IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 11, 2022, accepted April 19, 2022, date of publication April 28, 2022, date of current version May 5, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3170897

A Deep Attentive Multimodal Learning Approach for Disaster Identification From Social Media Posts

EFTEKHAR HOSSAIN[1], **(Graduate Student Member, IEEE), MOHAMMED MOSHIUL HOQUE**[2], **(Senior Member, IEEE), ENAMUL HOQUE**[3], **(Member, IEEE), AND MD. SAIFUL ISLAM**[1], **(Member, IEEE)**

[1]Department of Electronics and Telecommunication Engineering, Chittagong University of Engineering and Technology, Chattogram 4349, Bangladesh
[2]Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chattogram 4349, Bangladesh
[3]School of Information Technology, York University, Toronto, ON M3J 1P3, Canada

Corresponding author: Mohammed Moshiul Hoque (moshiul_240@cuet.ac.bd)

This work was supported by the Natural Sciences and Engineering Research Council (NSERC), Canada.

ABSTRACT Microblogging platforms such as Twitter have become indispensable for disseminating valuable information, especially at times of natural and man-made disasters. Often people post multimedia contents with images and/or videos to report important information such as casualties, damages of infrastructure, and urgent needs of affected people. Such information can be very helpful for humanitarian organizations for planning adequate response in a time-critical manner. However, identifying disaster information from a vast amount of posts is an arduous task, which calls for an automatic system that can filter out the actionable and non-actionable disaster-related information from social media. While many studies have shown the effectiveness of combining text and image contents for disaster identification, most previous work focused on analyzing only the textual modality and/or applied traditional recurrent neural network (RNN) or convolutional neural network (CNN) which might lead to performance degradation in case of long input sequences. This paper presents a multimodal disaster identification system that utilizes both visual and textual data in a synergistic way by conjoining the influential word features with the visual features to classify tweets. Specifically, we utilize a pretrained convolutional neural network (e.g., ResNet50) to extract visual features and a bidirectional long-term memory (BiLSTM) network with attention mechanism to extract textual features. We then aggregate both visual and textual features by leveraging a feature fusion approach followed by applying the softmax classifier. The evaluations demonstrate that the proposed multimodal system enhances the performance over the existing baselines including both unimodal and multimodal models by attaining approximately 1% and 7% of performance improvement, respectively.

INDEX TERMS Natural disasters, multimodal deep learning, social media, twitter, natural language processing, attention mechanism.

I. INTRODUCTION

In times of disaster events such as earthquake, flood, and hurricane, social media platforms can play a critical role in spreading a large volume of important information [1]–[3]. People frequently use these social media platforms to communicate at different hierarchies such as from individual to individual, individual to government, individual to community and government to people [3], [4]. Victims often share information about disaster events on Twitter, such as reporting about injured or deceased people, and infrastructural damages. Affected people also inquire for urgent aids by posting images, tweets, and videos. Analyzing such social media posts and extracting actionable insights in real-time can be very helpful for humanitarian organizations to assist the affected people [5], [6]. However, it is very difficult and time-consuming task to manually analyze and extract actionable insights from large amount of crisis-related tweets.

The humanitarian computing community has attempted to address the above challenge by developing automated

The associate editor coordinating the review of this manuscript and approving it for publication was Joanna Kołodziej.

systems that can extract and classify crisis-related social media posts [7]–[9]. For example, researchers have develop classifiers to identify event types (e.g., flood, hurricane) [10], whether a post is informative or not [11], as well as humanitarian information types (e.g., types of damages) [12]. Despite such recent progress, existing works are primarily limited in two ways. First, most works on for damage or disaster response from social media posts have mainly concentrated on textual or image content analysis independently. However, recent studies suggest that information from both texts and images often provides valuable insights about an event and thus leads to more precise inferences than the learning from unimodal content [13]. Second, a very few works that utilize multimodal features focus on applying CNN or RNN models for text feature representation [7], [8], which might not work well for longer sentences.

In this work, our goal is to develop an effective computational model for identifying disaster-related information by synergistically integrating features from visual and textual modalities. More specifically, we extract the image features using pre-trained visual (i.e., ResNet50) model. We also extract the textual features by integrating an attention mechanism with the BiLSTM network to address the long-range dependency problem with traditional RNN and CNN architecture. We then aggregate both types of features using the Deep level fusion, followed by applying the softmax layer to classify the given tweet. We perform extensive experiments on a multimodal damage dataset, where the goal is classify damage type (e.g., fire, floods, infrastructure damage) from an image-tweet pair. We compare our models with several baselines that do not utilize multimodal features or do not apply attention mechanisms (Section IV). The key findings from the these experiments are: *(i)* utilizing multimodal features is more effective that uni-modal features, and *(ii)* the RNN model with an attention mechanism can be very effective in improving the performance compared to its counterpart that does not incorporate such mechanism.

The primary contributions of our work are:

- We propose a multimodal architecture that utilizes ResNet50 and BiLSTM recurrent neural network with attention mechanism to classify the damage-related posts by exploiting both visual and textual information.
- We compare the performance of the proposed model with a set of existing unimodal (i.e., image, text) and multimodal classification techniques.
- We empirically evaluate the proposed model on a benchmark dataset and demonstrated how introducing attention could enhance the system performance through an intrinsic evaluation.
- We perform both quantitative and qualitative analysis to get deeper insights about the error types which provide future directions for improving the model.

The remainder of the article is organized as follows. First, we provide an overview of related research on disaster tweet classification in Section II. Next, we present our proposed method in Section IV. We then present our experimental setup, key findings, and errors analysis in Section V. Finally, we conclude the paper with the possible future directions in Section VI.

II. RELATED WORK

A significant amount of work has been done to classify, extract, and summarize disaster-relevant information from social media, see [14] for a detailed survey. Here, We broadly categorize computational models that are closely related to our damage/disaster classification task in two ways: *(i)* unimodal approaches which consider either text or images, *(ii)* multimodal approaches which consider both type of information. We discuss both types of approaches below.

A. UNIMODAL APPROACHES

1) TEXT-BASED DISASTER IDENTIFICATION

Many previous studies have utilized social media texts, and leveraged it for damage or disaster identification [15]. Early works focused on feature-engineering based approaches and used models such as support vector machine (SVM) [16], random forest [17], and logistic regression classifiers [18]. Later, researchers have widely used deep learning-based architectures such as CNN [19], and BiLSTM [20] for classifying the disaster-related tweets. Caragea *et al.* [21] and Nguyen *et al.* [19] proposed CNN-based models to classify the tweets into informative and not-informative categories which provides significant improvements over feature engineering-based approaches. Aipe *et al.* [22] also proposed a CNN-based approach but they focus on multi-label classification rather that simple binary classification to label disaster-related tweets. Similarly, Yu *et al.* [23] used CNN, logistic regression, and SVM to classify the tweets related to different Hurricanes into multiple categories. Their CNN-based model outperformed SVM and LR. In contrast to CNN-based approaches we consider BiLSTMs with attention mechanisms with an aim to better captures dependencies between word tokens.

Some researchers have focused on domain adaption and cross-domain classification [24], [25]. Li *et al.* [24] studied the feasibility of domain adaption for analyzing the disaster tweets by applying the naive Bayes classifier on the Boston Marathon bombing and Hurricane Sandy dataset. Graf *et al.* [25] focused on cross-domain classification so that the classifier can be used across different types disaster events. They employed a cross-domain classifier and utilized emotional, sentimental, and linguistic features extracted from the damage-related tweets. Others have focused on text mining and summarization approaches [26], [27]. For example, Rudra *et al.* [26] assign tweets into different situational classes and then summarizes those tweets. Cameron *et al.* [27] proposed an Emergency Situation Awareness-Automated Web Text Mining (ESA-AWTM) system that detects informative damage-related Twitter messages to inform charitable organizations about the incidents of a disaster. Unlike these systems that broadly focused on text mining and summarization, we only focus specifically on a multi-class classification problem on disaster-related tweets.

2) IMAGE-BASED DISASTER IDENTIFICATION

Most works on identifying disasters from social media images have applied CNN-based classifier. For example, Chaudhuri and Bose [28] used CNN-based model to locate the human body parts from the wreckage images. Nguyen *et al.* [29] developed a deep CNN architecture to label the social media images into multiple disaster categories (i.e., severe, mild, and no-damage). Similarly, Alam *et al.* [30] proposed a pre-trained CNN (VGG16) based framework that can identify the disaster images uploaded on the online platforms. Daly and Thom [31] culled flicker images to detect the fire event using pretrained classifiers. Finally, Lagerstrom *et al.* [32] developed a system to classify whether the image indicates a fire event or not. In contrast to these works that broadly developed binary classifier for classifying disaster vs. non-disaster images using CNN approach, we focus on identifying multiple disaster categories from the disaster-related images.

B. MULTIMODAL APPROACHES

In recent years, researchers have used multimodal data (i.e., image and text) to find disaster related information from social media, as information from both modalities often provide valuable insights for disaster classification. Most of the works employed fusion-based [33] approach to aggregate the multimodal features. Chen *et al.* [34] studied the relation between the images and texts and utilize visual features along with socially relevant contextual features (e.g., time of posting, the number of comments, re-tweets) to identify disaster information. Mouzannar *et al.* [7] explored damage detection by focusing on human and environmental damage related posts. They used the Inception pre-trained model for visual feature extraction and designed a CNN architecture for textual features. Similarly, Rizk *et al.* [35] proposed a multimodal architecture to classify the Twitter data into infrastructure and natural damage categories. Ofli *et al.* [8] also presented a multimodal approach for classifying the tweets into two categories: informative task (e.g., informative vs. non-informative) and humanitarian task (e.g., affected individuals, rescue volunteering or donation effort, infrastructure and utility damage). They used CNN based approach for extracting the visual and textual features. Gautam *et al.* [36] showed a comparison between unimodal and multimodal methods on CrisisMMD [37] dataset. They utilized the late fusion [38] approach for combining the image-tweet pairs. All the works reported significant performance improvement using multimodal information in contrast to their counterparts that utilize uni-modal information.

Motivated by the success of these multimodal approaches we focused on effectively utilizing features from text and images using BiLSTM and CNN models and then fusing them to form a joint representation for the classification. However, unlike the above multimodal-based approaches which used simple CNN/RNN models or n-gram features, we extract the textual features using the BiLSTM network with attention mechanism to address the long-range dependency problem.

TABLE 1. Number of samples in train, and test set for each class.

Class Names	Train	Test
Damage nature (DN)	459	55
Damage infrastructural (DI)	1246	144
Human damage (HD)	219	21
Fires	309	37
Flood	348	36
Non damage (ND)	2666	291
Total	**5247**	**584**

III. PROBLEM FORMULATION AND THE DATASET

In this work, our goal is to automatically classify disaster types such as floods, fires, earthquake etc. from social media posts. Formally, we are given a dataset with M examples, thus the i^{th} sample can be represented as $\{X^i = (P^i, Y^i)\}$ where $i \epsilon [1, \ldots M]$ and $Y^i \epsilon (1, ..K)$. Here, P^i, and Y^i denotes the post, and the associated class label for the i^{th} data point. Each post P^i consists of two modalities: visual (v^i) and textual (t^i). Our model utilizes both v and t simultaneously to classify the P^i into one of the K classes. We discuss the disaster types and analyze the dataset below.

A. DISASTER TYPES

We experiment with a benchmark multimodal damage dataset[1] from Mouzannar *et al.* [7], which consists of damage-related images along with their associated tweets. The dataset contains following five different categories of disaster image-tweet pairs as well as one category of non-damage (ND) image-tweet pairs.

- Damage to infrastructure (DI): Posts that contain information about wrecked buildings, damaged cars, and destroyed bridges.
- Damage to nature (DN): Posts that contain icehouse, landslides, and falling trees related information.
- Fires (F): Posts that conveys forest and building fires related information.
- Floods (Fl): Posts that contain flood related images and tweets occurred in rural, urban, and cities.
- Human damage (HD): Posts that provide information about injuries and deceased people.

B. DATASET ANALYSIS

The dataset from Mouzannar *et al.* [7] consists of a total of $5,831$ image-tweet pairs, were the training and test sets contain $5,247$ and 584 samples, respectively. The class-wise breakdown of the train and test sets is reported in Table 1. We observe that the Non damage class has the highest number of samples $(2,957)$ and the human damage consists of the

[1] https://github.com/eftekhar-hossain/Disaster_IEEE-Access

TABLE 2. Training set statistics for the textual data.

Class	Total words	Unique words	Max tweet length (words)	Avg. no. of words per tweet
DN	13230	5417	345	28.82
DI	35202	10244	363	28.25
HD	8800	1902	354	40.18
ND	111686	25978	382	41.89
Fires	11809	4055	382	38.21
Flood	11774	4105	290	33.83

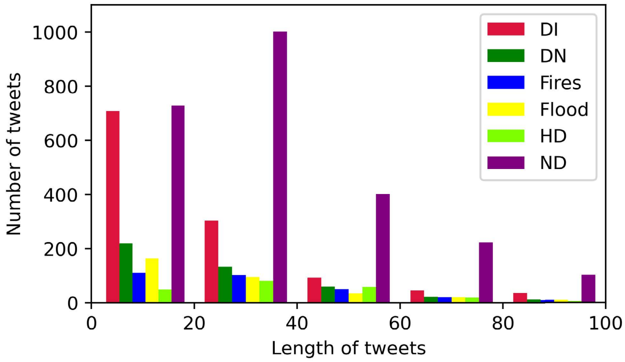


FIGURE 1. Tweet length frequency distribution of different classes.

lowest number of samples (240). The overall data distribution indicates that the dataset is somewhat imbalanced.

We have also analyzed the basic linguistic statistics including token statistics and tweet lengths. Table 2 shows that the average number of words per tweet are over 28 in all classes. We also notice that the *ND* class contained the highest number of total and unique words as this class has the maximum number of instances (2, 666) in the dataset. On the other hand, *HD* has lowest number of total words and unique words. Figure 1 further shows how the tweet length varies across the different classes. We observe that generally there are more shorter tweets than the longer ones and the most tweets contained less than 100 word. Overall, this distribution provides an idea of choosing the input text length during the training phase.

IV. METHODOLOGY

Figure 2 depicts our proposed multimodal architecture for disaster identification. The model consists of two parallel networks: one for visual feature extraction and another for textual feature extraction. We apply a pre-trained convolutional neural network (i.e., ResNet50) to extract the visual features and a BiLSTM model with the attention mechanism to obtain the textual features from the tweets. The features from both modalities are then aggregated to form a combined representation and passed into a softmax layer for classification. A brief description of each constituent part of the architecture is described in the following subsections.

A. DATA PREPROCESSING

We pre-process the image by resized it into 150 × 150 × 3 so that all images have the same size and also in this way we can then process these resized images more efficiently. Furthermore, the image pixels are scaled between 0 and 1 to reduce the computational complexity during the classifier model training. Concerning the textual modality, we discard all the hyperlinks in a tweet as well as all some special characters (e.g. !,@,$,%,&), punctuation symbols, and emoticons.

B. VISUAL FEATURE EXTRACTION

We apply the transfer learning [39] technique to obtain the visual features from the image. To this end, we use the pre-trained ResNet50 [40] model mainly because it can address the vanishing gradient problem by utilizing skip connections across different layers [41]. In order to adjust the ResNet50 for our task, we exclude the top two layers of the default model. We freeze initial 40 layers to use only the weights of the higher level visual features that were previously learned from the ImageNet [42] task. For the last 10 layers of the ResNet50 model, a global average max-pooling layer, and a dense layer, we retrain the model with new weights. The dense layer compute the visual features according to the following equation.

$$V_f^{(v)} = Relu\left(\sum_{j}^{d} W_{kj} * G_j + b_k\right) \tag{1}$$

Here, $V_f^{(i)} \in \mathbb{R}^{1\times d}$ represents the visual semantic expression extracted by the ResNet50 for v^{th} image and d denotes the number of hidden neurons in the dense layer. Also, G_j indicates j^{th} feature map generated by the global average max-pooling layer, W_{kj} denotes the weight matrix, b_k represents the bias vector for k^{th} dense node, and *Relu* is the activation function, respectively.

C. TEXTUAL FEATURE EXTRACTION

For textual feature extraction, we first transform the tweet into a vector representation and then use an embedding layer to obtain semantic representations (embedding features) of the words. We then feed the embedding features to the BiLSTM network, which produces the context-level feature vector for individual words. Finally, the attention layer finds the most significant textual features from this feature vector. We now describe each of these steps in details.

1) TEXT TO VECTOR REPRESENTATION

In order to generate an initial vector representation of the tweet, we first generate a numeric mapping of the words of $\tau[] = \{t_1, t_2, \ldots, t_M\}$, where t_i represents a tweet/text. To get this mapping, we first create a vocabulary (V) consisting of v unique words as $V = \{uw_1, uw_2, \ldots, uw_v\}$. The i^{th} words in a tweet $t_j = [w_1, w_2, \ldots, w_{l'}]$ is substituted by the corresponding index number $((i))$ of the words in V. By doing so, a tweet (t_j) is transformed into a sequence vector, $s' = [i_1, i_2, \ldots, i_{l'}]$. However, at this point, the obtained sequence vectors, $S' = \{s'_1, s'_2, \ldots, s'_M\}$ have variable lengths (l'), which is not appropriate for feature extraction and training.

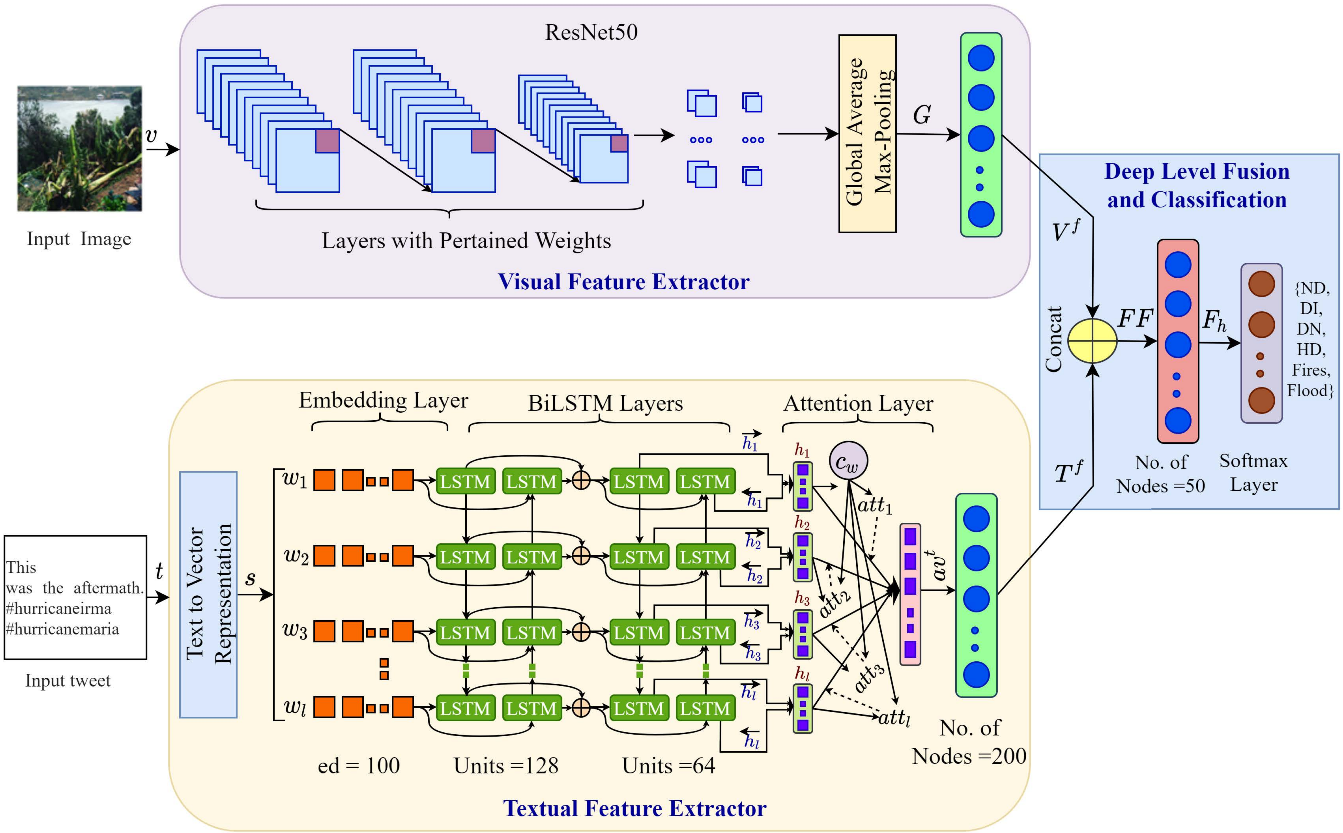


FIGURE 2. Our proposed multimodal architecture for disaster identification: the upper block represents the visual feature extractor module and the bottom block is the textual feature extractor module. Here, v and t indicates the preprocessed image and text respectively. The features extracted from the two modules are passed through the deep level fusion and classification layer to classify the sample.

Therefore, we transform S' into fixed-length sequences, $S = \{s_1, s_2, \ldots, s_M\}$, where a sequence (s_k) of S is a vector of size l. We choose $l = 150$ empirically based on the observation that most tweets in the dataset contain less than 100 words; therefore choosing such large number of dimensions for the vector would allow us to sufficiently capture the important information from different tweets.

2) EMBEDDING LAYER

After creating the initial vector representation S, it is necessary to encode the semantic information of the words (w_i) of a tweet to a global vector s_k^e. For this purpose, we first pass each sequence vector s_k in S into the Keras embedding layer to obtain word embedding vectors (we_i). We then simply concatenate these word embedding vectors according to the equation 2 to preserve the sequence of words.

$$s_k^e = |\vec{we}_1||\vec{we}_2||\vec{we}_3||\ldots.||\vec{we}_l|| \tag{2}$$

Here, $we_i \in \mathbb{R}^{1\times ed}$ represents the embedding vector of the i^{th} word. We keep the size of the embedding dimension large enough to capture the relationship between words $(ed = 100)$.

3) BiLSTM LAYERS

We apply the Bidirectional LSTM to generate a contextual representation of the input text from both backward and froward directions. Bidirectional LSTM [43] is an extension of long-established LSTM RNN architecture which is suited for abating the vanishing gradient problem that occurs due to the long context size. The model process the tweet from we_1 to we_l by the forward LSTM and from we_l to we_1 by the backward LSTM. For each word w_i, a forward LSTM generates the word feature $\vec{h_i}$, and a backward LSTM generates the word feature $\overleftarrow{h_i}$ using its memory blocks. The combined features h_i is calculated by Eq. (3).

$$h_i^{[p]} = \vec{h_i} \oplus \overleftarrow{h_i} \tag{3}$$

Here, $h_i^{[p]} \in \mathbb{R}^{1\times 2N}$ denotes the BiLSTM feature generated for i^{th} word in the p^{th} layer, where N represents the number of hidden units in the LSTM cell. The $\oplus$ is the concatenation symbol.

4) ATTENTION LAYER

Generally, all words in a tweet do not contribute equally in deciding whether the tweet should belong a particular class. Therefore, we utilize the attention mechanism [44] to emphasize on the most important words during the classification. The attention mechanism assigns a weight att_j to each individual word feature h_j of the BiLSTM layer with a focus on the output labels. Finally, we perform a weighted sum operation to generate an attentive feature vector av^t of the

t^{th} tweet. More formally,

$$e_i = tanh\left(W * h_i^{[2]} + b\right) \tag{4}$$

$$att_i = \frac{exp\left(e_i^T c_w\right)}{\sum_i^l exp\left(e_i^T c_w\right)} \tag{5}$$

$$av^t = \sum_i att_i h_i^{[2]} \tag{6}$$

Here, l is the length of a tweet and $h_i^{[2]}$ is the word feature vector obtained in the second BiLSTM layer, which is passed to a two-layer neural network to get the e_i as a hidden representation of $h_i^{[2]}$.

The weight matrix W and bias vector b is initialized during the neural network training. The influence of the words can be measured by calculating the similarity between e_i and a randomly initialized word-level context vector c_w. Afterward, by using the softmax function a normalized weight att_i is obtained for each word (i) in a tweet (t). The attention weights for a tweet, $\sum_i^l att_i = 1$. The larger the weight of att_i is, the more significant the word for classification. Finally, the attentive feature av^t for a tweet is fed to a dense layer consisting of d number of neurons. The output can be represented as in Eq. (7).

$$T_f^{(t)} = Relu\left(\sum_j^d W_{kj} * av_j^t + b_k\right) \tag{7}$$

Here, $T_f^{(t)} \in \mathbb{R}^{1\times d}$ represents a d-dimensional feature vector that resembles the t^{th} tweet feature, where d is the number of hidden neurons in the dense layer. Correspondingly, W_{kj}, b_k, and $Relu$ are represented as the weight matrix, bias vector, and the activation function.

D. DEEP LEVEL FUSION AND CLASSIFICATION

In order to create a shared representation of both modalities, we concatenate the output's of the dense layer obtained from the visual (V_f) and textual (T_f) modalities. To attain the deep level representation, we utilize an early fusion approach [45] which concatenate the visual and textual features. We use the same number of hidden nodes (d) in the last dense layer of both modalities. We select the same size to have an equal contribution from both the visual and textual sides. We set $d = 200$ empirically based on the highest accuracy on the validation set.

Suppose, the dataset contains M number of posts, where each post (P^i) consists of two types of information: visual (v^i), and textual (t^i). The visual, and textual feature extractor finds the visual, and textual feature as vectors $V_f^{(i)}$ and $T_f^{(i)}$, respectively. Then, the fusion of these vectors is computed as in Eq. (8).

$$FF^{(i)} = V_f^{(i)} \oplus T_f^{(i)} \tag{8}$$

where $FF^{(i)} \in \mathbb{R}^{1\times 2d}$ is the concatenation ($\oplus$) output of i^{th} visual and textual features. We then pass the fusion feature vector through the final hidden layer of n hidden neurons followed by a softmax layer for the classification. To mitigate the effect of overfitting a dropout [46] layer is added before the hidden layer. The process is illustrated in Eqs. (9) and (10).

$$F_h^{(i)} = Relu\left(\sum_j^n W_{qj} * FF_j + b_q\right) \tag{9}$$

$$Softmax(\vec{F_h})_r = \frac{e^{(F_h)_r}}{\sum_{j=1}^K e^{(F_h)_j}} \tag{10}$$

Here, $F_h^{(i)} \in \mathbb{R}^{1\times n}$ represents the final hidden layer output, where $n = 50$. The parameters W_{qj} and b_q are the weights and biases of the hidden layer and K represents the number of classes for the classification task.

TABLE 3. Hyperparameters value utilized for training of the proposed model. Here, L-1, and L-2 represents the layer-1 and layer-2 respectively.

Hyperparameters	Optimum value
Input text length	150
Embedding dimension (ed)	100
LSTM hidden units	128 (L-1), 64 (L-2)
Neurons (Dense layers)	200
Dropout	0.1
Loss function	*Categorical crossentropy*
Optimizer	*Adam*
Learning rate	$3e^{-3}$
Batch size	64
Epochs	100

E. MODEL HYPERPARAMETERS

We use the Keras tuner [47] to optimize hyperparameters including learning rate and batch size. We first configure the search space with different values for each hyperparameter (e.g. optimizer, learning rate, etc.) and then leverage the Hyperband [48] search algorithm to find the best hyperparameter values for the proposed model. The values are adjusted based on their impact on the validation set performance (i.e., accuracy). However, to reduce the computational cost, other hyperparameters such as number of hidden units, number of LSTM cells, dropout rate, and embedding dimension are not considered as those are empirically selected. Table 3 shows the optimized hyperparameter values of the proposed model.

The proposed model is compiled using the *categorical cross-entropy* loss function and *adam optimizer* with a learning rate of $3e^{-3}$. Furthermore, the model's training is performed for 100 epochs with 64 instances at each iteration. Additionally, the Keras checkpoint method has been utilized to stop the over training of the model by observing the validation accuracy up to five consecutive epochs.

V. EXPERIMENTS AND ANALYSIS

In this section, we first describe the baseline models that we compared with. We then present the comparative performance analysis of the proposed approach with these baselines. Finally, we provide an in-depth error analysis along with intrinsic performance analysis.

A. BASELINES

We consider three types of baselines based on the features they use: (i) visual only, (ii) textual only, and (iii) visual + textual.

1) VISUAL ONLY

For the visual (i.e., image) modality, we consider two state-of-the-art pertained CNN architectures: VGG19 and InceptionV3 along with the ResNet50 (Described in Section IV-B). These architectures are used for a wide range of image classification tasks. A variant of the VGG [49] model, VGG19 consists of 19 convolutional layers using a fixed kernel of size 3 × 3 at each layer. In contrast, InceptionV3 [50] is an advanced version of GoogLeNet [51], having several inception modules. Each module is associated with a series of stacked convolutional filters (1 × 1, 3 × 3, 5 × 5), making the architecture more robust in learning with fewer parameters. We excluded the top layers from both architectures and froze the initial layers except the last 10 layers of the networks to accomplish the task. We used the pre-trained weights of the initial layers while we retrained the last 10 layers and a global average max-pooling layer with new weights. Finally, a softmax layer is added for the classification.

2) TEXTUAL ONLY

We apply the following deep learning models for classifying the damage types using only the textual features: BiLSTM [52], CNN [53], BiLSTM + CNN [54], and BiLSTM + attention (A) [55]. We utilize the word embedding features with each model. The Keras embedding layer is initialized with the embedding dimension of 100 and settled the input text length at 150. The calculated features are then passed to every model.

The *BiLSTM* network consists of one layer with 128 hidden units. The final hidden state output of the BiLSTM layer is transferred to a softmax layer for the classification. Similarly, *CNN* architecture is constructed, having one convolutional layer with 128 filters of kernel size 2 and a max-pooling layer of window size 2. A flattening layer is added before perform the softmax classification. Subsequently, *BiLSTM* + *CNN* network is configured by stacking the BiLSTM and CNN architecture with the same parameters. Eventually, the output of the stacked network is passed to the softmax layer. In another architecture, an attention layer is added after the BiLSTM layer and creates *BiLSTM* + *Attention* model (Described in Section IV-C4). The obtained attention vector is then passed into a dense layer of 20 neurons, followed by a softmax layer for the classification.

3) MULTIMODAL (TEXTUAL + VISUAL) BASED MODELS

We experimented with 11 different models for combining text and image modalities, namely VGG19 + BiLSTM, VGG19 + CNNText, VGG19 + BiLSTM + CNN, VGG19 + BiLSTM + Attention, Inception + BiLSTM, Inception + CNNText, Inception+ BiLSTM + CNN, Inception + BiLSTM + Attention, ResNet50 + BiLSTM, ResNet50 + CNNText, and ResNet50 + BiLSTM + CNN-Text. Instead of using a softmax layer at the end of each model, a hidden layer of 200 neurons is placed. Subsequently, the hidden layers from both visual and textual sides are concatenated using the early fusion approach (Described in Section IV-D) to produce a shared representation of both modalities. We then pass the joint representation into a dense layer of 64 neurons, followed by a softmax layer. After the concatenation operation, we add a dropout layer (dropout rate = 10%) to abate the chance of layer overfitting.

B. IMPLEMENTATION SETTINGS

All the visual and textual models are compiled using the 'Adam' optimizer with a learning rate of $1e^{-5}$ and $1e^{-4}$, respectively. In cotrast, for multimodal case, the models having VGG19 and Inception are utilized 'RMSProp' optimizer, where the learning rate is settled at $1e^{-3}$. In contrast, multimodal models having ResNet50 are complied using 'Adam' (learning rate = $3e^{-3}$) optimizer. Other hyperparameters (i.e., loss function, batch size, epochs) and training configuration (i.e., Keras checkpoint) kept the same as described in Section IV-E.

Training and testing of the models are conducted on the Google Colab platform using Python = 3.6.9. Models are implemented using Keras = 2.4.0 with TensorFlow = 2.3.0 framework. For data preparation and evaluation, pandas =3.6.9 and Scikit-learn=0.22.2 packages have been used. We use 10% of the training dataset for validation and the remaining data for training. Finally, the trained models are evaluated using the test set instances.

C. RESULTS

For performance comparison, we use precision (P), recall (R), and weighted F1-score. For efficient comparison of the model's performance across different classes, the misclassification rate has been used as one of the measures. We use the weighted F1-score metric to determine the superiority of the models. However, we also report P, R, and MR for deeper analysis of the model's performance on the individual classes.

Table 4 shows the results of both unimodal (textual only and visual only) and multimodal (textual+visual) models. We observe that among the models that utilize the visual modality only, VGG19 and ResNet50 perform slightly better than the Inception model in terms of weighted F1. Textual models perform better than their visual only counterparts. Among them, the CNNText and BiLSTM+CNNText perform similarly. Interestingly, the performance is dramatically increased by 3.18% when the Attention is incorporated with BiLSTM (BiLSTM+Attention) compared to the BiLSTM only model. Overall, this suggests the usefulness of incorporating attention mechanism for our disaster type classification task.

TABLE 4. Performance comparison of different unimodal and multimodal models on the test set. Here, P, R, and WF denotes the precision, recall, and weighted F1-score respectively.

Approach	Models	P(%)	R(%)	WF(%)
Visual	VGG19 [49]	81.06	81.51	81.21
	Inception [50]	77.41	77.91	77.38
	ResNet50 [40]	81.88	81.51	81.63
Textual	BiLSTM	85.92	85.45	85.57
	CNNText	84.97	84.25	84.45
	BiLSTM+CNNText	85.54	84.42	84.70
	BiLSTM+Attention	89.14	88.87	88.75
Multimodal	VGG19+BiLSTM	81.98	76.20	78.14
	VGG19+CNNText	74.39	73.46	72.57
	VGG19+BiLSTM+CNNText	78.24	77.74	77.67
	VGG19+BiLSTM+Attention	89.54	89.38	89.19
	Inception+BiLSTM	82.21	74.48	77.01
	Inception+CNNText	79.66	79.10	78.28
	Inception+BiLSTM+CNNText	77.29	78.08	77.38
	Inception+BiLSTM+Attention	81.18	80.82	80.48
	ResNet50+BiLSTM	84.22	81.34	81.90
	ResNet50+CNNText	77.68	78.42	77.45
	ResNet50+BiLSTM+CNNText	80.30	79.62	79.84
	ResNet50+BiLSTM+Attention **(Proposed Method)**	93.35	93.15	**93.21**

TABLE 5. 5-fold cross validation results on the training set. Here, WF and Std denotes the weighted F1-score and standard deviation respectively.

Models	WF (Mean ± Std)
ResNet50 [40]	81.34 ± 0.816
BiLSTM+Attention	88.58 ± 0.688
ResNet50+BiLSTM+CNNText	79.45 ± 0.322
Inception+BiLSTM+Attention	79.96 ± 0.572
ResNet50+BiLSTM	81.42 ± 0.233
VGG19+BiLSTM+Attention	89.55 ± 0.422
ResNet50+BiLSTM+Attention **(Proposed)**	93.11 ± 0.360

Among the models that aggregate both visual and textual features, only two models performed better than the best unimodal counterpart (BiLSTM+Attention). In particular, the multimodal model (VGG19 + BiLSTM + Attention) showed a noticeable rise in WF-score (89.19%). Our proposed method (ResNet50 + BiLSTM + Attention) achieves state-of-the-art result by achieving the highest WF-score of 93.21% with a margin of 4.02% over VGG19 + BiLSTM + Attention.

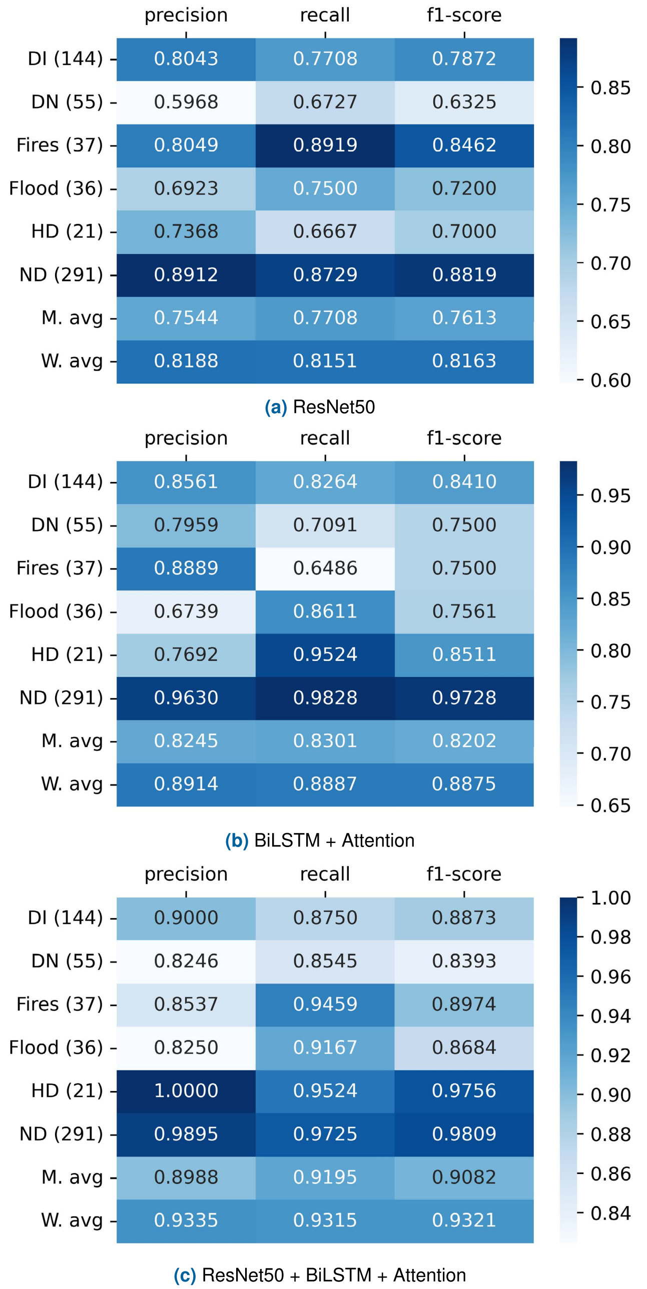
To verify the efficacy of the of the models' performance, we performed 5-fold cross validation [56]. The cross validation has been performed with seven models including the proposed (ResNet50+BiLSTM+Attention), best visual (ResNet50), best textual (BiLSTM+Attention), and best four multimodal models (such as ResNet50+BiLSTM+CNNText, Inception+BiLSTM+Attention, ResNet50+BiLSTM, and VGG19+BiLSTM+Attention). The results of cross-validation for these models are shown in Table 5. The results exhibits that the proposed model achieved the highest mean weighted f1-score of 93.10% with a deviation of 0.36031. The standard deviation and mean values of the models revealed that the different split of the dataset has not significant impact on the model's performance.

1) INVESTIGATING CLASSIFICATION REPORTS

To obtain deeper insights about the performance of different models on the individual classes, we examine their classification reports (Figure 3). In the case of the best visual model (ResNet50), DI, HD, and ND classes obtained the high precision values (0.804, 0.736, 0.891) and low recall scores (0.770, 0.666, 0.872) (shown in Figure 3(a)). These results indicate that some instances of each class are incorrectly identified as other classes. In contrast, the DN, Fires, and Flood classes attain high recall and low precision scores. We also notice that the classes such as DN, Flood, and HD have relatively low f1-scores of 0.632, 0.72, and 0.70 respectively, compared to the DI (0.787), Fires (0.846), and ND (0.881). In the case of the best textual model (BiLSTM + Attention), the overall performance is increased across various classes as depicted in Figure 3(b). However, he f1-score of the *Fires* class is surprisingly reduced from 0.846 to 0.75, which suggests that visual information is more critical than the textual one for this particular class. Finally, in the case of the proposed model, the precision and recall score of all the classes are improved by considerable margins compared to the best textual model (BiLSTM+Attention).

Overall, The results showed that the proposed approach (ResNet50 + BiLSTM + Attention) outperformed all the visual, textual, and multimodal models in classifying the disaster information. ResNet50 obtained the highest WF score among the visual models, while BiLSTM+Attention attained the maximum WF score among the textual only models. We also notice that models that utilize both textual and visual features do not necessarily improve the performance over the textual only models. This indicates that the superior performance of our model was primarily due to the incorporation of attention in the textual modality side which effectively captures the input text.

To analyze further the cases where our model makes a difference, Figure 4 shows the confusion matrices of different models. We notice that the visual only model (ResNet50) wrongly classified 18 instances as 'Non Damage class' (ND) out of 144 instances of 'Damage Infrastructure' (DI) in Figure 4(a). In contrast, the textual and proposed models incorrectly predicted 6 and 3 instances, respectively (Figures 4(b) and 4(c)). These results indicate out that the

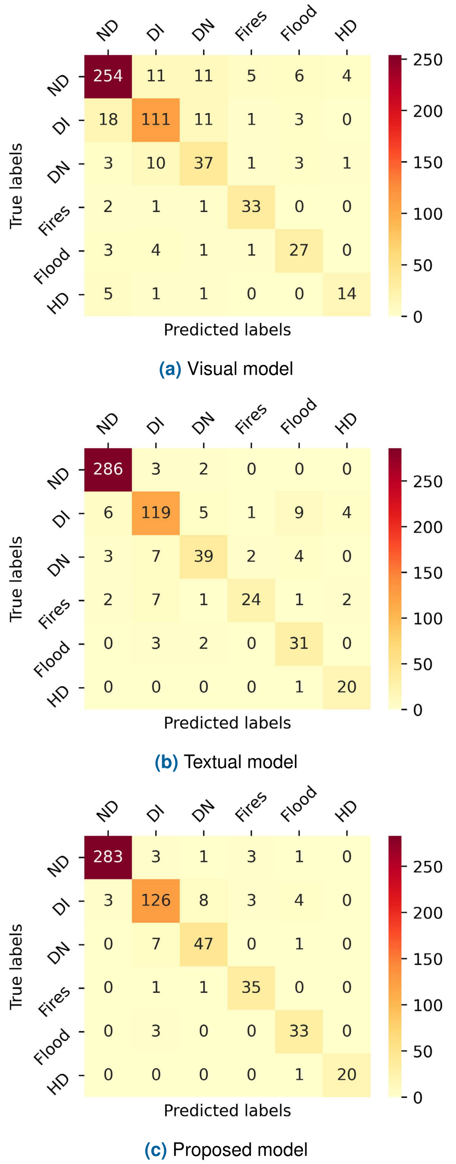


(a) ResNet50

(b) BiLSTM + Attention

(c) ResNet50 + BiLSTM + Attention

FIGURE 3. Classification report of the best visual (ResNet50), textual (BiLSTM+A) and proposed multimodal (ResNet50+ B + A) models. M. avg and W. avg denotes the macro and weighted average values, respectively.



(a) Visual model

(b) Textual model

(c) Proposed model

FIGURE 4. Confusion matrices of the best visual (ResNet50), textual (BiLSTM + Attention) and proposed multimodal (ResNet50 + BiLSTM + Attention) models.

fusion of both modalities' information aids the proposed model to comprehend the "Damage Infrastructure (DI)" category better, thus significantly curtail the prediction errors. A different phenomenon is noticed where the predicted label is "Fires" but the actual label is "Non Damage". In particular, the textual model did not misclassify a single instance in Figure 4(b) whereas the visual and multimodal models wrongly classified 5 and 3 instances in Figure 4(a) and 4(c), respectively. This suggests that for certain categories, unimodal models can be more effective and further investigation is required to address the noise that maybe introduce when two modalities are combined.

Figure 5 compares among the proportion of instances that are misclassified in different classes. We notice that most of the mistakes made by the best visual model (ReNet50) with "HD" (33%), "DN" (32.72%), "Flood" (25%), and "DI" (23%) categories. In contrast, the misclassification rate for "ND" (12.71%) and "Fires" (10.8%) classes are comparatively lower than others, which is also evident from Figure 4(a) where the number of misclassified instances are shown. Concerning the textual model (BiLSTM + Attention), we observed that the MR is reduced for almost every class except the "Fires" class (rise from 10.8% to 35%). Nevertheless, MR for the textual model is decreased by a more

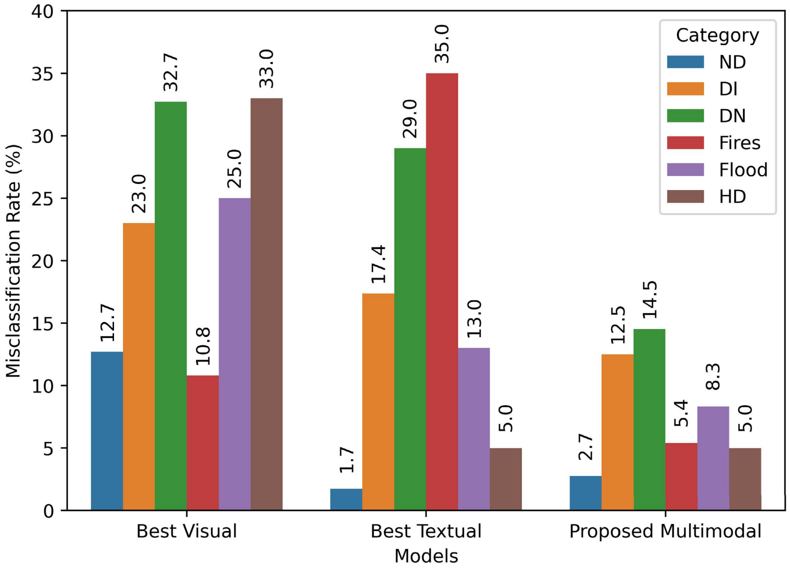


FIGURE 5. Error rate analysis of the individual classes with different approaches.

significant margin of approximately 10%, 12%, and 29% respectively for "ND", "Flood", and "HD" classes. Finally, the proposed model produced fewer mistakes across different classes (also shown in 4(c)). While with the "ND" category, MR increased by approximately 1% (from 1.74% to 2.74%) compared to the textual model, other classes experienced a considerable drop of approximately 5% (DI), 15% (DN), 30% (Fires), and 5% (Flood), respectively.

2) QUALITATIVE ANALYSIS

Table 6 shows example samples along with outputs from different models. Overall, these outputs illustrate the need for combining two modalities. For example, in Table 6 sample (1), the visual model wrongly classifies the image into the "Damage Nature (DN)" class since the image contains some trees and leaves. The text only model also classify the image as DN because of the presence of words like '#fallentree' and '#treebranch'. However, when features from these two modalities are combined together, they do not provide any insightful evidence for the model to infer this image-text data pair as "DN" anymore. Likewise, in Table 6 sample(4), the visual model classifies the image as "Damage Infrastructural (DI)", and the textual model also makes incorrect prediction due to the presence of the word '#buildingcollapse'. However, fusing information from both modalities leads the predict the "Fires" category correctly. Finally, in Table 6 sample (5), the visual model consider the image as "DI" because it shows broken roads like patterns, whereas the textual model reckons the text is from the "Flood" class as it mentions words related to flood such as '#flood' and '#tsunami'. However, the proposed multimodal model conjoined the information coming from both modalities and yielded the correct prediction (i.e., "Damage Nature"). Overall, these analyses confirm the suitability of the proposed multimodal approach over the other models in classifying the damage information.

3) INTRINSIC PERFORMANCE ANALYSIS

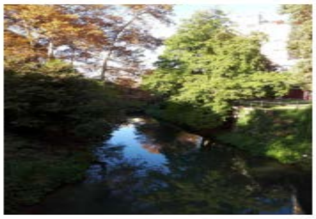
To further understand the possible reason for this superior performance of our proposed approach over other models, we performed an intrinsic performance analysis. In this analysis, we focus on how the attention layer impacts the performance of the proposed method by comparing with its counterpart (ResNet50+BiLSTM), where the only difference is the absence of the attention mechanism.

Figure 6 shows the feature visualization of the two multimodal models (with and without attention). The projected data points are obtained after applying the principal component analysis (PCA) [57] on the extracted hidden features. We observe that the multimodal model without attention (Figure 6 (a)) did not separate all the classes accurately as plenty of overlaps are visible among the classes such as "Non-Damage (ND)", "Damage Infrastructural (DI)", and "Damage Nature (DN)". On the other hand, when the model adds the attention layer in the text modality side of the same multimodal model, a noticeable difference is observed (Figure 6 (b)). Incorporation of the attention layer made the classes like "ND", "Fires", "DI", and "DN" are more separable and thus enhances the performance across different classes.

4) PROPOSED VS EXISTING METHODS

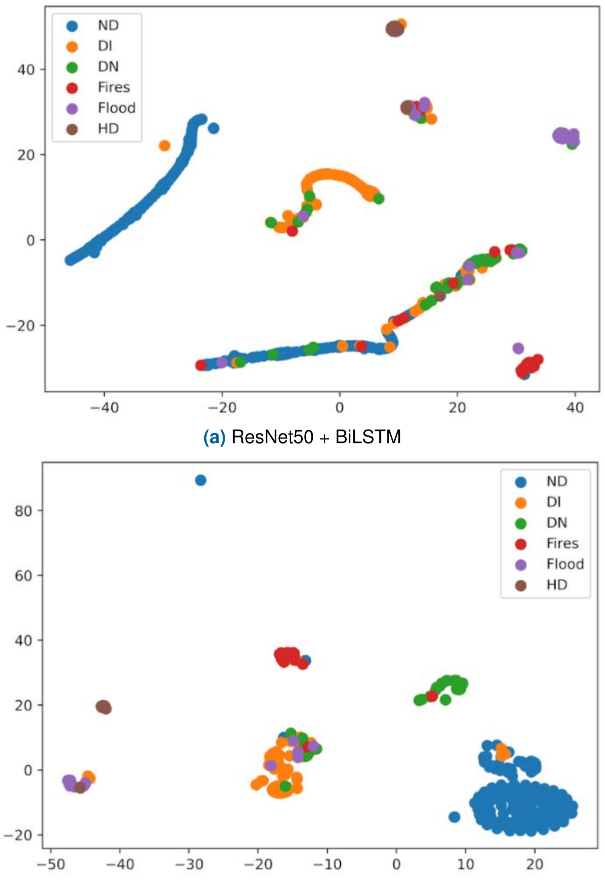
As per this work exploration, no significant work has been conducted over the multimodal dataset used in this research except the work done by [7]. However, the past study is not exactly comparable with the proposed method due to the differences in evaluation measures and dataset distribution. Therefore, for the comparison, we adopt several recent

TABLE 6. Example image and tweet text pairs where model aggregation of the input modalities produce better results. The symbol (✓) and (✗) indicates the correct and incorrect prediction respectively.

Sample	Image	Tweet	Actual label	Predicted label
(1)		MooseMonday with my favorites! A couple #bullmoose from the weekend! #moose #wildlife #wildlifephotography #mammal #wilderness #wildernessculture	ND	**Visual Modality:** DN (✗) **Visual Modality:** DN (✗) **Proposed Multimodal:** ND(✓)
(2)		#sandy #youwhore massive #treebranch fell and took out two 8 foot sections of the fence in the pic.#fallentree #30ftdrop #sandydamage	DI	**Visual Modality:** DN (✗) **Textual Modality:** DN (✗) **Proposed Multimodal:** DI(✓)
(3)		Please curtail this hazardous 20+ year practice.#csi #uci #bordertown #newportbeach #mudslide #caution #landslide #smashingpumkins	DN	**Visual Modality:** DI(✗) **Textual Modality:** DI(✗) **Proposed Multimodal:** DN(✓)
(4)		#terriblefire #plascobuilding #nostalgia #tragedy #buildingcollapse	Fires	**Visual Modality:** DI(✗) **Textual Modality:** DI(✗) **Proposed Multimodal:** Fires(✓)
(5)	LIVE TRACKING SANDY CRANE HANGING FROM BUILDING AT 57TH ST. & 7TH AVE. IN MANHATTAN TUESDAY NEW JERSEY TURNPIKE AUTHORITY WILL C	#hurricane #sandy #hurricanesandy #sandydamage #nyc #nj #crane.	DI	**Visual Modality:** ND(✗) **Textual Modality:** Flood(✗) **Proposed Multimodal:** DI(✓)
(6)		#landslide #naturaldisaster #earthquake #tsunami #flood #nature #destroy #environment #disaster #tornadoes #huricane #volcaniceruption #draught.	DN	**Visual Modality:** DI (✗) **Textual Modality:** Flood (✗) **Proposed Multimodal:** DN (✓)

techniques that have been explored on similar tasks. For uniformity, existing methods [7], [8], [11], [18], [21]–[23], [29] have been implemented on the same dataset including the proposed method.

Table 7 shows the result of the comparison. Mouzannar *et al.* [7] developed a multimodal model, where they used a pre-trained Inception model for image modality, and a convolutional neural network [58] for textual modality. By replicating their architecture, we obtained a WF-score of 92.21%. Ofli *et al.* [8] utilized VGG16 + CNNText which has achieved a WF-score of 75.11%. Kumar *et al.* [11] applied VGG16 (for image) + LSTM (for text) and obtained a WF of 77.84%. Three other works [21]–[23] are implemented considering only the textual modality, where custom and



(a) ResNet50 + BiLSTM

(b) ResNet50 + BiLSTM + Attention

FIGURE 6. Scatter plots of test input features extracted by the multimodal models (a) without attention layer and (b) with attention layer.

TABLE 7. Results of comparison concerning WF-score.

Method	Modality	WF(%)
Mouzannar et al. [7]	Image+Text	92.14
Ferda et. al [8]	Image+Text	75.11
Kumar et. al [11]	Image+Text	77.84
Nguyen et al [29]	Image-only	75.17
Caragea et al. [21]	Text-only	75.23
Aipe et. al. [22]	Text-only	76.76
Yu et. al. [23]	Text-only	78.47
Xiao et. al [18]	Text-only	86.05
Proposed	**Image+Text**	**93.21**

pre-trained CNN's are applied. These three methods also achieved the WF-scores of below 80%. Another work [18] employed logistic regression classifier for the text-based classification and achieved an 86.05% WF-score. The comparative analysis illustrates that the proposed technique outperformed the existing works by achieving the highest WF-score of 93.21%. In particular, it is almost 1% higher WF-score than the multimodal (92.14%) [7] method and 7% higher than the unimodal (86.05%) [18] technique.

VI. CONCLUSION

We have presented a multimodal approach that can effectively learn from the image and text data to classify the damage-related contents from Twitter. We utilize the pre-trained ResNet model for visual feature extraction and the attention mechanism with a BiLSTM model to extract the tweet features. The early fusion approach is used to aggregate both modalities' features. Besides, this work investigated various visual (i.e., VGG19, Inception) and textual (i.e., BiLSTM, CNN, BiSTM+CNN, BiLSTM+Attention) approaches for the baseline evaluation and constructed several multimodal models by exploiting them. The evaluation results revealed that the proposed model outperforms the baseline unimodal (i.e., image, text) and multimodal models by acquiring the highest weighted F1-score of 93.21%. Moreover, the comparative analysis illustrated that the proposed method outcome is approximately 1% and 7% ahead of the existing start-of-the-art models. Thus, the results confirmed the effectiveness of the proposed method in identifying the disaster content based on multimodal information. The error analysis further showed that it is difficult to identify the damage and non-damage contents by analyzing only one modality. At the same time, intrinsic performance analysis elucidated that incorporating an attention mechanism boosts the overall performance.

Despite achieving better performance than unimodal approaches, there are still rooms for improving the proposed method. In the future, we would like to explore different multimodal fusion approaches along with multitask learning technique for the disaster identification task. Besides, we aim to capture the combination of visual and textual features more effectively by employing the state of the art visual (i.e., Vision transformer [59]), textual (i.e., BERT [60], XLM-R [61]), and multimodal (i.e., VL- BERT [62], Visual BERT [63]) transformer models.

REFERENCES

[1] K. K. Kapoor, K. Tamilmani, N. P. Rana, P. Patil, Y. K. Dwivedi, and S. Nerur, "Advances in social media research: Past, present and future," *Inf. Syst. Frontiers*, vol. 20, no. 3, pp. 531–558, Jun. 2018.

[2] J. Kim and M. Hastak, "Social network analysis: Characteristics of online social networks after a disaster," *Int. J. Inf. Manage.*, vol. 38, no. 1, pp. 86–96, 2018.

[3] J. Son, H. K. Lee, S. Jin, and J. Lee, "Content features of tweets for effective communication during disasters: A media synchronicity theory perspective," *Int. J. Inf. Manage.*, vol. 45, pp. 56–68, Apr. 2019.

[4] A. Elbanna, D. Bunker, L. Levine, and A. Sleigh, "Emergency management in the changing world of social media: Framing the research agenda with the stakeholders through engaged scholarship," *Int. J. Inf. Manage.*, vol. 47, pp. 112–120, Aug. 2019.

[5] R. Dubey, A. Gunasekaran, S. J. Childe, D. Roubaud, S. F. Wamba, M. Giannakis, and C. Foropon, "Big data analytics and organizational culture as complements to swift trust and collaborative performance in the humanitarian supply chain," *Int. J. Prod. Econ.*, vol. 210, pp. 120–136, Apr. 2019.

[6] S. Akter and S. F. Wamba, "Big data and disaster management: A systematic review and agenda for future research," *Ann. Oper. Res.*, vol. 283, no. 1, pp. 939–959, 2019.

[7] H. Mouzannar, Y. Rizk, and M. Awad, "Damage identification in social media posts using multimodal deep learning," in *Proc. ISCRAM*, 2018, pp. 1–15.

[8] F. Ofli, F. Alam, and M. Imran, "Analysis of social media data using multimodal deep learning for disaster response," 2020, *arXiv:2004.11838*.

[9] F. Ofli, P. Meier, M. Imran, C. Castillo, D. Tuia, N. Rey, J. Briant, P. Millet, F. Reinhard, M. Parkan, and S. Joost, "Combining human computing and machine learning to make sense of big (aerial) data for disaster response," *Big Data*, vol. 4, no. 1, pp. 47–59, Mar. 2016.

[10] P. Jain, B. Schoen-Phelan, and R. Ross, "Automatic flood detection in Sentinei-2 images using deep convolutional neural networks," in *Proc. 35th Annu. ACM Symp. Appl. Comput.*, Mar. 2020, pp. 617–623.

[11] A. Kumar, J. P. Singh, Y. K. Dwivedi, and N. P. Rana, "A deep multi-modal neural network for informative Twitter content classification during emergencies," *Ann. Oper. Res.*, Jan. 2020, doi: 10.1007/s10479-020-03514-x.

[12] T. G. Mondal, M. R. Jahanshahi, R.-T. Wu, and Z. Y. Wu, "Deep learning-based multi-class damage detection for autonomous post-disaster reconnaissance," *Struct. Control Health Monitor.*, vol. 27, no. 4, 2020, Art. no. e2507.

[13] M. Imran, F. Ofli, D. Caragea, and A. Torralba, "Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions," *Inf. Process. Manage.*, vol. 57, no. 5, 2020, Art. no. 102261. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457320306002

[14] L. Dwarakanath, A. Kamsin, R. A. Rasheed, A. Anandhan, and L. Shuib, "Automated machine learning approaches for emergency response and coordination via social media in the aftermath of a disaster: A review," *IEEE Access*, vol. 9, pp. 68917–68931, 2021.

[15] C. Castillo, M. Imran, P. Meier, J. Lucas, J. Srivastava, H. Leson, F. Ofli, and P. Mitra, "Together we stand-supporting decision in crisis response: Artificial intelligence for digital response and micromappers," in *OCHA and Partners*. Istanbul: Tudor Rose, World Humanitarian Summit, 2016, pp. 93–95.

[16] C. Caragea, N. J. McNeese, A. R. Jaiswal, G. Traylor, H. W. Kim, P. Mitra, D. Wu, A. H. Tapia, C. L. Giles, B. J. Jansen, and J. Yen, "Classifying text messages for the Haiti earthquake," in *Proc. ISCRAM*. Princeton, NJ, USA: Citeseer, 2011, pp. 1–10.

[17] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, "AIDR: Artificial intelligence for disaster response," in *Proc. 23rd Int. Conf. world wide web*, 2014, pp. 159–162.

[18] Q. Huang and Y. Xiao, "Geographic situational awareness: Mining tweets for disaster preparedness, emergency response, impact, and recovery," *ISPRS Int. J. Geo-Inf.*, vol. 4, no. 3, pp. 1549–1568, 2015.

[19] D. Nguyen, K. A. Al Mannai, S. Joty, H. Sajjad, M. Imran, and P. Mitra, "Robust classification of crisis-related data on social networks using convolutional neural networks," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 11, no. 1, 2017, pp. 1–4.

[20] O. Sharif, E. Hossain, and M. M. Hoque, "Combating hostility: COVID-19 fake news and hostile post detection in social media," *CoRR*, vol. abs/2101.03291, 2021.

[21] C. Caragea, A. Silvescu, and A. H. Tapia, "Identifying informative messages in disaster events using convolutional neural networks," in *Proc. Int. Conf. Inf. Syst. Crisis Response Manage.*, 2016, pp. 137–147.

[22] A. Aipe, N. Mukuntha, A. Ekbal, and S. Kurohashi, "Deep learning approach towards multi-label classification of crisis related tweets," in *Proc. 15th ISCRAM Conf.*, 2018, pp. 1–13.

[23] M. Yu, Q. Huang, H. Qin, C. Scheele, and C. Yang, "Deep learning for real-time social media text classification for situation awareness–using Hurricanes Sandy, Harvey, and IRMA as case studies," *Int. J. Digit. Earth.*, vol. 12, no. 11, pp. 1230–1247, Nov. 2019.

[24] H. Li, N. Guevara, N. Herndon, D. Caragea, K. Neppalli, C. Caragea, A. C. Squicciarini, and A. H. Tapia, "Twitter mining for disaster response: A domain adaptation approach," in *Proc. ISCRAM*, 2015, pp. 1–7.

[25] D. Graf, W. Retschitzegger, W. Schwinger, B. Pröll, and E. Kapsammer, "Cross-domain informativeness classification for disaster situations," in *Proc. 10th Int. Conf. Manage. Digit. EcoSyst.*, Sep. 2018, pp. 183–190.

[26] K. Rudra, P. Goyal, N. Ganguly, M. Imran, and P. Mitra, "Summarizing situational tweets in crisis scenarios: An extractive-abstractive approach," *IEEE Trans. Computat. Social Syst.*, vol. 6, no. 5, pp. 981–993, Oct. 2019.

[27] M. A. Cameron, R. Power, B. Robinson, and J. Yin, "Emergency situation awareness from Twitter for crisis management," in *Proc. 21st Int. Conf. companion World Wide Web (WWW) Companion*, 2012, pp. 695–698.

[28] N. Chaudhuri and I. Bose, "Application of image analytics for disaster response in smart cities," in *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, 2019, pp. 1–10.

[29] D. T. Nguyen, F. Ofli, M. Imran, and P. Mitra, "Damage assessment from social media imagery data during disasters," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Jul. 2017, pp. 569–576.

[30] F. Alam, M. Imran, and F. Ofli, "Image4act: Online social media image processing for disaster response," in *Proc. 2017 IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining 2017*, 2017, pp. 601–604.

[31] S. Daly and J. A. Thom, "Mining and classifying image posts on social media to analyse fires," in *Proc. ISCRAM*. Princeton, NJ, USA: Citeseer, 2016, pp. 1–14.

[32] R. Lagerstrom, Y. Arzhaeva, P. Szul, O. Obst, R. Power, B. Robinson, and T. Bednarz, "Image classification to support emergency situation awareness," *Frontiers Robot. AI*, vol. 3, p. 54, Sep. 2016.

[33] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, 2016, pp. 284–288.

[34] T. Chen, D. Lu, M.-Y. Kan, and P. Cui, "Understanding and classifying image tweets," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 781–784.

[35] Y. Rizk, H. S. Jomaa, M. Awad, and C. Castillo, "A computationally efficient multi-modal classification approach of disaster-related Twitter images," in *Proc. 34th ACM/SIGAPP Symp. Appl. Comput.*, 2019, pp. 2050–2059.

[36] A. K. Gautam, L. Misra, A. Kumar, K. Misra, S. Aggarwal, and R. R. Shah, "Multimodal analysis of disaster tweets," in *Proc. IEEE 5th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2019, pp. 94–103.

[37] F. Alam, F. Ofli, and M. Imran, "CrisisMMD: Multimodal Twitter datasets from natural disasters," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 12, no. 1, 2018, pp. 1–9.

[38] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *Proc. IEEE 23rd Int. Conf. Inf. Fusion (FUSION)*, Jul. 2020, pp. 1–6.

[39] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. Hershey, PA, USA: IGI Global, 2010, pp. 242–264.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.

[41] D. Wu, Y. Wang, S.-T. Xia, J. Bailey, and X. Ma, "Skip connections matter: On the transferability of adversarial examples generated with ResNets," 2020, *arXiv:2002.05990*.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[43] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, no. 5, pp. 602–610, 2005.

[44] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[45] C. T. Duong, R. Lebret, and K. Aberer, "Multimodal classification for analysing social media," 2017, *arXiv:1708.02099*.

[46] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[47] T. O'Malley. (2019). *Keras Tuner*. [Online]. Available: https://github.com/keras-team/keras-tuner

[48] L. Li, K. G. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Efficient hyperparameter optimization and infinitely many armed bandits," 2016, *arXiv:1603.06560*.

[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations (ICLR)*. San Diego, CA, USA, May 2016. [Online]. Available: https://arxiv.org/abs/1409.1556

[50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015, *arXiv:1512.00567*.

[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[52] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," *IEEE Access*, vol. 7, pp. 51522–51532, 2019.

[53] A. Jacovi, O. S. Shalom, and Y. Goldberg, "Understanding convolutional neural networks for text classification," 2018, *arXiv:1809.08037*.

[54] O. Sharif, E. Hossain, and M. M. Hoque, "TechTexC: Classification of technical texts using convolution and bidirectional long short term memory network," in *Proc. 17th Int. Conf. Natural Lang. Process. (ICON)*. Patna, India: NLP Association of India (NLPAI), Dec. 2020, pp. 35–39. [Online]. Available: https://aclanthology.org/2020.icon-techdofication.8

[55] Y. Zhang, J. Wang, and X. Zhang, "YNU-HPCC at SemEval-2018 task 1: BiLSTM with attention based sentiment analysis for affect in tweets," in *Proc. 12th Int. Workshop Semantic Eval.*, New Orleans, LA, USA: Association for Computational Linguistics, Jun. 2018, pp. 273–278. [Online]. Available: https://www.aclweb.org/anthology/S18-1040

[56] T.-T. Wong and P.-Y. Yeh, "Reliable accuracy estimates from k-fold cross validation," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1586–1594, Apr. 2019.

[57] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.

[58] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Oct. 2014, pp. 1746–1751. [Online]. Available: https://aclanthology.org/D14-1181

[59] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and J. Uszkoreit, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, Oct. 2020, pp. 1–21.

[60] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[61] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2019, *arXiv:1911.02116*.

[62] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "VL-BERT: Pre-training of generic visual-linguistic representations," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–15.

[63] L. Harold Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," 2019, *arXiv:1908.03557*.

EFTEKHAR HOSSAIN (Graduate Student Member, IEEE) is currently pursuing the M.Sc. degree in electronics and telecommunication engineering with the Chittagong University of Engineering and Technology (CUET), Bangladesh. He is a Lecturer with the Department of Electronics and Telecommunication Engineering, CUET. His research interests include natural language processing, computer vision, and data science.

MOHAMMED MOSHIUL HOQUE (Senior Member, IEEE) received the Ph.D. degree from the Department of Information and Computer Sciences, Saitama University, Japan, in 2012. He is currently a Distinguish Professor with the Department of Computer Science and Engineering (CSE), Chittagong University of Engineering and Technology (CUET). He is also serving as the Dean of the Faculty of Electrical and Computer Engineering (ECE), CUET, and the Director of the Natural Language Processing Laboratory, CUET. He is also serving as the Chair of IEEE Bangladesh Section. He has published more than 155 publications in several international journals, books, and conferences. His research interests include computer vision, human–computer interaction, and natural language processing. He is a fellow of the Institute of Engineers, Bangladesh, and a Senior Member of IEEE RAS, IEEE SPS, IEEE CS, and IEEE WIE Affinity Group. He has served as the Award Coordinator, from 2016 to 2017, a Conference Coordinator, from 2017 to 2018, and a Vice-Chair Technical, from 2019 to 2021 for IEEE Bangladesh Section. He has also served as the TPC Chair for IEEE r10 HTC 2017, ECCE 2019, and ACMI 2021, the TPC Co-Chair for ICISET 2018/22, IEEE TenSymp 2020, IEEE WIECON-ECE 2021, and ICREST 2021, the Publication Chair for IEEE WIECON-ECE 2018/2019 and IEEE TenSymp 2020, and a TPC member for several international conferences.

ENAMUL HOQUE (Member, IEEE) received the Ph.D. degree in computer science from the University of British Columbia. He was a Postdoctoral Fellow in computer science at Stanford University. He is an Assistant Professor at York University, Canada, where he directs the Intelligent Visualization Laboratory. His research has been funded by the Natural Sciences and Engineering Research Council of Canada, the National Research Council Canada, and the Canada Foundation for Innovation. His research interests include natural language processing, information visualization, and human–computer interaction.

MD. SAIFUL ISLAM (Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering from the University of Ulsan, South Korea. He is an Assistant Professor with the Department of Electronics and Telecommunication Engineering, Chittagong University of Engineering and Technology (CUET), Bangladesh. His research interests include artificial intelligence, signal, and image processing.