A Novel Weighted Ensemble Transferred U-Net based Model (WETUM) for Post-Earthquake Building Damage Assessment from UAV Data: A Comparison of Deep Learning- and Machine Learning-based Approaches

Ehsan Khankeshizadeh, Ali Mohammadzadeh, Hossein Arefi, Amin Mohsenifar, Saied Pirasteh, *Senior Member IEEE,* En Fan, Huxiong Li, Jonathan Li, *IEEE Fellow*

***Abstract*— Nowadays, unmanned aerial vehicle (UAV) remote sensing data are key operational sources used to produce a reliable building damage map (BDM), which is of great importance in instant response and rescue operations after earthquakes. The present study proposes a novel weighted ensemble transferred U-Net-based model (WETUM) consisting of two major steps to create a reliable binary BDM using UAV data. In the first step of the proposed approach, three individual initial BDMs are predicted by three pre-trained U-Net-based composite networks. In the second step, these three individual predictions are linearly integrated through a proposed grid search technique so that an optimized hybrid BDM (OHBDM) incorporating complementary damage information is made. The proposed WETUM was then compared with several conventional deep learning (DL) and machine learning (ML) models. The models were compared across two pivotal scenarios, addressing the impact of diverse feature sets on model performance and generalizability. Specifically, the first scenario focused solely on spectral features, while the second incorporated both spectral and geometrical features. To make the comparisons, this study conducted empirical analyses using UAV spectral and geometrical data acquired over Sarpol-e Zahab, Iran. The experimental findings showed that the synergic use of spectral and geometrical data boosted both DL- and ML-based approaches in damage detection. Moreover, the proposed WETUM with DDR values of 65.22 and 78.26 (%), respectively, for the first and second scenarios, outperformed all the compared methods. Notably, WETUM with only spectral data outperformed the random forest (RF) classifier equipped with many hand-crafted spectral and geometrical features, indicating the highest potential and generalizability of the proposed WETUM for building damage evaluation in a new unseen earthquake-affected area.**

***Index Terms*— Building damage map, deep transfer learning, UAV data, U-Net, machine learning, spectral and geometrical features.**

I. INTRODUCTION

NATURAL disasters, including wildfires, floods, cyclones, landslides, and earthquakes, continuously pose threats to the world [1], [2]. Of these, the most catastrophic is the earthquake, causing many fatalities and severe damage to buildings within urban environments worldwide [3], [4]. Hence, diagnosing the earthquake-induced damages to these areas is crucial in immediate response and rescue operations. Building damage assessment relying on the field survey is costly and time-consuming and is nearly impossible in some cases due to road closures within urban regions after earthquakes. Lately, remote sensing (RS) technology and data sources, due to their synoptic view, have been broadly used in an operational way to produce the building damage map (BDM). Overall, the existing BDM generation approaches generally use two main RS data sources, including satellite imagery [5]–[8] and unmanned aerial vehicle (UAV) data. Particularly, UAVs, also called drones, play a vital role in post-earthquake rapid damage evaluation due to their capability to cover inaccessible areas and real-time and high-resolution

(*Corresponding authors*: spirasteh71@gmail.com; a_mohammadzadeh@kntu.ac.ir; efan@usx.edu.cn, jsj_lhx@126.com)

E. Khankeshizadeh and A. Mohammadzadeh are remote researcher, working jointly with S. Pirasteh, En Fan, and Huxiong Li, and are with the Institute of Artificial Intelligence, Shaoxing University, 508 West Huancheng Road, Yuecheng District, Shaoxing, Zhejiang Province, Postal Code 312000, China (email: spirasteh71@gmail.com; sapirasteh1@usx.edu.cn; efan@usx.edu.cn, jsj_lhx@126.com). In addition, E. Khankeshizadeh, A. Mohammadzadeh and A. Mohsenifar are with the Department of Photogrammetry and Remote Sensing, Geomatics Engineering Faculty, K. N. Toosi University of Technology, Tehran 15433-19967, Iran (e-mail:Seyedehsan.Khankeshizadeh@email.kntu.ac.ir;a_mohammadzadeh@kntu.ac.ir; a.mohsenifar@email.kntu.ac.ir). Also, S. Pirasteh is an adjunct faculty with the Department of Geotechnics and Geomatics, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India, 602105.

H. Arefi is with the School of Surveying and Geospatial Engineering, University of Tehran, Tehran 1417466191, Iran, and the Hochschule Mainz - University of Applied Sciences, Mainz 55128, Germany (e-mail: hossein.arefi@hs-mainz.de).

J. Li is a professor of Geomatics and LiDAR at University of Waterloo, Canada (email:junli@uwaterloo.ca)

data acquisition. In general, drone data-based building damage assessment approaches could be investigated from two time and methodological viewpoints. From the time perspective of the used drone data, previous methods employed either a uni-temporal post-event image [9]–[11] or bi-temporal pre- and post-event images [12], [13]. Using a single post-event drone data is more practical and preferred over applying bi-temporal images as the latter contains a number of challenges, such as shadow artifacts, geometrical displacements, and co-registration issues. Moreover, access to an appropriate pre-event image for bi-temporal damage evaluation is not always guaranteed [14].

In terms of the methodology employed, two major categories for producing BDM in previous drone data-based works are (1) Machine Learning (ML) and (2) Deep Learning (DL)-based methods. Two main steps are overall pursued in the ML approaches. Firstly, hand-crafted features are extracted from input data, and secondly, the extracted features are classified into different damage categories by a classifier. ML methods need more accurate manual/expert-dependent feature engineering in which the features and ML algorithms must be adjusted before classification, which may differ in various scenarios, applications, and geographical areas [15]. In recent years, DL-based methods have substantially addressed the aforementioned challenges of the ML algorithms in damage evaluation works [16]–[18]. Contrary to ML, automatically extracting and classifying many complex abstract features are coincidently carried out in DL models [19]. Among the existing DL structures used for damage assessment purposes, convolutional neural networks (CNNs) have been proven to be more effective and accurate in BDM generation due to their automatic feature extraction allowing for the effective extraction of multi-level features of buildings, such as colors, edges, and problem-specific features [20], [21]. Notably, amongst CNN-based structures, U-Net architecture [22] as an appropriate choice has attracted much attention in the context of building damage detection [23]–[28]. Because the U-Net structure can work efficiently even with a small amount of labeled data [22], which is especially suited to building damage assessment where annotated data is challenging to be obtained. Yet, the lack of a strong feature extractor mechanism within U-Net for capturing complex image features may lead to losing accurate earthquake damage information. To fill this gap, a very limited number of studies employed some CNN-based architectures as backbones added to the simple U-Net model to generate more abstract damage features. For instance, Irwansyah et al. [29] embedded the ResNet-50 backbone in the U-Net model and trained the composite structure using xBD dataset. They finally employed the trained model to generate BDMs. On the other hand, the pre-trained backbones, such as deep transfer learning (DTL) techniques, can help U-Net generalize well in a new earthquake scenario where the training data might be small and insufficient [30]. In fact, DTL is a strategy that involves pre-training deep neural networks on large datasets and then fine-tuning them on related smaller ones to improve deep models on smaller unseen datasets. In this way, transferred deep models can generalize to new scenarios through learning more general features that are applicable across new datasets [31]–[33]. Lately, coupling pre-trained backbones with the U-Net structure, which is regarded as composite deep models throughout the present paper, has already been found effective in medical image processing fields for raising the generalizability [34]–[36], but further research works are also to be carried out to peruse the efficiency of DTL in the building damage detection context.

Though the aforementioned and previous earthquake-related damage assessment studies led to satisfactory results, they still face some remarkable issues and gaps that should be well tackled. Firstly, none of the existing U-Net-based DL building damage detection methods have employed backbone structures and DTL techniques simultaneously, and they also lack utilizing ensemble predictions derived from several transferred models to evaluate building damages. Considering any individual pre-trained backbone can reflect some particular damage information in its prediction, integrating the predictions of several transferred models is assumed to reflect and accumulate distinct aspects of building damage information within an integrated individual prediction. Secondly, to the best of the authors' knowledge, the earthquake-induced building damage assessment literature lacks a comprehensive comparison between ML- and DL-based approaches in terms of input data (features) combination and generalizability. From the input features combination perspective, none of the previous works specifically evaluated the impact of spectral features (i.e., RGB bands) and their combinations with geometrical features such as digital surface model (DSM) on ML- and DL-derived BDMs. Notably, using DSM data as 3D geometrical information alongside spectral data can better reflect the physical properties of the buildings [37]. Moreover, in terms of generalizability, a comparison between ML and DL approaches for a new earthquake-related damaged area remains unexplored, where test and training data lack overlap. To overcome the issues and gaps mentioned for the previous efforts in the building damage examination field, a novel ensemble DTL algorithm was proposed here, and a detailed comparison among the proposed model and several ML- and DL-based approaches was conducted with several spectral and geometrical input features. Indeed, the present study has two major research significance from the viewpoints of data and methodology. Firstly, regarding data, the use of UAV remote sensing data is highlighted as a key operational source for generating reliable BDMs despite satellite data. This underscores the practical relevance of the study in the context of instant response and rescue operations after earthquakes. Secondly, the methodological significance involves introducing a novel DL-based architecture, analyzing the synergic integration of spectral and geometrical features, and comprehensively comparing the proposed DL approach and the state-of-the-art algorithms. The primary emphasis in this research is on localizing damaged buildings rather than the evaluation of the extent and severity of damage. Hence, the main contributions and methodological novelties of the present

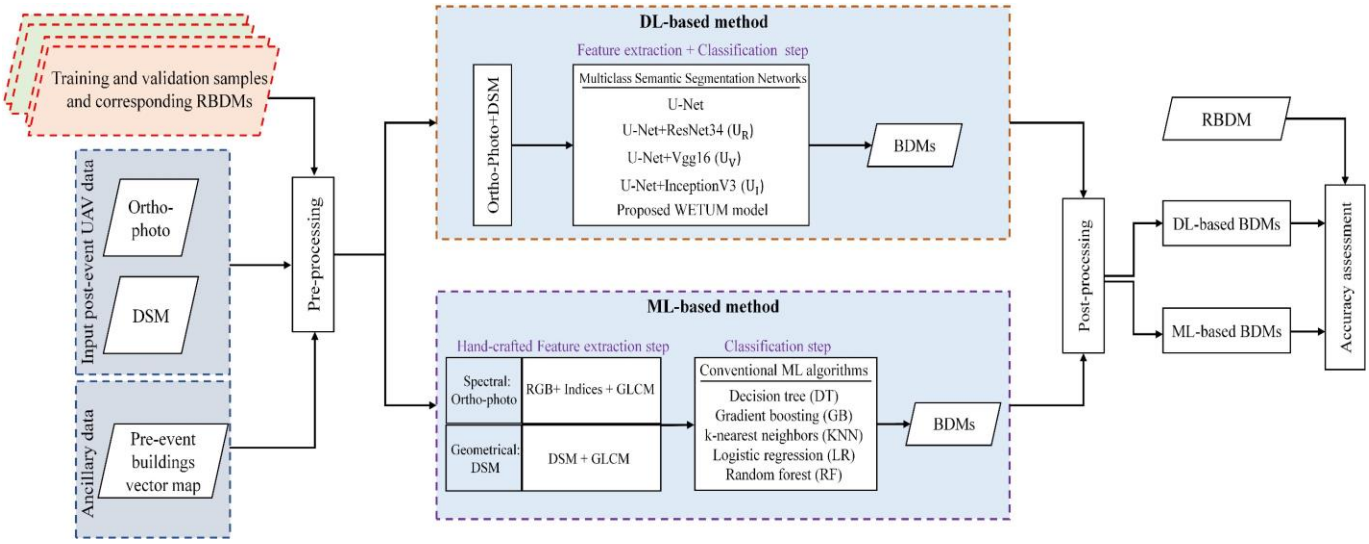

Fig. 1. The general comparison framework of the present study for producing ML- and DL-based BDMs, where BDM and RBDM are abbreviations for *building damage map* and *reference building damage map*, respectively.

study are summarized as follows:

1) A novel weighted ensemble transferred U-Net-based model (WETUM) with three various pre-trained backbone structures alongside a reliable integration approach is proposed to take advantage of the predictions of several models for generating an optimized BDM.

2) Various combinations of spectral and geometrical features as inputs given to both ML- and DL-based models are evaluated and compared to investigate their impact on recognizing damage/non-damage.

3) For the first time, both ML- and DL-based BDM generation approaches are compared in generalizability level for a new unseen earthquake-affected area non-overlapped with the training areas.

The remainder of this paper is presented as follows: Section II details the proposed DL-based WETUM and explains the comparative ML-based approaches. In Section III, the study area and datasets chosen for this work are introduced. The experiments are detailed, and the assessment criteria considered for evaluating the results attained in this research are also described in Section IV. Then, the analysis and discussion of the quantitative and qualitative results for both DL- and ML-based models are presented in Section V. Lastly, Section VI summarizes the key findings drawn from the results and concludes the paper.

II. Methodology

The general framework of the present comparative building damage detection study with four main steps is shown in Fig. 1: (1) pre-processing of the used drone data, (2) developing a novel DTL-based approach named WETUM based on optical and DSM data for binary BDM generation, (3) implementing various ML algorithms trained by different feature combinations, including optical, DSM and other hand-crafted features to produce binary BDMs, and (4) post-processing of the BDMs obtained by the DL and ML algorithms.

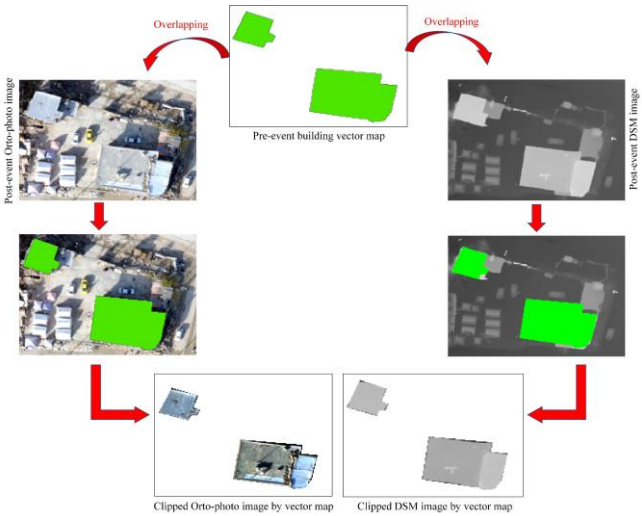

Fig. 2. Removing the non-building pixels from the input data in the pre-processing phase.

A. Pre-processing

The pre-processing of the data used in this study falls into two different stages: (1) non-building pixels removal for both ML- and DL-based methods and (2) patch extraction and data augmentation for DL-based approaches. At the first stage, in order to reduce the negative effects of non-building pixels on all training, validation and testing procedures in both ML and DL approaches, the building vector layer as one of the input data was overlaid on both input optical and DSM data to

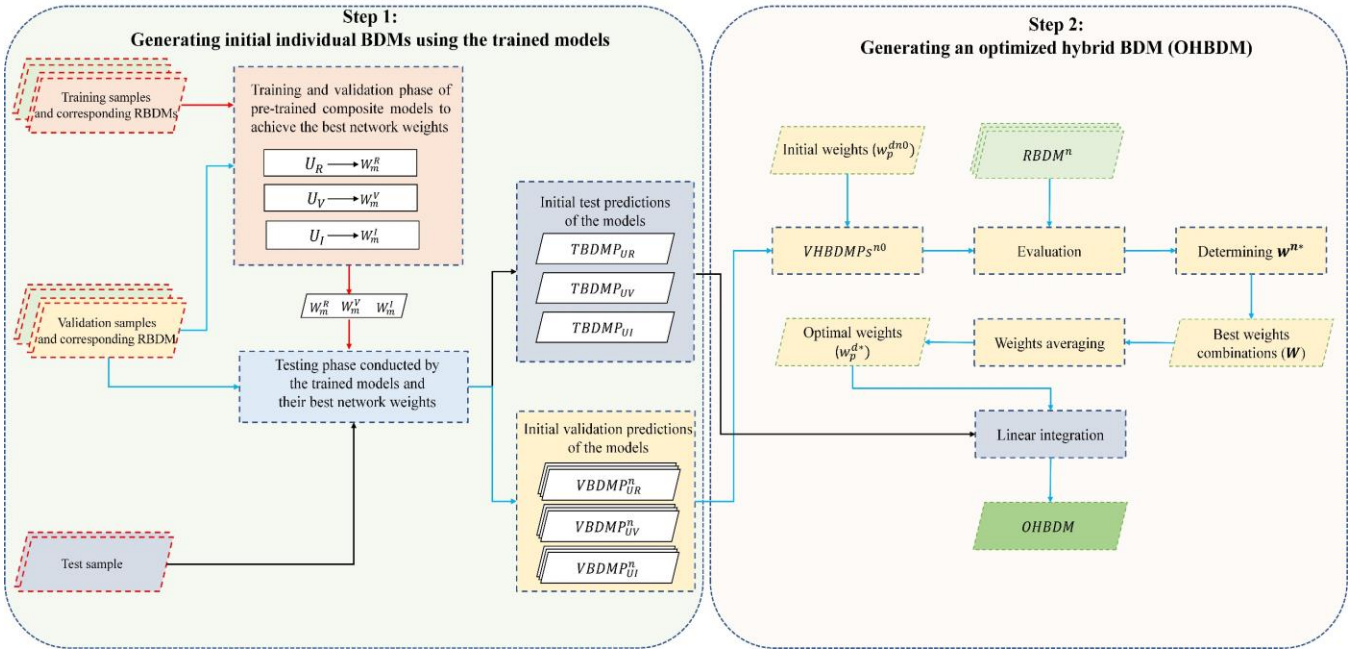

Fig. 4. The flowchart of the proposed weighted ensemble transferred U-Net-based model (WETUM) for generating a reliable BDM.

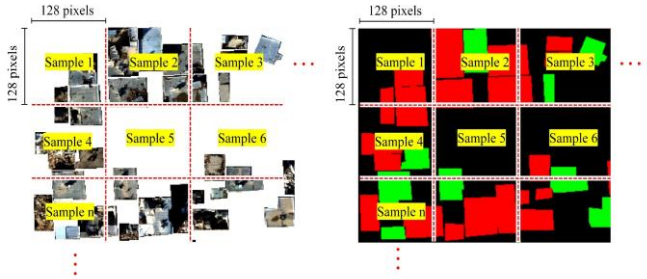

Fig. 3. Illustrating the patch extraction method used to prepare adequate datasets for deep learning (DL) algorithms, in which the input and their corresponding output (label) patches are respectively shown at the left and right sides.

separate the built-up and non-built-up areas as depicted in Fig. 2. The building vector map as an ancillary data, due to the lack of updated vector data for the Sarpol-e Zahab region, were manually prepared by precisely visualizing the corresponding pre-event Google Earth images. It is worth nothing that the presumption in the proposed methodology is that the vector data is of trusted quality, even though there might be instances of data being outdated or unavailable – circumstances irrelevant to the functionality of the developed algorithm. However, due to the removal of non-building objects that may negatively affect damage detection, this pre-processing step raises the reliability of detecting the damaged buildings, as only the building pixels are inspected, and the irrelevant background objects are ignored.

In the second stage, to convert the input data used for DL models to a correct format (i.e., a four-dimensional array containing the number of samples, rows, columns, and channels) suited to the Keras Python library [38] and to increase the amount of both training and validation data, all training, validation, and testing data and their corresponding labels were partitioned into nonoverlapping square patches of size 128×128 pixels (see Fig. 3). Additionally, the patches were next augmented with rotation with an angle of 90˚ and left/right and up/down flipping to sufficiently provide training, validation, and testing datasets were derived as (1148, 128, 128, n), (341, 128, 128, n), and (21, 128, 128, n), respectively, where n represents the number of input data bands, which is 3 (red, green, and blue bands) and 4 (red, green, blue, and DSM bands) for ortho-photo and stacked ortho-photo and DSM data, respectively.

B. DL-based WETUM

Fig. 4 depicts the 2-step architecture of the proposed weighted ensembled approach for the generation of BDM from input RGB and DSM images. In the first step, three individual initial probabilistic BDMs are produced using three different multi-class semantic segmentation pre-trained models. In the second step, to make an optimized ensembled BDM prediction from these individual predictions, they are linearly integrated by three different weights determined using a grid search combination strategy. Each step mentioned above is detailed in the following paragraphs.

1) Generating initial individual BDMs using the trained models: Even though the traditional U-Net model can achieve satisfactory outcomes in building damage recognition, it suffers from the existence of a limited number of convolutional layers as feature extractors in its architecture, leading to damage and non-damage information loss. To cope with this drawback, embedding useful backbones to the U-Net architecture can increase its potential to extract meaningful and discriminative features from the input image to precisely detect damage/non-

Algorithm 1: The Pseudo Code of the Proposed Grid Search Technique for Estimating the Best Integration Weights

Inputs:

$\boldsymbol{N}$ (Number of validation patches; n=1, … , N)

$\boldsymbol{VBDMP_{UR}^{n}}$, $\boldsymbol{VBDMP_{UV}^{n}}$, and $\boldsymbol{VBDMP_{UI}^{n}}$ (Individual VBDM patches respectively derived by U_R, U_V, and U_I for the n[th] validation patch)

$\boldsymbol{RBDMP^{n}}$ (Reference BDM patch for the n[th] validation patch)

Outputs:

$\boldsymbol{w_p^{R*}}$, $\boldsymbol{w_p^{V*}}$, and $\boldsymbol{w_p^{I*}}$ (The most optimal weight values of w_p^{Rn0}, w_p^{Vn0}, and w_p^{In0})

1: **for** n: = 1 to N, step 1 do
(Loop 1 on N)
for w_p^{Rn0} := 0 to 1, step 0.1 **do**
(Loop 2 on the number of candidate values for w_p^{Rn0})
for w_p^{Vn0}:= 0 to 1, step 0.1 **do**
(Loop 3 on the number of candidate values for w_p^{Vn0})
for w_p^{In0}:= 0 to 1, step 0.1 **do**
(Loop 4 on the number of candidate values for w_p^{In0})
if $w_p^{Rn0}+w_p^{Vn0}+w_p^{In0} = 1$ (Condition 1)
$\boldsymbol{w^{n0}}$=[w_p^{Rn0}, w_p^{Vn0}, w_p^{In0}];
$\boldsymbol{wv^{n0}} \leftarrow \boldsymbol{w^{n0}}$;

$HBDMP^{n0}$=$w_p^{Rn0}\times VBDMP_{UR}^{n}+w_p^{Vn0}\times VBDMP_{UV}^{n}+w_p^{In0}\times VBDMP_{UI}^{n}$;
k^{n0} =AccEval ($RBDMP^{n}$, $HBDMP^{n0}$);
where AccEval(.) is a function taking $RBDMP^{n}$and $HBDMP^{n0}$ to calculate kappa for the n[th] validation patch, i.e., k^{n0}.
$\boldsymbol{KV^{n0}} \leftarrow k^{n0}$;
where $\boldsymbol{KV^{n0}}$ is a vector in which the calculated k^{n0} are collected.
ID^{n*}=argmax (KV^{n0});
where argmax(.) is an operator returning the ID of the highest kappa value in KV^{n0}.
$\boldsymbol{w^{n*}}$=$\boldsymbol{wv^{n0}}$[ID^{n*}]; where $\boldsymbol{w^{n*}}$= [w_p^{Rn*}, w_p^{Vn*}, w_p^{In*}];
$\boldsymbol{W} \leftarrow \boldsymbol{w^{n*}}$;
where w^{n*} is the best form of w_p^{Rn0}, w_p^{Vn0}, and w_p^{In0} for the n[th] validation patch.
end if (End condition 1)
end for (End loop 4)
end for (End loop 3)
end for (End loop 2)
end for (End loop 1)
2: w_p^{R*}=$avg(\boldsymbol{W}[:,1])$;
3: w_p^{V*}= $avg(\boldsymbol{W}[:,2])$;
4: w_p^{I*}= $avg(\boldsymbol{W}[:,3])$;
where $\boldsymbol{W}[:,1]$, $\boldsymbol{W}[:,2]$, and $\boldsymbol{W}[:,3]$ denote the first, second, and third columns of $\boldsymbol{W}$, respectively.
5: return w_p^{R*}, w_p^{V*}, and w_p^{I*}

damage categories. Moreover, using the pre-trained variant of a backbone can help U-Net generalize better to new unseen data as a pre-trained backbone is already trained on large datasets with diverse images. On the other hand, since pre-trained backbone weights already contain beneficial low-level damage/non-damage building features like edges and textures, adding a pre-trained backbone to U-Net and fine-tuning it on damage detection task can enable U-Net to exploit the learned features to result in faster convergence and improved BDMs. Considering each backbone can capture some unique multi-level abstract features, the use of several different pre-trained backbones could reflect a broad range of feature sets, each of which could contain distinct particular aspects of damage/non-damage information. Hence, three well-known CNN-based backbones, including ResNet-34 [39], Vgg16 [40], and InceptionV3 [41], respectively abbreviated as R, V, and I, were considered in this research work. The backbones R, V, and I were pre-trained on the ImageNet database [42], a large-scale dataset of labeled images, which has been widely used to pre-train DL networks as an efficient transfer learning technique. The three pre-trained backbones were separately embedded into U-Net in order to form three different composite DL structures, namely U-Net+ResNet34 (U_R), U-Net+Vgg16 (U_V) and U-Net+InceptionV3 (U_I) within the Python Keras library. U_R, U_V, and U_I models were then fine-tuned by the available training and validation patches to reach the best network training weights (i.e., w_m^R, w_m^V, and w_m^I) for the composite models. After the training phase, the validation data with N patches (i.e., n=1, … , N) and the test data patches were given to the trained U_R, U_V, and U_I models to produce (1) validation data-based BDM patches ($VBDMP_{UR}^{n}$, $VBDMP_{UV}^{n}$, and $VBDMP_{UI}^{n}$) and (2) test data-based BDM patches ($TBDMP_{UR}$, $TBDMP_{UV}$, and $TBDMP_{UI}$), respectively. These two categories of initial single BDM results were then kept for the subsequent analyses.

2) Generating an optimized hybrid BDM (OHBDM): Each of the previously produced $TBDMP_{UR}$, $TBDMP_{UV}$, and $TBDMP_{UI}$ reflects unique damage/non-damage information as three different composite models produced them. Accordingly, their integration as a key solution to the damage detection problem can provide complementary valuable information due to accentuating the merits of the individual BDM results. To this end, a linear integration consisting of three unique weights and the predictions of U_R, U_V, and U_I was proposed in this study to produce an integrated individual prediction named optimized hybrid BDM (OHBDM). The optimal weight values of this linear integration were estimated through a proposed validation patches-based grid search approach, detailed in the following.

Assuming w_p^{Rn0}, w_p^{Vn0}, and $w_p^{In0}$are respectively the initial weights associated with $VBDMP_{UR}^{n}$, $VBDMP_{UV}^{n}$, and $VBDMP_{UI}^{n}$, where each weight ranges from 0 to 1 with a step of 0.1, and n also denotes the number of validation patches as already mentioned. In the proposed grid search approach, for the n[th] validation patch, many possible linear combinations according to the (1) were first created between the weights and the $VBDMP_{UR}^{n}$, $VBDMP_{UV}^{n}$, and $VBDMP_{UI}^{n}$ to generate initial BDM patches ($RBDMP^{n}$) to determine the best linear combination ($HBDMP^{n*}$) and its corresponding weights ($\boldsymbol{w^{n*}}$= [w_p^{Rn*}, w_p^{Vn*}, w_p^{In*}]).

$$HBDMPs^{n0} = \sum_d VBDMP_{Ud}^{n} \times w_p^{dn0};$$
$$\sum_d w_p^{dn0} = 1,\ d: R, V, I \text{ and } n=1, \dots, N \quad (1)$$

TABLE I
TEXTURAL AND OPTICAL HAND-CRAFTED FEATURES USED FOR ML ALGORITHMS

	Feature	Description	Equation
GLCM	Mean	-Sensitive to variations in surface roughness, cracks, or other textural changes that may indicate structural damage.	$\sum_{i=1}^{m}\sum_{j=1}^{n} j \times GLCM(i,j)$
	Variance	-Highlighting high damage potential by detecting pixel intensity variations due to changes in texture caused by cracks, debris, or structural alterations.	$\sum_{i=1}^{m}\sum_{j=1}^{n} (j-\mu)^2 \times GLCM(i,j)$
	Homogeneity	-Noise-resistant damage detection due to quantifying the local similarity of neighboring pixels. -Useful when dealing with complex scenes.	$\sum_{i=1}^{m}\sum_{j=1}^{n} \frac{GLCM(i,j)}{1+\lvert i-j\rvert}$
Optical index	PAN	-Identifying small cracks or structural. -Enhancing the contrast between damaged and undamaged buildings. -Detecting severe building damages due to multi-spectral analysis.	$0.2989 \times \text{Red} + 0.587 \times \text{Green} + 0.114 \times \text{Blue}$

After the best weights ($\boldsymbol{w^{n*}}$) determination for all N validation patches, all the w_p^{Rn*}, w_p^{Vn*}, and w_p^{In*} were averaged on N to estimate the most optimal weights, namely w_p^{R*}, w_p^{V*}, and w_p^{I*}. The further details of the proposed grid search algorithm are also well represented in Algorithm 1. Once the validation data patches led to the ideal weights, the reliable OHBDM patches (OHBDMPs) for the test data patches were achieved by linearly integrating $TBDMP_{UR}$, $TBDMP_{UV}$, and $TBDMP_{UI}$ produced in the previous section and the optimal weights as formulated in (2). Eventually, OHBDMPs were connected horizontally and vertically to form the final entire OHBDM. It's worth noting the training and testing processes in this work were performed within the Google Colaboratory Pro Python environment with 16 GB GPU (i.e., NVIDIA Tesla V100) and 25 GB of RAM.

$$OHBDMPs = \sum_d TBDMP_{Ud} \times w_p^{d*};\ d: R,V,I \qquad (2)$$

C. ML-based methods

In the present study, ML-based methods are implemented in two stages. Firstly, some hand-crafted ancillary features were extracted, including the grey level co-occurrence matrix (GLCM) features for both input ortho-photo and DSM data and RGB indices only for the RGB ortho-photo. Secondly, the conventional ML algorithms are applied to classify the two main inputs (i.e., ortho-photo and DSM) and the extracted hand-crafted features to damage and non-damage classes.

1) Hand-crafted features extraction: In order to provide adequate ancillary training features, as an essential stage of ML-based methods implementation, a number of hand-crafted features, including textural information and an optical index, are extracted for the input data. In detail, the overall principle behind exploiting the textural features is that the buildings damaged after earthquakes appear with rougher and more differentiable texture than that of the intact ones in RS data [43]. Consequently, extracting texture features for giving to MLs could be considered straightforward, leading to producing accurate BDMs in these methods. Accordingly, in the present work, three well-known Haralick [44] GLCM textural features, including mean, variance, and homogeneity, extracted from both the blue band of the ortho-photo and the DSM data were employed to better discriminate the damaged/non-damaged buildings. Besides using textural features, the use of optical indices could also be a good choice in ML-based building damage recognition due to their multi-spectral analysis capability. To this end, the optical panchromatic (PAN) index was also extracted here from the three-band drone ortho-photo data. The further advantages of the used ancillary GLCM-derived textural features and the PAN index alongside their formulae are reported in Table I where i and j are the row and column numbers, μ denotes the mean value of GLCM.

2) Damage/non-damage classification using ML algorithms: The classification stage in ML-based damage detection approaches plays a crucial role in identifying and categorizing post-earthquake damages. In this research, to generate ML-based BDMs, five widely used algorithms, including (1) gradient boosting (GB), (2) logistic regression (LR), (3) decision tree (DT), (4) k-nearest neighbors (KNN), and (5) random forest (RF) were used to categorize the previously mentioned input main and hand-crafted features to damage/ non-damage. The performance characteristics of each ML algorithm are briefly described in Table II, where further details on the algorithms can also be found in their relevant references.

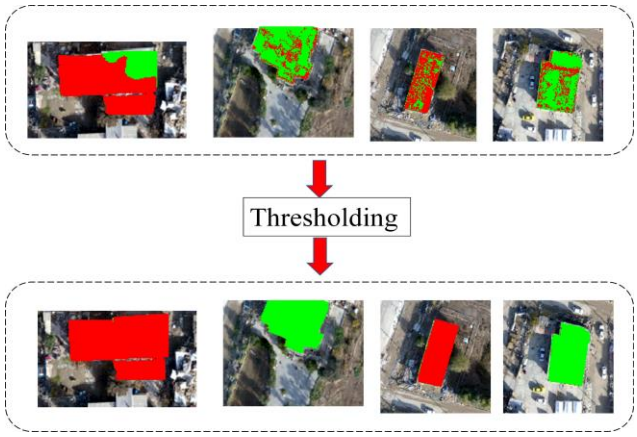

Fig. 5. Post-processing procedure used in this study

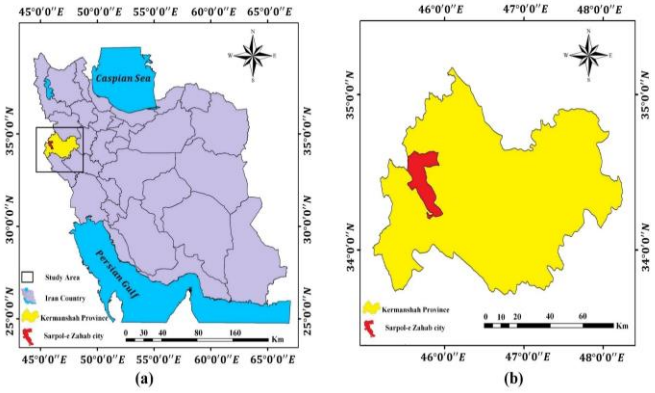

Fig. 6. Location of the study area in Iran and Kermanshah Province. (a) Location of Kermanshah Province in Iran. (b) Location of Sarpol-e Zahab County in Kermanshah Province

D. Post-processing

The initial BDMs previously obtained by both DL- and ML-based damage detection models contained individual damage/ non-damage labeled pixels within the building polygons, which leads to ambiguity about whether each building is entirely intact or destructed. In order to resolve this vagueness and to fully attribute each building to one of the two mentioned classes, a post-processing approach was employed to the initially obtained BDM results. This approach calculated a ratio where the number of damaged pixels is divided by the number of non-damage pixels for each candidate building. Afterward, buildings with ratio values equal to and greater than 0.5 were labeled as damaged, but the remaining were labeled as undamaged (Fig. 5).

III. Study Area and Dataset

The Sarpol-e Zahab earthquake, also known as the Kermanshah earthquake, occurred on November 12, 2017, in the western part of Iran. The earthquake had a magnitude of 7.3 and took place near the town of Sarpol-e Zahab in Kermanshah Province. This earthquake also caused significant loss of life and damage to many buildings and infrastructures in the region.

TABLE II
DESCRIPTIONS OF THE COMPARED ML ALGORITHMS

ML algorithm	Description
GB [45]	The main idea behind this algorithm is to create several sequential models in a forward stage-wise fashion, each of which attempts to reduce the errors of the previous one. This strategy also allows for the optimization of arbitrary distinct loss functions.
LR [46]	Logistic regression uses a logistic function to model some dependent variables that are dichotomous in nature, which could be considered as several possible classes.
DT [47]	This classifier creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, and each branch descending from that node also corresponds to one of the possible values for that attribute.
KNN [48]	The k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning classifier, which uses vicinity to make decisions about the grouping of individual samples.
RF [49]	A random forest is a meta classifier fitting a number of decision trees on various sub-samples of the dataset and averaging several results to improve the final single prediction and to control over-fitting.

TABLE III
THE CHARACTERISTICS OF THE UAV SENSOR USED IN THE PRESENT STUDY

UAV Device	Phantom 4 Pro
Flight altitude	98.8 m
Camera	FC6310
Focal length	8.8 mm
Image dimension	3648 ×5472 pixel
Pixel size	2.41 μm

It is estimated that over 600 people were killed, and thousands were also injured due to the earthquake [50]. The location map of the study area is shown in Fig. 6. In this study, the post-earthquake optical ortho-photo and DSM datasets related to four sub-areas of Sarpol-e Zahab were used to train and validate the comparative models, and a test study area without any overlapping with the train and validation data was also considered to independently evaluate the performance of the models in building damage detection. Both orthophoto and DSM data with the same spatial resolution of 25 cm were derived from stereo UAV (drone) data captured by a sensor with characteristics listed in Table III. Moreover, the training, validation, and test sub-datasets used for training and testing the models are depicted in Fig. 7, and their specifications are also reported in Table IV.

As for the reference BDM (RBDM) data required for training, validating, and testing the algorithms, due to the lack of updated and accurate pre-event building vector data for the Sarpol-e Zahab region, buildings footprints as a number of samples were manually delineated from pre-event Google Earth

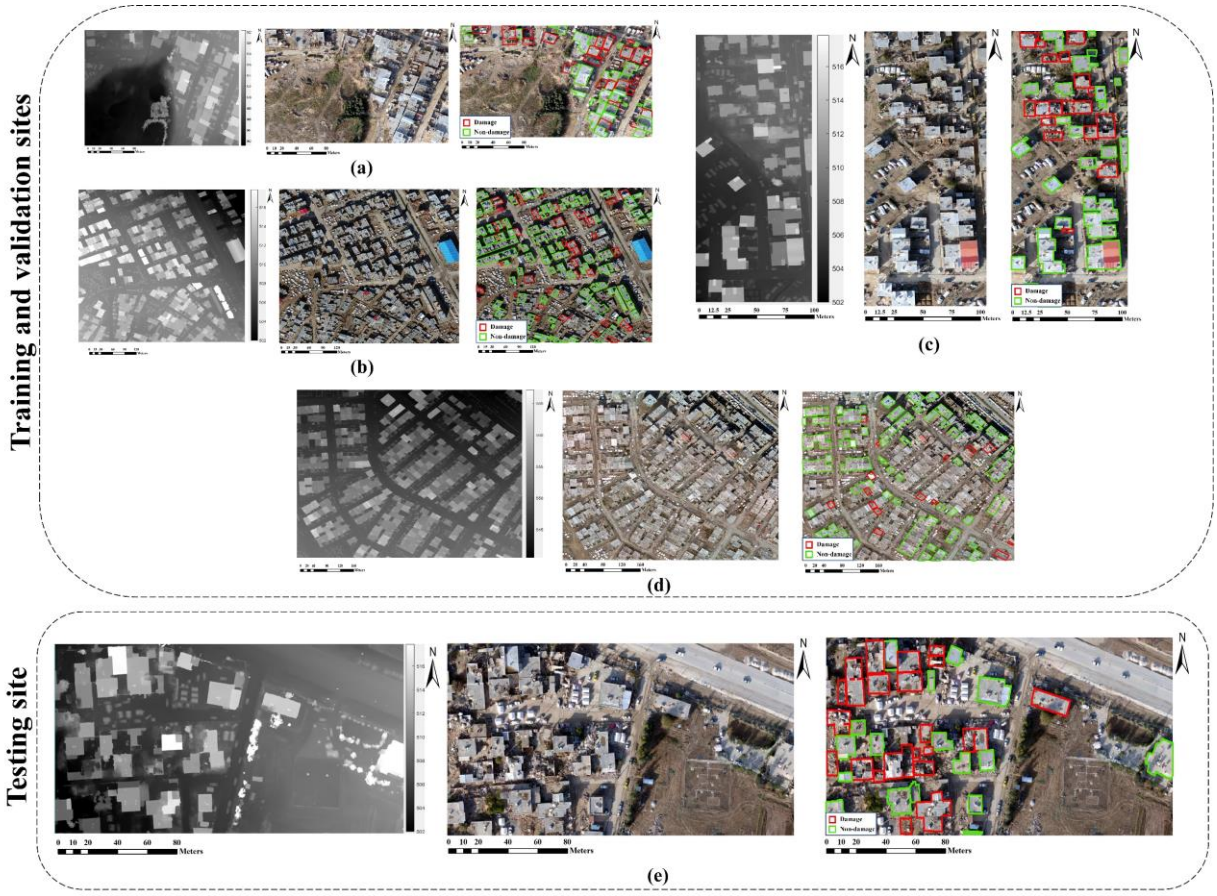

Fig. 7. The train, validation, and test sub-datasets selected in the present research. (a), (c) and (d) are the training sites, (b) is the validation site, and (e) is the testing site, where the first, second, and third columns show DSM, ortho-photo, and corresponding reference BDMs, respectively.

TABLE IV
THE SPECIFICATIONS OF THE SUB-DATASETS EMPLOYED FOR TRAINING, VALIDATING, AND TESTING THE MODELS

Dataset-labeled in Fig.7	GSD (m)	Area (km^2)
Train 1-(a)	0.25	0.03
Validation-(b)	0.25	0.11
Train 2-(c)	0.25	0.02
Train 3-(d)	0.25	0.14
Test-(e)	0.25	0.03

images within a precise visual interpretation procedure. Afterward, these collected samples were divided into two classes, namely (1) damage and (2) non-damage, where buildings with intact roofs were labeled as 'Non-damage' [1], while roof-damaged buildings with varying degrees of destruction were labeled as 'Damage'. Table V lists the number of samples gathered for training, validation, and testing for the two classes of damage and non-damage.

In the case of ML algorithms, all the data needed for the training and testing phases were selected as individual pixels, contrary to the DL models capturing building polygons in all levels of training and testing stages. In fact, to train the ML algorithms, %20 of the pixels of each training polygon with a specific class of interest were randomly considered as training data samples, the number of which can also be seen in Table VI.

IV. EVALUATION METRICS

To quantitatively evaluate the performance of the comparative ML and DL methods in building damage detection, several metrics, namely damage detection rate (DDR), non-damage detection rate (NDR), overall accuracy (OA), kappa coefficient (KC), and F-score (FS) were

respectively estimated by equations (3)-(7), where true positive (TP) and true negative (TN) as two agreement measures refer to the number of correctly recognized damaged and non-damage buildings, respectively. On the other hand, false positive (FP) and false negative (FN) are two disagreement metrics, respectively, signifying the number of non-damaged and damaged buildings detected wrongly as damaged and non-damaged in the predicted BDM. The values of these four individual criteria were derived based on a confusion matrix formed by

TABLE V
DETAILS OF THE SUB-DATASETS SELECTED FOR DL MODELS (UNIT IS POLYGON)

Dataset-labeled in Fig.7	Total Sample	Non-damage	Damage
Train 1-(a)	56	29	27
Validation-(b)	121	81	40
Train 2-(c)	45	27	18
Train 3-(d)	67	52	15
Test-(e)	38	15	23
Non-damage training polygons: **189** Damage training polygons:**100**			

TABLE VI
DETAILS OF THE SUB-DATASETS SELECTED FOR ML ALGORITHMS (UNIT IS PIXEL)

Dataset-labeled in Fig.7	Non-damage	Damage
Train 1-(a)	1971	1917
Train 2-(b)	12313	3642
Train 3-(c)	2527	1243
Train 4-(d)	13716	1312
Non-damage training pixels: **30527** Damage training pixels: **8114**		

TABLE VII
EXPLANATION OF THE CONFUSION MATRIX

		Predicted BDM labels	
RBDM (Actual labels)	Category	damage	non-damage
	damage	TP	FN
	non-damage	FP	TN

$$DDR = \left(\frac{TP}{TP + FN}\right) \times 100\ \% \tag{3}$$

$$NDR = \left(\frac{TN}{FP + TN}\right) \times 100\ \% \tag{4}$$

$$OA = \left(1 - \frac{FN + FP}{FN + TP + FP + TN}\right) \times 100\ \% \tag{5}$$

$$KC = \left(\frac{P_0 - P_e}{1 - P_e}\right) \times 100\ \%$$

$$s.t. \begin{cases} P_e = \frac{(TP+FN)\times(TP+FP)}{(TP+FP+FN+TN)^2} + \frac{(FN+TN)\times(FP+TN)}{(TP+FP+FN+TN)^2} \\ P_0 = \frac{TP+TN}{TP+FP+FN+TN} \end{cases} \tag{6}$$

$$F_S = \left(\frac{2TP}{2TP + FP + FN}\right) \times 100\ \% \tag{7}$$

TABLE VIII
THE CHARACTERISTICS OF THE COMPARATIVE DL MODELS

Model	Backbone	Pre-trained	No. of parameters ($\times 10^6$)
U-Net	None	Randomly initialized	2.164
MRU-Net	None	Randomly initialized	7.263
U-Net++	None	Randomly initialized	4.870
Link-Net	None	Randomly initialized	20.325
U_R	ResNet34	ImageNet	24.456
U_V	Vgg16	ImageNet	23.752
U_I	InceptionV3	ImageNet	29.933

TABLE IX
THE VALUES OF HYPERPARAMETERS USED FOR THE DL MODELS

DL Hyperparameters	Value
Learning rate	0.0001
Batch size	16
Maximum epochs	100
Optimizer	'Adam'
Loss	'Categorical cross entropy'

comparing the BDM labels predicted by a model with the actual labels of an RBDM, as demonstrated in Table VII [2].

V. RESULTS AND DISCUSSION

Experimental results obtained in this study are presented in three separate parts. The first part examines the performance of the proposed DL-based WETUM along with other U-Net-based models for different spectral and geometrical input data. The aforementioned widely used ML techniques are evaluated in the second part for different combinations of spectral- and geometrical-based input features. Finally, the best results of both DL-and ML-based approaches are analyzed and compared in the third part.

1) Comparing the proposed WETUM with other DL-based approaches: To investigate the performance efficacy of the proposed WETUM for two input data combinations, including RGB and RGB+DSM, its qualitative and quantitative outcomes derived in the two different schemes were compared to those outputted by some state-of-the-art multi-class semantic segmentation DL (MCSS-DL) models, including typical U-Net, U-Net+ResNet34 (U_R) [51], U-Net+Vgg16 (U_V) [52], U-Net+InceptionV3 (U_I) [53], multi-residual U-Net (MRU-Net)

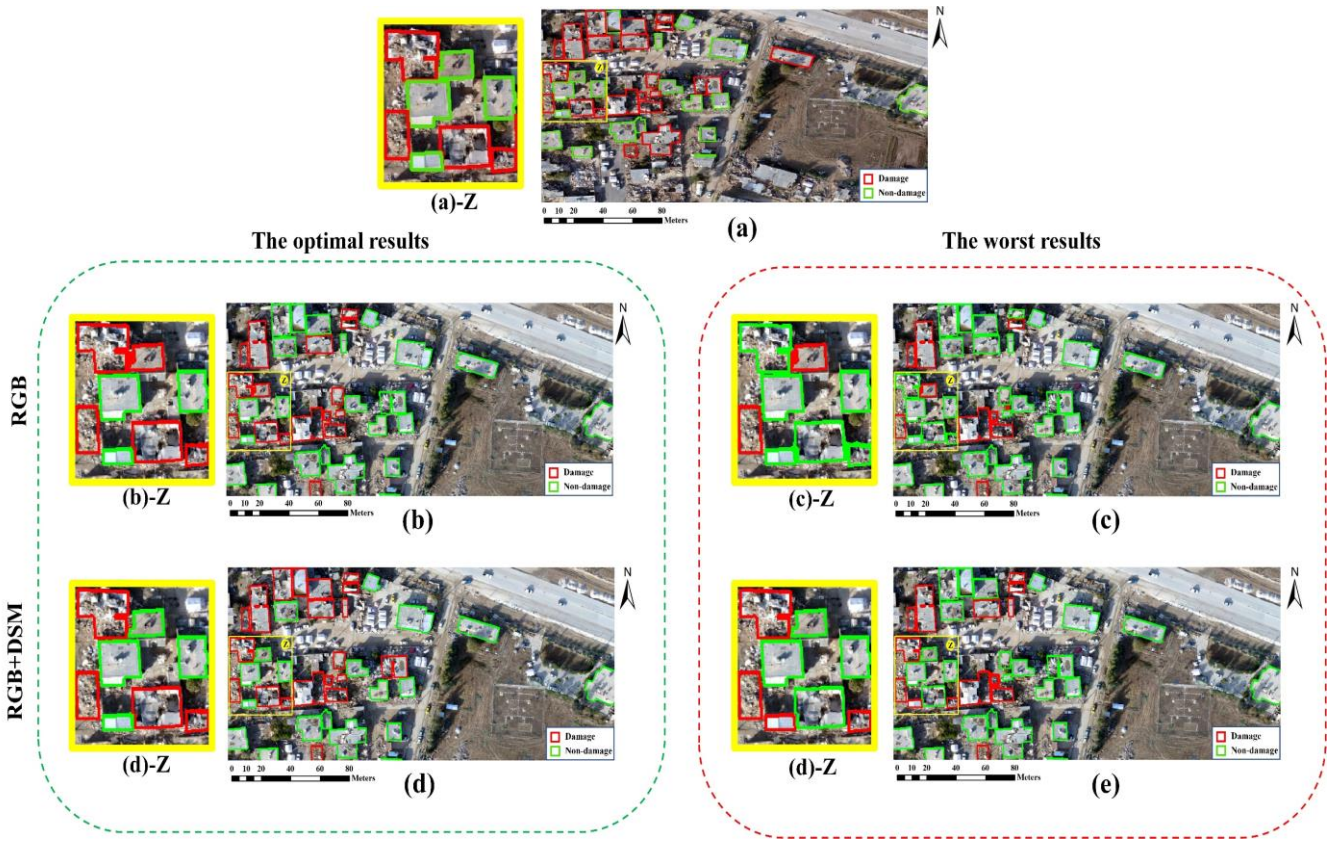

Fig. 8. The best and worst BDMs produced using the DL models for the two types of input data, RGB and RGB+DSM: (a) Reference BDM, (b) and (d) WETUM BDMs as the best results, (c) and (e) U-Net BDMs as the worst results.

TABLE X
QUANTITATIVE PERFORMANCE EVALUATION OF THE FIVE U-NET-BASED DL MODELS IN TWO CASES OF INPUT DATA, INCLUDING RGB AND RGB+DSM, WHERE THE BOLD TEXTS SIGNIFY THE BEST RESULTS

Input data	Model	TP (No.) (RBDM=23)	TN (No.) (RBDM=15)	Evaluation Metrics (%)		
				OA	KC	F-score
RGB	U-Net	8	14	57.89	24.19	50.00
	MRU-Net	10	14	63.16	32.32	58.82
	U_R	11	**15**	68.42	41.98	64.71
	U-Net++	12	14	68.42	40.78	66.67
	U_V	13	14	71.05	45.14	70.27
	U_I	13	14	71.05	45.14	70.27
	Link-Net	13	**15**	73.68	50.56	72.22
	WETUM	**15**	14	**76.32**	**54.16**	**76.92**
RGB+DSM	U-Net	12	13	65.79	35.17	64.84
	MRU-Net	11	**15**	68.42	41.98	64.71
	U_R	14	14	73.68	49.60	73.68
	U_V	13	**15**	73.68	50.65	72.22
	U_I	14	14	73.68	49.60	73.68
	Link-Net	14	14	73.68	49.60	73.68
	U-Net++	14	**15**	76.32	52.12	75.68
	WETUM	**18**	13	**81.58**	**62.75**	**83.72**

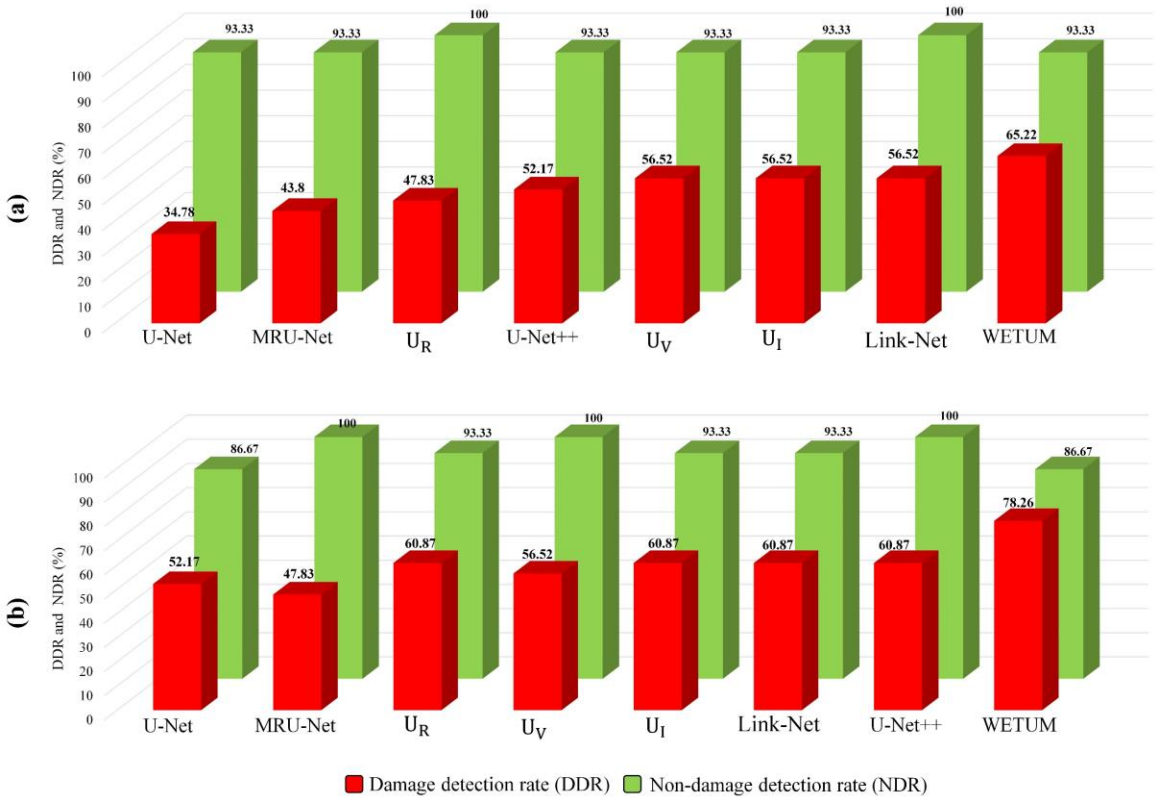

Fig. 9. The damage and non-damage detection rates obtained from the DL models for two types of input data: (a) RGB and (b) RGB+DSM

[54], U-Net ++ [55], and Link-Net [56] in BDM generation, reported in Fig. 8 and Table X. As for the MCSS-DL comparative models, their specifications are summarized in Table VIII. It is worth noting that all hyperparameters were considered the same for all the comparative models to ensure an impartial comparison among the results. The values set for these hyperparameters can be found in Table IX. Moreover, an early stopping parameter was used during the training stage to minimize excessive training and achieve the best accuracy for each deep network.

From the quantitative results reported in Table X and Fig. 9, it can be noticed that the proposed WETUM outperformed the other comparative MCSS-DL models in both input data cases. In detail, in the first case where only RGB drone data was given to the models, WETUM led to the highest OA/F- score values of 76.32/76.92 (%) with improvements of 18.43/26.92, 13.16/18.10, 7.9/12.21, 7.9/10.25, 5.27/6.65, 5.27/6.65, 2.64/4.7 (%) compared to the U-Net, MRU-Net, U_R, U-Net++ U_V, U_I, and Link-Net models, respectively. In addition, contrary to the NDR measure values that were nearly similar, DDR values in the first case varied remarkably for the compared MCSS-DL structures. Notably, the proposed WETUM with a DDR value of 65.22 (%) managed to improve the DDR values of the U-Net, MRU-Net, U_R, U-Net++ U_V, U_I, and Link-Net networks by 30.44, 21.42 17.39, 13.05, 8.7, 8.7 and 8.7 (%), respectively (Fig. 9a). Similarly, in the second case in which both the RGB and DSM data were used, a maximum KC value of 62.75 (%) was achieved when employing WETUM to produce the BDM, which was also more than those of U-Net, MRU-Net, U_R, U-Net++ U_V, U_I, and Link-Net by 27.58, 20.77, 13.15, 12.1, 13.15, 13.15, and 10.63 (%), respectively. As for correctly detecting the damaged buildings, the highest value of DDR metric, i.e., 78.26 (%), was associated with the proposed WETUM and contained drastic increments of 26.09, 30.43, 17.39, 21.74, 17.39 and 17.39 (%) compared to the aforementioned models, respectively (Fig. 9b). The superiority of the WETUM introduced in this study over the compared MCSS-DL networks is generally due to several reasons stated in the following. Firstly, U-Net-based networks used in the proposed model for resulting individual predictions are supported by the three different pre-trained backbone structures, each of which, as an efficient feature extractor, leads to particular discriminative damage information. Secondly, the outcome of WETUM made from linearly combining the individual predictions of the three composite models involves advantageous complementary building damage and non-damage information. Moreover, the proposed robust grid search approach used in WETUM efficiently estimates the weights required for integrating the individual predictions. In addition to the WETUM ranked as the first superior method, the composite structures U_R, U_V, and U_I with similar quantitative results performed better than the state-of-the-art DL architecture. For instance, in the first case, the DDR value of U-

Net as the worst model increased by 13.05, 21.74, and 21.74 (%) when U_R, U_V, and U_I were used, respectively, which indicates adding pre-trained backbones to U-Net can dramatically improve its capability in building damage detection. Even though the three composite models individually improved the performance of the ordinary U-Net network, no significant differences among their performances were observed in the BDM generation. Yet, integrating the three predictions of U_R, U_V, and U_I within the proposed ensemble model drastically increased the accuracy of BDM extraction accuracy. This again reveals the high potential of the grid search strategy embedded in WETUM for efficiently integrating several individual predictions. As well, other MCSS-DL models (i.e., MRU-Net, U-Net ++, and Link-Net) performed well compared to traditional U-Net. For example, in the second case, when applying MRU-Net, U-Net ++, and Link-Net, the OA value of U-Net improved by 13.05, 5.27, 10.53, and 15.79 (%), respectively. Nonetheless, they could not achieve better outcomes than WETUM in all cases. On the other hand, from the viewpoint of the impact of input data (features), in the second case in which the drone data-derived DSM as geometrical information was added to the first case (RGB), the DSM data positively influenced all the compared DL-based approaches. Notably, using DSM alongside the RGB data respectively raised the F-score values of U-Net, U_R, U_V, U_I, MRU-Net, U-Net ++, Link-Net, and WETUM by 18.84, 8.97, 1.95, 3,41, 5.89, 9.01, 1.46, and 6.8 (%), implying DSM provides precise height information and geometrical variations facilitating the detection of some damaged buildings not detectable by only spectral information, i.e., 2D RGB data.

Regarding the obtained BDMs, the yellow box in Fig. 8 demonstrates that the BDM produced by the typical U-Net structure, as the worst result, missed many damaged buildings in both the first and second cases. The reason is that U-Net solely relies on an individual prediction and also lacks an efficient feature extractor component to capture low- and high-level damage features. Contrarily, the proposed structure resulted in the optimal BDMs in which most of the damaged buildings were identified correctly. This is due to the reliable integration of the predictions of the three pre-trained composite models in WETUM, which provides complementary damage information. In addition, the presence of DSM data as a geometrical source alongside the spectral RGB data led to an improvement in detecting the damaged buildings in all outputs, especially. In particular, the improvement was more considerable in the results of the typical U-Net model, which indicates the addition of altitudinal data to 2D RGB information largely compensated for the lack of strong feature extractors in its architecture.

2) Performance evaluation and comparison of ML algorithms: To peruse and analyze the impact of spectral, geometrical, and hand-crafted features on the results of the ML approaches and also individually compare the performance of the models, the outcomes of the compared five ML algorithms, i.e., GB, LR, DT, KNN, and RF were obtained in two cases in terms of the features input to them. In the first case, their performances were evaluated for drone data-derived spectral features (SFs),

TABLE XI
THE SET HYPERPARAMETERS FOR EACH ML ALGORITHM

ML algorithm	ML Hyperparameters
GB	loss='log_loss', learning_rate=0.1, n_estimators=100, subsample=1.0, criterion='friedman_mse', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_depth=3
LR	penalty='l2', tol=0.0001, C=2.0, random_state=0, max_iter=100
DT	max_depth=15, random_state=0, min_samples_leaf=10
KNN	n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski'
RF	n_estimators=200, max_depth=15, random_state=0

including RGB bands, RGB-derived GLCM textures, and the optical PAN index. On the other hand, in the second case, the drone data-based DSM, DSM-derived GLCM features, and SFs as a combination of spectral and geometrical features (SGFs) were fed to the compared ML approaches. Hence, for ease of discussion, the results of the GB, LR, DT, KNN, and RF machine learning models for the two mentioned input data SFs/SGFs categories were respectively termed as GB-SF/GB-SGF, LR-SF/LR-SGF, DT-SF/DT-SGF, KNN-SF/KNN-SGF, and RF-SF/RF-SGF. Regarding the set hyperparameters listed in Table XI for the compared algorithms, their values were either set experimentally or considered the same as the default values suggested by the Python Sklearn library [57]. Furthermore, the quantitative and qualitative ML results attained for the two data cases are presented in Table XII, Fig. 10 and Fig. 11.

Based on the quantitative results described in Table XII and Fig. 11, the outcomes could overall be discussed in two input data cases SFs and SGFs. In the first case (SFs) where only spectral data was fed to the compared ML approaches, all the algorithms generally produced poor results and could not properly distinguish more damaged buildings. In more detail, though all the methods substantially tended to recognize undamaged buildings, the RF-SF with an OA of 47.37 (%) performed the best and improved all the GB-SF, LR-SF, DT- SF, and KNN-SF methods by 5.26 (%). In fact, the latter four algorithms had identical poor performances in BDM generation compared to RF-SF. In the second case (SGFs) in which the geometrical DSM data was exploited besides the SFs, the DSM data and its GLCM features led to significant accuracy increments in the results of all the compared ML approaches. Notably, the DDR values of GB-SGF, LR-SGF, DT-SGF, KNN-SGF, and RF-SGF were overall enhanced by an average of 35.69 (%) compared to the SFs case. However, substantial variations were not found among the results of the comparative ML models when giving SGFs to them for BDM generation. Totally, the RF classification method revealed rather promising results in comparison to the other four ML approaches in both SFs and SGFs cases. Because RF is known to be robust against input data outliers and, more importantly, is an ensemble classifier

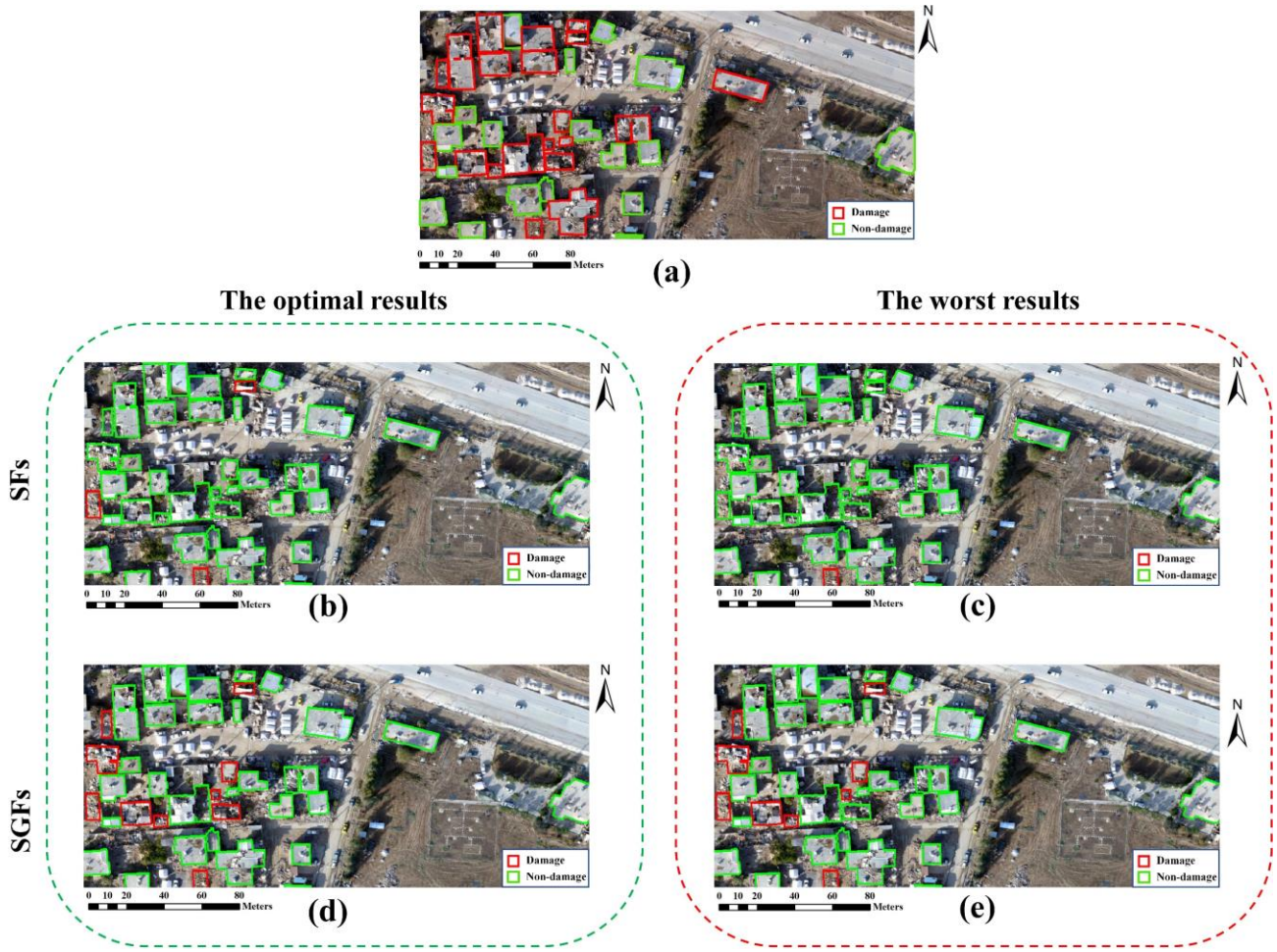

Fig. 10. The best and worst BDMs produced using the ML-based classifiers for the two types of input data, SFs and SGFs: (a) Reference BDM, (b) RF BDM as the best result, (c) GB, LR, KNN and DT BDMs jointly obtained as the worst results, (d) RF, KNN and DT jointly obtained as the best results, (c) GB and LR jointly obtained as the worst results.

TABLE XII
QUANTITATIVE PERFORMANCE EVALUATION OF THE FIVE ML ALGORITHMS FOR THE TWO TYPES OF INPUT DATA, SFS, AND SGFS, WHERE THE BOLD TEXTS SIGNIFY THE BEST RESULTS

Input data	Model	TP (No.) (RBDM=23)	TN (No.) (RBDM=15)	Evaluation Metrics (%)		
				OA	KC	FS
SFs	GB	1	**15**	42.11	3.46	8.33
	LR	1	**15**	42.11	3.46	8.33
	DT	1	**15**	42.11	3.46	8.33
	KNN	1	**15**	42.11	3.46	8.33
	RF	**3**	**15**	**47.37**	**10.59**	**23.08**
SGFs	GB	9	**15**	63.16	33.67	56.25
	LR	9	**15**	63.16	33.67	56.25
	DT	10	**15**	65.79	37.78	60.61
	KNN	10	**15**	65.79	37.78	60.61
	RF	10	**15**	65.79	37.78	60.61

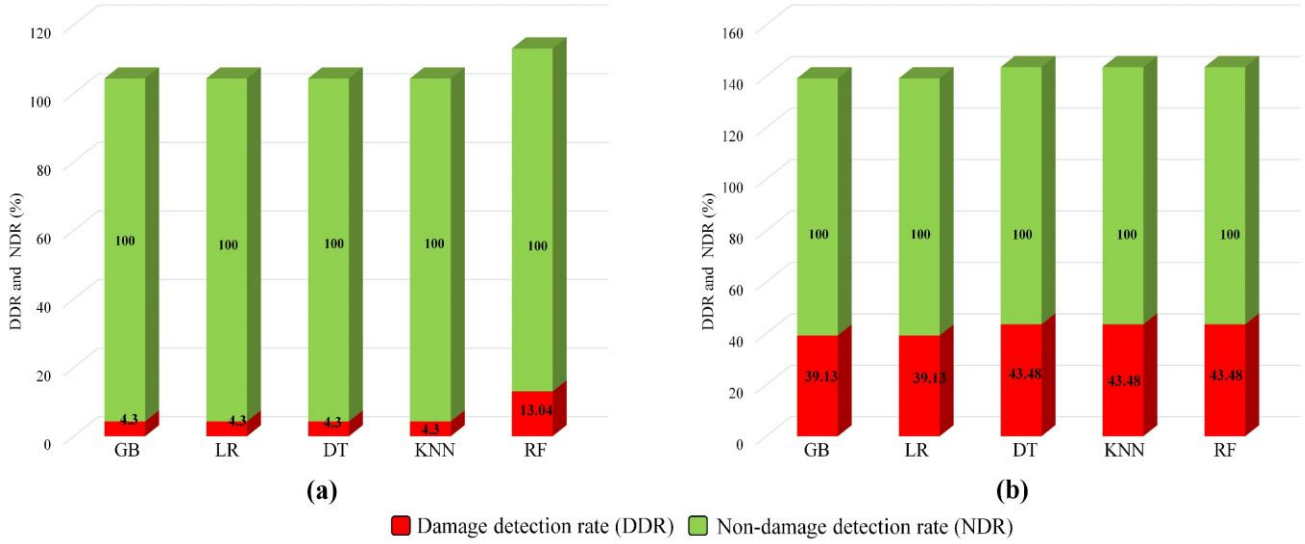

Fig. 11. The damage and non-damage detection rates obtained from the ML algorithms for the two types of input data: (a) SFs and (b) SFFs

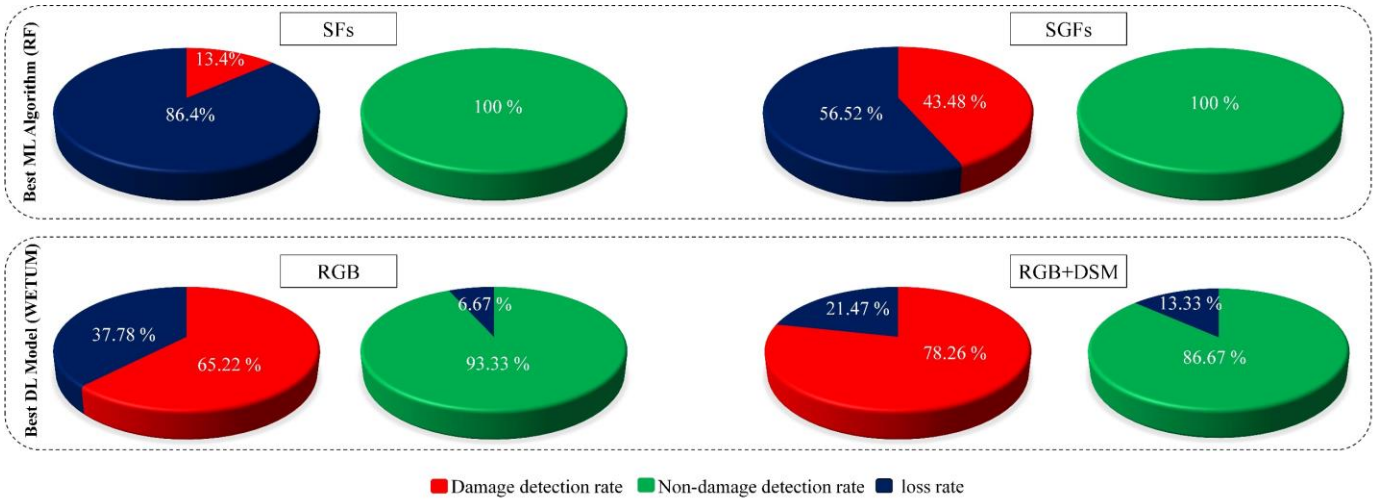

Fig. 12. Quantitative comparison of the BDMs produced using the best DL model (WETUM) and the best ML algorithm (RF)

averaging several DT-based predictions to improve the final single prediction.

On the other hand, the BDMs shown in Fig. 10 as qualitative outcomes confirm the RF-SGFs generated the best BDM where some of the damaged buildings were recognized due to the use of DSM-related features and the aforementioned merits of RF. In detail, in the first SFs case, GB-SF, LR-SF, DT-SF, and KNN-SF jointly with detecting only 1 damaged building led to the worst BDMs. But RF outputted relatively better BDM where 3 damaged buildings were correctly distinguished. In the second SGFs case, RF-SGF alongside the two DT-SGF and KNN-SGF models generated acceptable BDMs in which a fairly higher number of buildings were extracted than the first case BDMs. Consequently, it could be deduced ensemble classification can generate better damage detection results, and the synergic use of spectral and geometrical information can also boost the performance of ML-based building damage detection methods. Of course, the ancillary hand-crafted features extracted from the training area are merely suited and limited to the same particular region and not generic to other unseen test areas.

3) Comparison of the produced BDM using DL and ML-based method: To investigate how generalizable both ML-based and DL-based building damage detection approaches are to a new unseen earthquake-damaged region nonoverlapped with the training area, the BDMs resulted from both the RF model as the best ML algorithm and the proposed ensemble DL approach named WETUM were compared together. The comparison was made in two different scenarios in terms of the

input features. In the first scenario, the RF algorithm took 'SFs', but only the 'RGB' bands were fed to WETUM. Contrarily, in the second scenario, 'SGFs' and 'RGB+DSM' were given to RF and WETUM, respectively. For ease of comparison and discussion of the results obtained in the first scenario, the RF and the WETUM models were regarded as RF-SFs and WETUM-RGB, respectively, in the following. Likewise, the two compared models were also termed RF-SGFs and WETUM-RGBDSM in the second scenario.

From Fig. 12, the WETUM, as the best DL approach, performed best compared to RF as the best representative of ML algorithms in all the scenarios. In detail, in the first scenario that merely relies on spectral information, a DDR value of 65.22 (%) was attained when using WETUM-RGB for generating BDMs, which raised that of RF-SFs by 51.82 (%), implying WETUM detected the damaged buildings much better than the RF in this scenario. In the second scenario where both spectral and geometrical features were employed, WETUM-RGBDSM with a maximum DDR value of 78.26 (%) led to the best BDM in this research work and also increased the DDR values of WETUM-RGB, RF-SFs, and RF-SGFs by 13.04 (%), 64.86 (%), and 34.78 (%). Additionally, the damage loss rate, which demonstrates the amount of missed damaged buildings, contained the least value of 21.47 (%) for WETUM-RGBDSM compared to all other cases. Regardless of the comparison made for each scenario between the two methods, although many diverse hand-crafted features as ancillary information were injected into RF, the ML approach could not extract the damaged buildings even in comparison with the WETUM-RGB. It is therefore deduced that even though many of the aforementioned spectral and geometrical features were considered in the ML-based approaches, even the synergic use of all these features were not efficient for identifying the buildings damaged in another unseen earthquake-influenced area. On the contrary, only RGB-based multi-level deep features automatically extracted within the proposed DL-based WETUM were useful enough to be employed in a new earthquake scenario. Indeed, the proposed WETUM, even with RGB bands, generalizes to new unseen test data remarkably better than the ML-based RF equipped with many ancillary features. This superiority is mainly because the proposed DL model employs a reliable integrated prediction that is obtained by combining several predictions of transferred individual U-Net-based powerful composite models with the three optimal weights determined by the proposed effective grid search approach.

VI. Conclusion

In this research, three main objectives in the field of binary building damage map (BDM) generation were pursued. The first objective was to propose a novel weighted ensemble transferred U-Net-based model named WETUM linearly integrating the predictions of three pre-trained ResNet34, Vgg16, and InceptionV3 backbones. The second objective of this study was to compare the results of machine learning (ML) and deep learning (DL) approaches for two different input data sets based on (1) 'RGB' data and (2) the combination of RGB data and digital surface model (DSM). The third objective of the present work was to explore the generalizability level of both DL- and ML-based approaches for detecting damaged buildings in a new unseen earthquake-affected sub-region. To carry out experiments required for these objectives, drone data captured over Sarpol-e Zahab earthquake in Iran was utilized in this paper. The study's findings emphasize the operational relevance and potential impact on enhancing post-earthquake response efforts. In detail, the foremost remarks discovered in this study are therefore listed below:

1) The proposed WETUM, due to exploiting the transfer learning technique and integrating beneficial aspects of damage information, outperformed the other compared DL networks.
2) As for the impact of the input feature sets, the DSM data containing height information into substantially boosted both ML and DL approaches in building damage examination.
3) Regarding the generalizability level of the ML and DL approaches, the synergic use of main and hand-crafted features was solely suited to the training sub-regions in MLs. On the other hand, WETUM with and without adding DSM to RGB data revealed the highest generalizability among all the compared models.

Since remote sensing data, specifically ortho-photo UAV images observe the terrestrial objects with an orthogonal (nadir) view, the proposed approach cannot detect the damage connected to the underground and aboveground utilities. Furthermore, though the present study only focused on localizing damaged buildings regardless of analyzing building damage parameters such as damage extent and severity [57], this study significantly contributes to timely and accurate rescue operations. Anyway, as future research works, the potential of WETUM could be evaluated in higher levels of building damage assessment, such as characterizing damage extent and severity for further earthquake events, especially when outdated or incomplete building vector data for the target study area is available.

Acknowledgment

The authors would like to express their most sincere gratitude to Sahand Tahermanesh and Dr. Milad Janalipour for their support. We would also like to appreciate the LiDAR research laboratory of K. N. Toosi University of Technology for making the UAV-derived orthophoto and DSM available to the authors of the present paper (https://www.researchgate.net/lab/Laser-Scanners-Ali-Mohammadzadeh, accessed on January, 3rd 2024). This work is a part of international collaboration research and supported by the Natural Sciences Foundation of Zhejiang Province under Grant No. LY23F020006 for publication.

References

[1] S. T. Seydi, H. Rastiveis, B. Kalantar, A. A. Halin, and N. Ueda, “BDD-Net: An End-to-End Multiscale Residual CNN for Earthquake-Induced Building Damage Detection,” *Remote Sens (Basel)*, vol. 14, no. 9, 2022, doi: 10.3390/rs14092214.

[2] A. Mohsenifar, A. Mohammadzadeh, A. Moghimi, and B. Salehi, “A novel unsupervised forest change detection method based on the

integration of a multiresolution singular value decomposition fusion and an edge-aware Markov Random Field algorithm," *Int J Remote Sens*, vol. 42, no. 24, 2021, doi: 10.1080/01431161.2021.1995075.

[3] S. Pirasteh, J. Li, and I. Attarzadeh, "Implementation of the damage index approach for the rapid evaluation of earthquake resistant buildings," *Earth Sci Inform*, vol. 8, no. 4, 2015, doi: 10.1007/s12145-014-0204-0.

[4] S. Pirasteh *et al.*, "Cloud-based geospatial platform in support of sustainable development goals 2030: How to be prepared for earthquake disasters?," in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 2020. doi: 10.5194/isprs-archives-XLIII-B3-2020-1705-2020.

[5] M. Moradi and R. Shah-Hosseini, "Earthquake Damage Assessment Based on Deep Learning Method Using VHR Images," 2021. doi: 10.3390/iecg2020-08545.

[6] D. Omarzadeh, S. Karimzadeh, M. Matsuoka, and B. Feizizadeh, "Earthquake aftermath from very high-resolution worldview-2 image and semi-automated object-based image analysis (Case study: Kermanshah, sarpol-e zahab, Iran)," *Remote Sens (Basel)*, vol. 13, no. 21, 2021, doi: 10.3390/rs13214272.

[7] G. Abdi and S. Jabari, "A Multi-Feature Fusion Using Deep Transfer Learning for Earthquake Building Damage Detection," *Canadian Journal of Remote Sensing*, vol. 47, no. 2, 2021, doi: 10.1080/07038992.2021.1925530.

[8] X. Yuan *et al.*, "Automated building segmentation and damage assessment from satellite images for disaster relief," in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 2021. doi: 10.5194/isprs-archives-XLIII-B3-2021-741-2021.

[9] Y. Li, W. Hu, H. Dong, and X. Zhang, "Building damage detection from post-event aerial imagery using single shot multibox detector," *Applied Sciences (Switzerland)*, vol. 9, no. 6, 2019, doi: 10.3390/app9061128.

[10] N. Takhtkeshha, A. Mohammadzadeh, and B. Salehi, "A Rapid Self-Supervised Deep-Learning-Based Method for Post-Earthquake Damage Detection Using UAV Data (Case Study: Sarpol-e Zahab, Iran)," *Remote Sens (Basel)*, vol. 15, no. 1, 2023, doi: 10.3390/rs15010123.

[11] S. Tilon, F. Nex, N. Kerle, and G. Vosselman, "Post-disaster building damage detection from earth observation imagery using unsupervised and transferable anomaly detecting generative adversarial networks," *Remote Sens (Basel)*, vol. 12, no. 24, 2020, doi: 10.3390/rs12244193.

[12] D. Duarte, F. Nex, N. Kerle, and G. Vosselman, "DAMAGE DETECTION on BUILDING FAÇADES USING MULTI-TEMPORAL AERIAL OBLIQUE IMAGERY," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2019. doi: 10.5194/isprs-annals-IV-2-W5-29-2019.

[13] A. C. Loerch, D. A. Stow, L. L. Coulter, A. Nara, and J. Frew, "Comparing the Accuracy of sUAS Navigation, Image Co-Registration and CNN-Based Damage Detection between Traditional and Repeat Station Imaging," *Geosciences (Switzerland)*, vol. 12, no. 11, 2022, doi: 10.3390/geosciences12110401.

[14] M. Janalipour and A. Mohammadzadeh, "A novel and automatic framework for producing building damage map using post-event LiDAR data," *International Journal of Disaster Risk Reduction*, vol. 39, 2019, doi: 10.1016/j.ijdrr.2019.101238.

[15] E. Khankeshizadeh, A. Mohammadzadeh, A. Moghimi, and A. Mohsenifar, "FCD-R2U-net: Forest change detection in bi-temporal satellite images using the recurrent residual-based U-net," *Earth Sci Inform*, vol. 15, no. 4, 2022, doi: 10.1007/s12145-022-00885-6.

[16] S. T. Seydi, M. Hasanlou, J. Chanussot, and P. Ghamisi, "BDD-Net+: A Building Damage Detection Framework Based on Modified Coat-Net," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 16, 2023, doi: 10.1109/JSTARS.2023.3267847.

[17] S. Karimzadeh, M. Ghasemi, M. Matsuoka, K. Yagi, and A. C. Zulfikar, "A Deep Learning Model for Road Damage Detection After an Earthquake Based on Synthetic Aperture Radar (SAR) and Field Datasets," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 15, 2022, doi: 10.1109/JSTARS.2022.3189875.

[18] Y. Wang, W. Feng, K. Jiang, Q. Li, R. Lv, and J. Tu, "Real-Time Damaged Building Region Detection Based on Improved YOLOv5s and Embedded System From UAV Images," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 16, 2023, doi: 10.1109/JSTARS.2023.3268312.

[19] B. Jena, S. Saxena, G. K. Nayak, L. Saba, N. Sharma, and J. S. Suri, "Artificial intelligence-based hybrid deep learning models for image classification: The first narrative review," *Computers in Biology and Medicine*, vol. 137. 2021. doi: 10.1016/j.compbiomed.2021.104803.

[20] Q. Lin, T. Ci, L. Wang, S. K. Mondal, H. Yin, and Y. Wang, "Transfer Learning for Improving Seismic Building Damage Assessment," *Remote Sens (Basel)*, vol. 14, no. 1, 2022, doi: 10.3390/rs14010201.

[21] J. Yang, W. Xiong, S. Li, and C. Xu, "Learning structured and non-redundant representations with deep neural networks," *Pattern Recognit*, vol. 86, 2019, doi: 10.1016/j.patcog.2018.08.017.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015. doi: 10.1007/978-3-319-24574-4_28.

[23] C. Wu *et al.*, "Building damage detection using u-net with attention mechanism from pre-and post-disaster remote sensing datasets," *Remote Sens (Basel)*, vol. 13, no. 5, 2021, doi: 10.3390/rs13050905.

[24] L. Deng and Y. Wang, "Post-disaster building damage assessment based on improved U-Net," *Sci Rep*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-20114-w.

[25] P. Mangalraj, S. P. Duddela, P. Kirubanantham, and S. Iniyan, "Post-earthquake Building Damage Detection Using Deep Learning," 2022. doi: 10.1007/978-981-16-1249-7_13.

[26] S. May, A. Dupuis, A. Lagrange, F. De Vieilleville, and C. Fernandez-Martin, "BUILDING DAMAGE ASSESSMENT WITH DEEP LEARNING," in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 2022. doi: 10.5194/isprs-archives-XLIII-B3-2022-1133-2022.

[27] Y. Shen *et al.*, "BDANet: Multiscale Convolutional Neural Network with Cross-Directional Attention for Building Damage Assessment from Satellite Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022, doi: 10.1109/TGRS.2021.3080580.

[28] S. A. Ahmadi, A. Mohammadzadeh, N. Yokoya, and A. Ghorbanian, "BD-SKUNet: Selective-Kernel UNets for Building Damage Assessment in High-Resolution Satellite Images," *Remote Sens (Basel)*, vol. 16, no. 1, p. 182, Dec. 2023, doi: 10.3390/rs16010182.

[29] E. Irwansyah, H. Young, and A. A. S. Gunawan, "International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING Multi Disaster Building Damage Assessment with Deep Learning using Satellite Imagery Data," 2023. [Online]. Available: https://xview2.org/,

[30] W. Zhao, "Research on the deep learning of the small sample data based on transfer learning," in *AIP Conference Proceedings*, 2017. doi: 10.1063/1.4992835.

[31] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in Neural Information Processing Systems*, 2014.

[32] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10. 2010. doi: 10.1109/TKDE.2009.191.

[33] F. Yu, X. Xiu, and Y. Li, "A Survey on Deep Transfer Learning and Beyond," *Mathematics*, vol. 10, no. 19. 2022. doi: 10.3390/math10193619.

[34] C. Maarouf, M. L. Benomar, and N. Settouti, "Pre-trained Backbones Effect on Nuclei Segmentation Performance," in *Communications in Computer and Information Science*, 2022. doi: 10.1007/978-3-031-04112-9_8.

[35] N. Sharma *et al.*, "U-Net Model with Transfer Learning Model as a Backbone for Segmentation of Gastrointestinal Tract," *Bioengineering*, vol. 10, no. 1, 2023, doi: 10.3390/bioengineering10010119.

[36] L. D. Huynh and N. Boutry, "A U-Net++ with pre-trained efficientnet backbone for segmentation of diseases and artifacts in endoscopy images and videos," in *CEUR Workshop Proceedings*, 2020.

[37] S. Pirasteh *et al.*, "Developing an algorithm for buildings extraction and determining changes from airborne LiDAR, and comparing with R-CNN method from drone images," *Remote Sens (Basel)*, vol. 11, no. 11, 2019, doi: 10.3390/rs11111272.

[38] F. Chollet, "Keras (2015)," *URL http://keras. io*, 2017.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016. doi: 10.1109/CVPR.2016.90.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016. doi: 10.1109/CVPR.2016.308.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2010. doi: 10.1109/cvpr.2009.5206848.

[43] F. Eslamizade, H. Rastiveis, N. K. Zahraee, A. Jouybari, and A. Shams, "Decision-level fusion of satellite imagery and LiDAR data for post-earthquake damage map generation in Haiti," *Arabian Journal of Geosciences*, vol. 14, no. 12, 2021, doi: 10.1007/s12517-021-07293-y.

[44] R. M. Haralick, K. Shanmugam, and I. Dinstein, "TexturalFeaturesHaralickShanmugamDinstein," *IEEE Trans Syst Man Cybern*, vol. 3, 1973.

[45] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann Stat*, vol. 29, no. 5, 2001, doi: 10.1214/aos/1013203451.

[46] C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *Journal of Educational Research*, vol. 96, no. 1, 2002, doi: 10.1080/00220670209598786.

[47] L. Rokach and O. Maimon, "Decision Trees," in *Data Mining and Knowledge Discovery Handbook*, Springer-Verlag, 2006, pp. 165–192. doi: 10.1007/0-387-25465-x_9.

[48] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2888, 2003, doi: 10.1007/978-3-540-39964-3_62.

[49] Breiman L, "Random Forests," *Machine Learning, Springer*, 2001, doi: https://doi.org/10.1023/A:1010933404324.

[50] M. Miyamjima *et al.*, *Site Investigation of the Sarpole-Zahab Earthquake, Mw 7.3 in SW Iran of November 12, 2017*. 2018.

[51] J. L. Arrastia *et al.*, "Deeply supervised unet for semantic segmentation to assist dermatopathological assessment of basal cell carcinoma," *J Imaging*, vol. 7, no. 4, 2021, doi: 10.3390/jimaging7040071.

[52] S. Ghosh, A. Chaki, and K. Santosh, "Improved U-Net architecture with VGG-16 for brain tumor segmentation," *Phys Eng Sci Med*, vol. 44, no. 3, 2021, doi: 10.1007/s13246-021-01019-w.

[53] N. S. Punn and S. Agarwal, "Inception U-Net Architecture for Semantic Segmentation to Identify Nuclei in Microscopy Cell Images," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 16, no. 1, 2020, doi: 10.1145/3376922.

[54] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, 2020, doi: 10.1016/j.neunet.2019.08.025.

[55] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018. doi: 10.1007/978-3-030-00889-5_1.

[56] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE Visual Communications and Image Processing, VCIP 2017*, 2017. doi: 10.1109/VCIP.2017.8305148.

[57] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, 2011.

Ehsan Khankeshizadeh received a B.Eng. degree in geomatics (surveying) engineering from the University of Tabriz, Iran, in 2020 and an M.Eng. degree in remote sensing from Geomatics Engineering faculty of K.N. Toosi University of Technology, Tehran, Iran in 2022.

He is currently a Ph.D. student in remote sensing and a member of the LiDAR laboratory of the faculty of geomatics engineering at K. N. Toosi University of Technology, working on building detection, building change detection, and damage assessment topics under the supervision of Dr. Ali Mohammadzadeh.

His main research interests comprise LiDAR and very high-resolution remote sensing data processing, machine learning, deep learning, and remote sensing applications.

Ali Mohammadzadeh received a Ph.D. degree in remote sensing from Geomatics Engineering faculty of K.N. Toosi University of Technology, Tehran, Iran, in 2009.

He is currently an Associate Professor and Head of the LiDAR Laboratory at K.N. Toosi University of Technology. He has more than 60 published journal papers, and his research interests are LiDAR, artificial intelligence, image processing and pattern recognition, physics of remote sensing, optimization, sensor calibration, disaster management in dust, earthquake, and flooding. He has supervised more than 40 graduate students and actively involved in national industrial projects.

Hossein Arefi received an M.Eng. degree from HFT University of Stuttgart, Stuttgart, Germany, in 2003 and the Ph.D. degree from UniBW Munich University, Munich, Germany, in 2009.

He is currently an Assistant Professor with the School of Surveying and Geospatial Engineering, University of Tehran, Tehran, Iran. He is also a Professor with the Mainz University of Applied Sciences, Mainz, Germany. From September 2008 to August 2013, he was with the Institute of Remote Sensing Technology, German Aerospace Center (DLR), Cologne, Germany.

His research interests include 3-D point cloud analysis, 3-D object and surface modeling, machine learning, and digital elevation modeling.

Amin Mohsenifar received a B.Eng. degree in geomatics engineering from the University of Tabriz, Iran, in 2019 and M.Eng. degree in remote sensing from K. N. Toosi University of Technology (KNTU), Tehran, Iran, in 2021.

He is currently a member of the LiDAR laboratory of KNTU, pursuing a Ph.D. degree in remote sensing under the supervision of Dr. Ali Mohammadzadeh.

His main research interests comprise machine learning, change detection, computer programming, flood disaster management, image processing, and pattern recognition.

Saied Pirasteh (*Senior Member, IEEE*) is a professor with the Geomatics and Geospatial Artificial Intelligence and the Associate Dean at the Institute of Artificial Intelligence, Shaoxing University. He is also a visiting professor and scientist at the Geospatial Intelligence and Mapping Lab, University of Waterloo. He received his first Ph.D. degree in Remote Sensing and GIS (Geology) in 2004. He also received a Ph.D. degree in geomatics and GeoAI (Geography) from the University of Waterloo in 2018. His research interests include Geospatial Data Science & GeoAI, Remote Sensing Computer Vision including satellite, drone, LiDAR data processing, GIS and Geospatial Information analysis and modelling and developing algorithms for their novel applications in hazards & disasters beyond. He has co-authored over 200 Publications.

He is currently an Advisory Board and expert member of the United Nations Global Geospatial Information Management (UN-GGIM) Academic Network and the Chair of the ISPRS ICWG III/IVa on Disaster Management for 2022-2026. He founded the Geospatial Infrastructure Management Ecosystem (GeoIME) (www.geoime.ca) and invented and commercialized this Technology System product and software tools. He has supervised and co-supervised many bachelor students' projects, particularly more than 100 master/Ph.D. students, postdoctoral fellows and visiting scholars from various countries. He has delivered many keynotes at international events and also organized many events worldwide.

EN FAN received the B.S. degree in electronic information science and technology from Hubei Engineering University, in 2006, the M.S. degree in signal and information processing from Nanchang Hangkong University, in 2009, and the Ph.D. degree from Xidian University, China, in 2015. From 2016 to 2019, he held a Postdoc-toral position with the ATR Key Laboratory, Shen-zhen University, China. From 2022 to 2023, he was a visiting scholar with Murdoch University, Australia.

He is currently a Lecturer with Shaoxing University. His research interests include multisensor data fusion, multitarget tracking, and uncertain infor-mation processing.

HuXiong Li received the MS degree in software engineering major of Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree in pattern recognition and intelligent system from Northwest University of Technology, Xian, China, in 2012.

He is currently a Professor with the Department of Computer Science and Engineering, Shaoxing University, Shaoxing, China. His current research interests include network control, complex network, knowledge graph, and cloud computing.

Jonathan Li (*Fellow, IEEE*) received the Ph.D. degree in geomatics engineering from the University of Cape Town, Cape Town, South Africa, in 2000. He is currently a professor of geomatics and systems design engineering with the University of Waterloo, Waterloo, ON, Canada. He has coauthored almost 600 publications, more than 150 of which were published into remote sensing journals, including Remote Sensing of Environment, International Society for Photogrammetry and Remote Sensing (ISPRS) Journal of Photogrammetry and Remote Sensing, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and International Journal of Applied Earth Observation and Geoinformation (JAG). He has also published papers in flagship conferences in computer vision and AI, including Computer Vision and Pattern Recognition Conference (CVPR), Association for the Advancement of Artificial Intelligence (AAAI), and International Joint Conferences on Artificial Intelligence (IJCAI). He has supervised nearly 200 master's/Ph.D. students as well as post-doctoral fellows/visiting scholars to completion. His main research interests include AI-based information extraction from Earth observation images and LiDAR point clouds, photogrammetry and remote sensing, GeoAI and 3-Dvision for digital twin cities, and autonomous driving. Dr. Li is a fellow of the Canadian Academy of Engineering, the Royal Society of Canada (Academy of Science), and the Engineering Institute of Canada. He is also the Editor-in-Chief of JAG and an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.

.