risk-informed automated damage assessment outcomes using AI technology.

KEYWORDS

Bayesian inference, disaster damage, deep learning, remote sensing, uncertainty quantification

1 | INTRODUCTION

Accurate, rapid, and reliable post-disaster structural damage assessment is key to successful disaster risk assessment, mitigation, and response. Disaster reconnaissance and damage assessment, and particularly windshield surveys, is one of the critical first steps in seeking a presidential disaster declaration, by providing pivotal information and narratives of impact in the aftermath of disasters.[1] Traditionally, human crews (e.g., building inspectors, engineers, and volunteers) are deployed to the field to conduct a thorough assessment of damages to structural components (e.g., walls, columns, and roofs) and evaluate the extent of damage to structures. However, this process can be time-consuming and labor-intensive and hinder the ability of field crews to collect and process ephemeral disaster data. For example, after Hurricane Dorian in 2019, two reconnaissance units took three days to complete damage assessment for residential and commercial buildings in Marsh Harbor and Treasure Cay in the Bahamas.[2]

The emergence of remote sensing technologies and computer vision techniques has paved the way for more rapid and scalable post-disaster damage reconnaissance.[3] Optical imagery data collected from satellites and aerial platforms can be adopted for visual damage assessment immediately after disasters where these photos are processed and analyzed to recognize the damage extent or patterns of infrastructures. In this context, roof and façade damage are commonly observed building impacts under high-wind events (e.g., hurricanes and tornados), and remote sensing is capable of taking top-view photos that show roof and building appearance damages while keeping costs low and efficiency high. As shown in Figure 1, satellite photos have the advantage of covering larger areas in a single image. Moreover, the global position information (i.e., latitude and longitude) of buildings in satellite imagery as well as corresponding pre-disaster imagery are usually available, which can assist in improving the quality of damage assessment by comparing post- and pre-disaster images (i.e., change analysis).[4] For unmanned aerial vehicles (UAVs), the camera flight path and elevation are flexible when flying UAVs, and the resulting images are of sufficient quality for object detection and building-level damage classification.[5] In particular, the views of UAVs can depict building roofs as well as capture damage features on doors, which provides sufficiently representative data for detailed damage assessment. This study employs both satellite and UAV visual data to demonstrate the application of the proposed method for artificial intelligence (AI)-assisted disaster damage classification.

To expedite the damage assessment process, past research has introduced deep learning, particularly convolutional neural networks (CNNs), for automated damage assessment using aerial imagery.[6–14] Given aerial input visual data, these models have shown promises in performing visual damage assessment and speeding up the evaluation process. However, the adoption rate of these techniques in practice has been historically low partially due to the unknown model reliability or reluctance of humans to concede the responsibility to the machine (e.g., deep learning) when making life-critical decisions.[15] The erroneous predictions by deep vision models may potentially results in fatal accidents in recent years. For instance, in 2017, an autonomous vehicle driving system failed to distinguish between a white trailer and the white sky backdrop,[16] resulting in severe injuries to the driver. This and similar accidents have led researchers to investigate ways to improve the reliability and prediction capability of autonomous systems. McAllister et al[17] emphasize the great need for modeling the uncertainty of autonomous driving systems to prevent safety accidents. In addition, while deep learning has been shown to achieve nearly perfect (100%) accuracy for medical diagnoses using X-ray images, a tiny chance (<1%) of erroneous prediction can lead to fatal and catastrophic outcomes.[18,19] In the context of disaster management, uncertainty-aware and trustworthy damage assessment is critical to first responders

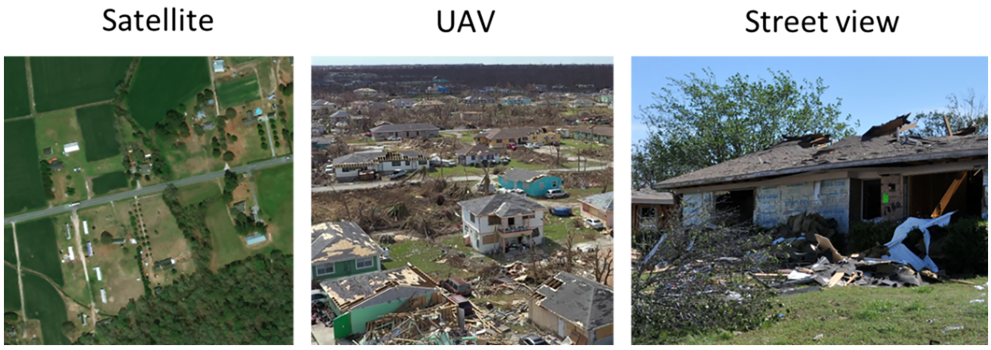

FIGURE 1 Post-disaster photos (captured from satellite, UAV, and human on the ground)

(e.g., NGOs, firefighters, engineers, and rescue teams) and decision-makers (e.g., government agencies) as it enables more informed operational decisions with real-world consequences.[1,20] Moreover, inaccurate prediction of disaster damage by AI models can mislead the allocation of recovery and rebuild resources and government assistance funds.[1] Incorporating risk information and context into model prediction can greatly contribute to the quality of deep vision-assisted post-disaster reconnaissance. These motivations necessitate investigating and exploring ways to quantify the propagated uncertainty and use this outcome to guide the reliability of AI-based disaster damage assessment.

The successful adoption of AI models in disaster management can be hampered by the lack of human trust in AI predictions due to the inherent uncertainty in the outcome, which may stem from uncertainty in models or data. This uncertainty may hinder the ability to make decisions with enough confidence. An important step to address this issue is to rigorously quantify and communicate confidence in the outcome. This requires a suitable framework for quantifying and propagating uncertainty in model predictions. The uncertainty-aware AI models in this research are a response to this need. Estimating the probabilistic descriptions of quantities of interest corresponding to the predictions of the AI model provides the ingredients of risk-informed decision making related to building damage severity in the aftermath of a disaster. Additionally, the quantified uncertainty output can represent the value of information (VOI) to determine whether other sources of data or sensing, such as human intervention, is necessary to resolve the uncertainty. This capability can pave the way for effective human-AI cooperation that leverages the collective strengths of humans and machines to reduce uncertainty in automated damage assessment.

The main contribution of this study is threefold: (1) A probabilistic framework for the AI-assisted classification of post-disaster damage to buildings is established. The framework uses deep CNN-based multiclass classification and variational Bayesian inference to estimate the probabilistic descriptions of damage level predictions. Predictive confidence and model reliability are analyzed through appropriate measurements of uncertainty. With quantified uncertainty, the AI model for disaster damage predictions generates outcomes that are more informative and representative of human-generated assessment; (2) the degree by which the prediction uncertainly is influenced by the actual (ground truth) damage state, and thus, the difficulty of the classification task, is analyzed; (3) the causability and explainability of AI model is investigated using the combination of uncertainty quantification and estimation of the model's attention zone when predicting the damage level. On the whole, the results of this study pave the way for more reliable, explainable, and risk-informed results in AI-assisted disaster assessment.

The paper is organized as follows. Section 2 provides a review of the literature on deep learning for structural health monitoring (SHM) and, in particular, its applications on post-disaster reconnaissance and damage assessment. Section 3 presents the methodology of this study, including data description, uncertainty quantification approach using Bayesian approximation, and CNN model architectures for two application database under study. Section 4 presents model implementation and calibration process. Section 5 presents the main results on two database scenarios, the discussion about quantification, and representation of uncertainty and model confidence. Finally, the conclusion section summarizes the main findings, emphasizes key contributions, and describes potential directions for future work.

2 | RELATED WORKS

2.1 | Deep learning applications in SHM

Computer vision and deep neural network have been introduced in recent years to the applications of SHM in structure and infrastructure systems, in general, and post-disaster reconnaissance, in particular. In the SHM domain, for example, Chen and Jahanshahi[21] proposed a deep learning-based approach, based on a CNN and a Bayesian-based data fusion methodology, for vision-based crack detection on metallic surfaces using inspection videos. Modarres et al[22] proposed a CNN-based approach (>90% accuracy) to identify the presence of concrete damage, which successfully assists in preventative structure maintenances and inspections. Ni et al[23] utilized convolutional features in CNN to detect visual concrete cracks at pixel level with nearly 80% precision and recall. Kim and Cho[24] devised a Mask R-CNN-based approach to conduct concrete crack assessment where crack widths can be accurately estimated (approximately 10% error). Jana and Nagarajaiah[25] employed computer vision for real-time measurements of tension in cable-stayed bridges. Mondal et al[26] developed a region-based CNN (R-CNN) for multi-class damage detection to reinforced concrete (RC) buildings. Bhowmick et al[6] proposed a CNN-based approach that can detect cracks on the

concrete surface as well as quantify their geometric properties such as length, width, and area using UAV-based videos. Zhang et al[27] introduced a Bayesian dynamic linear mode for online anomaly identification on SHM data. This approach enabled near real-time sensor-based structural condition assessment and decision making. Yan et al[13] leveraged CNN and lidar data from UAVs to automate detection and quantification of concrete cracks. Zou et al[28] explored deep neural network approaches for detecting pixel-wise patterns of ancient buildings in China. Sajedi and Liang[29] proposed the use of deep generative Bayesian optimization for searching optimal sensor placement in the sensor-based SHM. Case studies showed that generative machine learning achieves approximately 10% better performance compared to a benchmark genetic algorithm (GA). Hughes et al[30] presented a risk-based solution of active learning by leveraging probabilistic classifiers for determining the health of structures. Jana et al[31] utilized deep learning models to detect fault in sensor data and locate the faulty sensor and consequently reconstructed the correct sensor data in real-time.

In the context of disaster reconnaissance, Yeum et al[32] adopted CNN for visual structural damage classification in post-event building assessment. Yeum et al[33] proposed a CNN-based approach to automatically recognize and organize partial structure drawing images and reconstruct them into full drawing images for timely post-disaster reconnaissance. Pi et al[11] designed deep neural netwok models to detect damaged and undamaged roofs in post-hurricane footages. Chen et al[8] explored deep neural networks for damage evaluation in the aftermath of a tornado event. For model training, they created a visual damage assessment dataset by manually annotating unpiloted aerial system (UAS) tiles following the 26 June 2018 Eureka Kansas tornado. Cheng et al[10] proposed a stacked CNN architecture to detect damaged buildings and differentiate between various damage states in UAV footages after 2019 Hurricane Dorian with 60% accuracy (6 damage level classes). Most of these studies employed deep learning in a deterministic setting, and they did not account for the uncertainty in model predictions for risk-informed decision making.

2.2 | Applications of Monte Carlo (MC) dropout in deep learning

Uncertainty quantification plays an important role in predictive science as it provides the ingredients to evaluate the reliability of predictive models.[34] In the context deep learning, Bayesian approximation using MC dropout sampling is one of the main mechanisms to propagate and estimate the epistemic uncertainty (i.e., uncertainty originated from models). Harper and Southern[35] introduced a Bayesian deep learning framework (based on MC dropout) to explore the reliability of human emotion prediction using CNN and recurrent neural network (RNN) architectures. Since uncertainty is extremely important in machine-assisted medical diagnoses, CNNs with MC dropout for uncertainty propagation have been also utilized for magnetic resonance imaging (MRI) segmentation.[36–39] In the SHM domain, the reliability of AI-assisted structural monitoring and early warning systems are crucial because mis-predictions might cause impactful or fatal outcomes. In this context, Sajedi and Liang[40] proposed a CNN-based approach that uses MC dropout to incorporate uncertainty in the image-based SHM for damage detection of bridge structures. Vashisht et al[41] utilized both a Bayesian CNN and Long Short-Term Memory (LSTM) approach for vibration-based damage detection and localization in beam structures. While there are several recent studies adopting MC dropout for uncertainty estimation, the applications in SHM and post-disaster reconnaissance domains, particularly for post-disaster damage assessment, are still under-explored.

3 | METHODOLOGY

This section presents the framework used in this study to develop uncertainty-aware CNNs for probabilistic disaster damage assessment. The main approach can be divided into three steps: data preparation, model development, and uncertainty quantification. The methodology for the uncertainty quantification primarily relies on variational Bayesian inference where the fundamental theory will be overviewed in Section 3.2. The schematic diagram of the overall methodology is shown in Figure 2. Two different datasets, based on drone view and satellite view, are employed in this work for model development. A Bayesian inference framework is implemented using MC dropout sampling techniques to propagate the uncertainty in neural networks. The probabilistic descriptions of modes' output corresponding to prediction of damage scales are estimated to quantify the confidence in model predictions.

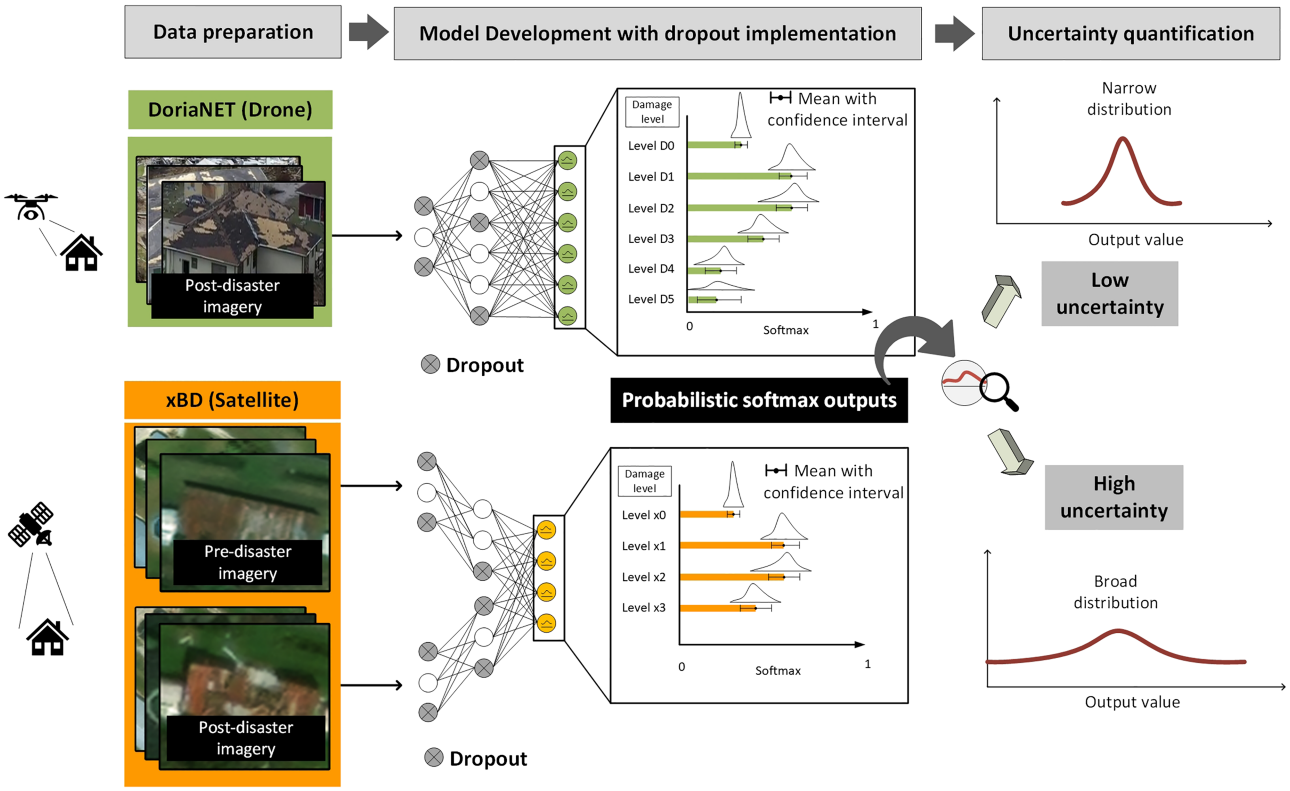

FIGURE 2 Schematic diagram of the methodology including data preparation, model development, and uncertainty quantification approach

FIGURE 3 Samples of building images with different levels of damage in DoriaNET

3.1 | Dataset preparation

The authors have recently created an published an in-house dataset, DoriaNET,[42] for aerial damage assessment. DoriaNET is a post-hurricane database that consists of visual annotations of UAV footages in the aftermath of the 2019 Hurricane Dorian. Figure 3 presents several sample images of damaged buildings in DoriaNET. The damage level descriptions and annotation guidelines used in DoriaNET are based on a wind-induced damage scale by Federal Emergency Management (FEMA) HAZUS model,[43] as shown in Table 1. A brief description of the dataset is presented in Tables 2 and 3. As shown in Table 2, the dataset contains 1501 annotated undamaged (level 0) and damaged (levels 1–5) building samples captured by UAV.

Another dataset of building damage that is utilized in this work for training the CNN model is xBD.[14] xBD contains satellite photos (pre- and post-disaster) covering a total area of over 45,000 square kilometers with damage level annotations of 850,736 buildings in recent worldwide disasters. Note that building photos may look blurred or hazy as shown in Figure 4. This is due to the low quality of xBD satellite images when zooming in to individual building locations (as presented in Figure 2) or as a result of smoke or could cover. It is expected that the lower quality imagery data similar to the ones in xBD will lead to increased uncertainty in model performance. Since this paper focuses on the wind-

induced damage assessment, we manually select samples from wind-related disasters (e.g., hurricanes and tornados) for CNN training as shown in Table 3. We use undersampling in the dataset to rectify data imbalance as well as eliminate very low-quality building imagery (e.g., images with less than 30 pixels and post-disaster image covered by clouds). Overall, a balanced dataset with 8056 samples is created for developing a deep learning model.

TABLE 1 Damage type description (wind-induced damage) in DoriaNET and xBD

DoriaNET damage level	Damage building description
D0 (No damage or very minor damage)	Little or no visible damage from the outside. No broken windows or failed roof deck. Minimal loss of roof over.
D1 (Minor damage)	Maximum of one broken window, door, or garage door. Moderate roof cover loss that can be covered to prevent additional water entering the building.
D2 (Moderate damage)	Major roof cover damage, moderate window breakage. Minor roof sheathing failure.
D3 (Severe damage)	Major window damage or roof sheathing loss. Major roof cover loss.
D4 (Destruction)	Complete roof failure and/or failure of wall frame. Loss of more than 50% of roof sheathing.
D5 (Destroyed and under-construction)	Completely destroyed or under the early stages of construction.
xBD damage level	**Damage building description**
x0 (No damage)	Undisturbed. No sign of structural or shingle damage.
x1 (Minor damage)	Roof element missing.
x2 (Major damage)	Partial wall or roof collapse.
x3 (Destroyed)	Completely collapsed or no longer present.

TABLE 2 Data description

Dataset name	Disaster name	Event date	Location	Type	Sample size
DoriaNET	Hurricane Dorian	Aug 28–31, 2019	The Bahamas	Drone	1501
xBD	Hurricane Florence	Sep 10–19, 2018	Carolinas, U.S.	Satellite	6446
xBD	Hurricane Harvey	Aug 17–Sep 2, 2017	Texas, U.S.	Satellite	23,013
xBD	Hurricane Mathew	Sep 28–Oct 10, 2016	The Bahamas	Satellite	13,937
xBD	Hurricane Michael	Oct 7–16, 2018	Florida, U.S.	Satellite	22,679
xBD	Joplin MO Tornado	May 22, 2011	Missouri, U.S.	Satellite	15,351
xBD	Moore OK Tornado	May 20, 2013	Oklahoma, U.S.	Satellite	22,957
xBD	Tuscallosa AL Tornado	Apr 27, 2011	Alabama, U.S	Satellite	15,001

TABLE 3 Distribution of annotation damage levels in DoriNET and xBD

Dataset Name	Disaster name	Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
DoriaNET	Hurricane Dorian	342	556	392	956	542	241
xBD	Hurricane Florence	356	228	265	54	N/A	N/A
xBD	Hurricane Harvey	266	256	338	201	N/A	N/A
xBD	Hurricane Mathew	193	347	223	146	N/A	N/A
xBD	Hurricane Michael	333	283	302	257	N/A	N/A
xBD	Joplin MO Tornado	327	279	307	374	N/A	N/A
xBD	Moore OK Tornado	368	254	249	384	N/A	N/A
xBD	Tuscallosa AL Tornado	264	289	316	597	N/A	N/A

FIGURE 4 Damage building samples in xBD

In xBD, the damage rating system is established according to a joint damage scale proposed by Gupta et al[14] as presented in Table 1. This joint damage scale integrates existing damage guidelines such as FEMA HAZUS (the same system used in DoriaNET), Kelman,[44] and EMS-98.[45] It should be emphasized that building samples in DoriaNET and xBD were annotated using two distinct damage level rating systems with different damage level ranges (DoriaNET: D0-D5, xBD: x0-x3), as shown in Figures 3 and 4. Additionally, in the xBD dataset, pre- and post-disaster aerial photos are both available, whereas DoriaNET only contains post-disaster images (Figure 3). For the purpose of model development, each dataset is further split into training (60%), validation (20%), and testing sets (20%).

3.2 | Uncertainty quantification in CNN using Bayesian approximation

The use of deep learning, particularly CNN, in performing computer vision tasks (e.g., classification and object detection) has been growing in recent years. Most CNN applications have been developed and used in a deterministic setting where the weights or kernels (expressed as $\mathbf{W}$) in the standard CNN architecture are deterministic. The trainable weights are tuned by minimizing the loss function (e.g., cross-entropy loss) obtained based on the training dataset which contains N_p visual inputs, that is, RGB values, denoted as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_p}\}$ and their assigned labels denoted as $\mathbf{Y} = \{y_1, y_2, \ldots, y_{N_p}\}$. However, quantification of epistemic uncertainty in the prediction of CNN models requires a stochastic setting that allows estimating the probabilistic characteristics of the model predictions. Bayesian inference is a common way to achieve this.[46] In this setting, the kernels $\mathbf{W}$ in CNN are treated as random variables for which the probabilistic descriptions are represented as probability distribution functions (PDFs). The resulting CNN can be deemed as a complicated function with tons of kernel distributions. The focal point is to obtain posterior distributions of network weights given training data, which is expressed as $P(\mathbf{W}|\mathbf{X},\mathbf{Y})$ or simply stated as $P(\mathbf{W}|data)$. It has been shown that obtaining analytical $P(\mathbf{W}|data)$ is not feasible since it is intractable to calculate the analytical posterior using the Bayes rule,[47] expressed in Equation (1):

$$P(\mathbf{W}|\mathbf{X},\mathbf{Y}) \propto P(\mathbf{Y}|\mathbf{X},\mathbf{W})P(\mathbf{W}) \tag{1}$$

As an alternative, Graves[48] proposed the use of the variational Bayesian method where a simplified distribution known as approximate variational distribution, $q(\mathbf{W})$, is sought to approximate the target posterior kernel distributions. The variational distribution $q(\mathbf{W})$ can be calculated by minimizing the Kuller–Leibler (KL) divergence,[49] a measure of the difference between two probability distributions, described as follows in Equation (2):

$$D_{\mathrm{KL}}(\, q(\mathbf{W}) \| P(\mathbf{W}|\mathbf{X},\mathbf{Y})) \tag{2}$$

Recently, Gal and Ghahramani[47] proposed the use of a Bernoulli distribution-based variational distribution denoted as $\hat{q}(\mathbf{W})$ to approximate the posterior $P(\mathbf{W}|data)$. Particularly, the process of minimizing the KL divergence between

$\hat{q}(\mathbf{W})$ and $P(\mathbf{W}|data)$ is shown to be equivalent to networks trained with dropout regularization.[47] For a given input $\mathbf{x}_i$, let $\hat{y}_i$ be the corresponding model prediction. The minimization of Equation (2) can be expressed as the training loss for CNN as shown in Equation (3):

$$\underset{\mathbf{W}}{\text{minimize}} \; D_{\text{KL}}(\hat{q}(\mathbf{W})\|P(\mathbf{W}|\mathbf{X},\mathbf{Y})) = \frac{1}{N_p}\sum_{i=1}^{N_p} Er(\hat{y}_i,y_i) + c\sum(\|\mathbf{W}\|^2 + \|\mathbf{b}\|^2) \tag{3}$$

where $Er(\hat{y}_i,y_i)$ represents the error between model prediction and the ground truth, that is, the assigned label, $\mathbf{b}$ is the vector of bias, and c is a scalar constant. Equation (3) shows that $\hat{q}(\mathbf{W})$ can be estimated following the standard deep learning training procedure (i.e., training, validation, and testing). The implication of this formulation is that the dropout function is now active in all phases instead of being active merely during training. With the dropout activated at the inference time, CNN outputs a prediction where the target classes' softmax outputs are deemed random variables characterized by their PDF. Having the probabilistic descriptions of predicted softmax outputs in hand, suitable metrics (e.g., entropy and standard deviation) can be calculated to measure the CNN epistemic uncertainty. Since the posterior is estimated by the approximate variational distribution $\hat{q}(\mathbf{W})$, the softmax output of a testing input image $\mathbf{x}^*$ belongs to k^{th} class, $P(\hat{y}_k|\mathbf{x}^*,\mathbf{X},\mathbf{Y})$, can be computed using Equation (4):

$$P(\hat{y}_k|\mathbf{x}^*,\mathbf{X},\mathbf{Y}) = \int P(\hat{y}_k|\mathbf{x}^*,\mathbf{X},\mathbf{Y},\mathbf{W})P(\mathbf{W}|\mathbf{X},\mathbf{Y})d\mathbf{W} \simeq \int P(\hat{y}_k|\mathbf{x}^*,\mathbf{X},\mathbf{Y},\mathbf{W})\hat{q}(\mathbf{W})d\mathbf{W} \tag{4}$$

Note that the left-hand side of Equation (4), that is, $P(\hat{y}_k|\mathbf{x}^*,\mathbf{X},\mathbf{Y})$, is the expectation of $P(\hat{y}_k|\mathbf{x}^*,\mathbf{X},\mathbf{Y},\mathbf{W})$ with respect to the probability measure of $\mathbf{W}$. In the damage assessment scenario shown in Figure 5, this quantity represents the softmax value associated with each damage severity classes k ($k = 0$–3 for xBD, $k = 0$–5 for DoriaNET). For a given realization W, let us denote the vector of softmax output as $\mathbf{S} = \{S_k\}_{k=0}^{n_c-1}$, where n_c is the total number of classes ($n_c = 4$ for xBD and $n_c = 6$ for DoriaNET). Using the CNN with Bayesian approximation, for a given image input, we can now construct the probabilistic description of softmax values S_k for each class k. In particular, with sufficiently large number of MC samples N_{MC}, the expected value of S_k, $\mathbb{E}[S_k]$, can be estimated, which provides a numerical estimation of the integral in Equation (4). Let us denote the j^{th} MC samples as S_k^j. Then this expectation is expressed in Equation (5) as follows:

$$P(\hat{y}_k|\mathbf{x}^*,\mathbf{X},\mathbf{Y}) = \mathbb{E}[S_k] \simeq \frac{1}{N_{MC}}\sum_{j=1}^{N_{MC}} S_k^j \tag{5}$$

Figures 2 and 5 present the schematic diagrams of deep learning UQ using Bayesian inference with MC dropout sampling. The upper right part of Figure 2 presents a scenario where the output values are within a narrow range even though the neurons are randomly dropped. This scenario suggests that the network is confident and has low

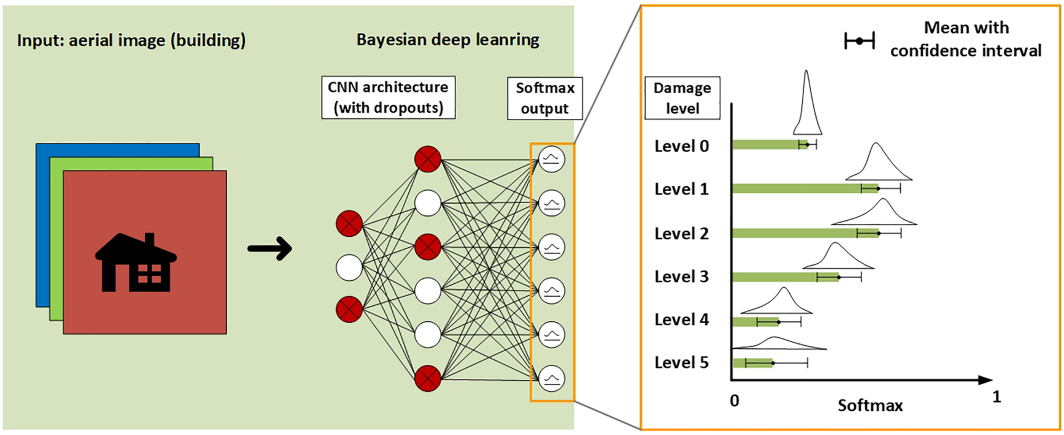

FIGURE 5 A schematic diagram of neural network classifiers with softmax distribution outputs

uncertainty in this predictive outcome. On the other hand, the lower right part of Figure 2 demonstrates a high variance with a broad range of output values, showing a high uncertainty of the model prediction. Moreover, as shown in Figure 5, the dropout mechanism in the model is activated during the inference. The model output in terms of softmax probabilities associated with each damage severity class is treated as random variables characterized by their probability density functions.

It is worth noting that the traditional deterministic classifiers usually output a softmax probability vector which is the output of the softmax function calculated from the last layer of the neural network.[50] While the softmax value can represent an estimate of model performance,[10] it may not fully represent the extent of model uncertainty.[47] A more rigorous quantification of uncertainty requires a stochastic setting where the model uncertainty is taken into account. Bayesian inference provides such a setting, and the Bayesian deep learning approach presented in this paper enables a consistent and systematic quantification and propagation of uncertainty and model confidence.

The overall uncertainty in model predictions can be aggregated and represented using proper measures based on statistics and information theory. Kendall and Gal[51] proposed using Entropy[52] to quantify the epistemic uncertainty of CNN predictions. For the problem in hand the Entropy (H) can be expressed as follow in Equation (6):

$$H(\mathbf{S}) = -\sum_{k=0}^{n_c-1} \mathbb{E}[S_k] \log(\mathbb{E}[S_k]) \tag{6}$$

Sajedi and Liang[40] employed the mean of the per-class softmax standard deviation (MCSSD) for quantifying the prediction outcome uncertainty. MCSSD takes into account the variance of MC samples by computing and aggregate the standard deviation of MC samples in each class, which is expressed in Equation (7):

$$MCSSD = \frac{1}{n_c} \sum_{k=0}^{n_c-1} \sqrt{\frac{\sum_{j=1}^{N_{MC}} \left(S_k^j - \mathbb{E}[S_k]\right)^2}{N_{MC} - 1}} \tag{7}$$

These two measures will be used in this work as aggregated measures of uncertainty in model predictions.

3.3 | CNN architectures

This study designs two CNN architectures for DoriaNET and xBD separately since the characteristics of the database and the damage scale classes are different. Here, we refer to these two CNN architectures as DoriaNET model and xBD model. As shown in Figure 6, we architect DoriaNET model based on MobileNet,[53] a fully trained CNN (pre-trained on ImageNet[54]) that can classify more than 1000 object classes with only 28 network layers. In order to explore the model

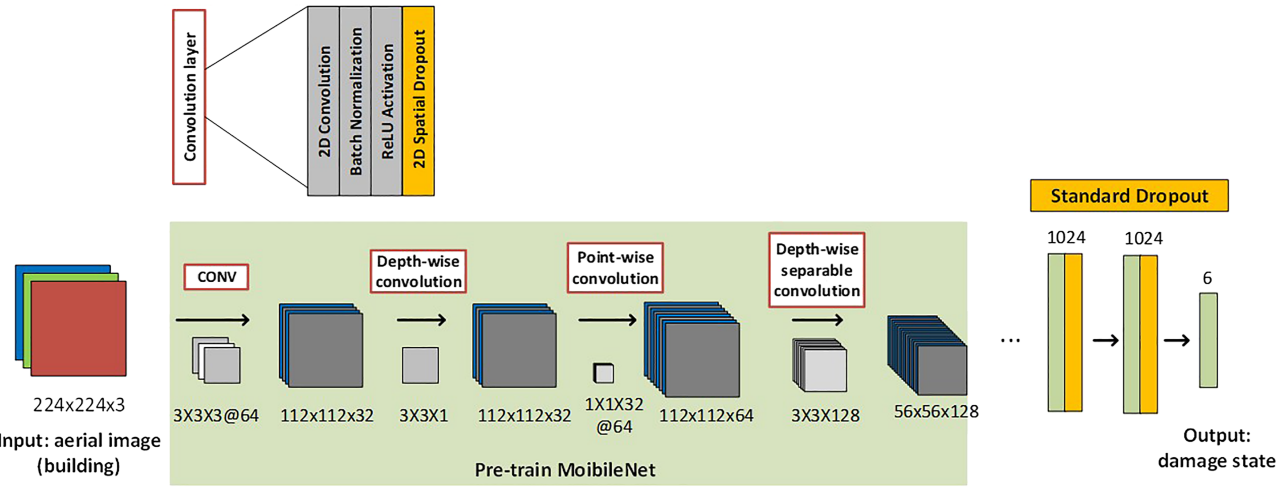

FIGURE 6 DoriaNET CNN architecture

reliability, we insert additional dropout layers (1% dropout rate) into every existing layer (both convolutional and dense layers) where the range of 0.5–1.5% dropout rates are suggested by the sensitivity analysis (Section 4.2). These dropout layers, in particular, are active both during training and inference phases. Similar to the DoriaNET model, the xBD model is also built upon MobileNet, and additional dropout layers with a 1% dropout rate are added to all existing layers. However, in xBD, each data point is a set of two images (pre-and post-disaster). As shown in Figure 7, this study designs a two-stream architecture following the concept proposed by Presa-Reyes and Chen.[12] The idea is to extract two image inputs independently using two pre-trained CNNs (same architecture with unshared weights) in stage 1. Next, stage 2 fuses the extracted information and allows the integrated neurons to go through two dense layers before outputting a prediction.

4 | MODEL IMPLEMENTATION AND CALIBRATION

4.1 | Model fine-tuning

The DoriaNET and xBD models (based on a pre-trained MobileNet CNN) were re-trained and validated on 60% and 20% of the datasets, respectively. For network training, the square of Earth Mover's Distance (also known as Wasserstein distance), denoted as EMD^2, was used as the loss function. It provides a measure for quantifying distance or dissimilarity between two probability distributions.[55] Let us denote $\hat{\mathbf{A}}$ and $\mathbf{A}$ as the model output and ground-truth vectors, respectively. The EMD^2 measure is mathematically expressed in Equation (8) as follows

$$EMD^2 = \sum_{k=0}^{n_c-1} \left(CDF_k(\hat{\mathbf{A}}) - CDF_k(\mathbf{A})\right)^2 \tag{8}$$

where $CDF(\cdot)$ denotes the cumulative distribution function and k is the damage level ranging from 1 to $n_c - 1$. It has been shown that EMD^2 is a suitable loss function for classification problems where there are strong inter-class relationships in the literature.[10,56] It should be, however, noted that there are other alternative distance measures that can be used as classification loss functions. One example of those is the Bhattacharyya distance,[57] which is suitable for measuring data separability in classification tasks.

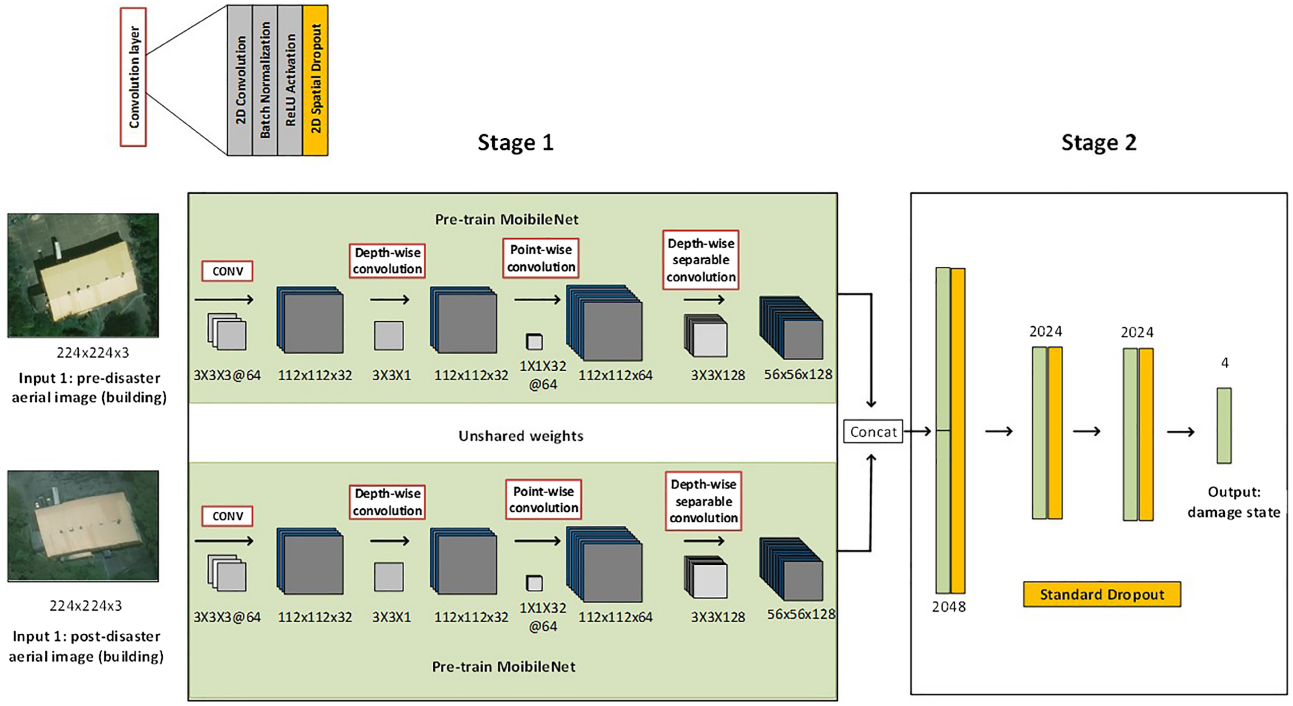

FIGURE 7 xBD dual-stream CNN architecture

The DoriaNET model was trained using Adam optimizer[58] with a learning rate of 10^{-5} and a batch size of 32, for 150 epochs. The xBD model was trained with Adam optimizer with a learning rate of 10^{-5} and a batch size of 64, for 150 epochs. During both training processes, each input was randomly augmented by scaling up or down by $\pm$ 10%, horizontally and vertically flipped from the selected 50% images. Note that the hyperparameters, such as learning rate, batch size, and epoch number, are selected empirically based on our previous works where we tuned and found the optimal (or suitable) hyperparameters for model training in post-disaster visual damage detection and classification applications. The CNN training time on DoriaNET was 3.07 h based on Intel Xeon E5-2680 v4 2.40GHz 14-core CPU, 128GBRAM, and NVIDIAK80 (12 GB) GPU. The second CNN was trained (on xBD) based on Intel Xeon E5-2680 v4 2.40GHz 28-core CPU, 128GBRAM, and NVIDIAK80 (16 GB) GPU for 26.4 hours.

4.2 | Sensitivity analysis with respect to the dropout rate

Based on previous research, suitable dropout rates for CNN were observed to be smaller than the one for fully connected networks.[51,59,60] In these references, 20–50% dropout rates were used for fully connected networks, and 1–20% rates were adopted for CNNs. Moreover, the suitable dropout rate also depends upon the characteristics of selected pretrained CNN architectures (e.g., ResNet, MobileNet, and DenseNet). Hence, we performed sensitivity analysis to determine a suitable dropout rate in our CNN architectures. Following the training procedures described in Section 5.1, we evaluated the fully trained CNN prediction accuracies on the testing portions where the dropout rate is adjusted in the range of 0–5%. Note that we used a sufficiently large MC sample size (100,000) in this sensitivity analysis to ensure that the prediction quality does not deteriorate. As can be seen in Figure 8, the accuracy slightly increases when using low dropout rates in the range of 0–1.2%. When the dropout rate becomes larger than 1.5%, the prediction performance significantly deteriorates. Based on this observation, the dropout rate of 1% has been selected for both DoriaNET and xBD models since it generally produces predictions with the best performance.

4.3 | Sensitivity analysis with respect to the MC sample size

In Section 4.2, a relatively large MC sample size of 100,000 was used to determine the optimal dropout rate of 1% for the CNN models. However, this large sample size can inadvertently result in a lengthy inference time, potentially slowing down post-disaster damage and impact evaluation. A sensitivity analysis is performed to explore the effect of the MC sample size on the accuracy in the results of the trained mode and determine the optimal number of samples for more efficient simulation. We select the fully trained CNN models with 1% dropout and analyze their testing accuracies with various sample size numbers as shown in Figure 9. It should be noted that inference results obtained from N_{MC} samples might fluctuate due to the stochasticity of the network dropout layers. To reduce this fluctuation, for each case with a particular MC sample size, we perform the model inference accuracy 5 times using MC sampling and report the mean of the obtained 5 accuracies. It is observed that for the sample size of $N_{MC} = 30$, the accuracy in the prediction using MC integration in Equation (4) is sufficiently converged as shown in Figure 9.

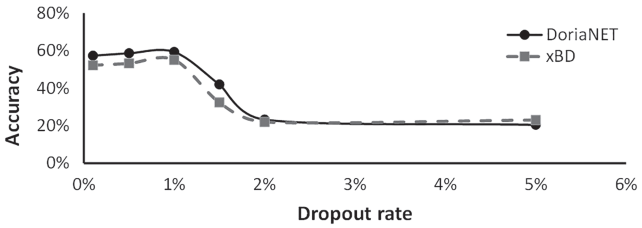

FIGURE 8 Sensitivity analysis of dropout rate based on 100,000 MC samples

5 | RESULTS AND DISCUSSION

5.1 | Model testing results

The standard metrics to evaluate the classification CNN performance include overall accuracy, precision, recall, and F1 values. The number of true-positive (TP_k), false-negative (FN_k), and false-positive (FP_k) cases in model predictions of each damage state k is required to compute these metric values. Precision is defined as $Pr_k = TP_k/(TP_k + FP_k)$, and recall is defined as $Re_k = TP_k/(TP_k + FN_k)$. The F1 score is the aggregation of precision and recall values, which ranges from 0 to 1. It is defined as follows in Equation (9):

$$(\mathrm{F1})_k = \frac{2 \times Pr_k \times Re_k}{Pr_k + Re_k} \tag{9}$$

The trained CNNs are evaluated on the testing portion (20%) of the datasets. To evaluate model performance, we select the damage level with the maximum mean of softmax values (based on $N_{MC} = 30$ samples) as the model prediction class *Pred*. That is expressed in Equation (10):

$$Pred = argmax_k(P(\hat{y}_k|\mathbf{x}^*, \mathbf{X}, \mathbf{Y})) = argmax_k(\mathbb{E}[S_k]) \tag{10}$$

The accuracy is defined as follows in Equation (11):

$$Accuracy = \frac{\sum_{k=0}^{n_c-1} TP_k}{\sum_{k=0}^{n_c-1}(TP_k + FP_k)} \tag{11}$$

Table 4 shows the confusion matrix of the DoriaNET model. Overall, DoriaNET model achieves 59.4% accuracy on test data. The precision, recall, and F1 scores are presented in Table 5. The presented confusion matrices provide insights into the scenarios where data-driven models are prone to make erroneous predictions. For example,

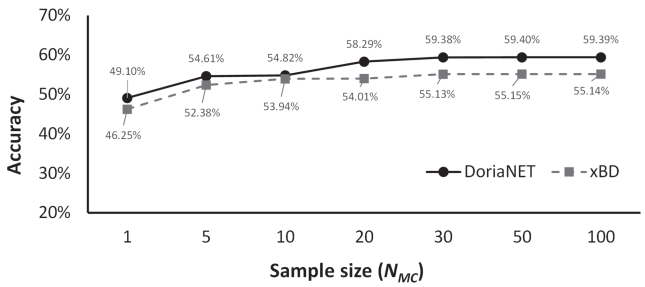

FIGURE 9 Sensitivity analysis of Monte Carlo sample size (N)

TABLE 4 Confusion matrix for DoriaNET model (MC sample size: 30)

		Predictions					
		D0	**D1**	**D2**	**D3**	**D4**	**D5**
Ground truth	D0	24 (71%)	10 (29%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	D1	12 (23%)	35 (66%)	3 (6%)	2 (4%)	0 (0%)	1 (2%)
	D2	5 (13%)	8 (21%)	16 (41%)	8 (21%)	0 (0%)	2 (5%)
	D3	1 (1%)	16 (17%)	8 (9%)	48 (52%)	12 (13%)	8 (9%)
	D4	1 (0%)	0 (0%)	0 (0%)	9 (17%)	30 (58%)	13 (25%)
	D5	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (5%)	21 (95%)

TABLE 5 DoriaNET model F1 score, Precision, and Recall

		F1 score	Precision	Recall
Damage level	D0	0.6316	0.5714	0.7059
	D1	0.5737	0.5072	0.6604
	D2	0.4849	0.5926	0.4103
	D3	0.6000	0.7164	0.5161
	D4	0.8333	0.9091	0.7692
	D5	0.8511	0.8000	0.9091

TABLE 6 Confusion matrix for xBD model (MC sample size: 30)

		Predictions			
		x0	x1	x2	x3
Ground truth	x0	266 (66%)	18 (5%)	40 (10%)	76 (19%)
	x1	83 (21%)	129 (32%)	69 (17%)	119 (29%)
	x2	30 (8%)	52 (13%)	204 (50%)	114 (29%)
	x3	30 (8%)	21 (5%)	66 (17%)	283 (70%)

TABLE 7 xBD model F1 score, Precision, and Recall

		F1 score	Precision	Recall
Damage level	x0	0.6576	0.6504	0.6650
	x1	0.4161	0.5864	0.3225
	x2	0.5237	0.5383	0.5100
	x3	0.5706	0.4780	0.7075

undamaged building samples are more easily mispredicted as minor damage than major damage samples since image features and patterns are similar between undamaged and minor damaged buildings. Similarly, xBD CNN reaches overall 55.1% accuracy, and Table 6 presents the confusion matrix when the trained xBD CNN is tested on the xBD testing set. Table 7 shows the overall precision, recall, and F1 values of each damage level. Note that the accuracy is not calculated as the simple average of the diagonal values from the confusion matrix. Instead, a better way of calculating the average accuracy is by considering the weight of each class since the datasets contain different numbers of instances in each damage level. Moreover, the lower accuracy of xBD compared to DoriaNET is due to the relatively low quality of satellite data. This aspect is critical for the damage assessment since roof, walls, and other building elements must remain visible in the process of assigning damage labels to a building.

5.2 | Uncertainty analysis

5.2.1 | Quantification of confidence in model predictions for uncertainty-informed decisions

The uncertainty-aware CNN models developed in this work enable estimating the probabilistic descriptions of quantities of interest corresponding to the model predictions. In particular, they can be used to estimate the distribution of softmax values for each class representing a damage level of a building. This information provides a means to quantify the confidence and reliability in AI-enabled disaster damage prediction. In addition, aggregated measures of uncertainty, such as Entropy (H) and MCSSD, can be used to comment on the reliability of the predictions. Figures 10 and 11 present examples of such outcomes in form of marginal PDFs of softmax values obtained from DoriaNET model and

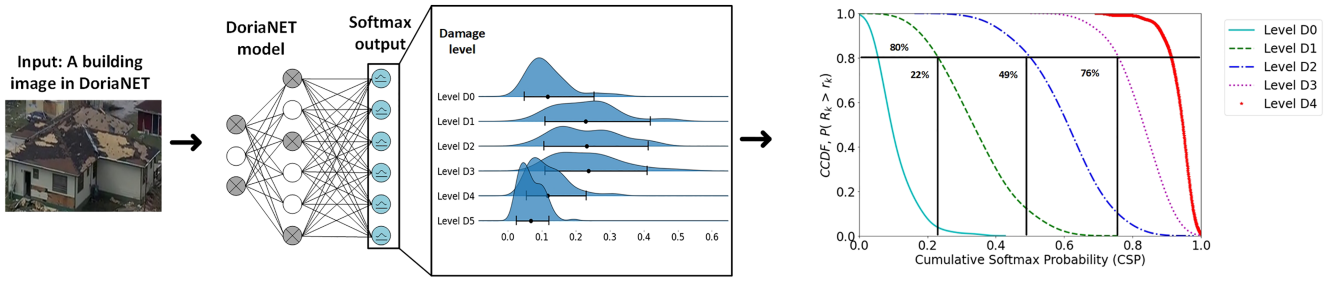

FIGURE 10 Example of DoriaNET model prediction with MC dropout sampling

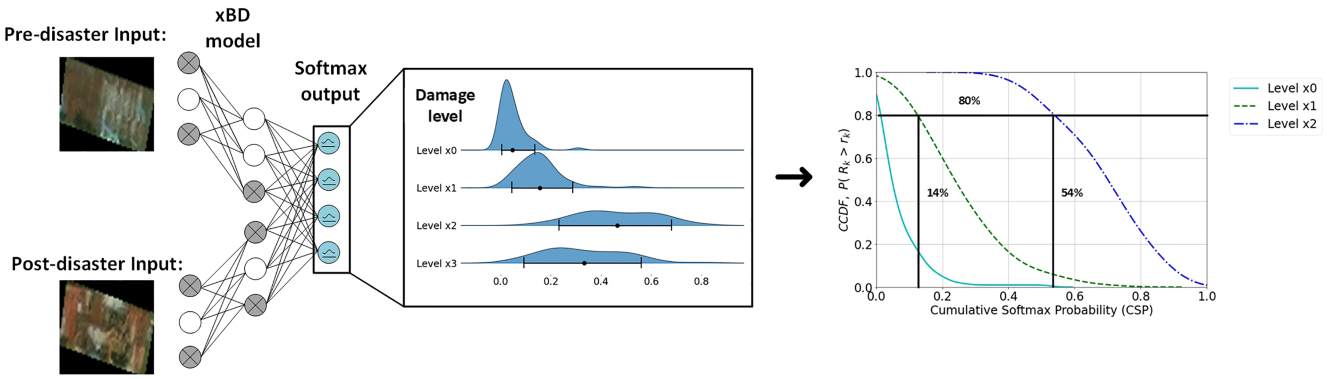

FIGURE 11 Example of xBD model prediction with MC dropout sampling

xBD dual-stream model, respectively. The mean of softmax values along with their 90% confidence bounds is also shown in these plots. The confidence bounds in Figure 10 are relatively broad, indicating high uncertainty for damage levels D1, D2, and D3, which also turn out to have the highest mean softmax values of about 30%. On the other hand, narrow confidence bounds (low uncertainty) are noticed for damage levels D0, D4, and D5, in which the mean softmax values are relatively smaller. These observations indicate that the model predicts with higher confidence that the actual damage is less likely to be within the groups of damage levels D0, D4, or D5. At the same time, the model remains uncertain about the particular damage level within higher likely classes of damage levels D1, D2, and D3 since the confidence bounds are highly overlapped between these three classes. It is worth noting that one can pick a different confidence level to observe and deduce conclusions from the probabilistic nature of the prediction outcome. Thus, the choice of confidence level (e.g., 90% for Figures 10 and 11 in this study) is a user-defined decision parameter indicative of the user's risk perception and tolerance in the choice of confidence bounds to communicate the results based on the intended use of the model. Clearly, the confidence bound will become narrower by applying lower confidence levels. In this case, smaller overlapping areas will be observed between the variations of softmax values across different classes leading to different outcomes and decisions by the end users. It should be noted that the marginal distributions of softmax values reported in Figures 10 and 11 are estimated from the generated MC samples using Kernel density estimation (KDE). Thus, there has not been any assumptions or predefined type of distributions. From what we have observed, there is no specific distribution type that can collectively describe the probabilistic nature of the output. In some cases, the distributions are close to Gaussian; however in many cases, they are skewed and the lognoraml or gamma distribution be more representative. Even in some instances, the distributions are relatively flat (close to uniform distribution) or even slightly bimodal. Figure 11 demonstrates this behavior, which, more or less, can be observed across different predictions. It is also important to note that the distributions are estimated based on limited samples. As such, they are subjected to statistical bias and numerical approximations associated with KDE. This also explains the negative values that are reported in the marginal pdfs of softmax values, which by definition should take values between 0 to 1.

The model predictions can also be represented in form of CDF of softmax values. This information can provide a different perspective of model reliability and facilitate uncertainty-informed decision making regarding building damage severity in the aftermath of a disaster. To that end, we introduce a new quantity referred to as cumulative softmax

probability (*CSP*), which represents the softmax probability that the damage state is equal to or less than a particular damage state k. In other words, CSP presents a softmax value corresponding to the prediction of an upper bound for damage state. Considering all the damage states, the collection of cumulative softmax probabilities form a vector of dimension n_c. Since our CNN models are stochastic and softmax values are random variables, the vector of CSP is also a random variable. Let us denote such vector-valued random variable as $\{R_k\}_{k=0}^{n_c-1}$. Using MC dropout sampling, realization of R_k can be generated and used to estimate its probabilistic description. Let us denote a particular realization as R_k^j, that is, the j^{th} MC sample corresponding to the softmax output of damage state equal or less than class k. This can be expressed by Equation (12) as follows:

$$R_k^j = \sum_{i=0}^{k} S_i^j \tag{12}$$

where $k = 0, \ldots, n_c - 1$. Given all the MC samples for a building input, the marginal complementary cumulative distribution function (CCDF) R_k, that is, probability of exceedance, can be calculated as follows in Equation (13).

$$P(R_k > r_k) = 1 - CDF_{R_k}(r_k) \tag{13}$$

These curves are plotted on the right side of Figures 10 and 11 for two building examples. The CSP values in the horizontal axis represent the softmax values of the upper bound of the ordinal damage levels. It is worth noting that softmax values represent the normalization of raw class scores into positive values that sum to 1. Thus, they are interpreted as the probabilities of damage belonging to different classes. However, their probabilistic notion does not reflect the uncertainty or confidence in model predictions.[47] The vertical axis, on the other hand, is the probability of exceedance of softmax values estimated from the realizations of model predictions. Thus, it represents the uncertainty in the probabilistic CNN models, and as such, it reflects the confidence in model predictions. For example, let's plot a horizontal solid line in Figure 10 with the CCDF value of 80%. Points of intersection between curves and this horizontal solid line present as follows: There is 80% chance that the value of CSP exceeds 22% when predicting an upper bound of D1 for the damage level. With the same 80% likelihood, the value of CSP exceeds 49% when predicting the damage level is at most D2. Similarly, in the right side of Figure 11, xBD model's predictions can be interpreted as follows: There is 80% likelihood that the cumulative softmax value exceeds 14% when predicting x1 as the upper bound of incurred damage. With the same 80% probability of exceedance, the CSP value is 54% when predicting the damage level is at most x2. It should be noted that one can choose a different CCDF value to infer the outcome of the probabilistic CNN based on decision maker's tolerance and perception of risk. For instance, the lower bound of CSP value will become smaller if user decides to increase the confidence by using a larger probability of exceedance value (e.g., 90%). Overall, these plots provide more complete representation of the perdition and the associated uncertainty, which can lead to a more informed and reliable AI-assisted post-disaster damage assessment by improving the communication between model predictions and decision making.

5.2.2 | Comparing model prediction uncertainty for various ground-truth classes

The purpose of this section is to analyze the degree by which the prediction uncertainty is affected by the actual incurred damage state as this somehow reflects the level of difficulty of the classification task. The human-based damage assessment of buildings in the DoriaNET dataset is relatively straightforward when buildings have minor damage or when the damage is obvious and severe. In contrast, for buildings with intermediate damage states, the annotator must spend more time cautiously observing visible features of buildings. Thus, the same behavior is expected when the trained models make damage state predictions. It can be fairly hypothesized that the model prediction uncertainty for more extreme ground-truth cases (i.e., levels D0 and D5) are lower compared to the cases that the ground truth belong to the intermediate classes (i.e., levels D1–D4). To examine this hypothesis, we group the buildings with the same ground truths and compute the average values of MCSSD and Entropy (H) obtained from model. These results are shown in Figure 12 for DoriaNET database. Applying a one-tailed z test with 90% confidence reveals that both average values of MCSSD and Entropy of damage level D0 (no damage) and D5 (destroyed) groups are statistically lower than the ones of damage level D1 through D4 groups. Applying a two-tailed z test with 90% confidence also reveals that there is no statistical difference between the average values of MCSSD and Entropy of damage level D0 (no damage) and D5

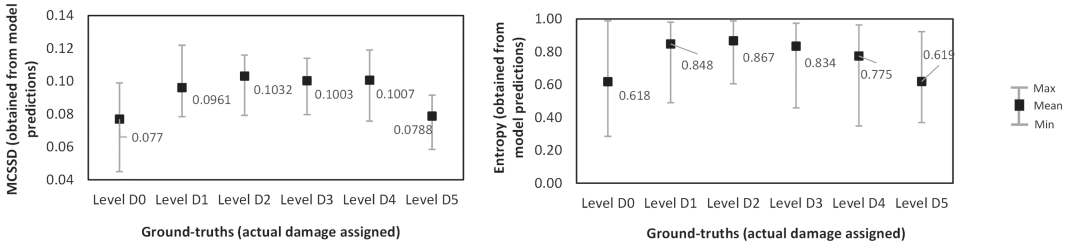

FIGURE 12 MCSSD and Entropy (H) of model predictions on buildings with various ground truths (DoriNET database)

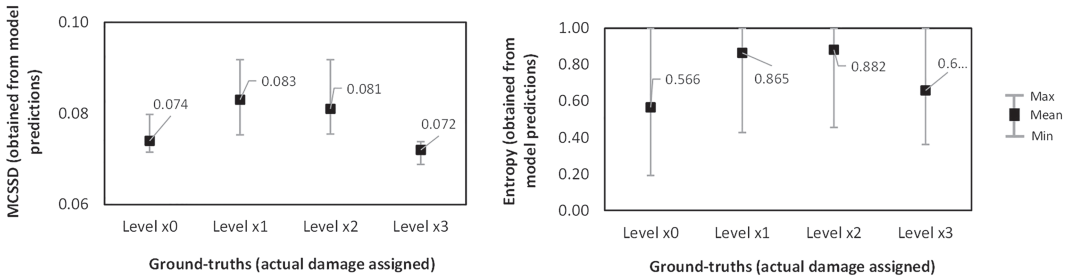

FIGURE 13 MCSSD and Entropy (H) of model predictions on buildings with various ground truths (xBD database)

(destroyed) groups. A similar pattern is also found in the xBD model in Figure 13. The average values of MCSSDs and Entropy of damage levels 0 (no damage) and 3 (destroyed) groups are statistically lower than the ones corresponding to damage levels x1 and x2 groups using a one-tailed z test with 90% confidence. These observations imply that model prediction behavior is consistent with how annotators label buildings for damage state by showing that the model has lower uncertainties when making predictions on the cases where the actual damage (ground truth) belongs to more extreme damage classes. It should be noted that the values of MCSSD and Entropy present an aggregated measure of epistemic uncertainty of predictions of the trained data-driven models. The high epistemic uncertainty might be either because of low quality of this input data point or could be due to the fact that the trained model has not seen this kind of data during the training, which can be improved by furthering the training dataset.

5.3 | Improving model explainability by visualizing CNN attention zone

The main goal of this study is to design reliable, uncertainty-informed, and ultimately, explainable neural networks for rapid and high-accuracy aerial post-disaster damage assessment. In the general AI domain, explainability can be defined as the degree to which humans can understand the model prediction or output.[15,61] Explainability is critical to decision-makers who use model outcomes to make impactful decisions. Beyond the benefits of understanding and trusting AI models, enhancing AI explainability provides opportunities to create responsible AI algorithms that consider the needs of the end users as well as sufficiently incorporate interpersonal and intergroup differences in the inference process.[62]

Towards achieving an explainable AI-assisted disaster damage assessment, this section explores adopting a course localization map, attention zone, that highlights the critical regions of the input image for model classification. This adoption improves the explainability of deep learning models since it explains why the classifiers make a particular prediction based on specific regions of the input images. In particular, we try to get insights into the specific regions of each aerial image that are the focus of the model when predicting damage levels. This can be achieved by employing class activation mapping (CAM) techniques to generate CNN heat maps highlighting the regions in input images that contain essential information for the model to make predictions. Specifically, in this work, we use gradient-weighted class activation mapping (Grad-CAM)[63] that enables extracting the gradients of any target concept from the final convolutional layer when processing the image. Figure 14 shows some sample results obtained using this technique. It is seen the trained CNN tends to classify the damage levels based on the roof conditions (e.g., roof cover damage and roof

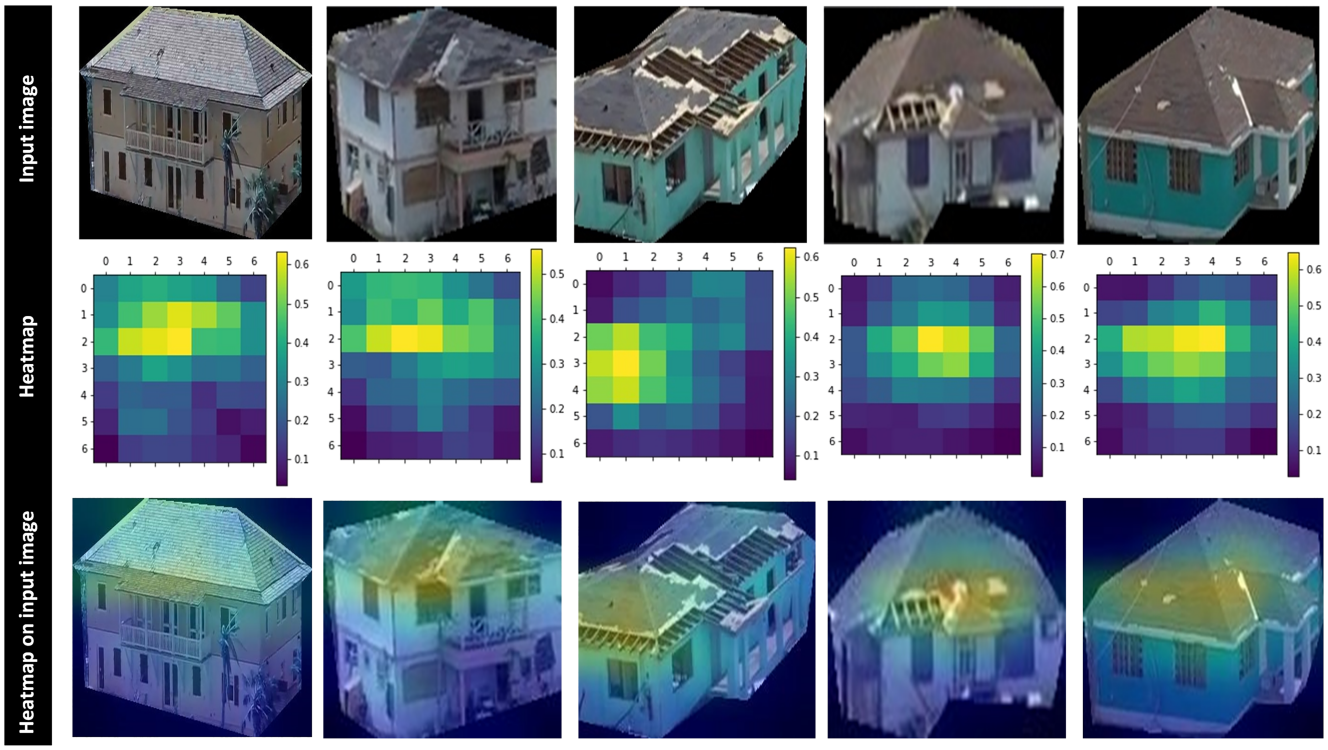

FIGURE 14 Attention zones corresponding to building input images in DoriaNET (generated by Grad-CAM)

deck damage). This observation is in agreement with the fact that this study focuses on wind-induced disaster damages where roofs are more vulnerable than other building components such as walls and windows.[64] It is also worth noting that this behavior is consistent with how the annotators labeled damage levels in the aerial building images in DoriaNET by first glancing at the building roof in each sample since, based on previous experiences, the majority of samples' damage states can be determined merely using the roof damage severity instead of looking into other structural components like windows and walls.

5.4 | Consistency of model predictions and crowdsourced annotation results

In DoriaNET, building damage severities were merely annotated by one annotator (one of the coauthors). While the DoriaNET annotation quality was monitored and well-controlled,[42] multiple humans may label the same building differently, particularly when the building condition does not exhibit a clear damage state, such as a common case where the damage appears to be in the range of several damage levels).[65] In order to explore this further, we design a small-scale experiment and conduct a comparative analysis of our uncertainty-aware CNN predictions and collective human decisions from multiple annotators. The experiment involves 20 human annotators who assign damage labels based on the FEMA rating system to five selected building images from DoriaNET. Similarly, we select 4 building images (pre- and post-disaster photos) from xBD and invite 33 human annotators to assign their damage severities. As shown in Figures 15 and 16, results from human annotations (the last column) clearly indicate the various opinions on the damage level determination from aerial imagery. Moreover, these figures generally show a strong consistency between CNN predictions and overall human decisions, evidenced in the last two columns. For example, considering the top building sample (the first row in Figure 15), human annotators have diverse decisions on the damage state between levels 3 (50%) and 4 (40%). Compared to humans, the trained CNN also exhibits a high uncertainty of choice between damage levels 3 and 4. This consistency between machine and human decisions is a promising indicator indicative of the fact that uncertainty-aware CNNs are informative and understandable. Moreover, this consistency helps decision-makers be more confident and willing to trust the model's predictive outcomes. The detailed exploration of this topic is beyond the scope of this paper and a potential direction for future research.

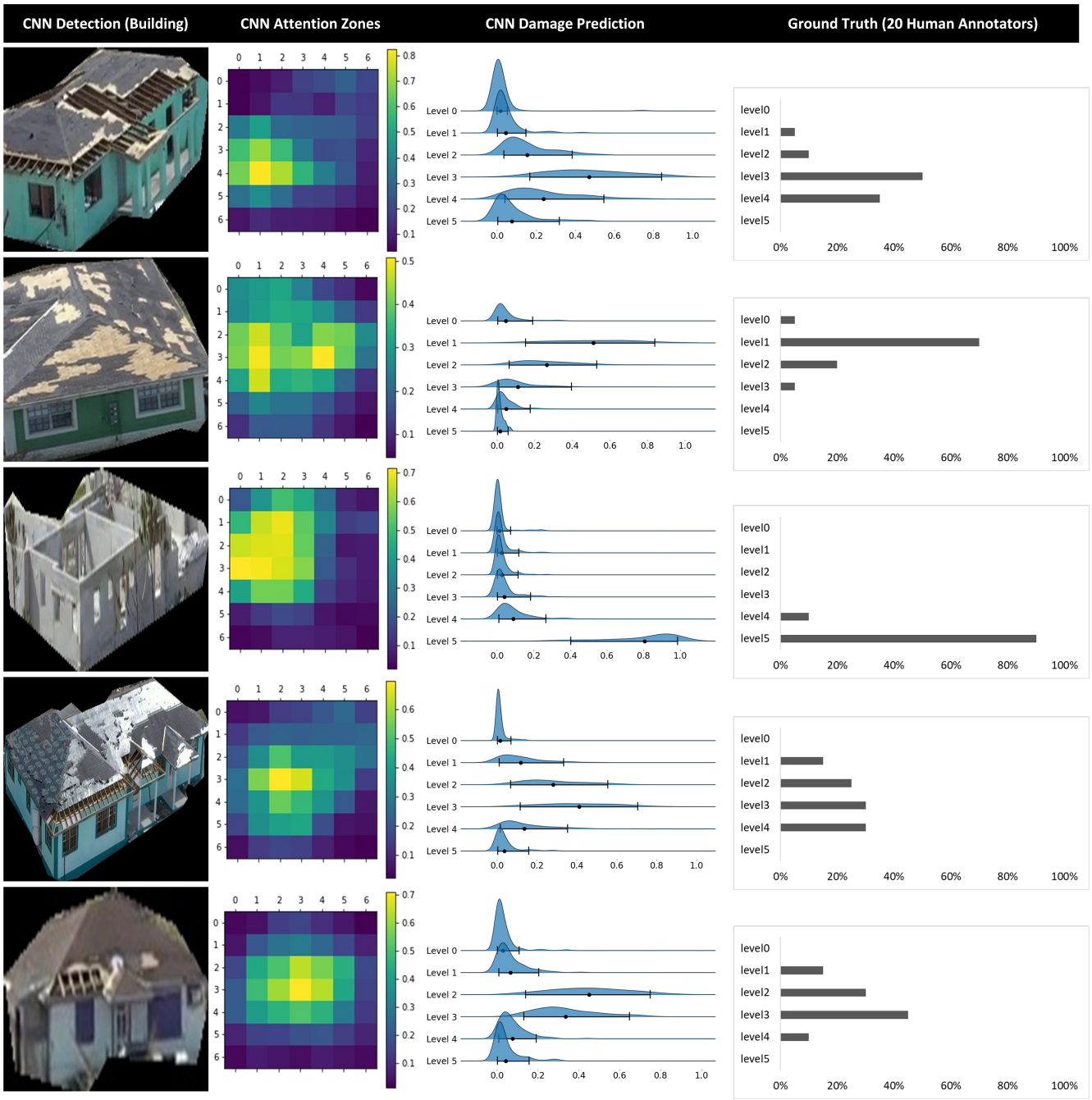

FIGURE 15 Comparison of CNN (trained on DoriaNET) predictions and human annotations

In the research domain of explainable AI (XAI), there is an ongoing debate on whether enhancing machine explainability deteriorates its accuracy. While some past research[62,66] point out a negative correlation between model explainability and prediction accuracy, Rudin[67] argues that enhancing machine explainability does not necessarily lose accuracy. Machines with a high level of explainability provide users with more insights on the cause and effect of the model and thus offer more leverage for improving the accuracy. Arrieta et al[62] review several Bayesian deep learning applications and argue that these Bayesian models can enable more rigorous quantification of prediction confidence and model reliability, which actually makes deep learning explainable. This study reported that the DoriaNET model achieves 59.4% accuracy, and the xBD model reaches 56.3% mean precision, 55.13% mean recall, and 54.2% mean F1 score. Compared with the baseline CNN models in the literature (without dropout activated at inference time), Cheng et al[10] reported a 60% accuracy when the model was tested on the DoriaNET dataset. Moreover, Gupta et al[14] provided a ResNet 50-based baseline CNN model with an overall 27% mean precision, 57% mean recall, and 32% mean F1 score

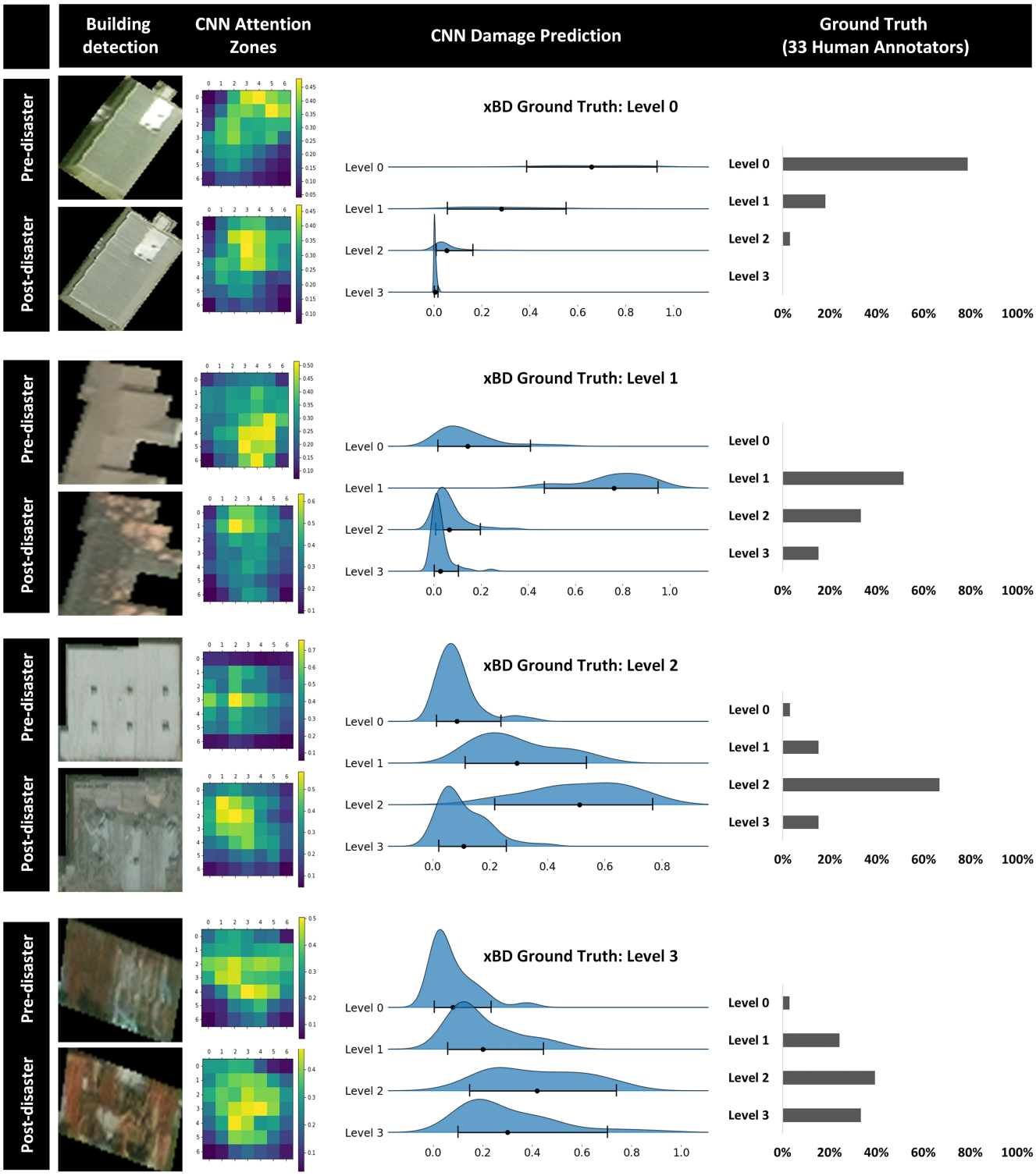

FIGURE 16 Comparison of CNN (trained on xBD) predictions and human annotations

on xBD dataset. It can be generally concluded that the uncertainty-aware CNNs presented in this paper are at least as accurate as the standard baseline models. This observation is generally in agreement with Rudin[67] by indicating that these CNNs do not need to sacrifice their prediction accuracies to produce additional reliability outputs. Also, it is worth noting that, due to the limited data, the model accuracy in this work appears to have an upper limit of approximately 60%. However, given the upward trend in the amount of publicly available disaster-related damage assessment data in recent years, further improvement of model performance is feasible through incorporating more heterogeneous and high-quality building data for model training. Moreover, the model can be developed through transfer learning on

a heavier neural network (e.g., ResNet-152[68]) with more neurons and layers that can extract more useful features to classify images. This would also potentially lead to better-performing CNN models for predicting building damage states.

The objective of developing uncertainty-aware damage assessment networks is not only to fully automate the post-disaster damage impact estimation but to enhance the better human-machine interaction by making machines more trustworthy, reliable, and informed. It should be, however, noted that this objective comes at a cost. A major challenge is the long inference time due to the MC sampling. For example, $N_{MC} = 30$ is selected to perform MC integration where the inference time is approximately 2.24 s per sample based on NVIDIA Quadro P2000 GPU where each input image needs to be processed by the network 30 times in order to generate 30 MC samples. This will indeed affect the real-time implementation of the CNN architecture. A practical remedy is to this challenge is to conduct neural network pruning by eliminating less useful layers, which will lighten the network and simultaneously increase the speed of model operation.[69,70]

6 | CONCLUSION

Accounting for uncertainty is particularly essential to deep learning applications related to critical decision making such as SHM. While past research has introduced several AI-assisted post-damage assessments leveraging aerial photos captured by UAVs or satellites, these techniques cannot inform the uncertainty of the prediction. This gap significantly increases the reluctance of human users to rely on machines to make critical decisions. The objective of this study was to develop uncertainty-aware deep neural networks for identification and classification of damage following a disaster. A Bayesian inference framework and MC dropout sampling techniques were used to propagate uncertainties in neural networks. The networks were developed and trained using image data from two aerial damage assessment datasets, namely, DoriaNET and xBD. DoriaNET consists of UAV-based post-disaster buildings with damage level annotation based on HAZUS and FEMA. xBD, on the other hand, contains worldwide pre- and post-disaster building satellite images with damage evaluations by humans. Using these two databases, two CNN models are developed and trained with dropout layers active both in training and inference phases. The xBD model, in particular, was designed as a dual-stream neural network that fuses pre and post-disaster image features and makes predictions. The performance of the trained models was evaluated on the testing (20%) set of the dataset, and it was reported that the DoriaNET model achieved 59.4% accuracy, and xBD model reached 55.1% accuracy with a 30 MC sample size. The stochastic CNN models developed in this work enable estimating the probabilistic descriptions of prediction outcomes leading to quantification of confidence in model predictions. Moreover, the correlation between the predicted uncertainly and the actual (ground truth) damage state is analyzed as well as the models' attention zone to gain insights about the explainability of the AI models. Consistency with human-generated labels are discussed by comparing the outcome with participatory human assessment. The observed consistency helps decision-makers to be more confident in the results of predictions and can pave the way to improve the existing human–AI partnership.

A potential direction for future work will be to improve the performance of uncertainty-aware models through the integration of multi-view building image predictions leveraging a 3D CNN architecture. With multi-view image entries, the trained classifiers are expected to achieve better performance as additional building visual features are provided from multiple angles.[71] We can also post-process the multi-view prediction outputs with Bayesian data fusion and produce a more reliable prediction based on this aggregated information.[21]

Finally, it is worth noting that the explored idea of creating uncertainty-aware deep learning models can also be applied to other computer vision-based SHM concerning, for instance, damage classification and crack segmentation.

ACKNOWLEDGMENT

The authors would like to thank Texas A&M University's High Performance Research Computing (HPRC) for providing necessary computing infrastructures 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.

AUTHOR CONTRIBUTIONS

Chih-Shen Cheng performed the conceptualization, methodology, writing–original draft, visualization, and data analysis. Amir H. Behzadan conducted the conceptualization, methodology, writing–review & editing, and supervision. Arash Noshadravan performed the conceptualization, methodology, writing-review & editing, and supervision.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in xView2 and DesignSafe-CI at https://xview2.org/dataset and https://doi.org/10.17603/ds2-gqvg-qx37, reference numbers 14 and 42, respectively.

ORCID

Arash Noshadravan https://orcid.org/0000-0001-6467-5689

REFERENCES

1. FEMA. Fema preliminary damage assessment guide. *Report*, FEMA; 2020. https://www.fema.gov/
2. Marshall J, Smith D, Lyda D, et al. Steer - hurricane dorian: Field assessment structural team (fast-1) early access reconnaissance report (earr). DesignSafe-CI; 2019.
3. Akter S, Wamba SF. Big data and disaster management: a systematic review and agenda for future research. *Ann Oper Res*. 2019;283(1):939-959. https://doi.org/10.1007/s10479-017-2584-2
4. Ghaffarian S, Kerle N, Pasolli E, Jokar Arsanjani J. Post-disaster building database updating using automated deep learning: An integration of pre-disaster openstreetmap and multi-temporal satellite data. *Remote Sensing*. 2019;11(20):2427.
5. Adams S, Friedland C, Levitan M. Unmanned aerial vehicle data acquisition for damage assessment in hurricane events. In: Proceedings of the 8th international workshop on remote sensing for disaster management, tokyo, japan, Vol. 30; 2010.
6. Bhowmick S, Nagarajaiah S, Veeraraghavan A. Vision and deep learning-based algorithms to detect and quantify cracks on concrete surfaces from uav videos. *Sensors*. 2020;20(21):6299.
7. Cao QD, Choe Y. Post-hurricane damage assessment using satellite imagery and geolocation features. arXiv preprint arXiv:08624; 2020.
8. Chen Z, Wagner M, Das J, Doe RK, Cerveny RS. Data-driven approaches for tornado damage estimation with unpiloted aerial systems. *Remote Sensing*. 2021;13(9):1669.
9. Nex F, Duarte D, Tonolo FG, Kerle N. Structural building damage detection with deep learning: assessment of a state-of-the-art cnn in operational conditions. *Remote Sensing*. 2019;11(23):2765.
10. Cheng C-S, Behzadan AH, Noshadravan A. Deep learning for post-hurricane aerial damage assessment of buildings. *Comput-Aided Civil Infrastruct Eng*. 2021;36(6):695-710.
11. Pi Y, Nath ND, Behzadan AH. Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. *Adv Eng Inf*. 2020;43:101009.
12. Presa-Reyes M, Chen S-C. 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); 2020:43-48.
13. Yan Y, Mao Z, Wu J, Padir T, Hajjar JF. Towards automated detection and quantification of concrete cracks using integrated images and lidar data from unmanned aerial vehicles. *Struct Control Health Monit*. 2021:e2757.
14. Gupta R, Hosfelt R, Sajeev S, et al. xbd: A dataset for assessing building damage from satellite imagery. ArXiv preprint arXiv:191109296; 2019.
15. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Mach Intell*. 2019;1(1):20-23.
16. NHTSA. Pe 16-007. *Report*, Technical report, U.S. Department of Transportation, National Highway Traffic Safety Administration, Jan 2017. Tesla Crash Preliminary Evaluation Report; 2017.
17. McAllister R, Gal Y, Kendall A, Van Der Wilk M, Shah A, Cipolla R, Weller A. Concrete problems for autonomous vehicle safety: advantages of bayesian deep learning. In: International joint conferences on artificial intelligence, inc.; 2017.
18. Bakator M, Radosav D. Deep learning and medical diagnosis: a review of literature. *Multimodal Technol Interact*. 2018;2(3):47.
19. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. *Ann Review Biomed Eng*. 2017;19:221-248.
20. Lindell MK, Prater CS. Assessing community impacts of natural disasters. *Assess Community Impacts Natural Disasters*. 2003;4(4):176-185.
21. Chen F-C, Jahanshahi MR. Nb-cnn: Deep learning-based crack detection using convolutional neural network and naïve bayes data fusion. *IEEE Trans Ind Electron*. 2017;65(5):4392-4400.
22. Modarres C, Astorga N, Droguett EL, Meruane V. Convolutional neural networks for automated damage recognition and damage type identification. *Struct Control Health Monit*. 2018;25(10):e2230.
23. Ni F, Zhang J, Chen Z. Pixel level crack delineation in images with convolutional feature fusion. *Struct Control Health Monit*. 2019;26(1):e2286.
24. Kim B, Cho S. Image based concrete crack assessment using mask and region based convolutional neural network. *Struct Control Health Monit*. 2019;26(8):e2381.
25. Jana D, Nagarajaiah S. Computer vision-based real-time cable tension estimation in dubrovnik cable-stayed bridge using moving hand-held video camera. *Struct Control Health Monit*. 2021;28(5):e2713.
26. Ghosh Mondal T, Jahanshahi MR, Wu R-T, Wu ZY. Deep learning-based multi-class damage detection for autonomous post-disaster reconnaissance. *Struct Control Health Monit*. 2020;27(4):e2507. https://doi.org/10.1002/stc.2507
27. Zhang Y-M, Wang H, Wan H-P, Mao J-X, Xu Y-C. Anomaly detection of structural health monitoring data using the maximum likelihood estimation-based bayesian dynamic linear model. *Struct Health Monit*. 2021;20(6):2936-2952.

28. Zou Z, Zhao P, Zhao X. Automatic segmentation, inpainting, and classification of defective patterns on ancient architecture using multiple deep learning algorithms. *Struct Control Health Monit*. 2021;28:e2742.
29. Sajedi S, Liang X. Deep generative bayesian optimization for sensor placement in structural health monitoring. Computer-Aided Civil and Infrastructure Engineering; 2021.
30. Hughes AJ, Bull LA, Gardner P, Barthorpe RJ, Dervilis N, Worden K. On risk-based active learning for structural health monitoring. *Mech Syst Sig Process*. 2022;167:108569.
31. Jana D, Patil J, Herkal S, Nagarajaiah S, Duenas-Osorio L. Cnn and convolutional autoencoder (cae) based real-time sensor fault detection, localization, and correction. *Mech Syst Signal Process*. 2022;169:108723.
32. Yeum CM, Dyke SJ, Ramirez J. Visual data classification in post-event building reconnaissance. *Eng Struct*. 2018;155:16-24.
33. Yeum CM, Lund A, Dyke SJ, Ramirez J. Automated recovery of structural drawing images collected from postdisaster reconnaissance. *J Comput Civil Eng*. 2019;33(1):04018056.
34. McClarren RG. *Uncertainty quantification and predictive computational science*: Springer; 2018.
35. Harper R, Southern J. A bayesian deep learning framework for end-to-end prediction of emotion from heartbeat. IEEE Transactions on Affective Computing; 2020.
36. Norouzi A, Emami A, Najarian K, Karimi N, Soroushmehr SMReza, et al. Exploiting uncertainty of deep neural networks for improving segmentation accuracy in mri images. In: Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp) IEEE; 2019:2322-2326.
37. Roy AG, Conjeti S, Navab N, Wachinger C, Initiative ADN, et al. Bayesian quicknat: model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage*. 2019;195:11-22.
38. Tousignant A, Lematre P, Precup D, Arnold DL, Arbel T. Prediction of disease progression in multiple sclerosis patients using deep learning analysis of mri data. In: International conference on medical imaging with deep learning; 2019:483-492.
39. Wang G, Li W, Aertsen M, Deprest J, Ourselin S, Vercauteren T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*. 2019;338:34-45.
40. Sajedi S, Liang X. Uncertainty assisted deep vision structural health monitoring. *Comput-Aided Civil Infrastruct Eng*. 2020;36(2):126-142.
41. Vashisht R, Viji H, Sundararajan T, Mohankumar D, Sumitra S. Structural health monitoring of cantilever beam, a case study using bayesian neural network and deep learning. *Structural integrity assessment*: Springer; 2020:749-761.
42. Cheng CS, Behzadan AH, Noshadravan A. Dorianet: A visual dataset from hurricane dorian for post-disaster building damage assessment). 2021. DesignSafe-CI.
43. FEMA. Multi-hazard loss estimation methodology: Hurricane model hazus-mh mr3, technical manual. *Report*, FEMA; 2003. https://www.hsdl.org/
44. Kelman I. Physical flood vulnerability of residential properties in coastal, eastern england. *Thesis*; 2003.
45. Grnthal G. European macroseismic scale 1998. *Report*, European Seismological Commission (ESC); 1998. http://lib.riskreductionafrica.org/
46. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, Fieguth P, Khosravi A, Acharya UR, Makarenkov V. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. arXiv preprint arXiv:06225; 2020.
47. Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. PMLR; 2016:1050-1059.
48. Graves A. Practical variational inference for neural networks. In: Advances in neural information processing systems. Citeseer; 2011: 2348-2356.
49. Kullback S, Leibler RA. On information and sufficiency. *The Ann Math Stat*. 1951;22(1):79-86.
50. Chollet F. *Deep Learning with Python*: Simon and Schuster; 2021.
51. Kendall A, Gal Y. What uncertainties do we need in bayesian deep learning for computer vision? 2017.
52. Shannon CE. A mathematical theory of communication. *The Bell Syst Tech J*. 1948;27(3):379-423.
53. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. ArXiv preprint arXiv:170404861; 2017.
54. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst*. 2012;25:1097-1105.
55. Levina E, Bickel P. The earth mover's distance is the mallows distance: some insights from statistics. In: Proceedings eighth ieee international conference on computer vision. iccv 2001, Vol. 2 IEEE; 2001:251-256.
56. Hou L, Yu C-P, Samaras D. Squared earth mover's distance-based loss for training deep neural networks. ArXiv preprint arXiv: 161105916; 2016.
57. Kailath T. The divergence and bhattacharyya distance measures in signal selection. *IEEE Trans Commun Technol*. 1967;15(1):52-60.
58. Kingma DP, Ba J. Adam: A method for stochastic optimization. ArXiv preprint arXiv:14126980; 2014.
59. Lee S, Lee C. Revisiting spatial dropout for regularizing convolutional neural networks. *Multimed Tools Appl*. 2020;79(45):34195-34207.
60. Ko B, Kim H-G, Oh K-J, Choi H-J. Controlled dropout: a different approach to using dropout on deep neural network. In: 2017 ieee international conference on big data and smart computing (bigcomp); 2017:358-362.
61. Chen Z, Bei Y, Rudin C. Concept whitening for interpretable image recognition. *Nature Mach Intell*. 2020;2(12):772-782.
62. Arrieta AB, Daz-Rodrguez N, Del Ser J, et al. Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai. *Inf Fusion*. 2020;58:82-115.

63. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the ieee international conference on computer vision; 2017:618-626.
64. Aghababaei M, Koliou M, Paal SG. Performance assessment of building infrastructure impacted by the 2017 hurricane harvey in the port aransas region. *J Perfor Construct Facil*. 2018;32(5):04018069.
65. Khajwal AB, Noshadravan A. An uncertainty-aware framework for reliable disaster damage assessment via crowdsourcing. *Int J Disast Risk Reduct*. 2021;55:102110.
66. Gunning D. Broad agency announcement explainable artificial intelligence (xai). *Report*. DARPA-BAA-16-53, DARPA; 2016.
67. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Mach Intell*. 2019;1(5):206-215.
68. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the ieee conference on computer vision and pattern recognition; 2016:770-778.
69. Nath N, Behzadan AH. Deep generative adversarial network to enhance image quality for fast object detection in construction sites. In: 2020 winter simulation conference (wsc); 2020:2447-2459.
70. Wu RT, Singla A, Jahanshahi MR, Bertino E, Ko BJ, Verma D. Pruning deep convolutional neural networks for efficient edge computing in condition assessment of infrastructures. *Comput-Aided Civil Infrastruct Eng*. 2019;34(9):774-789.
71. Seeland M, Mder P. Multi-view classification with convolutional neural networks. *Plos one*. 2021;16(1):e0245230.

How to cite this article: Cheng C-S, Behzadan AH, Noshadravan A. Uncertainty-aware convolutional neural network for explainable artificial intelligence-assisted disaster damage assessment. *Struct Control Health Monit*. 2022;29(10):e3019. doi:10.1002/stc.3019